Learning Causal AMP Chain Graphs

Jose M. Peña JOSE.M.PENA@LIU.SE
Linköping University
Linköping (Sweden)

Abstract

Andersson-Madigan-Perlman chain graphs were originally introduced to represent independence models. They have recently been shown to be suitable for representing causal models with additive noise. In this paper, we present an algorithm for learning causal chain graphs. The algorithm builds on the ideas by Hoyer et al. (2009), i.e. it exploits the nonlinearities in the data to identify the direction of the causal relationships. We also report experimental results on real-world data.

Keywords: Causality; learning; chain graphs.

1. Introduction

This paper deals with causal discovery under the assumption that the noise is additive, i.e. each observed random variable Y is of the form $Y = g(Pa(Y)) + U_Y$, where $Pa(Y)$ denotes its observed causes (also called parents) and U_Y represents its unobserved causes (also called noise or error or residual).[1] Additive noise is a rather common assumption in causal discovery (Hoyer et al., 2009; Mooij et al., 2016; Peters et al., 2014), mainly because it produces tractable models which are useful for gaining insight into the system under study. Note also that linear structural equation models, which have extensively been studied for causal effect identification (Pearl, 2009, Chapter 5), are additive noise models.

Consider two random variables X and Y that are causally related as $Y = g(X) + U_Y$. Assume that there is no confounding, selection bias or feedback loop, which implies that X and U_Y are independent. Hoyer et al. (2009) prove that if the function g is nonlinear, then the correct direction of the causal relationship between X and Y is generally identifiable: X and U_Y are independent for the correct direction, whereas Y and U_X are dependent for the incorrect direction. This leads to the following causal discovery algorithm: If X and Y are independent then they are not causally related because we assumed no confounding, selection bias or feedback loop. If they are dependent then first construct a nonlinear regression of Y on X to get an estimate $\hat{g}$ of g, then compute the error $\hat{u}_Y = y - \hat{g}(x)$, and finally test whether X and $\hat{U}_Y$ are independent. If they are so then accept the model $X \rightarrow Y$, otherwise repeat the procedure for the model $Y \rightarrow X$. When no model is accepted, it may be indicative that the assumptions do not hold. Hoyer et al. (2009) also propose a generalization to more than two variables: Given a directed and acyclic graph (DAG) over the observed random variables, first construct a nonlinear regression of each node on its parents, then compute each node's error, and finally test whether these errors are mutually

[1] Without loss of generality, the unobserved causes can be summarized by a unidimensional random variable (Mooij et al., 2016, Proposition 4).

```
**Input**: A CG $G$.
**Output**: The magnified CG $G'$.

Set $G' = G$
For each node $X$ in $G$
    Add the node $U_X$ and the edge $U_X \to X$ to $G'$
For each edge $X - Y$ in $G$
    Replace $X - Y$ with the edge $U_X - U_Y$ in $G'$
Return $G'$
```

Table 1: Algorithm for magnifying a CG.

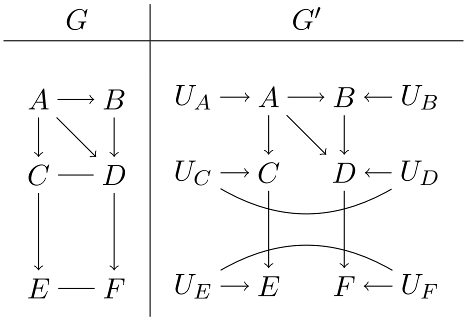

Figure 1: Example of the magnification of a CG.

independent. If they are so then accept the DAG, otherwise reject it. The algorithm performs well in practice (Peters et al., 2014). In this paper, we further generalize this idea by dropping the assumption that the errors are independent. Specifically, we use Andersson-Madigan-Perlman chain graphs (CGs) instead of DAGs to represent causal models. These CGs were originally introduced to represent independence models (Andersson et al., 2001). They have recently been shown to be suitable for representing causal models with additive noise (Peña, 2016). Specifically, we can interpret the parents of each node in a CG as its observed causes. Its unobserved causes are summarized by an error node that is represented implicitly in the CG. We can interpret the undirected edges in the CG as the correlation relationships between the different error nodes. The causal structure is constrained to be a DAG, whereas the correlation structure can be any undirected graph as long as it does not create semidirected cycles. This may be best understood by making the error nodes explicit by magnifying the CG as shown in Table 1. The magnification basically consists in adding the error nodes to the CG and connect them appropriately. Figure 1 shows an example.

The rest of this paper presents a novel learning algorithm for CGs that builds on the ideas by Hoyer et al. (2009). The paper also includes experimental results on real-world data. The paper ends with a discussion on follow-up questions worth investigating.

```
   **Input**: A dataset $D$ over $V$ with $M$ instances, and an integer $L$.
   **Output**: A CG over $V$.

1  Let $\mathcal{G}$ be a random sample of $L$ CGs over $V$
2  For each $G$ in $\mathcal{G}$
3     For each $X \in V$
4        $\hat{g}(x|pa_G(X)) = GP(X, Pa_G(X), D)$
5     For $m = 1, \ldots, M$
6        $\hat{u}_X^m = x^m - \hat{g}(x^m|pa_G^m(X))$
7     Let $D_U$ denote the dataset over $U$ created in the previous line
8     For each $X \in V$
9        $\hat{h}(\hat{u}_X|\hat{u}_{Ne_G(X)}) = GP(\hat{u}_X, \hat{u}_{Ne_G(X)}, D_U)$
10    For $m = 1, \ldots, M$
11       $\hat{r}_X^m = \hat{u}_X^m - \hat{h}(\hat{u}_X^m|\hat{u}_{Ne_G(X)}^m)$
12    Let $D_R$ denote the dataset over $R$ created in the previous line
13    $pvalue(G) = HSIC(D_R)$
14 Return the simplest CG in the set $arg\,max_{G \in \mathcal{G}}\ pvalue(G)$
```

Table 2: Algorithm for learning CGs.

2. Learning Algorithm

The learning algorithm can be seen in Table 2. It receives as input a dataset D with M instances over the observed random variables V, and an integer L. The algorithm consists in sampling L random CGs over V (line 1), scoring each of them with respect to D (lines 2-13), and returning the best one (line 14). Scoring a CG G starts in lines 3-4 pretty much like the algorithm by Hoyer et al. (2009), i.e. obtaining an estimate $\hat{g}$ of g by constructing a nonlinear regression of each node X on $Pa_G(X)$ using Gaussian processes (GPs) (Rasmussen and Williams, 2005). This estimate is used in lines 5-6 to compute the errors. We use a superscript to indicate the value of a set of variables in a particular instance of D, e.g. x^m and $pa_G^m(X)$ represent the value of X and its parents in the m-th instance of D. To test that the errors U have the correlation structure dictated by the CG, we have to test that each error U_X is independent of the rest of the errors given its neighboring errors $U_{Ne_G(X)}$. This is what lines 8-11 intend to do. Specifically, they construct a nonlinear regression of each error U_X on $U_{Ne_G(X)}$ in order to compute the residuals of the errors. Finally, line 13 scores the whole model by testing the independence of these residuals. The null hypothesis is joint independence. Specifically, the function $HSIC$ returns the p-value of the Hilbert Schmidt independence criterion, which is a kernel statistical test of independence (Gretton et al., 2008). Strictly speaking, lines 8-11 do not test that U_X is independent of the rest of the errors given $U_{Ne_G(X)}$. These lines test that the expectation of U_X is independent of the rest of the errors given $U_{Ne_G(X)}$. As we will see later, this solution works well in practice.

Note that several CGs may score the highest p-value, e.g. every supergraph of a CG with the highest p-value may also receive the highest p-value. Therefore, line 14 applies the Occam's Razor principle and returns the simplest best CG.

Using GPs and the HSIC test in lines 4, 9 and 13 are choices shared with Hoyer et al. (2009). Other choices are also possible.

3. Experiments

In this section, we run the learning algorithm introduced above on the DWD dataset, which contains climate data from the Deutscher Wetterdienst (`www.dwd.de/EN`) and has been used before for benchmarking causal discovery algorithms (Mooij et al., 2016). We use the data provided by this last reference.[2] The data consists of 349 instances, each corresponding to a weather station in Germany. Each instance consists of measurements for six random variables. We use only four of them since the number of CGs over four nodes is manageable and, thus, our learning algorithm has a chance to test most if not all of them. Specifically, there are 1688 CGs over four nodes (Steinsky, 2003), and we let our algorithm sample 10000 CGs in line 1. The four random variables that we consider are altitude (A), temperature (T), precipitation (P), and sunshine duration (S). The last three variables represent annual mean values over the years 1961-1990. Mooij et al. (2016, Appendix D.1) argue that the causal relationships $A \rightarrow T$, $A \rightarrow P$ and $A \rightarrow S$ are true. Their arguments are meteorological, i.e. not based on the data. We confirmed that these decisions make sense according to Wikipedia (entries for the terms ”rain” and ”precipitation”).

The learning algorithm is implemented in `R`.[3] We use the packages `kernlab` and `dHSIC` for the GPs and the HSIC test. Unless otherwise stated, we use the packages' default parameters. This means that we use the default Gaussian kernel for the GPs. However, we do not use the default automatic method for estimating the kernel width. Instead, we take a sort of Bayesian approach to set its value. Specifically, we use the three ground truth relationships $A \rightarrow T$, $A \rightarrow P$ and $A \rightarrow S$ as prior knowledge and choose a kernel width that detects the correct direction of these relationships. The default automatic estimation method misses one. So does setting the width to 1. Setting it to 0.5 misses none. So, we choose this value. Of course, more elaborated procedures may be devised, e.g. based on cross-validation or on a full Bayesian approach.

3.1 Results

Figure 2 shows the CG learned. This CG is clearly preferred (p-value = 8.246e-8) over a CG with only the three ground truth relationships (p-value = 1.733e-46). The p-values should be interpreted with caution. Their relative values are informative. However, their absolute values may not for the simple reason that they can be made arbitrarily close to 1 by overfitting the GP corresponding to each variable (since this would drag the residuals towards 0). Therefore, p-values may be used to compare models but not to reject models.

Note that there are many other CGs that represent the same independence model as the CG learned, i.e. they are Markov equivalent. However, they all represent different causal

[2] The data is also available at `https://www.dropbox.com/s/rs4q8oeutgqfcdn/D1.csv?dl=0`.

[3] The code is available at `https://www.dropbox.com/s/vxir1997w5hiiz6/AN.R?dl=0`.

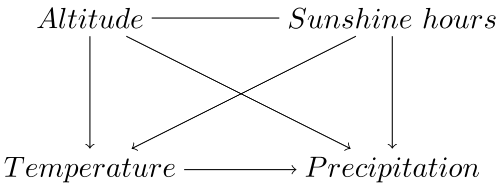

Figure 2: CG learned from the DWD dataset.

models and, thus, they may receive different p-values. For instance, the best and second best CGs found by our algorithm are Markov equivalent but they receive p-values 8.246e-8 and 3.431e-8, respectively.

In the rest of this section, we argue that the CG learned is plausible. The relationship $A \rightarrow T$ is confirmed by both Mooij et al. (2016) and Wikipedia. We can also think of an intervention where a thermometer is taken to higher and higher places. We expect that the higher the colder.

The relationships $A \rightarrow P$ and $A \rightarrow T \rightarrow P$ are confirmed by Wikipedia. Rain is produced by the condensation of atmospheric water vapor. Therefore, increasing water vapor in the air and/or decreasing the temperature are the main causes of precipitation. One way water vapor gets added to the air is due to lifting air over mountains. Mooij et al. (2016) also confirm $A \rightarrow P$ due to the mediator T.

The relationship $S \rightarrow T$ seems natural. We can also think of an intervention where we install new suns. We expect that the more suns the warmer.

The relationship $A - S$ is confirmed by Mooij et al. (2016), because all the mountains in Germany are in the south and the south is typically sunnier. The authors actually say that latitude is a confounder of A and S. However, we can think of an intervention where a piece of land is moved to the south. We do not expect the land to become more mountainous. So, we prefer to say that the unobserved causes of A and S are just correlated. This is what the CG learned actually represents.

Finally, the relationship $S \rightarrow P$ is unconfirmed.

Mooij et al. (2016) also confirm the relationships $A \rightarrow S$ due to clearer skies at higher altitudes, and $A - T$ again due to latitude. These relationships are not included in the CG learned, because a CG cannot contain a subgraph $A \rightarrow S - A$ or $A \rightarrow T - A$ due to the semidirected acyclity constraint. Our learning algorithm decided to include $A - S$ and $A \rightarrow T$ and not the others, because they may be more beneficial for the model as a whole. Further evidence on this behavior can be obtained by studying the relationships $T \rightarrow P$ and $S \rightarrow P$, which are included in the CG learned. If we just provide our learning algorithm with the data over T and P, then it prefers $P \rightarrow T$ (p-value = 0.007) over $T \rightarrow P$ (p-value = 7.748e-11). Likewise, if we just provide our learning algorithm with the data over S and P, then it prefers $P \rightarrow S$ (p-value = 0.021) over $S \rightarrow P$ (p-value = 0.011). So, our algorithm manages to get the apparently correct directions $T \rightarrow P$ and $S \rightarrow P$ just because it evaluates a whole model and not just individual relationships. In other words, by scoring a whole mode, relationships may introduce beneficial constraints (e.g. to avoid semidirected cycles) on the direction of other relationships. Note however that scoring a whole model can also impose harmful constraints. For instance, as discussed above, our algorithm is forced to decide between $A \rightarrow S$ and $A - S$ because both cannot be represented in a CG. Hence, the need to extend our learning algorithm to more general causal models, e.g. alternative acyclic directed mixed graphs which relax the semidirected acyclity constraint of CGs so as to forbid only directed cycles (Peña, 2016).

4. Discussion

We have proposed an algorithm for learning causal CGs. The algorithm exploits the nonlinearities in the data to identify the direction of the causal relationships. The algorithm uses a brute-force approach, i.e. it simply samples some CGs and returns the one with the highest score. The score is computed by (i) regressing each variable on its parents, (ii) computing the residual, (iii) regressing each residual on the neighboring residuals, (iv) computing the residual of the residual, and (v) computing the p-value of a test of independence of the latter residuals. Our experiments on real-world data have shown that this score is able to identify a plausible CG.

Our experiments have also helped us to identify some questions worth investigating further. For instance, it may be interesting to consider learning causal models that are more general than CGs, e.g. alternative acyclic directed mixed graphs (aADMGs) (Peña, 2016). These are graphs with directed and undirected edges, up to two different edges between any pair of nodes, and no directed cycle. Note that the algorithm in Table 2 is readily applicable to learn aADMGs: Just sample aADMGs instead of CGs in line 1. While aADMGs allow for a richer family of causal models to be represented, they also imply a larger search space, e.g. whereas there are 1688 CGs over four nodes, there are 34752 aADMGs (543 DAGs times 64 undirected graphs). This is a major problem for our brute-force learning algorithm. A solution is to develop a greedy hill-climbing version of the algorithm that evaluates all the models that differ from the current one by one edge and then moves to the best of them. Note that we do not have to compute the score from scratch for each candidate model to evaluate, as steps (i)-(iv) are the same for most nodes in the current and candidate models. This approach performed well for learning causal DAGs (Peters et al., 2014). It should be noted that the authors added an ad-hoc penalty for complexity to the score in order to return a minimal DAG, i.e. to avoid its supergraphs. This addition is also worth investigating, as larger search spaces usually imply larger risk of overfitting. Recently, Nowzohour and Bühlmann (2016) have proposed a more principled score, namely a penalized maximum likelihood score, as part of a new algorithm for learning causal DAGs. We end this section presenting an algorithm for learning causal CGs that is inspired by the score of Nowzohour and Bühlmann (2016).

The learning algorithm in Table 2 makes use of independence tests to assess if the independence model over the error terms induced by the CG being evaluated fits the learning data. Table 3 presents an alternative that makes use of a penalized maximum likelihood score to assess the fit. This learning algorithm assumes that the error terms follow a joint Gaussian distribution. This a relatively common assumption in causal discovery (Bühlmann et al., 2014; Imoto et al., 2002; Peters and Bühlmann, 2014) for mainly two reasons. First, if the noise is actually the summation of the noise due to different independent sources, then the noise approximately follows a Gaussian distribution by the central limit theorem

```
    **Input**: A dataset $D$ over $V$ with $M$ instances, and an integer $L$.
    **Output**: A CG over $V$.

 1  Let $\mathcal{G}$ be a random sample of $L$ CGs over $V$
 2  For each $G$ in $\mathcal{G}$
 3    For each $X \in V$
 4      $\hat{g}(x|pa_G(X)) = GP(X, Pa_G(X), D)$
 5    For $m = 1, \ldots, M$
 6      $\hat{u}_X^m = x^m - \hat{g}(x^m|pa_G^m(X))$
 7    Let $D_U$ denote the dataset over $U$ created in the previous line
 8    $\hat{\mu} = mean(D_U)$
 9    $\hat{\Sigma} = IPF((G')_U, D_U)$
10    $score(G) = \log p(D_U|\hat{\mu}, \hat{\Sigma}) - \frac{\log(M)}{2}|E(G)|$
11  Return $arg\, max_{G \in \mathcal{G}}\ score(G)$
```

Table 3: Algorithm for learning CGs under the assumption of Gaussian noise.

of probability theory. Second, additive Gaussian noise produces tractable models which are useful for gaining insight into the system under study. In our case, for instance, we can first compute the error terms in line 7 and then obtain the maximum likelihood estimate of the covariance matrix over the error terms in line 9 via the iterative proportional fitting (IPF) procedure (Wainwright and Jordan, 2008). In line 9, $(G')_U$ denotes the subgraph of the magnified CG G induced by the error nodes (recall Section 1). This enables us to devise a BIC-inspired score in line 10, where the penalty for model complexity is simply the number of edges in the CG G, i.e. $|E(G)|$. Of course, other penalties are also possible. In line 10, note that $p(D|\hat{\mu}, \hat{\Sigma}) = p(D_U|\hat{\mu}, \hat{\Sigma})$ because V is determined by U.

As mentioned, our algorithm in Table 3 is inspired by the work of Nowzohour and Bühlmann (2016). However, there are two main differences. Unlike them, we do not assume that the individual error terms are jointly independent. Like us, they consider additive noise but, unlike us, they do not assume that it is Gaussian. This may actually be a strong assumption in some domains. For the DWD dataset in Section 3, for instance, we did not find any value for the kernel width that gave higher scores to the correct directions of the three ground truth relationships than to the wrong directions. We interpret this as an indication that the Gaussian noise assumption is not reasonable. Therefore, we plan to generalize the algorithm in Table 3 by dropping the Gaussian noise assumption.

References

S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative Markov Properties for Chain Graphs. *Scandinavian Journal of Statistics*, 28:33–85, 2001.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. *Annals of Statistics*, 42:2526–2556, 2014.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, 2008.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear Causal Discovery with Additive Noise Models. In *Advances in Neural Information Processing Systems 21*, pages 689–696, 2009.

S. Imoto, T. Goto, and S. Miyano. Estimation of Genetic Networks and Functional Structures Between Genes by Using Bayesian Networks and Nonparametric Regression. In *Proceedings of the Seventh Pacific Symposium on Biocomputing*, pages 175–186, 2002.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

C. Nowzohour and P. Bühlmann. Score-Based Causal Learning in Additive Noise Models. *Statistics*, 50:471–485, 2016.

J. M. Peña. Alternative Markov and Causal Properties for Acyclic Directed Mixed Graphs. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 577–586, 2016.

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

J. Peters and P. Bühlmann. Identifiability of Gaussian Structural Equation Models with Equal Error Variances. *Biometrika*, 101:219–228, 2014.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2005.

B. Steinsky. Enumeration of Labelled Chain Graphs and Labelled Essential Directed Acyclic Graphs. *Discrete Mathematics*, 270:267 – 278, 2003.

M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.