Latent Causation: An algorithm for pairs of correlated latent variables in Linear Non-Gaussian Structural Equation Modeling*

Arnaud Pollaris and Gianluca Bontempi

Université Libre de Bruxelles, ULB CP212, boulevard du Triomphe, 1050 Bruxelles, Belgium Arnaud.Pollaris@ulb.ac.be and Gianluca.Bontempi@ulb.ac.be
https://mlg.ulb.ac.be

Abstract. This paper addresses the problem of inferring causation in a pair of linearly correlated continuous latent variables. We first discuss the limitations of the Direction Dependance Analysis (DDA) approach and then introduce the Latent Causation (LC). Five variants (in terms of dependency statistic) of the LC algorithm are assessed with ROC curves, then we consider the case of a latent confounder (uniform or chi-square distributed). While the distribution and the correlations of the latent confounder influence the accuracy, experimental results show the robustness of the method using bootstrapped p-values. Implications and limits of the experimental results are then discussed together with future directions.

Keywords: SEM · Latent Variables · Causal inference · Observational data · Latent Confounder · Non normality · Simulations.

1 Introduction

An observed dependency between two variables A and B may have four different explanations assuming no feedback loop: (1) A is a (direct or indirect) cause of B, (2) B is a (direct or indirect) cause of A, (3) there is a hidden common ancestor U of A and B, (4) a common descendant of A and B is kept fixed in the observed dataset.

This paper lies at the crossroad between Structural Equation Modeling (SEM) and causal inference literature. Widely used in psychology and in management research, SEM is a family of techniques which allows the analysis of relationships between continuous latent variables. Whereas the capacity of SEM to support causal inference has been discussed during decades (Bollen and Pearl, 2013), we consider here classical SEM as a set of confirmatory techniques since the causal graph specified by the user can (and should) be drawn before the data collection. In that perspective, the fitting of the data on one or more competing causal models should allow to reject wrong models and then inform the scientist that at least one of its related causal assumptions is wrong (see Bollen and Pearl,

* We would like to thank the anonymous reviewers for their helpful comments.

2013). However, for equivalent models (i.e., *alternative models that fit any data to the same degree* (MacCallum & Austin, 2000 p. 213)) it is not always easy to retrieve information about causation. So additional tools are needed.

In machine learning, causal discovery is not confirmatory but exploratory: its goal is to build a causal model based on available data (data collection comes before the learning of a causal graph). Though in causal discovery most algorithms make the assumption of causal sufficiency (also in cause-effect pairs (e.g., Guyon, 2014)), some of them address latent variables, e.g., the BPC algorithm (Silva, Scheine, Glymour, and Spirtes, 2006), the FindOneFactorClusters (FOFC) algorithm (Kummerfeld and Ramsey, 2016) or, more recently, the LSTC algorithm (Cai, Xie, Glymour, Hao & Zhang, 2019). Shimizu, Hoyer and Hyvärinen (2009) also show that a linear acyclic model for latent factors is identifiable when the data are non normal. However, our work in this paper differs from this literature because we assume the structure of the measurement model already known (i.e.: each indicator has been specified as measuring exactly one latent variable of interest in our models) and we do not focus on building large causal graphs from data.

In particular we focus on causal inference in pairs of latent variables. In a confirmatory perspective, assuming linearity and non normality, the Direction Dependent Analysis project (DDA project, 2020) offers an interesting starting point to infer causation in a pair of latent variables since indications for a latent confounder can also be detected using its independence component. However, as stressed in (Wiedermann, Merkle, & von Eye, 2018), there is still a need for improving the trustworthiness of the DDA approach in presence of meaningful confounding. For this reason, in this paper we focus on improving the independence component of the DDA approach by focusing in particular on discriminating between causal and spurious confounding latent configurations.

The paper is structured as follows: First, we present the causal inference setting we are interested in. Next, the DDA approach is introduced. Then, limitations for using DDA with latent variables are presented. Next, we propose the Latent Causation (LC) algorithm, grounded on the third DDA component "Independence properties of predictor and error term" (see Wiedermann & Li, 2018). Then, we present some experiments on simulated data: benchmarking LC with respect to to state-of-the-art DDA and sensitivity study of LC.

2 Problem setting

Let us consider two continuous correlated latent variables, denoted ξ and η and some observable children variables called "indicators" (e.g., Kline, 2011) which are functions of a latent variable plus an additive independent noise. Figure 1 visualizes a causal and a confounding topology we want to discriminate between. As an example, values and distributions specified in Figure 1 are possible instances which are used below as assignations for parameters in our simulations. The number of indicators can also differ from instances in Figure 1. While we want to confirm the correct causal direction $\xi \rightarrow \eta$ (and not $\xi \leftarrow \eta$) in causal

models like Model 1, we also want to make sure we will not conclude in favor of a causal direction for pairs (ξ, η) only correlated due to a latent confounder (called U below) like in Model 0. Throughout the paper, linearity is assumed, variables are continuous and all coefficients in theoretical models are presented for standardized variables.

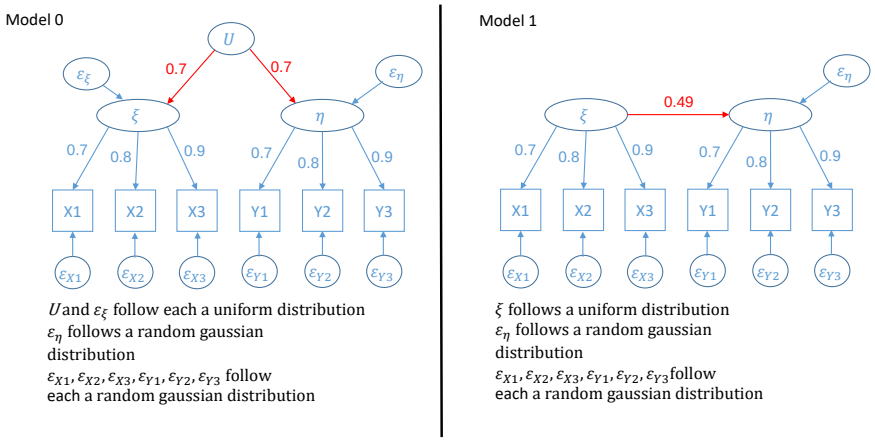

Fig. 1. Latent variables are represented by ellipses or circles. Observed variables are represented by rectangular boxes. Both models are completely standardized. In Model 1, there is a causation between ξ and η but not in Model 0. Note that the distributions of ξ and η also differ between both models.

3 The DDA approach

The DDA project regroups the techniques which address the inference of causation by considering 3 aspects: 1. distributional properties of observed variables, 2. distributional properties of error terms of competing models and 3. independence properties of predictor and error terms of competing models. In this paper we will focus on the third aspect. The rationale of our work resides in this consideration by Wiedermann & Sebastian (2019a, p. 15): *"Considering the behavior of DDA components under confounding, the DDA independence component is the most important criterion to confirm that no strong confounders are present (or, at least, that the influence of confounders is minimal). Thus, HSIC- and dCor-tests are crucial DDA procedures. When these tests indicate the presence of meaningful confounding, results of the remaining DDA procedures are no longer trustworthy."*

The algorithm used for the (third) DDA independence component is already well-established in machine learning (see Peters, Janzing & Schölkopf, 2018, p. 62):

1. Fit a regression model $\hat{f}_Y$ of Y on X
2. Test whether the residual $Y - \hat{f}_Y(X)$ is independent of X.
3. Repeat the procedure by exchanging the roles of X and Y.
4. If the independence is accepted for one direction and rejected for the other, infer the former one as the causal direction.

However, while the algorithm presented by Peters et al. (2018) is commonly used to determine the causal direction for data with relationships strictly non-linear, in the DDA framework, this algorithm is used assuming linearity and non-normality (Thoemmes, 2019).

Here is the rationale behind the algorithm for DDA. For instance, let us then assume that X and Y are two related continuous random variables such that :

$$Y = aX + \epsilon_Y \qquad \text{with } a \neq 0 \tag{1}$$

where ϵ_Y is the error term from a regression where Y is explained linearly as a function of X. And let us assume that either X or ϵ_Y is not normally distributed. In this context, we get the falsehood of the expression

$$X \perp\!\!\!\perp \epsilon_Y \qquad \text{AND} \qquad Y \perp\!\!\!\perp \epsilon_X \tag{2}$$

where ϵ_X is the error term from an alternative linear regression having X explained linearly as a function of Y:

$$\epsilon_X = X - bY \qquad \text{with } b \neq 0 \tag{3}$$

The above reasoning relies on the corollary of the Darmois-Skitovich theorem (see e.g.: Eberhardt, 2017, p.86):

Corollary 1. *Let $X_1, ..., X_n$ be independent, non-degenerate random variables. If for two linear combinations :*

$$l_1 = a_1X_1 + ... + a_nX_n \qquad \textit{with } a_i \neq 0 \tag{4}$$

$$l_2 = b_1X_1 + ... + b_nX_n \qquad \textit{with } b_i \neq 0 \tag{5}$$

at least one X_i is not normally distributed, then l_1 and l_2 are not independent.

After substitution of Y in (3) by its expression from (1):

$$\epsilon_X = X - b(aX + \epsilon_Y) = (1 - ab)X - b\ \epsilon_Y \tag{6}$$

it appears both ϵ_X and Y are linear combinations of X and ϵ_Y. Applying Corollary 1, it can then be affirmed that if X and ϵ_Y are independent, non-degenerate random variables that are not normally distributed in (1), then Y and ϵ_X can not be independent in (3).

Then, as nicely illustrated in Spirtes & Zhang (2016) and in Wiedermann & Li (2018), this asymmetric pattern of the causality can leave a footprint in the data. Furthermore, the shape of the distribution does not matter, since it is not a normal distribution.

3.1 Limitations of the current DDA approach

Here we address the limitations of the current DDA approach for causal inference in pairs of latent variables:

1. **DDA does not exploit all the available information in measures of dependencies**. The DDA independence component approach relies on a combination of two statistical tests of independence (see e.g., in equation (2), one test for each possible causal direction). Four conclusions are possible: (a) rejection of both independences (i.e., suspicion of confounder), (b) no rejection of both independences, and rejection of only one of the two independences which gives either (c) $X \to Y$ or (d) $Y \to X$. It is worth remarking what follows:
   (a) **A non-significant result for a test of independence is recommended** to infer a causal direction. But, the lack of independence rejection is not a proof of independence. So, in DDA, how to be sure we are not missing a confounder because of a lack of power in the test?
   (b) **Insufficient use of the continuum of dependencies.** Given a pair of variables, there is not only simplistically either "independence" or "dependence" but a whole range of dependence strengths. Even if DDA concludes in favour of the presence of a confounder (i.e, both independences are rejected), the comparison of statistics of dependence from both tests may convey additional information to favour one of the two causal directions, under the assumption there is also causation in the pair of variables of interest (beyond a simple spurious correlation).
2. **Limit of the DDA distinction between "presence of a confounder" and "causation".** In the DDA framework, the presence of a confounder may be revealed by rejection of the independence for both directions whereas a causation should be revealed by the rejection of the independence for only one direction. But some theoretical models can include both causation and confounder. Considering pairs of latent variables like in the example of the causal Model 1 (Fig. 1, right) with no confounder between ξ and η, the latent variable ξ can also be considered itself as a latent confounder between the two groups of observed indicators: $(X1, X2, X3)$ and $(Y1, Y2, Y3)$, where each indicator is a linear combination including the latent variable ξ. So, since ξ can be both a cause of η and a confounder between the two groups of indicators, maybe it would be better not to rely on the DDA independence component if we want to confirm there is no additional confounder between latent ξ and latent η when it concludes in favor of a causation. And, if the DDA independence component concludes there is a confounder, the question of the direction for a possible causation remains still open after.

So, to infer a causation in a pair of non-Gaussian linearly related latent variables, the point is maybe not to make sure first there is no latent confounder but to focus instead on a direct comparison of the strength of each dependence (one for each possible causal direction) to try to infer (if possible and with an associated level of confidence) a causal direction like LC does.

4 The Latent Causation algorithm

The main difference between the LC and the DDA independence component is related to the computation of differences between the statistics of dependency, which is an essential element to identify a causal direction. The Latent Causation (LC) pseudocode is detailed below. After the computation of factor scores (F_ξ, F_η) in steps (2a) and (2b) representing latent variables ξ and η respectively, the steps (2c), (2d), (2e) and (2f) implement the DDA independence component. However, if there is a true causal link between ξ and η, a difference of dependence should be observed between values computed in (2e) and (2f). Unlike classical DDA (yet inspired to a sensitivity analysis performed by Wiedermann and Sebastian (2019b) using bootstrap), the difference of values in (2f) and (2e) is always saved by LC in step (2g). While a positive score in (2g) is favouring one causal direction, a negative score is favouring the opposite causal direction. A non-parametric bootstrap (B resamples) is then adopted to assess the significativity of the consensus. LC may reach 3 conclusions in step (4): "infer $\eta \to \xi$", "infer $\xi \to \eta$" or "data do not allow to conclude." Unlike in classical DDA, we do not need to have a non significant p-value to conclude for a causal direction.

The Latent Causation algorithm

Input:

- An observed dataset with indicators divided in 2 pre-defined groups (with no overlap): $\mathbf{X}$ for the indicators of ξ, $\mathbf{Y}$ for the indicators of η.
- A metric to rate the strength of a bivariate dependence
- α: a threshold (to define acceptable type I error rate)
- B: number of bootstrap datasets

Output: A decision taken by the algorithm:
"infer $\eta \to \xi$" OR "infer $\xi \to \eta$" OR "data do not allow to conclude."

1. From the original sample of size n, draw B bootstrap samples ($size = n$, with replacement).
2. For each bootstrapped sample do :
   (a) Compute the factor scores "F_ξ" to represent ξ using $\mathbf{X}$ (exclusively)
   (b) Compute the factor scores "F_η" to represent η using $\mathbf{Y}$ (exclusively)
   (c) Regress linearly F_η *as a function of* F_ξ and save the residuals ($resid_{F_\eta}$)
   (d) Regress linearly F_ξ *as a function of* F_η and save the residuals ($resid_{F_\xi}$)
   (e) Measure how strong $dependence(resid_{F_\eta}, F_\xi)$ is
   (f) Measure how strong $dependence(resid_{F_\xi}, F_\eta)$ is
   (g) Save the difference between both measures from (f) and (e)
3. Based on the B saved differences (in 2g), select a percentile confidence interval based on probabilities ($\alpha/2$; $1 - \alpha/2$).
4. Select a conclusion:
   - If 0 is not included in the confidence interval:
     - If a majority of bootrapped samples gave:
       $dependence(resid_{F_\eta}, F_\xi) > dependence(resid_{F_\xi}, F_\eta)$:
       "infer $\eta \to \xi$"

    - Else:
      "infer $\xi \rightarrow \eta$"
- Else:
  "data do not allow to conclude."

Some considerations on the LC algorithms follow:

- Factor scores computation: since the information about the latent variables ξ and η is only available through noisy indicators, the question about their representation naturally arises. While each indicator is assumed to be a linear descendant of a specific latent variable of interest, we choose Principal Component Analysis (PCA) (Husson, Lê & Pagès, 2009) to compute factor scores separately for each latent variable ξ and η.
- Dependency measures: we considered 5 measures in our LC experiments:
  - Spearman's correlation (in absolute value).
  - Brownian distance correlation (Szekely, Rizzo, & Bakirov, 2007): it returns the dCor statistic (a score between [0;1] where 0 stands for independence).
  - dCor's p-value : it estimates a p-value using permutation bootstrap.
  - Hilbert-Schmidt Independence Criterion (HSIC; Gretton, Fukumizu, Teo, Song, Schölkopf & Smola, 2008): The closer HSIC is with 0, the weaker is the dependence.
  - HSIC's gamma-approximated p-value: The smaller the p-value is, the more we are in independence rejection.
- LC relies on some assumptions:
  - Two correlated non normal latent variables (i.e., ξ and η).
  - All the relationships are linear (measurement model included).
  - There is no cycle in the causations.
  - Two distinct groups of indicators are available for ξ and for η respectively. Each indicator is strongly correlated (e.g., Pearson's correlation ≥ 0.7 in our simulations) with its corresponding latent variable (either ξ or η).
  - Each indicator is linearly function of its latent variable + an independent random Gaussian noise.
  - If there is a causal effect between ξ and η, it is assumed the effect is the same for every individual (causal effect homogeneity).
  - There is no unusual or influential observations.
  - All the variables are continuous.

5 Experimental results

The experimental results are divided in two subsections: first, we use simulations to compare LC and DDA. Second, additional analysis are provided to further explore the performances of five LC variants.

5.1 Benchmarking LC vs DDA

Data generation. In this section, to compare LC and DDA, we used the causal structures from Model 0 and Model 1 to generate datasets by Monte Carlo (1000 datasets for each model). The distributions and the standardized values specified for the different coefficients (at the Population level) are available in Fig. 1. Since we want to infer causation beyond correlation, we arbitrary specify in every simulation in this paper the same theoretical Pearson's correlation of 0.49 between ξ and η.

To generate our datasets and compare the methods, we implemented a simulator in the R language (R Core Team, 2019)[1]. We consider two groups of 3 indicators: X_1, X_2, X_3 and Y_1, Y_2, Y_3 measuring latent ξ and η respectively. Since we cannot directly apply the DDA independence component on observed indicators, factor scores representing ξ and η are first computed using the first axis in two separated PCA (i.e., first axis build on X_1, X_2, X_3 and other first axis build on Y_1, Y_2, Y_3). So, we used PCA[2] in a similar way in LC and before applying DDA.

Accuracy assessment. Table 1 reports the comparison DDA vs LC in terms of accuracy. In DDA, since two p-values are used for taking decisions, 4 conclusions are possible: ξ causes η, η causes ξ, suspicion of a latent confounder (both p-values are sig.) and the "do not conclude" option (none of the 2 p-values is sig.). In LC, based on a bootstrapped confidence interval, 3 conclusions are possible: ξ causes η, η causes ξ and the "do not conclude" option.

Concerning DDA for Model 0 using HSIC gamma p-value or HSIC p-value bootstrap (Sen and Sen, 2014) as independence test, it appears that $(72 + 43)/1000 = 11.5\%$ and $(100 + 84)/1000 = 18.4\%$ of the conclusions are false positives (FP) (indicating wrongly causation) which exceeds in both cases the maximum $\alpha = 5\%$ allowed. In contrast, DDA's dCor p-values seem to work fine on Model 0 (FP rate: 0.4%), though this method shows less statistical power (only 303 on 1000 datasets were true positive (TP) causal conclusions) than LC variants (401 TP using Spearman as independence statistic, 510 TP using dCor-stat, 343 TP using dCor-p-value, 594 TP using HSIC-stat and 608 TP using HSIC's gamma-approximated p-value).

In Table 1, all LC variants show a FP rate under the expected $\alpha = 5\%$ (e.g., the observed total FP rate using HSIC-stat $= (1 + 2)/1000 = 0.3\%$) and outperform in power (i.e., number of TP in Model 1) DDA's dCor which is the only considered DDA variant with a FP rate below $\alpha = 5\%$.

Discussion of results. To infer causation in a pair of correlated latent variables, we are looking for an algorithm with low type I error rate (i.e a proportion of FP below the specified value for α) when there is no true causation between ξ

[1] Code in https://github.com/apollaris/LatentCausation

[2] In experiments we use the PCA function (using default option "scale.unit = TRUE") from the R package FactoMineR (Le, Josse & Husson, 2008).

Table 1. The DDA independence component applied on factor scores from PCAs vs. the LC algorithm (DNC stands for "Do not conclude")

DDA independence component	**Model 1:** true causation $\xi \to \eta$				**Model 0:** no causation but a latent confounder			
Independence statistic	$\xi \to \eta$ (TP)	$\eta \to \xi$ (FN)	Con-founder (FN)	DNC (FN)	$\xi \to \eta$ (FP)	$\eta \to \xi$ (FP)	Con-founder (TN)	DNC (TN)
dCor	303	0	0	697	2	2	0	996
HSIC - gamma	860	1	28	111	72	43	34	851
HSIC - bootstrap	870	1	71	58	100	84	62	754
LC algorithm	**Model 1:** true causation $\xi \to \eta$				**Model 0:** no causation but a latent confounder			
Independence statistic	$\xi \to \eta$ (TP)	$\eta \to \xi$ (FN)	DNC (FN)		$\xi \to \eta$ (FP)	$\eta \to \xi$ (FP)	DNC (TN)	
Spearman	401	0	599		19	0	981	
dCor - stat	510	0	490		1	0	999	
dCor - p-value	343	0	657		0	0	1000	
HSIC - stat	594	0	406		1	2	997	
HSIC - p-value gamma	608	0	392		0	3	997	

Parameters specification:
- **For both DDA and LC:** 1000 samples generated for Model 1 and 1000 samples for Model 0; sample size=500; α=0.05
- **For DDA only:** the number of replicates used for the estimation of each dCor-pvalue and the number of resamples used to compute each bootstrap's HSIC p-value were both set equal to 500.
- **For LC only:**, we used $B = 1000$ (bootstrap datasets) and the number of replicates used for the estimation of each dCor-pvalue was always set equal to 300.

and η (e.g., Model 0) and with good ability to retrieve the correct causal direction (i.e., statistical power represented here by the number of TP for datasets from Model 1). Looking only in our results at methods with an acceptable type I error rate, it appears all the five variants of LC are more powerful than the only acceptable DDA variant (using dCor as independence statistic). Furthermore, Table 1 shows the impact of the dependence statistic on the results (see e.g., for LC: 343 TP using dCor - p-value but 608 TP using HSIC's gamma-approximated p-value). The next section will present a LC sensitivity study to assess the role of the five statistics.

5.2 LC sensitivity study

Here we perform additional simulations to study the sensitivity of the LC accuracy to its parameters. Good accuracy means low type I errors (i.e., the proportion of FP can not be larger than α in the absence of causation between ξ and η) and good statistical power (i.e., a large number of TP is expected under causation between ξ and η). First, we show the statistical power of LC increases for larger sample sizes. Next, α's value is manipulated to show that LC does not exceed the allowed the type I error rate. Then we show that the TP rate increases with α. Finally, ROC curves visualize the ability of the five LC variants to discriminate beween a spurious correlation (i.e., Model 0) and a causation (i.e., Model 1). Because many different confounders can make a pair of latent variable (ξ,η) correlate, we conclude this section by a robustness anal-

ysis to answer the question: "Can a latent confounder U (due to its distribution and its correlations with ξ and η) increase LC's number of FP in the absence of causation between ξ and η?"

Sample size can increase the power of LC. 1000 datasets of different sample sizes ($n = 200$, $n = 300$, $n = 400$, $n = 500$) were simulated according to Model 1 (Fig. 1, right), i.e. with a true latent causation $\xi \to \eta$.

As observed in Fig. 2 upper left, the larger the sample size, the better is LC in retrieving $\xi \to \eta$. Furthermore, comparing the five variants of independence statistics, the methods based on HSIC appear here to be the most powerful.

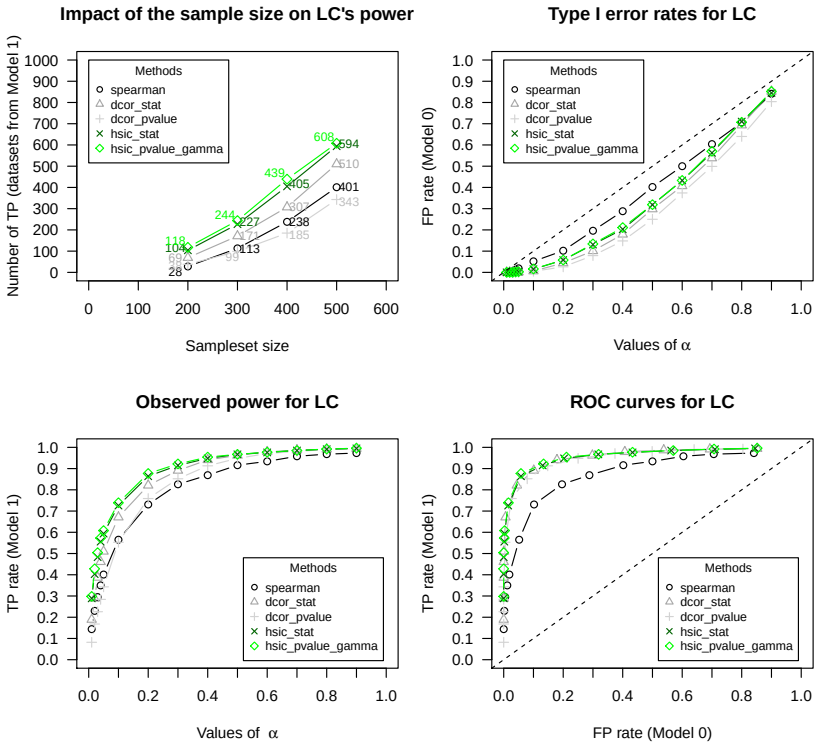

Fig. 2. Upper left: Power of the LC algorithm for retrieving the true $\xi \to \eta$ as a function of the sample size and the statistic used to measure independence; 1000 simulated datasets based on Model 1 for each sample size. Upper right: observed type I error rate (% of FP) for LC as a function of α. Lower left: observed power rate (% of TP) for LC as a function of α. Lower right: ROC curves for the 5 variants to measure dependence using LC ; for each curve, the different points correspond to different values assigned to α. For each last 3 plots: Sample size=500, 1000 simulated datasets for each estimation.

Manipulations on α. Using 1000 datasets generated from each model (Model 0 and Model 1) with a sample size $n = 500$, we manipulate the specified value of α to get the number of FP (for data generated from Model 0) and the number of TP (for data generated from Model 1). Here are the different values assigned to α: .01, .02, .03, .04, .05, .1, .2, .3, .4, .5, .6, .7, .8, .9.

Results in Fig. 2, upper right, show as expected for Model 0 that observed type I error rate (i.e., the proportion of FP) is always lower than the specified value of α in our simulations (whatever the method or the value of α).

Fig. 2 lower left shows that the number of correct causal direction (TP) increases with higher values of α and that the number of TP also differs between the five variants (measures of dependence). Notably, using HSIC's gamma-approximated p-values seem to get more TP compared with other observed methods. Then, by increasing α the number of both TP and FP increases as well. In Fig. 2 lower right, ROC curves to discriminate between Model 0 and Model 1 show that the methods (apart from Spearman correlation) present similar good abilities to discriminate between Model 1 (causation) and Model 0 (confounder).

Robustness: distribution and correlation with a latent confounder. Since estimated scores F_ξ, $resid_{F_\eta}$, F_η and $resid_{F_\xi}$ in Model 0 are all linear combinations of a sum of terms including the latent confounder U, according to Corollary 1 (of the Darmois-Skitovich theorem), $F_\xi \not\perp\!\!\!\perp resid_{F_\eta}$ and $F_\eta \not\perp\!\!\!\perp resid_{F_\xi}$ are expected together because U is not normally distributed (see also, Wiedermann et al., 2018). So, using additional simulated datasets (Monte Carlo), an empirical analysis of the robustness for LC is now performed to know if inflated type I errors (i.e, FP) can be avoided. While keeping constant the theoretical Pearson's correlation between ξ and η (as a reminder it was arbitrary set equal to 0.49 for our simulations), we generate additional datasets after manipulation of the distribution (symmetric uniform VS asymmetric chi-square) and the Pearson's correlation between U and ξ (and then also the Pearson's correlation between U and η) (see Fig. 3: Model 0 and its 4 variants : 0a, 0b, 0c and 0d for correlations assigned to the confounder). Because all methods to measure dependence were also compared here, we have now a "2 distributions of U $\times$ 5 models $\times$ 5 methods for statistics of dependence" design.

In the different plots in Fig. 4, the correlations of U influence the number of FP: for a strong correlation between U and ξ, the risk to conclude wrongly that ξ causes η increases ; on the opposite, when U mainly correlates with η, the risk to conclude wrongly that η causes ξ increases. However, the impact of the correlation of U is strongly reduced when U is symmetrical uniform (plots on the left) compared with an asymmetrical chi-square U ($df = 1$) (plots on the right). Fortunately, the problem of the distribution and correlations of U seems possible to overcome: looking at the dCor p-value method, the number of FP never exceeds 60, even when a very strong correlation of 0.875 between a chi-square U ($df = 1$) and η has been specified.

The last results seem to favour the method based of differences of dCor's p-values under the assumption of an influent latent confounder. However, a deeper

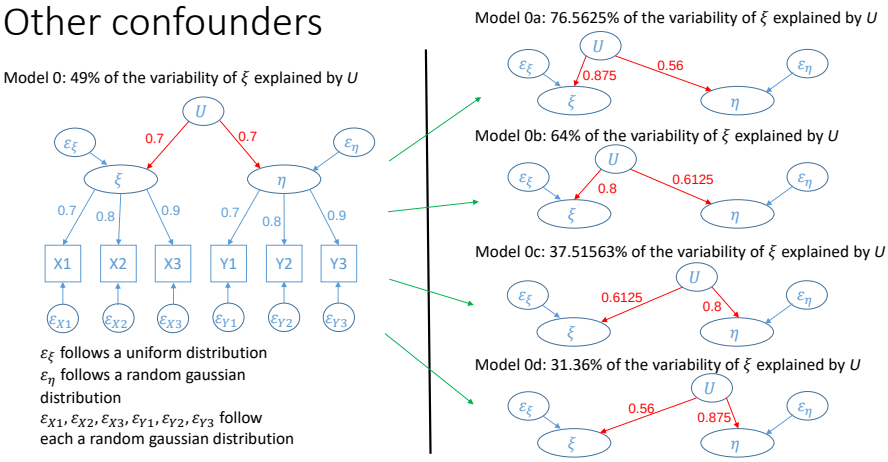

Fig. 3. In order to study the impact of the correlations of the confounder, simulations rely on variations of Model 0. While the population level's correlation between ξ and η is kept constant and equal to an arbitrary set value of 0.49, the confounder U is made more correlated with ξ in Models 0a and 0b and more correlated with η in Models 0c and 0d. In Models 0a, 0b, 0c, 0d, the measurement model (same as in Model 0) is not displayed to save space.

look at the results from extreme Models 0a and 0d with a chi-square U, reveals that for the variant using differences of HSIC's p-values, despite of the very high number of FP, the median of differences scores (-4.77e-3 for simulations based on Model 0a ; 2.14e-2 for simulations based on Model 0d) is not far from the expected 0. So, methods based on differences of p-values might have a small recurrent bias due to the presence of a (very) asymmetric, strongly correlated U with either ξ or η. Using difference of dCor's p-values, this bias can be hidden due to the random bootstrapped estimation of each p-value.

6 Conclusions and future directions

In this paper, we propose LC, an algorithm for causal inference in a pair of latent variables for confirmatory analysis. In this specific context, LC appears to be better suited than classical DDA to differentiate causation and confounder patterns from data. The resulting recommendation is then to enrich DDA analysis with bootstrapped differences of independence statistics (possibly also outside the context of latent variables).

Directions for LC improvement may also be considered. For instance, promising research directions to extend the current work are:

– **Considering alternative ways to compute factors scores.** There are indeed alternatives to PCAs to represent latent ξ and η.

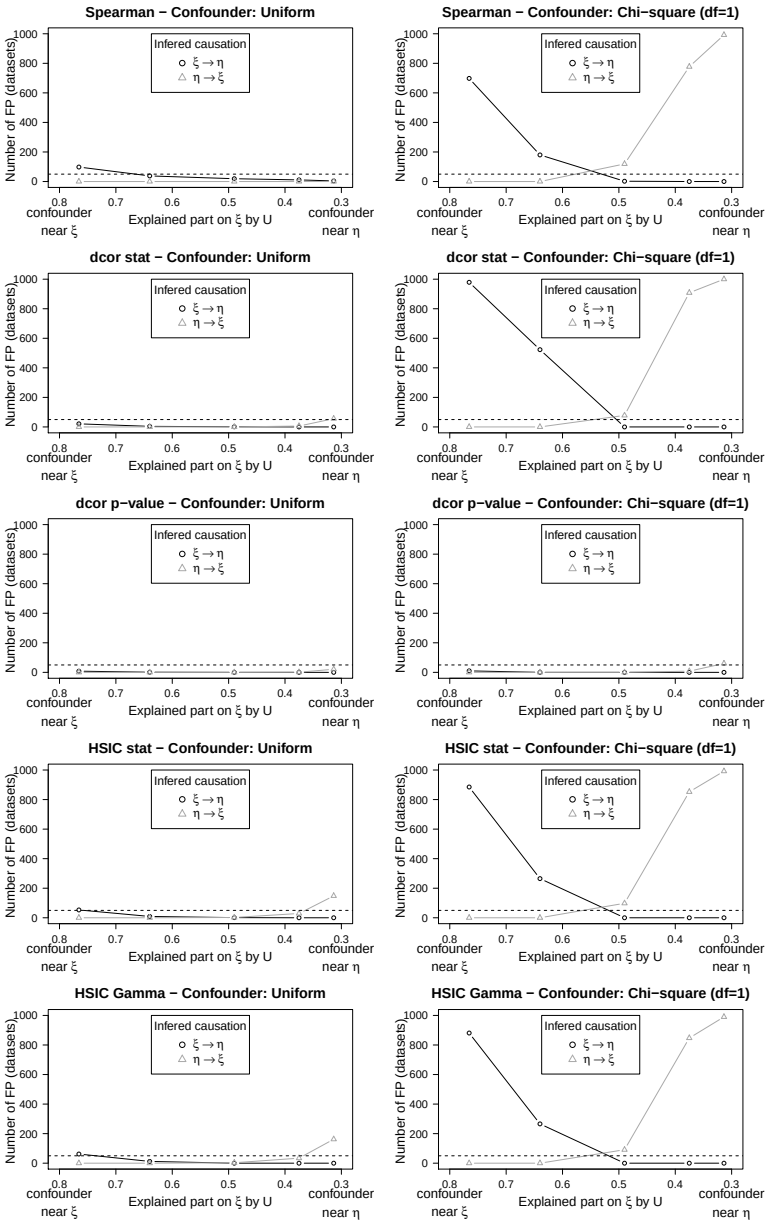

Fig. 4. Kinds of type I errors (FP) for each variant to measure dependence as a function of the distribution and the correlations of the latent confounder U (see Model 0 and variants 0a, 0b, 0c, 0d). Theoretical Pearson's correlation between ξ and η at the population level is always set equal to 0.49 in our simulation; sample size = 500, 1000 simulated datasets for each ten couples (5 models $\times$ 2 distributions of confounder)

- **Inclusion of an additional parameter allowing the user of LC to round to 0 each absolute difference of HSIC's p-values below a given specified threshold.** Whereas the variant based on differences of HSIC's gamma-approximated p-values gives some very good results, differences of bootstrap approximated dCor's p-values show more robustness in the presence of some specific latent confounders. However, about HSIC's gamma-approximated p-values, we remark that wrong conclusions in the robustness analysis come with a small persistent bias in the difference of p-values. So, rounding to 0 the very small observed differences in HSIC's gamma approximated p-value could improve its robustness.
- **Relaxing some assumptions of LC and comparison with other algorithms.** For instance, relaxing some assumptions, LC could be compared with parts of Cai et al. (2019)'s method. More widely speaking, future works should study the performances of LC under relaxed assumptions.
- **Manipulation of other parameters and additional comparisons using other models.** For instances, in our models, some distributions could be changed and other models could be considered. For instance, whereas Model 0 and Model 1 differed by the causal pattern and by the marginal distributions assigned to latent ξ and η respectively, an alternative for Model 0 would be to set exactly the same distributions for ξ and η than in Model 1 but with a specified theoretical correlation of 0. Furthermore, other models should also include confounder and causation together.
- **Presence of an observed confounder.** Corrections in each bootstrapped dataset could be included in LC to increase the accuracy.

Last but not least, future work should also test LC on real data as benchmark.

References

1. Bollen K. A. & Pearl, J.: Eight Myths about Causality and Structural Equation Models. In S.L. Morgan (Ed.), Handbook of Causal Analysis for Social Research, Chapter 15, 301-328. Springer. (2013) www.springer.com/lncs. Last accessed 24 Aug 2020
2. Cai, R., Xie, F., Glymour, C., Hao, Z., & Zhang, K.: Triad Constraints for Learning Causal Structure of Latent Variables. NeurIPS. (2019)
3. DDA Project Homepage, https://www.ddaproject.com/. Last accessed 14 July 2020
4. Eberhardt, F.: Introduction to the foundations of causal discovery. International Journal of Data Science and Analytics, **3**, 81–91 (2017)
5. Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. A Kernel Statistical Test of Independence. In Advances in Neural Information Processing Systems, **20**, 585–592. (2008)
6. Guyon, I.: Results and analysis of the 2013 ChaLearn cause-effect pair challenge. In Proceedings of NIPS 2013 Workshop on Causality: Large-scale Experiment Design and Inference of Causal Mechanisms (2014)
7. Husson F., Lê S., Pagès J. Analyse de données avec R, Rennes : Presses Universitaires de Rennes (2009).

8. Kline R. B. Principles and Practice of Structural Equation Modeling. (3rd ed.) New-York : The Guilford Press. (2011)
9. Kummerfeld, E., Ramsey, J.: Causal Clustering for 1-Factor Measurement Models. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1655-1664. ACM, 2016.
10. Le S., Josse J., Husson F.. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, **25**(1), 1-18. (2008). 10.18637/jss.v025.i01
11. MacCallum, R. C., & Austin, J. T.: Applications of Structural Equation Modeling in Psychological Research. Annual Review of Psychology, 51, 201–226. (2000)
12. Peters, J., Janzing, D., & Schölkopf, B.: Elements of causal inference - Foundations and learning algorithms. Cambridge, Massachusetts: MIT Press. (2018)
13. Pfister N. and Peters J. dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion. R package version 2.1. (2019). https://CRAN.R-project.org/package=dHSIC
14. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
15. Rizzo M. and Szekely G. energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-7. (2019). https://CRAN.R-project.org/package=energy
16. Sen, A., & Sen, B. Testing independence and goodness-of-fit in linear models. Biometrika, **101**, 927–942. (2014).
17. Shimizu S., Hoyer P. O., and Hyvärinen A.. Estimation of linear non-Gaussian acyclic models for latent factors. Neurocomputing, 72, 2024-2027. (2009)
18. Silva R., Scheine R., Glymour C., and Spirtes P.: Learning the Structure of Linear Latent Variable Models. Journal of Machine Learning Research, 7, 191-246. (2006).
19. Spirtes, P., Zhang, K.: Causal discovery and inference: concepts and recent methodological advances. Applied Informatics, **3**(3), 1–28 (2016)
20. Szekely, G. J., Rizzo, M. L., & Bakirov, N. K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics, **35**(6), 2769–2794 (2007)
21. Thoemmes, F.: The assumptions of direction dependence analysis. Multivariate Behavioral Research, 1-7. (2019)
22. Wiedermann W. & Li X.: Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS. Behavior Research Methods, **50**, 1581–1601 (2018)
23. Wiedermann, W., Merkle, E. C., & von Eye, A.: Direction of dependence in measurement error models. British Journal of Mathematical and Statistical Psychology, **71**(1), 117–145. (2018)
24. Wiedermann, W., Sebastian, J.: Direction Dependence Analysis in the Presence of Confounders: Applications to Linear Mediation Models Using Observational Data. Multivariate Behavioral Research, (2019a)
25. Wiedermann, W., Sebastian, J.: Sensitivity Analysis and Extensions of Testing the Causal Direction of Dependence: A Rejoinder to Thoemmes (2019), Multivariate Behavioral Research. (2019b)