Causal discovery using clusters from observational data

Sepideh Pashami [1] **Anders Holst** [2] **Juhee Bae** [3] **Sławomir Nowaczyk** [1]

Abstract

Many methods have been proposed over the years for distinguishing causes from effects using observational data only, and new ones are continuously being developed – deducing causal relationships is difficult enough that we do not hope to ever get the perfect one. Instead, we progress by creating powerful heuristics, capable of capturing more and more of the hints that are present in real data.

One type of such hints, quite surprisingly rarely explicitly addressed by existing methods, is inhomogeneities in the data. Clusters are a very typical occurrence that should be taken into account, and exploited, in the process of identifying causes and effects. In this paper, we discuss the potential benefits, and explore the hints that clusters in the data can provide for causal discovery. We propose a new method, and show, using both artificial and real data, that accounting for clusters in the data leads to more accurate learning of causal structures.

1. Introduction

In general it is not possible to reconstruct the causal structures purely from the observational data, however, in many situations there are important hints available that can be used to learn it in an approximate fashion. To this end, a number of methods have been proposed for distinguishing causes from effects. What hints are available varies from case to case, and thus we do not expect a single approach to work perfectly, nor to be the best choice in all contexts. Instead, our research community acknowledges that different methods are needed for different situations.

An important direction of scientific progress in the area is to identify heuristics that can be used to discover causal relationships under realistic conditions. The idea of explicitly considering cluster data as latent variables previously proposed by Sgouritsa et. al (Sgouritsa et al., 2013). They used a kernel method to infer a common cause of a sets of variables by clustering the variables. However, there is no hint for selecting these variables efficiently which makes their method computationally expensive. There is also no grantee that they can find confounders that makes a set of variables jointly conditionally independent. In this paper we show that inhomogeneities in the data, appearing due to a confounding factor, can hide existing causal relationship between two variables. We explore the hints based on structure of discovered causal network that can reduce ambiguity in the causal structure. We also propose a new method that, as shown by experiments performed on synthetic and real data, is able to improve the quality of learned causal structures.

Correlation among variables can manifest in many different ways. It is often implicitly assumed that the relationship between two continuous-valued variables will take some functional form, such that the value of one variable can be expressed as a (possibly approximate or noisy) function of the other. For example, the simulated data in (Mooij et al., 2016) follows this assumption. When such a relationship is linear, or at least monotonous, correlation between the variables can be detected by looking at Pearson correlation coefficient, or other related measures. However, existence of clusters in the data, while a very common occurrence, leads to a very different manifestation of the correlation between variables. We claim that clusters should be addressed explicitly.

Figure 1 presents an example of this concept. Plots (a) and (b) show functional relationships, the first linear with Pearson correlation coefficient of $+0.9$ and the second nonlinear, with Pearson correlation coefficient of $+0.5$. Both of those will lead to correct causal discovery results. However, the correlation coefficient is not suitable to detect cases with clusters, such as Figure 1(c). In this case the linear part of the correlation can be small, or zero. Even more seriously, the "global" correlation may have quite different direction than the relationship within individual clusters of the data, as hinted in this toy example. The Pearson correlation coefficient for data in Figure 1(c) is $+0.1$, while the two variables rather appear to have a clear negative relation.

[1]Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden [2]RISE SICS, Sweden [3]School of Informatics, University of Skövde, Sweden. Correspondence to: Sepideh Pashami <sepideh.pashami@hh.se>.

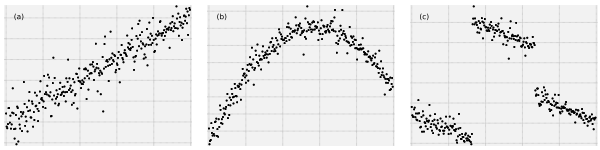


Figure 1. Different kinds of data: (a) linear correlation (b) non-linear functional correlation (c) correlation with clusters.

When there are clear clusters in data, then, they need to be taken care of before continuing with the causal analysis. Otherwise, both the sign and the magnitude of any real effects between the variables are likely to be greatly misjudged. Often such clusters represent some discrete physical states which may affect many of the other variables. In such cases, it makes sense to continue the analysis by considering correlations and causal effects within each cluster, rather than on the total data set. It means that all the variables that are affected by such a state should be conditioned on the value of the discrete state.

Conversely, finding inhomogeneities in the data that cannot be explained by the available attributes is an indicator of possible existence of a confounding factor. It means that if clusters exist in the data as a whole, or in a subset of features, there might be some latent variables involved that cause this clustered structure. In this study, we aim to detect the existence of a confounding factor as a common cause of two or more variables. We represent the confounding factor as a latent variable, with values corresponding to the clusters which express the unexplained structure, or "lumpiness", of the data.

The intuition behind our idea lies in the notion that "lumpiness in the data must have a reason". This is similar to the principle used in the Information Geometric Causal Inference (IGCI) class of methods (Janzing et al., 2012), where entropy is used to estimate whether an attribute is likely a cause or an effect. In our case, it corresponds to noticing that a variable with varying density has lower entropy than a uniform variable of the same variance. However, IGCI only considers pairs of variables in the calculations, while our idea is to explicitly consider clusters among larger sets of attributes.

Our contribution is to extend existing causal inference algorithms with a cluster-based conflict resolution mechanism. When a causal graph initially produced by the algorithm contains ambiguities, in the form of mutually correlated variables for which it is unclear in what direction the causality goes, we check whether their joint distribution consists of several clusters. If so, a clustering algorithm is used to identify these clusters, and a discrete latent variable is added. This new variable uses cluster identities as values, and then the original causal inference algorithm is run again, now capable of finding the causal structure undisturbed by the inhomogeneities. Furthermore, we can assume that the edges connecting the latent variable with the variables whose clustered structure it is supposed to explain have a causal direction towards the latter. Thus, the existence of these clusters provide clues that are propagated and improve the entire causal graph.

This paper is organized as follows. In the next section we relate our contribution to the existing results in the area. Section 3 explains the proposed methods, followed by Section 4, with experimental evaluation using both artificial and real-world data. Finally, conclusions and discussion of future directions are presented in Section 5.

2. Related work

Causal inference is challenging for many reasons: missing values, sampling error, measurement error, complicated probability distributions and enormous search space, to name just a few. One especially problematic issue is related to discovery of hidden (latent) variables. Performing the correct causal inference with unmeasured confounders is challenging since they are, by definition, not found in the data itself.

At the same time, in real world applications, it is almost always impossible to measure all of the relevant variables, and therefore a lot of research has been devoted to this issue. The fast causal inference (FCI) and the really fast causal inference (RFCI) algorithms (Spirtes et al., 2000; Zhang, 2008) are the variations of the PC algorithm that allow latent variables. These methods perform additional conditional independences tests compared to the PC algorithm. FCI and RFCI can find a latent variable affecting two variables, but they do not allow for detecting latent variables that affect more than two variables. Meganck et al. (Meganck et al., 2007) propose graphical modeling techniques allowing for latent variables, transforming a complete partial ancestral graph (CPAG) (Zhang, 2008) into a semi-Markovian causal model (SMCM) (Pearl, 2000). Each latent variable is represented with a bi-directed edge in the SMCM graph. Hoyer et al. (Hoyer et al., 2008) shows how to estimate linear causal models when hidden variables exist in the linear non-Gaussian acyclic model (LiNGAM). It uses independent component analysis (ICA) to find the latent variables. However, it is limited in determining the direction of the influence if the distribution is symmetric. Buchanan et al. (Buchanan et al., 2010) propose a generative model of causation called causal edge replacement process, where causal knowledge representations are constructed by a particular set of rules to create causal graphs. They claim that the edge replacement better explains causal representations and the reasoning than minimality, and take advantage of stream location effect in

which causal relationships can differ based on the moment when interventions occur during the causal stream.

An important direction is graphical causal modeling, aiming to explain and better understand the causal inference. Spirtes et al. (Spirtes et al., 2000) give an introduction to graphical causal modeling and describes the differences between causal inference and ordinary machine learning classification and prediction problems. Yet, the possible presence of unmeasured causes and latent variables still remain challenging, especially with a large number of variables and smaller sample sizes (Meganck et al., 2007).

More generally, learning of causal structures is based on one of the three broad approaches: constraint-based, score-based, or Bayesian model averaging.

Constraint-based structure learning is the closest to the method proposed in this paper, being based on tests for conditional independences in the data. Generally, the result is an equivalence class of structures, typically represented as a partially directed acyclic graph (PDAG). Most approaches are extensions of the PC method (Spirtes et al., 2000). For example, to address order sensitivity, Colombo et al. (Colombo & Maathuis, 2014) proposed a modification of the PC algorithm which removes the order-dependence limitation to a large extent. Regarding Markov equivalence classes, Ali et al. (Ali et al., 2009) state and prove conditions under which two maximal ancestral graphs are Markov equivalent, leading to an algorithm for determining Markov equivalence that runs in polynomial time.

Alternatively, causal structures can be learned using score-based methods (Triantafillou & Tsamardinos, 2016). A common approach is to define an optimization problem for finding the structure optimizing a given score, typically using heuristic search. The search procedure picks the best scoring from the two general scoring functions such as likelihood and Bayesian scoring functions until it converges to a local optimum. Finally, Bayesian model averaging approaches (Hoeting et al., 1999) can generate an ensemble of possible structures, including explicit model uncertainty, to address over-fitting issues.

Discovery of causal relationships from purely observational data have also been investigated for pairs of variables, focusing on distinguishing between cause and effect. Stegle et al. (Stegle et al., 2010) propose a probabilistic latent variable model (GPI:Gaussian Process Inference) to distinguish between cause and effect using standard Bayesian model selection. They treat the "noise" as a latent variable that summarizes the influence of all other unobserved causes of the effect. In general, two main methods for distinguishing between cause and effect are Additive Noise Methods (ANM) and Information Geometric Causal Inference (IGCI) methods (Mooij et al., 2016). However, in the real world, it is common that we see bi-directions, clusters, and cases with confounding variables, not only the direct causal relations that these methods are designed for.

3. Method

The proposed method starts by using an existing algorithm for finding an initial causal structure. For this, we have here used the PC algorithm (Spirtes et al., 2000; Kalisch et al., 2012). The PC algorithm builds a Markov equivalence class which contains the underlying causal graph and represents it by a Completed Partially Directed Acyclic Graph (CPDAG). The calculations are based on conditional independence tests. In many cases, the graph produced contains bi-directed edges. Such bi-directed edges are undesirable, as they imply the existence of a confounding variable, i.e., an unknown factor which influences both of the signals (nodes). Similarly, fully connected nodes (cliques) within a causal network are undesirable due to the ambiguity of identifying cause and effect.

The next step of the method is to look for clustered data in the subset of variables constituting such cliques of ambiguous causal direction. Note that it is not critical that it is a true clique, we just want to select a large enough set of variables that makes any clusters easily distinguishable, but not too many irrelevant variables not involved in the same clusters that confuses the picture. Therefore, a set of tightly correlated variables regarding which we can suspect a latent variable because the causal directions are ambiguous is our best bet. In this way we relax the sufficiency assumption by adding a latent variable that can explain the clusters in the data.

The clustering method used here is a Gaussian Mixture Model (GMM) (Reynolds, 2008) trained by Expectation Maximization (EM). This is combined with a test for "homogeneity", in this case, defined as being similar to a (multivariate) Gaussian distribution, since it is the distribution (in Euclidean space) which has the highest entropy for a given variance. However, most tests for Gaussian distributions is based on checking moments of the distribution. Here, we are not interested in the higher moments (actually we will happily accept both skew and fat-tailed distributions as homogeneous), but rather in the local density of the distribution. The approach taken here is therefore to model the distribution of pairwise distances between samples in the data and compare this to the expected distribution of distances from samples from a multivariate Gaussian with the same covariance matrix as the actual data. If the local density is much higher than expected, this indicates a clear clustered structure.

The homogeneity measure is used both to check whether there are any clusters for a clique of variables and to determine the number of clusters required. Training of the GMM is started with a small number of clusters (two) and the resulting components are then tested for homogeneity to see whether the number of clusters should be increased. The model continues to be retrained until all components are considered sufficiently homogeneous.

In the final step, the PC algorithm is run again, this time with the added new variable, and values corresponding to cluster indices. If successful, the added variables will resolve the cliques of ambiguous causal direction. Furthermore, when determining the causal direction in the last phase of the PC algorithm, we can use the assumption that there is probably a cause behind the clustered structure in the data, to deduce that the causal impact is from the clustered index to the variables showing traces of this clustering. This may resolve more ambiguities, also further away from the clustered variables.

The above steps are then repeated until no more unexplained clusters in the data are left. In this way, the existence of clusters in data will help resolve the causal directions. We now turn to the evaluation of the proposed method.

4. Empirical evaluation

The empirical evaluation of adding a latent variable when there are clear clusters in the data set is illustrated using artificial data where the underlying causal network is known.

Moreover, the proposed method is tested on the real-world data set collected by a fleet of six city buses. In particular, the aim is to identify the causal relation between the set of signals influencing fuel consumption.

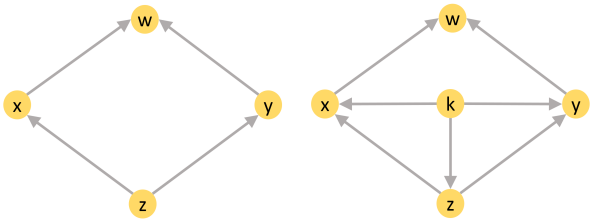
4.1. Artificial data

We use the synthetic data to show that adding the latent variable corresponding to underlying clusters in the data improves the accuracy of the discovered causal structure.

4.1.1. Data generation

The artificial data is generated by a linear model based on the dependency between the variables expressed as edges in a given underlying causal graph.

The underlying data is an exhaustive set of all the non-isomorphic DAGs with 5 vertices. To generate all these graphs, first, all non-isomorphic DAG graph with four vertices are generated. Then, a cluster vertex is connected to each subset of the four vertices with a directed edge. There are 30 distinct non-isomorphic DAGs with four vertices and there are 15 different ways of connecting the cluster vertex to the other vertices. In total, 450 different graphs are generated. The vertices in each graph are arranged in a causal order in a way that earlier variables are causes of the later variables.



(a) without k (b) with k

Figure 2. Examples of underlying causal graphs with and without the cluster vertex (k).

The set of 5 variables are selected corresponding to each vertex. The value of each vertex is a linear combination of the values of its parents with some random noise, i.e., the value assigned to each variable (x_i) is a linear combination of the values of already assigned variables ($x_1, ..., x_{i-1}$) plus a Gaussian error term (e_i). The error terms are Gaussian with mean zero and standard deviation between 0 and 1 in such a way that they are independent of each other. The parameters of the linear model are selected randomly in the range of 1 and 2. The root variables are either a Gaussian random term or a cluster variable. The equation for the cluster variable is given in Equation 1.

Figure 2 presents two example DAGs with and without cluster vertex k. To generate data for these examples the following equations has been used (with $k = 0$ when no clusters are used):

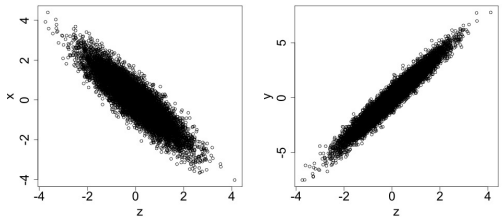
$$k_j = \begin{cases} 2 & \text{for } j \in [1, n/4), \\ 1 & \text{for } j \in [n/4, n/2), \\ -1 & \text{for } j \in [n/2, 3 * n/4), \\ -2 & \text{for } j \in [3 * n/4, n]. \end{cases} \quad (1)$$

$z = k + e_z \qquad e_z \sim \mathcal{N}(0,\, 1)$
$x = -1 * z + -10 * k + e_x \qquad e_x \sim \mathcal{N}(0,\, 0.5)$
$y = 2 * z + 10 * k + e_y \qquad e_y \sim \mathcal{N}(0\, 0.5)$
$w = 3 * x + 5 * y + e_w \qquad e_w \sim \mathcal{N}(0\, 0.5)$

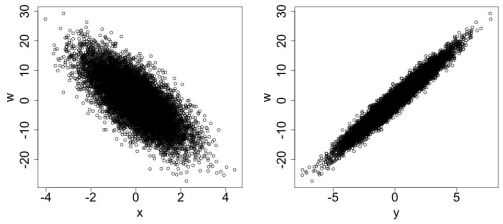
The relation between variables generated from the above equation have been shown for the causal graph with and without the cluster variable in the Figures 3 and 4, correspondingly.

4.1.2. Example result

Figure 5 shows the result of the PC algorithm for the example data shown in Figure 3, i.e., generated from the graph shown in Figure 2a. In this example, the data is homogeneous and there are no clusters in the data. In this case,

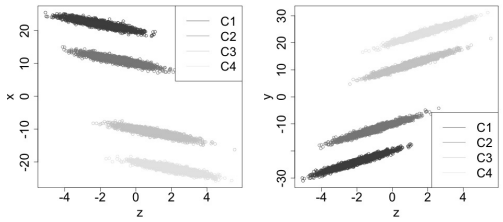


(a) z-x plot (b) z-y plot

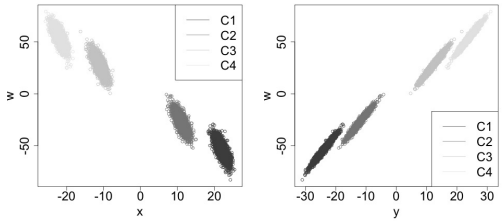


(c) x-w plot (d) y-w plot

Figure 3. Plots showing the relation of the variables for the causal graph without the cluster variable, as shown in Figure 2a.

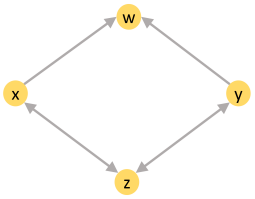


(a) z-x plot (b) z-y plot



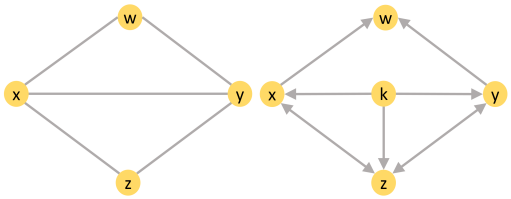
(c) x-w plot (d) y-w plot

Figure 4. Plots showing the relation of the variables for the causal graph with the cluster variable (k), as shown in Figure 2b.



(a) result of PC

Figure 5. The causal graph without the latent variable for the simulated data.



(a) excluding k (b) including k

Figure 6. The result of the PC algorithm for the example shown in 2b.

the PC algorithm is able to find Markov equivalence of the underlying graph.

If a latent cluster variable is added to the data (Figure 2b), finding the Markov equivalence of the underlying graph is not possible for the PC algorithm. Figure 6a shows the result where the cluster variable is not visible to the algorithm. In this case, the PC algorithm is not able to find any directions, nor the existing v-structure. On the other hand, making the cluster variable available will reveal the desired v-structure (Figure 6b). This exemplifies the importance of identifying and adding such latent variables in the analysis.

4.1.3. Results of M-vertex graphs

The focus of section is to investigate how including or excluding the cluster variable affects the causal discovery in graphs with five vertices. The evaluation is performed for all the non-isomorphic DAGs with 5 vertices, where one of them is the cluster vertex.

Structural Hamming distance (SHD) (Tsamardinos et al., 2006) has previously been used as a performance measure to compare the structure of the result of the PC algorithm and the underlying causal graph. SHD is defined as a number of additions or deletions of an undirected edge and additions, deletions or reversing the orientation of a directed edge.

In order to understand whether adding a cluster variable improves the result or not, we calculate the difference between SHD of the result of the PC algorithm and the underlying causal graph in two cases: including and excluding

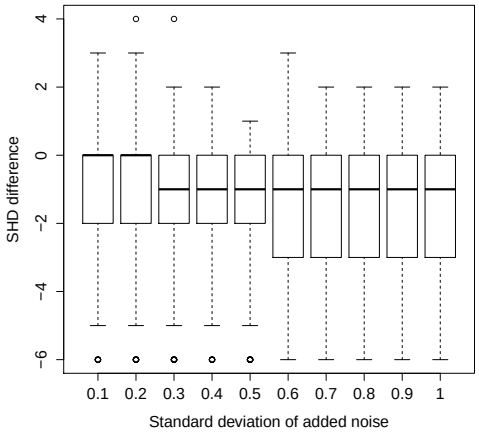


Figure 7. The difference between the Structural Hamming Distance (SHD) calculated between the result of PC algorithm and the underlying causal graph in the presence and absence of the cluster variable for different noise levels. The negative values indicate that including the cluster variable improved causal discovery.

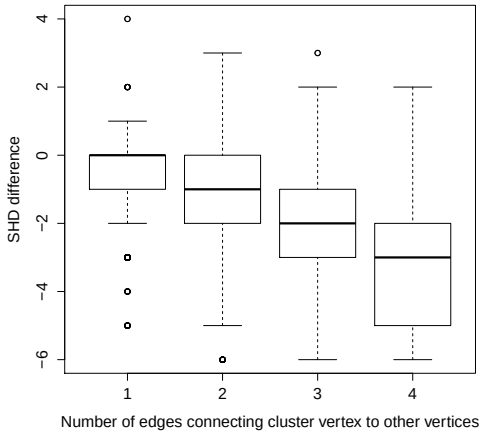


Figure 8. The difference between the Structural Hamming Distance (SHD) calculated between the result of PC algorithm and the underlying causal graph in the presence and absence of the cluster variable compared to the number of edges connecting the cluster vertex to other vertices.

the cluster variable. The score is calculated by subtracting SHD in the case of excluding the cluster variable from SHD in the case of including it. Since the number of vertices in these cases are not equal, comparing SHDs is not fair. Therefore, we don't consider the cluster vertex and its connection to the rest of the vertices in the calculation of the SHD when the latent variable is included in the analysis, to make the SHDs comparable. Figure 7 shows the SHD differences for two mentioned cases when the standard additive noise varies between 0.1 and 1. In this plot, negative numbers show that the SHD of the result of the PC algorithm with the underlying causal graph including cluster variable is smaller than the case of excluding it. The negative values of the mean, first and third quartiles in the SHD differences, in Figure 7, show that for a big portion of the generated graphs adding a cluster variable improves the discovery of causal relations independent of noise level.

Positive SHD differences mainly happens when the PC algorithm fails to find the correct connections between the cluster variable and the rest of the variables.

Note that adding a cluster variable becomes more important when the new vertex is connected to more than one variable. Figure 8 shows how SHD difference explained previously changes with respect to the number of edges connecting the cluster vertex to other vertices. As it is clear from the Figure 8, adding a latent variable affects the result more when the cluster variable is connected to more vertices.

In many cases, the PC algorithm doesn't automatically find the correct directions of the cluster variable to other variables correctly. Figure 9 shows the SHD calculated between the result of PC and the underlying causal graph only for edges connected to/from cluster vertex compare to the number of edges connecting cluster vertex to other vertices in the underlying graph. As shown in Figure 9, the absolute value of the mean SHD is equal to the number of edges connecting cluster vertex to other vertices in the underlying graph. This can be explained by the fact that PC algorithm correctly finds the edges between the cluster vertex and rest of the vertices, however, it often incorrectly marks it as bi-directed edge or gives it a reverse direction.

4.2. Fleet of city buses

One application domain on which this approach has been evaluated is to identify the causal relations between signals measured on-board heavy duty vehicles. A causal network is useful to provide the overall picture of how various parameters are affecting the vehicles performance.

The attributes in the fleet of city buses, that we used, are AcceleratorPedalPos, AmbientAirTemperature, BrakePedalPos, EngineCoolantTemperature, EngineSpeed, FuelRate, RelativeSpeedFrontLeft, RelativeSpeedFrontRight, SelectedGear, SteeringWheelAngle, VehicleSpeed.

In this dataset, there is one discrete variable included which appears to cause clear clusters in the data, and that is the selected gear. Figure 10 shows the relation between the engine speed and the vehicle speed. In this example, selected

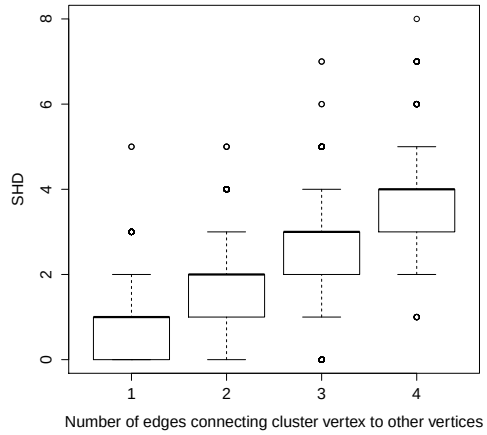


Figure 9. It shows the Structural Hamming Distance (SHD) calculated between the result of PC and the underlying causal graph connected to/from cluster vertex compare to the number of edges connecting the cluster vertex to other vertices in the underlying graph.

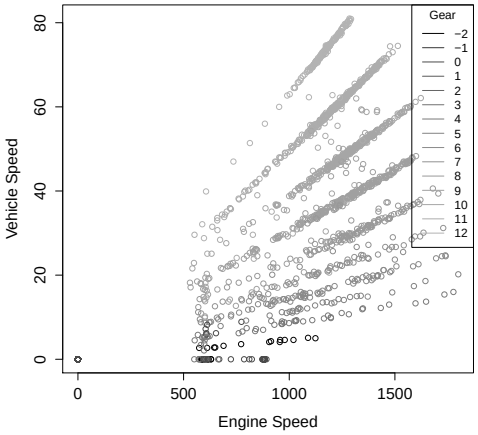


Figure 10. The clusters in the engine vs. vehicle speed plot can be explained by the selected gear.

gear causes the grouping into clusters of different slopes in the figure. Obviously, these kinds of clusters must be taken into account when trying to distinguish causes from effects in the data. There is actually another clustering vaguely indicated in the figure: A vertical streak of higher density for all gears around of engine speed of 600 rpm. Figure 11 shows a histogram of the engine speed, in which this is clearly visible as a peak to the left of the main mass. This corresponds to an engine that is idling. There is, on close inspection, actually two peaks, due to somewhat different idling speed in the individual buses. The small peak around 0 corresponds to engines that have not yet started. There is no variable in the data corresponding to idling, so this is one candidate for a latent variable. It is very likely that the causal effects in the engine differ between idling and normal operation.

We have added two latent variables to the data to account for clustered structures, “LatentVar1” and “LatentVar2”. The first added cluster variable (“LatentVar1”) represents “idle run” (engine speed below or around 620 rpm), and the second (“LatentVar2”) indicates whether the bus is an old or new generation (for example, the “Steering Wheel Angle” signal is not measured in the old generation).

We then run the PC algorithm, first without the explained two latent variables, then with them. The causal relation of the 11 measured attributes as a result of the PC algorithm is shown in Figure 12. Note that the number of bi-directed edges (ambiguous causal direction) differs in the resulting graphs: there are 5 fewer bi-directed edges when the latent variables are added. This shows that adding

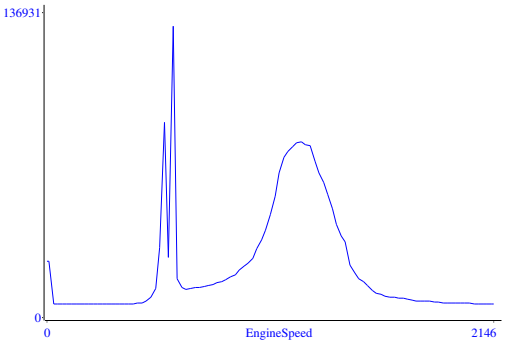


Figure 11. Histogram over engine speeds, showing a clustered structure.

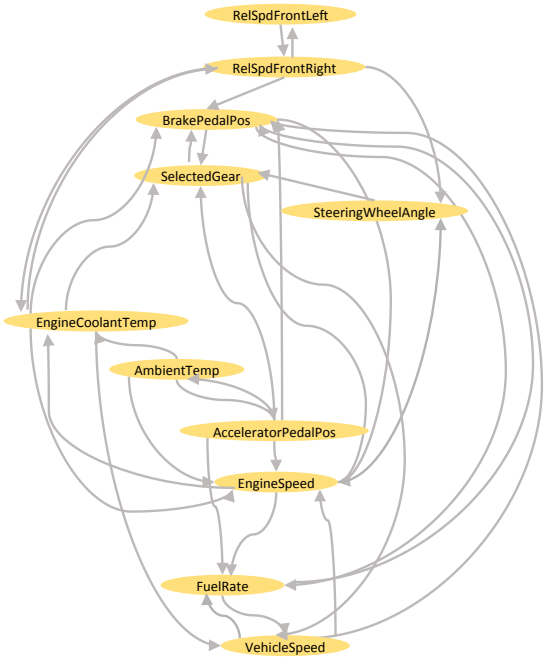
these variables helps in identifying causal directions between the other variables. Moreover, some of the remaining bi-direction edges in the graph with latent variable (Figure 12b) make sense. One example is the bi-direction edge between relative speed of the left and right front wheels ("RelSpdFrontLeft" and "RelSpdFrontRight" attributes) which are also very tightly connected to each other but none of them can be said to be the cause of the other. On the other hand, the bi-directional edges which are disambiguated include, for example, the connection between "AmbientTemperature" and "AcceleratorPedalPosition".

5. Discussion and Conclusion

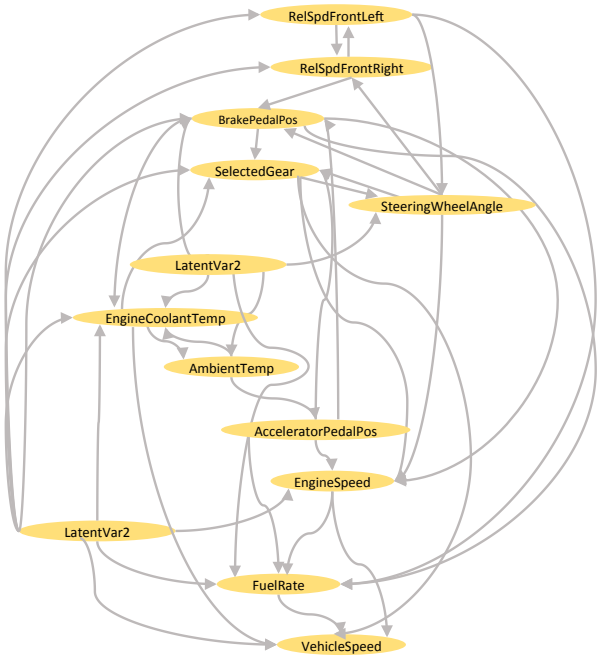
We propose and discuss the idea that clusters in the observational data should be used as a clue for causal discovery. There are many methods that attempt to distinguish the cause and effect, and we have shown that clusters are a potential catalyst that can disambiguate the causal directions. We demonstrate that adding latent variables, corresponding to clusters in the data, helps to learn the causal structure more accurately. Our results, both on synthetic and real-world data, show clear promise of the proposed method.

The underlying idea is that inhomogeneities in the data are unlikely to appear spontaneously; instead, they probably have a cause. If none of the variables existing in the data can explain the cluster structure, it is an indication that a latent variable may be behind it. However, it should be noted that this is only a heuristic. There are many counterexamples, where inhomogeneities in the data can be created from continuous processes. Consider, as case in point, a sinus wave sampled at random points in time. There will be an accumulation of points close to the upper and the lower extremes, and lower density in the middle. Nevertheless, it would be wrong to assume a necessary binary value as a cause behind these two resulting "clusters".

Still, this is the same for most methods based on, for example, entropy differences; that it is only an indication and not the absolute truth. Deducing causal direction from data is hard enough that we must use all available hints provided, and actively look for more ideas. Hints related to cluster structure have, to the best of our knowledge, not been previously addressed. Clusters are abundant in real world data, which makes this an important addition to the causal discovery tool box.



(a) without latent variables



(b) with latent variables

Figure 12. Results of the proposed algorithm for the bus fleet data.

References

Ali, R. Ayesha, Richardson, Thomas S., and Spirtes, Peter. Markov equivalence for ancestral graphs. *Ann. Statist.*, 37(5B):2808–2837, 10 2009. doi: 10.1214/08-AOS626. URL http://dx.doi.org/10.

1214/08-AOS626.

Buchanan, D., Tenenbaum, J., and Sobel, D. Edge replacement and nonindependence in causation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32, 2010.

Colombo, Diego and Maathuis, Marloes H. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, January 2014. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2627435.2750365.

Hoeting, Jennifer A., Madigan, David, Raftery, Adrian E., and Volinsky, Chris T. Bayesian model averaging: A tutorial. *STATISTICAL SCIENCE*, 14(4):382–417, 1999.

Hoyer, P.O., Shimizu, S., and Kerminen, A.J. Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

Janzing, Dominik, Mooij, Joris, Zhang, Kun, Lemeire, Jan, Zscheischler, Jakob, Daniuis, Povilas, Steudel, Bastian, and Schlkopf, Bernhard. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1 – 31, 2012. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2012.01.002. URL http://www.sciencedirect.com/science/article/pii/S0004370212000045.

Kalisch, Markus, Mchler, Martin, Colombo, Diego, Maathuis, Marloes, and Bhlmann, Peter. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software, Articles*, 47(11):1–26, 2012. ISSN 1548-7660. doi: 10.18637/jss.v047.i11. URL https://www.jstatsoft.org/v047/i11.

Meganck, Stijn, Leray, Philippe, and Manderick, Bernard. *Causal Graphical Models with Latent Variables: Learning and Inference*, pp. 5–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-75256-1. doi: 10.1007/978-3-540-75256-1_4.

Mooij, Joris M., Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob, and Schölkopf, Bernhard. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17(1):1103–1204, January 2016. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2946645.2946677.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

Reynolds, D. A. Gaussian mixture models. *Encyclopedia of Biometric Recognition*, 2008. URL https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf.

Sgouritsa, Eleni, Janzing, Dominik, Peters, Jonas, and Schölkopf, Bernhard. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. *arXiv preprint arXiv:1309.6860*, 2013.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Stegle, Oliver, Janzing, Dominik, Zhang, Kun, Mooij, Joris M, and Schölkopf, Prof. Bernhard. Probabilistic latent variable models for distinguishing between cause and effect. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 1687–1695. Curran Associates, Inc., 2010.

Triantafillou, Sofia and Tsamardinos, Ioannis. Score-based vs constraint-based causal learning in the presence of confounders. In *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), Jersey City, USA, June 29, 2016.*, pp. 59–67, 2016.

Tsamardinos, Ioannis, Brown, Laura E., and Aliferis, Constantin F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006. ISSN 1573-0565. doi: 10.1007/s10994-006-6889-7. URL http://dx.doi.org/10.1007/s10994-006-6889-7.

Zhang, Jiji. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873 – 1896, 2008. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2008.08.001.