Journal of Machine Learning Research 1 (2018) 1-48 Submitted 4/00; Published 10/00

Evaluation of Causal Structure Learning Methods on Mixed Data Types

Vineet K. Raghu VINEET@CS.PITT.EDU
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15213, USA

Allen Poon ALP170@PITT.EDU
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15213, USA

Panayiotis V. Benos BENOS@PITT.EDU
Department of Computational and Systems Biology
University of Pittsburgh
Pittsburgh, PA 15213, USA

Editor:

Abstract

Causal structure learning algorithms are very important in many fields, including biomedical sciences, because they can uncover the underlying causal network structure from observational data. Several such algorithms have been developed over the years, but they usually operate on datasets of a single data type: continuous or discrete variables only. More recently, we and others have proposed new causal structure learning algorithms for mixed data types. However, to-date there is no study that critically evaluates these methods' performance. In this paper, we provide the first extensive empirical evaluation of several popular causal structure learning methods on mixed data types and in a variety of parameter settings and sample sizes. Our results serve as a guide as to which method performs the best in a given context, and as such they are a first step towards a "method selection guide" for those running causal modeling methods on real life datasets.

Keywords: Causal Discovery, Mixed Data, Empirical Evaluation

1. Introduction

Causal discovery from observational data has been a topic of growing interest for several years (Spirtes et al., 2000; Pearl, 2009). The general problem of causal discovery is to infer a graphical structure from data where nodes in the graph correspond to random variables in the data, and edges in the graph depict direct (causal) relationships among the variables. This problem is crucial to many domains such as economics, social sciences, biology, and biomedicine, as causal knowledge is what allows researchers to understand the effectors of different variables or outcomes and generate hypotheses on how interventions will change the system under study. For example, in biomedicine, understanding causal relationships

allows a physician to predict the outcome of a treatment and which variables most affect it, a task which cannot be done using correlations alone.

Automated approaches to causal discovery from observational data fall into two main categories: constraint-based (Spirtes et al., 2000) and score-based (Chickering, 2002). Generally, constraint-based approaches identify the causal structure by starting with a fully connected undirected graph and using conditional independence tests to eliminate edges. Finally, the graph is oriented via a series of rules (identifying v-structures or colliders, avoiding cycles, etc.). The score-based approaches use a score specifying the goodness of fit of the model to the data subject to a sparsity penalty to avoid overfitting. Both approaches have demonstrated success in the past, and hybrid approaches have also been proposed (Sokolova et al., 2014; Tsamardinos et al., 2006).

Currently, there are many algorithms for causal discovery from observational data, but empirical evaluations and comparisons in different experimental settings are lacking. Thus, it is unclear to researchers attempting to use these methods, which algorithm is appropriate for their particular dataset and domain. In this work, we perform a thorough empirical evaluation of several causal discovery techniques across datasets with different properties. We provide the reader with a better understanding of the practical properties of each of these techniques in learning causal structure from simulated data.

In the past, there were few comparative studies of algorithms for causal structure learning from observational data. In Tsamardinos et al. (2006), the authors conduct a thorough test of the state of the art methods for causal structure learning. However, this evaluation is limited by the methods available at the time at which it was conducted, and as such it missed several modifications to those algorithms that have improved their performance (Ramsey et al., 2006; Ramsey, 2015; Colombo and Maathuis, 2014). A more recent benchmarking of causal discovery techniques was performed on both simulated and real biological data (Singh et al., 2017). This evaluation goes beyond learning the structure of causal graphs to determining whether these causal graphs are able to accurately predict the result of interventions. However, this evaluation is limited by the small size of the simulated datasets (10 variables) compared to what would actually be encountered in practice. In addition, this evaluation is limited by the nature of the real biomedical datasets used. Though real data is good for evaluations in the sense that it can give better indication of how algorithms will generalize to real data in the future, it is unclear whether the "ground truth" in these cases is actually the truth in nature or just our current understanding of the truth.

Further, one of the challenges faced by these algorithms is that often times the observational datasets contain mixed data types (continuous and discrete variables). In these cases, constraint-based algorithms require an independence test suitable for determining the independence of a mixed set of variables, and score-based algorithms require a likelihood score that can handle these sorts of data. Recently, attention has been focused on causal structure learning from mixed data, and both constraint- and score-based algorithms have been developed for this purpose (Andrews et al., 2017; Sedgewick et al., 2017; Tsagris et al., 2018); however, to our knowledge no hybrid constraint and score based approaches have been developed and tested. Other methods have been developed that are able to address mixed data using answer-set programming (Hyttinen and Jrvisalo, 2014; Borboudakis and Tsamardinos, 2016); however, these methods are not able to scale to the high-dimensional

datasets studied in this work that are of interest to many domain experts (e.g. biomedicine, biology).

Performing a thorough evaluation of causal structure learning techniques presents a number of challenges. Experimental parameters for data generation, algorithmic parameters, and number of replicates can result in a complexity too large to both run the experiment and interpret the results. In addition, finding suitable metrics for a particular experiment are a challenge as different causal structure learning algorithm can provide different information. In addition, ensuring that the simulation parameters match up with the expectations of real data is a persistent challenge. Doing this evaluation with mixed datasets increases these challenges, as runtime of algorithms tends to be longer, and there are more experimental parameters to generate the data and evaluate the results.

Here, we address this challenge by focusing upon datasets with similar properties to those that we expect to have in biomedical domains. In particular, we present a comparative empirical evaluation of algorithms for causal structure learning from simulated high-dimensional mixed data. Our specific contributions are the following:

- We evaluate several popular causal structure learning algorithms on continuous datasets of varying sample and variable sizes.
- We compare causal structure learning algorithms for mixed data against methods that treat these datasets as fully continuous or fully categorical data.
- We compare several causal structure learning algorithms specifically for mixed datasets across different sample and variable sizes and different algorithmic parameters.

2. Methods Evaluated

In this section, we briefly discuss the methods compared in this study. For full details of each method we refer the reader to their original publications. First, we discuss algorithms for learning causal structure from data, then we discuss the independence tests and scores used for evaluation, and finally we discuss methods for parameter selection.

2.1 Algorithms to learn Causal Structure

2.1.1 PC and Variants

PC has been one of the most popular constraint-based algorithms for causal structure learning from observational data (Spirtes et al., 2000). The algorithm begins with a fully connected graph and performs conditional independence tests with increasing conditional set size in order to remove edges. For example, for an edge between X and Y the algorithm first performs an unconditional independence test between X and Y, and then conditional independence tests of X and Y given S with $|S| = 1$ to $|S| = N$, where N is the number of nodes adjacent to X or Y. The edge between X and Y is removed if they are independent given some set S. After determining the edges in the graph, the algorithm then orients them according to a set of rules. First, it orients unshielded colliders (i.e. when X and Y are independent, but dependent conditional on a third variable Z, then X and Y are both parents of Z). Then, it orients edges to avoid cycles and prevent creating extra unshielded

colliders. The algorithm requires an appropriate conditional independence test for the type of data being analyzed.

Since its induction, many variants have been proposed for the PC algorithm, and we choose to not include PC itself in our evaluation, as the modifications have better theoretical properties and empirical performance. PC-Stable is a modification that avoids the order dependence problem present in PC, as the output of PC depends upon the order in which independence tests are performed (Colombo and Maathuis, 2014). Conservative-PC (CPC) uses a conservative strategy when determining whether to orient an edge as a collider. In particular, the algorithm orients an edge X-Y-Z as a collider only if Y appears in *none* of the possible conditioning sets which separate X and Z (Ramsey et al., 2006). Unlike CPC, PC-Max uses the conditioning set with the largest p-value in its conditional independence test to determine if a collider should be oriented (Ramsey, 2016).

Copula-PC is a relatively new modification of the PC algorithm that is designed for mixed continuous and ordinal datasets (Cui et al., 2016). The method first infers a rank correlation matrix using a projected inverse-wishart distribution as a prior distribution on the correlation matrix, and then using a Gibbs sampling approach to generate samples from this distribution to finally infer a posterior distribution and an estimated correlation matrix. Then this estimated correlation matrix along with a modified effective sample size is fed to the PC algorithm as usual to infer a causal graph. Though the algorithm is designed for continuous and ordinal data (monotonic relationships) we test the algorithm on continuous and categorical datasets.

2.1.2 Greedy Equivalence Search

Greedy Equivalence Search (GES) is a popular score-based algorithm to learn causal structure. Unlike the PC variants, GES does not use conditional independence tests and instead greedily optimizes a likelihood score subject to a sparsity penalty to avoid overfitting (Chickering, 2002). In this work, we use the modification called the Fast Greedy Search (FGES), as this method was shown to give accurate causal predictions while achieving significantly better runtime (Ramsey, 2015). FGES has recently been extended to mixed data via a new Conditional Gaussian scoring function (Andrews et al., 2017). This method uses the standard Bayesian Information Criterion (BIC) score where the likelihood is computed by modeling each continuous variable as a unique Gaussian distribution for each setting of its discrete parents.

2.2 Independence Tests for Mixed Datasets

All constraint-based causal discovery algorithms require an independence test to learn causal structure. In this work, we evaluate two independence tests for mixed datasets: a Multinomial Logistic Regression test (Multinomial LRT), and a Conditional Gaussian test (CG).

2.2.1 Multinomial Logistic Regression Test

The Multinomial LRT test performs different conditional independence tests depending upon the type of variables involved. Assume that we are testing the conditional independence of variables X and Y given a set of variables $\boldsymbol{S}$. If X and Y are both continuous, then the test is simply a linear regression of X given Y and $\boldsymbol{S}$. Any categorical variables in

$\boldsymbol{S}$ are converted into binary indicator variables for each category, which is necessary when using a regression based approach since the categories do not necessarily have a well-defined ordering and scale (e.g. Ethnicity). X is determined to be dependent on Y if the coefficient is significantly different than zero by a t-test. If X or Y is categorical, then the test is a likelihood ratio test. Here, we use logistic regressions to test whether $P(X|Y, \boldsymbol{S}) = P(X|\boldsymbol{S})$, which is expected to be true under the null hypothesis of independence. In particular, the log of the likelihood ratio between these two regressions is known to follow a chi-square distribution, so the significance of this ratio can be computed accordingly.

2.2.2 Conditional Gaussian Test

Unlike the multinomial test, the Conditional Gaussian (CG) test assumes that the data is generated from a distinct multivariate Gaussian distribution for each setting of the discrete variables (Andrews et al., 2017). In order to perform a conditional independence test of X and Y given S, a likelihood ratio test is performed between $P(X|Z)$ and $P(X|Y, Z)$ as well as $P(Y|Z)$ and $P(Y|X, Z)$. The null hypothesis of independence is rejected if either direction gives evidence of dependence. To compute the likelihood for the test, the conditional Gaussian approximation is used:

$$P(X|Y, Z) = \frac{P(X_c, Y_c, Z_c|X_d, Y_d, Z_d)P(X_d, Y_d, Z_d)}{P(Y_c, Z_c|Y_d, Z_d)P(Y_d, Z_d)} \tag{1}$$

Here, X_c, Y_c, and Z_c denote the continuous variables among X, Y, and Z, whereas X_d, Y_d, Z_d denote the discrete variables in this group. The likelihoods involving only discrete variables are computed via a multinomial distribution, whereas the conditional likelihoods are computed using a conditional Gaussian distribution. For full details of this test, we refer the reader to the original publication.

2.3 Mixed Graphical Models (MGM)

A Mixed Graphical Model (MGM) is an undirected graphical model which characterizes the joint distribution over a dataset with both continuous and discrete variables, and it is given by the following expression (Lee and Hastie, 2013):

$$p(x, y; \theta) \propto exp\Big(\sum_{s=1}^{p} \sum_{t=1}^{p} -\frac{1}{2}\beta_{st} x_s x_t + \sum_{s=1}^{p} \alpha_s x_s + \sum_{s=1}^{p} \sum_{j=1}^{q} \rho_{sj}(y_j) x_s + \sum_{j=1}^{q} \sum_{r=1}^{q} \phi_{rj}(y_r, y_j) \Big) \tag{2}$$

where θ represents all parameters of the model, x_s represents the s^{th} of p continuous variables and y_j represents the j^{th} of q discrete variables. β_{st} represents the potential for an edge between continuous variables s and t, α_s represents the potential for a node of a continuous variable, ρ_{sj} represents the potential for an edge between continuous variable s and discrete variable j, and finally ϕ_{rj} represents the potential for an edge between discrete variables r and j. This joint distribution can be decomposed into conditional distributions given by

Gaussian linear regression and Multiclass Logistic Regression for continuous and discrete variables respectively.

$$\widetilde{l}(\Theta|x,y) = -\sum_{s=1}^{p} \log p(x_s|x_{/s}, y; \Theta) - \sum_{r=1}^{q} \log p(y_r|x, y_{/r}; \Theta) \tag{3}$$

Learning this model over high dimensional datasets directly is computationally infeasible due to the computation of the partition function, so to avoid this, a proximal gradient method is used to learn a penalized negative log pseudolikelihood form of the model. This negative log pseudolikelihood is given in Equation 3, and the penalized form is presented in Equation 4 and described in (Sedgewick et al., 2016). For both of these, Θ refers to all parameters of the model collectively.

$$\underset{\Theta}{\text{minimize}}\, l_\lambda(\Theta) = \widetilde{l}(\Theta) + \lambda_{CC} \sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \lambda_{CD} \sum_{s=1}^{p} \sum_{j=1}^{q} ||\rho_{sj}||_2 + \lambda_{DD} \sum_{j=1}^{q} \sum_{r=1}^{j-1} ||\phi_{rj}||_F \tag{4}$$

Here, the algorithm is used as specified in (Sedgewick et al., 2016), (called CausalMGM). As in previous publications, the algorithm is terminated when the MGM graph remains unchanged for three consecutive iterations, and the learned model is used as an input graph to causal structure learning algorithms by starting with an undirected sparse model instead of a fully connected graph, which has shown promise in causal discovery (Loh and Bühlmann, 2014; Sedgewick et al., 2017). In this work, we test the impact of using CausalMGM with constraint-based algorithms on mixed data.

2.4 StARS and StEPS for Parameter Selection

Both the constraint and score based structure learning algorithms require the selection of a parameter. Constraint-based algorithms require a parameter α, which determines the p-value cutoff for making independence test decisions, whereas score-based algorithms require a penalty discount parameter to determine how heavily including edges in the model should be penalized (or equivalently how sparse the final output graph should be). In many cases, there is no clear way to determine the values of these parameters, so one method that we choose to evaluate in this paper is StARS (Liu et al., 2010), as it outperformed several competing parameter selection metrics on simulated and real data. Essentially, StARS selects the parameter value that gives the most stable (insensitive to variations in the data), yet sparse graph when subject to random subsampling without replacement. For CausalMGM, we used a related method called StEPS, which performs a similar selection as StARS, except for the three edge-type dependent sparsity parameters as it is required by the CausalMGM method (Sedgewick et al., 2016). The benefit of using this approach is that one selects a threshold for stability that has an interpretable meaning, as opposed to choosing a parameter value arbitrarily for causal discovery, and in addition, using a standard instability threshold of 0.05 has shown good success in causal discovery (Sedgewick et al., 2016).

3. Results

Next, we present the main contributions of the paper. We begin by discussing an evaluation of algorithms for causal structure learning from entirely continuous data across a variety of parameters. We then evaluate different independence tests for mixed data against tests which treat the data as purely continuous or purely categorical. Finally, we evaluate methods for learning causal structure from mixed continuous and categorical data using the best performing independence test against the modification of FGES for mixed data.

3.1 Experimental Procedure

Simulated data. In all of the subsequent experiments, simulated data was generated in a similar manner. First, a ground truth graph was generated uniformly at random from the set of all directed acyclic graphs with N edges, where N was normally distributed with mean equal to either 1.5 (referred to as graph density 3) or 2.5 (referred to as graph density 5) times the number of variables, and standard deviation equal to half the number of variables. This graph was then parametrized with edge weights selected uniformly at random from the range $(-1.5, -0.5), (0.5, 1.5)$. Using this parametrized graph, samples were generated independently using a linear Gaussian model for continuous data, and using both a Lee and Hastie model (Lee and Hastie, 2013) and a Conditional Gaussian (Andrews et al., 2017) model for mixed data with a 50-50 split between four-category categorical and continuous variables, resulting in 25% of the edges being between continuous variables, 25% of the edges between categorical variables, and 50% of the edges being mixed type. To ensure reasonable statistical power, each categorical variable was constrained to have at least four samples for each category.

All algorithms were tested using a variety of parameters. The ground truth causal graphs consisted of 50 or 100 variables, and the datasets had sample sizes in the set: (100, 1000, 3000, 5000) for continuous data, and (100, 300, 500) for mixed data. For constraint-based causal discovery algorithms, α values were taken from the set: (1E-4, 0.001, 0.01, 0.03, 0.05, 0.08, 0.1). For FGES, penalty discount values were taken from the set: (0.5, 1, 2, 4, 8, 10, 20) for continuous data, and the binomial structure prior was taken from the set (1,1.5,2,3,4,5) for mixed data. In addition, StARS was used to automatically select a stable parameter value from these sets, with the stability threshold set to the value suggested in the original publication, 95%.

Evaluation metrics and specifications. Several evaluation metrics were used to determine the accuracy of the Partially Directed Acyclic Graph (PDAG) estimated by the algorithms in comparison to the ground truth data generating graph. Adjacency precision and recall refer to the correctness of the edges in the graph estimated by the search algorithms, with precision and recall taking the standard definition from the literature. Arrowhead precision is computed by taking the number of correctly placed arrowheads (causal orientations) divided by the number of correct arrowheads plus false positive arrowheads. Recall is defined in a similar way but with false negative (missed) arrowhead placements. Finally, structural hamming distance (SHD) refers to the number of changes that must be made to an estimated causal graph to recreate the ground truth graph. For mixed continuous and discrete datasets, these metrics are further split by edge type: i.e., edges between two

Table 1: Evaluation of causal structure learning methods from continuous data. The results are shown for the parameter which gives the best performance (Oracle). *AP* and *AR* are adjacency precision and recall; *AHP* and *AHR* are arrowhead precision and recall; *SHD* is structural hamming distance. The highest values for each sample size, metric category are bolded.

Algorithm	SS	Parameter	AP	AR	AHP	AHR	SHD
CPC	100	0.03	**0.9290**	0.5953	**0.9537**	0.2547	187.5
PC-Max	100	0.03	0.9090	0.6013	0.7635	0.3533	185.25
PC-Stable	100	0.03	**0.9290**	0.5953	0.5494	0.4277	203.2
FGES	100	1	0.8864	**0.6943**	0.8054	**0.5267**	**150.8**
CPC	1000	0.01	**0.9372**	0.7557	**0.9400**	0.4927	129.2
PC-Max	1000	0.01	0.9111	0.7687	0.8326	0.5833	127.75
PC-Stable	1000	0.01	**0.9372**	0.7557	0.6224	0.6010	155
FGES	1000	1	0.8316	**0.9033**	0.7660	**0.7470**	**115.35**

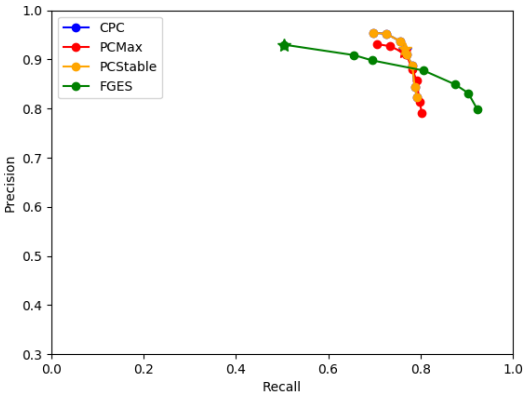
continuous variables (CC), a continuous and a discrete variable (CD), and two discrete variables (DD).

All combinations of causal structure learning algorithm, parameter value, number of variables in the data, and number of samples in the data were repeated across 20 graphs, and the results are reported as averages across these graphs. For StARS, the parameter selection was done independently for each graph, and we only report the most often chosen parameter value for each of these combinations. All experiments were performed on an 8 core machine with a 256 GB SSD, 256 GB of RAM, and a 2.4 GHz processor. Parallelized versions of causal structure learning algorithms were used, and all runtime measurements are given in CPU time.

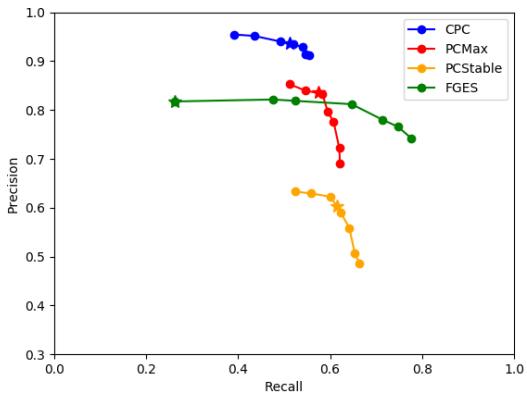
3.2 Evaluation of Structure Learning from Continuous Data

We first examined how causal structure learning algorithms perform on continuous datasets using the parameter with the best results (Oracle) in terms of SHD shown in Table 1. Two sample sizes (100 and 1,000) are displayed here. *Parameter* refers to α level for constraint-based methods and the penalty for FGES. Each entry is an average over 20 graphs with 100 variables each. In this experiment, all algorithms used the Fisher Z test of conditional independence (Spirtes et al., 2000) for consistency.

In terms of reliability, all three constraint-based approaches tend to give high precision for adjacencies, which could be attributed to the repeated independence tests that must be performed for each edge under varying conditioning sets, where even a single p-value above the α threshold will result in an edge deletion. As a drawback, this results in significantly poorer recall when compared to FGES, which we find to hold regardless of sample size. In terms of causal orientations, CPC and PC-Max appear to differ only upon where they fall in terms of the tradeoff between precision and recall, whereas PC-Stable has a large drop in precision for nearly the same recall, suggesting the poorer performance of this method. In terms of causal orientations and SHD, FGES appears to have a significant advantage when the oracle parameter value is chosen, as determined by recall and SHD.



a Adjacency Recovery for 100 Variable, 1000 Sample Data



b Arrowhead Recovery for 100 Variable, 1000 Sample Data

Figure 1: Precision-Recall Plots for 100 variable, 1000 sample, continuous datasets. The star refers to the average parameter values chosen by StARS. The CPC results are identical to the PC-Stable results in (a).

Oracle results are difficult to be obtained in practice because we do not generally know the ground truth (and if we did, we wouldn't need a causal search method!). So, it is important to see how algorithms perform using a wide range of parameters. Figure 1 depicts the effect of such parametrization of each algorithm on its performance in terms of adjacency recovery (Figure 1a) and orientation recovery (Figure 1b) for networks with 100 nodes and 1,000 samples each. In these plots, it is clear that FGES's precision for both adjacencies and arrowheads remains unaffected by the choice of penalty in this setting; its recall, however, is affected the most compared to constraint-based methods. In addition, CPC has the best arrowhead recovery performance for all parameterizations. We note that the effects of parameter choices tend to be smaller in larger sample sizes. In terms of algorithm efficiency, all algorithms show good performance with runtime of less than a second on 100 variable, 5000 sample continuous data.

Overall, on continuous data we find that the choice of parameter can have a significant impact on precision for constraint-based algorithms, and on recall for FGES. StARS tends to select parameters with a good balance of precision and recall for constraint-based algorithms; however, it tends to produce very sparse graphs when run in conjunction with FGES. More work must be done to select the instability threshold suitable for FGES to achieve accurate results in these cases.

3.3 Evaluation of Independence Tests for Mixed Data

Before evaluating different algorithms for learning causal structure, we first evaluate how independence tests, specifically designed for mixed data perform. As baselines, we compare these tests to a test that treats mixed data as entirely continuous, and a test that treats mixed data as categorical.

Two mixed data independence tests were evaluated in this study: a Conditional Gaussian independence test (CG), a Multinomial Logistic Regression Test (Multinomial LRT). We

compared those to the Fisher Z test of independence for continuous data, and a Chi-Square test for categorical data (Spirtes et al., 2000). Due to runtime constraints, these tests were evaluated only for datasets with 100, 300, and 500 samples. In these experiments, CPC was used in order to compare these tests in the context of causal structure learning, because it gave reliable structure results in the continuous data simulations, and it was highly efficient.

Table 3 depicts the accuracies of the oracle parameter selection for each independence test on 100 variable, 100 sample datasets with the results split by different edge types: CC are edges between continuous variables, DD are edges between discrete variables, and CD are mixed variable edges. Many of these results are as expected, the FisherZ score is unable to find connections involving categorical variables, as it expects linear relationships, though due to its accuracy with continuous variables the overall result is reasonable. Due to the loss of information when discretizing the dataset, the ChiSquare test suffers on recall for all edge types, and when run with CPC orientations on low sample sizes, the ChiSquare test produces very few arrowheads as it detects faithfulness violations.

The Multinomial LRT test is the only test we found able to achieve decent adjacency recall, and the CG test achieves very high adjacency precision. No independence test is able to detect arrowheads involving categorical data, which is expected for such low sample size. However, the fact that the precision remains high suggests that the mixed data independence tests can be used on this high-dimensional data regardless.

Table 2b gives results when the sample size of the data is increased to 500. The patterns here are largely the same as the 100 sample data, except for a few striking differences. First, the Multinomial LRT test most effectively uses the larger sample size to detect both edges and arrowheads involving categorical variables, whereas the other methods still have very low recall, especially for arrowheads. In terms of precision, all methods remain reliable for both adjacency and arrowhead recovery. Overall, it is clear that the Multinomial test performs the best regardless of edge type due to its superior edge and arrowhead recall.

Figure 3 shows the impact of parameter selection and simulation methodology for each independence test on 500 sample data. It is clear from these plots that the selection of the parameter has a large impact on adjacency and orientation precision for both the FisherZ test and the Multinomial test, but regardless of the selection of the parameter the Multinomial test provides superior recall for similar precision on Lee and Hastie data. For arrowhead recovery, the results appear to be parameter independent, but very dependent upon the simulation methodology, as Conditional Gaussian has superior recall from CG simulated data, and the Multinomial test has superior recall from LH simulated data. However, all methods have very poor recall in CG simulations, and there isn't a significant difference between treating the data as mixed or treating the data as fully continuous or fully discrete. In these experiments, StARS selects the parameter that favors recall the most (the densest graphs), indicating that the instability threshold of 95% is not appropriate in this setting. For causal structure learning, it is not clear whether setting a stability threshold in this way is the best methodology to do parameter selection for mixed datasets.

Finally, Figure 4 shows the average runtime of CPC using each of the four independence tests. Though, the Multinomial test gives substantial better graph estimation performance, we find that the runtime could be prohibitively expensive in some cases. For large mixed datasets, it may be more appropriate to use the CG test to avoid this expensive computation at the expense of detection of edges and orientations.

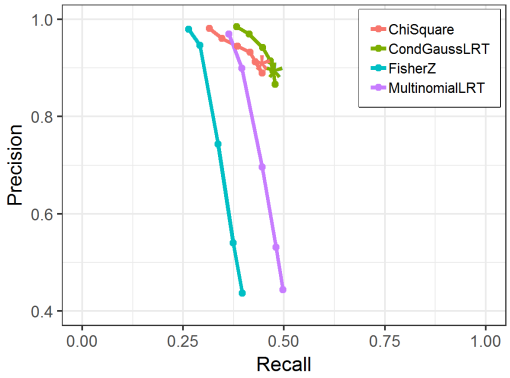
Table 2: Comparison of mixed data independence tests for 100 variable graphs with average graph density 1.5 times the number of variables. Data was generated using the Lee and Hastie simulation method. A * indicates that no arrowheads were predicted by the method.

(a) 100 variable, 100 sample data

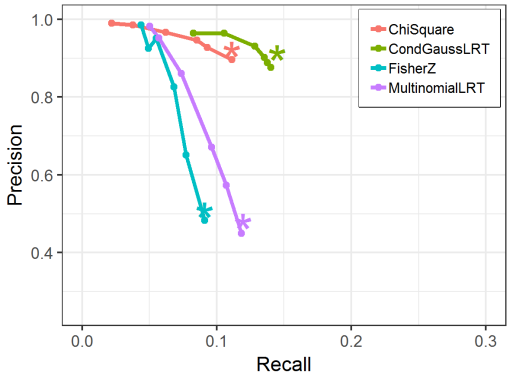
Type	Independence Test	α	AP	AR	AHP	AHR	SHD
CC	FisherZ	0.001	0.9887	0.6703	**1.0000**	0.1676	43.6
	ChiSquare	0.1	0.9411	0.4676	**1.0000**	0.0135	58.4
	Multinomial LRT	0.01	**0.9889**	0.7081	0.9833	0.2027	41
	CG	0.08	0.9370	**0.7595**	0.9750	**0.2703**	**39.9**
CD	FisherZ	0.001	0.9689	0.1682	*	0.0000	121.7
	ChiSquare	0.1	0.9324	0.2758	0.8333	0.0212	115.1
	Multinomial LRT	0.01	0.8539	**0.4818**	0.8519	0.0303	**109.2**
	CG	0.08	**0.9857**	0.0848	**1.0000**	0.0030	126.4
DD	FisherZ	0.001	0.8798	0.0787	*	0.0000	91.3
	ChiSquare	0.1	0.9482	0.3894	0.6429	0.0128	**77.4**
	Multinomial LRT	0.01	0.8251	**0.4915**	**0.7188**	**0.0426**	79.4
	CG	0.08	**1.0000**	0.2149	*	0.0000	83.9
All	FisherZ	0.001	**0.9710**	0.2640	**1.0000**	0.0413	256.6
	ChiSquare	0.1	0.9404	0.3587	0.8102	0.0167	250.9
	Multinomial LRT	0.01	0.8834	**0.5407**	0.9175	**0.0767**	**229.6**
	CG	0.08	0.9563	0.2920	0.9750	0.0680	250.2

(b) 100 variable, 500 sample data.

Type	Independence Test	α	AP	AR	AHP	AHR	SHD
CC	FisherZ	0.001	0.9877	0.9286	**1.0000**	0.5810	21.6
	ChiSquare	0.1	0.9143	0.7881	0.9622	0.2190	48.2
	Multinomial LRT	0.01	0.9974	**0.9452**	0.9934	**0.7190**	**14.5**
	CG	0.03	**1.0000**	0.9262	0.9923	0.5738	21.2
CD	FisherZ	0.001	0.9240	0.4015	0.8817	0.0574	109.6
	ChiSquare	0.1	0.9325	0.5603	0.8729	0.0529	100.5
	Multinomial LRT	0.01	0.9215	**0.8118**	0.9852	**0.2824**	71.1
	CG	0.03	**0.9957**	0.5250	**1.0000**	0.0050	**59**
DD	FisherZ	0.001	0.9189	0.3625	0.6250	0.0075	68
	ChiSquare	0.1	0.9787	0.8775	0.8333	0.0575	44.4
	Multinomial LRT	0.01	0.8938	**0.9575**	0.9537	**0.3050**	**39**
	CG	0.03	**0.9957**	0.5250	**1.0000**	0.0050	59
All	FisherZ	0.001	0.9528	0.5387	0.9759	0.1907	199.2
	ChiSquare	0.1	0.9410	0.7087	0.9210	0.1007	193.1
	Multinomial LRT	0.01	0.9341	**0.8880**	0.9833	**0.4107**	**124.6**
	CG	0.03	**0.9967**	0.6020	**0.9934**	0.1873	139.2

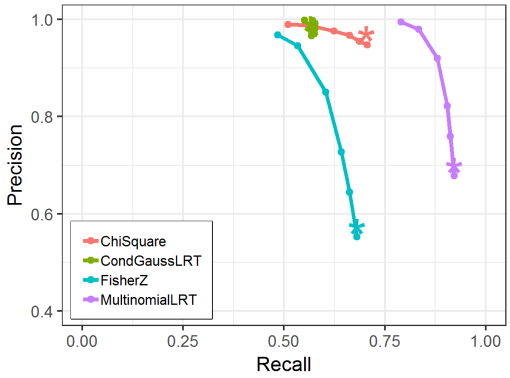


a Adjacency Recovery for 100 Variable, 500 Sample CG Simulated Data

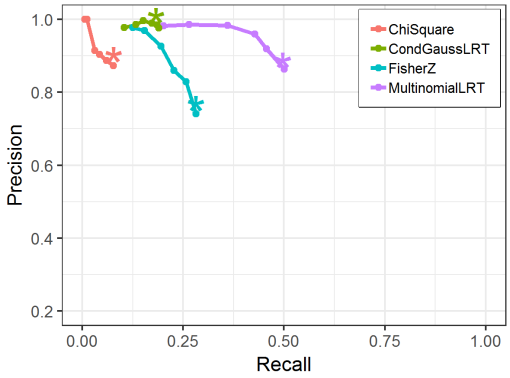


b Arrowhead Recovery for 100 Variable, 500 Sample CG Simulated Data

Figure 2: Precision-Recall Plots for 100 variable, 500 sample, mixed continuous and discrete datasets. The star refers to the average parameter values chosen by StARS.



a Adjacency Recovery for 100 Variable, 500 Sample LH Simulated Data



b Arrowhead Recovery for 100 Variable, 500 Sample LH Simulated Data

Figure 3: Precision-Recall Plots for 100 variable, 500 sample, mixed continuous and discrete datasets. The star refers to the average parameter values chosen by StARS.

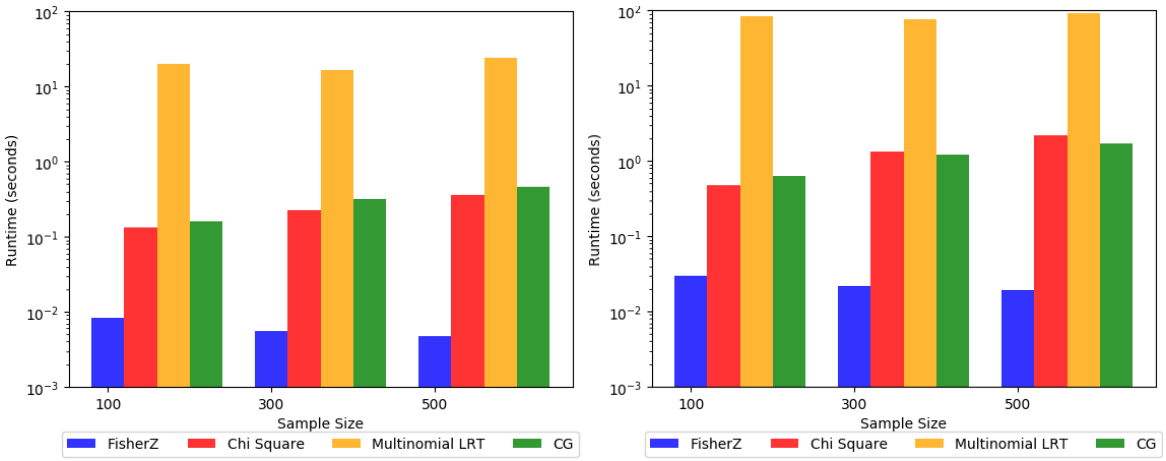
Table 3: Oracle parameter selection performance with algorithms suitable for mixed data types separated by edge type. All datasets contain 100 variables and 300 samples with average graph density 1.5 times the number of variables.

(a) Data simulated using the Lee and Hastie model.

Type	Algorithm	Parameter	AP	AR	AHP	AHR	SHD
CC	Copula-PC	0.01	0.913	0.906	**0.986**	0.399	27.950
	MGM-CPCStable	0.08	0.966	0.898	0.959	0.423	25.750
	FGES	1	0.924	**0.950**	0.837	0.625	**22.400**
	MGM-PCMax	0.03	**0.972**	0.889	0.423	0.344	41.450
	MGM-PCStable	0.08	0.966	0.898	0.479	**0.710**	39.850
CD	Copula-PC	0.001	0.938	0.431	0.949	0.054	122.850
	MGM-CPCStable	0.08	**0.990**	0.762	**0.967**	0.218	83.050
	FGES	1.5	0.974	**0.778**	0.716	0.379	**80.950**
	MGM-PCMax	0.08	**0.990**	0.762	0.421	0.284	105.250
	MGM-PCStable	0.08	**0.990**	0.762	0.486	**0.599**	101.600
DD	Copula-PC	0.001	0.926	0.234	0.250	0.001	65.200
	MGM-CPCStable	0.08	0.926	**0.529**	**0.947**	0.157	**52.750**
	FGES	1	**1.000**	0.490	0.616	0.141	55.100
	MGM-PCMax	0.01	0.954	0.526	0.323	0.141	60.100
	MGM-PCStable	1E-4	0.976	0.522	0.457	**0.336**	57.300
All	Copula-PC	0.001	0.954	0.484	**0.969**	0.103	217.200
	MGM-CPCStable	0.08	0.976	0.735	0.960	0.251	161.550
	FGES	1	0.963	**0.745**	0.747	0.377	**158.650**
	MGM-PCMax	0.08	0.976	0.735	0.408	0.266	207.900
	MGM-PCStable	0.03	**0.982**	0.727	0.481	**0.558**	199.950

(b) Data simulated using a Conditional Gaussian model.

Type	Algorithm	Alpha	AP	AR	AHP	AHR	SHD
CC	Copula-PC	1E-4	0.918	0.297	**1.000**	0.078	70.500
	MGM-CPCStable	0.03	**0.972**	0.374	0.975	0.104	65.350
	FGES	2	0.862	**0.498**	0.765	**0.314**	**58.850**
	MGM-PCMax	0.03	0.969	0.374	0.261	0.087	73.550
	MGM-PCStable	0.03	**0.972**	0.374	0.384	0.194	71.850
CD	Copula-PC	1E-4	0.859	0.159	**0.983**	0.029	164.400
	MGM-CPCStable	0.08	0.944	0.225	0.965	0.048	155.550
	FGES	1	0.825	**0.374**	0.719	**0.198**	**144.100**
	MGM-PCMax	0.01	**0.958**	0.222	0.319	0.049	164.300
	MGM-PCStable	0.01	0.956	0.222	0.432	0.122	161.300
DD	Copula-PC	1E-4	0.885	0.204	**1.000**	0.042	81.450
	MGM-CPCStable	0.05	**0.980**	0.296	0.940	0.064	75.050
	FGES	1	0.893	**0.408**	0.773	**0.212**	**68.500**
	MGM-PCMax	0.05	**0.980**	0.296	0.321	0.070	80.750
	MGM-PCStable	0.03	**0.980**	0.296	0.421	0.156	79.700
All	Copula-PC	1E-4	0.902	0.203	**0.994**	0.045	316.350
	MGM-CPCStable	0.08	0.960	0.279	0.952	0.069	296.000
	FGES	1	0.854	**0.410**	0.750	**0.227**	**271.600**
	MGM-PCMax	0.03	0.966	0.278	0.310	0.065	319.050
	MGM-PCStable	0.03	**0.967**	0.278	0.429	0.149	313.000



a Runtime for 50 Variable Mixed Data

b Runtime for 100 Variable Mixed Data

Figure 4: Algorithm runtime on mixed datasets with varying sample size. Each bar is averaged over all parameter values for the algorithm.

3.4 Evaluation of Structure Learning Algorithms from Mixed Data

Finally, we examine the ability of causal structure learning algorithms to estimate a causal graph from mixed data, by using the CausalMGM method as a preprocessing step. In particular, our goal is to compare the downstream constraint-based algorithms against one another and against the score-based FGES, and not to demonstrate the usefulness of CausalMGM, as other empirical evaluations have shown this (Sedgewick et al., 2017). For all algorithms, we use the Multinomial LRT as the conditional independence test based upon its performance in the prior section. To select parameters for CausalMGM, StEPS was used, as it is known to select accurate parameter values (Sedgewick et al., 2016). We were unable to test StARS to select downstream orientation algorithm parameters due to runtime constraints. For FGES, we use the recently described Conditional Gaussian score with the binomial structure prior. For a fair comparison, we simulate data using both the Lee and Hastie model which suits the assumptions of both algorithms, and the Conditional Gaussian model which only fits the assumptions of FGES.

First, we describe the performance of each algorithm on datasets with only 300 samples when its optimal parameter is chosen, split by each type of edge (Table 3). We focus on this high-dimensional setting for mixed data experiments because researchers are frequently posed with this problem when trying to generate hypotheses on real datasets. We first note that all algorithms perform well in terms of adjacency precision and recall (Table 3a). In general, MGM-CPC-Stable has excellent precision in both adjacencies and orientations regardless of which simulation method is used. In terms of recall, especially when datasets are mostly continouus, FGES is the best performing method, and can give high recall (0.858) with decent precision when its modeling assumptions are met (Conditional Gaussian data, Table 3b). The major difference between FGES and MGM-CPCStable is just a tradeoff between better recall (FGES) vs. better precision (MGM-CPCStable). For all results across parameter settings, we refer the reader to the supplementary material.

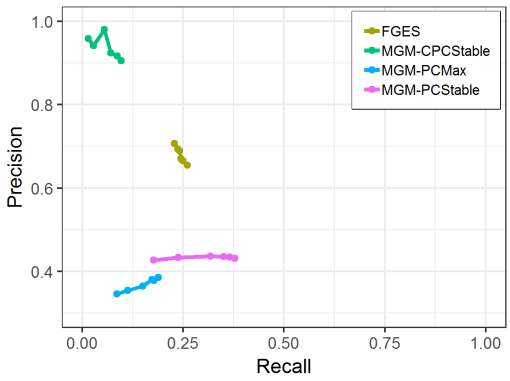
Next, we investigate how changing parameter settings (α for constraint-based searches and the binomial structure prior for FGES) affects the performance of algorithms across sample sizes (Figure 5). We first note that Copula-PC is unable to run in experimental settings where the number of variables is greater than or equal to the number of variables. In 500 sample size settings, Copula PC tends to be greatly affected by the choice of parameter, but with a conservative parameter choice, the algorithm gives good precision but low recall. When using CausalMGM as a first step, the constraint-based algorithms do not appear to be affected much by the choice of α with datasets of more than 100 samples, and FGES appears to be consistent across parameter choice in nearly all experimental settings. It appears that MGM-PC-Stable outperforms MGM-PC-Max in orientation recovery, and since these methods have nearly identical adjacencies in all cases, the results suggest that it is never preferable to run MGM-PC-Max. In addition, with Conditional Gaussian simulated datasets, FGES outperforms both of the aforementioned methods regardless of sample size, whereas with Lee and Hastie assumptions, FGES tends to have worse recall with improved precision when compared to MGM-PC-Stable. Finally, CPC-Stable maintains high precision in all instances with low recall in all cases except for low-dimensional Lee and Hastie data. This suggests that MGM-CPC-Stable could be used as a first pass to gather reliable orientation information before trying another method if more predictions are desired (e.g. FGES). We have repeated all experiments with denser graphs (number of edges is on average 2.5 times number of variables),and found that all results remain the same except adjacency recall is lower for all methods (Supplementary Material). For an investigation of mixed data methods with and without CausalMGM we refer the reader to (Sedgewick et al., 2017). Finally, we note that in terms of runtime efficiency all algorithms complete in less than a second on 100 variable size data.

4. Discussion

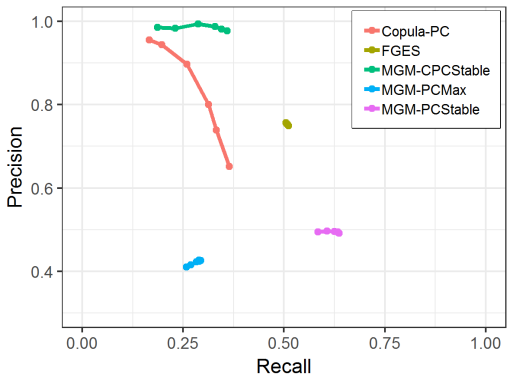
In this work, we have presented a thorough empirical evaluation of algorithms for causal structure learning on continuous and mixed datasets. We specifically look at high dimensional settings (low sample size), because this is the nature of most biomedical problems. Our results indicate that even in these settings, constraint-based causal discovery algorithms are more precise in edge predictions at the expense of recall which could be attributed to the repeated independence tests that are performed on particular edge. FGES can achieve superior causal orientation performance on continuous data but the selection of the penalty parameter is very important to its recall. For continuous datasets, we find that using StARS to do parameter selection gives reasonable results at the 95% stability threshold.

On mixed datasets, we find that using tests specifically tailored to mixed data gives a significant boost in recall, especially when the data has enough statistical power to detect associations among categorical variables. We further find that the Multinomial LRT test gives the best recall for equal precision; however, it is much more computationally demanding than treating the data as all continuous or discrete or using a CG test. We further find that parameter selection can have a large impact on the results, and StARS with a 95% threshold does not give a good tradeoff between precision and recall on mixed data. Therefore, it remains an open question how to set this parameter to achieve good accuracy.

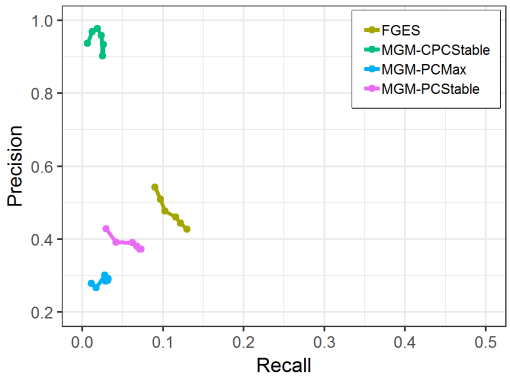
Figure 5: Arrowhead recovery for mixed data algorithms on datasets with different simulation parameters. All graphs had 100 variables and average density 1.5 times the number of variables. The first row depicts Lee and Hastie simulated data, while the second row depicts Conditional Gaussian. The first column is 100 sample size, and the second column is increased to 500 samples.



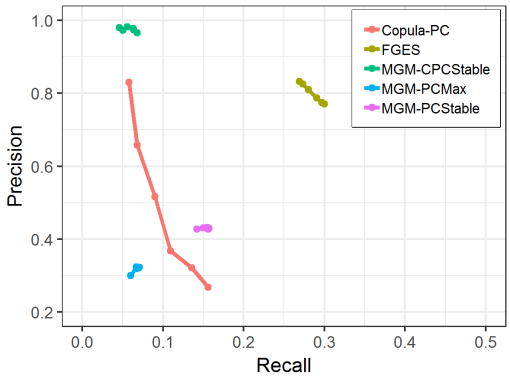
a Arrowhead recovery from mixed 100 variable 100 sample Lee and Hastie data



b Arrowhead recovery from mixed 100 variable 500 sample Lee and Hastie data



c Arrowhead recovery from mixed 100 variable 100 sample Conditional Gaussian data



d Arrowhead recovery from mixed 100 variable 500 sample Conditional Gaussian data

Finally, as expected, we find that when the sample size is limited, both edge discovery and edge orientation suffer. We find that when using CausalMGM, PC-Stable has a significant advantage over using PC-Max in terms of recall, and CPC maintains high precision in all cases. The choice of parameter for all of these algorithms has a large impact in the low sample size setting (100 variables, 100 samples). FGES achieves the best recall in both adjacencies and orientations; however, this benefit is only marginal in low sample size cases. Depending upon the use case, the nearly perfect precision of MGM-CPC-Stable may be preferable to using FGES unless more predictions are required (in tasks like hypothesis generation for cheaper experiments).

For future work, we intend to explore the effect of latent confounding, for which we have done some preliminary analysis (Raghu et al., 2018). Our major experimental question is whether using an algorithm such as Fast Causal Inference (FCI) which gives theoretical correctness guarantees in the presence of confounding is always better than a simpler constraint-based algorithm that does not account for latent variables. In addition, we hope to include stability selection approaches like Bootstrapping or Stability Selection to see whether these can alleviate the problem of parameter selection, as StARS does not appear to have great benefit on mixed datasets.

Acknowledgments

We would like to thank the creators of the Tetrad VI software, as the methods evaluated in this study were implemented in this package. This work was supported by NIH grants: T32CA082084 (VKR) and R01LM012087, U01HL137159 (PVB).

References

Bryan Andrews, Joseph Ramsey, and Greg Cooper. Scoring bayesian networks of mixed variables. In *Proceedings of the 2017 ACM SIGKDD Workshop on Causal Discovery*, 2017.

Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1435–1444. ACM, 2016.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

Ruifei Cui, Perry Groot, and Tom Heskes. Copula pc algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer, 2016.

Frederick Eberhardt Hyttinen, Antti and Matti Jrvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Jd Lee and Tj Hastie. Learning the Structure of Mixed Graphical Models. *Stanford.Edu*, pages 1–32, 2013. ISSN 1061-8600. doi: 10.1080/10618600.2014.900500.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009. ISSN 1935-7516. doi: 10.1214/09-SS057.

Vineet K. Raghu, Joseph D. Ramsey, Alison Morris, Dimitrios V. Manatakis, Peter Sprites, Panos K. Chrysanthis, Clark Glymour, and Panayiotis V. Benos. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, Feb 2018.

Joseph Ramsey. Improving Accuracy and Scalability of the PC Algorithm by Maximizing P-Value. *arXiv*, pages 1–11, 2016. URL `https://arxiv.org/abs/1610.00378`.

Joseph Ramsey, Jiji Zhang, and Peter L. Spirtes. Adjacency-Faithfulness and Conservative Causal Inference. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 401–408, 2006.

Joseph D. Ramsey. Scaling up greedy equivalence search for continuous variables. *arXiv*, 2015. URL `http://arxiv.org/abs/1507.07749`.

Andrew J. Sedgewick, Ivy Shi, Rory M. Donovan, and Panayiotis V. Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, 17(S5):175, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1039-0.

Andrew J. Sedgewick, Joseph D. Ramsey, Peter Spirtes, Clark Glymour, and Panayiotis V. Benos. Mixed graphical models for causal analysis of multi-modal variables. *CoRR*, abs/1704.02621, 2017. URL `http://arxiv.org/abs/1704.02621`.

Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. Comparative benchmarking of causal discovery techniques. *CoRR*, abs/1708.06246, 2017.

Elena Sokolova, Perry Groot, Tom Claassen, and Tom Heskes. LNAI 8754 - Causal Discovery from Databases with Discrete and Continuous Variables. pages 442–457, 2014.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Michail Tsagris, Giorgos Borboudakis, Vincenzo Lagani, and Ioannis Tsamardinos. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, pages 1–12, 2018.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.