Constructing Variables that Support Causal Inference

Stephen E. Fancsali

May 2013

Acknowledgements:

After roughly seven years at Carnegie Mellon University, I owe thanks to many individuals.

David Danks has been an ideal adviser and dissertation committee co-chair. Despite a busy schedule and many advisees, he always found time to read drafts and provide careful feedback. His understanding ear and sage advice (both personal and professional) I value immensely. Thank you.

Richard Scheines, my other co-chair, provided the intellectual inspiration for much of this work. His long-standing interests in applied causal inference and matters of variable construction and definition in the social sciences led him to ask challenging questions, seriously improving this work.

My committee members have provided guidance and inspiration in a variety of ways. Daniel Neill, some time ago, stimulated my interest in developing interpretable models for predictive and causal inference that might prove useful for real-world policymakers. Partha Saha and Steven Ritter, in different settings and over the course of several years, have demonstrated that the type of questions we ask in this work are important in real education settings. Also, they have provided various forms of assistance as well as valuable feedback about causal models of educational data featured in the case study of this work.

This work was funded in part by Apollo Group, Inc. and Carnegie Learning, Inc. Any positions or opinions expressed in this work are my own and do not reflect those of either organization.

I owe special thanks to Joseph Ramsey for programming advice and computational/technical assistance. Several other people provided patient assistance in collecting, interpreting, and/or analyzing data: Ryan S.J.d. Baker, Sujith Gowda, Tristan Nixon, and Mike Sharkey.

Other faculty in the Philosophy Department have also influenced my thinking and this work in a variety of ways. Clark Glymour, Mara Harrell, Teddy Seidenfeld, and Peter Spirtes have provided valuable advice and other insights at various points in the development of this dissertation. Errors, failures of imagination, and other shortcomings of this work are attributable to me alone.

The staff members of the Philosophy Department have always provided great conversation and been a vital resource for accomplishing practical tasks necessary to progress through graduate school. Thank you: Mauren Antkowski, Ro Commisso, Jackie DeFazio, and Jan Puhl.

Several (current and former) fellow graduate students deserve special thanks for their assistance with this work or their friendship over the years (in many cases both): Ed Dean, Frederick Eberhardt, David Grey, Ben Jantzen, Conor Mayo-Wilson, Ruth Poproski, Rob Tillman, and Sarah Wellen. Ed, Conor, and Ruth deserve special praise for putting up with me as an office-mate for so many years. I was lucky to have found them in my office in Baker Hall on my first day at Carnegie Mellon.

Many friends have been sources of support while writing this dissertation, especially Derrick Monahan, Will Mulligan, Jessica Scott, Stephanie Selover, Luke Stamper, and Shawn Standefer. Thank you.

I am grateful to Dr. Jim Petraglia, for having sympathy for a poor student and providing generous and excellent dental care in several times of serious need during my graduate school career.

Finally, my mother (Kim), father (Paul), and sister (Susan) are ever-present sources of unconditional love and support. I will never be able to thank them enough.

For my brother, Vincent. He left us far too early.

Contents

Chapter 1

Complex, "Raw" Data and Causal Discovery

1.1 Introduction

Scientists and policy makers frequently require causal knowledge to solve the problems that they face[1]. Economists, rather than merely predict inflation rates, seek to know what effect a change in monetary policy will have on inflation. Psychiatrists, rather than merely predict that a new patient suffers from depression, require knowledge of what (and how) pharmaceuticals effectively rid patients of depression. While educators may find it useful to know how likely it is that a student will fail a course, they are likely to be more interested in interventions that will keep a student from failing. Merely predictive questions can be answered by numerous statistical and algorithmic methods. To make predictions about systems we are not passively observing but rather changing by intervening, we must move beyond predictive inference (e.g., relying on "correlations that do not imply causation") to learn causal structure and provide causal explanations of scientific phenomena.

Typically, scientists infer causal structure by performing experiments, e.g., randomized controlled trials (RCTs). However, RCTs and other experiments are often impossible, unethical, cost-prohibitive, or for some other reason cannot be carried out. Work over the past 20+ years in philosophy, computer science, and statistics has been devoted to reliably discovering causal structure from observed, non-experimental data. Such work generally assumes appropriate variables have been defined. For example, to learn

---

[1] Portions of this chapter appear in Fancsali (2011a) and Fancsali (2012).

the causes of depression, we presumably have appropriate measurements of the extent to which a subject is depressed. Without appropriate measured or observed variables (i.e., without variables at an appropriate level or unit of analysis), we must first construct them.

Examples are legion in both the sciences and everyday life in which data do not "cleanly" manifest as variables that support or underwrite causal explanations and inferences. Such data are thus initially un-suited to the goal of discovering causal relationships. Consider, for example, "raw" sense-data with which we are bombarded at every moment. Complex cognitive processes allow us to quickly process multitudes of visual, auditory, and tactile sense data such that when instructed to sit down, for example, we can understand words and the command that has been uttered, distinguish an object from its surroundings that corresponds to our concept of a chair, and take action to sit down. Very roughly, our brains have managed to take fine-grained (sense) data and transform them appropriately (into corresponding concepts, natural kinds, objects in our visual field, etc.) so that we have a modicum of control over our environment and can take appropriate actions. These transformations provide an efficient representation of the world around us, and, more importantly, these representations allow us to causally navigate and control the world around us. So it goes with the natural and social sciences; we seek to appropriately transform fine-grained, "raw" data into variables from which we can learn causal information about (and acquire causal control over) systems of interest.

1.1.1 Examples

Problems posed by fine-grained, "raw" data for predictive and causal inference vary in degree and scope. We consider several examples, all of which share the feature that data as collected are at an intuitively "wrong" level/unit of analysis (or perhaps aggregation) for important scientific, causal questions:

- Therapists and social scientists often assess levels of subjects' depression (or some other mental health/psychiatric condition) with diagnostic/survey instruments composed of multiple presumed measures or indicators (effects) of the underlying condition. Investigators must decide which measured indicators are appropriate to include in a variable constructed from a survey instrument and what function should be applied to the measures. In practice, social scientists frequently construct scales as the sum of positive, integer valued measures of underlying "latent" phenomena, but other transformations could be appropriate.

- Economists model macroeconomic phenomena like inflation by constructing indices (e.g., various forms of the Consumer Price Index) from prices of "baskets of goods" assessed over disparate geographic regions. To construct such indices, one must choose which goods and which geographic regions ought to be considered (i.e., the level or scope of the index), as well as the aggregate function or transformation that should be applied to these prices.
- Cognitive neuroscientists construct variables to represent regions of interest (ROIs) in the brain from thousands of indirect measurements of neuronal activity. Which voxels (i.e., "volumetric pixels" representing small regions of neuronal activity) should be included in ROIs? What are the relevant characteristics of these voxels and how do we aggregate over them to learn about interactions among ROIs?
- Education researchers frequently work from fine-grained logs generated by computer-based courseware; they construct variables to represent aspects of student behavior to predict and explain learning outcomes. Such logs track many characteristics and features of thousands of individual student interactions with courseware, making problems of aggregation level and scope as well as appropriate functions with which to construct aggregate variables difficult (even compared to the examples above in which investigators at least generally have some notion of appropriate aggregation functions).

The algorithmic approach we propose is largely inspired by problems like those from courseware logs (the last example above); given the complexity of these logs and the nature of fine-grained data they track, we have little domain knowledge to expect that extant methods have constructed aggregate/student-level variables optimally, especially for purposes of causal explanation. Thus, the significant focus of this work is to develop methods suitable for constructing variables that support causal inference (and to adjudicate success in doing so) from data sources like education courseware logs.

1.1.2 Proxies vs. Aggregates

These examples illustrate that sometimes we seek variables to represent causally relevant, underlying phenomena that are, in principle, not directly observable, but for which we only have indicators (i.e., presumed effects of "latent" phenomena, like depression). From these indicators (e.g., symptoms), we hope to construct "proxies" that support causal inferences by obtaining in the same inferred causal relationships as the underlying phenomena they represent (e.g., we infer that a proxy for depression roughly stands in the same causal relationships we would infer if depression were directly observable). Importantly, proxies so constructed do not have any necessary semantic relationship to the underlying phenomenon of interest. A well-constructed proxy establishes the extent of a patient's depression from her symptoms and will be a good "stand-in" for the unobservable phenomenon for statistical and causal inference, but we do not think that the unobservable phenomenon is identical to the proxy. When a psychiatrist talks about a patient's depression, she is presumably not talking about the proxy variable she constructed but rather the actual underlying phenomenon.

In other cases, causally relevant variables and phenomena are aggregates of directly (or indirectly) observable, but fine-grained measures (e.g., aggregate brain activity in a ROI that could play an explanatory role for thought). In contrast to situations in which we seek variables as proxies for latent phenomena, in many cases in which we seek aggregate constructed variables, there is a necessary semantic link between constructed aggregates and the phenomena we seek to represent. For example, activity at an ROI in the brain really just is the aggregate of neuronal activity (even if indirectly measured) over a particular region, so there is a clear semantic link between variables constructed and the phenomenon we seek to represent.

In cases like education courseware logs, we likely face both challenges. Further, sometimes distinctions between these different situations are approximate at best; despite differences in the semantic content of constructed variables (latent variables vs. constructed aggregates), we can methodologically proceed in a similar fashion across situations. We explore these ideas further in Chapter 2.

1.1.3 (Fine-grained) Predictions vs. (Aggregate) Causal Explanations

To predict the value or distribution of an outcome or target variable from passive observation, any set of variables constructed from "raw" data that underwrite accurate, reliable predictive models generally suffice. Consider, for example, a credit card company that collects "raw" data about its consumers' transactions (e.g., location, amount, type of purchase, etc.) and puts them to a variety of uses. First, consider predicting which future transactions are likely to be fraudulent. Any set of variables constructed from characteristics of transactions that helps to successfully identify fraud will be sufficient to help the company decline appropriate transactions. That a single transaction is at a distant geographic location compared to the last transaction, for example, might indicate that a card credit has been stolen. Second, at a higher-level of aggregation, consider predicting which consumers are likely to default on their debt so that appropriate steps might be taken to restructure their debt or provide other advice. Perhaps higher variability in transaction type (i.e., buying many different kinds of products on credit) is indicative of a consumers risk for default. Again, any set of aggregate, consumer-level variables constructed from transaction characteristics that helps to successfully identify consumers at risk of default provides important insight to the company.

However, constructed variables suited to the task of prediction need not be well-suited to causal discovery. The same credit card company may, for example, seek to understand the impact a change in interest rates they offer will have on aggregate consumer (credit card) spending. This task is different than the task of passively determining whether one particular transaction is likely to be fraudulent or whether a particular consumer is likely to default. Aggregate causal explanations of consumer behavior are required to predict the effects of such an intervention (e.g., to predict the distribution of aggregate consumer spending following an intervention on a system, like a credit card company changing the interest rate it charges consumers). Given different goals, we must figure out not only what aspects of "raw" data are important (e.g., transaction amount vs. type, etc.) but also what transformations or functions (e.g., average, variance, etc.) must be applied to construct variables. For most interesting applications, optimistically, there are hundreds or thousands of possibilities.

Much research is devoted to dimensionality reduction with feature selection methods (from sets of variables presumably at an appropriate level of analysis or description) and feature extraction (or "feature engineering") from "raw" data for predictive purposes. While some attention has been paid to dimensionality reduction via feature selection (again, from large sets of appropriately-aggregated variables) for purposes of causal inference, less research has been directed at dimensionality reduction involving variable construction (i.e., feature extraction or "feature engineering") for explicitly causal purposes.

1.1.4 Proposal: Search for Variables that Support Causal Inference

We propose methods for variable construction that simultaneously search for constructed variables from "raw" data (i.e., data initially unsuited to causal explanation) and causal structure over these constructed variables. Our objective is to discover variables that support inferences about qualitative causal structure and strength of causation. We introduce several notions of "causal predictability" to capture predictive access (i.e., notions of qualitative and quantitative "strength" of causation) to a target from its inferred causes. Thus, the goal of search for constructed variables is to maximize causal predictability.

While we focus on the causal desideratum of constructing variables that contribute to learning causal structure and strength, there are other possible (and important) desiderata (e.g., constraints on interpretability, possible interventions, etc.) for constructing variables useful for causal purposes, some of which might be integrated into search for constructed variables. In the case study of Chapter 5, we attempt to delimit a search space of interpretable constructed variables, but further work is necessary to determine how background and expert knowledge can provide constraints for causally interpretable constructed variables that are targets of conceivable interventions.

Notably, any proposed notion of causal predictability should incorporate the idea that causal structure is, in general, underdetermined by observational data. Consider, for example, the causal graph of Figure 1.1. Nodes represent variables, and directed edges represent direct causal relationships. Without background knowledge and under generally reasonable assumptions, we can infer, from data sampled from this system, that X and Y are causes of W (we explain how in §1.2), so we find ourselves in a fortunate position to maximize causal predictability if W is our target variable. However, if our target variable is Z, we face greater uncertainty with respect to causal structure, as observational data alone will not generally allow us to infer whether Y → Z or Z → Y; that is, from non-experimental data, we cannot distinguish in which direction causal influence "flows." We develop an approach to assessing causal predictability that is sensitive to this type (and other types) of uncertainty.

Further, as problems posed by different types of data vary in degree and scope, so does the space of constructed variables over which we search. In some cases, we have a great deal of background and domain knowledge to delineate and place constraints on a search space of constructed variables.

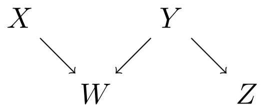

Figure 1.1: Example causal structure: with observational data sampled from this structure alone we can infer that X and Y are causes of W, but we cannot determine direction of causal influence between Y and Z. We explain why in §1.2.

Consider the example above of psychometric "scales" used to produce proxies of underlying latent phenomena like depression. Ideally, survey and other instruments have been validated (i.e., they measure intended underlying phenomena) and are composed of relatively few items, so there is a relatively small number of possible constructed variables/scales to consider to find the scale(s) that maximize causal predictability (assuming domain knowledge provides the function/transformation we should apply to the survey items). In cases like these, scientists seem to have (at least a modicum of) success in constructing variables they put to a variety of purposes. However, it is far from clear exactly how it is that scientists are successful (e.g., whether via general principles, heuristics, domain knowledge, intuition, or some combinations of all of these, etc.).

In other cases like education courseware logs, we have less background/domain knowledge to delineate the search space for constructed variables, so we may need to consider a wider range of transformations of different aspects of underlying "raw" data. In general, computational approaches are necessary because we will face a large search space (on the order of at least hundreds or thousands of possible constructions).

We investigate ways to construct causally-interpretable, scientifically meaningful variables in situations in which causal knowledge is required to drive strategy and interventions from measurements or observations that are inappropriate for causal purposes (e.g., because of technological constraints, because we do not know how best to, or simply did not, measure phenomena at a level of analysis best-suited for causal explanation). To do so, we provide means by which to assess causal predictability and propose algorithmic schemata to discover sets of constructed variables from "raw" data. We begin by reviewing the framework of probabilistic, graphical causal models we use in our investigation.

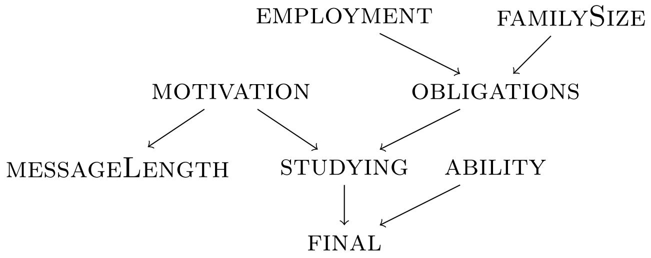

Figure 1.2: Directed acyclic graph representation of hypothetical causal relationships among attributes of students enrolled in an online course.

1.2 Graphical Causal Models

1.2.1 Directed Acyclic Graphs and Basic Formalism

Aforementioned research on causal discovery from observational datasets centers on the causal interpretation of probabilistic graphical models, especially directed graphs without cycles[2] (direct acyclic graphs or DAGs) and associated probability distributions, called Bayesian networks (Spirtes, et al. 2000; Pearl 2009). Within this formalism, variables are represented by nodes in a graph with edges between nodes representing causal relationships between variables. Consider the DAG in Figure 1.2, modeling qualitative, hypothetical causal relationships among attributes of students in a hypothetical online course.[3]

We model relationships between (hypothetical) measures of EMPLOYMENT, size of family (FAMILYSIZE), time obligations not related to a student's education (OBLIGATIONS), time spent STUDYING, MOTIVATION, length of messages in an online discussion forum (MESSAGELENGTH), academic ABILITY, and final exam scores in a course (FINAL). In this hypothetical model, student final exam performance has two direct causes: student ABILITY and STUDYING. Further, the model represents relationships among the determinants of a student's STUDYING behavior.

---

[2] Feedback cycles (in time) can be modeled within the Bayesian network framework by, for example, deploying variables indexed by time. Some literature (Richardson 1996a, 1996b; Pearl and Dechter 1996, Lacerda, et al. 2008) focuses on the discovery of cyclic graphical models, though work on this topic is less developed than the Bayesian network/DAG formalism for causal discovery.

[3] Several attributes might be particularly salient for non-traditional students to whom online education and degree programs are appealing. Such an example is motivated by the nature of data in the case study in Chapter 5.

1.2.2 The Causal Markov Condition and Causal Faithfulness Condition

We assume two conditions to connect causal structure encoded by a DAG and an associated probability distribution to a set of probabilistic (conditional) independence relations: the Causal Markov Condition and the Causal Faithfulness Condition (Spirtes, et al. 2000).[4] The Causal Markov Condition asserts[5] that, assuming there are no unmeasured common causes of variables we consider, a variable is probabilistically independent of its non-descendants (non-effects) conditional on its direct causes. The Causal Faithfulness Condition provides that all observed probabilistic (in)dependence relations (i.e., those that obtain in the distribution) are entailed by the Causal Markov Condition applied to the DAG. Conditional independence between variables, for example, does not occur by accident (e.g., via "canceling out" settings of parameters), but only because of the lack of a (direct) causal relationship.

To illustrate how a violation of the Causal Faithfulness Condition might occur, consider a slight modification to hypothetical causal relations among three variables from Figure 1.2 in Figure 1.3, in which the association represented by each arrow is positive or negative. Suppose we posit that increased FAMILYSIZE has a negative impact on EMPLOYMENT; the student is likely to work less as the size of his or her family increases but that EMPLOYMENT and FAMILYSIZE both contribute to increased non-educational time OBLIGATIONS. The negative effect of FAMILYSIZE on EMPLOYMENT combined with the positive effect of EMPLOYMENT on OBLIGATIONS, given appropriate (perhaps unlikely) parameter values (representing strength of causal relations), may exactly "cancel out" the positive effect of FAMILYSIZE on OBLIGATIONS. Such "canceling out" parameter values could lead us to believe that FAMILYSIZE and OBLIGATIONS are independent, despite the fact that there is a direct causal relationship between the two. This hypothetical judgment of independence despite a direct causal relation is a violation of the Causal Faithfulness Condition.

---

[4] There is substantial philosophical literature about the Causal Markov Condition and the Causal Faithfulness Condition (e.g., Sober 2001; Cartwright 1999, 2006; Hoover 2003; Steel 2005; Hausman and Woodward 1999, 2004). We omit controversy arising in that literature, as the assumptions are standard in the causal learning framework deployed here (and to which this work contributes).

[5] assuming it is possible to represent the underlying causal structure (or data generating process) as a directed acyclic graph

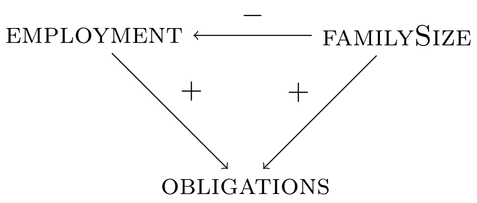

Figure 1.3: Hypothetical illustration of causal relationships that could lead to a violation of the Causal Faithfulness Condition

1.2.3 Constraint-Based Structure Search: Causal Sufficiency vs. Causal Insufficiency

While a causal Bayesian network implies conditional independencies, this "graphs → independencies" mapping is many-one: for any particular graph G, there will typically be other graphs, though not all graphs, that predict or entail the same (conditional) independencies, and so are observationally indistinguishable from G. To illustrate: consider the maxim that "correlation does not imply causation." If verbosity in online message forums (messageLength) and studying are correlated, this can be explained by messageLength → studying, messageLength ← studying, or studying and messageLength sharing a common cause (as they do in Figure 1.2, motivation), or a combination of these explanations. Thus, multiple graphs can imply the same observed correlations and/or independencies.

We can use observational data to learn a set of (observationally indistinguishable) causal structures: namely, exactly the set of possibilities that could have produced the observed pattern of independencies in the data. Assuming there are no unmeasured common causes (hereafter, "causal sufficiency"), causal Bayesian network structure learning algorithms, e.g., the PC algorithm (Spirtes, et al. 2000), under suitable assumptions, will identify the set of observationally indistinguishable graphs containing the correct DAG in the large sample limit.

Two rough intuitions illustrate basic principles of search for causal structure from conditional independence relations. The first concerns so-called "screening off" relations whereby, to consider a simple three variable example, two variables, e.g., messageLength and studying, are correlated but are independent conditional on a third variable, motivation; this conditional independence tells us that messageLength and studying are not directly causally related. Assuming causal sufficiency for message-

Length, studying, and motivation, this conditional independence claim is explained by three possible causal structures: messageLength → motivation → studying; messageLength ← motivation ← studying; or messageLength ← motivation → studying. Lacking background knowledge, these three DAGs are indistinguishable from observational data. However, if we assume student motivation to be inherent or at least temporally prior to a student's enrollment in a program and behavior in a course, then we can infer that motivation is a common cause of messageLength and studying (i.e., messageLength ← motivation → studying).

The second intuition has us consider two independent variables that share a common effect. Suppose that a student's level of motivation and non-educational, time-consuming obligations are independent. We expect that each of these student attributes share a common effect in the amount of time a student devotes to study (studying). Unconditionally an instructor cannot infer anything about a student's motivation level from the knowledge that a student has many time-consuming obligations outside of the course in which they are enrolled; the instructor, however, can make inferences about a student's motivation when the student (honestly) reports to the instructor how much they study. If we know that a student is studying a lot while juggling many obligations, we infer something about the student's motivation level, namely that it is high. We can similarly infer from a student's report that they are highly motivated and yet are not studying as much as they would like that they are likely dealing with many outside obligations. In both cases, when we condition on a common effect two otherwise independent variables now provide information about each other.

In the DAG of Figure 1.2, this appears as what is called an "unshielded collider," where directed edges from motivation and obligations meet at studying (motivation → studying ← obligations; the "unshielded" modifier stems from the fact that motivation and obligations are not adjacent). Assuming (however unlikely) that we omit no common causes, there are no other graphical structures that explain this constraint on conditional independencies (or dependencies). That is, we can orient "unshielded colliders" when such circumstances arise as we search over conditional independence relations that obtain in a distribution or dataset.

A constraint-based algorithm for causal structure search, assuming causal sufficiency, like PC systematizes search over (sets of) variables and conditional independence relations judged to obtain among them. PC searches over graphical objects called patterns (Verma and Pearl 1990) that can contain undirected and directed edges (and represent sets of DAGs). An undirected edge X - Y in a pattern indicates that either X → Y or X ← Y.

The algorithm proceeds in two "phases": beginning with a complete undirected graph, in the "adjacency" phase PC iterates over pairs of variables in the graph, seeking a set of variables (of which neither in the pair under consideration is a member) conditional upon which the pair are judged independent. If such a set is found, the undirected edge between the two variables is eliminated from the graph.

The second phase of search is the "orientation" phase in which, first, unshielded colliders entailed by the data and inferred adjacencies are oriented. Also, since we assume that graphs are acyclic, we can often orient edges that oriented the opposite way would create a directed cycle. Further, having oriented unshielded colliders, we can often orient edges in such ways that avoid creating unshielded colliders not entailed by the data. The full set of orientation rules for patterns (and thus for PC) is provided by Meek (1995). In addition, one can also provide background knowledge (e.g., temporal ordering of variables) that the algorithm can use to help orient edges. The pattern resulting from PC search thus captures causal structure uncertainty by representing the set of observationally indistinguishable DAGs compatible with conditional independence constraints present in a distribution/data set.

So far we have been considering situations in which there are no unobserved (i.e., latent) common causes of measured variables. However, in most practical (social) science and policy applications, it is unlikely that we will deal with causally sufficient sets of measured variables.

We return to the example of hypothetical causal relationships among attributes/variables measured over students enrolled in an online course from Figure 1.2. Suppose, contrary to the graph of Figure 1.2, that MOTIVATION is an unmeasured, "latent" variable and that EMPLOYMENT and OBLIGATIONS share an unmeasured common cause in addition to their direct causal connection, the level of responsibility in a student's job (or possibly how "responsible" the student is in general). A DAG representation of this scenario (with latent variables *italicized*) is provided in Figure 1.4.

To reason about causally insufficient systems (i.e., those with unmeasured common causes of measured variables), we use graphical objects called Partial Ancestral Graphs (PAGs) (Richardson 1996a, 1996b; Spirtes, et al. 2000). To represent further causal structure uncertainty due to causal insufficiency, PAGs can contain one new type of edge mark, an oval (o) (in addition to "tail" and "arrow" edge marks found on edges in patterns and DAGs). The FCI algorithm (Spirtes, et al. 2000; Zhang 2008b) infers a PAG that represents the equivalence class of causal graphs compatible with a set of conditional independence relations among measured variables allowing for

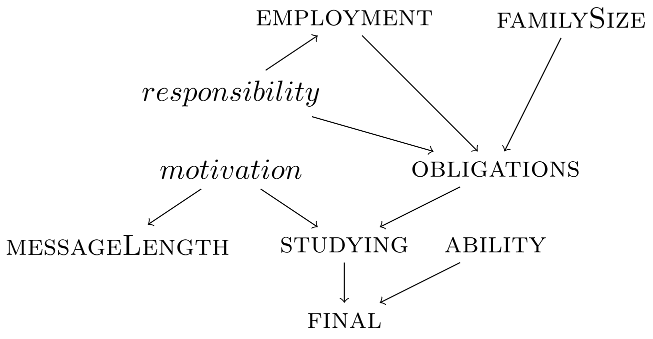

Figure 1.4: DAG representation of hypothetical causal relationships among attributes of students enrolled in an online course, including two unmeasured or "latent" variables (italicized)

causal insufficiency.

Edges between two variables in a PAG are interpreted in the following way:

1. X o→ Y: X is either a cause of Y, or X and Y share a latent common cause in every graph in the equivalence class; in no graph is Y an ancestor (i.e., cause) of X.
2. X o−o Y: In every graph in the equivalence class, (1) X is a cause of Y, (2) Y is a cause of X, (3) X and Y share a latent common cause, or (4) some combination of (1) and (3) or (2) and (3).
3. X ↔ Y: There is a latent common cause of X and Y in every graph in the equivalence class. Neither X is an ancestor of Y, nor Y an ancestor of X.
4. X → Y: X is an ancestor (i.e., cause) of Y in every graph in the equivalence class.[6]

The last type of edge indicates that it is sometimes possible, even with latent confounders, to make unambiguous, positive causal inferences. We assume there is no selection bias throughout this work, and this assumption excludes X − Y and X o− Y as edge possibilities in a PAG.

---

[6] Importantly, according to the ancestral interpretation of PAG edges, X is not necessarily a direct cause of Y given this edge in a PAG. X may only be an indirect cause of Y despite their adjacency in the PAG.

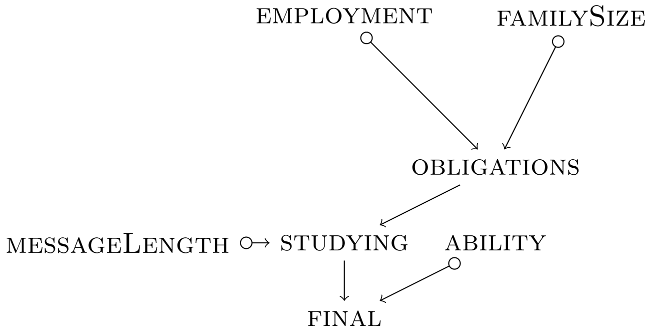

Figure 1.5: PAG inferred by FCI algorithm applied to conditional independence relationships entailed by hypothetical causal graph in Figure 1.4.

In broad outline, FCI is nearly identical to PC, with both an adjacency phase and an orientation phase. Some orientation rules are similar to those used in PC (e.g., a rule that has us avoid the creation of directed cycles), but others are different.[7] We provide a step-by-step trace of the FCI algorithm, including several of its orientation rules, for a particular example, in the Appendix (§7.2).[8] Such an example illustrates patterns of conditional independence relationships that might obtain among measured variables to allow, for example, unambiguous edge orientations indicating a causal relationship between two variables.

Figure 1.5 provides the PAG inferred from conditional independence relationships among measured variables of the example in Figure 1.4. We see that causal relationships between obligations and studying and studying and final are unambiguously inferred in this case. Whether, for example, EMPLOYMENT is a cause of OBLIGATIONS (or MESSAGELENGTH a cause of STUDYING, etc.) or they share a common cause (or both) cannot be resolved from observational data alone. We see that relaxing the assumption of causal sufficiency introduces more uncertainty about causal structure.

1.2.4 Score−Based and Other Search Procedures

Constraint−based algorithms like PC and FCI infer causal structure from conditional independencies among a set of measured variables (with linear

---

[7] Zhang (2008b) provides a complete set of orientation rules for FCI.

[8] Scheines (2005) also provides clear exposition of parts of the FCI algorithm and the similarity of causal inference from observational data using algorithms like FCI and one traditional practice in (social) science using instrumental variables.

relationships and multi−variate Normal distributions or discrete, multi−nomial distributions). Rather than explicitly rely on conditional independence constraints, GES (Greedy Equivalence Search) (Meek 1997; Chickering 2002) is a score−based method to infer a pattern from data, assuming causal sufficiency. Starting with the disconnected pattern, in the "forward" phase of GES, single edges are added to the pattern until the BIC score of the resulting model can no longer be improved. At this point, single edges are removed from the pattern, in the "backward" phase, until the BIC score can no longer be improved.

Recent work seeks to relax distributional and linearity assumptions. If, for example, variables are linearly related with non−Gaussian distributions, other conditions, rather than the presence of patterns of conditional independence relationships, allow for unambiguous orientations and causal inferences. The LiNGAM algorithm (Shimizu, et al. 2005, 2006), for example, infers DAGs (as we are left with less structure uncertainty than in linear, Gaussian cases) in causally sufficient systems in such cases based on Independent Components Analysis (Hyvärinen 2001). Subsequent research extends these models, providing a score−based method (Hoyer and Hyttinen 2009), methods to deal with latent variables (Hoyer, et al. 2006, 2008), cyclic models (Lacerda, et al. 2008), non−linear models with additive noise (Hoyer, et al. 2009), and using experiments (Eberhardt 2007) to learn linear cyclic models with latent variables (Hyttinen, et al. 2012). Alas, this outline of contemporary algorithmic research is far from exhaustive, but provides an inkling as to state−of−the−art work in the field.

1.3 Structure of the Dissertation

Our strategy is to develop semi-automated search procedures for constructed variables from fine-grained data that support causal inference(s) about particular, fixed target variables of interest. After we consider several in depth examples of how constructed variables can support causal inference in Chapter 2, we review literature on natural kinds and argue that our approach is aligned with (perhaps an extension of) a contemporary view of natural kinds as property clusters that focuses on the importance of causal mechanisms in the delineation of natural kinds and the role of natural kinds in causal explanation. In Chapter 3, we illustrate novel quantitative metrics to assess causal predictability afforded by sets of (constructed) variables. Chapter 4 details results of simulation studies investigating the performance of these metrics. In Chapter 5, we deploy these methods on data from an intelligent tutoring system, Carnegie Learning's Algebra Cognitive Tutor, to develop causal models of "gaming the system," off-task, and other behaviors and their influence on aggregate student learning. Chapter 6 outlines future research.

Chapter 2

Constructing Variables in Practice

2.1 Introduction: Three Situations

Fundamental questions arise about the nature of causally interpretable variables we seek to construct from fine-grained, "raw" data. We can distinguish variable construction problems along at least two dimensions: the degree to which constructed aggregate variables have (or do not have) a necessary semantic link to features of interest and the degree to which "raw," measured data provide direct access to features of interest. Along these two dimensions, we provide detailed examples for three situations that obtain in scientific practice:

- Situation #1: Data are collected as variables that are indirect measures or indicators of unobservable features of interest. In such cases, we seek to construct variables that serve as proxies for unobservable phenomena that have no necessary semantic connection to underlying phenomena. While the investigator does not have direct access (generally, in principle) to the phenomenon for which she wishes to construct a proxy (e.g., depression or relationship satisfaction), she does have relatively direct access to phenomena out of which she constructs a proxy (e.g., survey responses about frequency of suicidal ideation or verbal disputes in a relationship).
- Situation #2: Data are collected as variables that directly measure features of interest but are too "fine-grained" or "low level." Despite

having direct access to observable phenomena, we must construct variables to learn causal relationships of interest (e.g., sea-surface temperature measurements from buoys intended to construct aggregate ocean/climate indices in global climate modeling). We expect such aggregate constructed variables to have necessary semantic links to the phenomena they are constructed to represent.

- Situation #3: Data are collected as variables that are indirect measures of features of scientific interest and "fine-grained" (e.g., indirect, voxel−level measures of physiological activity, serving as indirect measures of neuronal activity, used to construct variables representing brain regions of interest [ROIs] in functional Magnetic Resonance Imaging [fMRI]). Despite having only indirect measurements of features of interest (e.g., neuronal activity), we assume we can still meaningfully impose a semantic connection between aggregate constructed variables (e.g., those constructed from measures of appropriate physiological activity) and scientific phenomena they are intended to capture (e.g., brain ROIs).

These situations are neither mutually exclusive nor exhaustive. They are meant to be roughly illustrative, but, in practice, scientists often find themselves in all three situations. Other situations might arise, and there are also likely to be "borderline" cases.

2.1.1 Aggregate Variable Construction vs. Latent Variable Modeling

Situation #1 describes data as collected frequently in studies of "latent" variables. For example, data may be collected about subjects' levels of various indicators of depression, and some mathematical function of these indicators may be calculated to serve as a new, proxy variable for depression. When a scientist makes recommendations about treatment for depression, she is not referring to the depression proxy variable, but rather the underlying depression phenomenon (i.e., the latent variable). There are no necessary semantic relationships between a constructed proxy and a latent variable; that is, we do not identify a latent variable by such a proxy. The scientist hopes that the proxy is a good "stand−in" for the latent, having appropriate statistical connections to the latent variable. Specifically, our focus is that a constructed proxy stands in approximately the same causal relationships as the intended latent variable. We explicate this construction problem for latent variable modeling and propose one solution in greater detail later in this chapter, following work of Fancsali (2008).

Situations #2 and #3 present problems of aggregate variable construction. In these cases, we construct aggregate variables from fine-grained measured data. One critical contrast between aggregate variable construction and modeling latent phenomena with proxy variables is that there is a necessary semantic connection between constructed aggregates and the target phenomena of interest; we identify phenomena under consideration by aggregate constructed variables, but we do not identify latent phenomena by constructed proxies.

Consider an example of Situation #2 from climate phenomena. Local observations of atmospheric pressure can lead a meteorologist to posit that a "high-pressure system" is located over a particular geographic area. Infinitely many possible instantiations of local readings might lead to a similar supposition of a high-pressure system over a particular region; the feature of interest for (many) purposes of explanation and inference, the aggregate "high-pressure system," supervenes on local measurements of atmospheric pressure. We identify the high-pressure system by an aggregate function of a set of relatively fine-grained measurements. The aggregate is not just a proxy for the high-pressure system; rather, what we mean semantically by the high-pressure system just is that aggregate. Of primary concern in this work are scientific and policy situations frequently arising when data manifest as in Situations #2 and #3.

2.1.2 Direct vs. Indirect Measurements and Aggregation

The distinction between Situation #2 and Situations #3 arises out of the measurement of fine-grained, "raw" data collected. In some cases, like "high-pressure systems" and other climate examples we consider, we have direct, fine−grained measurements or observations. In other cases, we have only indirect ways of measuring and observing fine−grained, low−level phenomena.

Consider the case of brain imaging studies in cognitive neuroscience. Brain imaging (e.g., fMRI) data provide indirect access to neuronal activity (i.e., to correlates of neuronal/brain activity) at small regions of the brain (i.e., volumetric pixels or voxels) in the form of measurements of Blood Oxygenation Level Dependent (BOLD) response (Lazar 2008, 14). That is, BOLD response is generally not an important scientific feature of interest, but it provides means by which to infer (and thereby indirectly measure) neuronal activity. However, the level at which we frequently find such activity interesting is that of macroscopic regions of interest (ROIs) in the brain, for example, to study causal relationships and neural connectivity between different areas of the brain. In such studies, investigators take the ROI constructed variable to identify (at least roughly) the anatomical brain ROI; thus, such constructed variables have an important semantic connection to the anatomy of the brain, despite only indirect access to fine-grained, neuronal activity provided by BOLD response data. We return to this example later in this chapter.

Whether measurements are "direct" or "indirect" is in many domains a question of degree rather than kind. The distinction (if one can or need be clearly drawn) may be important in some domains for the interpretability of aggregate constructed variables, but we treat both cases roughly the same methodologically in what follows.

2.1.3 Aggregation Level

To construct aggregate variables, we must decide a suitable "level" of aggregation. That is, we must decide over which of available "raw" or fine-grained variables (measurements, indicators, etc.) we will aggregate to construct higher level variables. For survey data, which items will we retain in constructing proxies? Over which atmospheric pressure readings do we define the "high-pressure" system? Which voxels are relevant to a particular brain ROI?

2.1.4 Aggregation Function

Further, we must decide the appropriate mathematical function to apply to "raw" variables to capture salient scientific phenomena. Proxies constructed from survey data generally are constructed as the sum of numeric, "raw" responses to survey items. Perhaps high average atmospheric pressure over a particular geographic reason is appropriate in defining a "high-pressure" system. In other settings, perhaps variability over fine-grained, "raw" features is most important to capture salient behavior or phenomena. Still in other statistical dimensionality reduction settings, e.g., principal components analysis, more complicated functions of "raw" variables/features are deployed to construct new variables.

2.1.5 Domain Knowledge, Theory, and Heuristics for Variable Construction

In many scientific domains and application areas, questions of aggregation level and aggregation function have relatively standard answers based on background knowledge and established theory. In some cases, e.g., construction of proxy variables from survey data, background knowledge and theory are combined with heuristics to assist in variable construction. Nevertheless, in many disciplines, we lack sufficient background knowledge and theory to be confident that aggregate variables constructed are capturing features of interest well; indeed, in some cases we may have only rough knowledge of what those features might be. Thus, we seek to develop data-driven methods to help us work through problems of aggregation level and aggregation function to find constructed variables that contribute to causal models and explanations. As data storage becomes less expensive and "big data" scientific and policy applications proliferate, these problems will only become more widespread and therefore important.

In the following three sections, we explore examples of real-world scientific applications with varying degrees of background knowledge and domain theory to inform the construction of variables and the judgment of resulting causal models. In the next section, we summarize prior work due to Fancsali (2008) that studies the combination of background knowledge and a commonly used heuristic to develop proxy variables from survey data. This work provides a computational justification for the heuristic that helps to motivate work on aggregate variable construction developed later in this dissertation. In two subsequent sections, we consider applications in climate science and cognitive neuroscience in which aggregate variables are constructed and consider the quality of resulting causal models. Finally we motivate data-driven search for aggregate constructed variables from intelligent tutor data and consider some philosophical issues arising out of this work.

2.2 Psychometric Scales, Survey Data, and Latent Variable Models

Figure 2.1 illustrates a situation in which an unobserved, latent variable L has three measured indicators X1, X2, and X3, each of which is a (linear) function of L and an independent noise term.[1] Linear structural equations corresponding to the measured indicators of the underlying latent in this model are:[2]

$$X1 = \alpha L + \epsilon_1$$
$$X2 = \beta L + \epsilon_2$$
$$X3 = \gamma L + \epsilon_3$$

Suppose an investigator constructs a variable to serve as a proxy (often called a "scale") for the latent variable from measured indicators. In the case of Figure 1, the variable L_Scale would be constructed as the sum of the measured indicators, i.e.:

$$\text{L\_Scale} = X1 + X2 + X3$$

Consider a concrete example. Therapists seek to diagnose and treat depression while also investigating relationships between depression and other latent phenomena like satisfaction in a romantic relationship. The therapist cannot measure a patient's depression or relationship satisfaction directly, like she could measure a patient's height or weight. Rather, therapists measure indicators or (presumed) effects of these phenomena.

Cramer (2004), for example, studies the effects of depression, support, and conflict on romantic relationship satisfaction. One unobserved variable, "relationship satisfaction" (RS), is measured with the Relationship Assessment Scale (Hendrick 1988). The scale, as deployed by Cramer (2004), consists of seven survey items, answered on a Likert scale of 1 to 5 (greater valued responses corresponding to greater satisfaction) with respect to the subject's current romantic partner:

- s1. How well does s/he meet your needs?
- s2. In general, how satisfied are you with your relationship?

---

[1] This section largely summarizes Fancsali (2008) and is relatively autonomous from the rest of this work. We provide conditions under which multi-item "scales," constructed from measured indicators of latent variables, can serve as appropriate proxies for latent variables in graphical causal models. The reader can skip this section with little impact for understanding the rest of this work.

[2] For simplicity, we assume that each indicator is a "pure" measure of L; that is, each indicator is only a function of the underlying latent and an independent noise term. This is a common, and important, assumption for such models, but other, more complicated situations (e.g., when measured indicators are causes of others or are the effects of other latent variables) can and do occur in practice. We discuss one possible way to deal with such situations shortly.

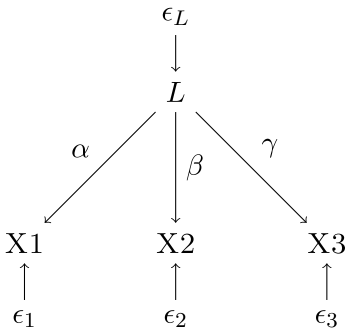

Figure 2.1: Graph depicting causal relationships among a latent variable L and its three measured indicators $X1$, $X2$, and $X3$.

- S3. How good is your relationship compared to most?
- S4. How often do you wish you hadnt got into this relationship?
- S5. To what extent has your relationship met your original expectations?
- S6. How much do you care for her/him?
- S7. How many problems are there in your relationship?

A variable RS_SCALE is constructed from these measured indicators to serve as a proxy for *RS*:

$$\text{RS\_SCALE} = \text{S1} + \text{S2} + \text{S3} + \text{S4} + \text{S5} + \text{S6} + \text{S7}$$

Under certain conditions, constructed variables like L_SCALE or RS_SCALE will have appropriate statistical and causal connections to intended latent variables (e.g., L or RS) and thereby serve as appropriate proxies for these latent variables. Specifically, such conditions are those under which variables like L_SCALE and RS_SCALE preserve conditional independence relationships into which L and RS enter, respectively. We explicate these conditions in terms of a heuristic commonly used in psychometrics and elsewhere for assessing the so−called reliability of a scale and provide a computational justification of a particular version of this heuristic. We then provide a data−driven search procedure for multi−item scales (i.e., constructed variables) based on this heuristic and look at an example with real−world data collected for the Relationship Assessment Scale.

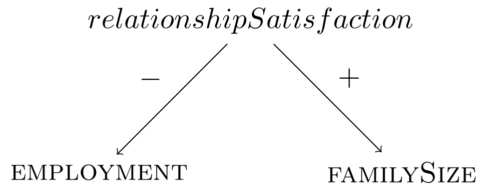

Figure 2.2: Introducing a latent common cause of employment and familySize.

2.2.1 "Screening Off" and Conditional Independence: An Example

Suppose we augment the causal relations of two variables from our online education example in Chapter 1 by including a latent common cause of employment and familySize. Consider satisfaction in a student's primary romantic relationship (*relationshipSatisfaction*) as a cause of both; suppose increased relationship satisfaction causes people to decline over−time and/or spend less time at work (decreasing hours of employment) while also tending to increase the size of a person's family. This is illustrated as Figure 2.2, with *relationshipSatisfaction* italicized to indicate that it is latent.

Suppose we have three presumed, "noisy" effects of *relationshipSatisfaction*, (X1, X2, and X3) we measure via a survey instrument. We represent this scenario graphically in Figure 2.3.

Figures 2.2 and 2.3 imply that familySize and employment are negatively associated. Noting their dependence, a sensible investigator ponders whether this association arises because they are directly causally related or because they share a common cause, suspecting that *relationshipSatisfaction* is a common cause.

Given direct access to *relationshipSatisfaction*, we would judge that employment and familySize are independent conditional on *relationshipSatisfaction*. However, we have no such direct access. The investigator must find a way to "control" for or condition on this unmeasured construct. In general, measured indicators of latent phenomena do not preserve the conditional independence relations of their unmeasured causes or parents. For example, conditioning on the three measured indicators (X1, X2, and X3) is not sufficient for us to judge that employment and fam-

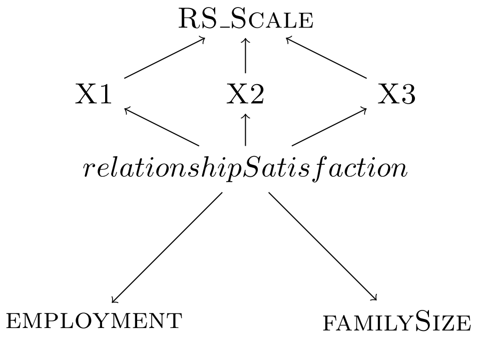

Figure 2.3: Graph introducing causal relations among *relationshipSatisfaction*, three "noisy" measured indicators (independent error−terms omitted), and a deterministically constructed scale (RS_Scale).

ilySize are conditionally independent.[3] The investigator constructs a proxy for *relationshipSatisfaction* from the measured indicators available:

$$\text{RS\_Scale} = \text{X1} + \text{X2} + \text{X3}.$$

The investigator hopes that when he conditions on his proposed proxy, RS_Scale, familySize and employment will become independent, and he will have provided evidence that there is a common cause. However, graphically it is not clear why RS_Scale should screen off employment from familySize. Just as for X1, X2, and X3, the graph does not imply such a "screening off" (i.e., conditional independence) relation. Can (and under what conditions will) the investigator succeed given a larger set of measured indicators from which to choose in constructing a scale for *relationshipSatisfaction*? How can heuristics and a data-driven procedure for variable construction help?

2.2.2 Constructed Scale Variables and "Screening Off"

A natural condition under which a variable like RS_Scale will preserve the appropriate "screening off" or conditional independence relationships is

---

[3] This follows from a straightforward application of Pearl's d-separation criteria (Pearl 1988). Else we might demonstrate this via regression on data simulated from such a model. In the regression model, we regress, for example, familySize on X1, X2, X3, and employment, and still find that employment is statistically significant, suggesting that there may be a direct causal relation.

when it is highly correlated with the underlying phenomenon it is intended to measure, here $relationshipSatisfaction$. In a computational setting with simulated data, we can estimate the degree of correlation necessary between a scale and an underlying latent for screening off relations to be preserved. Fancsali (2008) demonstrates that a very high degree of correlation is necessary for the preservation of screening off relations. Only in simulated situations in which this correlation is roughly 0.9 or greater is it likely that appropriate conditional independence relations are judged to obtain. That is, in simulations, scales have appropriate statistical and causal connections to (and thereby serve as satisfactory proxies for) latent variables when they are correlated with them at a level of 0.9 or greater.

However, outside a simulated setting we are unable to measure latent phenomena such that we could estimate the degree of correlation between a multi-item scale and a latent construct; we must resort to other means. Here, we turn to what is called the reliability of a scale. A common measure of reliability turns out suitable, given several important assumptions, for incorporation in heuristics that we can deploy for data-driven search for constructed scales. Before exploring the notion of reliability and heuristics for data-driven variable construction, we briefly explicate these (statistical and causal) assumptions.

2.2.3 Validity

Validity is fundamental if we are to treat a scale as a proxy for an underlying "true" latent variable. Items (i.e., "raw" variables like survey items) we take to be measured indicators of some underlying phenomenon must really be measures of that phenomenon and not be only measures of something else. For example, in attempting to measure a latent phenomenon like relationship satisfaction, we want measured indicators to be genuine measures of such satisfaction; while indicators may have many causes, a valid measure of relationship satisfaction counts as at least one of its causes such satisfaction. By assuming validity, we essentially assume that we are (at least roughly) measuring the things we intend to measure; raw variables we consider really are (partly) effects of the phenomena we are investigating. Thus, when we construct a scale from them we provide a genuine, albeit necessarily indirect, measure of the underlying phenomenon.

Social science methodology literature, especially in social psychology, psychometrics, and similar disciplines, provides several different ways of cashing out notions of validity. For our purposes we are not concerned with any particular type of validation methodology. We assume that practitioners deploy accepted data collection methods and well-validated instruments, necessarily depending upon background knowledge and domain theory to do so. To illustrate, when considering measured indicators of $relationshipSatisfaction$ for inclusion in a scale, we assume that a practitioner includes such indicators (e.g., in their survey instrument) because the indicators have as at least one plausible cause, $relationshipSatisfaction$.

2.2.4 Measurement Model Purity

Assuming measured items are valid, we consider causal relations holding among latent variables and measured indicators available as raw variables. The "measurement model" is the portion of a latent variable model that provides causal relationships between latent variables and their measured indicators. Contrast this with the "structural model" that specifies causal relationships among latent variables (Spirtes, et al. 2000). Social scientists frequently assume what is called "local independence," referring to the idea that members of a set of measured indicators for a latent variable become independent of each other after conditioning on the variable they measure. The source of co-variation or correlation among all of the measured indicators of a latent construct is the latent itself, not causal relations between measured indicators or measured indicators and other latent constructs.

Notably, a measured indicator we call valid for a particular construct may have additional constructs of which it is an effect. A "pure" indicator has only one cause among variables under consideration, and is not the cause of any other variables under consideration, measured or latent. We rely on this assumption so that we can assume that correlation among measured indicators of a latent variable arises because the indicators are effects of the same underlying latent, not from other potential sources of correlation. When our purposes are the discovery of causal relationships between latent variables, with currently available discovery algorithms, this assumption is necessary. Spirtes, et al. (2000) refer to this assumption as "purity" of a measurement model.

Whether the assumption of "local independence" is met or the practitioner is working with a pure measurement model might be adjudicated in several ways. Social scientists usually rely upon some combination of background knowledge about the measured indicators and the presumed underlying phenomena as well as measures of "goodness-of-fit," well known to the factor analysis and structural equation modeling communities. Different numbers of latent variables (or "factors") may be posited and appropriate measured indicators grouped together as effects of these factors, based on background knowledge, until a model that fits the data well is found.

However, simply because a certain specified model fits the data does not eliminate alternatives as plausible possibilities. Silva, et al. (2006) provide an automated search procedure called BPC (Build Pure Clusters) that searches over raw variables and, given certain assumptions, returns sets of "pure" indicators which are effects of only one latent and meet the assumptions for measurement model purity. We can assume "local independence" or measurement model purity, taken as if it were simply an axiom of latent variable modeling, or we can adopt data-driven procedures to discover sets of indicators that plausibly satisfy the assumption.

2.2.5 Reliability and Cronbach's Alpha

Assuming validity and "purity" of a set of measured indicators for a latent variable, we still need some quantitative measure of the quality (i.e., reliability) of a set of indicators that would comprise a proxy constructed variable like a scale. Here, we roughly link a measure of quality with the correlation between a scale and an underlying latent variable. Since we cannot, in principle, measure the correlation of a scale with an unobserved variable or construct, we look to methods deployed by social scientists for similar purposes and their assessment of the reliability of scales composed of multiple items.

The concern for reliability is essentially that measured indicators "respond" to change in the underlying latent they measure. That is, variation in the underlying latent leads to variation among its measured indicators and thus co-variation and correlation among its indicators. Such reliability is (in one way) manifested by the internal consistency of a set of measured items for a particular latent variable. The extent to which the indicators "respond" to change in the underlying latent is affected by two factors: (a) the strength of the causal relation between latent and indicator, quantified by a "factor loading" for each indicator, and (b) the level of "noise" or measurement error for each indicator. High levels of factor (a) and low levels of factor (b) provide us with measured indicators that when composed together as a constructed variable (scale) provide us with one that can be used repeatedly across subjects and situations to measure levels of an underlying latent variable (assuming that the construct or phenomenon is relatively fixed and not drastically changing). Such measured indicators will also have a high degree of unconditional inter-item correlation.

In the social science literature, reliability is defined roughly as "the degree to which a measure is consistent or dependable; the degree to which it would give you the same result over and over again, assuming the underlying phenomena is not changing" (Trochim 2005). We focus on an approach to assessing reliability via the internal consistency of a set of measured indicators of a latent variable. To the extent that measured indicators are highly responsive to changes in levels of an underlying phenomenon (with relatively low levels of noise or measurement error), we will have reliable measures of latent variables via constructed scales.

Fancsali (2008) focused on Cronbach's alpha (Cronbach 1951), a commonly deployed quantitative measure of internal consistency and therefore reliability. One formulation of Cronbach's alpha is (Trochim 2005):

$$\alpha = \frac{nr_{aii}}{1+[(n-1)r_{aii}]},$$

where n is the number of items in the set of measured indicators under consideration and r_{aii} is the average inter-item correlation over pairs of measured indicators. Thus, if the set of measured indicators comes from, for example, a survey instrument, alpha is a function of the length of the instrument and the average inter-item correlation among instrument items.

In models we consider, the average inter-item correlation of a set of indicators, r_{aii}, is a function of the factor loading for each indicator as well as the error term of each indicator, the latter of which we can normalize in terms of the factor loading of the indicator; this makes the average inter-item correlation depend only on factor loadings. Fancsali (2008), with simulated data, demonstrated that Cronbach's alpha taking on a value greater than or equal to roughly 0.8 is both necessary and sufficient for the correlation of a multi-item scale and an underlying latent construct to be greater than or equal to roughly 0.9. Recall that in the same simulation setting, a correlation of about 0.9 was determined to be necessary for a scale to preserve conditional independence relations of its intended latent variable. Other assumptions in the analysis are those considered above with respect to the items that compose a scale. Thus, given these (perhaps strong) assumptions, a constructed scale possessing a value of Cronbach's alpha greater than or equal to roughly 0.8 will preserve conditional independence for its intended latent variable.

The result in Fancsali 2008 is demonstrated for cases in which factor loadings, quantifying the strength of causal relations between the latent construct and measured indicators, are fixed and equal, an assumption equivalent to what is referred to in the literature as "essential tau-equivalence" (Miller 1995) and for cases in which factor loadings take on a wide variety of different values. Thus, a connection is provided between estimates of internal consistency and reliability via Cronbach's alpha and the necessary level of correlation, found via simulation, to preserve "screening off" and conditional independencies for causal discovery.

Nunnaly and Bernstein (1994) provide a heuristic that an instrument (i.e., a set of items comprising a scale) be used when it has an alpha value greater than or equal to roughly 0.8.[4] Fancsali (2008) confirms this, given the relatively strong but crucial assumptions provided above (and frequently made in practice), via the novel computational argument outlined above. This heuristic allows us to deploy scales as constructed proxies for latent variables that preserve probabilistic (in)dependence relations.

To make this discussion more concrete, consider again the graph in Figure 2.3, wherein $relationshipSatisfaction$ has a pure measurement model consisting of indicators X1, X2, and X3. If the calculated value for Cronbach's alpha for RS_SCALE, composed of measured indicators X1, X2, and X3, is greater than or equal to roughly 0.8, then we will judge that EMPLOYMENT and FAMILYSIZE are independent given RS_SCALE. Since $relationshipSatisfaction$ is unobservable, this is the best situation we could hope for to detect a lack of causal connection between EMPLOYMENT and FAMILYSIZE. We move from an estimate of the internal consistency and reliability of a set of observed raw variables composing a constructed variable to a causal conclusion concerning a latent variable as a common cause of EMPLOYMENT and FAMILYSIZE.

2.2.6 Data-Driven Search for Constructed Scale Variables

Having established the virtues of a scale that preserves (probabilistic) conditional independence relations for a latent variable, we now consider procedures for discovering the appropriate indicators or items to include in such scales. We suggest both an exhaustive search procedure and a "leave-one-out" procedure. Other heuristic procedures may prove useful. Background knowledge and domain theory play an important role in meeting the assumptions elucidated above. We elaborate upon the deployment of Cronbach's alpha as a crucial ingredient (i.e., an objective function) for data-driven search. We conclude §2.2 with a concrete example of scale construction for $relationshipSatisfaction$.

---

[4] The appropriate threshold for Cronbach's alpha is not without some controversy; several values are suggested in the literature. For example, 0.7 is also often recommended.

2.2.7 Scale Items and Function: Background Knowledge + Heuristic Search

As presented, scales are deterministically constructed variables, constrained to be linear combinations of raw variables (here, measured indicators) with all coefficients either 0 or 1. Each scale is just the sum of the items we include in it (e.g., RS_SCALE = X1 + X2 + X3).[5] Standard methodological practices thus inform the solution to the problem of the appropriate mathematical function to construct a scale as a proxy variable; this problem is analogous to that of "aggregation function" introduced for constructed aggregate variables in §2.1.4.

To choose items to be included in each scale, the practitioner starts with some set of candidate indicators/items. The problem the investigator must solve is analogous to that of aggregation level for constructed variables introduced in §2.1.3.

The practitioner might deploy a survey with sets of items intended to measure different latent phenomena; we assume that practitioners deploy a survey instrument that has been previously validated for the construct intended. Starting from the set of candidate indicators, we must decide which to include in the final constructed scale and which to discard. Here, semantic background knowledge can allow us to eliminate indicators which are not "pure" as described above or that might have otherwise become problematic in the data collection process. Survey items can, for example, be eliminated if there are expected "anchoring" effects, when a subject's response to one item might influence responses to other items.

Alternatively, a data-driven method like the BPC algorithm (Silva, et al. 2006) might be deployed on the full set of indicators (for all latent constructs or phenomena) to provide us with "pure" clusters of indicators; each cluster would then correspond to a set of candidate indicators for a particular latent phenomenon.[6] Background knowledge, domain theory, and possibly data-driven methods help us to delimit the space of possible variables from which to construct proxies while also helping us to insure that assumptions of validity and measurement model purity (local independence) are plausible. We now must consider the reliability of scales constructed from subsets of

---

[5] We assume that the indicators (raw variables) are all measured on the same scale or have been normalized in such a way that they are on the same scale.

[6] Of course, the number of pure clusters found by such an algorithm may not match the number of latent phenomena for which the practitioner has set out to construct proxies, so careful thought must be put into the reification of latent variables/phenomena of which the pure clusters are indicators.

candidate items for each latent construct to develop a data-driven, heuristic (proxy) construction search procedure.

2.2.8 Cronbach's Alpha as a Heuristic

Having a candidate set of indicators for which we find assumptions of validity and measurement model purity plausible, we need to discover the subset of indicators that maximizes reliability. Having demonstrated necessary conditions for a constructed scale to preserve conditional independence relations in terms of (internal consistency) estimates of reliability, we provide a heuristic (i.e., an objective function to maximize) for construction search in terms of Cronbach's alpha.

The objective function is simple: for each candidate set of indicators from which we can construct a scale, calculate Cronbach's alpha. Our objective is to find a subset[7] of indicators from the set of candidates that either maximize Cronbach's alpha or to find a set by incrementally adding items that increase the value of Cronbach's alpha until the set of indicators is of a desired cardinality. We suggest two search procedures for these objectives.

2.2.9 Two Heuristic-Based Search Procedures

An exhaustive search procedure considers, for each candidate set of indicators (each of which are intended as measures of a particular latent phenomenon), each possible subset (above a minimum cardinality) of indicators and calculates Cronbach's alpha for that subset. We find the subset that maximizes the value of Cronbach's alpha. From the subset that maximizes Cronbach's alpha, we construct a scale as a sum of its members. This scale should serve as an appropriate proxy for the underlying latent variable/phenomenon of interest (given that the appropriate assumptions obtain) and preserve conditional independence relations for that construct.

A less computationally intensive procedure for search involves simple "leave-one-out" reasoning.[8] We start with a set of candidate items meeting the assumptions laid out earlier in this section. We calculate Cronbach's

---

[7] It is sensible to impose a constraint on the minimum cardinality of the set of indicators from which we will construct a scale. For example, we might say that a minimum of three or four indicators for each scale is necessary before we can sensibly reify our scale as a proxy for an underlying latent variable in an application domain.

[8] Indeed, the statistical package Minitab, for example, in its "Item Analysis" module that calculates Cronbach's alpha for a set of items, provides the results of calculating Cronbach's alpha leaving each item in the set out one at a time to suggest ways in which the internal consistency of the set of items might be increased.

alpha for the complete set of candidate items, and then we calculate Cronbach's alpha for each set of items that results when we omit one item in the set at a time. We then select the set that yields the largest value of Cronbach's alpha. This process iterates until we reach a minimum cardinality of the set of items or no item omission increases Cronbach's alpha. We might also have a pre-set threshold for Cronbachs alpha that, when reached, would provide a stopping criteria for the "leave-one-out" process. The resulting subset of candidate items provides us with components from which to construct a scale.

2.2.10 Example: Relationship Satisfaction Scale

We conclude §2.2 with a real-world example. We consider the construction of a scale to serve as a proxy for a latent variable intended to capture satisfaction in a romantic relationship (a la $relationshipSatisfaction$ of our prior example). This illustration uses data from an analysis of the determinants of romantic relationship satisfaction in Cramer 2004 and follows portions of the (re)analysis of this data in Fancsali 2008.

As noted, Cramer (2004) deploys the Relationship Assessment Scale (Hendrick 1988) as part of a survey that includes instruments to assess a total of four latent constructs (including relationship satisfaction). Seven items, numbered S1 through S7 (provided earlier in this section), in the Relationship Assessment Scale serve as candidates for our scale. Responses to these seven items are on a Likert scale of 1 to 5.

Cramer analyzes data from 111 student respondents to these questions and others, providing a structural equation model of the influence of depression, "conflict satisfaction," and "support satisfaction" on overall satisfaction in a romantic relationship, the latter of which is measured using the above instrument. Further analysis as to the broader causal model and other features of this study are explored in depth by Fancsali (2008). We focus here on the construction of the scale for relationship satisfaction based on data for the above items.

Our objective is to maximize the value of Cronbach's alpha using a "leave-one-out" strategy, so we begin by calculating Cronbach's alpha for the items S1 through S7; we find an alpha value of 0.697. This does not meet our minimum heuristic threshold of 0.8, suggesting data-driven search for a better scale. We consider the omission of one particular item and find that omitting S7 leads to the largest increase in the value of Cronbach's alpha. Alpha calculated for items S1 through S6 is 0.82. This set of indicators thus meets our threshold of 0.8 and so would be suitable as a set of indicators from which to construct a scale. We find, however, that omitting another item, S4, will lead to a further increase in Cronbach's alpha. Alpha calculated for the set $\{\text{S1}, \text{S2}, \text{S3}, \text{S5}, \text{S6}\}$ reaches a level of 0.835. Omitting any single item at this stage does not lead to an increase in the value of Cronbach's alpha. Thus, we might adopt this as a plausible set of indicators from which to build a scale. Indeed, the reanalysis provided by Fancsali (2008) proceeded with exactly this set of items as the components of the scale for Relationship Satisfaction. Deploying this constructed scale with graphical causal discovery methods suggests plausible alternative causal models that explain the data as well as (or better than) the published model and lead to different causal explanation of some of the latent phenomena of interest. Notably, some features of the published structural equation model are also supported by this reanalysis.

We have not necessarily maximized the value of Cronbach's alpha for all given subsets of the candidate items, but we have found a sensible stopping point in accordance with one of the heuristic objectives above, retaining five of the seven candidate items. We thus illustrate combining background knowledge and a data-driven heuristic within a search procedure to discover constructed variables fit for our purposes. Here we search for a construction of a scale that will preserve probabilistic (in)dependence relationships for the underlying phenomenon for which it is a proxy.

While this example is with respect to a particular scale and latent phenomenon of interest, the over-arching theme is that this sort of procedure can be followed for any given latent variable or phenomenon we propose to measure in similar ways. Following a procedure akin to this for other latent variables makes it more likely that search procedures deployed over such constructed variables will reliably infer features of the true latent variable causal structure. From a combination of background knowledge and the preservation probabilistic (in)dependence relations by constructed variables, we are in a better position to treat constructed scales as proxies for latent variables that are the underlying (social) scientific features we intend to investigate. Consequently, we are in a better position to make causal inferences about these latent phenomena and thus to provide better recommendations for policy and interventions.

2.3 Global Climate Data

We now consider a real-world example of Situation #2, involving coarse (or macroscopic), directly measureable aggregates over fine-grained (perhaps microscopic) measurements in data. In contrast to the case of latent variables in Situation #1, aggregate features of interest are functions of these fine-grained measurements and have necessary semantic connection(s) to these "raw" data, namely that we identify aggregate phenomena by appropriate aggregate constructions, like a "high-pressure system."

On a larger scale, consider measurements of regional sea surface temperatures and atmospheric pressure (e.g., from ocean buoys and other sources) that are frequently aggregated into indices so that regional climate phenomena can be modeled and correlated with other, distant phenomena; these correlations are commonly referred to as climate "teleconnections" (Glymour 2007). One well-known "teleconnection" concerns El Niño, characterized by unusually warm temperatures at the surface of the equatorial Pacific Ocean and implicated in a variety of weather and climate phenomena across the globe (NOAA 2013).[9]

Data are fine-grained, direct measurements of climate phenomena and, in many situations, best used to develop explanations of broader, aggregate phenomena (i.e., Situation #2). Macroscopic structure of climate teleconnections and relationships among such indices are relatively simple compared to any model that attempts to model fundamental physical principles that drive such macroscopic relationships (Glymour 2007).

Aggregates (indices) are frequently calculated over canonical, climatologically relevant geographic regions.[10] An alternative strategy is to deploy data-driven procedures to discover regions over which aggregates or indices are constructed (e.g., Steinbach, et al. 2003). Notably, data-driven efforts have not focused on aggregation and index construction to support causal inference or to preserve "screening off" relationships vital to causal modeling and discovery. We compare results of efforts to develop causal models of global climate teleconnection phenomena from indices constructed both ways. Unfortunately, deploying new methods developed later in this work on climate data remains a topic for future research; we develop this example

---

[9] We emphasize that El Niño is not a latent phenomenon for which we seek a constructed variable proxy; rather, we identify El Niño with aggregated measurements of sea surface temperatures over the appropriate ocean region.

[10] While we consider an example from climate modeling in which well developed domain theory provides canonical geographic regions and aggregation functions (though the question remains whether optimal), we are not always so fortunate, even within this discipline. Parker (2011) notes, in a discussion of sources of uncertainty in simulated global climate models, that "for processes that are believed to influence climate in important ways but that occur on scales finer than those resolved in today's models... rough representations in terms of larger-scale variables must be developed, and it is rarely obvious how this can best be done" (Parker 2011, pg. 581-582).

because it is instructive philosophically and to better motivate applications on which we deploy new methods.

2.3.1 Domain Knowledge vs. Data-Driven (Non-Causal) Climate Index Construction

We describe two attempts at aggregate variable construction (i.e., index construction) with climate data and causal modeling with these variables. In one case, efforts are informed heavily by domain knowledge, and in the other case data-driven clustering methods are deployed.

Domain Knowledge Climate Index Construction

Glymour (2007) reports results of Chu, et al. (2005) analyzing relationships between the following climate indices, constructed from well-established domain knowledge that provides both relevant aggregation levels (i.e., geographic regions) and aggregation functions:

- QBO (Quasi Biennial Oscillation): Regular variation of zonal stratospheric winds above the equator.
- SOI (Southern Oscillation): Sea level pressure (SLP) anomalies between Darwin and Tahiti.
- WP (Western Pacific): Low frequency temporal function of the ‘zonal dipole’ SLP spatial pattern over the North Pacific.
- PDO (Pacific Decadal Oscillation): Leading principal component of monthly sea surface temperature (SST) anomalies in the North Pacific Ocean, poleward of 20° N.
- AO (Arctic Oscillation): First principal component of SLP poleward of 20° N.
- NAO (North Atlantic Oscillation): Normalized SLP differences between Ponta Delgada, Azores, and Stykkisholmur, Iceland (Chu, et al. 2005, pg. 4).

From time series[11] data for each index, available for hundreds of months, three lagged series (i.e., three lagged variables) are constructed. For exam-

---

[11] Climate index variables “are not functions of features of any particular set of objects, but rather of features of whatever objects occupy a certain volume of space; and those objects and the values of their relevant variables are continually changing” (Glymour

ple, time series data for the Southern Oscillation determine four total variables:

$SOI_0 = \{s_1, s_2, ..., s_{500}\}$,
$SOI_1 = \{s_2, s_3, ..., s_{501}\}$,
$SOI_2 = \{s_3, s_4, ..., s_{502}\}$,
$SOI_3 = \{s_4, s_5, ..., s_{503}\}$,

and so on for the other indices. Causal graph search deployed over these variables results in the graph of Figure 3, reproduced from Chu, et al. 2005.[12] Glymour (2007) notes that the graph is what we should expect in the situation at hand:

> Despite the fact that the indices do not determine the microstate of a region, the indices screen one another off exactly as in a causal sequence: the southern oscillation is independent of the Pacific decadal oscillation conditional on the spatially and temporally intermediate Western Pacific measure; WP_t is independent of SOI_{t-1} conditional on SOI_t and WP_{t-1}. These independence relations are exactly what we should expect if the arrows in the [inferred causal] diagram [see Figure 2.4] represent relatively direct causal inferences and if there are no significant unobserved common causes of represented variables (Glymour 2007, pg. 340).

Chu, et al. (2005, pg. 16) further report that relationships among PDO, AO, and NAO are in agreement with expert opinion and that other relationships, if novel, are at the least "not controverted" by such opinion. That modeling results coincide with expert opinion provides good evidence that well-established domain theory provides suitable constructed variables to represent climate phenomena of interest. However, one might consider other, data-driven ways of constructing indices or aggregates from such data to provide better results.

2007, pg. 340). That the data for each aggregate climate index form a time series of many aggregated measurements (i.e., that the aggregate units of analysis are time-stamped measures over particular geographic regions) is not crucial to ideas about aggregation from fine-grained data upon which we focus.

[12] The same results are also reported and discussed by Glymour (2007).

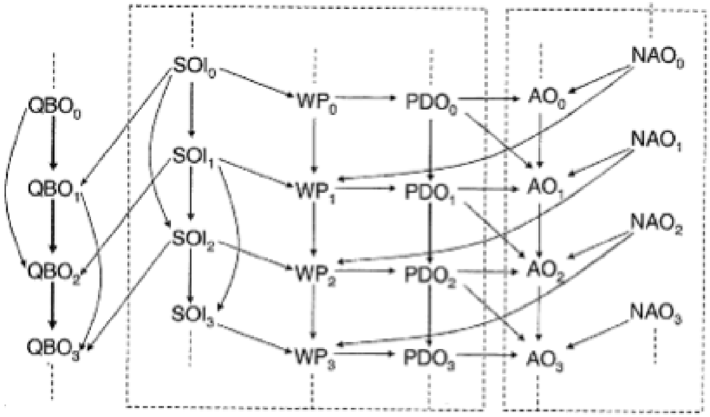

Figure 2.4: Causal graph for time-series of climate indices reproduced from Chu, et al. 2005 and Glymour 2007.

Data-Driven (Non-Causal) Climate Index Construction

Chu, et al. (2005) and Glymour (2007) also report attempts to learn causal structure among climate indices constructed by data-driven clustering methods. Steinbach, et al. (2003) deploy a shared-nearest neighbor clustering algorithm and propose hundreds of candidate climate indices based on aggregates of sea surface temperature and sea level pressure measurements. Notably, such clustering techniques do not attempt to constrain search for clusters to those that are causally interpretable, those that would preserve conditional independence relationships, or those that otherwise have characteristics important for causal modeling. Causal structure learning from these candidate indices does not produce satisfying results:

> The time series of these indices do not... generally form robust structures like those of Figure [2.4]... for which a connection that occurs between an index at time k and another index at time $k+1$ occurs again at times $k+1$ and $k+2$, respectively, and so on. That stability does not happen with most of... [Steinbach, et al.'s] climate indices: there is no stable screening off; the conditional independence relations are unstable and change over relatively brief times. The automated clusters are functions of the underlying energies all right, but the wrong functions in the wrong places (Glymour 2007, pg. 342).

Glymour provides an over-arching lesson: "[N]ot every way of aggregating to form macroscopic variables will yield screening off relations that reflect a causal structure; indeed many will not" (Glymour 2007, pg. 342). For the case of these climate indices, domain knowledge and expert opinion provide "good" ways of forming aggregate constructed variables (e.g., appropriate aggregation levels and functions) to represent global climate phenomena and causal structures that obtain among them. However, we are able to judge them as "good" exactly because of this domain knowledge and the spatio-temporal nature of such climate systems. That is, in this case, domain theory provides guidance as to the appropriate qualitative causal structure we ought to infer from appropriately constructed variables.

2.4 fMRI Data

We now consider an example of Situation #3. We face indirect, "raw" measures we seek to aggregate to construct variables of scientific interest. Specifically, we consider data from fMRI studies in cognitive neuroscience. As noted, such data provide us indirect measures of neuronal activity at relatively small, three-dimensional segments (i.e., voxels) of the brain from BOLD response (Lazar 2008). While in some cases, cognitive neuroscientists and other investigators may be interested in modeling relationships among individual or small groups of voxels, at other times we seek (causal) explanations of relationships between, or "effective connectivity" (Friston 1994) of, relatively coarse(r) brain ROIs.

We hypothetically illustrate the case of measuring neuronal activity via voxel level BOLD response to develop aggregate ROIs in Figure 2.5. Dashed arrows represent individually insignificant, if present, (causal) relationships among neuronal activity variables that are not directly observable ($NA1$ through $NA2000$ in Figure 2.5). Variables VOXEL1 through VOXEL2000 are measured variables for BOLD response at particular volumetric pixels, providing indirect access to $NA1$ through $NA2000$. Ellipses indicate omitted variables (and concomitant omitted dashed arrows). Further, it need not be the case that every neuronal activity variable is (insignificantly) causally related to every other neuronal activity variable associated with a particular ROI, but it is likely that such links are relatively "dense."

Of particular scientific interest in our hypothetical scenario is the causal link between ROI1 and ROI2. While individual causal relationships among instances of fine-grained neuronal activity are relatively insignificant, aggregate coarse-grained ROIs can have significant causal dependencies important

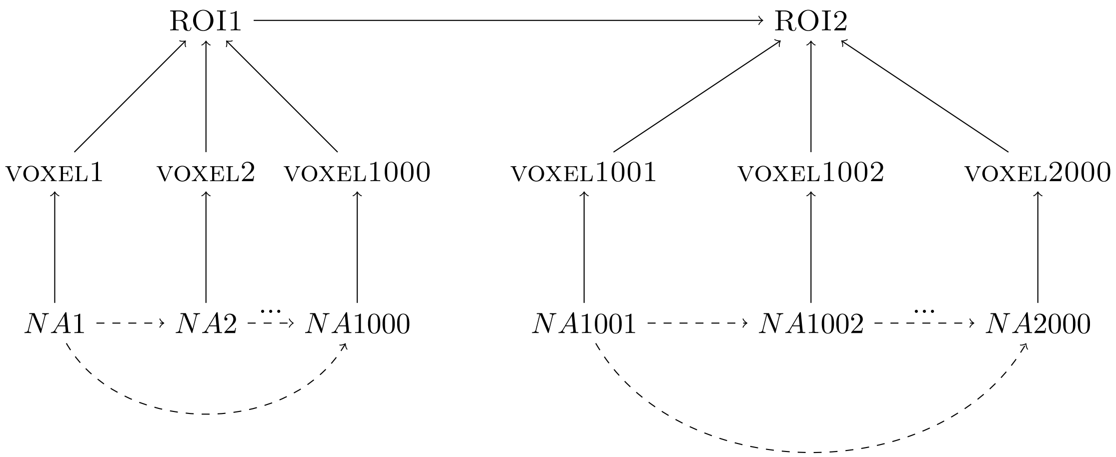

Figure 2.5: Hypothetical illustration of indirect measurement of neuronal activity ($NA1$ to $NA2000$, ommitting many) at volumetric pixels (VOXEL1 to VOXEL2000, ommitting many) to construct brain ROIs.

to the explanation of mental phenomena and human behavior.

Along these lines, Glymour (2007) argues that aggregation akin to that deployed for global climate teleconnections is the appropriate way to conceive of mental causation. Thoughts or mental processes are to be conceived of as the aggregation of a multitude of microscopic physical processes; the salient level of interest (scientifically and philosophically) to explain mental phenomena such as how beliefs cause our actions is that of brain ROIs.

Brain activity at a region (indirectly) measured at a particular voxel and sea-surface temperature (directly) measured at a particular buoy, while immediately available in data, are not salient features for developing explanations, especially causal explanations, of certain phenomena of interest (e.g., mental causation and large-scale climate phenomena, respectively). Specifically, just as temperature or pressure at one particular buoy is insufficient to causally explain the effects of El Ni*ñ*o, so is activity at a particular (small group of) neuron(s) insufficient to explain mental phenomena such as how beliefs cause our actions. Only at higher-levels of aggregation do collections of individually insignificant causal relationships among micro-phenomena (or micro-variables) become significant for a variety of instances of scientific explanation.[13]

---

[13] Sinnott-Armstrong (2005) provides an ethical variation on this theme in the context

In practice, cognitive neuroscientists supply domain expertise about the anatomy of the brain relevant to defining ROIs. Ramsey, et al. (2010) deploy a modified version of the GES algorithm to discover causal relationships that are plausible given state of the art domain knowledge among canonically constructed ROIs. On-going work is aimed at improving these algorithms for use with fMRI data (e.g., Ramsey, et al. 2011). However, data-driven approaches for inferring better ROIs are wanting, as much is still unknown about how exactly such regions should be delineated. Indeed, the problem of inferring ROIs has been cited as a quintessential example of an open question for researchers in causal discovery from observational data:

> In many domains, such as fMRI research, there are thousands of variables, but the measured variables do not correspond to functional units of the brain. How is it possible to define new variables that are functions of the measured variables, but more useful for causal inference and more meaningful? (Spirtes 2010, pg. 1659).

Notably, Mumford, et al. (2010) propose a data-driven method for the discovery of ROIs based on independent components analysis. It is an open question whether such methods are appropriate to infer better ROIs that preserve causal (i.e., conditional independence or "screening off") relationships that obtain in the brain or that are in some other way more meaningful.

In sum, domain knowledge and theory, in the case of discovering effective connectivity from fMRI, provide us fewer resources (than those available from climate science) for determining whether we have discovered the "best" ways of constructing aggregate ROIs (especially if judging constructed aggregate variables based on inferred causal structure among ROIs). While climate science and cognitive neuroscience practitioners have good ideas about appropriate aggregation functions and levels, in other domains, we find ourselves with even less domain knowledge and theory to guide us.

Again, methods we propose to search for constructed aggregate variables treat direct and indirect fine-grained measurements (Situation #2 and Situation #3) in data methodologically the same, but differences arise, for

---

of personal versus aggregate (i.e., institutional, national) responsibility for the effects of global warming. While Sinnott-Armstrong concludes that at best it is unclear whether it is morally wrong for an individual to go on a joyride in a gas-guzzling S.U.V. (insofar as such behavior contributes to or causes global warming), he admits that man-made global warming is a problem for which governments and large-scale institutions are morally responsible. Insignificant, individual contributions, in aggregate, produce phenomena for which governments and international organizations must develop better policy to combat.

example, in possible interpretations of resulting causal models (e.g., how best to consider interventions on aggregate variables that might "close the loop" from the results of such models). We now introduce data that are the focus of an in-depth case study in which we deploy methods we develop in the next two chapters. These complex data exhibit aspects of all three situations we have considered.

2.5 Education Data: Intelligent Tutor and Courseware Logs

2.5.1 Intelligent Tutor Logs

In Chapter 5, we consider data collected from education courseware deployed in courses at the University of Phoenix and in (junior) high schools nationwide. Our case study centers on search for constructed aggregate variables from logs of student interactions with intelligent mathematics tutoring software. Such tutor logs provide fine-grained, transaction-by-transaction access (over thousands of transactions per student) to variables that are direct measures of aspects of student interactions like transaction timing (e.g., time per problem per student), whether transactions are requests for help (e.g., whether a transaction is a hint request), and a variety of other transaction characteristics. Other characteristics immediately available in tutor logs may be interpreted as indirect, fine-grained measures of other underlying phenomena over which we should construct aggregate variables. Further, we deploy data-driven "detectors" to indirectly measure (or infer) fine-grained student behaviors like being "off-task" or "gaming the system," behavior associated with students taking advantage of tutor features to advance through a course without genuinely attempting to master course material. As such, our case study and search for constructed aggregate variables depends heavily on well-established domain theory and background knowledge surrounding these software "detectors."

However, extant work on such tutor log data either analyzes it at the transactional (or some other fine-grained) level, seeking, for example, predictive models of whether particular transactions will be correct, or (in fewer cases) at a high-level of aggregation (e.g., aggregating over all student transactions in a particular [portion of a] course). We propose that aggregating over such fine-grained data is appropriate to develop student-level models of behavior, but that both the appropriate level of aggregation and the appropriate function to transform fine-grained variables into aggregate constructed variables (i.e., what we have called the aggregation level and aggregation function problems) have not been explored in sufficient depth.

Still other constructed (aggregate) variables we construct from fine-grained (direct and indirect) measures may be indicators of underlying latent phenomena (a la Situation #1), and we might deploy methods for scale construction to construct variables. We provide a deeper look into the data collected by automated intelligent tutoring systems as well as "detectors" deployed to provide input to search procedures for aggregate constructed variables in Chapter 5.

2.5.2 Online Courseware Forum Messages

Fancsali (2011a, 2011b) considers data collected from message forums from an online graduate MBA economics course at the University of Phoenix. Online message forums were the primary means of communication and interaction between students and between students and course facilitators in online courses like this over the time period in which these data were collected. Fancsali (2011a) presents difficulties inherent to attempts to construct ad hoc aggregate variables from such data, lacking guidance from domain theory and background knowledge. From such data, one might seek insight into relationships between aggregate functions of raw characteristics of messages (e.g., word and character count, subject line word/character count, etc.) and learning outcomes in the course; from message-level raw log data, one can construct student-level variables to represent students' aggregate messaging behavior. One can then interpret variables out of which we build aggregate constructed variables as either direct or indirect measures of messaging behavior. Alternatively, constructed aggregate variables may be plausible indicators of broader, underlying, latent phenomena (e.g., "student engagement"). We do not pursue this work as a case study within this dissertation, but merely point to it as a particularly vexing problem for variable construction and causal modeling with education data.

2.5.3 Further Remarks: Judging Constructed Variable "Quality"

Whether we seek variables to identify aggregate phenomena (Situations #2 and #3) or proxies/instruments for latent phenomena (Situation #1), a central concern is whether variables are (at least approximately) "correct." We ask: How do we judge "goodness" or "quality" of constructed variables or features like climate indices or an instrument for relationship satisfaction?

Glymour's (2007, pg. 342, footnote) remarks regarding the success of ocean indices constructed from expert/domain knowledge point to one partial answer: constructed variables should enter into screening off relations that reflect causal structure. In the case of ocean indices, for example, times series of the indices provide for stable screening off relationships (i.e., causal connections and screening off relationships persist through time, as Glymour notes: a causal connection, between one index at k and another at $k+1$ persists between the two indices at time $k+1$ and $k+2$). Proposed alternative variables provide for no such stable screening off relations.

Other characteristics of constructed variables might be important; we suggest the importance of their contribution to (causally) predictive models of a fixed target variable. The analog of stability of the sort Glymour suggests is important for climate indices will be apparent in some applications but possibly less obvious for others.

Especially in domains in which we lack significant theory and/or background knowledge to construct aggregate variables, we suggest that a sensible strategy to overcome this deficit is to develop data-driven search procedures for sets of aggregate constructed variables. We conjecture that useful (possibly "correct") constructed variables exhibit patterns of conditional independence that allow us to infer causes of a target outcome and to best predict a target based on (interventions on) inferred causes.

In the following two chapters, we develop methods to adjudicate the quality of collections of aggregate constructed variables and search for "high quality" sets of variables based on such ideas. However, what have we discovered if and when we find plausible, useful variables for causal inference?

To partially answer this question, we consider recent literature in the philosophical tradition of "natural kinds" and suggest that constructed variables that support causal inference (like canonical climate indices) are candidate natural/scientific kinds, worthy of further scientific investigation in cases in which we have little domain knowledge to (a priori) support their status as plausible natural/scientific kinds.

2.6 Natural Kinds

We largely follow an explication of natural kinds due to Machery (2005; 2009, Chapter 8), who aimed to establish a formulation of natural kinds appropriate for psychology. Machery focuses on the import of causal relationships for natural kinds, synthesizing a great deal of the contemporary literature on the topic. Further, general features of his account seem to broadly apply outside of psychology, so it seems (at least approximately) appropriate for the study of aggregate behavior in education courseware we take up in Chapter 5. Finally, state-of-the-art work on behavior in education courseware explicitly seeks to model the type of aggregate behavior with which we are concerned in Chapter 5 alongside student affect and emotion, topics that necessarily invoke psychological kinds.[14]

There are two criteria that an account of natural kinds must meet to be appropriate for psychology: (1) that it apply to psychological kinds, and (2) that it be "broad, meaning that many classes have to qualify as natural kinds under this account" (Machery 2009, pg. 231).

2.6.1 Essentialist and Nomological Approaches to Natural Kinds

One approach to natural kinds that goes at least as far as back as Locke ([1690] 1979) and that is still prominent in contemporary literature (Kripke [1972] 1980; Putnam 1975) identifies those kinds as natural that possess an "essence, that is, a set of intrinsic, causally explanatory properties that are necessary and jointly sufficient for belonging to the kind" (Machery 2009, pg. 231). Examples of natural kinds by this tradition include chemical kinds such as lead or that of the identification that "water = H_2O" (Putnam 1975). The essentialist program appears insufficient as an account of psychological kinds, because psychological kinds (i.e., emotions, desires, etc.) are not identified by intrinsic properties but rather by their causal roles and thus relational properties (Machery 2009, pg. 231-232).[15]

The next approach on Machery's chopping block is that in which natural kind terms take part in laws of nature, the "nomological" account of natural kinds. Laws of nature refer to "generalizations that are temporally and spatially unrestricted and that support counterfactuals" (Collier 1996; Machery 2009, pg. 232). This account is inappropriate for psychological kinds, because it is too restrictive and because psychological kinds only feature in *ceteris paribus* generalizations (Fodor 1974).[16] Noting that

---

[14] We discuss contemporary research on student affect and emotion in education courseware in Chapter 6.

[15] Machery resists the idea that we simply include relational properties in the notion of an "essence," arguing that such an account "would not distinguish the properties that determine membership from the causally explanatory properties. However, for some natural kinds, the properties that determine the membership in these kinds are not identical to those that causally explain why the members of these kinds have many properties in common" (Machery 2009, pg. 232).

[16] If one were to adopt an approach wherein laws of nature include *ceteris paribus* gen-

it is a step in the right direction to consider loosening the requirement of underwriting laws (of nature) to merely underwriting *ceteris paribus* generalizations, Machery moves on to consider his (and our) preferred account of natural kinds: a causal account in which natural kinds are a particular type of property cluster.[17]

2.6.2 The Causal Notion of Natural Kinds

Machery's "causal notion of natural kinds" is roughly based on ideas due to Richard Boyd (1990, 1991, 1999; Griffiths 1997, Chapters 6-7; Machery 2005, 2009):

> A class C of entities is a natural kind if and only if there is a large set of scientifically relevant properties such that C is the maximal class whose members tend to share these properties because of some causal mechanism. ... The core idea ... is the following. A natural kind is a class about which many generalizations can be formulated: its members tend to have many properties in common. These generalizations are not accidental: there is at least one causal mechanism that explains why its members tend to have those properties. Finally, this class is not a subset of a larger class about which the same generalizations can be formulated (Machery 2009, pg. 232-233).

Boyd's (1990, 1991, 1999) original account posits that natural kinds are Homeostatic Property Clusters (HPCs): underlying causal mechanisms responsible for shared properties within a natural kind are homeostatic. That is, the "instantiation of a property causes the occurrence of other properties and is caused by their instantiation" (Machery 2009, pg. 233). Boyd's account has become popular[18] in the contemporary philosophical literature. Indeed, Samuels and Ferreira (2010, pg. 222) declare that "philosophers of science have, in recent years, reached a consensus or as close to consensus as philosophers ever get according to which natural kinds are Homeostatic Property Clusters... ."

---

eralizations (possibly necessarily), one could (possibly) provide an amended nomological account appropriate for our purposes, but such an inquiry is inappropriate for our present purposes.

[17] Again, one might provide some sort of amended nomological causal account of natural kinds, but we do not pursue this line of thought. We emphasize the causal characteristics of any account of natural kinds.

[18] and widely discussed (see, e.g., Hacking 1991; Millikan 1999; Craver 2009; Williams 2011; Magnus 2011)

Dogs illustrate the paradigmatic example of species[19] as a natural kind. Dogs generally possess a set of properties many of which are explained by underlying causal mechanisms (examples below) that allow us to make (generally reliable) inductive inferences about dogs we will encounter in the future.

Dogs count as a natural kind, but "white dogs" do not comprise a natural kind because relevant properties of dogs that are white are shared by all dogs (i.e., Machery's account of natural kinds is not vacuous). Machery's account also allows for a greater variety of mechanisms and that multiple mechanisms can explain different properties of a natural kind. For the example of species alone, mechanisms can include common descent (e.g., as a causal mechanism for dogs that explains that their body plan is inherited from their mother's species or in bats that explains the structure of their wings), selective pressures, and social causes (e.g., that the social properties of dogs likely result from "artificial selection pressures" exerted by humans) (Hare, et al. 2002; Griffiths 1997; Machery 2009, pg. 234-235). Beyond the paradigmatic example of species, such an inclusive approach is appropriate for Machery's purposes as it allows for a wide range of natural kinds, including psychological kinds, but also including natural kinds ranging from chemical elements to (social) artifacts.[20]

Both Machery's account and Boyd's account are appropriately deemed "causal accounts" of natural kinds, and much literature devoted to Boyd's HPC account focuses on the nature of the underlying causal mechanism(s) that explain clusters of properties that we use to identify natural kinds. Millikan (1999), for example, provides an account of Boyd-style HPC kinds like species, and those frequently of interest to social scientists, like membership in ethnic, social, economic, and other groups, as historical. "Inductions made from one member of the kind [i.e., a particular species] to another are grounded because there is a certain historical link between the members of the kind that causes the members to be like one another" (Millikan 1999, pg. 54-55). Boyd goes on to deny that a historical link among members of an HPC kind is the crucial aspect for kind membership, arguing that species,

---

[19] Significant literature on natural kinds (and on the HPC account of natural kinds) concerns the problem of "species" in biology. Given the paradigmatic status of species, any account of natural kinds must count them as natural kinds (Boyd 1999; Millikan 1999; Magnus 2011; Devitt 2011; Rheins 2011; Crane and Sandler 2011). Literature so directed generally aims at clarifying the nature of HPC natural kinds (or natural kinds in general) and demonstrates how a particular account accommodates biological species (presumably better than a previous account).

[20] Machery's example is that of tools in paleoanthropology (Machery 2009, pg. 234).

for example, "are defined, according to the HPC conception, by those shared properties [phenotypic characteristics] and by the mechanisms ... which sustain their homeostasis" (Boyd 1999, pg. 81). Importantly, Boyd further argues that restricting HPC kinds to historical kinds would exclude a variety of non-historical phenomena that Boyd intends to capture under the wide umbrella as HPC natural kinds, including examples like storm patterns in meteorology, mineral species in mineralogy, and economic categories like capitalism in the social sciences.[21]

Magnus (2011) defends the HPC account of natural kinds, but takes issue with what he calls the "similarity fetishism" of Boyd's position,[22] the "basic idea ... that a natural kind is primarily a set of similar things" (Magnus 2011, pg. 863). Instead, Magnus emphasizes the fundamental importance of underlying (homeostatic[23]) causal processes for HPC kinds, arguing that the HPC account is separable from similarity fetishism. Magnus demonstrates this separation by providing some exotic (perhaps pathological) examples of recently discovered species, members of which are drastically different depending upon their sex, arguing that we need not have, for any particular HPC kind, "a single list of properties which are all typical for members of the kind" (Magnus 2011). Causally explanatory underlying mechanisms and processes are what really matter.

We will focus on the importance of causal explanations in the discovery of natural kinds, as emphasized in this recent work by Magnus, but also present in early work by Boyd on his account of natural kinds. We are not concerned in this work to provide an exact characterization of natural kinds or of a

---

[21] A broader issue in the recent literature on natural kinds (especially HPC-style kinds) is the dearth of in-depth deployment for kinds that are not species, e.g., social kinds, outside of some psychological examples (Griffiths 1997). Most literature treats situations in which relevant individuals are assumed or already well-known. For the case of capitalism, presumably the individuals which qualify to take part of the kind are countries or economic systems. Williams (2011) discusses this concern for the case of diseases as HPC kinds. For cases we consider (especially cases of kinds that pick out things like student behavior in courseware in Chapter 5, but also in examples already illustrated in this chapter), it is perhaps less clear what the relevant individuals even are. Indeed, one might see the discovery problem as being that of both natural kinds and scientifically appropriate individuals.

[22] One notes a hint of what Magnus deems "similarity fetishism" in the formulation of natural kinds provided by Machery as it involves "a large set of scientifically relevant properties such that ... members tend to share these properties because of some causal mechanism" (Machery 2009, pg. 232). Magnus presumably agrees about the importance of the underlying causal mechanism, but they diverge about the necessity of kind members sharing a "large set of scientifically relevant properties."

[23] though perhaps we need not be so restrictive, following Machery's account of natural kinds

particular formulation of the underlying mechanisms necessary to underwrite the co-occurrence of relevant property clusters; rather we propose that the data-driven search for variables that support and underwrite broader causal inferences provides means of discovering candidate natural kinds from (fine-grained) observational data not already "carved at (scientific or natural) joints." We focus on disciplines where relevant (scientific/natural) kinds are not well-established or as yet well-known (i.e., paradigmatic examples are lacking or the "best" way in which to construct appropriate variables is less developed).

While much of the literature on natural kinds and the HPC account focuses on how such an account (suitably modified and explicated) can accommodate the paradigmatic example of biological species as natural kinds, Boyd intended for the account to encompass a multitude of other natural kinds, extending beyond examples above like storms to social kinds. We consider Boyd's example of storms[24] briefly.

Considering storms as an HPC kind, we can make inductions about characteristics of storms[25]: we can identify them as arising from similar underlying causal mechanisms or processes and list their common characteristics. Storms also, like other HPC kinds, figure in a variety of broader causal explanations. Storms, for example, have effects, like making the grass in a field wet. A common sense causal explanation of the source of wet grass invokes the presence of a storm (when appropriate). Explanatory practices frequently involve providing causal explanations. Magnus emphasizes that "[s]ystems of classification are implicated in both induction and explanation, and there is no reason to think that induction is always doing the heavy lifting" (Magnus 2011, pg. 867). Magnus argues that an HPC natural kind is best seen as "an explanatory instrument, rather than a narrowly inductive one" (ibid, pg. 869) and, as is typically in the literature, briefly discusses how methodological considerations from biology inform such a view of biological species. We similarly emphasize the importance of causal explanations, for the discovery of candidate natural kinds (and useful scientific variables) from data.

---

[24] Boyd (1999) provides the (relatively vague) example of storms as a natural kind that the HPC account should accommodate; perhaps a more specific example of some meteorological phenomenon would be more appropriate (e.g., cumulonimbus clouds, low pressure systems, etc.). That he provides such a vague example without considering details is symptomatic of the lack of consideration provided for natural kinds other than the paradigmatic case of species (e.g., social kinds).

[25] One natural question to ask is that of the means by which we identify storms. What are the relevant individuals that take part of this natural kind? Do we build the identity of a "storm" up from fine-grained meteorological data?

Boyd's early conception already appreciates the importance of causal explanation for the demarcation of natural kinds. The strategy we propose to discover natural kinds *qua* variables that support causal inference, by preserving conditional independence relationships and allowing for better predictions of the effects of interventions, aligns with Boyd's broad injunction:

> [N]atural kinds reflect a strategy of deferring to nature in the making of projectability judgments: we define such kinds *a posteriori* in ways which reflect actual causal structure precisely because we are unable to identify or specify projectable generalizations without doing so. ... Kinds useful for induction or explanation must always 'cut the world at its joints' in this sense: successful induction and explanation always requires that we accommodate our categories to the causal structure of the world (Boyd 1991, pg. 139).

Our strategy defers to nature by providing a data-driven procedure to search among "possible cuts" (i.e., possible constructed variables) to find "nature's joints." There are legion possible "joints" in the applications we attack, so such a data-driven approach is vital.

2.6.3 Craver on Perspective and Levels of Aggregation

Further, in our approach we assume that a fixed target variable has been provided. From a computational standpoint, this provides one way to define an objective function for a search procedure but also provides a scientific and investigative goal. Craver (2009), considering the nature of underlying mechanistic accounts of HPC kinds, argues that human perspectives (i.e., investigative goals, etc.) frequently enter into the project of defining natural kinds. Since natural kinds are those which allow us to better explain, control, and predict, a natural question is: what are we trying to explain, control, and predict?

Craver provides an example relevant to Situation #3, of the hippocampus of the brain. Depending on investigative goals, one might "carve" the hippocampus at a number of different levels of description (all of which have important implications for the type of mechanisms underlying proposed HPC kinds). Importantly, one aspect of our search for constructed variables includes considering a variety of different levels of aggregation to determine which works best for a particular target variable. From such considerations, Craver concludes:

> This raises a challenge to defenders of the HPC account either to find an objective basis for taxonomizing the mechanistic structure of the world, or to argue that these perspectival intrusions into the accommodation process do not threaten the realist objectives that motivate the belief in natural kinds in the first place (Craver 2009, pg. 591).

We remain agnostic with respect to the exact nature of underlying mechanisms (a la Machery's account of natural kinds), but Craver's challenge remains. Perhaps nature is multiply-jointed at various levels of salience for scientific investigation. This may not be any serious threat to realism about natural kinds, as interventions in science and policy can have a variety of grain-sizes (or levels of aggregation). With the sort of approach we propose, given a particular target we hope to provide objective criteria (based on statistical criteria and the discovery of variables that support causal inference) upon which particular "carvings" of nature can be judged better or worse. Thus, perspectival intrusions need not threaten realistic objectives that have us seek possible targets for future interventions.

2.7 Toward Discovery: (Candidate) Natural Kinds and Causal Models

Our approach to discovering variables and thus candidate natural kinds is to search systematically through sets of constructed variables to find those that allow us to best infer causes of a particular target and make predictions for the target from (hypothetical interventions on) those causes. We diverge from the standard approach of literature on natural kinds by not focusing on the underlying causal mechanisms that underwrite property clusters but rather focusing on broader causal structure into which candidate kinds (i.e., constructed variables) fit. Providing scientifically meaningful (or useful) interpretations of constructed variables that result from search remains a possible difficulty, depending on the application domain and available background/domain knowledge.

Since causal structure is usually not uniquely determined by observed, non-experimental data, we develop ways to take such uncertainty into consideration so that we can find variables which (if possible) help mitigate such uncertainty. Some variables will help mitigate structure uncertainty better than others, for example those that better preserve particular conditional independence relationships relative to other variables; we propose that these variables are more likely to be located at or near “nature’s joints” (or at least they may point us in their direction). We “defer to nature” and observed data as Boyd guides us.

Climate indices provide good examples of variables or kinds (or perhaps individuals) with which scientists can make systematic, reliable predictions and inferences about global climate phenomena, and they play an important explanatory and causal role in such predictions. Indices constructed from data, as explored above, have been successfully deployed in statistical and causal models that accord with our best contemporary science by preserving expected screening off relationships (Glymour 2007). Nevertheless, it remains unclear whether such indices are optimally defined or whether data-driven discovery procedures may improve them for particular types of investigative purposes. In some domains, we will have the luxury of well-developed theory and background/domain expertise to provide guidance as to the demarcation of variables, kinds, or individuals for study. The state of the art in other domains provides less guidance.

In the following chapter, we develop several notions of “causal predictability,” roughly expected causal control we achieve for a particular target of interest from a set of candidate constructed variables and a causal model learned from those variables. To do this, we also develop several means of characterizing (and dealing with) causal structure uncertainty to determine the extent to which constructed variables support causal inference (and contribute to expected causal control). In Chapter 4, we provide the results of simulation studies to motivate an algorithm to discover sets of constructed variables that support causal inference that we deploy on real world data in Chapter 5. These sets of constructed variables constitute candidate natural/scientific kinds for further study. Real world data in Chapter 5 consists of aforementioned data from educational courseware logs, where the state of the art provides little guidance as to appropriate, aggregate variable constructions.

Chapter 3

Causal Predictability and Structure Uncertainty

In this chapter, we consider ways to judge the "quality" of sets of constructed variables for causal inference about a target variable. We begin by introducing two quantitative notions of "predictability," our ability to predict a target variable from a statistical model. Next, we consider two ways to specify and estimate statistical models according to the structure of graphical causal models. Combining notions of predictability with model specification from causal information, we introduce three types of metrics to quantify "causal predictability." We explicate these notions assuming causal sufficiency and using DAGs, but we seek to focus in practice on situations in which we do not assume causal sufficiency and learn a PAG. Generally, PAGs represent many causal explanations that are indistinguishable from observational data. As such, we combine notions of "causal predictability" with methods to deal with structure under-determination to provide concrete, practical metrics (e.g., in linear regression and structural equation models) for judging the support for causal inferences underwritten by a set of constructed variables. After explaining metrics in this chapter, we explore their efficacy and broader issues of variable construction for predictive and causal inference via simulation studies in the next chapter.

3.1 Two Notions of Predictability

We consider two ways to quantify the extent to which we can predict a target variable from a statistical model. Since we often use linear regression models, we provide a brief review.

3.1.1 Simple Linear Regression

Wasserman (2004, Chapter 13) provides a standard, concise overview of the simple linear regression model to investigate the relationship between a target (or response) variable Y and a single predictor variable X. The regression function summarizes the relationship we seek to estimate from data of the form $(Y_1,\, X_1)$, ... , $(Y_n,\, X_n)$ with some distribution $F_{X,Y}$:

$$r(x) = E(Y|X = x) = \int yf(y|x)dy \text{ (ibid, pg. 209; Eq. 13.1).}$$

Assuming that $r(x)$ is linear with only one predictor variable (a.k.a. regressor, co-variate) X:

$$r(x) = \beta_0 + \beta_1 x \text{ (ibid, pg. 209).}$$

Wasserman then provides the simple linear regression model, assuming that $V(Y|X = x) = \sigma^2$ and that $V(Y|X = x)$ does not depend on x:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ (ibid, pg. 210; Eq. 13.2),}$$

where $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$. Let $\hat{\beta_0}$ and $\hat{\beta_1}$ be estimates of parameters β_0 and β_1 that determine a fitted line,

$$\hat{r(x)} = \hat{\beta_0} + \hat{\beta_1} x \text{ (ibid, pg. 210; Eq. 13.3).}$$

Parameters estimated for predictors are important because they quantify the magnitude of relationships between co-variates and the variable we predict with the model as well as the individual "contributions" of changes in the value of co-variates to change in the expected value of Y. From these parameter estimates, we calculate predicted values for Y as:

$$\hat{Y_i} = r(\hat{X_i}).$$

Residual values are calculated as the difference between predicted values and those that actually obtain in our sample:

$$\hat{\epsilon_i} = Y_i - \hat{Y_i} = Y_i - (\hat{\beta_0} + \hat{\beta_1} X_i) \text{ (ibid, pg. 210; Eq. 13.4).}$$

It is also important to assess the goodness of fit of our estimated linear model to the data. Residual values help us to do this when we calculate the residual sum of squares (RSS):

$$RSS = \textstyle\sum_{i=1}^{n} \hat{\epsilon_i}^2.$$

Least squares estimate values of $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize RSS and the fraction of variability not explained by predictors in the model.[1] The fraction of variability explained by a regression model is one way we quantify model fitness.

3.1.2 Notion #1: Fraction of Variability Explained / Goodness of Fit

We use these ideas to assess the extent to which we can account for variability in a target variable from the predictors included in a model. Roughly, this provides us with a measure of whether predictors we include in the model explain substantial sources of variation (arising in principle from causes, random noise, and/or measurement error) in our target. To do this in a linear regression model, we let the total sum of squares $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (where $\bar{Y}$ is the sample mean) and consider one minus the fraction of the sample variance of the target not explained by predictor variables in our model; this is the regression R^2 value (Stock and Watson 2003, pg. 174):

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Other goodness of fit criteria for statistical models are widely used, including adjusted-R^2 values, Akaike Information Criteria (AIC; Akaike 1973), Bayesian Information Criteria (BIC; Schwarz 1978), and others that balance the trade-off between the goodness of fit of a predictive model and its complexity (e.g., number of predictors and corresponding parameters to estimate). Other measures of predictive accuracy such as area under the ROC curve (for classification accuracy) or mean squared error are also often used, depending on the problem being considered.

3.1.3 Notion #2: Effects on Expected Value of Target

In addition to fitting data well and explaining variability in a target, we consider the effect of changes in the value of predictors or co-variates on the expected value of a target. In a linear regression model with one predictor, its parameter estimate corresponds to the change in the expected value of the target variable given a one unit change in its value; if there is more than one predictor/co-variate in the model, then parameter estimates for each predictor correspond to the change in the expected value of the target given

---

[1] The derivation of these estimates and the extension of these ideas to the case of multiple predictor variables (i.e., multi-variate linear regression) is straight forward and can be found in most statistics textbooks, including Wasserman 2004.

a one unit change in that predictor variable, holding all other co-variates fixed.

If, for example, we include both X_1 and X_2 as co-variates in a linear regression model for Y and estimate corresponding $\hat{\beta}_1$ and $\hat{\beta}_2$ coefficients, respectively, we say that given a one unit change in the value of X_1, $\hat{\beta}_1$ is the estimated change in the expected value of Y holding X_2 constant. $\hat{\beta}_1$ is often called the partial effect of a unit change in X_1 on Y, holding X_2 constant (Stock and Watson 2003, pg. 150). This notion of predictability allows us to consider the magnitude of change to expect in a target variable given a change in value to one of its predictors.

Consider a hypothetical example of an early childhood education program that is estimated to have a large total effect on elementary school learning outcomes. Given a host of other causes that increase the variability in such learning outcomes (over, for example, the population of all school age children in the United States), we might find that, despite a large estimated total effect (i.e., high $\hat{\beta}$), we still only account for a small percentage of the variability in learning outcomes with the early childhood program alone (low R^2).

Contrast the above with the hypothetical early childhood education case involving a particular cognitive, verbal, or other task that most school children of a particular age in the United States can easily perform. Suppose investigators observe a particular strategy that is found to have a small effect on performance of this task among children that demonstrate difficulty with the task. Despite the relatively low $\hat{\beta}$ we hypothetically estimate, we could still find that estimated models have relatively high R^2 values; we account for a significant portion of the (little) variability in performance of the task over the population. Hypothetical numerical examples are also easily concocted in which small total effects are estimated in models with high R^2 values because target variables have low variance, and vice versa.

So far we have said nothing about whether predictors or co-variates in a statistical model are causes of the outcome. If, for example, a predictor is included in a model that is an effect of (i.e., is caused by) the target, we predict the value of the target conditioning on such an effect, but if we were to intervene upon the effect (e.g., by changing its value by one unit), our predictions are worthless; interventions on a target's effect cause no change in the value of the target.

We seek to know, in addition to how well we can predict a target from its causes, what to expect from interventions on causes of a target. For these purposes, we consider two ways to specify models according to the structure of graphical causal models.

3.2 Two Notions of Model Specification Based on Causal Information

Causal information provided by the structure of graphical models can be used in several ways; we provide two notions of how to specify certain types of statistical models based on knowledge about causes.

3.2.1 Notion #1: Predictive Models Specified According to Inferred Causes

A variable's Markov blanket (Pearl 1988) is the set of variables conditional on which a variable in independent of all other variables in the network. For DAG-based models, including Bayesian networks, a target variable's Markov blanket is its ideal set of predictors, comprised of its direct causes (parents), direct effects (children), and other direct causes of its direct effects (other parents of its children).

Consider, for example, the causal structure of the DAG in Figure 3.1. The set of parents or direct causes of target variable T is $\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}\}$; T has one child or effect, C, and C has one other parent or cause, S. Thus, the Markov blanket of T is the set of variables $\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}, \mathrm{C}, \mathrm{S}\}$. Further, T is independent of $\{\mathrm{G}, \mathrm{D}, \mathrm{L}\}$ conditional on $\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}, \mathrm{C}, \mathrm{S}\}$.

A variable's direct causes are a (generally proper) subset of its Markov blanket, and we might be concerned, under either notion of predictability, to investigate only the (direct) causes of a target, as we are motivated to investigate the extent to which we might predict and/or control T using its causes and discover efficacious intervention to impact T.

First, we consider the specification of a statistical model from only causes of the target variable. In the case of Figure 3.1, we would specify, for example, linear regression models for T with only three predictors or co-variates: P1, P2, and P3. After estimating such a model, we can ask questions with respect to the two notions of predictability we have introduced: what fraction of sample variation in T can be explained by its causes? What is the estimated partial effect for a unit change in each of T's causes holding its other causes fixed?

Answering the first question quantifies the extent to which we have successfully captured sample variation of T from causes upon which we might intervene; if the fraction of sample variation explained by causes of T is especially low, it may suggest that we consider the possibility of including other causes of T in the model, or that T is subject to large sources of random variation (i.e., "noise"), or both. Answers to the second question provide

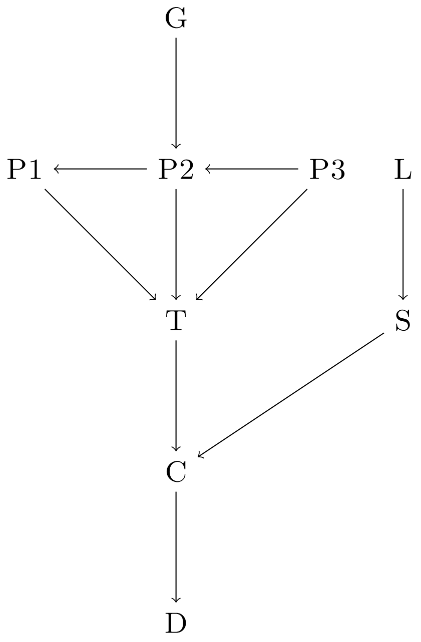

Figure 3.1: DAG illustrating Markov blanket of $\mathrm{T} = \{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}, \mathrm{C}, \mathrm{S}\}$.

estimates of direct (i.e., partial, but with a causal connotation) effects on the expected value of the target if the value of each cause were individually to change by one unit and other causes are held fixed. However, neither question directly addresses the impact hypothetical, ideal interventions have on a target. For that, we consider a second notion of "causal predictability."

3.2.2 Notion #2: Predictive Models Specified/Estimated Post-Intervention

We now consider the idea that we might specify and estimate predictive models after hypothetical interventions on one (or more) cause(s) of a target variable. Suppose we parameterized the DAG in Figure 3.1 by making each variable a linear function of its parents (with each edge in the graph parameterized by an edge coefficient corresponding to an appropriate regression coefficient) and an independent, normally distributed error term, following the common approach of structural equation modeling, or SEM (Bollen 1989).

Assuming each independent error term has mean 0, we estimate such a model from sample data and calculate its implied covariance matrix.[2] At this stage, the sub-matrix of the *pre-intervention* implied covariance for the target and its causes can be used to specify a regression model to answer questions with respect to our first notion of causal predictability; from data we estimate a regression function that provides $\hat{r}(\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}\}) = E(\mathrm{T}|\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}\})$. This estimated regression function is used to calculate predicted values for T conditional on values of variables P1, P2, and P3.

Next, we must consider the effects of hypothetical interventions (possibly more than one simultaneously) on causes of a target. For the causal structure of the DAG in Figure 3.1, we might consider an ideal (i.e., "surgical") intervention (Pearl 2009) on P2 that eliminates edges from both P1 and G into P2. Suppose we make the simplifying assumption that the hypothetical intervention sets the distribution of (the error term of) P2 to a Normal distribution with some mean and variance (later, we assume the standard Normal distribution: mean = 0 and variance/standard deviation = 1). Re-estimating the structural equation model given this intervention, we calculate the *post-intervention* implied covariance matrix. The relevant post-intervention sub-matrix can then be used to specify a statistical (e.g., linear regression) model, providing a corresponding estimated regression

---

[2] Comparing an estimated structural equation model's implied covariance matrix to the sample covariance matrix is one way in which to judge the overall "goodness of fit" of a structural equation model (Bollen 1989).

function for $\hat{r}(\{\text{P1}, \text{P2}, \text{P3}\}) = E(\text{T}|do(\text{P1}, \text{P2}, \text{P3}))$.[3] We can then ask the same questions we ask about our first notion of causal predictability (e.g., calculated predicted values of T) in the context of a post-intervention, estimated statistical model, providing us insight into the effects of hypothetical interventions on our ability to account for variability in a target and the partial/direct effects of a unit change in value, post-intervention, on causes of a target.

From these ideas and the idea of *total* effect (having only considered *partial/direct* effects so far), we describe three types of metrics to quantify these notions of causal predictability.

3.3 Three Types of Metrics for Causal Predictability

We consider three broad, but not exhaustive, categories or types of metrics for causal predictability, combining notions of predictability with different ways to specify statistical models and/or estimate them based on information about causes of a target. Since we learn a causal model over a particular set of variables, we associate a given measure/metric with the set of variables from which we learned a causal model. We thus have a basis upon which to compare a particular set of variables to other sets of variables. We provide concrete metrics that we use in practice, and suggest several possibilities for future work, in §3.6.

3.3.1 Naïve Fraction of Variability / Goodness of Fit

The first type of metric for causal predictability combines our first notion of predictability with our first notion of model specification from causal information. We consider measures or metrics of the fraction of variability explained by (or goodness of fit of) a statistical model, calculated from models specified pre-intervention to include only predictors that are inferred, direct causes of a target. We call this type of metric one of "naïve" causal predictability because models are specified according to inferred causes of a target, but models are specified pre-intervention.[4] Thus, we arrive at

---

[3] We use Pearl's (2009) *do* operator to denote that P1, P2, and P3 have been subjected to an intervention as opposed to mere conditioning. Spirtes, et al. (2000) would denote this by $E(\text{T}||(\text{P1}, \text{P2}, \text{P3}))$.

[4] Despite negative connotations associated with "naïve," we seek to emphasize that this is not a notion that invokes intervention, while still reflecting an important aspect of causal predictability. The notion might just as well be described as the "natural," or

an estimate of the sample variability in a target explained by observing its inferred causes. However, such metrics do not provide information about the extent to which we can explain sample variability in a target (or estimate models with good fit) following interventions on inferred causes.[5]

3.3.2 (Post) Intervention Fraction of Variability / Goodness of Fit

The second type of metric for causal predictability combines our first notion of predictability with our second notion of model specification based on causal information. These metrics are calculated from statistical models specified post-intervention, where such interventions are understood to be hypothetical (i.e., the product of appropriate re-calculation of implied covariance matrices from estimated, manipulated structural equation models, as explained in §3.2.2).

We suffer from the drawback that we must appropriately set the distribution of variables upon which we hypothetically intervene. In some application domains, we might have background knowledge for reasonable settings of these distributions, but in other cases we must make relatively arbitrary simplifying assumptions about manipulated distributions (i.e., setting variables upon which we intervene to the standard Normal distribution).

3.3.3 Total Effects

The third type of causal predictability metric corresponds to the second notion of predictability: the estimated change in expected value of a target from a change in the value of one (or more) of its causes. So far, we have only discussed partial effects of a change in value of a predictor on a target (holding all other predictors fixed). If the predictor variable is a cause of the target, then this partial effect is best described as its direct effect. For purposes of describing the effects of interventions on a cause of a target, it is also important to know its total effect, the sum of a cause's direct and indirect effects on the target. This value provides us a better estimate of the impact we have on a target by a particular intervention.

---

simply "pre-intervention," fraction of variability or goodness of fit.

[5] except possibly for the case in which individual interventions set each of a target's causes to its distribution in the sample (i.e., when variables are set by intervention to their "natural" distribution)

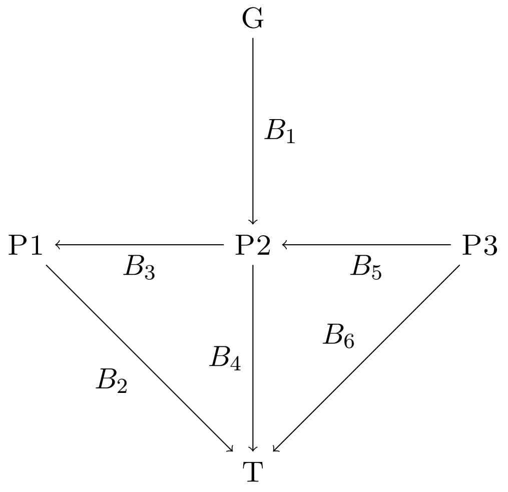

Figure 3.2: Parameterized sub-graph of Figure 3.1 DAG; independent error terms for each variable are omitted.

Direct, Indirect, and Total Effects in Regression Models and SEMs

Figure 3.2 is a parameterized sub-graph of Figure 3.1 (that omits an independent, Normally-distributed error term for each variable) and defines a linear structural equation model. Consider the structural equation for T:

$$\mathrm{T} = B_2\mathrm{P1} + B_4\mathrm{P2} + B_6\mathrm{P3} + \epsilon_{\mathrm{T}}.$$

As we have noted, edge parameters in such a structural equation model correspond to appropriate regression coefficients. Thus, if we regress T on $\{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}\}$, the ordinary least squares parameter estimate corresponding, for example, to P2 (call it $\hat{B_4}$) is an unbiased estimate of the true, population value of the parameter (i.e., the true partial effect) B_4. We predict that, holding P1 and P3 fixed, a unit change in P2 changes the expected value of T by $\hat{B_4}$. In standard regression parlance, this is the partial effect of P2 on T holding P1 and P3 fixed. Since P2 is also a direct cause of T, it is appropriate to refer to this as the direct effect of P2 on T; if we hold P1 and P3 fixed while intervening to change P2s value by one unit, we predict the same change in the expected value of T because we have an unbiased estimate of the strength of the direct causal link between P2 and T ($\mathrm{P2} \rightarrow \mathrm{T}$).

Inspecting the graph of Figure 3.2, P2 is both a direct and indirect cause of T; there are two directed paths from P2 to T: one direct (P2 → T) and one indirect (P2 → P1 → T). Accordingly, an intervention on P2 for which we do not hold other variables fixed (especially P1) will have an effect beyond the direct effect of P2 on T. Remember that, for additive parameterizations, the total effect is the sum of direct and indirect effects of one variable on the other:

$$\text{total effect} = \text{direct effect} + \text{indirect effects}$$

For linear models, the total effect corresponds to the sum of the products of edge coefficients over all directed paths, including a one-edge direct path if it exists, between two variables. We provide two sample calculations of total effect (TE) from the structure and edge coefficients/parameters of Figure 3.2.

The direct effect of P2 on T is B_4, the edge coefficient for the direct link between the two variables. To this, we add the product of coefficients along the other directed path from P2 to T (P2 → P1 → T), $B_3 B_2$, so:

$$TE(\text{P2}, \text{T}) = B_4 + B_3 B_2$$

There are two directed paths from P3 to T that pass through at least one intermediary variable, so there are two indirect effects to add to the direct effect of P3 on T, B_6:

$$TE(\text{P3}, \text{T}) = B_6 + B_5 B_4 + B_5 B_3 B_2$$

Parameters for each of the edge coefficients in a structural equation model can be estimated by various means to calculate estimates of total effects of variables in the model on a target. To estimate the total effect of a variable X on a target T using a linear regression model, we regress T on the set of predictors including only X and the parents of X (assuming we have inferred a DAG causal structure). Assuming we specify the model from correct causal structure, the least squares parameter estimate for X is an unbiased estimate of the true total effect of X on T.

Metrics Based on Total Effects

Investigators' goals inform a variety of metrics based on total effects of variables. One might develop metrics that consider total effects of multiple, simultaneous interventions on causes of a target, or alternatively only the total effects of individual interventions on causes of a target. The latter is appropriate for plausible, constrained situations in which an investigator seeks to decide upon a single intervention that will be most causally efficacious (i.e., have the greatest "impact" on the target); we focus on these situations. For other purposes, metrics might be calculated, for example, according to the minimum or maximum value of the total effect from a single intervention over a set of causes of target. The metric we deploy involves considering the maximum (expected) total effect for a single intervention on a cause of a target, leading us to the variable on which we should intervene to get the biggest "impact" on a target.

Pre-Intervention vs. Post-Intervention Total Effect

We provide two types of metrics for fraction of variability or goodness-of-fit, those that are "naïve" (pre-intervention) versus post-intervention, but we do not bifurcate total effect metrics. Since we restrict ourselves to interventions on a single variable, we need not split total effect metrics into pre-intervention and post-intervention. Consider, for example, our calculation of the total effect of P2 on T in Figure 3.2. If we were to intervene on P2 and "break" (i.e., by "surgical" or ideal intervention) the edges from G and P3 into P2, there is no change in the total effect of P2 on T. For multiple, simultaneous interventions, we would have to split these types of metrics into pre-intervention and post-intervention varieties. For example, the total effect of P3 on T would change given an intervention on P2 (as the edge from P3 to its effect P2 would be broken). We do not consider total effects of multiple simultaneous interventions.

3.4 How to Deploy Metrics to Judge "Quality" of a Set of Variables

Having explored several types of metrics to assess causal predictability for a given set of variables, we examine how to deploy specific metrics given causal structures we infer from data.

3.4.1 Relaxing Causal Sufficiency

So far, we have made reference only to DAGs. If we assume there are no unmeasured, common causes of measured variables, we generally learn a pattern from data, representing an equivalence class of DAGs, as causal structure is (almost always[6]) underdetermined by observational data. However, in realistic (social) scientific and policy settings, generally there are unmeasured common causes of measured variables. In these cases, we learn a Partial Ancestral Graph (PAG), perhaps by using the FCI algorithm. This increases the computational complexity of the problem of considering causal structures compatible with data. We must combine ways to "count" specific causal structures represented by a PAG with metrics we propose to adjudicate sets of variables.

3.4.2 Infer PAG from Set of Measured Variables

In Chapter 1, we introduced the FCI algorithm to learn (aspects of) causal structure over measured variables allowing for unmeasured common causes. In general, infinitely many underlying causal structures including latent variables are compatible with a PAG. Just as patterns represent equivalence classes of DAGs, PAGs represent equivalence classes of Maximal Ancestral Graphs (MAGs). Statistical methods for parameter estimation and principles of causal reasoning with MAGs have been developed (Spirtes and Richardson 2002; Zhang 2006, 2008a), but rather than work with the formalism of MAGs, we explore assumptions that permit us to consider sets of DAGs (including those with latent variables) that are compatible with PAGs. Importantly, while FCI can infer PAG causal models in the presence of selection bias, in what follows we assume there is no selection bias.

3.4.3 Calculate Metrics As Expectations Over Causal Models Represented by PAG

Particular DAGs represented by a PAG are likely to count different variables as causes of the target variable. Thus, they are likely to have different values for each causal predictability metric. For each set of variables and corresponding inferred PAG, we can calculate an expected value for each metric over the set of represented DAGs we choose to consider. The expected value of a metric M over a set of DAGs D can be calculated as

$$E(M_D|data) = \textstyle\sum_{d \in D} M(d|data)P(d)$$

corresponding to the sum, over a set of represented DAGs (D), of a metrics value over each DAG d given data (i.e., a model specified according to d and estimated from data over constructed variables) multiplied by the

---

[6] especially for cases of multi-variate linear, Normal distributions

probability of the particular DAG being the "correct" DAG over a particular set of measured variables.

This requires the investigator to provide prior probabilities for causal structures represented by a PAG. For example, the investigator might expect the true causal structure to have certain characteristics that skew the distribution of structures represented by a PAG. This is a prime point of entry for background and domain knowledge in our procedure. In the simulation studies in Chapter 4 and the case study of Chapter 5, we use a uniform distribution over causal structures represented by PAGs rather than impose constraints from background knowledge.

This also requires that for each PAG we provide the set of causal structures D over which we calculate this expectation. In the next section we consider three "counting" schemes whereby we provide sets of causal graphs represented by a PAG.

3.5 Counting Schemes for (Super) DAGs Represented by a PAG

In this section, we present three "counting schemes" for DAGs, as well as for DAGs we introduce that include latent variables, represented by a PAG, noting important limitations of each scheme. We explore their practical import and performance with respect to tracking true values of metrics we propose in simulation studies in the next chapter. Recall from Chapter 1 that edges in a PAG represent ancestral causal relationships among measured variables; there are three types of marks for edge endpoints in PAGs and four edge possibilities, as we assume there is no selection bias. For a review of the interpretation of edges in PAGs, see §1.2.3.

3.5.1 Scheme #1: Pattern counting scheme: ignore possible latent confounders

The simplest possible way to generate a set of DAGs that are represented by a PAG is to reduce the problem to that of generating the set of DAGs represented by an inferred pattern, ignoring the possibility of latent confounding. For example, there are three DAGs represented by the pattern X − Y − Z:

1. X → Y → Z
2. X ← Y ← Z

3. $X \leftarrow Y \rightarrow Z$

We deploy the PC (or GES) algorithm on data for a set of measured/constructed variables and a target, rather than FCI. For relatively small numbers of variables and sparse graphs, it is straightforward to generate the set of DAGs represented by a pattern[7]. Given its computational tractability and (perhaps despite) the relative simplicity of this counting scheme, we consider, in the next chapter, how model averaging for metrics according to this counting scheme fares in comparison to more computationally intense counting schemes that "throw out" less information.

3.5.2 Scheme #2: Heuristic PAG counting schema

A variety of heuristics can be developed to generate DAGs represented by a PAG, iterating over possible orientations of PAG edges. We employ a simple heuristic that takes a relatively narrow reading of PAG edges to generate DAGs represented by a PAG. For example, with this heuristic we make no attempt to account for the possibility that two variables have both a direct causal relationship and a confounding common cause.

First, we generate the set of all acyclic directed mixed graphs, acyclic graphs in which both directed and bi-directed edges may occur (Richardson 2003), according to every possible setting of edge marks in a given PAG. We then eliminate graphs that create unshielded colliders not already present in the given PAG. To limit our heuristic to DAGs, we drop bi-directed edges from the acyclic, directed mixed graphs that remain and obtain a set of DAGs.

Consider the PAG X o−o Y o−o Z. In this PAG Y is a non-collider. We start by generating the set of all graphs with directed and bi-directed edges (mixed, directed graphs) according to the six possible ways in which we can set edge marks in the PAG (allowing only one edge between any pair of variables):

1. $X \leftrightarrow Y \leftrightarrow Z$
2. $X \leftrightarrow Y \rightarrow Z$
3. $X \leftrightarrow Y \leftarrow Z$
4. $X \rightarrow Y \leftrightarrow Z$

---

[7] We return to the problem for generating an appropriate set of DAGs from a pattern for a large number of variables shortly.

5. X → Y → Z
6. X → Y ← Z
7. X ← Y ↔ Z
8. X ← Y → Z
9. X ← Y ← Z

Inspecting this set of mixed, directed graphs, we find that graphs (1), (3), (4), and (6) orient colliders at Y and are thus not represented by the PAG under consideration. In the five remaining directed mixed graphs we adopt the heuristic that disallows edges between two variables that share a common cause (i.e., dropping bi-directed edges). We interpret bi-directed edges to indicate that two variables share a latent common cause and have no direct causal link, so we resolve the problem of latent confounding by dealing with it implicitly, assuming no connection between two confounded variables. If, for example, X ↔ Y is an edge in a PAG (and thus in every mixed, directed graph we generate from the PAG), every member of the set of DAGs we generate will have no edge from X to Y. By this counting scheme, the following set of DAGs is represented by this PAG (i.e., are those over which we should average when calculating metrics for causal predictability):

1. X Y → Z
2. X → Y → Z
3. X ← Y Z
4. X ← Y → Z
5. X ← Y ← Z

There are several shortcomings of this "PAG Heuristic." Notably, if X → Y is an edge in a PAG, then X → Y is an edge (i.e., X is a direct cause of Y) in every DAG we generate. This is not necessarily a correct inference given the ancestral interpretation of edges in a PAG. We explore the extent to which this simplification will harm performance of metrics calculated with this counting scheme in Chapter 4.

Like the iteration of DAGs compatible with a pattern, this heuristic is not particularly computationally intense given a relatively small number of variables and sparse PAG connectivity.

3.5.3 Scheme #3: Exhaustive counting of "Super DAGs" including possible confounders

For our third counting scheme, we introduce "Super DAGs." We define Super DAGs as causal graphs over both measured and latent variables to provide more realistic representations of causally insufficient systems. We allow that every pair of measured variables may have one latent variable as a common cause. This simplifying assumption about the nature of latent confounding allows us, for small numbers of measured variables, to generate the set of Super DAGs represented by a PAG. Our third counting scheme generates the set of "Super DAGs" that are represented by a PAG we infer over a set of measured or constructed variables. By this counting scheme, we explicitly account for latent variables, in contrast to how we avoided accounting for latent confounding directly (if at all) in the previous two counting schemes. This counting scheme is the most principled of the three counting schemes we consider.

To illustrate, Figure 3.3 provides the set of nine Super DAGs from which we can infer the PAG X o−o Y o−o Z. In the limited case of this PAG on three measured variables, we go from considering five DAGs with the PAG Heuristic to nine Super DAGs. In practice, with larger numbers of measured variables, the difference is more substantial.

While Super DAG "counting" provides a principled means of surveying a large set of causal structures represented by a PAG, exhaustive counting is computationally expensive and intractable for more than four or five variables. Simulations in Chapter 4 pursue the exhaustive counting strategy for four measured variables (involving up to 6 latent variables); we compare results of model averaging for metrics for causal predictability over exhaustive sets of Super DAGs to averaging over sets of DAGs generated by the pattern counting scheme and the PAG Heuristic. We introduce these specific metrics for causal predictability in the following section and consider several practical problems (and partial solutions), especially for the Super DAG counting scheme.

3.6 Metrics in Practice

We provide three concrete metrics, each of which corresponds to a particular notion of causal predictability. Alternative metrics are legion. We consider one proposed alternative metric from related literature in §3.7.2; future work should explore alternatives.

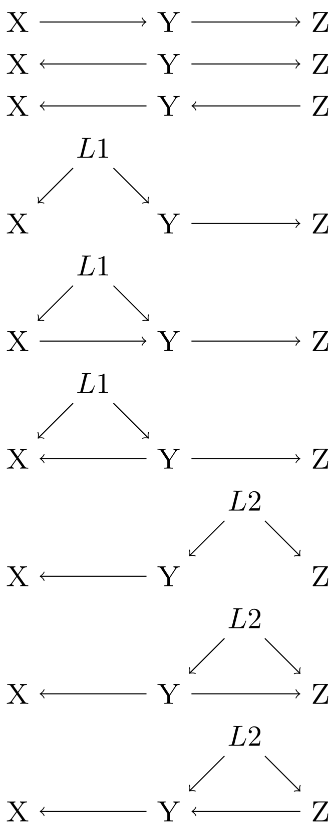

Figure 3.3: Super DAGs represented by PAG X o−o Y o−o Z.

3.6.1 Maximum Expected Naïve (Pre-Intervention) "Causal R^2"

To capture the notion of naïve or pre-intervention fraction of variability in practice, we calculate the R^2 value for linear regression models specified according to each DAG or Super DAG represented by a PAG, each of which may count different variables as direct causes of the target; for each compatible causal graph, we regress the target variable on its (measured) direct causes in that graph, calculate the model's R^2 value, and call it the model's "causal R^2" value. We seek the set(s) of variables from which we infer a PAG that maximizes the expectation of this value, calculated according to a chosen DAG/Super DAG counting scheme. In the next two chapters, we consider each DAG/Super DAG generated by a counting scheme to be equally probable, so the expectation is the average over these graphs.

3.6.2 Maximum Expected (Post-) Intervention "Causal R^2"

We capture the notion of intervention fraction of explained variability in essentially the same way as the naïve notion. However, we specify "post-intervention" regression models based on the direct causes of the target in each DAG or Super DAG represented by a PAG. To do this, we make the simplifying assumption that we simultaneously intervene on each inferred cause in a compatible causal graph to give it a standard Normal distribution (mean = 0, standard deviation = 1). We estimate a structural equation model from data according to the structure of each DAG or Super DAG, providing us with an implied covariance matrix. We manipulate the estimated, implied covariance matrix according to our simplifying intervention assumption, and from this manipulated covariance matrix we specify a regression model for the target with the manipulated, direct causes as predictors. We seek the set(s) of variables that maximize the expectation of this intervention causal R^2 over the set of DAGs or Super DAGs counted for a PAG. Again, in the next two chapters, given equally probable DAGs or Super DAGs, the expectation is the average over these graphs.

3.6.3 Maximum Expected Total Effect

We devise a metric to capture the single variable that we expect to have the greatest impact on the target were we to intervene and changes its value by one unit . Over a set of DAGs or Super DAGs represented by a PAG, we consider the expected total effect of each variable that is a direct cause of the target in at least one member DAG or Super DAG. The expected value of each variable's total effect is the sum, over all DAGs or Super DAGs represented by a PAG (according to a particular counting scheme), of the product of the variable's total effect on a target for the DAG or Super DAG and the probability of that DAG or Super DAG. For a set of (Super) DAGs D represented by a PAG, for each possible cause X_i of T, we assume we have estimated an appropriate linear regression or structural equation model such that:

$$E(TE(\mathrm{X}_i, \mathrm{T})) = \textstyle\sum_{d \in D} TE_d(\mathrm{X}_i, \mathrm{T})P(d)$$

Having calculated the expected value for each possible cause X_i of T, we take the variable that maximizes this expected value; the set(s) of variables in which it appears are adjudicated "best" by this metric.

Since an estimated linear regression or structural equation model is required to calculate R^2 values and total effects TE_d for each represented DAG or Super DAG d, respectively, we turn to practical considerations for calculating proposed metrics and model averaging over DAGs and Super DAGs represented by an inferred PAG.

3.7 Counting Schemes and Metrics in Practice

Pattern and heuristic PAG schemes "count" DAGs over measured variables; without latent variables, estimating a statistical model to calculate any of our metrics is relatively simple, especially compared to our "Super DAG" counting scheme. For expected naïve causal R^2 and maximum expected total effect, we need only estimate regression models from data according to the structure of DAG, as naïve causal R^2 is calculated from predicted values of a regression model with all direct causes in a particular DAG as predictors, while the total effect of a variable X on T is the parameter estimate (i.e., coefficient) from the regression of T on X and PARENTS(X) in a DAG. To calculate intervention causal R^2, we first estimate a structural equation model according to the structure of a DAG. We then manipulate the covariance matrix implied by this estimated SEM such that each direct cause has been set by an ideal intervention to the standard Normal distribution. Such calculations are relatively simple, and both regression and structural equation model estimation are easily carried out with freely available software (e.g., TETRAD). The Super DAG counting scheme is more complicated. We consider details for calculating these metrics for each counting scheme in practice and consider some related work.

3.7.1 Pattern Scheme

We propose in our first counting scheme to average over DAGs in the equivalence class represented by a pattern, discarding any information we have about latent confounding. For relatively small numbers of variables and sparse causal structures, it is straightforward to provide the set of DAGs represented by a pattern. Assuming linear, multi-variate Normal distributions, we use linear regression models specified according to the structure of a DAG to calculate our (naïve) causal R^2 metric and total effects for each possible cause of a target variable. Intervention causal R^2 can be calculated by specifying a linear regression from an appropriately manipulated implied covariance matrix.

While calculating (causal) R^2 values over regressions specified according to the structure of a DAG is relatively novel as a metric to judge the "quality" of a variable set in its support for causal inferences, calculating total (i.e., causal) effects of variables over DAGs represented by a pattern has been pursued as a means for dimensionality reduction in causally sufficient systems. We briefly consider this work.

3.7.2 Using Patterns for Dimensionality Reduction in Causally Sufficient Systems

While we seek to judge the quality of sets of constructed variables in application domains without causal sufficiency, Maathuis, et al. (2009) estimate lower bounds for variables' total (causal) effects in causally sufficient systems. They deploy their methods on data from biological systems in which causal sufficiency is a reasonable assumption and use estimated lower bounds as a means of dimensionality reduction.

Maathuis, et al. (2009) consider the problem of discovering genes that have a "knock-down" (intervention) effect on a particular phenotype, in their case study the production of riboflavin in the bacterium *Bacillus subtilis.* They consider observational expression data for thousands of genes over "essentially" the entire genome of this bacterium as co-variates for the continuous-valued response of riboflavin production.

To find those genes that "score" highest with respect to intervention/causal effects, they consider causal effects for each co-variate, estimated by linear regression models specified according to each DAG in a set of DAGs in the equivalence class represented by a pattern[8]. Importantly, their algorithm

---

[8] Maathuis, et al. refer to patterns as completed partially directed acyclic graphs (CPDAGs).

uses local graph information to avoid iterating over the full equivalence class of DAGs represented by a pattern, but its result contains the correct set of unique causal/total effect values for each co-variate for the equivalence class. The algorithm is thus useful when the underlying, unknown causal DAG is sparse but contains many variables. Co-variates are scored according to the minimum value in the resulting multi-set of causal effects (i.e., the estimated lower bound of each co-variate's causal/total effect). Presumably, genes with the greatest lower bound on their total effects are those toward which we should direct attention in future experiments. By considering only those genes/co-variates above a particular threshold, we reduce dimensionality of our problem and narrow the space of variables (genes) on which to target future interventions.

The approach of Maathuis, et al. (2009) to dimensionality reduction is similar in spirit to our approach to judging the "quality" of sets of variables for causal inference. They devise feasible algorithms to generate a set of DAG causal structures indistinguishable from observational data, a "counting scheme" different than those we explore, as we consider every DAG compatible with a pattern for graphs with a small number of variables. They also provide a different type of metric to score individual variables based on lower bounds for total/causal effects (rather than expected values/averages). This type of metric might be adopted given the risk/reward profile of a scientific investigator or policy maker concerned with the problem of constructing variables to support causal inferences.

Maathuis, et al.'s (2009) application is one in which causal sufficiency is a reasonable assumption; they consider nearly the entire genome of the bacterium in question. It remains unclear that the type of correctness results they provide are forthcoming (or possible) for causally insufficient systems, as total effects are frequently not identifiable without simplifying assumptions, as we soon discuss. Nevertheless, their ideas can be adapted to our purposes in future work, and the pragmatic impact of relaxing the causal sufficiency assumption can be explored via simulation studies. Indeed, we explore exactly this issue for our counting schemes and metrics in the following chapter.

3.7.3 Comparison of Heuristic PAG and Pattern Counting Schemes

Our proposed Heuristic PAG counting scheme is similar to the pattern scheme, but the former implicitly accounts for possible latent confounding by generating DAGs without edges between variables for PAG mark orientations that entail that two variables shared an unmeasured common cause. By accounting for latent confounding, the PAG Heuristic scheme "counts" more DAGs over which to average than the Pattern scheme. Beyond this, there are no significant differences in how we handle the calculation of causal predictability metrics under these two counting schemes.

3.7.4 "Super DAG" counting scheme

If the exhaustive "counting" of DAGs in the equivalence class represented by a pattern is computationally intractable for patterns defined over large numbers of variables, the problem of generating Super DAGs compatible with a particular PAG is even worse. We provide a simplifying assumption that each pair of variables in a PAG can have only one unique unmeasured common cause, and yet the exhaustive generation of Super DAGs (and their estimation) still becomes computationally intractable quickly.

Domain knowledge might provide other plausible assumptions; perhaps, for example, we place a limit on the total number of unmeasured common causes there may be in the true, underlying Super DAG generating causal structure. This will at the very least extend the feasibility of this counting scheme to somewhat larger sets of measured (or in our case, constructed) variables. Future work should explore simplifying assumptions under which iterating over Super DAGs represented by a PAG becomes computationally tractable for larger numbers of measured or constructed variables while still achieving good performance in estimating causal predictability metrics. One might also consider counting Maximal Ancestral Graphs in the equivalence class represented by a PAG.[9]

We now illustrate how we estimate statistical models and calculate causal predictability metrics from Super DAGs in the simulation studies that comprise Chapter 4.

Transformation of Super DAG to Directed, Mixed Graph

Starting with a given Super DAG (containing both measured and latent variables) in the set represented by a given PAG, we seek a simpler graphical object from which we can estimate a statistical model, in our case a linear structural equation model. We transform the given Super DAG into

---

[9] Richardson and Spirtes (2002) provide a method to generate the "canonical DAG" of a MAG. Future work should consider: (a) iterating over such DAGs from MAGs represented by a PAG, or (b) another way of "counting" MAGs to calculate metrics for causal predictability.

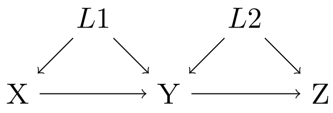

Figure 3.4: A Super DAG that is converted to directed, mixed graph in Figure 3.5 for SEM estimation

a directed, mixed (acyclic) graph on only measured variables. We eliminate latent variables from a Super DAG by introducing bi-directed edges between any two variables that share a latent variable as a common cause. All directed edges linking measured variables in a Super DAG remain in the mixed, directed graph. We allow that both a directed and bi-directed edge can simultaneously obtain between a pair of nodes, corresponding to variable pairs in the given Super DAG that have both a direct causal link and share an unmeasured common cause. For example, the Super DAG of Figure 3.4 is converted to the directed, mixed graph of Figure 3.5.

We now consider estimating structural equation models from graphs like those in Figure 3.5 to carry out our proposed Super DAG counting scheme in practice.

Estimation of SEMs from directed, mixed graph

From a Super DAG we have obtained a directed, mixed graph without cycles, possibly having both bi-directed and directed edges between two variables. Such graphs are common in the framework of structural equation models; bi-directed edges represent correlated error terms (e.g., due to an unmeasured common cause). Parameters associated with bi-directed edges represent error co-variances. Appropriately, we have framed our discussion of causal predictability metrics in terms of models like SEMs and now have a graph from which we can specify and estimate a statistical model with standard techniques. We can then calculate causal predictability metrics from a Super DAG and model average over an appropriate set represented by a PAG.

Difficulties with SEM estimation

Of course, estimating structural equation models in this way is not without its hazards. Since we estimate a SEM on data over measured variables, in many cases Super DAGs represented by a PAG will be converted into directed, mixed graphs that when parameterized for estimation (e.g., Figure

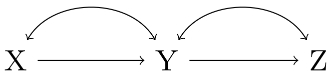

Figure 3.5: Directed, mixed graph resulting from Super DAG in Figure 3.4

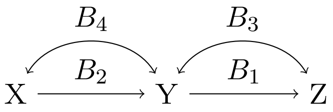

Figure 3.6: Parameterized SEM (correlated error terms omitted) specified from directed, mixed graph of Figure 3.5

3.6) have zero or negative degrees of freedom and are thus unidentifiable. Rather than simply discard all such unidentifiable parameterized models, in simulation studies we make a simplifying assumption to achieve identifiability. If any parameter on any directed path from any cause of the target variable to the target is not identified, then for any two variables that have both a directed and bi-directed edge between them, we constrain estimation such that the two parameters are equal. For Figure 3.6, we would estimate the model with the constraints that $B_1 = B_3$ and $B_2 = B_4$. This constraint will make models identifiable, but we still may encounter models for which we cannot stably estimate parameters due to local optima in estimation. Also, this constraint is arbitrary; other constraints ought to be explored or justified by domain knowledge.

Importantly, we only seek to make sure that all parameters/coefficients along directed paths from direct causes of the target to the target are estimable so that we can calculate causal predictability metrics. However, if one or more parameters along such paths do not have stable estimates, we discard the SEM and its corresponding Super DAG from consideration. Thus, generally, we do not (cannot) calculate an expectation, in practice, over every possible Super DAG represented by PAG.

Using SEMs to Calculate Metrics

After estimating a SEM from data, calculating each of the proposed causal predictability metrics is straightforward. The total effect of each direct cause of the target is just the product of estimated coefficients along each directed path from the direct cause to the target; naïve causal R^2 is calculated from the un-manipulated, implied covariance matrix of the estimated SEM, and the intervention causal R^2 (for the "standardized" intervention setting every direct cause by surgical intervention to a standard Normal distribution) is calculated from simple manipulations of the estimated SEM's implied covariance matrix.

3.8 Summary

In this chapter, we introduced broad notions of the extent to which we can predict a target from a statistical model and tied those notions to different ways in which statistical models can be specified according to causal models we can learn from data. Since learning causal models from observational data generally does not result in a single causal model, we introduced several "counting schemes" to deal with structure uncertainty represented by a PAG. We considered several shortcoming of the different ways in which we might handle this uncertainty before introducing concrete metrics of causal predictability calculated as expectations over models represented by PAG causal structures. We then consider various practical issues that will arise when deploying these metrics and counting schemes in simulation and with real-world data. In the following chapter, we pursue simulation studies to determine the extent to which different ways of "counting" causal structures represented by a PAG provide accurate estimates of true values of proposed causal predictability metrics.

We then consider a simulation study incorporating these metrics to adjudicate the "best" among sets of constructed variables. Further, we consider whether sets of variables that maximize causal predictability or expected causal control are the same that maximize prediction without concern for causation. An over-arching question for this latter simulation study is: does (should) our scientific inference goals (i.e., prediction vs. causal control) inform variable construction, or will the same set(s) of variables do equally well for each purpose?

Chapter 4

Causal Predictability, Structure Uncertainty, and Variable Construction: Simulation Studies

In this chapter, we explore, via simulation studies, aspects of causal predictability metrics and "counting schemes" for causal structures over which they are calculated. In §4.1 we determine the counting scheme that minimizes error for each causal predictability metric when data are generated from known generating processes. In §4.2, we deploy causal predictability metrics on problems of variable construction from "raw" data using counting schemes chosen in §4.1.

These two simulation studies provide evidence for the following conclusions:

- Conclusions #1 through #4 are summarized in Table 4.1. Conclusion #1, for example, provides the counting scheme (i.e., the PAG Heuristic) that best tracks the *variable* with maximum total effect when using the maximum expected total effect (M.E.T.E.) causal predictability metric. However, the Pattern counting scheme best tracks the true value of the total effect of the variable "chosen" by the M.E.T.E. causal predictability metric (for generating Super DAGs not represented by the complete PAG), whether or not that variable has the maximum total effect in the generating model.

#	**Causal Predictability Metric**	**Maximizing Counting Scheme**	**Section**
1	M.E.T.E. Variable	PAG Heuristic	4.1.3
2	Chosen M.E.T.E. Variable Total Effect Value	Pattern	4.1.4
3	Average Naïve Causal R^2	PAG Heuristic	4.1.5
4	Average Intervention Causal R^2	Super DAG	4.1.6

Table 4.1: Summary of Conclusions #1 through #4

- Conclusion #5: Overall, performance is better for inferred PAGs that are not complete compared to complete PAGs. In §4.2 and Chapter 5, we resolve to only consider constructed variables over which we learn PAGs that are not complete (i.e., not maximally qualitatively uninformative about causal structure).

From Conclusion #5, we establish a baseline criterion for judging constructed variables that any set of constructed variables over which we infer a complete PAG is discarded.

Understanding whether or not we must assume a target variable to be causally "final" (i.e., last in a temporal ordering or otherwise without effects among variables under consideration) is important for applying these methods in practice. The relationship between performance for counting schemes and whether the target variable has effects in the generating model is complicated.

- Conclusion #6: For qualitatively tracking the variable with maximum total effect, performance is far better when the target variable is "final" (i.e., has zero out-degree) (§4.1.3). Aside from the case of intervention causal R^2, when the inferred PAG is not complete, the target variable's status as final in the generating model generally improves performance.[1]

- Conclusion #7: Complex counting schemes work best when they are actually needed; performance usually improves as the number of la-

---

[1] When we infer a complete PAG, performance is better in cases with positive target out-degree. However, we resolve to discard sets of constructed variables over which we infer the complete PAG.

tent causes of the target variable in the generating model increases. (Appendix, Chapter 7)

In §4.2, we take up questions of variable construction, using causal predictability metrics to determine sets of constructed variables that "best" support causal inference. Three major conclusions are demonstrated in this section:

- Conclusion #8: Constructed aggregate variables that maximize causal predictability metrics generally are not the same as the aggregate variables in the data generating process. (§4.2.3)
- Conclusion #9: Constructed variables that maximize causal predictability metrics and those that maximize traditional prediction metrics (without respect to causal structure) are generally not the same, demonstrating that causal and predictive inference "come apart" with respect to variable construction. (§4.2.3)
- Conclusion #10: Inferred causal structures (i.e., inferred PAGs) for variables that maximize metrics of causal predictability in many cases do not match the PAG one expects given the structure of the data generating process. (§4.2.4)

We now turn to presenting data that support these conclusions. Within each section, we note explicitly where we take evidence and support for each conclusion to occur.

4.1 Performance of Counting Schemes for Causal Predictability Metrics

In this section, we consider the extent to which three (Super) DAG counting schemes allow us to estimate the true value of metrics for causal predictability when we learn a PAG from data generated by ground truth Super DAGs. We match each causal predictability metric with the counting scheme that minimizes (relevant ways of quantifying) error. In §4.2, we use each causal predictability metric with its "best" counting scheme in a simulation study that involves the construction of aggregate variables from "raw" variables.

4.1.1 Simulation Regime

To generate data from known causal structures, we consider the space of Super DAGs on four measured variables (X1, X2, X3, and X4) and zero to six latent variables ($L1$ through $L6$), comprised of 35,295 Super DAGs. Our target variable is always X4. Conditional independence relations among the four measured variables in 18,639 (52.8%) of these Super DAGs are represented by the complete PAG on X1, X2, X3, and X4. The complete PAG is qualitatively maximally uninformative with respect to causal structure, and we treat these Super DAGs separately in our simulation study. We consider three strategies, including two different ways of pursuing the Super DAG counting scheme in cases in which generating Super DAGs are represented by the complete PAG:

- Simulation Strategy #1: Randomly sample generating Super DAGs from the set of 16,656 Super DAGs that are not represented by the complete PAG.
- Simulation Strategy #2: Randomly sample from the set of Super DAGs represented by the complete PAG and consider a relatively small subset of Super DAGs represented by the complete PAG when deploying the Super DAG counting scheme (i.e., a "non-exhaustive" Super DAG counting scheme).
- Simulation Strategy #3: Randomly sample from the set of Super DAGs represented by the complete PAG and consider all Super DAGs represented by the complete PAG when deploying the Super DAG counting scheme (i.e., an "exhaustive" Super DAG counting scheme).

In the first simulation strategy, we randomly sample 32,000 Super DAGs from the set of Super DAGs that are not represented by the complete PAG, treating each Super DAG as equally probable. We then instantiate a linear structural equation model, "the generating SEM," specified according to the structure of the "generating Super DAG," choosing edge parameters and error variances randomly.[2]

We sample data (n=1,000) from the generating SEM and calculate the true value of each causal predictability metric:

---

[2] We roughly use the default settings for randomly instantiating the parameters of a "standardized" linear SEM found in the TETRAD suite of software with positive edge coefficients and unit variances for each variable's error term.

- Naïve Causal R^2: We generate predicted values for X4 based on sampled values of the direct causes of X4 in the generating Super DAG and their corresponding edge parameters in the generating SEM. We then calculate the R^2 value comparing the predicted values to the true sampled values of X4.[3]
- Intervention Causal R^2: We manipulate the implied covariance matrix of the generating SEM such that each direct cause is (ideally) intervened upon, with distribution set as Normal(0,1). We then specify a regression model for X4 from the sub-matrix of this implied covariance matrix containing only X4 and its direct causes in the generating Super DAG.
- Maximum Total Effect (on X4): We calculate and store the individual total effect of X1, X2, and X3 on X4 from the parameters of the generating SEM, and determine the variable that has the maximum total effect. We allow that no variable might have a direct causal relationship with X4 and thus the maximum total effect could be zero.

Next, we deploy the FCI algorithm to learn a PAG over measured variables in the sampled data. From each inferred PAG, we consider each of three counting schemes introduced in the previous chapter: the exhaustive Super DAG counting scheme, the PAG Heuristic counting scheme, and the Pattern counting scheme. We generate DAGs or Super DAGs represented by the inferred PAG according to each counting scheme and calculate the average naïve causal R^2, average intervention causal R^2, and maximum expected total effect variable and value for each of the sampled 32,000 generating Super DAGs/SEMs (our instantiation of Simulation Strategy #1). We discard Super DAGs for which we encounter any estimation problems (e.g., unstable parameter estimates due to local optima in estimation) for any path from a direct cause of the target to the target in their estimated linear SEMs. Recall that for Super DAGs with an unidentifiable parameter along a directed path from a cause of X4 to X4, in which a pair of variables have both a direct causal link and a latent common cause, we constrain parameter estimates for the edge representing the direct causal link and a double-headed edge between the pair to be equal.

To consider the 18,639 Super DAGs that are represented by the complete PAG, we pursue the same simulation regime by two different strategies.

---

[3] One way to think of the true naïve causal R^2 value is as the intervention causal R^2 when the intervention is one in which each direct cause of the target is set to its distribution in the sample.

With one strategy we randomly sample 18,000 Super DAGs from among the 18,639 represented by the complete PAG, assuming each equally probable, but we only consider 700 randomly sampled Super DAGs with the "Super DAG" counting scheme. This is how we instantiate Simulation Strategy #2.

For Simulation Strategy #3, we randomly sample 400 Super DAGs from among those represented by the complete PAG, again assuming each equally probable, and we consider all 18,639 Super DAGs represented by the complete PAG in the "Super DAG" counting scheme.

We can then consider the extent to which different counting schemes approximate the true value of each causal predictability metric, as well as the extent to which we qualitatively find the (correct) variable with maximum total effect for each of these three strategies (and characteristics of underlying causal structures).

4.1.2 Stratification of Results

There are many possible ways to stratify the characteristics of generating Super DAGs and the PAGs that represent them. For each causal predictability metric, we begin by summarizing results over all sampled generating Super DAGs, and then stratify results by salient features of generating Super DAGs:

- Out-degree of the target variable in the generating Super DAG: Is the target variable a cause of other variables under consideration, or is it causally "last?"
- Number of latent causes of the target; see Appendix (Chapter 7).

Stratifications inform us as to how we might expect counting schemes to perform for particular causal predictability metrics given domain knowledge and/or assumptions about the underlying dynamics of systems we are trying to model.

4.1.3 Maximum Expected Total Effect: Qualitatively Tracking Correct Variable

We begin with the simplest result: How often do the counting schemes coupled with the Maximum Expected Total Effect (M.E.T.E.) metric choose the "true" variable with maximum total effect (or judge correctly that no variable has a total effect greater than zero)? This is a rough assessment

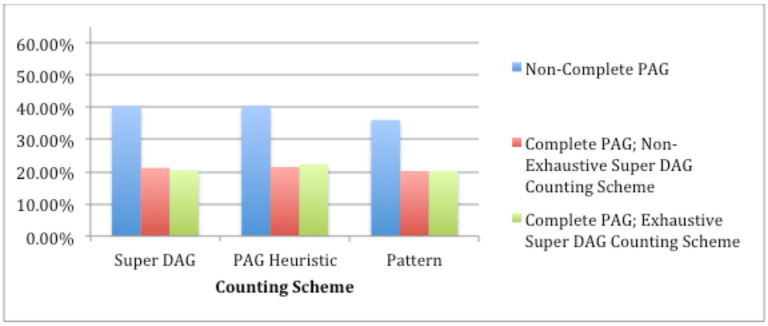

Figure 4.1: Percentage of Simulated Generating Super DAGs for which we recover the correct variable with Maximum Total Effect using the M.E.T.E. Causal Predictability Metric by Counting Scheme and Generating Super DAG structure; Summarizes Table 4.2

of the qualitative "correctness" (or error) of the three counting schemes for the M.E.T.E. metric.

Results for the three Super DAG simulation strategies are provided in Figure 4.1 and Table 4.2[4], as percentages (and counts) of instances in which there are "matches" between the chosen variable with M.E.T.E. and the true variable with maximum total effect. Note that, given four choices (X1, X2, X3, and "No Variable"), the probability of choosing the correct variable by chance, all other things equal, is 25%.

Overall, the PAG Heuristic counting scheme (narrowly) best tracks the individual variable with maximum total effect (cf. Conclusion #1). Notably, its performance in this task is similar to that of the "Super DAG" counting scheme, despite the greater complexity (and in one case exhaustiveness) of the latter approach. Further, performance for all counting schemes is better in cases in which the generating Super DAG is not represented by the complete PAG (support for Conclusion #5); in maximally uninformative cases (i.e., generating Super DAGs represented by complete PAGs), we perform poorly (indeed, worse than chance), but it is not clear that considering only a small subset (700 of 18,639) of Super DAGs represented by the complete PAG substantially hurts performance.

Figures 4.2-4.4 demonstrate that our ability to discover the correct vari-

---

[4] In Table 4.2, the column heading "All Super DAGs?" refers to whether the Super DAG counting scheme considers all Super DAGs represented by the PAG inferred from data simulated from the generating Super DAG.

Simulation Strategy				**Counting Scheme: Correct Count (%)**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	12,986 (40.57%)	12,998 (40.62%)	11,550 (36.1%)
2	Yes	18,000	No	3,816 (21.2%)	3,878 (21.5%)	3,361 (20.2%)
3	Yes	400	Yes	82 (20.5%)	89 (22.3%)	81 (20.3%)

Table 4.2: Overall Qualitative Match Counts for Maximum Expected Total Effect (M.E.T.E.) Variable and True Maximum Total Effect Variable.

able with maximum total effect is significantly influenced by whether our target is causally "final" (i.e., has zero out-degree). When the target has out-degree zero in the generating causal structure (Table 4.3), we choose the correct variable in the majority of cases in all simulation strategies. Again, the performance of the Super DAG and PAG Heuristic counting schemes are similar (nearly identical) across simulation regimes, and the Pattern counting scheme has roughly the same performance as the Super DAG and PAG Heuristic schemes in cases in which the generating Super DAG is represented by the complete PAG.

Contrast these results with those in Table 4.4, wherein the target variable is not causally "final," in which our only hope of performing (barely) better than chance are cases of generating Super DAGs not represented by the complete PAG on which we deploy the Super DAG or PAG Heuristic counting scheme. Performance is far worse than chance in cases in which the generating Super DAG is represented by the complete PAG. This indicates the importance of the target's status as last in the causal ordering of variables under consideration if we are to choose the variable with maximum total effect with the M.E.T.E. metric (cf. Conclusion #6).

The Appendix to this chapter (Chapter 7) provides data, for each causal predictability metric, stratifying generating Super DAG structures according to the number of latent causes of the target to provide support for Conclusion #7.

To summarize, the PAG Heuristic minimizes error in identifying the

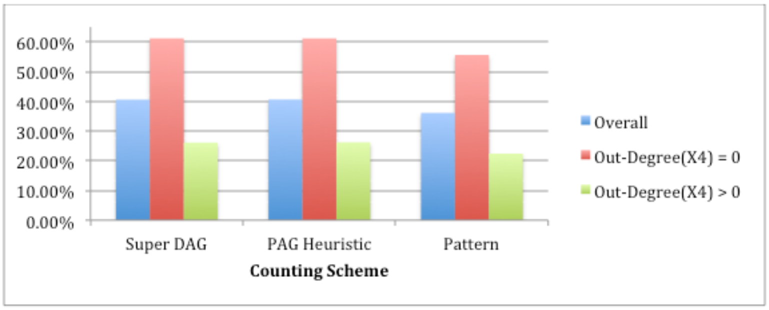

Figure 4.2: Qualitative Match Percentages for M.E.T.E. Variable and True Maximum Total Effect Variable when Generating Super DAG is Not Represented by Complete PAG by Target Out-Degree

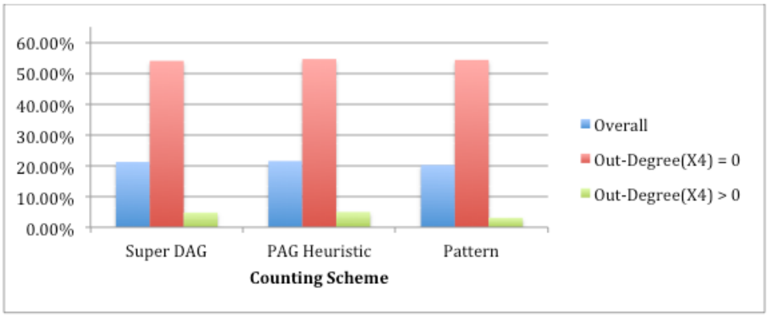

Figure 4.3: Qualitative Match Percentages for M.E.T.E. Variable and True Maximum Total Effect Variable when Generating Super DAG is Represented by Complete PAG with Non-Exhaustive Super DAG Counting Scheme by Target Out-Degree

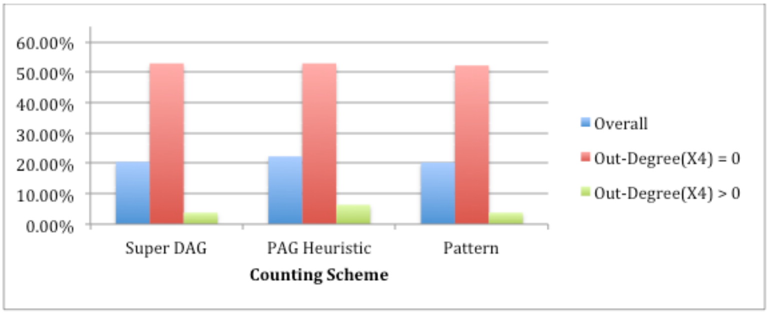

Figure 4.4: Qualitative Match Percentages for M.E.T.E. Variable and True Maximum Total Effect Variable when Generating Super DAG is Represented by Complete PAG with Exhaustive Super DAG Counting Scheme by Target Out-Degree

Simulation Strategy				**Counting Scheme: Correct Count (%)**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	8,075 (61.19%)	8,073 (61.18%)	7,337 (55.6%)
2	Yes	6,003	No	3,240 (54%)	3,279 (54.6%)	3,261 (54.3%)
3	Yes	136	Yes	72 (52.9%)	72 (52.9%)	71 (52.2%)

Table 4.3: Generating Super DAGs with Target (X4) Out-Degree = 0: Qualitative Match Counts for M.E.T.E. Variable and True Maximum Total Effect Variable.

Simulation Strategy				**Counting Scheme: Correct Count (%)**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	4,911 (26.1%)	4,925 (26.2%)	4,213 (22.4%)
2	Yes	11,997	No	576 (4.8%)	599 (5%)	370 (3.1%)
3	Yes	264	Yes	10 (3.8%)	17 (6.4%)	10 (3.8%)

Table 4.4: Generating Super DAGs with Target (X4) Out-Degree > 0: Qualitative Match Counts for M.E.T.E. Variable and True Maximum Total Effect Variable.

qualitatively correct variable with maximum total effect using the M.E.T.E. metric of causal predictability (Conclusion #1). Further, both maximally uninformative cases (i.e., those of generating Super DAGs represented by the complete PAG) and cases in which the target variable is not causally "final" pose a significant problem for identifying the correct variable with maximum total effect, supporting Conclusion #5 and Conclusion #6.

4.1.4 Maximum Expected Total Effect: Tracking Quantitative Value

Whether or not a counting scheme picks out the variable with maximum total effect in the underlying generating process, we are concerned that we track the total effect of whatever variable is determined to have M.E.T.E. That is, even if a variable has M.E.T.E. that is not the variable with maximum total effect in the generating SEM, we might still gain causal access to the target variable via the M.E.T.E. variable. Thus, we compare the expected total effect for the M.E.T.E. variable with its true total effect in the generating SEM. For these simulations, with edge parameters constrained to be positive, the maximum possible value for any variable's total effect is 1 while the minimum value is 0.

For quantitative results that follow, we present error/performance in two ways:

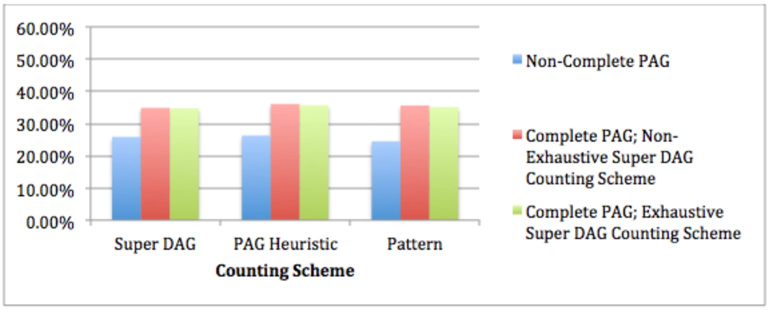

Figure 4.5: Overall Absolute Error of M.E.T.E. as Mean Percentage of Maximum Possible Error (for M.E.T.E. Variable); Summarizes Table 4.5

- Let $s = \#$ of generating SEMs for a simulation strategy (or a subset thereof);
  m_{ci} = calculated value of a causal predictability metric for generating SEM i according to a particular counting scheme; and
  m_{ti} = true value of a causal predictability metric for SEM i.
  Absolute Error as Mean Percentage of Maximum Possible Error[5] =

  $$\frac{1}{s}\sum_{i=1}^{s}\left(\frac{|m_{ci}-m_{ti}|}{max(1-m_{ti},m_{ti})}*100\right)$$

- *Mean Squared Error (MSE)* for calculated values of a causal predictability metric compared to their true values in the generating SEM.

Figure 4.5 summarizes Table 4.5, providing overall absolute error as mean percentage of maximum possible error (mean absolute error percentage for short) for inferred M.E.T.E. variables' values compared to the true total effects for the M.E.T.E. variables in each of the three simulation strategies. Table 4.6 provides overall mean squared error for inferred M.E.T.E. variables' values. Overall, we find that the Pattern counting scheme has lowest mean squared error in all simulation strategies but is narrowly outperformed for mean absolute error percentage by the Super DAG counting scheme for complete PAG simulation strategies. Notably, counting scheme performance is roughly similar within each simulation strategy for both ways of quantifying error.

---

[5] The first argument of the max function contains 1 as the number from which we subtract because it is the maximum possible value for the total effect of any given variable; it is also the maximum possible R^2 value for metrics we consider in sections to follow.

Simulation Strategy				**Counting Scheme: Absolute Error of Maximum Expected Total Effect as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	25.85%	26.28%	24.50%
2	Yes	18,000	No	34.90%	36.10%	35.63%
3	Yes	400	Yes	34.73%	35.70%	35.08%

Table 4.5: Overall Absolute Error of Maximum Expected Total Effect Variable's Estimated Value vs. True Value as Mean Percentage of Maximum Possible Error

Simulation Strategy				**Counting Scheme: MSE of Maximum Expected Total Effect Variable**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	.0687	.0692	.0621
2	Yes	18,000	No	.1020	.1039	.1016
3	Yes	400	Yes	.1004	.1038	.0995

Table 4.6: Overall Mean Squared Error of Maximum Expected Total Effect Variable's Estimated Value vs. True Value

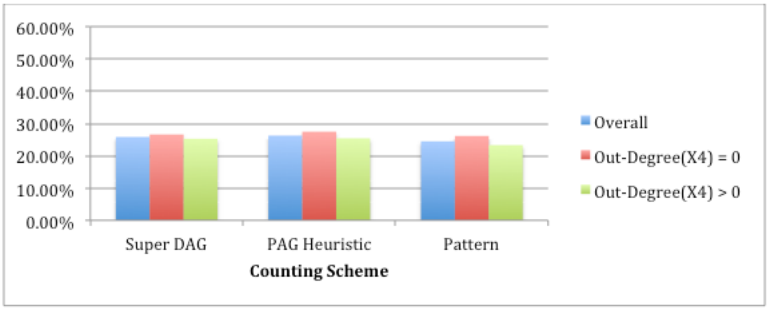

Figure 4.6: Absolute Error of M.E.T.E. as Mean Percentage of Maximum Possible Error (for M.E.T.E. Variable) for Generating Super DAGs Not Represented by Complete PAG By Target Out-Degree

Figures 4.6, 4.7, and 4.8 compare mean absolute error percentage for each simulation strategy, respectively, stratified by whether the target variable in the generating Super DAG has zero or positive out-degree. Un-stratified, overall error is also provided for comparison.

For generating Super DAGs not represented by the complete PAG, there is no substantial difference in performance based on the out-degree of the target, but we see that for both simulation strategies involving generating Super DAGs represented by the complete PAG performance is substantially worse when target has out-degree zero. Thus, for generating Super DAGs represented by the complete PAG the target having out-degree zero makes it easier to find the correct variable with M.E.T.E., but it is harder to discover quantitative causal “strength” of the M.E.T.E. variable.

Tables 4.7 and 4.8 provide mean absolute error percentage and mean squared error, respectively, for each counting scheme for generating Super DAGs in which the out-degree of the target is zero. As noted, there is a substantial decrease in relative performance when the generating Super DAG is represented by the complete PAG vs. non-complete PAG across counting schemes; however, for generating Super DAGs not represented by the complete PAG we find that performance is marginally better in terms of mean squared error and slightly worse in terms of mean absolute error percentage compared to overall performance. The Pattern counting scheme performs best on generating Super DAGs not represented by the complete PAG, and best performance is split across error measures between the PAG Heuristic and Pattern counting schemes for generating Super DAGs represented by the complete PAG.

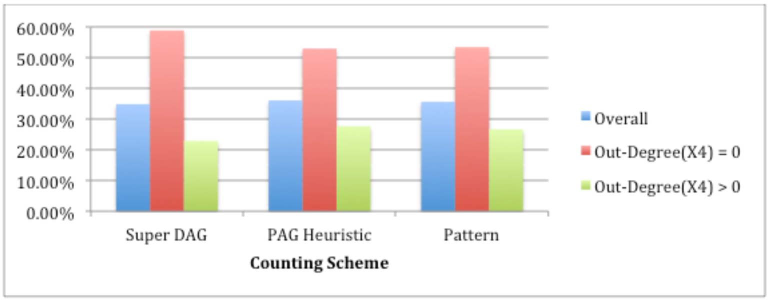

Figure 4.7: Absolute Error of M.E.T.E. as Mean Percentage of Maximum Possible Error (for M.E.T.E. Variable) for Generating Super DAGs Represented by the Complete PAG with Non-Exhaustive Super DAG Counting Scheme By Target Out-Degree

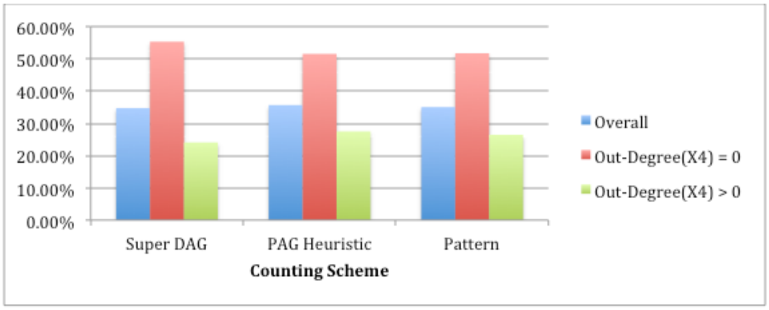

Figure 4.8: Absolute Error of M.E.T.E. as Mean Percentage of Maximum Possible Error (for M.E.T.E. Variable) for Generating Super DAGs Represented by the Complete PAG with Exhaustive Super DAG Counting Scheme By Target Out-Degree

Simulation Strategy				**Counting Scheme: Absolute Error of Maximum Expected Total Effect as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	26.62%	27.48%	26.17%
2	Yes	6,003	No	58.90%	52.98%	53.44%
3	Yes	136	Yes	55.37%	51.58%	51.77%

Table 4.7: Generating Super DAGs with Target (X4) Out-Degree = 0: Absolute Error of M.E.T.E. Value as Mean Percentage of Maximum Possible Error for Chosen Variable

For generating Super DAGs in which the target variable has effects (i.e., out-degree greater than zero), the Pattern counting scheme again performs best with generating Super DAGs not represented by the complete PAG (cf. Figure 4.6, Tables 4.9 and 4.10). The Super DAG counting scheme performs best for generating Super DAGs represented by the complete PAG (cf. Figure 4.7-4.8). Notably, performance for all counting schemes does not degrade as substantially for generating Super DAGs represented by the complete PAG compared to those not represented by the complete PAG when the target has positive out-degree in the generating structure. Indeed, performance for each class of generating Super DAGs is roughly comparable.

Importantly, our performance is roughly the same (and best) for generating Super DAGs not represented by the complete PAG, regardless of target out-degree. When we are in the situation in which inferred, qualitative PAG causal structure is informative, we are in the best position to infer the value of the expected total effect of putative causes of the target variable.

So far, these simulations suggest on one hand that for qualitative accuracy (with respect to the choice of variable with maximum total effect), we should use the PAG Heuristic counting scheme. On the other hand, if we are only concerned for the quantitative accuracy of the estimate of expected total effect, we should, perhaps surprisingly, deploy the Pattern counting scheme, essentially disregarding information about latent confounding. However, similar performance with respect to quantitative accuracy, especially between the Pattern and PAG Heuristic counting schemes, coupled with its su-

Simulation Strategy				**Counting Scheme: MSE of Maximum Expected Total Effect Variable**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	.0552	.0573	.0520
2	Yes	6,003	No	.1946	.1603	.1612
3	Yes	136	Yes	.1734	.1579	.1540

Table 4.8: Generating Super DAGs with Target (X4) Out-Degree = 0: Mean Squared Error of M.E.T.E. Variable's Estimated Value vs. True Value

Simulation Strategy				**Counting Scheme: Absolute Error of Maximum Expected Total Effect as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	25.30%	25.45%	23.33%
2	Yes	11,997	No	22.89%	27.66%	26.71%
3	Yes	264	Yes	24.10%	27.52%	26.48%

Table 4.9: Generating Super DAGs with Target (X4) Out-Degree > 0: Absolute Error of M.E.T.E. Value as Mean Percentage of Maximum Possible Error for Chosen Variable

Simulation Strategy				**Counting Scheme: MSE of Maximum Expected Total Effect Variable**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	.0783	.0776	.0692
2	Yes	11,997	No	.0556	.0756	.0718
3	Yes	264	Yes	.0628	.0760	.0714

Table 4.10: Generating Super DAGs with Target (X4) Out-Degree > 0: Mean Squared Error of M.E.T.E. Variable's Estimated Value vs. True Value

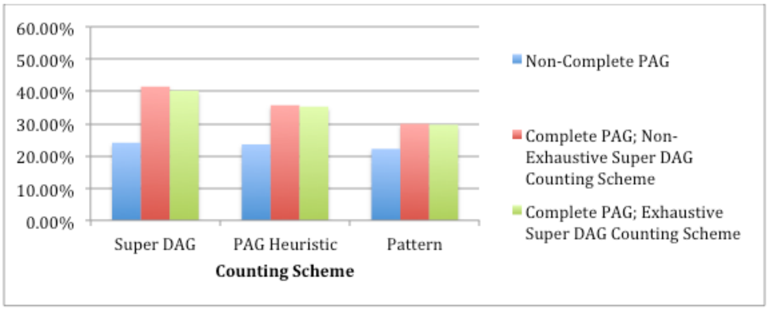

Figure 4.9: Overall Absolute Error of Average Naïve Causal R^2 as Mean Percentage of Maximum Possible Error; Summarizes Table 4.11.

perior qualitative performance lead us to favor and deploy the PAG Heuristic counting scheme for the M.E.T.E. causal predictability metric both in §4.2 and the practical case study in Chapter 5.

4.1.5 Naïve Causal R^2

We summarize results for the (Average) Naïve Causal R^2 causal predictability metric for three counting schemes and simulation strategies in the same way we do for the M.E.T.E. metric. Similar to results for the M.E.T.E. metric, in Figure 4.9 and Tables 4.11-4.12 we see that the Pattern and PAG Heuristic counting schemes perform best in tracking the true naïve causal R^2 of the generating SEM.

The Pattern counting scheme performs best regardless of whether the

Simulation Strategy				**Counting Scheme: Absolute Error of Average Naïve Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	24.09%	23.58%	22.18%
2	Yes	18,000	No	41.50%	35.72%	29.98%
3	Yes	400	Yes	40.26%	35.29%	29.67%

Table 4.11: Overall Error for Average Naïve Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Naïve Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Naïve Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	.1120	.0573	.0633
2	Yes	18,000	No	.1513	.1127	.0985
3	Yes	400	Yes	.1479	.1073	.0925

Table 4.12: Overall Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value

generating Super DAG is represented by the complete PAG in terms of mean absolute error percentage, but in terms of mean squared error, the PAG Heuristic modestly outperforms the Pattern counting scheme for generating Super DAGs not represented by the complete PAG. Again, performance degrades for generating Super DAGs represented by the complete PAG, but all three counting schemes have relatively similar performance for generating Super DAGs not represented by the complete PAG.

Considering the same stratifications we do for the M.E.T.E. metric, a pattern similar to the overall, un-stratified result emerges. Either the PAG Heuristic or Pattern counting scheme always achieves best performance. Figures 4.10-4.12 and Tables 4.13-4.14 show that the Pattern counting scheme does modestly better than the other two counting schemes for generating Super DAGs with zero out-degree for the target variable, but performance degrades for the Super DAG counting scheme in all cases and for the PAG Heuristic in both complete PAG generating Super DAG simulation strategies. Perhaps surprisingly, for generating Super DAGs with target out-degree greater than zero (cf. Figures 4.10-4.12 and Tables 4.15-4.16), performance of the Super DAG counting scheme (in all cases) and PAG Heuristic counting scheme (in most cases) improves compared to the case of zero out-degree for the target, while the Pattern counting scheme's performance degrades.

For the naïve causal R^2 metric of causal predictability, we find the assumption that the target variable has no effects is not as important as it is for our ability to choose the variable with maximum total effect. Notably, however, these results mirror those we find considering the quality of our quantitative estimates of the total effect of the M.E.T.E. variable.

The choice of a counting scheme to deploy in practice for this metric depends on investigator expectations, intuitions, and background knowledge; for generating Super DAGs not represented by the complete PAG, we consider the PAG Heuristic to have a slight advantage over the Pattern counting scheme. We use the PAG Heuristic in simulations in §4.2 and in the applied case study in Chapter 5 for the Naïve Causal R^2 metric.

4.1.6 Intervention Causal R^2

We present results for the (Average) Intervention Causal R^2 metric in the same manner. Both overall and stratified results point to the superiority of the Super DAG counting scheme for this metric. The Pattern counting scheme is substantially worse than both the PAG Heuristic and Super DAG counting schemes in nearly all cases; a quick inspection of the overall

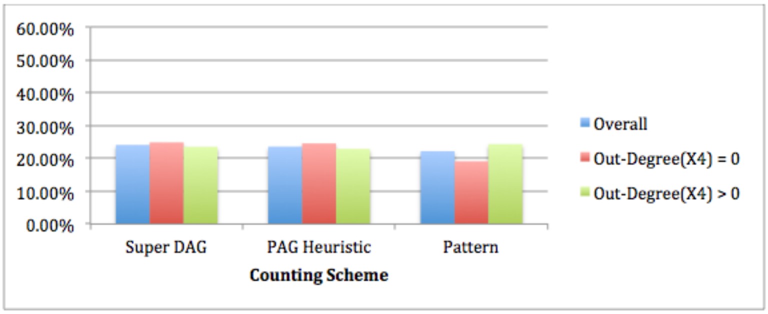

Figure 4.10: Overall Absolute Error of Average Naïve Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Not Represented by the Complete PAG By Target Out-Degree

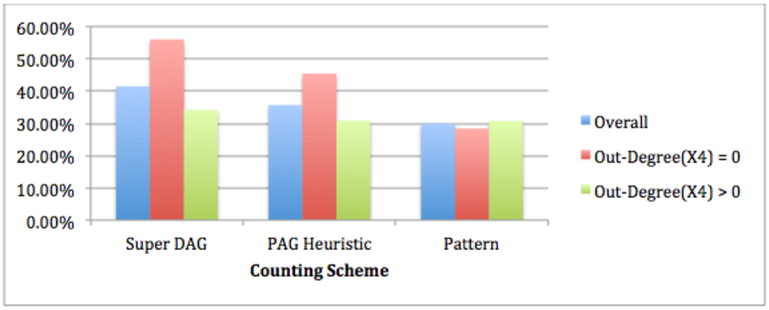

Figure 4.11: Overall Absolute Error of Average Naïve Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Represented by the Complete PAG; Non-Exhaustive Super Counting Scheme; By Target Out-Degree

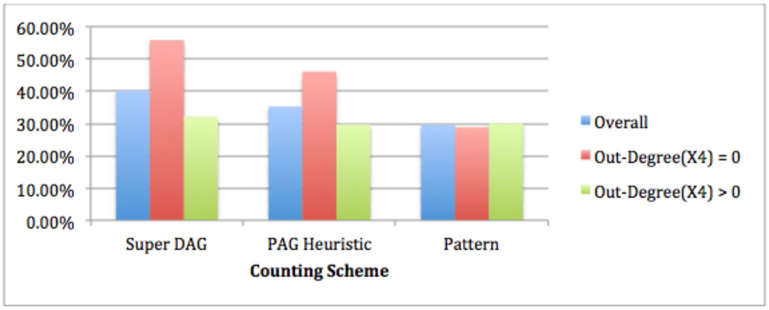

Figure 4.12: Overall Absolute Error of Average Naïve Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Represented by the Complete PAG; Exhaustive Super Counting Scheme; By Target Out-Degree

Simulation Strategy				**Counting Scheme: Absolute Error of Average Naïve Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	24.86%	24.55%	19.11%
2	Yes	6,003	No	56.08%	45.47%	28.41%
3	Yes	136	Yes	55.92%	46.14%	28.85%

Table 4.13: Generating Super DAGs with Target (X4) Out-Degree = 0: Error for Average Naïve Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Naïve Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Naïve Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	.1393	.0528	.0411
2	Yes	6,003	No	.2516	.1604	.0641
3	Yes	136	Yes	.2564	.1555	.0618

Table 4.14: Generating Super DAGs with Target (X4) Out-Degree = 0: Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value

Simulation Strategy				**Counting Scheme: Absolute Error of Average Naïve Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	23.54%	22.89%	24.34%
2	Yes	11,997	No	34.20%	30.85%	30.77%
3	Yes	264	Yes	32.20%	29.71%	30.09%

Table 4.15: Generating Super DAGs with Target (X4) Out-Degree > 0: Error for Average Naïve Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Naïve Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Naïve Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	.0929	.0606	.0789
2	Yes	11,997	No	.1011	.0888	.1157
3	Yes	264	Yes	.0919	.0825	.1083

Table 4.16: Generating Super DAGs with Target (X4) Out-Degree > 0: Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value

results in Figure 4.13 and Tables 4.17-4.18 as well as stratified results in Figures 4.14-4.16 and Tables 4.19-4.22 make apparent this stark contrast in performance.

Generally, the PAG Heuristic appears to roughly fall in the middle between the performance of the Super DAG counting scheme and that of the Pattern counting scheme. In a case in which the PAG Heuristic out performs the Super DAG counting scheme (i.e., target out-degree = 0 generating Super DAGs in Tables 4.19 and 4.20, Simulation Strategy #2), we see that the PAG Heuristic narrowly outperforms the Super DAG counting scheme. The performance gap between the PAG Heuristic and Super DAG counting scheme is smallest for generating Super DAGs represented by the complete PAG with zero target out-degree regardless of whether the Super DAG counting scheme is exhaustive or not.

From these results, we choose the Super DAG counting scheme to assess Average Intervention Causal R^2 in the simulations of §4.2. However, in §4.2 we only consider search over four variables at a time and disregard sets of variables from which we infer the complete PAG; this makes the Super DAG counting scheme feasible. However, in the applied case study of Chapter 5 we will consider larger sets of constructed variables for which the Super DAG counting scheme as explicated is no longer feasible, so we deploy the next best counting scheme, the PAG Heuristic.

4.1.7 Methodological/Practical Implications and Limitations

To summarize, in simulations we use to consider variable construction problems in §4.2, we use the PAG Heuristic counting scheme for the M.E.T.E.

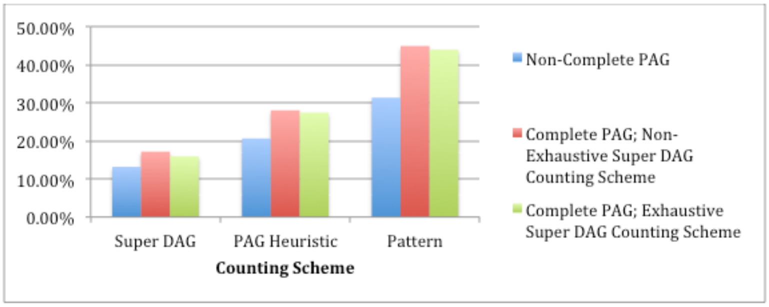

Figure 4.13: Overall Absolute Error of Maximum Average Intervention Causal R^2 as Mean Percentage of Maximum Possible Error; Summarizes Table 4.17

Simulation Strategy				**Counting Scheme: Absolute Error of Average Intervention Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	13.20%	20.65%	31.37%
2	Yes	18,000	No	17.20%	28.05%	44.94%
3	Yes	400	Yes	15.97%	27.44%	43.99%

Table 4.17: Overall Error for Average Intervention Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Intervention Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Intervention Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	32,000	Yes	.0192	.0498	.1122
2	Yes	18,000	No	.0260	.0757	.1692
3	Yes	400	Yes	.0238	.0712	.1621

Table 4.18: Overall Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value

Simulation Strategy				**Counting Scheme: Absolute Error of Average Intervention Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	15.19%	25.22%	37.29%
2	Yes	6,003	No	22.88%	21.61%	38.63%
3	Yes	136	Yes	20.83%	21.72%	37.84%

Table 4.19: Generating Super DAGs with Target (X4) Out-Degree = 0: Error for Average Intervention Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Intervention Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Intervention Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	13,196	Yes	.0206	.0563	.1261
2	Yes	6,003	No	.0388	.0387	.1025
3	Yes	136	Yes	.0363	.0394	.1027

Table 4.20: Generating Super DAGs with Target (X4) Out-Degree = 0: Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value

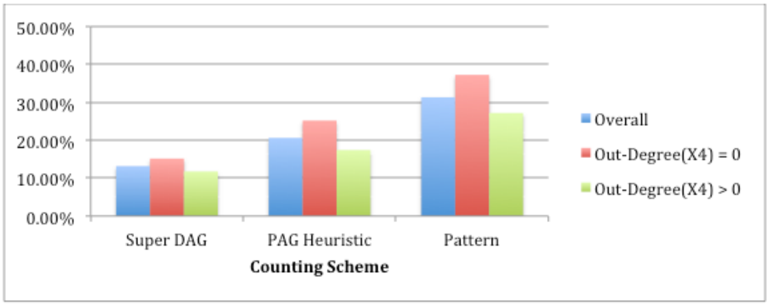

Figure 4.14: Overall Absolute Error of Average Intervention Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Not Represented by Complete PAG; By Target Out-Degree

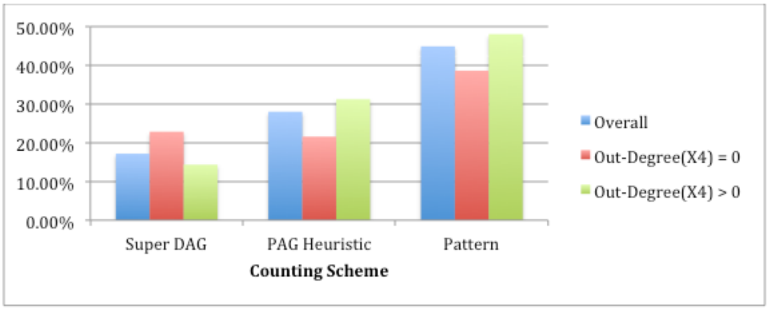

Figure 4.15: Overall Absolute Error of Average Intervention Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Represented by Complete PAG; Non-Exhaustive Super DAG Counting Scheme; By Target Out-Degree

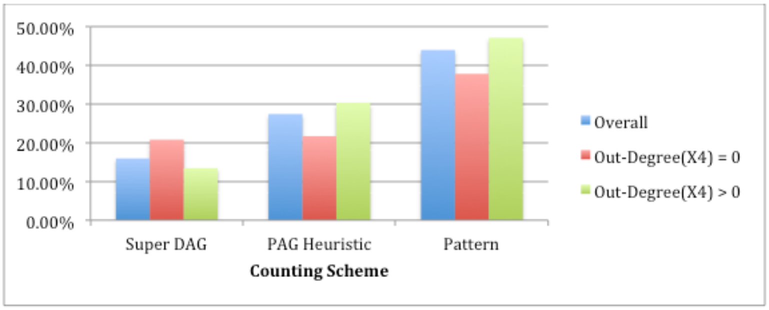

Figure 4.16: Overall Absolute Error of Average Intervention Causal R^2 as Mean Percentage of Maximum Possible Error: Generating Super DAGs Represented by Complete PAG; Exhaustive Super DAG Counting Scheme; By Target Out-Degree

Simulation Strategy				**Counting Scheme: Absolute Error of Average Intervention Causal R^2 as Mean % of Maximum Possible Error**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	11.79%	17.44%	27.22%
2	Yes	11,997	No	14.36%	31.27%	48.10%
3	Yes	264	Yes	13.46%	30.38%	47.16%

Table 4.21: Generating Super DAGs with Target (X4) Out-Degree > 0: Error for Average Intervention Causal R^2 Value as Mean Percentage of Maximum Possible Error Given True Intervention Causal R^2 Value

Simulation Strategy				**Counting Scheme: MSE of Average Intervention Causal R^2**		
#	*Complete PAG*	*# Super DAGs Sampled*	*All Super DAGs?*	*Super DAG*	*PAG Heuristic*	*Pattern*
1	No	18,804	Yes	.0182	.0453	.1024
2	Yes	11,997	No	.0196	.0941	.2025
3	Yes	264	Yes	.0174	.0875	.1927

Table 4.22: Generating Super DAGs with Target (X4) Out-Degree > 0: Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value

and Naïve Causal R^2 metrics and the non-exhaustive Super DAG counting scheme for the Intervention Causal R^2 metric. In the applied case study of Chapter 5 in which we consider sets of variables with more than four variables, we use the PAG Heuristic counting scheme for all three metrics, as it is not computationally feasible to iterate over the set of Super DAGs represented by a PAG on more than four variables.

Heuristic methods to iterate over appropriate subsets of Super DAGs represented by a PAG are a subject for future research. Ideally, we can improve on the performance of the relatively simple PAG Heuristic for all metrics with a computationally feasible counting scheme. Further, we might improve on the performance of the Super DAG counting scheme(s) for the Intervention Causal R^2 or find a way to minimally decrease performance while providing a computationally feasible heuristic to generate sets of Super DAGs over which to average (i.e., provide a better Super DAG counting scheme).

Beyond the limitations of the three counting schemes we consider, we only consider their application to cases with four variables. Both the PAG Heuristic and Pattern counting scheme can be feasibly applied in cases with at least several more variables, but it less clear what an appropriate simulation regime would look like to test the performance of the counting schemes for causally insufficient cases. On four variables, we can consider the entire space of generating Super DAGs with up to six unique latent variables. As this still involves a simplifying assumption with respect to possible latent variables, future work should consider what simplifying assumptions are reasonable to consider cases with latent causes of more measured variables. This would allow for the specification of possible generating causal structures and to pursue simulations similar to those we have considered so far.

Despite the limited causal structures these simulations cover, the relative performance of the counting schemes, especially that of the Pattern and PAG Heuristic counting schemes for the M.E.T.E. and Naïve Causal R^2 metrics, is noteworthy. Both the Pattern and PAG Heuristic schemes discard significant (if not all) information about latent confounding; in both cases, only DAGs (i.e., without latent variables) are generated as causal structures represented by a PAG. That either counting scheme out performs the more sophisticated counting of Super DAGs represented by a PAG is an interesting result worthy of further study. Further, they perform better than the Super DAG counting scheme in most cases even when there are several latent causes of the target variable, so their better performance in the aggregate is not due simply to cases in which the generating causal structure is relatively simple (e.g., in terms of the number of unmeasured causes of the target variable of interest).

4.2 "Cloud" Variable Construction Simulations

We now consider situations in which we construct aggregate variables. We develop a simulation regime in which we construct variables to maximize each causal predictability metric according to each metric's "best" counting scheme (i.e., those counting schemes matched to metrics in §4.1). We begin by describing the basic, underlying causal structure for data generating processes for simulation, including (hypothetically) "raw" variables and aggregate variables constructed from these raw variables. We simulate data from models including constructed aggregate variables, but then assume that we only have access to raw variables from which to construct aggregate variables that maximize causal predictability metrics.

Results of this simulation help us to understand circumstances under which variables that support causal inference may be different from variables that are optimal for predicting a particular target variable or that are present in the underlying data generating process. Understanding inferred causal structures over these different sets of variables helps us better understand how these two investigative goals (i.e., causal vs. purely predictive inference) can come apart.

4.2.1 Simulation Regime

In our simulations, data generating processes are linear structural equation models specified according to the basic causal structure of Figure 4.17 (in most cases augmented or changed in ways we note as we proceed). We suppose variables X0 through X8 and Y0 through Y2 are "raw" variables; X0 through X8 are variables with Normal distributions with zero mean and variance assigned randomly[6] for each variable in each simulation instance, and Y0 through Y2 have Normally distributed error terms with zero mean and randomly assigned variance in all generating linear structural equation models.

The model includes aggregate variables, "generating process aggregate variables": AVG_X0X1X2X3X4, AVG_X5X6X7X8, and AVG_Y0Y1Y2, determined[7] by raw variables, that are causally efficacious

---

[6] from a Uniform(1.0, 3.0) distribution

[7] In generating structural equation models, we consider a variable with exceedingly small (or zero) error variance and constructed as the appropriate aggregate function (here, average = AVG) of raw variables to be deterministically constructed.

(e.g., two aggregate variables are direct causes of T and the only cause of Y0, Y1, and Y2 is AVG_X0X1X2X3X4). We call the sets of variables that determine generating process aggregate variables "generating process X Cloud #1" ($\{X0, X1, X2, X3, X4\}$) and "generating process X Cloud #2" ($\{X5, X6, X7, X8\}$). T is a fixed target variable. While aggregate variables are in the generating causal structure, after generating data from a model, we assume no knowledge of their constituent raw variables; we consider simulated data for "raw" variables and T, seeking to construct variables from raw variables to infer causes of and/or predict T.

More generally, we refer to Figure 4.17 and similar models as "cloud" causal models, as there are many possible ways in which "raw" variables might be grouped to construct aggregate variables (the possible groupings being cloud-like). Assuming we do not know the variables that determine generating process aggregate variables, we search for a set of three constructed aggregate variables. For each set of constructed variables we consider in the search, we learn a PAG and calculate metrics of causal predictability to judge which set of aggregate constructed variables is "best" for each metric. We also construct multi-variate linear regression models for T to calculate the predictive R^2 value for each set of constructed variables so that we can compare sets of variables that maximize causal predictability metrics to those that are judged as best predictors (without respect to causal information).

Searching for constructed variables, we address a variety of questions. For example, do aggregate constructed variables that maximize metrics of causal predictability (e.g., M.E.T.E.) and those that maximize overall predictive accuracy match the generating process aggregate variables? How does the complexity of the generating causal structure influence the answer to this question and others we seek to answer? Do specific characteristics, other than complexity, of the generating structure matter?

Inspecting the structure of Figure 4.17, one notes the peculiarity of constructing aggregate variables from "raw" variables that are independent of each other. In most real world applications such aggregates seem unlikely to be salient causal features. To pursue questions of complexity of generating models (i.e., of the complexity or density of raw variable "clouds"), we consider augmenting the basic generating structure of Figure 4.17 by adding causal relationships among raw variables that make up what we call the two "X clouds." We augment the data generating causal structure by adding "within cloud" edges (e.g., X0 → X4, X5 → X7, etc.), "inter-cloud" edges (e.g., X0 → X8, X7 → X2, etc.), or both.

We consider a variety of "within cloud" and "inter-cloud" edge config-

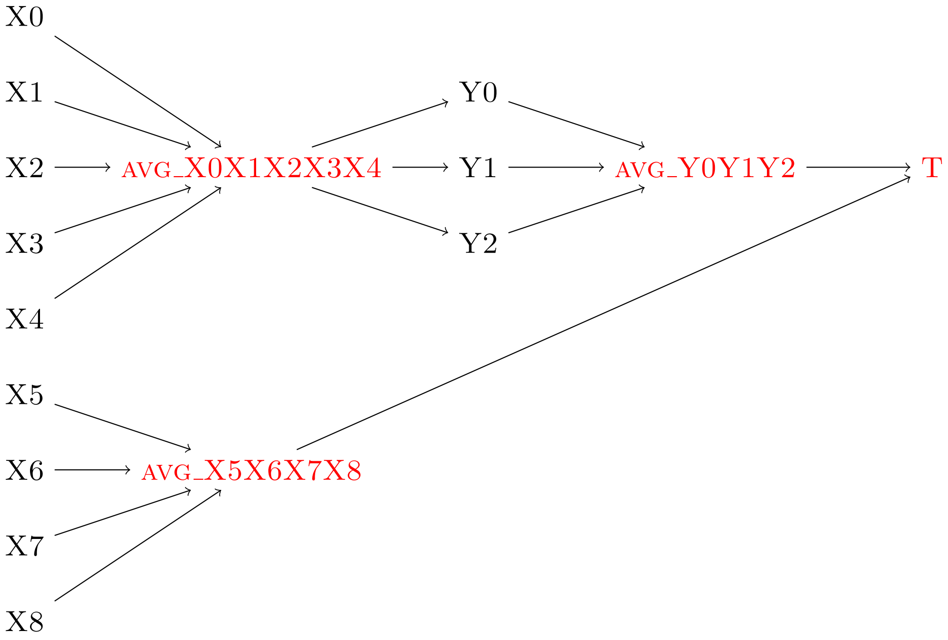

Figure 4.17: Basic Generating “Cloud” Causal Structure

Edge Configuration Shorthand	**“Inter-Cloud” Edges**
NONE	None
X4 → X5	X4 → X5
X5 → X4	X5 → X4
X0_X4 → X5	X0 → X5, X1 → X5, X2 → X5, X3 → X5, X4 → X5
X5_X8 → X4	X5 → X4, X6 → X4, X7 → X4, X8 → X4
X0_X4 → X5_X8	X0 → X5, X1 → X6, X2 → X7, X3 → X8, X4 → X8
X5_X8 → X0_X4	X5 → X0, X6 → X1, X7 → X2, X8 → X3, X8 → X4

Table 4.23: “Inter-Cloud” Edge Configurations for “X Clouds” in Generating Process Causal Structure for Simulation Study

urations for generating causal structures from which we simulate data. We consider instances in which there are zero, two, and four randomly placed “within cloud” edges in the two generating process “X Clouds,” coupled with each of seven possible “inter-cloud” edge configurations (see Table 4.23).

We randomly instantiate parameters of a linear structural equation model ten times for each combination of a number of “within cloud” edges and an “inter-cloud” edge configuration, for a total of 210 simulated data sets (n = 1,000 for each data set). For each data set, we search over constructed aggregate variables over “X Cloud” raw variables, assuming we know the structure of the “Y Cloud” so AVG_Y0Y1Y2 is always a variable in the set under consideration (along with the target variable T). We seek constructed variables that maximize causal predictability metrics and best support causal inferences about T. We now describe the search space for constructed aggregate variables.

4.2.2 “Cloud” Search Space

To pursue the search for aggregate constructed variables, we delimit a relatively limited space and exhaustively consider the entire space. For each simulated data set, we consider every partition of the X Cloud raw variables $\{X0, X1, ..., X8\}$ into at least two sets with a minimum cardinality of three each. In cases in which such partitioning produces three sets with three variables each, we drop one set. For simplicity, we only consider the average function; applying this function to the partitioned sets of raw variables produces 1,050 sets of two disjoint constructed "X Cloud" variables with at least three component raw variables. Examples include:

- $\{X0, X1, X2, X3, X4, X5\}$ and $\{X6, X7, X8\}$ [a (6, 3) partition]
- $\{X1, X4, X5, X7, X8\}$ and $\{X0, X2, X3, X6\}$ [a (5, 4) partition]
- $\{X0, X4, X7\}$ and $\{X2, X6, X8\}$ [a (3, 3, 3) partition from which we omit one set of 3 "raw" variables]

To these sets of two constructed aggregate variables we add AVG_Y0Y1Y2 and T and learn a PAG. We then calculate causal predictability metrics and predictive R^2 values by estimating linear SEMs over the Super DAGs and DAGs represented, according to each counting scheme, by each PAG. We then compare (sets of sets of) constructed variables that maximize particular causal predictability and predictive metrics to the aggregate generating process variables and describe the structure of PAGs learned over these variables. We contrast these causal structures with those expected over aggregate generating process variables.

4.2.3 Results: Constructed Variable Sets vs. Generating Process Variables vs. Best "Predictive" Variable Sets

Recall that the three causal predictability metrics were each matched to a counting scheme in §4.1. For intervention causal R^2, we deploy the non-exhaustive Super DAG counting scheme; for naïve causal R^2 and M.E.T.E., we deploy the PAG Heuristic counting scheme. We use a shortened name of causal predictability metrics to indicate the combination of this metric with its corresponding counting scheme. Specifically, we call the naïve causal R^2 metric "Naïve," the intervention causal R^2 metric "Intervention," maximum expected total effect "METE," and the best predictive (non-causal) R^2 "Predictive." Sample size n refers to number of simulation instances, rather than the sample size of simulated data (which is always 1,000).

The first set of results provide answers to two important questions about the comparison of variables that maximize causal predictability metrics (i.e., those that support causal inference) to generating process variables and those that are simply best predictors of the target variable. We consider results first for all 210 simulation-instances, and then we stratify the results based on characteristics of the generating causal structures from which we simulate data, specifically "within cloud" and "inter-cloud" edge configurations for these generating causal structures.

Figure 4.18 provides a stark demonstration that constructed variable sets that maximize each of the three causal predictability metrics infrequently match aggregate variables in the underlying data generating process (Conclusion #8). Meanwhile, those that maximize predictive R^2 without respect to causal information frequently match (in 69% of 210 simulation-instances) aggregate generating process variables. Together, these facts further demonstrate that variables that maximize causal predictability metrics are generally not the same as those maximize a purely predictive metric, e.g., linear regression R^2 (Conclusion #9). That is, predictive and causal inference can "come apart" with respect to aggregate variables one might construct for each inference task.

The pattern obtains regardless of differing levels of complexity in the "X Clouds" in the generating process (see Figure 4.19-4.20). One notable exception is the case of intervention causal R^2 for two relatively complex inter-cloud configurations, X5_X8 → X4 and X5_X8 → X0_X4. In both of these configurations we see a greater percentage of matches, and both involve all of the "X Cloud #2" variables having direct causal connections into the "X Cloud #1" variables. Nevertheless, the percentages are relatively small compared to those of the best predictive aggregate variables matching generating process variables.

M.E.T.E. variable sets appear to have the least (indeed, a very small) chance of matching aggregate generating process variables at least partially because we seek a single variable in a set with maximum expected total effect. We frequently encounter cases in which the single constructed aggregate variable that has M.E.T.E. is an "omnibus" variable with five or six raw variable components, including variables from both "X Cloud #1" and "X Cloud #2" in the generating process. Given the basic generating causal structure, this can be expected for particular simulated parameterizations in which AVG_Y0Y1Y2 does not have M.E.T.E.; we might include variables from "X Cloud #2" that have a direct effect on T via AVG_X5X6X7X8 (in the generating process) and some raw variables from "X Cloud #1" that have an indirect effect on T via the AVG_X0X1X2X3X4, Y0, Y1, Y2, and AVG_Y0Y1Y2. We would expect that in some cases this would produce a greater expected total effect than just including variables from "X Cloud #2."

However, this does not fully explain why the M.E.T.E. and other causal predictability metrics almost always lead to variables that do not match aggregate variables in the generating process. To further explain this non-

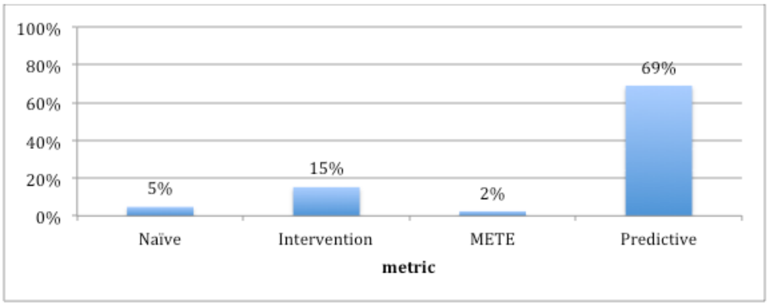

Figure 4.18: Overall Percentage of Instances By (Causal) Predictability Metric with Data Generating Process Variable Match (n = 210)

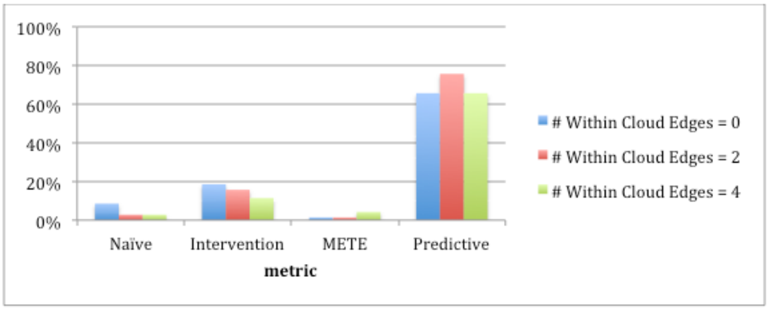

Figure 4.19: Percentage of Instances with Generating Process Aggregate Variable Match by "Within Cloud" Edge Configurations (n = 70 simulation-instances for each configuration)

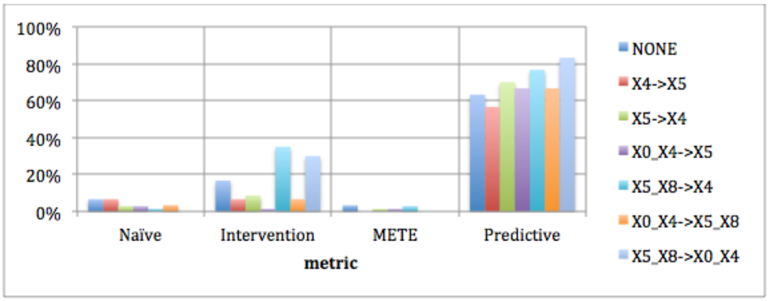

Figure 4.20: Percentage of Instances with Generating Process Aggregate Variable Match by "Inter-Cloud" Edge Configurations (n = 30 simulation-instances for each configuration)

matching phenomenon, we must consider how constructed aggregate variables support causal inference (in contrast to aggregate generating process variables) in learning PAG structures that represent conditional independence relationships among variables in our search.

4.2.4 Results: Inferred PAG Causal Structures vs. Generating Process PAG

Searching for constructed, aggregate variables that maximize causal predictability metrics calculated over inferred PAGs, we find several "exemplar" PAGs that occur frequently. We describe the relative frequency with which these exemplar PAGs occur and show that "generating process isomorphic" PAGs, those one might expect to infer given the basic generating causal structure, occur less frequently than might be expected.

We begin by describing the two "generating process" PAGs. For the basic causal structure of Figure 4.17, using the generating process aggregate variables (likely) leads to inferring the PAG of Figure 4.21. Because of the independence of raw X variables, we expect AVG_X5X6X7X8 to be independent of AVG_X0X1X2X3X4 and AVG_Y0Y1Y2. Further, we expect AVG_X0X1X2X3X4 to be independent of T conditional on AVG_Y0Y1Y2.[8] Because AVG_X5X6X7X8 and AVG_Y0Y1Y2 are independent but neither is independent of T, we infer a collider at T, but otherwise we infer no more edge orientations. If there were inter-cloud edges between raw variables in generating process X Cloud #1 and X Cloud #2, we would expect to infer the PAG of Figure 4.22, as AVG_X0X1X2X3X4 and AVG_X5X6X7X8 are no longer independent[9], but this allows us to infer no more edge orientations.

We call an inferred PAG a "generating process isomorphic PAG" if it has the same structure as either Figure 4.21 or Figure 4.22 and constructed aggregate variables are substituted for AVG_X0X1X2X3X4, AVG_X5X6X7X8, or both.[10] For example, for a particular instance of the underlying gener-

---

[8] Recall from Chapter 1 that the FCI algorithm we use to infer PAGs starts with a complete PAG (i.e., o−o edges between every pair of variables). These two unconditional independencies allow FCI to eliminate the following two edges: AVG_X0X1X2X3X4 o−o AVG_X5X6X7X8 and AVG_X5X6X7X8 o−o AVG_Y0Y1Y2. FCI eliminates the o−o edge between AVG_X0X1X2X3X4 and T based on the conditional independence of those two variables given AVG_Y0Y1Y2.

[9] Note that AVG_X0X1X2X3X4 and T will still be independent given AVG_Y0Y1Y2 and AVG_X5X6X7X8.

[10] If constructed aggregate variables happen to exactly match generating process variables and either of these two PAGs is inferred, we also call this an instance of a generating process isomorphic PAG.

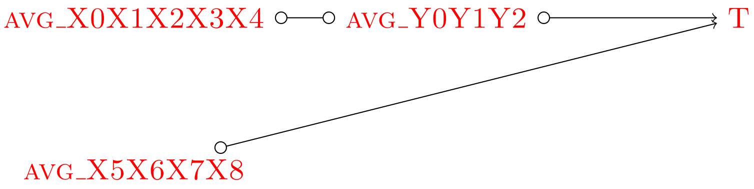

Figure 4.21: PAG for Generating Process Aggregate Variables with Independent Raw X Variables

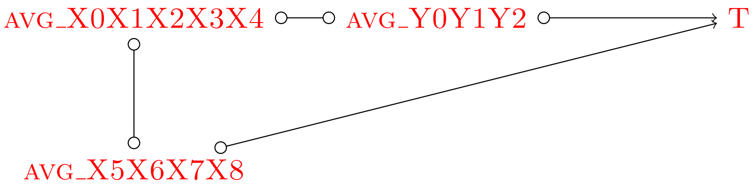

Figure 4.22: PAG for Generating Process Aggregate Variables without Independent Raw X Variables

ating process causal structure, we might infer the same PAG structure as in Figure 4.22 if we considered constructed variables AVG_X0X1X2 and AVG_X5X6X7. We would say that we inferred a generating process isomorphic PAG, but this would be an instance in which we did not find a match for generating process aggregate variables.

In what follows, we use X_CLOUD_1 and X_CLOUD_2 as "wildcard" variables to stand in for any (contextually appropriate) constructed, aggregate variables over X raw variables, and Y_CLOUD stands in for AVG_Y0Y1Y2. Specifically we use wildcards when we talk about exemplar PAGs that represent many possible inferred PAGs (e.g., in Figure 4.21 and 4.22, AVG_X0X1X2X3X4 would be replaced by X_CLOUD_1 and AVG_X5X6X7X8 by X_CLOUD_2 in exemplars for generating process isomorphic PAGs). In general, for PAGs we infer over constructed variables in one simulation instance, there are many possible constructed variables that take on the same conditional independence relationships (and thus the same PAG structure) in other simulation instances.

In "generating process isomorphic" PAGs, we find two o→ edges, leaving us with uncertainty about the status of Y_CLOUD (i.e., AVG_Y0Y1Y2) and X_CLOUD_2 (i.e., AVG_X5X6X7X8 in Figures 4.21-4.22) as causes of T. Recall that predictive R^2 is calculated without respect to structure uncertainty in the PAG (i.e., all three variables other than T are included in the predictive R^2 regression[11]). In the majority of simulation instances, constructed variables that maximize predictive R^2 match generating process aggregate variables, but causal relations among variables that maximize causal predictability metrics generally are represented by other PAGs that entail less uncertainty about causal structure.

We describe exemplars (and concrete examples) of these other PAGs now. Then we provide statistics about the relative frequency with which instances of these exemplars occur in our simulation results. This provides a broader picture of how predictive and causal inference can come apart when constructing aggregate variables to answer particular types of scientific questions.

Figure 4.23 provides the exemplar for a PAG structure that appears frequently in our simulation results; we call it "Y Structure #1," noting the so-called "Y Structure" formed by X_Cloud_1, X_Cloud_2, Y_Cloud, and T. In this PAG, a collider is oriented at Y_Cloud, and there are two unambiguously oriented edges: Y_Cloud → T and X_Cloud_2 → T. We provide a step-by-step explanation of how FCI infers Y Structure #1 in the Appendix (§7.2)[12], but provide three important factors here. First, X_Cloud_1 is independent of T given both Y_Cloud and X_Cloud_2. Second, X_Cloud_1 is unconditionally independent of X_Cloud_2. Third, X_Cloud_2 (however it is constructed) must not be independent of Y_Cloud (in contrast to the generating process isomorphic PAG). This could happen, in the example of independent raw X variables, by including one or more variables from the set $\{X0, X1, X2, X3, X4\}$ in the set of components for the X_Cloud_2 aggregate constructed variable.

Importantly, this shows that "grouping" raw variables for constructed aggregate variables in a way that is presumably unintuitive, given the generating causal structure, allows us to make unambiguous edge orientations when doing inference with the FCI algorithm. Similar "unintuitive" groupings of variables arise for other exemplar PAG causal structures that result from variables that maximize causal predictability metrics calculated over such structures. This raises important issues with respect to discussion from

---

[11] For larger graphs, we might only include the target's Markov blanket (i.e., the target's parents, children, and other parents of children in the PAG).

[12] Briefly, Y_Cloud → T is inferred by the "away from collider" rule after we have oriented the collider at Y_Cloud, and the X_Cloud_2 → T edge is inferred using the "away from ancestor" and "definite discriminating path" rules. For details about the logic of these orientations, see §7.2.

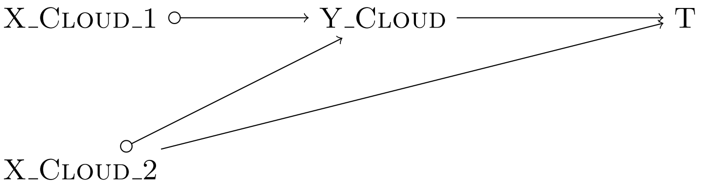

Figure 4.23: Y Structure #1 [Y #1]

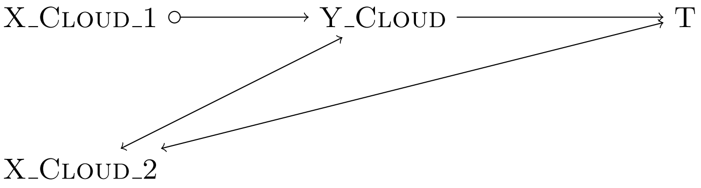

Figure 4.24: Y Structure #2 (Bi-Directed Edges) [Y #2]

Chapter 2 about the nature of scientific (natural) kinds and variables constructed to achieve particular scientific goals.

While the PAG Heuristic and Pattern counting schemes count Y_Cloud and X_Cloud_2 as causes of T in every DAG represented by "Y Structure #1" PAG, it is important to remember that the interpretation of PAG edges allows for the possibility that X_Cloud_2 is only an indirect cause of T. That is, we would infer this PAG also, for example, if there were a latent variable cause of X_Cloud_2 and T (with no direct edge between the two), X_Cloud_2 were a cause of Y_Cloud, and Y_Cloud were a cause of T. The Super DAG counting strategy includes such a PAG, as it iterates over all Super DAGs represented by a PAG.

We call the PAG structure in Figure 4.24 Y Structure #2. It is inferred under the same conditional independence conditions among constructed variables as Y Structure #1, except that in this case X_Cloud_1 is independent of T given Y_Cloud, but the two variables are not independent given both Y_Cloud and X_Cloud_2.

We call exemplar PAGs in Figures 4.25 and 4.26 the "X-Y Reversal Structure" and "Simple Y" Structure, respectively. They both occur relatively infrequently in our simulations, but both share, along with the other exemplars, the characteristic that they have unambiguously oriented edges into T.

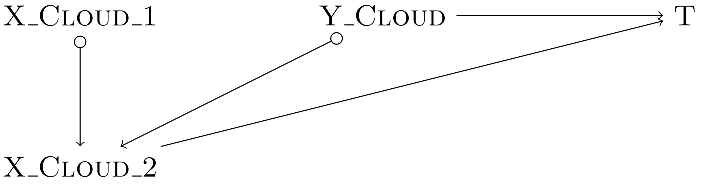

Figure 4.25: X-Y Reversal Structure [X-Y Reversal]

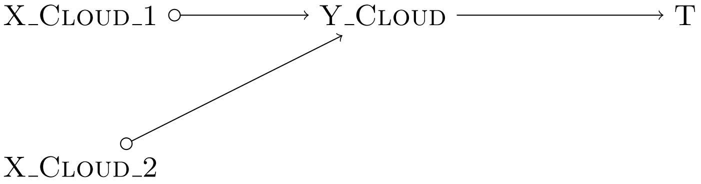

Figure 4.26: Simple Y Structure [Simple Y]

Figure 4.27 provides the frequency (as percentage of simulation instances) in which each of the exemplar causal PAG causal structures results from the set(s) of variables that maximize each of the three causal predictability metrics. In the Appendix to this chapter, we provide a breakdown, by "within cloud" and "inter-cloud" edge configurations in the generating process, of the frequency with which exemplar PAG causal structures result for each causal predictability metric.

In Figure 4.27, we see that generating process isomorphic PAGs result for the majority of simulation instances when considering the intervention causal R^2 metric and that Y Structures #1 and #2 are also prevalent for two causal predictability metrics each. This establishes Conclusion #10: inferred causal structures (i.e., inferred PAGs) for variables that maximize metrics of causal predictability in many cases do not match the PAG one might expect given the structure of the data generating process (i.e., generating process isomorphic PAGs).

The prevalence of the generating process isomorphic PAG for the intervention causal R^2 metric has an interesting explanation that calls for further investigation. Recall that we have paired this causal predictability metric with the Super DAG counting scheme for this simulation study. Iterating over Super DAGs on four variables represented by the PAG structure of Figure 4.21, for example, results in only one Super DAG (out of 23) in

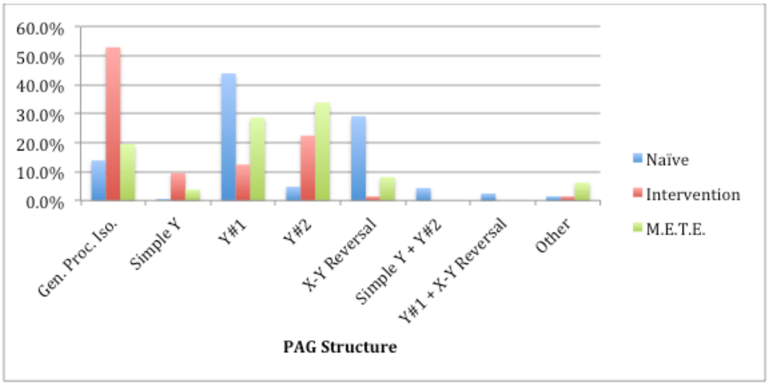

Figure 4.27: Percentage of Simulation Instances with Exemplar PAG Structure(s) by Causal Predictability Metric (n = 210)

which T does not have at least one of avg_Y0Y1Y2 or avg_X5X6X7X8 (Y_Cloud or X_Cloud_2) as a direct cause (though also confounded by a latent variable in some cases). Thus, the generating process isomorphic PAG, considered according to the computationally expensive Super DAG counting scheme, is not far removed, using the PAG Heuristic and Pattern counting schemes, from the other exemplar PAGs in "counting" direct causes of T in represented causal structure. Future research might shed light on when PAG causal structures (or particular edges in PAGs) might give rise to similar Super DAG counting situations.

4.2.5 "Cloud" Simulation Summary and Methodological / Practical Implications

This "cloud" simulation study establishes the unobvious conclusion that predictive and causal inference can "come apart" when constructing variables from raw data. Variables that maximize metrics of causal predictability need not match aggregate variables in the underlying data generating process, and causal structures that give rise to maximum values of our causal predictability metrics need not match the expected PAG structure over generating process aggregate variables.

Thus, when scientific questions call for inferring causal structure based on conditional independence relationships in observational data, if such data manifest as "raw" variables not well-suited to answer such questions, care should be taken to consider a variety of alternative, possible constructed variables relying on domain knowledge as well as the type of data-driven metrics (and modifications thereof) for causal predictability we have introduced. In some cases, it may turn out that variables constructed according only to their purchase of predictive access to a target variable are those that also maximize our ability to make causal inferences about the target, but these simulations demonstrate that this need not always be the case.

Further, these simulations raise interesting questions about the nature of differences between variables that support causal inference (e.g., help us to unambiguously orient PAG edges into a target variable) and variables that provide merely predictive access to the target, leaving greater uncertainty about causal structure. Recalling the discussion of natural kinds in Chapter 2, the reader may be tempted to interpret our results instrumentally. While such inferred variables might be useful for causal/statistical modeling, variables we infer to support causal inference cannot be candidate natural kinds because they do not match underlying generating process aggregate variables. However, it is not clear that such an inference is warranted.

While variables that support causal inference do not generally match aggregate generating process variables, they do reflect important aspects of the generating causal structure. For example, variables from which we infer Y Structure #1 are constructed from raw X variables in such a way that X_CLOUD_1 and X_CLOUD_2 are independent, but both are not independent of Y_CLOUD. Inspecting particular simulation instances reveals that the "groupings" of variables used in such constructions generally reflect the causal structure of generating processes (i.e., we expect that X_CLOUD_1 and X_CLOUD_2 would be independent but both dependent on Y_CLOUD given particular "groupings" of raw variables we find).

A priori, it is not clear we should privilege matching the generating process aggregate variables as the only way to judge having inferred useful scientific variables and/or natural kinds. Reflecting "true" characteristics of generating process causal structure might be just as (or more) important than capturing particular aggregates in the generating process to discover natural/scientific kinds. Alternatively, an instrumental approach that interprets our results to show that particular scientific questions call for particular (and different) ways of constructing variables raises important and interesting questions for future research. Indeed, we expect that differing choices of target phenomena or variables (e.g., at different levels of aggregation or salience) lead to different variables being appropriate for statistical and/or causal modeling. In summary, there is room for both realistic, instrumental, and hybrid interpretations of our results. Without further research in particular domains and more theoretical/simulation work, it is premature to stake a claim as to the best interpretation.

4.3 Chapter Summary

This chapter establishes ten substantive conclusions with respect to our proposed metrics of causal predictability and the use of those metrics to construct aggregate variables that support causal inference. We re-state these conclusions before considering several limitations of the present study and presenting future possibilities.

Our first four conclusions "match" each causal predictability metric to the counting scheme that best tracks the relevant "true" value in underlying generating processes we consider. In §4.1.3, we match the PAG Heuristic counting scheme to the M.E.T.E. metric in terms of best tracking the variable with maximum total effect. In §4.1.4, we match the Pattern counting scheme to the M.E.T.E. metric in terms of best tracking the value of the total effect of the variable chosen by the M.E.T.E. metric. In §4.1.5, we match the PAG Heuristic counting scheme to the naïve causal R^2 metric. In §4.1.6, we match the Super DAG counting scheme to the intervention causal R^2 metric.

In establishing these four conclusions, we establish our fifth that performance is better for inferred PAGs that are not complete (i.e., compared to complete, maximally qualitatively uninformative PAGs). Thus, we only consider constructed variables over which we learn PAGs that are not complete in §4.2 and in Chapter 5.

Further, we establish that performance is far better when the target variable is "final" (i.e., has zero out-degree) for qualitatively tracking the variable with maximum total effect in §4.1.3. We note that when the inferred PAG is not complete, the status of the target variable as final in the generating model generally improves performance, except for the case of the intervention causal R^2 metric.

Chapter 7 (this chapter's appendix) provides evidence for our seventh conclusion. With increasing numbers of latent causes of the target variable in the generating model, we generally see improved performance for all metrics; graphs over which we iterate by counting schemes tend to help us account for greater complexity in underlying generating models.

In §4.2, we provide evidence for three conclusions about constructed aggregate variables that support causal inference, judging such support using causal predictability metrics. In §4.2.3, we provide evidence for two. First, we find that constructed aggregate variables that maximize causal predictability metrics generally are not the same as the aggregate variables in the data generating process. Second, those variables that maximize causal predictability metrics also do not match those that maximize traditional prediction metrics, so causal and predictive inference sometimes "come apart" with respect to variable construction.

Finally, in §4.2.4, we show that inferred PAGs for variables that maximize metrics of causal predictability in many cases do not match the PAG expected given the structure of the data generating process.

4.4 Limitations and Future Work

Simulations in this chapter have several important limitations to address in future work. We briefly review several in this section before returning to most in Chapter 6.

First, we restrict ourselves in both §4.1 and §4.2 to PAG inference on only four (in §4.2, constructed) measured variables. We demonstrate in Chapter 5 that the PAG Heuristic counting schemes can be fruitfully deployed over a few more than four variables, but future work should address this limitation and test the limits of computational feasibility. One important issue is the problem of appropriately "counting" Super DAGs represented by a particular PAG. Scientifically reasonable simplifying assumptions (e.g., as to the number of possible latent variables and their configuration) should be sought to (possibly) make the problem tractable. Alternatively, Maximal Ancestral Graphs (MAGs; Richardson and Spirtes 2002) represented by a PAG might be considered to develop new metrics of causal predictability. We return to this idea and similar topics in Chapter 6.

In §4.2, we assumed that the target variable is causally "final," that the target variable does not have any effects among variables under consideration. In §4.1, we show that generally we do a better job of tracking causal predictability metrics when this is the case; performance, for most causal predictability metrics, degrades if the out-degree of the target is positive. We have run several ad-hoc simulations that indicate that the assumption is also important for the variable construction simulations of §4.2. If we alter the basic generating process such that T is a cause of the Y raw variables, we still tend to infer constructed variables for which we find edges into the target T from AVG_Y0Y1Y2. This is unfortunate because it does not reflect the underlying causal structure of the generating process. On the other hand, in the majority of cases we simulated we were not lead into complete error, as we found that, for example, the X_CLOUD_2 variable (in whatever way it was constructed) had the M.E.T.E., and, of course, the generating process aggregate variable with maximum total effect in these cases is AVG_X5X6X7X8. Future work might shed light on improvement to these metrics to better handle cases of target variables with positive out-degree. Moreover, the assumption of target variables that are causally "final," via background knowledge and time ordering constraints, is appropriate in a variety of applied settings (e.g., education where learning outcomes like post-tests, final exams, and standardized tests all take place after students' interaction with courseware, behavior in a course, and similar phenomena that we model as possible causes of learning outcomes).

Finally, among other limitations we consider in Chapter 6, in §4.2 we only consider one aggregation function (average), and the structure of the generating process is limited. In real world applications, we expect that salient aggregate phenomena may depend on a wide variety of functions or transformations of underlying raw data and that the underlying data will have possibly many thousands of variables. The present study merely scratches the surface of addressing important questions about causal modeling and large-scale data that are now common place in natural science, social science, and policy settings.

Chapter 5

Case Study: Causal Models of Gaming the System and Off Task Behavior in Intelligent Tutoring Systems

Researchers constructing student-level statistical and causal models from "raw" log data of education courseware must first construct variables to represent student-level, aggregate features of interest.[1] Rather than rely *solely* on investigators' intuitions and possibly ad-hoc choices, in this chapter we propose search for constructed variables to represent aggregate student behavior.

We use causal predictability metrics to simultaneously search for student-level variables constructed from intelligent tutoring system (ITS) log data and graphical causal models over those variables. Non-experimental, raw data are collected from adult learners enrolled in (online and on-campus) offerings of an algebra course at the University of Phoenix (UoP) to develop causal explanations of in-course behavior, including "gaming the system" and "off-task" behavior, much discussed in recent literature, in an ITS. Our results compare favorably to those we describe for baseline causal models of gaming the system and off-task behavior. We (qualitatively and quantitatively) compare causal models over: (1) variables described in previous studies (a baseline causal model), (2) previously described variables combined with a novel student-level, aggregate variable discovered by our method,

---

[1] Portions of this chapter appear in Fancsali 2012.

and (3) only novel variables discovered by our method. In Chapter 6, we suggest methodological avenues for future research as well as extensions for the online education and intelligent tutoring systems domains.

5.1 Related Work on Education Courseware and Intelligent Tutoring Systems

While the novelty of this work arises from the combination of search for constructed variables and graphical causal models, work in the educational data mining community has engaged both research programs individually. For example, recently, variable construction, or "feature engineering," techniques have been explored for modeling educational data (e.g., Arnold, et al. 2006; Baker, et al. 2008; Baker, et al. 2011; and others), including solutions to the KDD Cup 2010: Educational Data Mining Challenge (e.g., Pardos and Heffernan [in press]; Yu, et al. [in press]). Much of the focus of such work (though not without exception, e.g., Baker, et al. 2008) has been to develop predictive models for fine-grained target outcomes (e.g., whether a student will be successful at her next opportunity to practice a particular skill) rather than aggregate, student-level learning outcomes (e.g., course final exams or standardized test scores).

Similarly, graphical causal models have also been used to model a variety of educational data (e.g., Scheines, et al. 2005; Shih, et al. 2007; Rai and Beck 2010, 2011), both from ITSs and other educational courseware. Fancsali (2011a, 2011b) explored the roots of the approach adopted in this work. Generally, extant causal modeling of educational data assumes a set of aggregate, student-level features constructed from domain knowledge or (possibly data-driven) intuitions. We combine data-driven variable construction (i.e., "feature engineering") and algorithmic discovery of causal models to help explain student behavior in ITSs.

5.2 Cognitive Tutors

Our study focuses on student behavior and interactions, in a UoP algebra course, with the Carnegie Learning Cognitive Tutor (Ritter, et al. 2007), an ITS based on the ACT-R theory of cognition (Anderson 1990, 1993). Cognitive Tutor software probabilistically tracks student mastery of basic math skills using a framework called Bayesian Knowledge Tracing (Corbett and Anderson 1995). As students progress through course material, fine-grained logs are generated for thousands of individual interactions per student (e.g., hint requests, errors, whether student errors correspond to "known misconceptions" that trigger just-in-time feedback, etc.).

We seek to learn about aggregate student behavior in the Cognitive Tutor from its fine-grained logs. Specifically, we focus on "gaming the system" and off-task behavior in such environments. Such behavior has been the subject of a great deal of recent research and literature, but little work has been devoted to establishing whether relationships between gaming the system, off-task behavior, and learning outcomes are causal.

5.3 "Gaming the System" and Off-Task Behavior in Cognitive Tutors

Learners engage in "gaming the system" behavior in intelligent tutors by taking advantage of tutor properties to advance through course material without having to genuinely learn course material to provide correct answers (Baker, et al. 2006). While gaming the system behavior involves engaging the tutoring system (i.e., exploiting its properties), learners are off-task when they disengage from the tutoring system and behave in ways unrelated to learning tasks (Baker 2007). Both types of behavior have been associated with decreased learning (Baker, et al. 2004b). In general, gaming behavior falls into two categories (cf. Baker, et al. 2008):

1. "hint-abuse," when one rapidly proceeds through hints provided by the tutor, possibly until the "bottomed-out" (i.e., last available) hint provides the correct answer (Aleven and Koedinger 2000), and

2. systematic/rapid "guessing," when the student/user quickly enters a range of possible answers until the correct answer has been provided.

Some gaming is described as "non-harmful" and not associated with decreased learning (Baker, et al. 2004a, 2008). Much research on gaming uses the moniker "harmful" in a non-causal way; the term only denotes association with negative outcomes, without any implication of causation.[2] We seek to provide evidence for (or against) a causal relationship between gaming behavior (as well as off-task behavior) and learning.

---

[2] See, for example, discussion of "harmful" gaming the system behavior in Baker, et al. 2008: "The word 'harmful' is used for brevity and simplicity in discussion; there is still not conclusive evidence as to whether the relationship between harmful gaming and learning is causal or correlational" (Baker, et al. 2008, pg. 291, footnote).

With respect to "non-harmful gaming," one possible explanation for the lack of association between some gaming behavior and learning is that students have already mastered parts of course material (cf. Cocea, et al. 2009). Other work suggests that behavior that on the surface appears to be gaming may actually be associated with better learning; students may, for example, rely on "bottomed-out" hints to provide worked examples upon which they reflect to improve learning (Shih, et al. 2011).

5.3.1 Software "Detectors"

Several "detectors" of gaming the system and off-task behavior have been developed (e.g., Baker, et al. 2004a; Baker and de Carvalho 2008; Beck 2005; Walonoski and Heffernan 2006; Johns and Woolf 2006; Beal, et al. 2006). We deploy the gaming detector used in Baker and de Carvalho 2008 and the detector of off-task behavior used in Baker 2007. Both detectors use Latent Response Models (Maris 1995) applied to a variety of "engineered" or "distilled" features (Baker, et al. 2008) to code tutor transactions with a numerical value to which a threshold/step function is applied (values greater than 0.5 correspond to gaming or off-task behavior) to judge whether a transaction corresponds to gaming or off-task behavior. We treat output of such detectors as noisy, "fine-grained" observations of gaming and off-task behavior; we seek constructed variables to represent aggregate gaming and off-task behavior over these "fine-grained" observations, as (at least some of) the causal influence of such behavior on learning is presumably only in the aggregate.

5.3.2 Prior Work: Aggregate Predictive Models

Recent work (Cocea, et al. 2009) considers whether the influences of gaming the system and off-task behavior on learning are immediate, aggregate or both. We focus on their approach to the question of aggregate learning, as we seek student-level, aggregate causal models.

In Cocea, et al. 2009, linear regression models of aggregate learning are used to predict post-test scores for several Cognitive Tutor lessons. Variables are constructed by aggregating individual tutor transactions into steps (or "learning opportunities") that correspond to consecutive transactions associated with a single skill (VanLehn 2006). As with our gaming detector, the data in Cocea, et al. 2009 do not distinguish between "harmful" and "non-harmful" gaming, so they only work with transactions labeled as gamed or not and off-task or not. Steps are considered gamed if they contain at least one transaction that a detector labels gamed, and steps are considered off-task if they contain at least one transaction a detector labels off-task.

Aggregate variables are constructed for the number of gamed, non-gamed, off-task, and, in some cases, total steps over particular lessons. Using these variables and pre-test and post-test scores, they specify regression models to predict post-test results. For brevity, we omit their regression models; our primary concern is how they model aggregate gaming and off-task behavior and their overall conclusions. They report that gaming behavior is weakly associated with aggregate poorer learning and that off-task behavior is strongly associated with aggregate poorer learning (Cocea, et al. 2009, pg. 512).

5.4 Data

We focus on 102 non-traditional, adult learners who enrolled in 2010 in online and on-campus offerings of an algebra course at the University of Phoenix. We focus on data from the last module of the algebra curriculum used for this course. It is comprised of the following Cognitive Tutor Algebra units:

- Systems of Linear Equations
- Systems of Linear Equations Modeling
- Linear Inequalities
- Graphing Linear Inequalities
- Systems of Linear Inequalities

Our target learning outcomes are students' final exam scores for the course, where pre-test results are also available.

5.5 Variable Construction Search Space

The aggregate models of Cocea, et al. 2009 provide variables constructed over specific lessons of interest (e.g., "Scatterplots," "Percents") as counts of gamed and off-task steps. A plethora of alternative variables can be constructed, many of which plausibly represent student behavior in an intelligent tutor. More characteristics of students' steps (e.g., attributes calculated by gaming and off-task detectors; other characteristics tracked by Cognitive

Tutor logs) may provide insights into aggregate behaviors and causes of student learning. We consider alternative variable constructions to determine those that underwrite (better) causal and predictive models of aggregate, student-level learning. Having decided on a set of characteristics over which to construct variables, we must consider different levels of aggregation and aggregation functions.

5.5.1 Aggregation Level

We consider data from a module of an algebra course (broader in scope than individual lessons in Cocea, et al. 2009), ranging over several units (with corresponding sections) of material. Further there are 32 skills that students work to master. It is possible that student behavior in one of these units, sections, or skills is more important for aggregate student learning than behavior over the entire module. Identifying particular areas of course behavior that are causally related to learning would help direct future interventions on student behavior or course design.

Such an approach may allow us to pick apart aspects of “harmful” vs. “non-harmful” (vs. possibly “helpful”) gaming behavior, as differences in gaming may occur in different parts of a course. Students, for example, may have previously mastered early material and thus engage in “non-harmful” gaming early in the course, but as they progress similar behavior may have a detrimental effect on learning as the student engages new material. We investigate aggregation levels between individual transactions or steps and the entire module.

5.5.2 Aggregation Function

Aggregate models in Cocea, et al. 2009 only consider variables as *counts* of gamed or off-task steps. Other functions of characteristics of students’ steps might manifest important aggregate student behaviors. Rather than simply count a step as gamed or off-task because the detector labels one transaction within the step as such, “alternative aggregation functions... could be used, such as average and weighted average” (Cocea, et al. 2009, pg. 513). For example, one might consider the average number of gamed transactions per step (within a module, unit, section, or skill) as judged by a gaming detector. Other aspects of student behavior might be captured by other aggregation functions, such as the variance of “gamed” transaction (or step) counts. We consider several functions over a variety of characteristics detailed in §5.7.

5.6 Baseline Causal Model for UoP Data

We learn a causal model for our data using variables described by Cocea, et al. 2009. This provides a baseline for causal predictability to compare to that which we achieve with constructed variables discovered by our method.

We include the following student-level variables in our baseline causal model:

- MODULE4_PRETESTPERCENTAGE: percentage score on a pre-test for the module under consideration
- NUMBER_STEPS_OFFTASK: count of steps (i.e., "learning opportunities") in the module containing at least one off-task transaction
- TOTAL_NUMBER_STEPS: total count of steps through which a student worked in the module. This varies between students because students encounter different numbers of problems as the tutor adapts to their performance and assesses skill mastery.
- NUMBER_STEPS_GAMED: count of steps in the module containing at least one gamed transaction
- FINAL_EXAM: algebra course final exam score

We apply the FCI algorithm to these variables constructed over collected data, assuming that pre-test is prior in time to the rest of the variables and that the final exam comes after all other variables. The PAG from this search is provided as Figure 5.1. Average causal predictability is assessed by taking the average naïve causal R^2 value achieved over linear regression models specified according to each DAG compatible with this PAG. In this case, all compatible DAGs include NUMBER_STEPS_GAMED as a cause of FINAL_EXAM because of the unambiguously oriented edge between them, so we infer that the negative association between NUMBER_STEPS_GAMED (i.e., gaming behavior) and FINAL_EXAM (our learning outcome) is induced by a causal relationship between the two. The average naïve causal R^2 for this PAG is 0.5028.

Further, gaming behavior (i.e., NUMBER_STEPS_GAMED) "screens off" all other variables in our analysis from learning. Off-task behavior is independent of all the other variables under consideration, and students with better pre-test scores tend to have fewer overall steps. Gaming behavior is positively associated with the total number of steps through which students

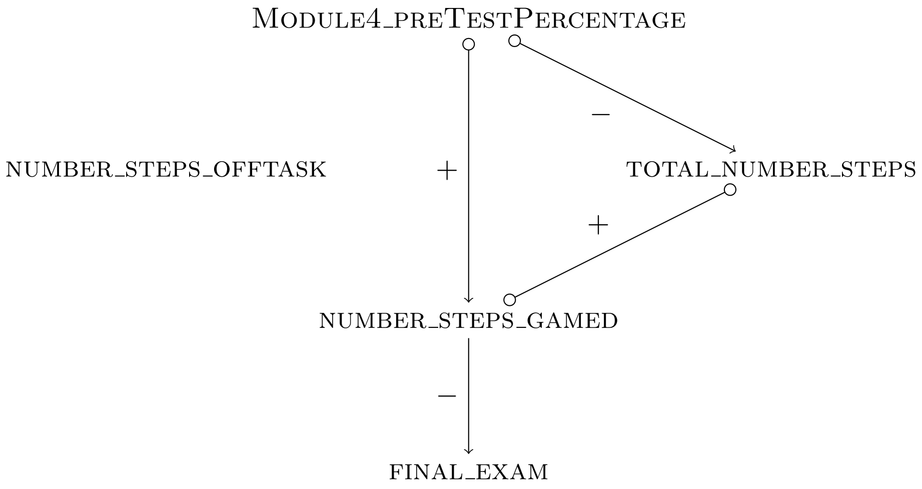

Figure 5.1: Baseline PAG model with variables described in Cocea, et al. 2009; edges are marked +/− according to the sign of parameter values in an estimated structural equation model.

work, suggesting that students that game the system also have to encounter more problems and/or skill-practice opportunities to master material.

This baseline model demonstrates a stronger association (due likely to a causal link) between aggregate gaming behavior and aggregate learning than that reported in Cocea, et al. 2009.[3] We do not find evidence that there is a link between aggregate off-task behavior and learning. Having established a baseline level of causal predictability, we now provide details of our approach to variable construction to underwrite causal discovery.

5.7 Constructed Variables and Search

Many characteristics of students' transactions and steps are tracked by Cognitive Tutor logs. Further, more characteristics are calculated by "detectors"

---

[3] The PAG of Figure 5.1 is different from that reported in Fancsali 2012 because it includes TOTAL_NUMBER_STEPS rather than NUMBER_STEPS_NOT_GAMED. Regression results reported by Cocea, et al. 2009 did not lead us to expect that gaming behavior would "screen off" non-gaming behavior from aggregate learning, but that is the result reported in Fancsali 2012. Despite small differences between the two models, results that gaming the system is a cause of decreased learning and that it "screens off" all other variables from FINAL_EXAM are common to both. Neither model provides evidence for a causal relationship between off-task behavior and learning.

as noisy, fine-grained measures from which we infer aggregate behaviors (e.g., gaming and off-task behavior). Rather than assume *a priori* that we know which characteristics we should use to construct variables to develop successful aggregate causal models, we consider search over variables constructed from a variety of characteristics. In our analysis, we consider the following step-level characteristics:

- counts of transactions: overall, correct, wrong, help request (i.e., hint request), "known bugs",[4] gamed, off-task
- proportions of transactions: correct, wrong, help request, "known bugs," gamed, off-task
- transaction time taken
- average gaming and off-task estimate (i.e., average over step transactions of raw inferred values from gaming and off-task detectors)

We also include counts of steps, gamed steps (+ proportions of gamed steps), and off-task steps (+ proportions of off-task steps), but we must ask questions such as: proportions relative to what? Over what part of the course are counts considered? To answer such questions, we construct aggregate variables by considering candidate solutions to the problems of aggregation level and aggregation function. We consider aggregating over the entire module, individual sections, units, and skills from the module, as well as two clusters of skills.[5] For each level of aggregation, we consider several candidate functions (when appropriate) of step-level characteristics to determine candidate constructed variables: sum, average, variance, max, and min.[6] For example, consider the variance of the count of help request transactions per step over steps associated with the "Systems of Linear Equations" unit; we represent such a constructed variable as:

UNIT_systems-of-linear-equations_var(count_help_request)

---

[4] "Known bug" transactions are those (tracked by the Cognitive Tutor) in which student exhibit known misconceptions with respect to course material. Such transactions trigger "just-in-time" hint responses from the tutor.

[5] The two skill clusters considered are for those skills that are rapidly learned and those that are unlikely to be learned; these categories/clusters are used by "feature distillation" techniques (Baker, et al. 2008) to prepare data for gaming and off-task detectors.

[6] Future work should provide means of determining whether particular transformations of constructed variables are appropriate (e.g., taking the logarithm of variables that are non-Normally distributed to possibly arrive at a constructed variable with a Normal distribution).

A general schema for constructed variable names is as follows:

LEVEL_LEVEL-NAME_FUNCTION(CHARACTERISTIC),

where LEVEL is the aggregation level (MODULE, UNIT, SECTION, SKILL, or a SKILL-CLUSTER's name), LEVEL-NAME is the name of unit, section, or skill (omitted if aggregation level is skill-cluster or module), FUNCTION is the (abbreviation for the) aggregation function, and CHARACTERISTIC provides the step-level characteristic or count of (gamed/off-task) steps.

By applying various functions to characteristics at different levels of aggregation, we "explode" a space of several hundred constructed variables.[7] To reduce this "exploded" set of constructed variables, we adopt a relatively simple approach to dimensionality reduction based on the maximization of causal predictability assuming the possible presence of unmeasured common causes. We first "prune" the "exploded" set down to a smaller, manageable set. Pruning removes uninformative and/or redundant variables; for highly correlated pairs of variables, we remove the variable with lower correlation to the target variable. We rank variables that "survive" based upon their correlation to the target. In our analysis, we "keep" the top twenty variables. Starting with a smaller set of top ranked variables, we iteratively, randomly select an even smaller set of variables for removal and insert other randomly selected variables from the larger post-pruning set. With each of several hundred iterations, we apply the FCI algorithm and calculate the value of causal predictability metrics afforded by the PAG inferred over the set of constructed variables. Our results describe the sets of constructed variables (and corresponding PAGs) that maximize each of the causal predictability metrics.

5.8 Results

Recall that in the last chapter, we resolved to pair each metric for causal predictability with the best, computationally feasible counting scheme. This resolution had us use the PAG Heuristic counting scheme for each of the three causal predictability metrics. We first present an improved, augmented version of the baseline causal model based on our results. We then specifically delve into details of results for each causal predictability metric, presenting constructed variables and corresponding PAGs for each. Recall that

---

[7] More sophisticated approaches are possible, some of which are considered in Fancsali 2011b. This is also an area for future research.

our target learning outcome is FINAL_EXAM, the final exam score for the UoP Algebra course under consideration.

5.8.1 Improving the Baseline Causal Model

One important variable we find by search in the manner we have described is an aggregate at the level of the entire module:

$$\text{MODULE\_SUM(COUNT\_KNOWN\_BUGS)}.$$

This variable represents a student's count of "known bug" or "misconception" transactions in the module. That the Cognitive Tutor tracks many misconceptions allows it to provide relevant, just-in-time feedback to the student about errors made. Before considering the full causal model for the constructed variables we discover, we first explore augmenting the baseline causal model by including MODULE_SUM(COUNT_KNOWN_BUGS) and using FCI for structure search with the same background knowledge deployed to infer the baseline model. The result is the PAG of Figure 5.2.

This augmented causal model provides us with greater causal predictability than the baseline model; the average[8] naïve causal R^2 value is 0.5816. We here identify a causal relationship between students' exhibition of misconceptions and learning. Further, we see that the causal relationship between gaming the system behavior and learning is mediated by the production of known misconceptions; the latter is a more proximate cause of learning in a causal chain. As may be expected, higher pre-test scores are associated with less off-task behavior, less gaming the system behavior, and fewer overall steps encountered as students progress through course material, suggesting it takes students less practice (e.g., fewer problems) to achieve skill mastery.

Variables at finer-grained levels of aggregation (i.e., at levels of individual skills, section, units, or skill-clusters), however, might help us to infer causal relationships among these in-course behaviors better.[9] We pursue this

---

[8] The astute reader notices that in every DAG compatible with the PAG of Figure 5.2, by the Pattern or PAG Heuristic counting scheme, MODULE_SUM(COUNT_KNOWN_BUG) is a cause of FINAL_EXAM, so the naïve causal R^2 value is constant across all compatible DAGs. Notably, this would also be the case for Super DAGs represented by the PAG according to the Super DAG counting scheme; we do not deploy the Super DAG counting scheme in our case study, however.

[9] Constructed variables at a finer-grained level of aggregation may also be easier (or more plausible) targets for conceivable future interventions to help "close the loop" from observationally inferred causal claims to effective interventions.

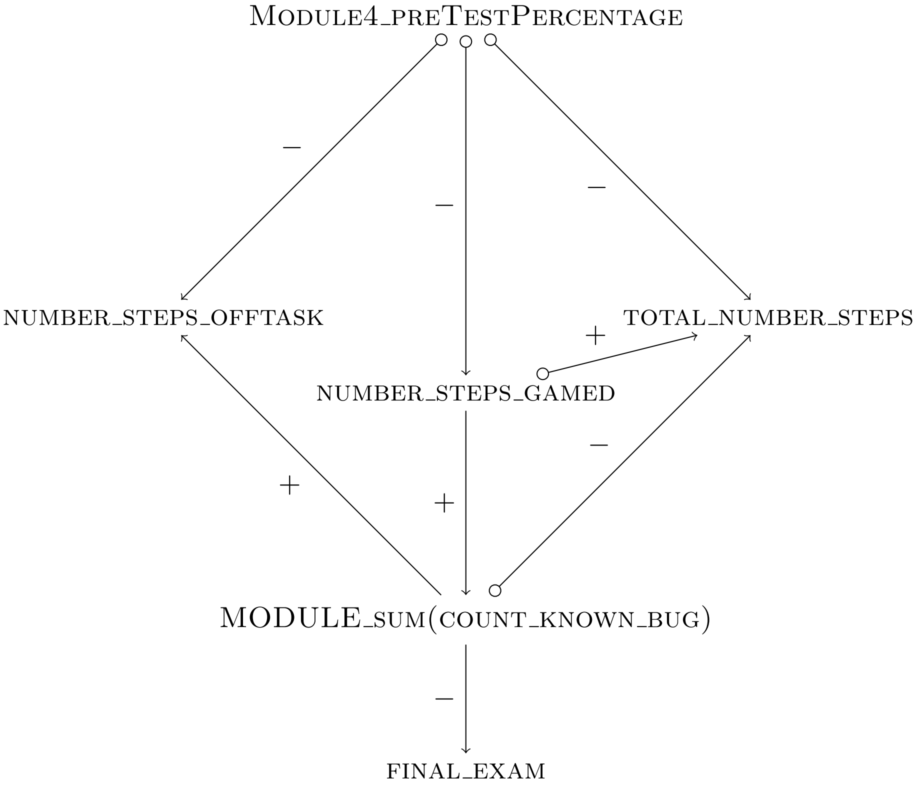

Figure 5.2: PAG resulting from FCI search over baseline constructed variables augmented by the module level count of student "known misconceptions" (i.e., "bugs").

question by considering the complete results of our search for constructed variables that maximize causal predictability metrics.

5.8.2 Naïve Causal R^2 Search Results

Before describing our results, we provide details of the training and test regime with which we test the extent to which models over constructed variables may over-fit the data. With only 102 learners in our sample, we decided not to "split" the data into a training set and test set based on members of the sample, but rather to split based on random sampling of students' steps. Our training set consists of variables constructed over 80% of steps randomly sampled over all 102 students. Our test set consists of those same constructed variables, but computed with the remaining 20% of steps for each student. We report training and test results for the naïve causal R^2 causal predictability metric, but only training set results for the other causal predictability metrics.

The PAG causal model, inferred on the training set, over the set of constructed variables that maximizes average naïve causal R^2 is illustrated in Figure 5.3 (average naïve causal $R^2 = 0.6107$). We assess model over-fit in two ways. Assuming the PAG structure learned from the training set, we calculate average causal R^2 achieved on the test set by averaging over regression models specified from compatible DAGs. We set parameters for these test set regression models in two ways: (1) parameters fixed as those learned in the regression for each DAG for the training set, and (2) parameters re-learned for test set data. With fixed parameters, we achieve average naïve causal $R^2 = 0.5431$ on the test set; re-learning parameters, we achieve average naïve causal $R^2 = 0.5504$. In either case, over-fitting does not appear to be a significant problem for our method.

Qualitative causal structure of the resulting model provides details as to possible causal linkages between actions and behaviors in the tutor and learning outcomes. We find one other possible direct cause of FINAL_EXAM,

$$\text{UNIT\_SYSTEMS-OF-LINEAR-EQUATIONS\_AVG(PCT\_KNOWN\_BUG)}.$$

This represents the average percentage of known bug/misconception transactions per step in the unit on systems of linear equations. The sign of its association is consistent with that for overall misconceptions, but this may point to an important unit of the course in which such behavior is prevalent. On statistical grounds, rather than on the basis of background knowledge or expert opinion (which were not provided as input to FCI), we

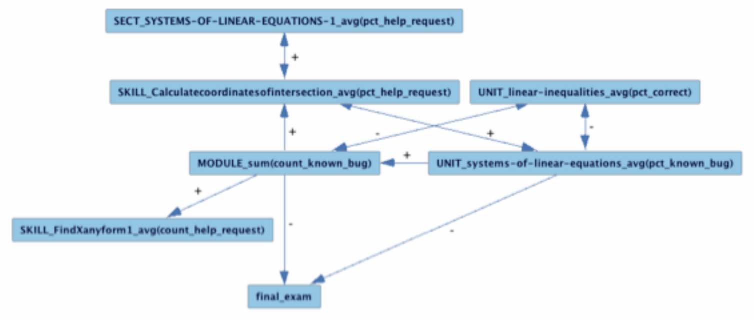
SECT_SYSTEMS-OF-LINEAR-EQUATIONS-1_avg(pct_help_request)
+
SKILL_Calculatecoordinatesofintersection_avg(pct_help_request)
UNIT_linear-inequalities_avg(pct_correct)
+
-
+
-
MODULE_sum(count_known_bug)
+
UNIT_systems-of-linear-equations_avg(pct_known_bug)
+
SKILL_FindXanyform1_avg(count_help_request)
-
-
final_exam

Figure 5.3: PAG and constructed variables that maximize average naïve causal R^2 using the PAG Heuristic counting scheme

orient MODULE_sum(count_known_bug) as an unambiguous cause of final_exam. This result is consistent with the improved baseline model.

5.8.3 Intervention Causal R^2 Search Results

Similar to the PAG model that maximizes average naïve causal R^2, we find MODULE_sum(count_known_bug) to be an unambiguously oriented direct cause of (and negatively associated with) final_exam in the PAG model, inferred without background causal knowledge or expert opinion, with variables that maximize average intervention causal R^2 (Figure 5.4). Maximum average intervention causal R^2 is 0.582.

Interestingly, we find that the average over steps of the average gaming estimate for transactions per step in a graphing skill, “Calculate Coordinates of Intersection,” is negatively associated with the following three variables:

- average number of help requests per step for those associated with the “Find X Any Form” skill,
- maximum number of misconceptions or “known bugs” over steps associated with the “Calculate Graph Characteristic” skill, and
- total overall exhibition of misconceptions in this module of the course.

On the surface this contrasts with our baseline and augmented baseline causal models, as we saw that “gaming the system” behavior in the aggregate is a cause of increased manifestation of “known misconceptions,” and this is a negative association. One possible explanation is that we have found an

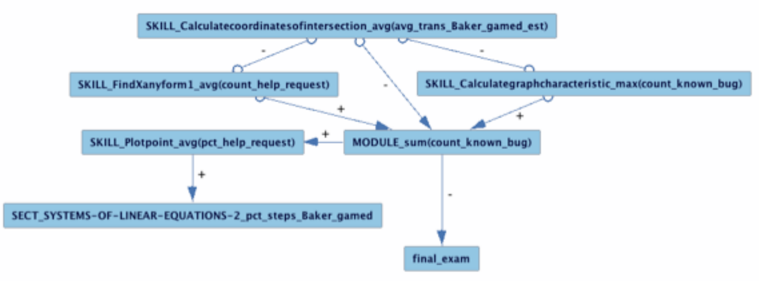
SKILL_Calculatecoordinatesofintersection_avg(avg_trans_Baker_gamed_est)
-
-
-
SKILL_FindXanyform1_avg(count_help_request)
SKILL_Calculategraphcharacteristic_max(count_known_bug)
+
+
+
SKILL_Plotpoint_avg(pct_help_request)
MODULE_sum(count_known_bug)
+
-
SECT_SYSTEMS-OF-LINEAR-EQUATIONS-2_pct_steps_Baker_gamed
final_exam

Figure 5.4: PAG and constructed variables that maximize average intervention causal R^2 using the PAG Heuristic counting scheme

example of "non-harmful" gaming; students may, for example, seek worked examples by reaching "bottomed out" hints (Shih, et al. 2011) in problems associated with this graphing skill that in turn decrease their reliance on hint/help requests for "Find X Any Form" and their incidence of manifesting "known misconceptions" in "Calculate Graph Characteristic."

5.8.4 Maximum Expected Total Effect Search Results

Our theory of non-harmful gaming is possibly bolstered by maximum expected total effect (M.E.T.E.) search results, presented in Figure 5.5. We find that the variable with M.E.T.E. is a variable that tracks the average over steps of the average gaming the system estimate[10] for transactions per step in a graphing skill, "Calculate Coordinates of Intersection." Its expected total effect = 0.626. As in Figure 5.3, we also find that a variable associated with "known misconception" dealing with systems of linear equations is a possible cause of FINAL_EXAM. Further, students' increased production of hint/help requests in another section on systems of linear equations are inferred as a cause of decreased learning. This negative association between increased hint requests and learning is consistent with recent findings (Shih 2011).

---

[10] Unfortunately, interpreting this variable is at best tricky. Baker provides a guideline whereby the gaming the system "estimate," generated by the latent response model at the heart of the detector, is used to say whether a particular transaction or step in the tutor is "gamed" or not, but there is no discussion of practically using the raw gaming estimate beyond the guideline that transactions with values ≥ 0.5 are judged "gamed."

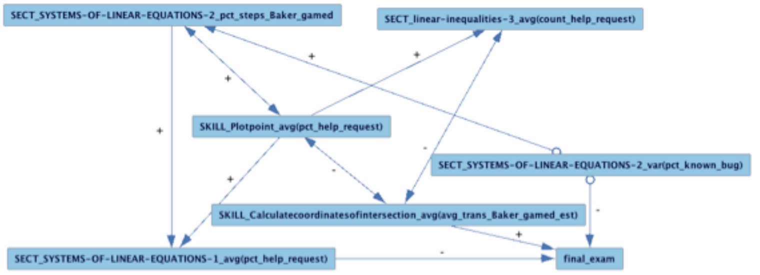
SECT_SYSTEMS-OF-LINEAR-EQUATIONS-2_pct_steps_Baker_gamed
SECT_linear-inequalities-3_avg(count_help_request)
SKILL_Plotpoint_avg(pct_help_request)
SECT_SYSTEMS-OF-LINEAR-EQUATIONS-2_var(pct_known_bug)
SKILL_Calculatecoordinatesofintersection_avg(avg_trans_Baker_gamed_est)
SECT_SYSTEMS-OF-LINEAR-EQUATIONS-1_avg(pct_help_request)
final_exam
+
+
+
+
+
+
-
-
-
-

Figure 5.5: PAG and constructed variables that provide maximum expected total effect using the PAG Heuristic counting scheme

5.8.5 Summary of Results

Our primary contribution establishes that at least one variable, directly tracked by Cognitive Tutor logs, mediates the likely causal relationship between "gaming the system" behavior and learning. Our results are thus consistent with the correlational findings in Cocea, et al. 2009, and they provide novel evidence for causal relationships using observational data. Finally, off-task behavior does not seem to be an important predictor of, or have a causal link to, aggregate learning, indicating that the association found in Cocea, et al. 2009 is not likely due to off-task behavior causing decreased learning.

These results raise a number of other questions. Our noted primary contribution comes from a causal model that integrates a variable found by a data-driven variable construction procedure with variables constructed from intuition and domain theory. This is perhaps a virtue. It will be important to improve the interpretability of variables in causal models that result from any data-driven variable construction procedure. Better procedures, perhaps agglomerating or clustering sections, units, and skills of course material, are likely to improve interpretability. We return to this problem in the next chapter.

Questions remain about how interventions might be directed at preventing "gaming the system" behavior that is harmful to student learning. Further, the constructed variable set for search can be refined in a number of ways that would enhance interpretability of our results. It seems likely that despite our efforts to aggregate over fine-grained logs at the levels of skills (+ two small clusters thereof), sections, units, and at the module level, there are likely other important, more interpretable aggregation levels that will be fruitful to pursue in future work.

Other important behavior, student affect (Baker, et al. 2012), and different target outcomes might also be integrated into a procedure similar to the one we outline to develop richer causal models of learning in ITSs. In the following chapter, we propose several extensions for future research, some of which are general and methodological and others that are specifically aimed at developing better causal models of learning in ITSs.

Chapter 6

Future Directions

In this chapter, we provide a host of future directions for this research program. We begin with several broad methodological concerns before briefly discussing how several other disciplines might benefit from this work. Finally, we propose several avenues for intelligent tutoring systems research and education research more generally.

6.1 Improved Counting Schemes

We presented three counting schemes in Chapter 3 that we evaluated in Chapter 4. We deployed the so-called PAG Heuristic counting scheme in Chapter 5. Despite the relatively comparable performance of the PAG Heuristic and Super DAG counting schemes under many circumstances, it is possible that other schemes may have improved performance with respect to ground-truth values of causal predictability metrics. We consider several.

6.1.1 MAGs + Canonical DAGs

One way in which the PAG Heuristic counting scheme is heuristic is that it counts DAGs "represented" by a PAG by a relatively rough interpretation of PAG edges. Of course, properly interpreted, PAGs do not represent sets of DAGs. Rather, they represent equivalence classes of Maximal Ancestral Graphs (MAGs). Spirtes and Richardson (2002) discuss causal and statistical reasoning with MAGs and also provide a procedure for inferring a "canonical DAG" from a MAG.

One possible new counting scheme generates the set of MAGs represented by a PAG and, for each MAG, generates the canonical DAG associated with it. We can then calculate causal predictability metrics over this set of (canonical) DAGs in the same way we do for the PAG Heuristic and Pattern counting schemes.

6.1.2 Local Algorithm for MAG Iteration

Maathuis, et al. (2009) provide an algorithm to iterate over a subset of DAGs represented by a pattern. The algorithm uses only local graph information, so it is computationally tractable, even on thousands of variables, for relatively sparse graphs. Such an algorithm might be extended to MAGs represented by a PAG to (possibly) make the proposal of §6.1.1 feasible even over large numbers of measured variables.

6.1.3 Graph Priors

We assumed a uniform distribution over DAGs and Super DAGs represented by a PAG, so the expected value of each causal predictability metric was just its average value over represented DAGs or Super DAGs. In many domains we have prior knowledge that allows us to place prior distributions over DAGs or Super DAGs represented by a PAG. Specifically, prior knowledge should at least provide graph characteristics that help to inform prior distributions for graphs. For example, we may know that some variable X is not a direct cause of Y because of an intermediary variable Z, so graphs with a direct edge between X and Y (and/or those without a directed path from X to Y passing through Z) might be assigned zero (or at least a relatively low) probability. This provides another way in which domain theory can inform the data-driven procedures we propose.

6.2 Causal Predictability Metrics

A multitude of alternative metrics (and modifications of proposed metrics) might usefully capture aspects of the extent to which a particular (set of) variable(s) provides causally predictive access to a target variable. We consider several.

6.2.1 Minimum Total Effect

On several occasions, we noted that Maathuis, et al. (2009) pursue a dimensionality reduction strategy based on inferred patterns over high dimensional data. They propose to choose variables from a large set of variables based on the value of their minimum total (causal) effect; variables with relatively large values for their minimum total effect over the set of values calculated from DAGs represented by a pattern are those that should be the focus of future interventions. The metric we proposed based on total effect had us take the set of variables with the single variable with maximum expected total effect, but we might have also considered the set of variables that provides the single variable with the largest minimum value for the total effect over DAGs or Super DAGs represented by a PAG. This strategy would seem to represent a relatively risk averse strategy for choosing sets of constructed variables.

6.2.2 Maximum Total Effect

The opposite of the minimum total effect causal predictability metric would have us seek the set of constructed variables containing the single variable that has the largest single value for the total effect over all DAGs or Super DAGs represented by a PAG. This represents a "high-risk, high-reward" strategy to judging sets of constructed variables.

6.2.3 Regression Model BIC Score

Proposed metrics that called for maximizing regression model R^2 values do so without judging whether small increases in (adjusted) R^2 are statistically significant. Raftery (1995) proposes the calculation of an approximate Bayesian Information Criterion (BIC) score (Schwarz 1978) from a regression model, suggesting that a difference of 10 in this BIC score is roughly equivalent to a statistically significant difference in model fit for an appropriate statistical test at the $\alpha = 0.05$ significance level. One could use this BIC score for regression models specified according to the causal structure of DAGs or Super DAGs represented by a PAG as an alternative causal predictability metric.

6.3 Aggregation Functions Other than Average

In Chapter 4, we only considered variables constructed as the average of hypothetical "raw" variables. Models in Chapter 5 included variables constructed according to other functions, e.g., sum, maximum, minimum, and variance. These functions capture important features of interest for a variety of domains and applications, but whether specific care need be taken to construct variables according to these functions remains a question for future research.

6.4 New Domains

6.4.1 Economics

Many commonly reported economic statistics (e.g., unemployment and inflation rates) are generally calculated from fine-grained raw and survey data. The United States Bureau of Labor Statistics' Consumer Price Index (CPI), for example, is calculated according to prices consumers pay for a basket of a wide variety of goods over disparate economic regions. The CPI "measures inflation as experienced by consumers in their day-to-day living expenses" (U.S. Bureau of Labor Statistics 2013). Such (often complex) calculations are determined by expert knowledge and long-standing convention, but insofar as causal inferences about the impacts of policy on aspects of inflation (or unemployment rates, etc.) are concerned, data-driven analysis of index construction is a target for future research.

Similarly, data from the United States Census pose many problems for construction of variables to represent features of (at least) economic, sociological, and political interest. Raw census data have fine (and irregular) granularity. Consider the representation of age in census data: "age brackets" are given at levels such as "number of individuals of age 21," "number of individuals aged 22 through 24," (in a particular census "tract" or voting districts comprised of multiple tracts) and so on. Including similarly fine-grained measures of other socio-demographic categories, the number of dimensions quickly becomes large relative to the sample size. Thus, the data are not immediately amenable to analysis by traditional statistical techniques. Further, it is not immediately apparent how variables at such a "fine" level could be meaningfully interpreted in a causal or predictive model.

We have explored the possibility of identifying predictors and causes of voting behavior from census data, particularly focusing on data from the 1990 U.S. Decennial Census for voting districts in Allegheny County, Pennsylvania, and the outcome of a referendum vote on the 1997 Regional Renaissance Initiative. We have made some progress in constructing variables that provide information about aggregate characteristics of interest, including age, education, and income levels as possible causes of voting behavior, but further work on such data (and its causal interpretation) is necessary. We omit detailed analyses for brevity, merely noting that the problem remains outstanding.

6.4.2 Ophthalmology

An open problem in ophthalmology is to infer predictors and causes of glaucoma by discovering salient regions of the eye from high-dimensional raw data. Optical coherence tomography (OCT) provides physicians with access to fine-grained, “raw” measures of a variety of ocular characteristics. Previous predictive modeling work focuses on characteristics of the optic nerve head (ONH) as well as the retinal nerve fiber layer (RNFL) of the eye (Burgansky-Eliash et al. 2005; Townsend et al. 2008).

One might also seek to discover (characteristics of) regions of the eye that cause glaucoma. After identifying key regions of the eye, including physiological and other factors in models might help determine whether some factors play greater causal roles (or perhaps mediate causal relationships) to better understand the disease. Even in the absence of possible direct interventions, identifying ocular regions and relevant characteristics that are causes of glaucoma will lead to important advances in early detection and treatment.

Previous attempts to predict glaucoma with machine learning classifiers modeled regions of the eye by dividing it into arbitrary regions (e.g., quadrants or “clock hours”) (cf. Burgansky-Eliash et al. 2005; Townsend et al. 2008). We might seek to discover less arbitrary, scientifically plausible regions of the eye using data-driven technique like those we have introduced, improving causal explanations of the disease and our ability to predict and treat it.

6.5 Refining Intelligent Tutoring System Construction Space

As noted, some constructed variables we considered in Chapter 5 lacked straightforward interpretability. Refining the space of constructed variables will enhance interpretability and allow us to represent a broader range of semantic features and student behavior. We consider two ways of making such refinements; many other possible refinements remain.

6.5.1 Skill Clustering

Ritter, et al. (2009) cluster skills from Cognitive Tutor mathematics curricula based on parameters that the tutor uses to assess student mastery of those skills. They find that 23 sets of parameters broadly represent the space of thousands skills tracked in the curricula. Rather than attempt to model behavior characteristics over particular skills, as we did in Chapter 5,[1] we might pursue a similar clustering strategy to represent student behavior over swaths of similar skills. We might also group together sections or units into clusters in a data-driven manner to better (interpretably) represent student behavior. In sum, future research should attempt to aggregate student behavior at broader levels of aggregation than we attempted in Chapter 5.

Such an approach is supported by high levels of inter-correlation among many behavioral variables representing particular sets of sections and skills; the pruning procedure we used for analysis in Chapter 5 eliminated highly correlated constructed variables by "keeping" variables with greater correlation to the target variable when a pair of constructed variables were found to be highly correlated. Rather than rely on such a procedure, clustering by skill, section, and/or unit should better capture interpretable behavior patterns over parts of the mathematics curricula.

6.5.2 Strategic Behavior

Shih (2011) describes clustering techniques based on Hidden Markov Models to infer instances of strategic behavior (e.g., with respect to hint-seeking and repeated attempts at particular problem steps) from response data for students in intelligent tutoring systems. From instances of strategic behavior, we can construct aggregate variables as different functions of these behaviors in tutoring systems (e.g., proportion of time rapidly attempting the same step of a problem or the variance in such behavior over a set of skills). Such techniques will provide for a broader range in the semantics of constructed aggregate variables that may be causes of learning and other outcomes.

6.6 Generalizing Intelligent Tutor Causal Models

The sample of adult learners considered in Chapter 5 is novel for analyses of Cognitive Tutor data; historically these tutors have been deployed in middle school and high school environments. Data we collected were from among the first deployments of such tutors for adult learners in a post-secondary (online and ground-campus) education environment. As such, inferences about gaming the system and off-task behavior may be limited to

---

[1] We considered two simple clusters of skills in the variable construction space in Chapter 5. Here we advocate a more sophisticated clustering of skills.

this particular population of students. Assuming we have data for appropriate learning outcomes (e.g., final exam scores) we can expand these models and analyses to large amounts of available data for students in grades 6-12 to get a sense of how well the models (and inferences about behavior based on them) generalize.

Further, we only consider one module of the Cognitive Tutor Algebra curriculum, so analyzing data for a broader portion of the curriculum is necessary. Finding early indicators (and/or causes) of decreased student learning would help to provide better remediation, either by identifying students in need of such remediation, or by actually suggesting specific interventions. We might also consider other curricula (e.g., Cognitive Tutor Algebra II or Geometry) and other intelligent tutoring systems (e.g., ASSISTments; Razzaq, et al. 2005).

6.7 Different Intelligent Tutor Targets

We only considered one target in the case study of Chapter 5, a final exam score for an algebra course at the University of Phoenix. Inferring causes of a variety of other behavioral targets and learning outcomes may be important in future education research.

6.7.1 Assistance Score

Noting large correlations among many constructed variables we considered at the module level of aggregation, we considered whether some module-level constructed variables might plausibly be indicators of latent variables as one first step toward the type of refinements of the construction space we advocate in §6.5. Contrast this attempt to infer indicators of latent variables with our approach in much of this work to infer aggregate variables representing (at least roughly) observable behavior.

Toward the goal of possibly discovering "pure" measurement model(s) (Spirtes, et al. 2000) for latent variables as discussed briefly in Chapter 2, we deploy the Build Pure Clusters (BPC) algorithm (Silva, et al. 2006) over the set of module-level, aggregate constructed variables defined by the sum function. The BPC algorithm judges that three variables may provide a pure measurement model for one latent variable: the module-level sum of the number of help/hint requests, the sum of incorrect/"wrong" transactions, and the sum of incorrect transactions that correspond to "known misconceptions" for which the tutor provides just-in-time feedback.

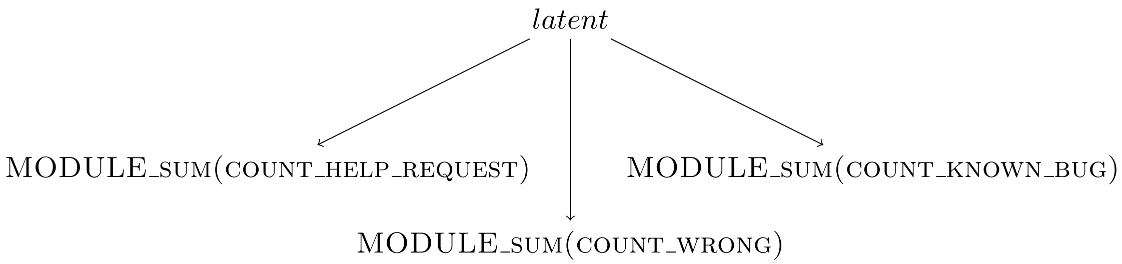

Figure 6.1: Graphical representation of pure measurement model for a single latent variable with three measured indicators.

For these three variables in the data we consider in Chapter 5, Cronbach's alpha is 0.97. Recall in Chapter 2 we summarized work (Fancsali 2008) that showed that, given such a large value of Cronbach's alpha and a pure measurement model, we could construct a new variable, by taking the sum of these indicators, that will preserve conditional independence relationships into which the underlying latent variable enters.

Figure 6.1 provides a graphical representation with variable *latent* as a stand-in for the latent variable that could explain the correlations among these three module-level, aggregate constructed variables. We posit that the variable (or scale) constructed as the sum of these three module-level sum variables is a good proxy for *latent*. We now consider one way to reify the variable *latent*.

The Pittsburgh Science of Learning Center's DataShop repository (Koedinger, et al. 2011) defines (i.e., its web application calculates) an "assistance score" over student practice opportunities. The assistance score is the sum of the number of student incorrect attempts and hint requests and has been deployed in intelligent tutoring research (cf. Hausmann and VanLehn 2010). The way we carved the variable construction space in Chapter 5, the total number of incorrect attempts is the sum of the number of incorrect/"wrong" transactions and the number of "known misconception" transaction. Thus, what we find in a purely data-driven fashion corresponds to the DataShop's assistance score.

While the assistance score has a natural (obvious) interpretation as quantifying the extent to which a student requires assistance in practicing a particular skill or working through a part of a course, we also provide data-driven justification that the assistance score would be a good proxy for an underlying latent variable for something like "assistance required," or possibly "difficulty experienced." This provides some evidence that carving the space of constructed variables as we did (i.e., into hints, "wrong" transaction, and "known bug" transactions) may be too fine-grained.[2] Using a variable like the assistance score provides a justifiable refinement to the search space. Future work can explore this insight further, perhaps deploying the assistance score as a target for causal modeling.

6.7.2 Gaming the System as Target

Having causally linked gaming the system behavior to decreased learning, another natural target to consider in future models is gaming the system behavior itself. This will be important to determine possible targets for future interventions aimed at decreasing such behavior, when it is deemed harmful. While there may be important behavioral causes of gaming the system behavior we can discover, another interesting area for future work will focus on student affect (e.g., boredom, confusion, etc.) as a cause of harmful gaming the system behavior and student learning.

6.8 Student Affect

Pardos, et al. (2013) point to research on how student affect is associated with differences in learning (cf. Craig, et al. 2004; Pekrun, et al. 2002; Rodrigo, et al. 2009). A natural extension of our work is to develop causal models of student affective states and emotions as causes of behavior and learning outcomes.

6.8.1 Detectors of Affect

Recent work introduces software detectors of student affect similar to those we deploy for gaming the system and off-task behavior. Such detectors link patterns in tutor logs to observations from field trials, using machine learned classifiers to predict whether students were coded by observers to be bored, confused, frustrated, or in a state of engaged concentration (Baker, et al. 2012). Such detectors have also been used to predict standardized test scores (Pardos, et al. 2013). Future work will deploy affect detectors to develop

---

[2] Whether or not a "carving" of the space of variables is too fine-grained is also likely to hinge upon goals for one's investigation. For example, if an investigator is only interested in hint-seeking behavior, then it seems appropriate to construct variables that count only hint requests (or possibly more fine-grained data about hint requests like student progression through 1st, 2nd, and 3rd hints, and so on).

causal models of affective states and their relations to gaming the system behavior, off-task behavior, and student learning outcomes.

6.8.2 Latent Variables vs. Aggregate Behavior Variables

Developing causal models of student affect and learning will have to consider contrasts between prior work on gaming the system and off-task behavior detectors and novel affect detectors. While the mechanics of software "detectors" (i.e., fine-grained, features "distilled" from tutor logs from which detectors infer behavior and affect) are similar, underlying phenomena they detect are fundamentally different. Gaming the system and off-task behavior are aggregate (if only indirectly measured or inferred) behaviors while student affective states are explicitly latent variables. As is well known, special care must be taken in constructing causal models involving latent variables; we advocated one particular approach in Chapter 2 and briefly again in §6.7.1. Given appropriate circumstances, we might justifiably construct scale variables as proxies for latent variables that are similar to variables we construct to represent aggregate behavior, or we might explicitly model student affect with latent variables in structural equation models. Determining the appropriate methodological course of action is the subject of future research.

6.9 Causal Models of Student Affect, Behavior, and Learning

The over-arching goal of future methodological and educational research we suggest is to develop integrated causal models of student behavior, affect, and learning. Such models will provide rich explanations of how student emotions like boredom and frustration, while working with intelligent tutoring systems, might cause particular types of behavior like gaming the system (or vice versa) and how both may (or may not) be related to learning outcomes. Such work necessarily combines all three situations introduced in Chapter 2 with respect to the types of "raw" data we increasingly face in real world social science applications. In this way, proposed future work brings the work of this dissertation "full circle." We provide one final important goal for future research, namely that we "close the loop" by putting research results into experimental (and everyday) practice.

6.10 Putting Causal Models to Work in Practice: "Closing the Loop"

We endeavor to explain causes of student learning so that we can improve systems designed to facilitate student learning. Knowledge of the causes of learning informs ways in which instructors, administrators, curriculum developers, and software engineers develop interventions to enhance student learning. Ideally, we "close the loop" from causal inferences made from observational data to develop experiments that test whether our inferences are correct and/or whether we have efficaciously intervened to enhance student learning. Some work (e.g., Baker 2005, Baker, et al. 2006) seeks to actively intervene by discouraging gaming the system behavior with an animated character that reacts to student interactions with a tutor.

Future work might develop interventions that, in addition to impacting particular student behaviors, will help to manage student affect and emotion (e.g., alleviate frustration, decrease boredom, etc.), assuming research points to various affective states as causes of decreased learning or of behavior that is detrimental to learning.

Another important area for experiments that "close the loop" will be to identify important sections / units / skills on which to focus curriculum development attention. The framework we have developed for thinking about different aggregation levels for constructed variables (combined with refined constructed variable search spaces) might help illuminate places in which faults in course material lead to sub-optimal student outcomes. For example, we might find that a particularly flawed portion of problems associated with a particular set of skills creates a great deal of student frustration. Data-driven guidance for curriculum improvement is yet another important area for future research.

6.11 Closing Thoughts

Answers to questions of how to improve student learning generally turn on matters of causation. This work scratches the surface of key problems vital to inferring causal relationships from observed, fine-grained, "raw" log data that is increasingly available from intelligent tutoring systems and other (educational) systems. Important questions about modeling and interpreting such data, and developing appropriate interventions to enhance learning (or other outcomes), will provide years of work for education researchers, computer scientists, and others.

Beyond education, questions we ask about constructing aggregate variables from fine-grained data, and especially inferring causes from such data, will only become more important as various technological innovations provide for "big data" applications in a plethora of natural and social science disciplines and public policy settings. We look forward to continuing this necessary work.

Chapter 7

Appendix to Chapter 4

7.1 Stratifying Error by Latent Causes of Target in Generating Super DAG

7.1.1 Maximum Expected Total Effect: Tracking Qualitative Performance

Table 7.1 provides qualitative counting scheme performance, stratified by the number of latent causes of X4 in the generating Super DAG, for the sampling strategy of generating Super DAGs not represented by the complete PAG. We see that the performance of the Super DAG and PAG Heuristic counting schemes are similar (close to identical), with the Super DAG counting scheme barely outperforming the PAG Heuristic scheme in cases in which X4 has 2 and 3 latent causes. The reader may find it surprising that the Pattern counting scheme does not outperform the other two in cases in which there are no latent causes of X4.

Table 7.2 provides the same information as Table 7.1 for sampling from generating Super DAGs represented by the complete PAG without an exhaustive Super DAG counting scheme, as does Table 7.3 for generating Super DAGs represented by the complete PAG with an exhaustive Super DAG counting scheme. The relative performance of each counting scheme is similar in all three cases: the PAG Heuristic counting scheme in most cases outperforms (by a small margin) the Super DAG and Pattern counting schemes. Notably, in both simulation strategies for generating Super DAGs represented by the complete PAG, all three counting schemes perform similarly and relatively poorly (regardless of the degree to which the target is confounded by latent variables) compared to their performance on

Simulation Strategy #1: Confounding of Target in Generating Super DAG (Incomplete PAG)		**Counting Scheme: Correct Count (%)**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	6,911	3,404 (49.3%)	3,424 (49.5%)	3,140 (45.4%)
1	14,136	5,391 (38.1%)	5,418 (38.3%)	4,733 (33.5%)
2	9,444	3,265 (34.6%)	3,243 (34.3%)	2,835 (30%)
3	1,509	926 (61.4%)	913 (60.5%)	842 (55.8%)
0-3	32,000	12,986 (40.58%)	12,998 (40.62%)	11,550 (36.1%)

Table 7.1: Qualitative Match Counts for M.E.T.E. Variable and True Maximum Total Effect Variable: Generating Super DAGs Not Represented by Complete PAG, Stratified by Number of Latent Causes of Target (X4).

generating Super DAGs not represented by the complete PAG.

7.1.2 Maximum Expected Total Effect: Tracking Quantitative Value

Table 7.4 provides mean squared error for inferred M.E.T.E. values compared to the true maximum total effect for generating Super DAGs not represented by the complete PAG. Stratifying generating Super DAGs by the number of latent variable causes of the target, the results are consistent with the overall result that the Pattern counting scheme outperforms the other counting schemes, but each counting scheme has roughly similar performance at each level of stratification. Interestingly, error decreases as the number of latent variable causes of the target increases, even for the Pattern counting scheme.[1]

---

[1] We provide one (possibly counter-intuitive) observation that partially explains this phenomenon. The Pattern counting scheme averages over fewer DAGs when there are more latent causes of X4 because we are more likely to infer oriented edges (→) at X4 in

Simulation Strategy #2: Confounding of Target in Generating Super DAG (Complete PAG)		**Counting Scheme: Correct Count (%) (Non-Exhaustive Super DAG Scheme)**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	884	89 (10.1%)	89 (10.1%)	69 (7.8%)
1	5,592	1,568 (28%)	1,542 (27.6%)	1,487 (26.6%)
2	8,010	1,622 (20.3%)	1,691 (21.1%)	1,547 (19.3%)
3	3,514	537 (15.3%)	556 (15.8%)	528 (15%)
0-3	18,000	3,816 (21.2%)	3,878 (21.5%)	3,631 (20.2%)

Table 7.2: Qualitative Match Counts for M.E.T.E. Variable and True Maximum Total Effect Variable: Generating Super DAGs Represented by Complete PAG, Non-Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #3: Confounding of Target in Generating Super DAG (Complete PAG)		**Counting Scheme: Correct Count (%) (Exhaustive Super DAG Scheme)**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	17	2 (11.8%)	1 (5.9%)	2 (11.8%)
1	136	29 (21.3%)	32 (23.5%)	28 (20.6%)
2	171	41 (24%)	43 (25.2%)	40 (23.4%)
3	76	10 (13.2%)	13 (17.1%)	11 (14.5%)
0-3	400	82 (20.5%)	89 (22.3%)	81 (20.3%)

Table 7.3: Qualitative Match Counts for M.E.T.E. Variable and True Maximum Total Effect Variable: Generating Super DAGs Represented by Complete PAG, Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #1		**Counting Scheme: MSE of M.E.T.E. Variable T.E.**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	6,911	.0773	.0769	.0700
1	14,136	.0745	.0742	.0660
2	9,444	.0576	.0597	.0538
3	1,509	.0453	.0476	.0419
0-3	32,000	.0687	.0692	.0621

Table 7.4: Mean Squared Error for Maximum Expected Total Effect Variable's Estimated Total Effect Value vs. True Total Effect: Generating Super DAGs Not Represented by Complete PAG, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #2		**Counting Scheme: MSE of M.E.T.E. Variable T.E.**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	884	.0818	.0888	.0832
1	5,592	.1309	.1240	.1224
2	8,010	.0958	.1002	.0971
3	3,514	.0751	.0840	.0832
0-3	18,000	.1020	.1039	.1016

Table 7.5: Mean Squared Error for M.E.T.E. Variable's Estimated Total Effect Value vs. True Total Effect: Generating Super DAGs Represented by Complete PAG, Non-Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

Tables 7.5 and 7.6 demonstrate again that performance degrades for generating Super DAGs represented by the complete PAG. The Pattern counting scheme performs best overall in both "complete PAG" cases, but for individual levels of stratification we see that the Super DAG and, in one case, PAG Heuristic are sometimes best, while performance across counting schemes does not substantially differ within each level of stratification. As in Table 7.4 (i.e., with generating Super DAGs not represented by the complete PAG), we see that performance is better for generating Super DAGs with a greater number of latent variable causes of the target. This suggests that Super DAGs and/or DAGs we consider for each counting scheme (whether exhaustively or be random sampling for the Super DAG counting scheme) in the case of the complete PAG lead to a kind of bias toward accounting for greater confounding of the target variable. When the ground truth is relatively simple (e.g., zero or one latent causes of the target), performance is hampered.

---

corresponding patterns (and o→ edges in PAGs), as opposed to more un-oriented pattern edges (and o−o PAG edges) when there are zero latent causes of the target. When there are many latent causes of the target, we end up with more causal information about the target variable that appears to improve performance.

Simulation Strategy #3		**Counting Scheme: MSE of M.E.T.E. Variable T.E.**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	17	.0711	.0645	.0668
1	136	.1093	.1128	.1049
2	171	.1070	.1087	.1058
3	76	.0763	.0855	.0830
0-3	400	.1004	.1038	.0995

Table 7.6: Mean Squared Error for M.E.T.E. Variable's Estimated Total Effect Value vs. True Total Effect: Generating Super DAGs Represented by Complete PAG, Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4).

7.1.3 Naïve Causal R^2

Tables 7.7 through 7.9 provide performance results for mean squared error of average naïve causal R^2 compared to the generating SEMs naïve causal R^2 stratified by the number of latent causes of the target variable in the generating model. Again, performance generally improves (though not without exception) as the number of latent variables causes of the target increases, and the PAG Heuristic and Pattern counting schemes have relatively similar performance in most cases.

7.1.4 Intervention Causal R^2

Consistent with results for the other metrics, the PAG Heuristic and Pattern counting schemes do better than the Super DAG counting scheme in Tables 7.10-7.12 for generating Super DAGs in which the target has no latent causes. This is plausible, as simpler counting schemes (e.g., the Pattern counting scheme) should be appropriate in the absence of latent causes of the target variable. For calculating the Intervention Causal R^2 metric, it appears that introducing just one latent cause of the target is sufficient to require the use of the Super DAG counting scheme to achieve best performance. For the two metrics that do not model ideal interventions on causes of the target, this is not generally the case, demonstrating a possibly important difference between the intervention metric and the others. Notably,

Simulation Strategy #1		**Counting Scheme: MSE of Average Naïve Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	6,911	.1269	.0533	.0718
1	14,136	.1431	.0609	.0622
2	9,444	.0663	.0573	.0627
3	1,509	.0393	.0440	.0383
0-3	32,000	.1120	.0573	.0633

Table 7.7: Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value: Generating Super DAGs Not Represented by Complete PAG, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #2		**Counting Scheme: MSE of Average Naïve Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	884	.1687	.1053	.0697
1	5,592	.1805	.1227	.0853
2	8,010	.1419	.1095	.1004
3	3,514	.1217	.1058	.1222
0-3	18,000	.1513	.1127	.0985

Table 7.8: Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value: Generating Super DAGs Represented by Complete PAG, Non-Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #3		**Counting Scheme: MSE of Average Naïve Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	17	.1475	.1160	.1062
1	136	.1490	.1078	.0818
2	171	.1599	.1074	.0915
3	76	.1187	.1043	.1108
0-3	400	.1479	.1073	.0925

Table 7.9: Mean Squared Error for Average Naïve Causal R^2 Value vs. True Naïve Causal R^2 Value: Generating Super DAGs Represented by Complete PAG, Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

the Super DAG counting scheme's performance improves as the number of latent causes of the target variable increases. However, contrary to results for other metrics, the performance of the PAG Heuristic and Pattern counting scheme degrades as the number of latent causes of the target variable increases.

7.2 Inferring a PAG with FCI: An Example

Figure 7.1 provides an example of a "Y Structure #1" PAG with three constructed variables (avg_X2X3X4, avg_X5X7X8, and avg_Y0Y1Y2) and target variable T. We explain, step-by-step, how FCI infers this PAG structure from conditional independence relations among these variables.

FCI begins with the complete PAG, each pair of variables connected by an o−o edge (Figure 7.2).

Conditional independence relations among these variables are representative of the independencies among other sets of constructed variables that allow us to infer the same PAG structure (i.e., Y Structure #1). As such, avg_X2X3X4 corresponds to X_Cloud_1; avg_X5X7X8 corresponds to X_Cloud_2; and avg_Y0Y1Y2 corresponds to Y_Cloud_1, as always. Two (conditional) independence relations obtain among these four variables (where "X $\perp\!\!\!\perp$ Y|**Z**" reads "X is independent of Y conditional on the variables in set **Z**"):

Simulation Strategy #1		**Counting Scheme: MSE of Average Intervention Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	6,911	.0330	.0365	.0323
1	14,136	.0177	.0416	.0927
2	9,444	.0126	.0621	.1669
3	1,509	.0122	.1116	.3178
0-3	32,000	.0192	.0498	.1122

Table 7.10: Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value: Generating Super DAGs Not Represented by Complete PAG, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #2		**Counting Scheme: MSE of Average Intervention Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	884	.0988	.0343	.0792
1	5,592	.0370	.0563	.1363
2	8,010	.0167	.0814	.1817
3	3,514	.0116	.1038	.2156
0-3	18,000	.0260	.0757	.1692

Table 7.11: Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value: Generating Super DAGs Represented by Complete PAG, Non-Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

Simulation Strategy #3		**Counting Scheme: MSE of Average Intervention Causal R^2**		
# Latent Causes of X4	*# Super DAGs Sampled*	*Super DAG*	*PAG Heuristic*	*Pattern*
0	17	.0777	.0528	.1130
1	136	.0351	.0541	.1330
2	171	.0156	.0808	.1795
3	76	.0102	.0843	.1860
0-3	400	.0238	.0712	.1621

Table 7.12: Mean Squared Error for Average Intervention Causal R^2 Value vs. True Intervention Causal R^2 Value: Generating Super DAGs Represented by Complete PAG, Exhaustive Super DAG Counting Scheme, Stratified by Number of Latent Causes of Target (X4)

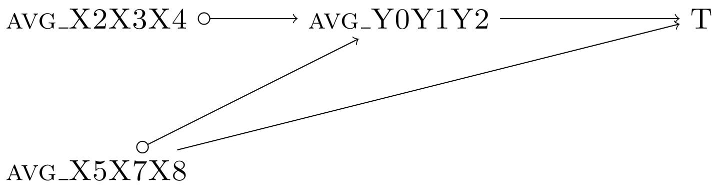

Figure 7.1: Example PAG: We illustrate the inference of this PAG from a set of conditional independencies among its four measured (of which three are constructed) variables.

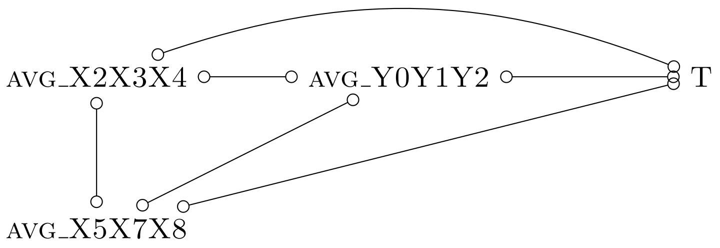

Figure 7.2: FCI Starting Point (Step #1): Complete PAG

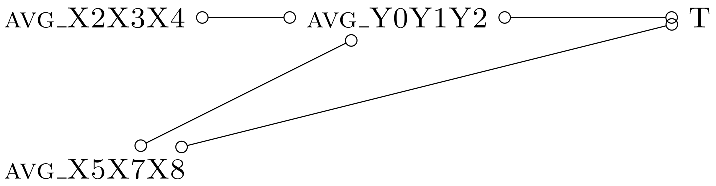

Figure 7.3: FCI Step #2: PAG after Adjacency Phase

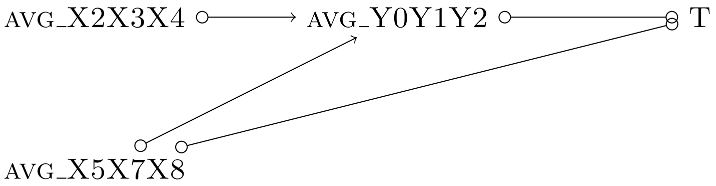

Figure 7.4: FCI Step #3: PAG after Unshielded Collider Orientation at AVG_Y0Y1Y2

- $\text{AVG\_X2X3X4} \perp\!\!\!\perp \text{AVG\_X5X7X8}$
- $\text{AVG\_X2X3X4} \perp\!\!\!\perp \text{T} | \{\text{AVG\_Y0Y1Y2}, \text{AVG\_X5X7X8}\}$

Because of these two independencies, the adjacency phase of FCI removes two corresponding edges: AVG_X2X3X4 o−o AVG_X5X7X8 and AVG_X2X3X4 o−o T (Figure 7.3).

Both AVG_X2X3X4 and AVG_X5X7X8 are adjacent to AVG_Y0Y1Y2, but the two variables are unconditionally independent; FCI's orientation phase begins, using the "collider orientation" rule, by orienting an unshielded collider at AVG_Y0Y1Y2 (Figure 7.4), i.e., orienting AVG_X2X3X4 o→ AVG_Y0Y1Y2 and AVG_X5X7X8 o→ AVG_Y0Y1Y2.

Next, note that if the AVG_Y0Y1Y2 o−o T edge in Figure 7.4 were oriented as AVG_Y0Y1Y2 ←o T, the resulting PAG would imply that AVG_X2X3X4 is not independent of T given AVG_Y0Y1Y2 and AVG_X5X7X8, since the collider AVG_X2X3X4 o→ AVG_Y0Y1Y2 ←o T would be active. This is contrary to independencies in our data, so we infer AVG_Y0Y1Y2 → T. This is an instance of FCI's "away from collider" rule (Figure 7.5). Unshielded colliders are detectable in data at the collider orientation step, so the "away from collider" rule ensures that FCI does not orient any unshielded colliders beyond those found at the collider orientation

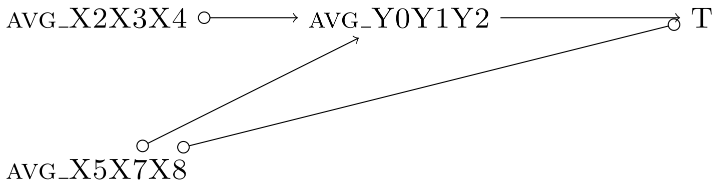

Figure 7.5: FCI Step #4: PAG after application of "Away from Collider" rule orients AVG_Y0Y1Y2 → T

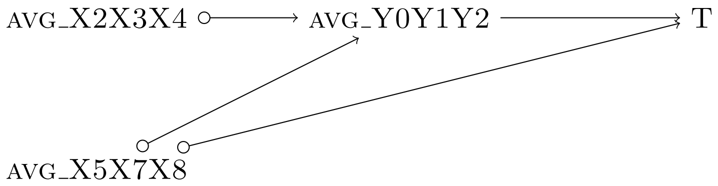

Figure 7.6: FCI Step #5: PAG after application of "Away from Ancestor" rule orients AVG_X5X7X8 o→ T

step (i.e., only those compatible with data).

Next, the arrowhead at AVG_Y0Y1Y2 on the AVG_X5X7X8 o→ AVG_Y0Y1Y2 edge in Figure 7.5 means (by the semantics of PAG edges) that AVG_Y0Y1Y2 is not an ancestor of AVG_X5X7X8. At the same time, AVG_Y0Y1Y2 is an ancestor (or cause) of T. Consider the edge: AVG_X5X7X8 o−o T. If this edge were oriented "out of" T, then both T and AVG_Y0Y1Y2 would be ancestors of AVG_X5X7X8. Since we know that AVG_Y0Y1Y2 is not an ancestor of AVG_X5X7X8, we orient the AVG_X5X7X8 o−o T edge "into" T. This is an instance of the FCI "away from ancestor" rule (Figure 7.6).

FCI's final orientation uses the "definite discriminating paths" rule to orient AVG_X5X7X8 → T (Figure 7.7). Spirtes, et al. (2000) and Zhang (2008b) formally describe this rule (as well as the rest of the FCI algorithm, including other orientation rules[2]). We consider the logic of this orientation. Given the PAG of Figure 7.6 and the fact that AVG_X2X3X4 is independent of T conditional on the set {AVG_Y0Y1Y2, AVG_X5X7X8}, AVG_X5X7X8 must be a non-collider. Suppose we orient the AVG_X5X7X8 o→ T edge as AVG_X5X7X8 ↔ T. There are two possible ways to orient the AVG_X5X7X8 o→ AVG_Y0Y1Y2 edge:

---

[2] Zhang (2008b) provides a complete set of orientation rules.

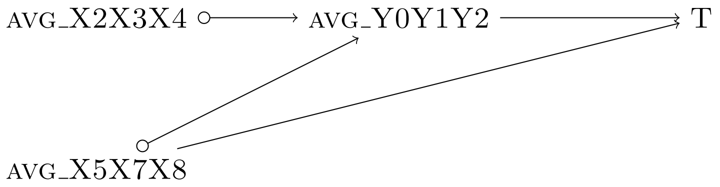

Figure 7.7: FCI Step #6: Final Example PAG (Figure 7.1) after application of "Definite Discriminating Path" rule orients AVG_X5X7X8 → T

1. AVG_X5X7X8 ↔ AVG_Y0Y1Y2 creates a collider at AVG_X5X7X8.
2. AVG_X5X7X8 → AVG_Y0Y1Y2 makes AVG_X5X7X8 an ancestor of T, but we have assumed AVG_X5X7X8 ↔ T, so AVG_X5X7X8 is definitely not an ancestor of T.

Since the first possibility is not compatible with data, and the second possibility violates the semantic of PAGs, the AVG_X5X7X8 ↔ T orientation is ruled out as a possibility. Thus, we orient AVG_X5X7X8 → T. We now have the final inferred PAG, corresponding to what we call Y Structure #1 in Chapter 4, over four variables from two observed (conditional) independence relations.

Chapter 8

References

1. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Second International Symposium on Information Theory*, 267-281.
2. Aleven, V., Koedinger, K. R. (2000). Limitations of Student Control: Do Students Know When They Need Help? *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, 292-303.
3. Anderson, J.R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.
4. Anderson, J.R. (1993). *Rules of the Mind.* Hillsdale, NJ: Erlbaum.
5. Arnold, A., Beck, J., Scheines, R. (2006). Feature Discovery in the Context of Educational Data Mining: An Inductive Approach. *Proceedings of the AAAI 2006 Workshop on Educational Data Mining*, 7-13.
6. Baker, R.S. (2005). *Designing Intelligent Tutors That Adapt to When Students Game the System.* Ph.D. Thesis, Human-Computer Interaction Institute, Carnegie Mellon University. CMU Technical Report CMU-HCII-05-104.
7. Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004a). Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.
8. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004b). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System." *Proceedings of ACM CHI 2004*, 383-390.

9. Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006). Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

10. Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 287-314.

11. Baker, R.S.J.d., de Carvalho, A. M. J. A. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays. *Proceedings of the 1st International Conference on Educational Data Mining*, 38-47.

12. Baker, R.S.J.d., Gowda, S.M., Corbett, A.T. (2011). Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179-188.

13. Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. (2012). Sensor-free Automated Detection of Affect in a Cognitive Tutor for Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126-133.

14. Baker, R.S.J.d. (2007). Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of the 2007 Conference on Human Factors in Computing Systems*, 1059-1068.

15. Beal, C.R., Qu, L., Lee, H. (2006). Classifying Learner Engagement Through Integration of Multiple Data Sources. *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 151-156.

16. Beck, J. (2005). Engagement Tracing: Using Response Times to Model Student Disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 88-95.

17. Bollen, K. (1989). *Structural Equations with Latent Variables.* Wiley.

18. Boyd, R. (1990). What Realism Implies and What It Does Not. *Dialectica*, 43, 5-29.

19. Boyd, R. (1991). Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds. *Philosophical Studies*, 61, 127-148.

20. Boyd, R. (1999). Kinds, Complexity and Multiple Realization. *Philosophical Studies*, 95, 67-98.

21. Burgansky-Eliash, Z., Wollstein, G., Chu, T., Ramsey, J., Glymour, C., Noecker, R., Schuman, J. (2005). Optical Coherence Tomography Machine Learning Classifiers for Glaucoma Detection. *Investigative Ophthamology and Visual Science*, 46, 4147-4152.

22. Cartwright, N. (1999). Causal Diversity and the Markov Condition. *Synthese*, 121, 3-27.

23. Cartwright, N. (2006). From Metaphysics to Method: Comments on Manipulability and the Causal Markov Condition. *British Journal for the Philosophy of Science*, 57, 197-218.

24. Chickering, D.M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3, 507-554.

25. Chu, T., Danks, D., Glymour, C. (2005). Data Driven Methods for Nonlinear Granger Causality: Climate Teleconnection Mechanisms. Technical Report CMU-PHIL-171, Department of Philosophy, Carnegie Mellon University.

26. Cocea, M., Hershkovitz, A., Baker, R.S.J.d. (2009). The Impact of Off-Task and Gaming Behavior on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

27. Collier, J. (1996). On the Necessity of Natural Kinds. In: P. Riggs (ed.) *Natural Kinds, Laws of Nature and Scientific Reasoning*, 1-10. Dordrecht: Kluwer.

28. Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User Adapted Interaction*, 4, 253-278.

29. Craig, S. D., Graesser, A. C., Sullins, J., and Gholson, B. 2004. Affect and learning: an exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241-250.

30. Cramer, D. (2004). Emotional Support, Conflict, Depression, and Relationship Satisfaction in a Romantic Partner. *The Journal of Psychology*, 138, 532-542.

31. Crane, J.K, Sandler, R. (2011). Species Concepts and Natural Goodness. In: Campbell, J.K., O'Rourke, M., Slater, M.H. (eds.) *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*, 289-312. Cambridge, MA: MIT Press.

32. Craver, C.F. (2009). Mechanisms and Natural Kinds. *Philosophical Psychology*, 22(5), 575-594.

33. Cronbach, L.J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297-334.

34. Devitt, M. (2011). Natural Kinds and Biological Realisms. In: Campbell, J.K., O'Rourke, M., Slater, M.H. (eds.) *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*, 155-174. Cambridge, MA: MIT Press.

35. Eberhardt, F. (2007). *Causation and Intervention.* Ph.D. Thesis, Department of Philosophy, Carnegie Mellon University.

36. Fancsali, S.E. (2008). *Cronbach's Alpha, Latent Variables, and Causal Inference.* M.S. Thesis, Department of Philosophy, Carnegie Mellon University.

37. Fancsali, S.E. (2011a). Variable Construction for Predictive and Causal Modeling of Online Education Data. *Proceedings of the First International Conference on Learning Analytics and Knowledge (LAK 2011)*, 54-63. New York: ACM.

38. Fancsali, S.E. (2011b). Variable Construction and Causal Modeling of Online Education Messaging Data: Initial Results. *Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011)*, 331-332.

39. Fancsali, S.E. (2012). Variable Construction and Causal Discovery for Cognitive Tutor Log Data: Initial Results. *Proceedings of the Fifth International Conference on Educational Data Mining (EDM 2012)*, 238-239.

40. Fodor, J.A. (1974). Special Sciences. *Synthese*, 28, 97-115.

41. Friston, K.J. (1994). Functional and Effective Connectivity in Neuroimaging: a Synthesis. *Human Brain Mapping*, 2, 56-78.

42. Glymour, C. (2007). When is a Brain Like the Planet? *Philosophy of Science*, 74, 330-347.

43. Griffiths, P.E. (1997). *What Emotions Really Are*. Chicago: Chicago UP.

44. Hacking, I. (1991). A Tradition of Natural Kinds. *Philosophical Studies*, 61, 109-126.

45. Hare, B., Brown, B., Williamson, C., Tomasello, M. (2002). The Domestication of Social Cognition in Dogs. *Science*, 298, 1634-1636.

46. Hausman, D.M., Woodward, J. (1999). Independence, Invariance and the Causal Markov Condition. *British Journal for the Philosophy of Science*, 50, 521-583.

47. Hausman, D.M., Woodward, J. (2004). Modularity and the Causal Markov Condition: A Restatement. *British Journal for the Philosophy of Science*, 55, 147-161.

48. Hausmann, R.G.M., VanLehn, K. (2010). The Effect of Self-Explaining on Robust Learning. *International Journal of Artificial Intelligence in Education*, 20(4), 303-332.

49. Hendrick, S.S. (1988). A Generic Measure of Relationship Satisfaction. *Journal of Marriage and the Family*, 50, 93-98.

50. Hoover, K. (2003). Nonstationary Time Series, Cointegration, and the Principle of the Common Cause. *British Journal for the Philosophy of Science*, 54, 527-551.

51. Hoyer, P.O., Hyttinen, A. (2009). Bayesian discovery of linear acyclic causal models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-2009)*, 240-248.

52. Hoyer, P.O., Shimizu, S., Kerminen, A.J. (2006). Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM '06)*, 155-162.

53. Hoyer, P.O., Shimizu, S., Kerminen, A.J., Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49, 362-378.

54. Hoyer, P.O., Janzing, D., Mooij, J., Peters, J., Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21 (NIPS 2008), 689-696.

55. Hyttinen, A., Eberhardt, F., Hoyer, P.O. (2012). Learning Linear Cyclic Causal Models with Latent Variables. *Journal of Machine Learning Research*, 13, 3387-3439.

56. Hyvärinen, A., Karhunen, J., Oja, E. (2001). *Independent Component Analysis.* Wiley Interscience.

57. Johns, J., Woolf, B. (2006). A Dynamic Mixture Model to Detect Student Motivation and Proficiency. *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 163-168.

58. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2011). A Data Repository for the EDM Community: The PSLC DataShop. In: C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.) *Handbook of Educational Data Mining*, 43-55. Boca Raton, FL: CRC Press.

59. Kripke, S.A. [1972] (1980). *Naming and Necessity.* Cambridge, MA: MIT Press.

60. Lacerda, G., Spirtes, P., Ramsey, J., Hoyer, P.O. (2008). Discovering Cyclic Causal Models by Independent Components Analysis. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 366-374. Corvallis, OR: AUAI Press.

61. Lazar, N.A. (2008). *The Statistical Analysis of Functional MRI Data.* New York: Springer.

62. Locke, J. [1690] (1979). *An Essay Concerning Human Understanding.* P.H. Nidditch (ed.) Oxford: Oxford UP.

63. Maathuis, M.H., Kalisch, M., Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37, 3133-3164.

64. Machery, E. (2005). Concepts Are Not a Natural Kind. *Philosophy of Science*, 72, 444-467.

65. Machery, E. (2009). *Doing Without Concepts.* New York: Oxford UP.

66. Magnus, P.D. (2011). Drakes, Seadevils, and Similarity Fetishism. *Biology and Philosophy*, 26, 857-870.

67. Maris, E. (1995). Psychometric Latent Response Models. *Psychometrika*, 60, 523-547.

68. Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 403-411. San Francisco: Morgan Kaufmann.

69. Meek, C. (1997). *Graphical Models: Selecting Causal and Statistical Models*. Ph.D. Thesis, Department of Philosophy, Carnegie Mellon University.

70. Mill, J.S. (1843). *A System of Logic*. London: J.W. Parker.

71. Miller, M.B. (2005). Coefficient Alpha: A Basic Introduction from the Perspective of Classical Test Theory and Structural Equation Modeling. *Structural Equation Modeling*, 2, 255-273.

72. Millikan, R.G. (1999). Historical Kinds and the "Special Sciences." *Philosophical Studies*, 95, 45-65.

73. Mumford, J.A., Horvath, S., Oldham, M.C., Langfelder, P., Geschwind, D.H., Poldrack, R.A. (2010). Detecting Network Modules in fMRI Time Series: A Weighted Network Analysis Approach. *NeuroImage*, 52, 1465-1476.

74. NOAA (National Oceanic and Atmospheric Administration). (2013). NOAA's El Niño Page. Retrieved 1 April 2013.
$< http://www.elnino.noaa.gov/ >$

75. Nunnaly, J.C., Bernstein, I.H. (1994). *Psychometric Theory*, Third Edition. New York: McGraw-Hill.

76. Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2013). Affective States and State Tests: Investigating How Affect Throughout the School Year Predicts End of Year Learning Outcomes. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge.*

77. Pardos, Z.A., Heffernan, N. T. (in press). Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. *Journal of Machine Learning Research, Workshops and Conference Proceedings.*

78. Parker, W.S. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, 78, 579-600.

79. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* 2nd Edition. Cambridge: Cambridge UP.

80. Pearl, J., Dechter, R. (1996). Identifying independencies in causal graphs with feedback. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, 420-442. San Francisco: Morgan Kaufmann.

81. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Francisco: Morgan Kauffman.

82. Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. (2002). Academic emotions in students self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist*, 37, 91-106.

83. Putnam, H. (1975). The Meaning of "Meaning." In: *Mind, Language, and Reality*, by H. Putnam, 215-271. Cambridge: Cambridge UP.

84. Quine. W.V.O. (1969). Natural Kinds. In: *Ontological Relativity and Other Essays*, by W.V.O. Quine, 114-138. New York: Columbia UP.

85. Raftery, A.E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25, 111-163.

86. Rai, D., Beck, J.E. (2010). Analysis of a Causal Modeling Approach: A Case Study with an Educational Intervention. *Proceedings of the 3rd International Conference on Educational Data Mining*, 313-314.

87. Rai, D., Beck, J.E. (2011). Exploring User Data from a Game-Like Math Tutor: A Case Study in Causal Modeling. *Proceedings of the 4th International Conference on Educational Data Mining*, 307-311.

88. Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C. (2010). Six Problems for Causal Inference from fMRI, *NeuroImage*, 49, 1545-1558.

89. Ramsey, J.D., Hanson, S.J., Glymour, C. (2011). Multi-subject Search Correctly Identifies Causal Connections and Most Causal Directions in the DCM Models of the Smith et al. Simulation Study. *NeuroImage*, 58, 838-848.

90. Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K.R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. *Proceedings of the 12th International Conference on Artificial Intelligence In Education*, 555-562. Amsterdam: ISO Press.

91. Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

92. Rheins, J.G. (2011). Similarity and Species Concepts. In: Campbell, J.K., O'Rourke, M., Slater, M.H. (eds.) *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*, 253-288. Cambridge, MA: MIT Press.

93. Richardson, T.S. (1996a). A Discovery Algorithm for Directed Cyclic Graphs. In: Horvitz, E., Jensen, F. (eds.) *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, 454-461. San Francisco: Morgan Kaufmann.

94. Richardson, T.S. (1996b). *Models of Feedback: Interpretation and Discovery*. Ph.D. Thesis, Department of Philosophy, Carnegie Mellon University.

95. Richardson, T.S. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1), 145-157.

96. Richardson, T.S., Spirtes, P. (2002). Ancestral Graph Markov Models, *Annals of Statistics*, 30(4), 962-1030.

97. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. (2007). Cognitive Tutor: Applied Research in Mathematics Education. *Psychonomic Bulletin and Review*, 14(2), 249-255.

98. Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B. (2009). Reducing the Knowledge Tracing Space. *Proceedings of the 2nd International Conference on Educational Data Mining*, 151-160.

99. Rodrigo, M.M.T., Baker, R.S., Jadud, M.C., Amarra, A.C.M., Dy, T., Espejo-Lahoz, M.B.V., Lim, S.A.L., Pascua, S.A.M.S., Sugay, J.O., Tabanao, E.S. 2009. Affective and Behavioral Predictors of Novice Programmer Achievement. *Proceedings of the 14th ACM-SIGCSE Annual Conference on Innovation and Technology in Computer Science Education*, 156-160.

100. Samuels, R., Ferreira, M. (2010). Why *don't* concepts constitute a natural kind? *Behavioral and Brain Sciences*, 33,222-223.

101. Scheines, R., Leinhardt G., Smith, J., Cho, K. (2005). Replacing Lecture with Web-Based Course Materials. *Journal of Educational Computing Research*, 32, 1-26.

102. Scheines, R. (2005). The Similarity of Causal Inference in Experimental and Non-Experimental Studies. *Philosophy of Science*, 72, 927-940.

103. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461-464.

104. Shih, B., Koedinger, K.R., Scheines, R. (2007). Optimizing Student Models for Causality. *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 644-646.

105. Shih, B., Koedinger, K.R., Scheines, R. (2011). A Response-Time Model for Bottom-Out Hints as Worked Examples. In: C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.) *Handbook of Educational Data Mining*, 201-211. Boca Raton, FL: CRC Press.

106. Shih, B. (2011). *Target Sequence Clustering*. Ph.D. Thesis, Machine Learning Department, Carnegie Mellon University.

107. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.J. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2003-2030.

108. Shimizu, S., Hyvärinen, A., Kano, Y., Hoyer, P.O. (2005). Discovery of non-gaussian linear causal models using ICA. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, 526-533.

109. Silva, R., Scheines, R., Glymour, C., Spirtes, P. (2006). Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research*, 7, 191-246.

110. Sinnott-Armstrong, W. (2005). It's Not My Fault: Global Warming and Individual Moral Obligations. *Perspectives on Climate Change: Science, Economics, Politics, Ethics (Advances in the Economics of Environmental Research, Volume 5)*, 293-315.

111. Sober, E. (2001). Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *British Journal for the Philosophy of Science*, 52, 1-16.

112. Spirtes, P., Glymour, C., Scheines, R. (2000). *Causation, Prediction, and Search.* 2nd Edition. Cambridge, MA: MIT Press.

113. Spirtes, P. (2010). Introduction to Causal Inference. *Journal of Machine Learning Research*, 11, 1643-1662.

114. Steel, D. (2005). Indeterminism and the Causal Markov Condition. *British Journal for the Philosophy of Science*, 56, 3-26.

115. Steinbach, M., Tan, P., Kumar, V., Klooster, S., Potter, C. (2003). Discovery of Climate Indices Using Clustering. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003)*, 446-455.

116. Stock, J.H., Watson, M.W. (2003). *Introduction to Econometrics.* New York: Addison Wesley.

117. Townsend, K. A., Wollstein, G., Danks, D., Sung, K. R., Ishikawa, H., Kagemann, L., Gabriele, M. L., Schuman, J. S. (2008). Heidelberg Retina Tomography III Machine Learning Classifiers for Glaucoma detection. *British Journal of Ophthalmology*, 92, 814-818.

118. Trochim, W.M.K. (2005). *Research Methods: The Concise Knowledge Base.* Thomson.

119. U.S. Bureau of Labor Statistics. (2013). Consumer Price Index: Frequently Asked Questions (FAQs). Retrieved 1 April 2013. $< http://www.bls.gov/cpi/cpifaq.htm >$

120. VanLehn, K. (2006). The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 16, 227-265.

121. Verma, T., Pearl, J. (1990). Equivalence and synthesis of causal models. *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, 220-227. New York: Elsevier Science.

122. Walonoski, J.A., Heffernan, N.T. (2006). Detection and Analysis of Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 382-391.

123. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* New York: Springer.

124. Williams, N.E. (2011). Arthritis and Nature's Joints. In: Campbell, J.K., O'Rourke, M., Slater, M.H. (eds.) *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*, 199-230. Cambridge, MA: MIT Press.

125. Yu, H., et al. (in press). Feature Engineering and Classifier Ensemble for KDD Cup 2010. *Journal of Machine Learning Research, Workshops and Conference Proceedings.*

126. Zhang, J. (2006). *Causal Inference and Reasoning in Causally Insufficient Systems.* Ph.D. Thesis, Department of Philosophy, Carnegie Mellon University.

127. Zhang, J. (2008a). Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9, 1437-1474.

128. Zhang, J. (2008b). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172, 1873-1896.