Causal Discovery and MIMIC Models

Alexander Murray-Watters

April 30, 2013

Contents

List of Figures

Acknowledgements

I'd like to thank my parents for putting up with my curiosity, and all that entails. I'd also like to thank Lizzie Silver for introducing me to causal graphs, and Dr. Cosma Shalizi for introducing me to both causal search and Professor Glymour. Finally, I'd like to thank my thesis advisor, Dr. Clark Glymour, for his advice, patience, and willingness to tell me when he thought I was wrong...

Abstract

This thesis presents an alternative method for the detection of MIMIC models. Previous methods (such as factor analysis) suffer from a number of significant flaws and limitations, which the new method (a causal search algorithm) doesn't suffer. A new algorithm is introduced, followed by a worked-through example of its application. Discussion focuses on some of the limiting assumptions the algorithm currently requires. Finally, recommendations for future work address improvements of the algorithm, as well as its applicability.

Chapter 1

Introduction

MIMIC (an acronym for Multiple Indicators, Multiple Causes) models are a class of models positing latent, unrecorded, common causes. The causal structure consists of three components: a set of latent variables (not observed), a set of inputs (observed variables that act as causes of latent variables), and a set of outputs (observed variables that are caused by latent variables). MIMIC models are often used in cases where there are believed to be unobserved variables acting as causes on either some observed variables (i.e., outputs) or other latent variables.

The applications of MIMIC models are widespread, ranging from economics, to psychology, and even public health. Advocates of MIMIC models emphasize their usefulness in simultaneously assessing multiple dimensions of complex social issues.

For example, Lester (2008) applied a MIMIC model to examine factors related to successful settlement of immigrants to Australia. Groups of interest were economic immigrants, non-economic immigrants, and those who were not labor force participants. The study included two waves of immigrants (4867 and

3538 individuals, respectively). Analyses included two latent constructs, representing successful settlement at each of two time periods. In the measurement model indicators of successful settlement included level of satisfaction with life in Australia, mental health, encouraging others to migrate to Australia, and believing that their decision to immigrate was right. The structural model included economic and labor market factors for each time period, and a formative model included time-invariant variables (e.g., gender and status of the person at time of entry).

Several studies have used MIMIC models to estimate the size of the hidden economy since factors such as GDP and unemployment rates do not give a comprehensive picture of a nation's economic conditions. Bühn and Schneider (2008) developed a MIMIC model and tested its ability to examine both the size and development of economic loss attributed to the shadow economy in France. Giles (1999) estimated a MIMIC model to uncover a time-series view of the hidden economy in New Zealand. Tedds (1998) developed a MIMIC model to estimate the hidden economy in Canada over the period of 1976 to 1995, suggesting the model represented approximately 15% of the GDP. Defending the use of MIMIC models in estimating the shadow economy, DellAnno and Schneider (2006) provided extensions of standard MIMIC models and argued that their potential is undervalued in the field of economics.

Ríos-Bedoya and colleagues utilized MIMIC models to determine the strength of association between current smoking status (i.e., current daily smokers or those who had exposure but no pattern of regular smoking) and two latent constructs (pleasant and unpleasant early smoking experiences) in a sample of 458 participants (Ríos-Bedoya et al., 2009). The overall MIMIC model included measurement and structural components. In the study's measurement model, categorical responses to items from the Early Smoking Experiences (ESE) ques-

tionnaire (which assessed response to early experimentation with cigarettes) were considered to be indicators of the two unobserved latent constructs. The structural model examined the association between covariates of interest (i.e., race, depression history, sex, age of first cigarette, and current smoking status) and the unobserved latent constructs.

MIMIC models are usually specified a priori. That is, an investigator assumes she knows all of the relevant causal relations and the only function of the data is to enable estimation of parameters and to confirm the hypothesis by statistical tests. When such tests fail to confirm the a priori model, another model may be proposed, more or less in one at a time fashion, roughly as conceived by Karl Popper. The chief automatic search aid used is factor analysis, which is supposed to locate the latent variables and which "output" variables they influence. As an alternative method for the detection of MIMIC models, I propose a causal search algorithm (the detect.MIMIC algorithm) that does not suffer from the concerns of previous ad hoc methods used in discovering MIMIC models.

Chapter 1 introduces the area of study. Chapter 2 provides a brief description of MIMIC models, followed by a discussion of the limitations existing methods suffer, as well as examples of the use of MIMIC models in various fields. In Chapter 3 I present a survey of existing work from which the new method is derived. Chapter 4 describes the assumptions underlying the detect.MIMIC algorithm and provides a detailed discussion of relevant concepts, as well as a description of the proposed algorithm. Results of tests performed on the algorithm are represented through a worked through example in Chapter 4, followed by a section on assumptions and limitations of the algorithm. Chapter 5 provides a summary of the project as well as identifying two areas for future work, that is, improvements in the algorithm and improvements in its applicability.

Chapter 2

Motivation

2.1 Place in the Method of Science

Frequently, when a scientist attempts to construct a model of how some natural (or manmade) phenomena functions, a number of a priori assumptions must be made. Some of these assumptions are unavoidable (such as using GDP as a measure of the size of the economy when no other data exists). Others, however, are avoidable, provided the data is allowed to speak for itself. As there is always some risk involved when making assumptions, it is (generally) preferable to minimize the number of unnecessary assumptions when creating a scientific model. Happily, this is exactly what a causal search algorithm does.

Another property of causal search algorithms, desirable in model construction, is the explicitness of its methodology. More ad hoc methods (whose modeling decisions are frequently opaque) tend to leave a great deal of information on the cutting-room floor. The assumptions and evidence that led a researcher to believe in the truth of a causal relation are left in the ether. In contrast, causal search algorithms follow an explicit (and public) process whose validity

can be thoroughly tested. Causal search algorithms provide a straightforward method for checking already existing causal models (such as those popular in the current literature). Data, both new and old, can be fed into the algorithm, producing results that can be used to determine the set of possible models that could pass a collection of statistical tests. Finally, the proper automation of the procedure invites, and often forces, clarity about which aspects of a model are provided only by an investigator's assumptions or theories or conjectures, and which parts are data driven.

2.2 MIMIC Models

MIMIC models are a kind of causal structure, that can be represented by a class of directed, acyclic graphs. Although use of the term "MIMIC" varies in the literature, I will assume all models considered have the following features:

1. A set of independent input variables whose values occur in the data.
2. A set of output variables whose values occur in the data.
3. If there exists a latent variable on a directed path between input and output variables, then it must have an outdegree of at least 2.

There are also several other additional (optional) characteristics:

4. Output variables are independent of one another conditional on their latent and input common causes.
5. Every input influences an output through a latent path.
6. Every latent has at least n outputs.

In addition, applications of MIMIC models typically make distributional assumptions, for example that the joint distribution of the variables is Gaussian,

the relation is linear, and each measured variable and each latent common cause has specific sources of variance that are independent of the sources of variance specific to other variables.

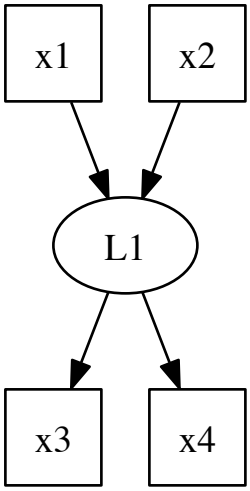

Figure 2.1: Illustration of a MIMIC model.

Chapter 3

Background

Predicting outcomes from the study of available data has been of interest to philosophers and scientists since the distant past (e.g., Hume, Bacon, etc). Such causal relationships require both a temporal order (i.e., a cause must precede its effect; Reichenbach, 1956) and consideration of spurious associations (i.e., covariance or correlation that merely appears to be causal). Sober (1998) while not employing MIMIC models per se, presented an argument that can help determine when such causal models are warranted. When considering the choice between models that address only cause and effect (black box inference models) or those that include intervening variables, he argued that the choice of models should be driven by the data, not simply by parsimony. Models that include unobserved intervening variables (i.e., those reflecting latent constructs) should be used when there is a claim of probabilistic dependence but the choice of an intervening variable also requires that it have at least two effects. Utilizing existing psychological studies Sober considered the latent constructs of stimulus generalization, response generalization, and the theory of mind (i.e., internal states) as explanations of how the effects from two experiments can be related.

Advances in computation, coupled with improved statistical techniques, have enabled a systematic consideration of requirements for a study of causal relationships. However, determining causal structures between variables, including unobserved (latent) variables, still poses challenges. To address these relationships and enhance causal discovery, scholars have proposed the use of causal graphs (e.g., Glymour et al. (2010); Pearl (2009); Scheines (1997); Spirtes et al. (2000)).

3.1 Causal Graphs

The major concepts necessary to understand causal graphs are elucidated below.

An undirected path (Figure 3.1) is a connection between two nodes (in the example below, x_1 and x_2).

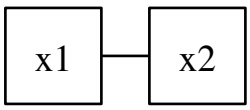

Figure 3.1: Depiction of an undirected path.

A directed path (Figure 3.2), is interpretable in causal terms. In the example below, node x_1 is read as causing x_2.

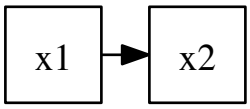

Figure 3.2: Depiction of a directed path.

It should also be noted that in the above Figure, x_2 is referred to as a child of x_1.

There are three basic structures used in causal graphs: Forks, Chains, and Colliders.

Forks (Figure 3.3) are made up of a center node with two or more children (or effects).

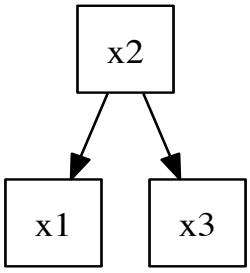

Figure 3.3: Illustration of a Fork.

Chains (Figure 3.4) are made up of a set of nodes, all of whom have a child, except for the last node in the chain.

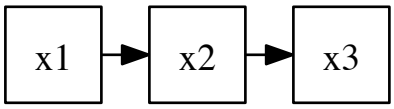

Figure 3.4: Illustration of a Chain.

Finally, Colliders (Figure 3.5) involve two (or more) nodes having a common child. This structure is particularly relevant, as this structure is a naturally occurring[1] instance of d-separation.

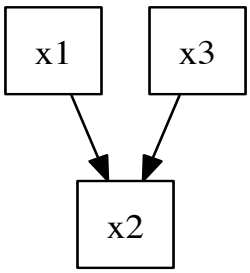

Figure 3.5: Illustration of a Collider.

Definition 3.1.1 (Causal Markov Condition)**.** A node in an acyclic directed causal graph is conditionally independent of its non-decedents, given its parents.

From the causal Markov condition, Pearl (2009) obtained an efficient method for determining whether or not two variables are said to be independent.

D-separation is defined as follows:

[1] In other words, without conditioning on any set of variables, x_1 is d-separated from x_3

Definition 3.1.2 (d-separation)**.** Two nodes, X and Y, in a DAG (i.e., directed acyclic graph) are d-separated by a set $\mathbf{S}$ of nodes not containing X or Y if and only if all paths between X and Y are blocked by $\mathbf{S}$. Otherwise, X and Y are d-connected.

A path is said to be blocked if either (a) the path contains an intermediate node that has been conditioned on and isn't a collider (or a descendant of a collider on the path), or (b) there is a collider on the path that hasn't been conditioned on and the collider hasn't had one of its decedents conditioned on.

Here's an example utilizing the implications of d-separation. Envision the following Figure:

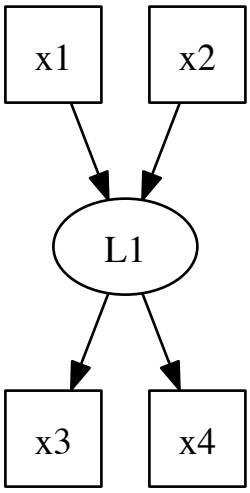

Figure 3.6: Illustration of d-separation.

As x_1 is d-separated from x_2 (the only path between x_1 and x_2 is through a collider), x_1 is independent of x_2 (i.e., $x_1 \perp\!\!\!\perp x_2$). However, conditioning on

x_3 (a decedent of the collider) causes x_1 to no longer be d-separated from x_2. As a result, x_1 is no longer independent of x_2 (i.e., $x_1 \not\perp\!\!\!\perp x_2|x_3$). Contrast this with x_3 and x_4 (which are not d-separated). x_3 is not independent of x_4, and conditioning on x_1, or x_2 won't change this.

3.2 Prior Work

Silva et al. (2006) created an algorithm intended to search for causal structures (involving latent variables) in more general circumstances than my own. However, one assumption made in their paper (and necessary for their method to work) requires that observed variables are not causes of unobserved variables. My algorithm, while not as general as the one in Silva et al. (2006), does not make such an assumption. Further, the Silva procedure is confined to linear systems and a few other special distributions. I offer some proposals for search methods that are essentially non-parametric.

3.3 Parametric Methods

A parametric method is any method that adds constraints to the relationships in data beyond those imposed by the graph.

3.3.1 SEM

SEM (or structural equation modeling) is a commonly used method of representing causal relations between variables. It uses a set of equations to represent causal relations. In these equations, each variable is represented by a function composed of non-random variables, with an additional collection of error terms. These errors are assumed to follow some joint probability distribution. The choice of distribution for the errors is left up to the researcher.

Using SEM to represent a model in causal search poses two important difficulties. The most important of these is the necessity of constructing a model by hand. Second, SEM models are difficult to modify for alternative purposes. As Judea Pearl mentions, graphical depictions (like circuit diagrams), can be taken-up and modified by other researchers for purposes never imagined by the model's originator (Pearl, 2009). As the intention of causal search is to either generate new models or check old models, flexible manipulation of a model is a very desirable characteristic.

3.3.2 Factor Analysis

Factor analysis is a method of modeling data ($\mathbf{X}$), by left multiplying a matrix of principle components ($\mathbf{w}$) with a matrix of data projected onto the principle components in $\mathbf{w}$. Any differences between $\mathbf{X}$ and the resulting matrix multiplication are resolved by the inclusion of an error term, ϵ.

$$\mathbf{X} = \mathbf{F}\mathbf{w} + \epsilon \tag{3.1}$$

Two primary motivations drive the use of factor analysis:

1. Dimension Reduction
2. Inferring Causal Structures

Dimension reduction is generally intended for use in high dimensional problems, such as genomics. For instance, imagine a researcher has one hundred different genomes (each corresponding to a single individual). Each of these genomes contains thousands of genes. The researcher attempts to predict whether or not an individual has a disease, using that individual's genes. In order to use standard methods (such as linear regression), the number of parameters (i.e., the number of genes) must be less than the number of observations (i.e., the

number of genomes). Under these conditions, factor analysis replaces the use of specific genes, with the most explanatory factors (in this case, amalgams of genes).

The other common motivation for using factor analysis is to infer causal structures when some number of unobserved variables (i.e., latent variables) act as causes on observed variables. This use started when Charles Spearman observed that a number of different variables (specifically, children's grades in different subjects) followed a specific pattern of constraints[2] on their correlations. He then used this pattern to justify his claim that there exists a common factor corresponding to general intelligence (which he named G). Spearman's observed pattern of correlations failed to hold in general, leading to modifications of Spearman's method by his students. These modifications enabled Spearman's method to account for more than a single factor; however, they were computationally unfeasible (Glymour et al., 1987). Later modifications by Thurstone (1934) created a version of factor analysis that didn't suffer the same computational limitations.

Factor analysis suffers from a rather serious theoretical difficulty (even when all of the assumptions necessary for factor analysis to work are met), often referred to as the rotation problem (Shalizi, 2012). It involves the use of an orthogonal matrix, used to rotate the original coordinate system. Illustrated below:

Let $\mathbf{m}$ be an orthogonal matrix[3]. Begin by inserting an orthogonal matrix, left multiplied by its transpose.

$$\mathbf{X} = \mathbf{F}\mathbf{m}^T\mathbf{m}\mathbf{w} + \epsilon \tag{3.2}$$

$$= \alpha\beta + \epsilon \tag{3.3}$$

[2]Known as the tetrad constraints

[3]a matrix ($\mathbf{m}$) is orthogonal if $\mathbf{m}^T\mathbf{m} = \mathbf{I}$. Note that $\mathbf{I}$ is the Identity matrix

Where α and β are the new versions of $\mathbf{F}$ and $\mathbf{w}$ (respectively).

This new factor model has the same number of latents, the same fit, and has the same distributions as before the rotation. However, the rotation has changed the structure reported by the previous factor model (Shalizi, 2012).

3.3.3 Tetrad Methods

Vanishing Tetrad differences (Spirtes et al., 2000)[pg. 149-150] can be used to determine the number of latent variables under some conditions (such as linearity) For instance:

If ρ_{ij} is the correlation between variables i and j, then

$\rho_{12}\rho_{34} - \rho_{13}\rho_{24} =$

$\rho_{14}\rho_{23} - \rho_{12}\rho_{34} =$

$\rho_{13}\rho_{24} - \rho_{14}\rho_{23} = 0$

holds in Figure 3.7 (left), however, only $\rho_{13}\rho_{24} - \rho_{14}\rho_{23} = 0$ holds in Figure 3.7 (right).

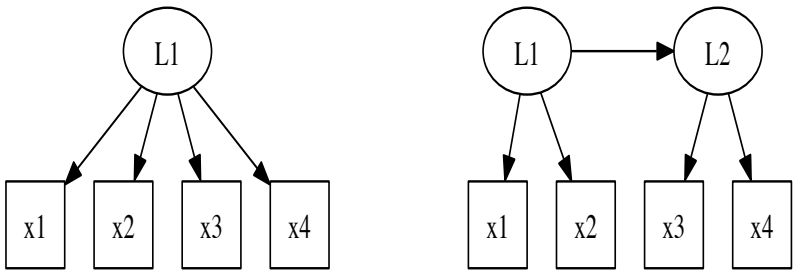

Figure 3.7: The structure on the left has three vanishing tetrad differences whereas the structure on the right only has one.

3.4 Nonparametric Methods

A method is nonparametric if it adds no constraints to the relationships in data beyond those imposed by the graph.

3.4.1 EM Algorithm

The EM algorithm begins by making an initial guess for each parameter or conditional probability in need of estimation. It then takes the guess for each parameter, and finds the posterior distribution for the latent variables (the E step). These posterior distributions are then used to calculate the complete-data log likelihood function. Next, it finds the parameter values that maximize the calculated log likelihood (M step). Then it repeats the E step using these results (and the M step, using the results of the E step) until the parameter estimates don't change by very much (Bishop, 2006)[pg. 616].

The first major drawback to using the EM algorithm for causal modeling is the amount of computing time needed to successfully complete the algorithm. This amount of time varies, depending on the particulars of the problem being attacked. The amount of time isn't known in advance, and it is difficult to be certain that the algorithm has successfully approximated the actual maximum likelihood[4].

This computational problem also becomes increasingly worse as the number of parameters in need of estimation increases (this is known as the Curse of Dimensionality). As a result, the more latents believed to be in need of estimating, the longer the algorithm takes to finish.

Finally, while the EM algorithm is generally used to cluster observations according to the effect of some estimated latent variable(s), there are some

[4]It is possible that the maximum reported by the algorithm is not the true maximum, but is instead merely a local maximum. There are some checks that can be done for this, (such as choosing very spread-out starting points for the algorithm); however, doing so increases the amount of computing time.

instances of EM being used to discover latent variable structures (Elidan et al., 2001). This version of the algorithm (referred to as the Structural EM), was found to be highly unreliable when used for causal discovery in the presence of latent variables (Tillman et al., 2008).

Chapter 4

Method

This chapter gives a description of the detect.MIMIC algorithm, as well as the assumptions necessary for the algorithm to function correctly.

4.1 Description of Method

4.1.1 Detect.MIMIC Algorithm

The following section contains a written description of the detect.MIMIC algorithm. A step-by-step example is also included at the end. While software applications often assume that the partition of measured variables into input and output is already given, when the MIMIC assumptions (mentioned in Section 2.2) are met that classification can be done automatically with the PC algorithm, which also estimates which input variables influence which output variables.

The PC algorithm (with depth set to zero) operates as follows:

1. Draw an undirected path from each variable to every other variable.
2. For every pair of connected variables, if the two variables are independent

of one another (without conditioning on any other variables) then remove the undirected path connecting them.

3. For every pair of variables that are connected after step 2, if the pair of variables remain independent of one another after conditioning on every other variable adjacent to them, then remove the path connecting those two variables.

Inputs and Outputs

Run the PC algorithm on the data with depth (i.e., the maximal size of the conditioning set) set to zero. In the resulting graph, all nodes (variables) with an indegree of zero are members of the input set. All other variables are members of the output set.

Clustering and Determining Arrow Structure

Using the graph found with the PC algorithm, hypothesize a latent variable between every subset of inputs and outputs, where the subset of outputs all have paths from the same subset of inputs. Next, draw a path from the input set of each latent to the latent. Having done this, we can now determine the arrow structure between the latents.

For each latent variable, if the input set of one latent ($IN(L_1)$) is a proper subset of the other latent's input set ($IN(L_2)$), then draw a path from the first latent to the second latent. Also, remove the proper subset from the input set of L_2. Then draw paths from each input to its respective latent.

Finally, only draw a path from a latent to an output variable if that output is not also a member of the output set of some latent descendent.

4.1.2 A Worked Through Example

The algorithm is perhaps best illustrated by walking through a simple example. Take the following graph (Figure 4.1), which the algorithm will attempt to discover:

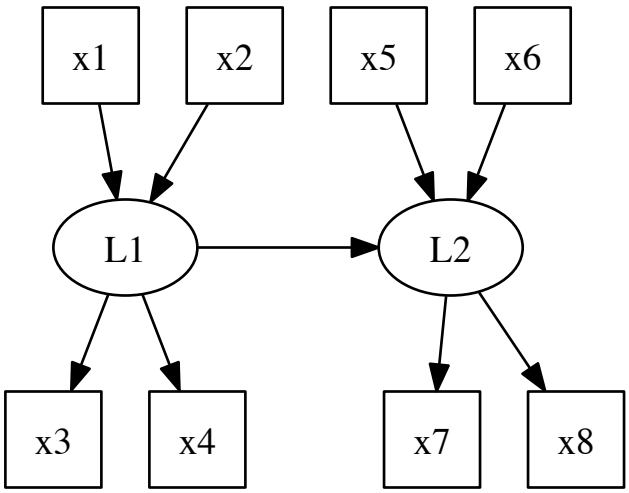

Figure 4.1: An example of a structure to be discovered.

Step 1: Finding Inputs and Outputs.

Running the PC algorithm produces the following graph:

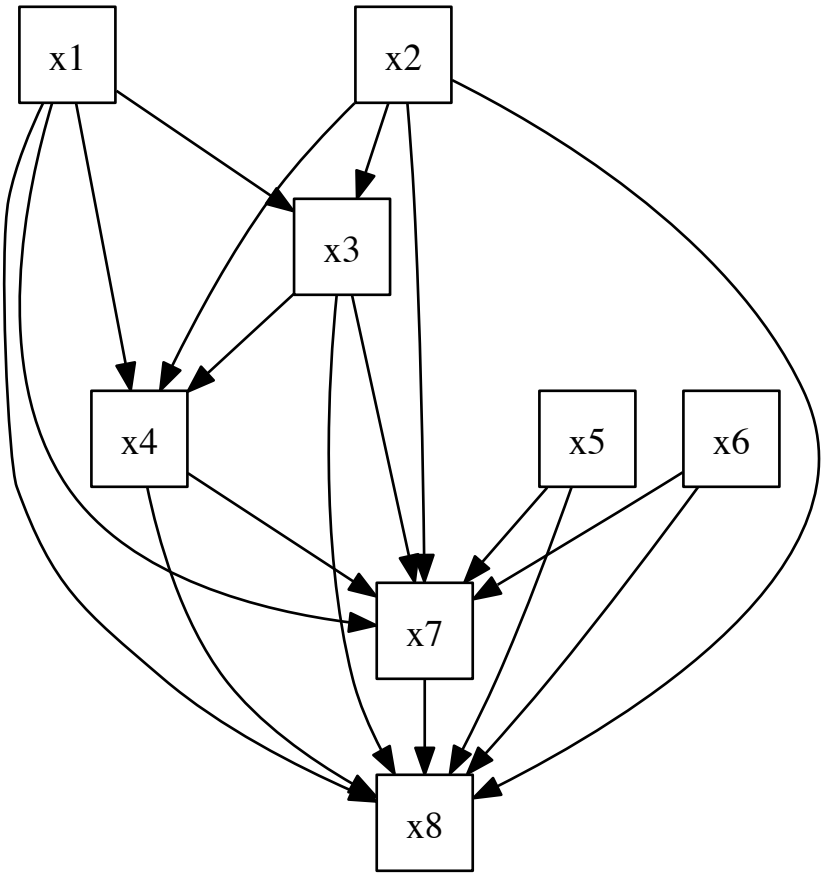

Figure 4.2: Structure reported by the PC algorithm.

At this point, we know that the inputs are x_1, x_2, x_5 and x_6, as each of these variables have an indegree of zero.

We also know that the outputs are x_3, x_4, x_7 and x_8, as each of these variables has an indegree greater than zero.

Step 2: Clustering and Determining Arrow Structure.

Continuing the example, we can see in Figure 4.2, there are two sets of outputs with common inputs. Namely, the set $[x_1, x_2, x_5, x_6]$, and the set $[x_1, x_2]$. Therefore we posit the existence of two latents, with each latent placed between one of the two input sets (see Figure 4.3).

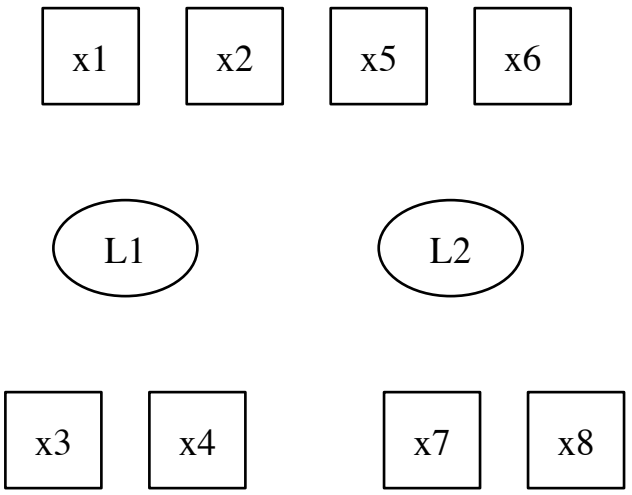

Figure 4.3: Introduction of the latents.

The algorithm now notes that $[x_1, x_2]$ is a proper subset of $[x_1, x_2, x_5, x_6]$. It, therefore, draws a path from L_1 to L_2, as well as paths from the input sets to their respective latents (see Figure 4.4).

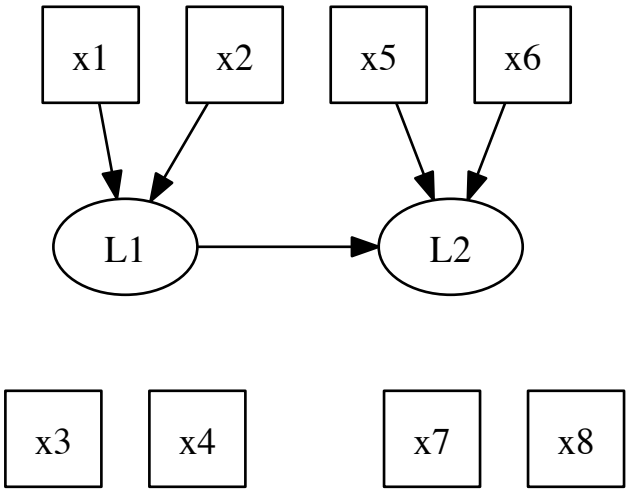

Figure 4.4: Drawing the paths amongst latents.

Finally, the algorithm notes that in Figure 4.2, the output set of [x_1, x_2] that is not part of the output set for [x_5, x_6] consists of [x_3, x_4]. It therefore draws paths from L_1 to [x_3, x_4]. As the only outputs left unclustered belong to the output set of [x_5, x_6], paths are drawn from L_2 to [x_7, x_8] (giving Figure 4.5).

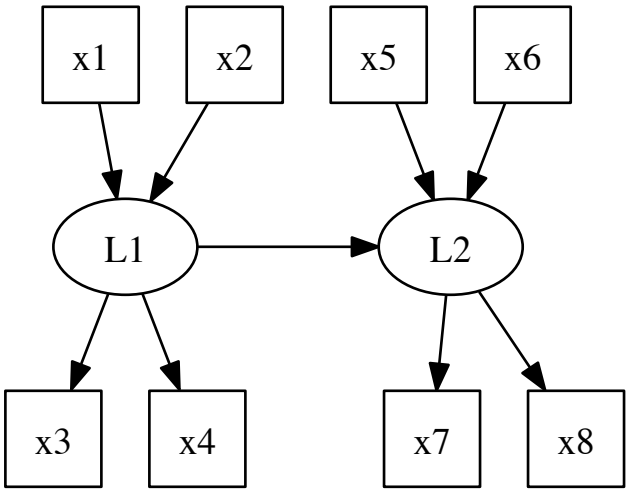

Figure 4.5: Final graph produced by the algorithm.

4.1.3 Assumptions and Limitations

Despite the search space being restricted to MIMIC models, there are some structures that are indistinguishable when using the detect.MIMIC algorithm. Unfortunately, the complete class of indistinguishability structures for MIMIC models is not presently known. Therefore, some restrictions have been made about allowable structures beyond those imposed by the definition of a MIMIC model, so as to further reduce the search space to a manageable size.

First, we attempt to identify only models with singly connected graphs, i.e., between any two variables there is at most one directed path.

Definition 4.1.1 (Minimal Model)**.** A model consisting of a graph and distribution pair satisfying the Markov condition is minimal provided the removal

of any edge from the graph results in a graph/distribution pair that does not satisfy the Markov condition.

We also add an additional constraint. Namely, the simplicity criterion.

Definition 4.1.2 (Simplicity Criterion)**.** Of any indistinguishable set of models, chose the one that has the fewest edges.

In practice, this means that although graphs in Figures 4.6 and 4.7 satisfy the same constraints, given data from either of these structures the DETECT procedure would identify the second graph, Figure 4.7, rather than the first, because a single latent is introduced for every class of output variables and each input has a unit outdegree. Although I do not pursue the question here it would seem straightforward to characterize the data-equivalent, singly connected graphs that differ in the respects illustrated by Figures 4.6 and 4.7, and I pose that as a research problem.

Finally, I also require that the structure be singly connected.

Definition 4.1.3 (Singly Connected)**.** A structure is singly connected if there at most a single path connecting any pair of nodes.

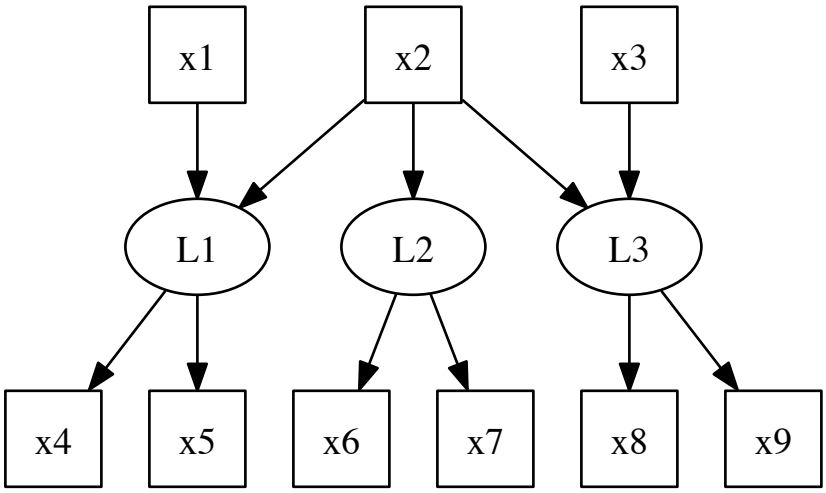

Figure 4.6: An example of a structure indistinguishable from Figure 4.7.

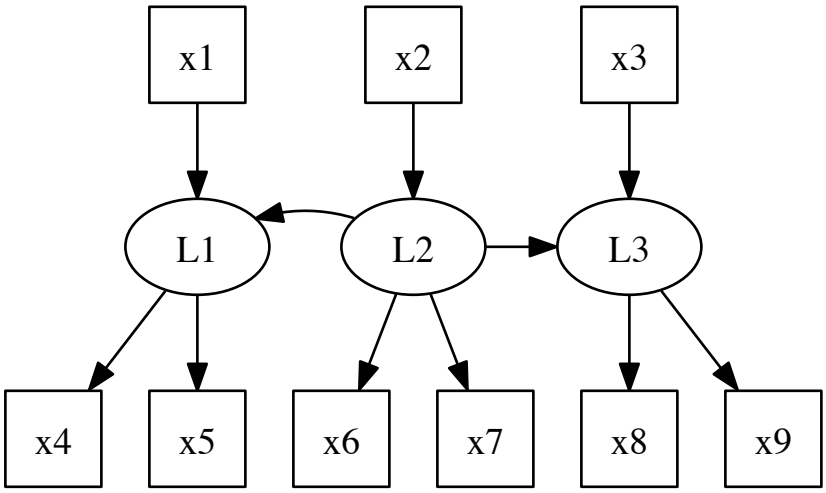

Figure 4.7: An example of a structure indistinguishable from Figure 4.6.

Additionally, the third restriction (i.e., the structure must be simply connected) allows us to exclude the structure in Figure 4.8, which has multiple paths between L_2 and x_8. These multiple paths create a graph whose observable independence relations are identical to those in Figure 4.6 (i.e., the graphs are both members of the same equivalence class), preventing us from determining which structure is correct.

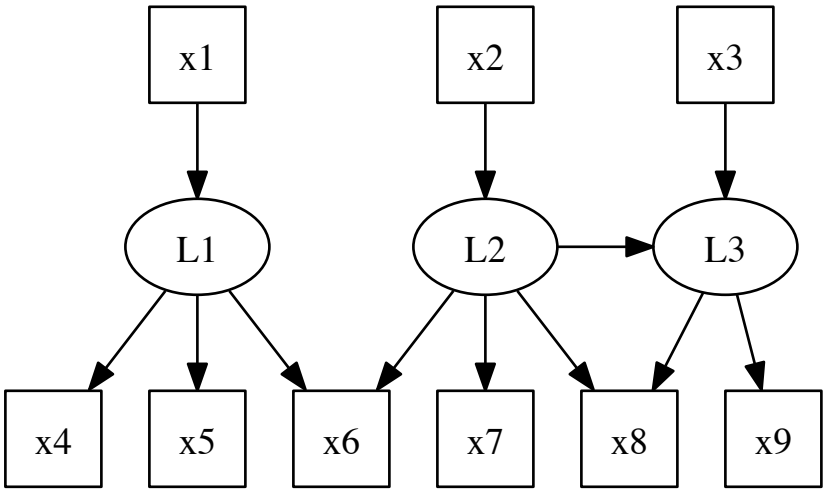

Figure 4.8: An example of a multiply connected structure

Chapter 5

Conclusion

5.1 Summary

At this point, the detection of MIMIC models is in a better state than it previously was. Instead of having no reliable methods, there is now an incomplete (but reliable within its restrictions) method for detecting MIMIC models. This new method is built on a thoroughly researched algorithm (the PC algorithm); hence, there is already a body of work that (with some adjustments) can be applied to understanding the limitations and capabilities of the new algorithm.

5.2 Future Work

Future work can be divided into two main subsections, improvements of the algorithm, and improvements in the applicability of the algorithm.

Currently, the proposed algorithm cannot handle all MIMIC models in their full generality, as the indistinguishability classes for MIMIC models are presently unknown. In order to create a more general form of the algorithm, it will be necessary to determine these classes.

Other work of interest involves the inclusion or combination of detect.MIMIC in (or with) other more general algorithms (such as the PC algorithm). Additionally, a formal estimate of algorithmic complexity, as well as systematic analyses of the algorithm's consistency and efficiency are rather important for understanding the applicability of the algorithm in practical data analysis.

Bibliography

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Bühn, A. and Schneider, F. (2008). MIMIC Models, Cointegration and Error Correction: An Application to the French Shadow Economy.

DellAnno, R. and Schneider, F. (2006). Estimating the Underground Economy by Using MIMIC Models: A Response to T. Breusch´s Critique. Technical report.

Elidan, G., Lotner, N., Friedman, N., Koller, D., et al. (2001). Discovering Hidden Variables: A Structure-based Approach. *Advances in Neural Information Processing Systems*, pages 479–485.

Giles, D. (1999). Measuring the Hidden Economy: Implications for Econometric Modelling. *The Economic Journal*, 109(456):370–380.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C., and Zhang, J. (2010). Actual Causation: A Stone Soup Essay. *Synthese*, 175(2):169–192.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press.

Lester, L. (2008). A Multiple Indicators and Multiple Causes (MIMIC) Model of Immigrant Settlement Success.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Ríos-Bedoya, C., Pomerleau, C., Neuman, R., and Pomerleau, O. (2009). Using MIMIC Models to Examine the Relationship Between Current Smoking and Early Smoking Experiences. *Nicotine & Tobacco Research*, 11(9):1035–1041.

Scheines, R. (1997). An Introduction to Causal Inference.

Shalizi, C. (2012). *Advanced Data Analysis from an Elementary Point of View*. Unpublished;http://www.stat.cmu.edu/∼cshalizi/uADA/12/lectures/ADAfaEPoV.pdf.

Shimizu, S. Hyvärinen, A. K. Y. and Hoyer, P. (2005). Discovery of Non-Gaussian Linear Causal Models Using ICA. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, 7:526–533.

Silva, R., Scheine, R., Glymour, C., and Spirtes, P. (2006). Learning the Structure of Linear Latent Variable Models. *The Journal of Machine Learning Research*, 7:191–246.

Sober, E. (1998). Black Box Inference: When Should Intervening Variables be Postulated? *The British Journal for the Philosophy of Science*, 49(3):469–498.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation Prediction & Search 2e*. Adaptive Computation and Machine Learning Series. MIT Press.

Tedds, L. (1998). Measuring the Size of the Hidden Economy in Canada: A Latent Variable/MIMIC Model Approach. *MA Extended Essay, Department of Economics, University of Victoria*.

Thurstone, L. (1934). The Vectors of Mind. *Psychological Review; Psychological Review*, 41(1):1.

Tillman, R., Danks, D., and Glymour, C. (2008). Integrating Locally Learned Causal Structures with Overlapping Variables. *Advances in Neural Information Processing Systems*, 21:1665–1672.