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

Arnaud Pollaris[1] & Gianluca Bontempi[1]

[1] *MLG, Departement d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium*

I. 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 (confounder case)
4. a common descendant of A and B is kept fixed in the observed dataset.

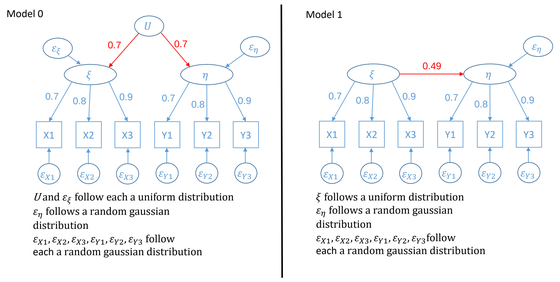
With a focus on the 3 first cases, assuming linearity and non Gaussian distributions, we introduce a new algorithm "Latent Causation" (LC) to help to infer a causal direction in a pair of latent variables (for each of them, observed children variables called "indicators" are available).
Assuming the measurement model (i.e. structure of relations between latents variables and its indicators) is already known, we present the state-of-the-art Direction Dependent Analysis (DDA) project [1], discuss about its limitations for pairs of latent variables and propose the LC algorithm. The goal of this research is twofold:

1. to compare LC and the DDA independence component applied on factor scores.
2. to provide an additional sensitivity study of LC.

II. Problem setting

Let us consider 2 continuous linearly correlated latent variables ξ and η and their indicators (which are linear functions of a latent variable + an independent noise).

Here is an instance of a causal and a confounding topology we want to discriminate between. Standardized values and distributions specified here are possible instances used below as assignations for parameters in our simulations. The number of observed indicators (rectangular boxes) can also differ.



III. State-of-the-art and limitations of the DDA

The DDA project regroups techniques for causal inference considering 3 components: 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.
Because of the confounder case, this last component is of the first importance [2]. Indeed, it relies on a combination of 2 statistical tests of independence and 4 conclusions are then possible: rejection of both independences (i.e., suspicion of confounder), no rejection of both independences, and rejection of only one of the two independences (i.e. causal direction).
However, concerning causal inference in pairs of latent variables, we address some limitations of the current DDA approach:

1. **DDA does not exploit all the available information in measures of dependencies**. First, a non-significant result for a test of independence is recommended while a lack of independence rejection is not a proof of independence. Second, the continuum of dependencies suffers from a discretization in "independence" VS "dependence" categories while a whole range of dependence strengths could be exploited to infer causation, even when both independences are rejected.
2. **Limit of the DDA distinction between "presence of a confounder" and "causation"**. Some theoretical models can include both causation and confounder. See e.g. the model 1 where the latent cause ξ can also be considered itself as a latent confounder between the two groups of observed indicators: $(X1, X2, X3)$ and $(Y1, Y2, Y3)$. So, maybe it would be better not to rely on the DDA independence component if we want to confirm there is no additional confounder when it concludes in favor of a causation.

So, even if LC shares important features with the DDA independence component, LC differs by focusing on direct comparisons of the strength of each dependence (one for each possible causal direction) to try to infer a causal direction (if possible and with an associated level of confidence).

IV. Latent Causation algorithm

Input:

- An observed dataset with indicators divided in 2 predefined 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 of (f) & (e)
3. Based on the B saved differences (in 2g), select a percentile conf. interval based on proba.: $(\alpha/2\ ;\ 1-\alpha/2)$.
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 $\xi \to \eta$"
   - Else: "data do not allow to conclude."

V. Results: LC VS DDA

1000 datasets were generated for each models (0 and 1) (see above for additional parameter specifications). Before to apply the DDA independence component, factor scores representing ξ and η are first computed in a similar way to LC, 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).
The table below shows the comparison of DDA VS LC in terms of accuracy (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)	Confounder (FN)	DNC (FN)	$\xi \to \eta$ (FP)	$\eta \to \xi$ (FP)	Confounder (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	

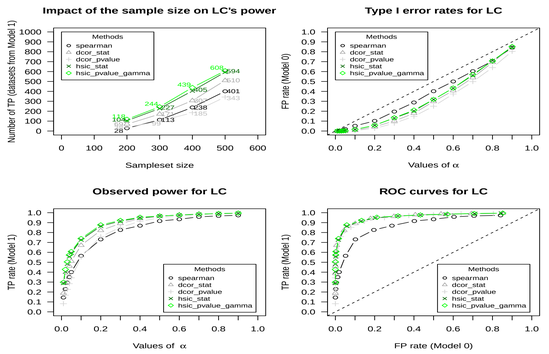
Parameters specification:
- **For both DDA and LC:** 1000 samples generated for Model 1 and 1000 samples for Model 0; sample size=500; α=0.05
- **For DDA only:** the number of replicates used for the estimation of each dCor-pvalue and the number of resamples used to compute each bootstrap's HSIC p-value 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.

We observe:

- Concerning DDA for Model 0 using HSIC gamma p-value or HSIC p-value bootstrap 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.
- 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 get more TP (i.e. correct causal direction) in Model 1 than DDA's dCor which is the only considered DDA variant with a FP rate below $\alpha = 5\%$

VI. Results: LC sensitivity study

We also compare 5 LC variants (differing by their measurement of the dependencies):



The figure beside shows:

- **upper left:** the sample size increases LC's ability to retrieve the true causation $\xi \to \eta$ in Model 1 for all 5 variants.
- **upper right:** as expected for Model 0, the observed proportion of FP is always lower than the specified value of α (whatever the variant of LC or the value of α)
- **lower left:** the number of correct causal directions (TP) increases with higher values of α. Notably, using HSIC's gamma-approximated p-values seems to get more TP compared with other observed variants.
- **lower right:** ROC curves show that the variants (apart from Spearman correlation) present similar good abilities to discriminate between Model 1 (causation) and Model 0 (confounder).

N.B.: inspired from Model 0, additional results involving manipulations of a latent confounder U (variations of its distribution and of its correlations with ξ and η) are also presented in the article. While the distribution of U and its correlations can influence the risk of FP, the dCor p-value variant shows a better robustness than the 4 other methods in our simulations.

VII. Conclusions

For causal inference in a pair of latent variables, 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). However, results presented here are only preliminar and the current work needs to be extended. It can concern e.g.:

- alternatives to compute factor scores
- inclusion of an additional parameter to neglect smallest differences when dependencies are compared
- the way to relax some assumptions of LC to make it more usable
- additional analysis involving other models or real data

VIII. References

[1] DDA Project Homepage, https://www.ddaproject.com/. Last accessed 14 July 2020

[2] Wiedermann, W., Sebastian, J.: Direction Dependence Analysis in the Presence of Confounders: Applications to Linear Mediation Models Using Observational Data. Multivariate Behavioral Research (2019)