Detecting Low-Complexity Confounders from Data

Maria Virginia Ruiz Cuevas [1 2] **Nataliya Sokolovska** [1] **Pierre-Henri Wuillemin** [3] **Jean-Daniel Zucker** [4]

Abstract

Statistical dependencies between two variables X and Y indicate that either X causes Y, or Y causes X, or there exists a latent variable Z which influences X and Y. In biology and medicine, an important problem is to find genetic or environmental unobserved causes of phenotypic difference between individuals. In this contribution, we introduce a novel approach to identify unobserved confounders in data. The proposed method is based on the state-of-the-art 3off2 causal network reconstruction algorithm, and on an evidence for a direct causal relation represented by purity of conditionals. The proposed method is implemented in Python, and it will be publicly available shortly. We discuss the results obtained on a real biomedical dataset.

1. Introduction

The decision what is a *"cause"* and what is an *"effect"* is generally made taking into account several principles based on moral aspects, background knowledge about normal cases to identify deviations, subject matter and importance. All these premises may not be found in observed data, making them insufficient (Spirtes et al., 2000; Pearl, 2000). In real-life applications, we decide what is a cause and what is an effect from static non-temporal data. Moreover, we aim to verfiy whether two observed variables are intermediated by an unobserved confounder.

We reconstruct a network using the 3off2 method, and we test the presence of latent confounders by conditions proposed by (Janzing et al., 2011). The 3off2 algorithm is a recently developed algorithm (Affeldt et al., 2016) which merges the principles of constraint-based and score-based approaches to reconstruct the causal graphical model from data. This algorithm exchanges the conditional independence tests (*i.e.* χ^2 and G^2) for a more robust raking function based on information theory. In particular, this algorithm does not try to directly find the conditional independence between variables: first, it tries to iteratively find the set z_n of the most important indirect contributors for each pair of nodes (x, y). For each contributor z_i found, the 3-point information $I(x; y; z_i)$[1] is subtracted or "taken off" from the 2-point information $I(x; y)$[2] eventually resulting in $I(x; y|z_n)$[3].

2. The 3off2 Algorithm: Learning Skeleton and Edge Orientation

The nodes (*x,y*) are conditionally independent if $I(x; y|z_n)$ is lower than 0, and in this case the edge will be removed from the graph. However, this is only valid for an infinite dataset ($N \to \infty$). In *real life*, there is not enough data available, so it is necessary to compare the mutual information to a factor *K* of $O(\frac{logN}{N})$ before drawing any conclusions on the conditional independence between x and y. This factor tries to limit the complexity of the model by favoring fewer edges while taking into account the finite data size. Then, the resulting *skeleton* is oriented based on the sign of the conditional 3-point information of unshielded colliders (Affeldt et al., 2016). The 3off2 has three main steps that are described below:

• **Step 0**. *Initiation*: the 3off2 starts with a complete undirected graph. Then, for each pair of nodes (*x,y*), the mutual information *I*(*x,y*) is calculated and compared to the factor *K* which is estimated using the Minimal Description Length (MDL) or the Normalized Maximum Likelihood (NML) score. If *I*(*x,y*) is inferior to $K_{x;y} / N$, the edge is removed. As in the PC algorithm, each pair of nodes (*x,y*) is associated to their individual and initially empty list - called separation set - that will contain all nodes that might justify the independence between *x* and *y*. In the case where *I*(*x,y*) is inferior to $K_{x;y} / N$, this set stays empty meaning that (*x,y*) are directly independent. Otherwise, the algorithm searches

*Equal contribution [1]Paris Sorbonne University, INSERM, France [2]IRIC, Canada [3]Paris Sorbonne University, LIP6, France [4]IRD, France. Correspondence to: Nataliya Sokolovska <nataliya.sokolovska@upmc.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1] $I(x; y; z_i) = H(x) + H(y) + H(z) - H(x, y) - H(x, z) - H(y, z) + H(x, y, z)$ where H is the Shannon entropy

[2] $I(x; y) = H(x) + H(y) - H(x, y)$

[3] $I(x; y|z_n) = I(x; y) - I(x; y; z_1) - I(x; y; z_2|z_1) - ... - I(x; y; z_n|z_{n-1})$

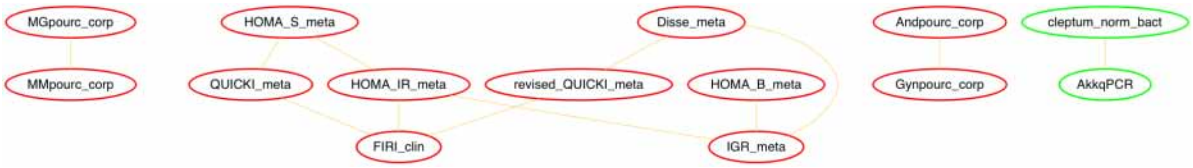

Figure 1. A graph illustrating dependencies between clinical parameters and bacteries constructed from the MicroObese data.

for the first contributor z within all nodes in the graph by computing the rank $R(x, y; z|\varnothing)$. The result of this step is the *first skeleton* graph.

• **Step 1**. *Iteration*: at this stage, the algorithm tries to remove as many edges as possible from the *skeleton* obtained in step 0. To do so, the algorithm will iteratively try to find more contributors using the rank *R* of a node. This step will stop when there are no remaining links left in *R* with a probability superior to 0.5. Otherwise, the link (*x,y*) with the highest probability in *R* is used to calculate the conditional mutual information $I(x; y|z_n)$, which is compared to the factor $K_{x;y|z_n}/N$ to determine whether x and y are conditionally independent or not. If $I(x; y|z_n) < K_{x;y|z_n}/N$, then (*x,y*) is not an essential link in the graph and can be removed. The result of step 1 is the *final skeleton* graph, and the separation sets for each couple (x, y) in the graph.

• **Step 2**. *Orientation/Propagation*. This stage uses the skeleton and the separation sets obtained from the step 1. First, all the unshielded colliders (*x,y,z*) are sorted in decreasing order of $|I'(x; y; z_i|z_n)|$. Then, for each triplet in the ordered list, one of two rules is applied depending on the sign of $I(x; y; z_i|z_n)$.

Rule 0: for cases where $I(x; y; z_i|z_n) < 0$, if (*x,y,z*) does not have a diverging orientation (*i.e.* both edges are not divergent from the center node), the triplet will be oriented as $x \rightarrow z \leftarrow y$.

Rule 1: for cases where $I(x; y; z_i|z_n) > 0$, if (*x,y,z*) has a converging edge $x \rightarrow z - y$ (i.e. one of the edges converges towards the center node), the triplet will be oriented as $x \rightarrow z \rightarrow y$.

3. Identification of Latent Confounders

According to the Reichenbach's common cause principle (Peters et al., 2017), if two random variables X and Y are statistically dependent, then there exists a third variable Z that causally influences both. Here we suppose that Z does not coincide neither with X, nor with Y, and we aim to identify Z.

Pairwise pure conditionals (Janzing et al., 2011). The conditional distribution $P(Y|X)$ is said to be pairwise pure if for any two $x_1, x_2 \in \mathcal{X}$ the following condition holds. There is no $\lambda < 0$ or $\lambda > 1$ for which

$$\lambda P(Y|X = x_1) + (1 - \lambda)P(Y|X = x_2) \tag{1}$$

is a probability distribution.

The purity is defined (Janzing et al., 2011; Peters et al., 2017) by the following condition:

$$\inf_{y \in \mathcal{Y}} \frac{p(y|x_1)}{p(y|x_2)} = 0, \text{ for all } x_1, x_2 \in \mathcal{X}. \tag{2}$$

In practice, to decide whether the pairwise purity holds, we can estimate

$$\min_{y \in \mathcal{Y}} \frac{\hat{p}(y|x)}{\hat{p}(y|x')} \text{ for all } x, x'. \tag{3}$$

If the conditional distribution is pure, then the path between X and Y is not intermediated by a confounder Z.

4. Experiments

The MicroObes corpus (Cotillard et al., 2013) contains heterogeneous biomedical data of obese patients. The data set contains information about 49 patients hired and examined at the Pitié-Salpêtrière hospital, Paris, France. The clinical parameters include standard clinical measurements such as BMI, sex, insuline sensitivity, etc. We have also access to the abundance matrices of gut flora genes, namely, bacterial quantification (qPCR), and abundance of bacterial clusters (MGS) of individual patients. Figure 1 shows a graph reconstructed by our Python implementation of the 3off2 algorithm. Since the number of observations is quite small, the 3off2 fails to estimate the edge orientations. The red nodes are the clinical variables, and the green ones are the bacteria. We tested the purity of the conditionals, and we concluded that there is an unobserved common cause between the two bacteria, the Akkermansia and the Cleptum. Further analysis will be done by researchers doing pre-clinical research to give a meaning to this common cause.

5. Conclusion

We proposed and tested a novel approach to discover latent variables. Currently we are developing a method to identify the nature of the latent variables (e.g., their domain size), and to extend the method for multivariate cases.

References

Affeldt, Séverine, Verny, Louis, and Isambert, Hervé. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, pp. 12. BioMed Central Ltd, 2016.

Cotillard, A., Kennedy, S. P., Kong, L. C., Prifti, E., Pons, N., Chatelier, E. Le, Almeida, M., Quinquis, B., Levenez, F., Galleron, N., Gougis, S., Rizkalla, S., Batto, J.-M., Renault, P., consortium, ANR MicroObes, Doré, J., Zucker, J.-D., Clément, K., and Ehrlich, S. D. Dietary intervention impact on gut microbial gene richness. *Nature*, 500:585–588, 2013. URL doi:10.1038/nature12480.

Janzing, D., Sgouritsa, E., Stegle, O., Peters, J., and Schölkopf, B. Detecting low-complexity unobserved causes. In *UAI*, 2011.

Pearl, Judea. The art and science of cause and effect. *Causality: models, reasoning and inference*, pp. 331–358, 2000.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference*. The MIT Press, 2017.

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