Revealing causality between heterogeneous data sources with deep restricted Boltzmann machines

Nataliya Sokolovska, Karine Clément, Jean-Daniel Zucker

Information Fusion 50 (2019) 139–147

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Revealing causality between heterogeneous data sources with deep restricted Boltzmann machines

Nataliya Sokolovska[a,*], Karine Clément[a,b], Jean-Daniel Zucker[a,c]

[a] Sorbonne University, NutriOmics team, Paris, France
[b] Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Nutrition Department, Paris, France
[c] Research Institute for Development, UMI 209, UMMISCO, Bondy, France

ARTICLE INFO

Keywords:
Probabilistic deep models
Causal inference
Heterogeneous data sources
Principal component analysis

ABSTRACT

In a number of real life applications, scientists do not have access to temporal data, since budget for data acquisition is always limited. Here we challenge the problem of causal inference between groups of heterogeneous non-temporal observations obtained from multiple sources. We consider a family of probabilistic algorithms for causal inference based on an assumption that in case where X causes Y, $P(X)$ and $P(Y|X)$ are statistically independent. For a number of real world applications, deep learning methods were reported to achieve the most accurate empirical performance, what motivates us to use deep Boltzmann machines to approximate the marginal and conditional probabilities of heterogeneous observations as accurate as possible.

We introduce a novel algorithm to infer causal relationships between blocks of variables. The proposed method was tested on a benchmark of multivariate cause-effect pairs. We show by our experiments that our method achieves the state-of-the-art empirical accuracy, and sometimes outperforms the state-of-the-art methods. An important part of our contribution is an application of the proposed algorithm to an original medical data set, where we explore relations between alimentary patters, human gut microbiome composition, and health status.

1. Introduction

An open problem in many real world applications is causal inference between two variables from observational data in absence of time series. In particular, in the medical domain, inferring causal relations between health status of a patient and treatment, or between results of some medical tests and nutritional habits, can help to improve the quality of diagnostics, and also motivate to develop methods of personalised medicine.

Even a more challenging task which we consider in this paper, is causal inference between blocks or modules of variables. As mentioned by [1], high dimensionality can increase computational complexity, and also it can reduce the predictive accuracy. Moreover, nowadays, the case of bivariate relationships is much better studied. The case of high-dimensional variables is considered in [2,3] only, and both papers introduce the IGCI-related independence between probability distributions and are based on the trace condition. The identifiability via the trace condition is proved [2,4] for relations without noise, and no theory exists for noisy cases which are much more relevant for real-life applications.

The relationships between alimentary habits and our way of life, the composition of the human gut metagenome and our clinical status, are among the most important problems studied by interdisciplinary scientific groups nowadays [5]. The human gut microbiota has a direct influence on health and well-being [6]. The emerging field of metagenomics allows to explore the human microbiome, to propose and to test novel hypotheses related to a number of diseases [7]. Very recent publications in leading biological and medical journals reveal relationships between human gut microbiota and cancer [8], gut microbiota and inflammatory colon diseases [9], gut microbiota and heart diseases [10], and gut microbiota and nutritional habits [5].

Biological mechanisms of metabolic changes and drug effects are examples of tasks we can tackle. Nowadays, a number of drugs and treatments are prescribed to patients, however, their mechanisms of action usually remain unclear. Metformin which is today the most prescribed drug for the type 2 diabetes, is a typical example. Medical doctors prescribe the metformin, since there is some empirical evidence that it has beneficial effects on blood glucose level, and on some cardiovascular parameters [11], and this treatment is also relatively cheap. There are several hypotheses of mechanisms of the metformin action. One of them states that the drug meditates its antihyperglicemic effects by pathways

* Corresponding author at: Team NutriOmics, Sorbonne University, INSERM, 91 blvd de l'Hopital Paris 75013 France
E-mail address: nataliya.sokolovska@upmc.fr (N. Sokolovska).

https://doi.org/10.1016/j.inffus.2018.11.016
Received 29 November 2017; Received in revised form 23 November 2018; Accepted 28 November 2018
Available online 29 November 2018

in the liver [12]. However, recent research in gut microbiome confirms another hypothesis that the metformin acts through pathways in the gut [13].

Heterogeneous data are data containing multiple types of variables such as e.g., unstructured text documents, multi-lingual data, images, and audios. We are interested, in particular, to develop methods for real medical and biological applications, and we focus on biomedical heterogeneous data such as clinical parameters (sex, gender, BMI, etc.), and "omics" data (transcriptomics, proteomics, metagenomics, and lipidomics).

A topical question is whether a diet and environment have an impact on the gut flora genes, and whether the gut flora in its turn influences the insuline sensitivity, and other clinical parameters, and therefore, our general health status. It can also occur that the mechanisms of the mediation are quite different from these hypotheses. Our goals are:

- to propose a novel robust causal inference method, suitable for heterogeneous multiple data sources, taking into consideration that real biological data are always limited, due to high acquisition cost, and noisy;
- to validate our approach on a widely used (by the causal inference community) cause-effect benchmark, to visualise causal graphs which we obtain, and to discuss which hypotheses can be verified on an original rich biomedical data.

Modern data sets are often high dimensional, and huge causal network reconstruction can be computationally intractable. A number of causal inference methods are focused on a case of two variables only [14–16]. Note that methods based on Markov equivalent graphs [17,18] will fail in the bivariate case, since $X \to Y$ and $Y \to X$ are Markov equivalent.

A recently introduced 3off2 method [19] seeks for all possible triplets in data to identify colliders in a graph. The 3off2 can not be applied to a bivariate case, since it needs at least three variables to infer causality.

In this work, we consider bivariate causal discovery approaches which are based on probabilistic inference [14–16]. This family of methods relies on the following postulate: if a cause X impacts an effect Y, then the marginal distribution of the cause and the conditional distribution of the effect given the cause are independent. The probabilities are supposed to be estimated from the non-temporal observational data.

Deep learning methods [20] were reported to outperform the state-of-the-art supervised and unsupervised methods in terms of empirical accuracy in computer vision applications, speech (signal) processing, natural language processing, and biology [21,22]. The deep models were also reported to be efficient to combine multiple sources of data [23].

We decided to focus on neural networks, since the deep probabilistic classifiers model the probabilities which are the key elements of our approach. So, the conditional probability $P(Y|X)$ will be modelled by a last (supervised) layer of the deep network; and the $P(X)$ can be also approximated by the deep approaches [24,25].

Deep restricted Boltzmann machines (DRBM) were introduced by [24] as a multi-layer stochastic approach. The standard DRBM has one layer of observed units and an arbitrary number of hidden layers. If a task is supervised, then the deep architecture has an additional layer containing the classes. A recent work of [26] discusses an important problem of overfitting in deep learning [26]. propose an efficient method which combines unsupervised learning in RBM with a supervised sparsity-inducing regularizer to get the best from two worlds: unsupervised learning using cheap unlabeled data, and sparse supervised training to fine-tune the model.

To our knowledge, the only existing method of inferring causal directions between blocks of variables is the trace method [3] which can be unstable in the presence of noise.

Our contribution is multifold:

- We introduce an original robust method of inferring causal directions between blocks of heterogeneous variables.
- We provide all the details to approximate the conditional and marginal probabilities using the deep RBM.
- The results of our experiments on multivariate benchmark data confirm that the proposed algorithm is computationally efficient and its empirical performance is highly competitive compared to the state-of-the-art causal inference methods.
- We illustrate the interest to use the proposed method by a real medical problem of revealing causality in rich original heterogeneous data. In particular, our goal is to explore causal relationships between human gut composition, nutritional habits, and clinical status, i.e., clinical parameters indicating whether a patient is healthy or not.

Our work is organized as follows. We discuss the state-of-art methods of bivariate causal inference in Section 2. These methods can be naturally divided into continuous and discrete methods for causal inference, and we consider them in details in Section 3. We introduce our method of causal inference between modules of heterogeneous data in Section 4. Section 5 is dedicated to the deep restricted Boltzmann machines, and we also provide all the details to compute the conditional and marginal probabilities. In Section 6, we discuss the results of our experiments on the multivariate cause-effect pairs, and on an original medical problem. We share our conclusions and perspectives at the end of the paper.

2. Related work

In this work, we consider two families of methods of causal inference focusing on bivariate relations, namely, the Additive Noise Models (ANM) and Information Geometric Causal Inference (IGCI) [27]. For a more general overview of causal structure learning see [4,28].

Additive noise models (ANM) were originally introduced by [29,30]. The ANM design causal relations between two variables. The ANM model a dependency between a cause X and an effect Y given some noise E: $Y = f(X) + E$. According to the hypothesis behind the approach, if X and E are independent, then it is possible to infer causal direction, and it is assumed that X causes Y. The ANM are usually used for continuous data, and the approach can not be applied directly to the categorical variables [31]. However, there exist a number of extensions of the additive noise models, and, for instance, a generalisation called post-nonlinear models was proposed by [32].

The ANM is not the only method to exploit the asymmetry between the cause X and the effect Y. The linear trace method introduced by [3] states that the trace condition is fulfilled in the causal direction, if X causes Y, and $P(X)$ and $P(Y|X)$ are independent. Note that the trace condition does not hold in the opposite direction. Another method exploiting the asymmetry is the information-geometric causal inference (IGCI) [15]. It verifies whether the density of X and the log slope of the function which transforms cause to effect are not correlated. It was shown [33] that the density of Y and the log slope of the inverse of the function are correlated. Recently, a generalisation of the IGCI for non-deterministic cases was proposed by [34].

The trace condition is proved under the assumption that the covariance matrix is drawn from a rotation invariant prior [14]. Recently, the method was generalized for non-linear cases [35], and it was shown that the covariance matrix of the mean embedding of the cause in reproducing kernel Hilbert space is free independent with the covariance matrix of the conditional embedding of the effect given cause.

Origo [36] is a causal inference method that is based on a postulate [14] that the factorisation of the joint probability of cause and effect in the causal direction has lower complexity than in the anti-causal direction. The result of [14] is formulated in terms of Kolmogorov complexity. Origo [36] uses the Minimum Description Length (MDL) as an approximation of Kolmogorov complexity. Another recently published approach based on the same principles was proposed by [37], and it is called Slope. The weakness of Origo and of other MDL-based methods is that the MDL only approximates Komolgorov complexity, and involves

unknown metric errors. The empirical performance is highly related to the dataset, and Origo was reported to reach the state-of-the-art performance on the multivariate benchmarks (Acute inflammation, ICDM abstracts, Adult data set), however, it performs less accurate than the ANM on the univariate benchmark of cause-effect pairs with known ground truth (the Tübingen data set).

MIIC (Multivariate Information Inductive Causation) is a recently introduced algorithm [38] with competitive performance on real biological data. It is an information-theoretic approach of learning causal networks. The MIIC algorithm is an advanced version of the 3off2 method [19], and was reported to be efficient both in terms of empirical accuracy and in terms of computational time. However, the MIIC is not adapted for causal inference between groups of variables, and we will show by our experiments that it fails in this challenging task.

We are particularly motivated by the recent, reported to be efficient causality discovery methods (see, e.g., [14–16]) which are based on the postulate of the independence of mechanisms. It tells that a causal direction can be inferred from estimated marginal and conditional probabilities of two random variables from a non-temporal observational data set. Below, we investigate this research avenue.

3. Bivariate causal inference in discrete and continuous data

This section is dedicated to methods which are based on the following postulate [14–16] and assumptions.

Postulate 1. If $X \to Y$, then the marginal distribution of the cause $P(X)$ and the conditional distribution of the effect given the cause $P(Y|X)$ are "independent", i.e., the cause distribution and the mechanism producing the effect distribution are independent.

Assumption 1. X and Y are observed, and it is assumed that there are not any confounders, any selection bias, and no feedback.

Assumption 2. According to the Reichenbach's principle of common sense [4], if there exists a statistical dependency between two observable variables X and Y, it indicates that there exists a variable Z which causes X and Y. We assume here that Z coincides either with X or with Y what leads directly to inferring causality $X \to Y$ or $Y \to X$.

3.1. Bivariate causal inference with regression

CURE (Causal inference with Unsupervised inverse REgression) introduced by [16] is one of the methods which rely on the Postulate 1. The intuition behind the approach is as follows. The CURE infers "X causes Y" if the estimation of the conditional probability $P(X|Y)$ which is done from samples from $P(Y)$ is more accurate than the estimation in the opposite direction. A natural question is how to quantify the accuracy of the conditional probabilities, and [16] propose to compare the difference between the unsupervised and supervised log-likelihoods obtained from N pairs $\{X_i, Y_i\}_{i=1}^N$ of observations:

$$D_{X|Y} = \mathcal{L}^{\text{unsup}}_{X|Y} - \mathcal{L}^{\text{sup}}_{X|Y} = \tag{1}$$

$$-\frac{1}{N}\sum_{i=1}^{N} \log p\big(X_i|Y_i, \mathbf{y}\big) + \frac{1}{N}\sum_{i=1}^{N} \log p\big(X_i|Y_i, \mathbf{x}, \mathbf{y}\big), \tag{2}$$

and

$$D_{Y|X} = \mathcal{L}^{\text{unsup}}_{Y|X} - \mathcal{L}^{\text{sup}}_{Y|X} = \tag{3}$$

$$-\frac{1}{N}\sum_{i=1}^{N} \log p\big(Y_i|X_i, \mathbf{x}\big) + \frac{1}{N}\sum_{i=1}^{N} \log p\big(Y_i|X_i, \mathbf{x}, \mathbf{y}\big). \tag{4}$$

The conditional probability $p(X_i|Y_i, \mathbf{y})$ is estimated from Y observed but X are not observed, $p(Y_i|X_i, \mathbf{x})$ is estimated using observed X only, and $p(X_i|Y_i, \mathbf{x}, \mathbf{y})$ are computed when both X and Y are observed.

The edge orientation in the CURE is decided as follows. The causal direction is set to $X \to Y$, if $D_{X|Y} < D_{Y|X}$, otherwise $Y \to X$ is inferred. The CURE method is based on the Markov Chain Monte Carlo (MCMC) to approximate the posterior distributions what is computationally consuming, and what can be computationally intractable in many real applications. Therefore, we decided to consider methods which are based on discrete data analysis.

3.2. Causal discovery with distance correlation

Let us consider another approach to discover causal relations which is also based on the Postulate 1. The causal discovery using distance correlation to measure the distance between probability distributions was proposed by [39]. In their approach, it is assumed that both X and Y are discrete, and it this case, it is straightforward to present the probability distributions as tables. Let us consider the measure of the distance correlation and how it can help to infer causal direction in details.

The dependence measures are defined as follows:

$$D_{Y|X} = dCor(P(X), P(Y|X)) \tag{5}$$

$$D_{X|Y} = dCor(P(Y), P(X|Y)), \tag{6}$$

where $dCor\ (a,\ b)$ is the distance correlation.

We use the distance correlation as the independence measure [39]. The method to estimate an empirical distance correlation from data was originally proposed by [40], and we summarize it briefly. Given two random one-dimensional or high-dimensional variables a and b, the empirical distance covariance $C\ (a,\ b)$ is defined:

$$C(a, b) = \frac{1}{N}\sqrt{\sum_{i,j=1}^{N} \tilde{A}_{ij}\tilde{B}_{ij}}, \tag{7}$$

where

$$\tilde{a}_{ij} = \|a_i - a_j\|, \tilde{a}_{i\cdot} = \frac{1}{N}\sum_{j=1}^{N} a_{ij}, \tag{8}$$

$$\tilde{a}_{\cdot j} = \frac{1}{N}\sum_{i=1}^{N} a_{ij}, \tilde{a}_{\cdot\cdot} = \frac{1}{N^2}\sum_{i,j=1}^{N} a_{ij}, \tag{9}$$

and

$$\tilde{b}_{ij} = \|b_i - b_j\|, \tilde{b}_{i\cdot} = \frac{1}{N}\sum_{j=1}^{N} b_{ij}, \tag{10}$$

$$\tilde{b}_{\cdot j} = \frac{1}{N}\sum_{i=1}^{N} b_{ij}, \tilde{b}_{\cdot\cdot} = \frac{1}{N^2}\sum_{i,j=1}^{N} b_{ij}, \tag{11}$$

and the matrices $\tilde{A}$ and $\tilde{B}$ take the following form:

$$\tilde{A}_{ij} = \tilde{a}_{ij} - \tilde{a}_{i\cdot} - \tilde{a}_{\cdot j} + \tilde{a}_{\cdot\cdot}, \tag{12}$$

$$\tilde{B}_{ij} = \tilde{b}_{ij} - \tilde{b}_{i\cdot} - \tilde{b}_{\cdot j} + \tilde{b}_{\cdot\cdot}. \tag{13}$$

Then the distance correlation $dCor(a,\ b)$ is defined as follows [39,40]:

$$dCor(a, b) = \frac{C(a, b)}{\sqrt{C(a, a)C(b, b)}}, \tag{14}$$

and it can be computed directly from data. Note that $dCor(a, b) = 0$, if $C(a, a) = 0$ or $C(b, b) = 0$.

It is easy to see from the definition of the distance correlation that it can be directly computed from observational data. The algorithm of [39] uses ϵ to assess how close $D_{X|Y}$ and $D_{Y|X}$ are. If the absolute difference between distance correlations ϵ is too small, the causal direction can not be inferred. It was reported [39] that the empirical accuracy is acceptable for $\epsilon > 0.05$, and this value was suggested to be a reasonable choice.

4. Causal inference between groups of variables

Here we suppose that $\mathbf{X}$ and $\mathbf{Y}$ are multivariate variables. A matrix of observations $\mathbf{X}$ is a matrix of size $N \times p$, $\mathbf{Y}$ is also a matrix $N \times q$, where N is the number of data points, and p and q are the numbers of features of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

We propose to use an eigenstructure decomposition to transform original data into a data set suitable for a block analysis. An eigenvector is a weighted average of variables of a group, and we summarize the information of each group in an eigenvector. This idea is actively exploited by systems biology and bioinformatics communities [41].

Given matrices $\mathbf{X}$ and $\mathbf{Y}$, a Principal Component Analysis (PCA) will produce a derived set of variables which are not correlated

$$\bar{\mathbf{X}}_{\cdot m} = \mathbf{X}\alpha_m, \ m = 1, \dots, p', \ p' < p, \tag{15}$$

$$\bar{\mathbf{Y}}_{\cdot l} = \mathbf{Y}\beta_l, \ l = 1, \dots, q', \ q' < q. \tag{16}$$

that are linear combinations of the original data, and that explain most of the variation in the original set. $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the projections of the data onto the principal components, $\alpha_1, \dots, \alpha_{p'}$ are the eigenvectors of $\hat{\Sigma}_{\mathbf{X}}$, the sample covariance matrix of $\mathbf{X}$, and $\beta_1, \dots, \beta_{q'}$ are the eigenvectors of $\hat{\Sigma}_{\mathbf{Y}}$, the sample covariance matrix of $\mathbf{Y}$.

The original high-dimensional data are projected (here using Principal Component Analysis) to a low-dimensional space. The first Principal Component encodes information about all features of the original space, and also has the largest variance (among the principal components). Inferring causality between original data can be approximated by discovering causal relations between the data projected to the Principal Components since the PCA is an optimal low-rank approximation [42].

We consider here 4 schemes to decide the direction between the groups.

1. Majority vote

$$\begin{cases} \mathbf{X} \to \mathbf{Y}, \text{ if } \sum_{m,l} \mathbb{1}_{\{\bar{\mathbf{X}}_m \to \bar{\mathbf{Y}}_l\}} > \sum_{m,l} \mathbb{1}_{\{\bar{\mathbf{Y}}_l \to \bar{\mathbf{X}}_m\}}, \\ \mathbf{Y} \to \mathbf{X}, \text{ otherwise.} \end{cases} \tag{17}$$

2. Majority vote weighted by the order of principal components

$$\begin{cases} \mathbf{X} \to \mathbf{Y}, \text{ if } \sum_{m,l} \frac{1}{p'}\frac{1}{q'} \mathbb{1}_{\{\bar{\mathbf{X}}_m \to \bar{\mathbf{Y}}_l\}} > \sum_{m,l} \frac{1}{p'}\frac{1}{q'} \mathbb{1}_{\{\bar{\mathbf{Y}}_l \to \bar{\mathbf{X}}_m\}}, \\ \mathbf{Y} \to \mathbf{X}, \text{ otherwise.} \end{cases} \tag{18}$$

3. Majority vote weighted by significance of causal decision

$$\begin{cases} \mathbf{X} \to \mathbf{Y}, \text{ if } \sum_{m,l} \mathbb{1}_{\{\bar{\mathbf{X}}_m \to \bar{\mathbf{Y}}_l\}} \ |D_{\mathbf{X}|\mathbf{Y}} - D_{\mathbf{Y}|\mathbf{X}}| > \\ \qquad\qquad \sum_{m,l} \mathbb{1}_{\{\bar{\mathbf{Y}}_l \to \bar{\mathbf{X}}_m\}} |D_{\mathbf{X}|\mathbf{Y}} - D_{\mathbf{Y}|\mathbf{X}}|, \\ \mathbf{Y} \to \mathbf{X}, \text{ otherwise.} \end{cases} \tag{19}$$

4. Double weighted majority vote

$$\begin{cases} \mathbf{X} \to \mathbf{Y}, \text{ if } \sum_{m,l} \frac{1}{p'}\frac{1}{q'} \mathbb{1}_{\{\bar{\mathbf{X}}_m \to \bar{\mathbf{Y}}_l\}} \ |D_{\mathbf{X}|\mathbf{Y}} - D_{\mathbf{Y}|\mathbf{X}}| > \\ \qquad\qquad \sum_{m,l} \frac{1}{p'}\frac{1}{q'} \mathbb{1}_{\{\bar{\mathbf{Y}}_l \to \bar{\mathbf{X}}_m\}} |D_{\mathbf{X}|\mathbf{Y}} - D_{\mathbf{Y}|\mathbf{X}}|, \\ \mathbf{Y} \to \mathbf{X}, \text{ otherwise.} \end{cases} \tag{20}$$

In many applications, and in biological applications in particular, we have multiple data sources, and usually their number is bigger than 2. Algorithm 1 generalizes the bivariate approach which we presented for two heterogeneous data sources only, $\mathbf{X}$ and $\mathbf{Y}$. Let us consider that we have K sources of information, and K matrices $\mathbf{X}^k$, $k \in \{1, \dots, K\}$ generated by them. We propose a procedure to establish causal relationships between the multiple data sources applying the pairwise causal inference. The learning procedure is drafted as Algorithm 1.

5. Deep restricted boltzmann machines

A deep restricted Boltzmann machine (DRBM) proposed by [24] is an energy-based model which contains a layer of observed variables $\mathbf{v} \in \{0, 1\}^D$, and an arbitrary number of layers of hidden units $\mathbf{h} \in \{0, 1\}^P$. A supervised DRBM includes also an output layer; in case of a binary task, this layer has two units. We aim to model conditional and marginal probabilities as accurate as possible, and we are not interested to design a joint probability distribution of X and Y. It explains our choice to use the DRBM which is an undirected graphical model and a special case of Markov random fields. It allows to model $P(Y|X)$ directly. In this contribution, we do not consider Deep Belief Networks (DBN) which can efficiently design $P(Y, X)$.

Algorithm 1 Causal Inference between Multiple Sources of Heterogeneous Observations.

Input: Matrices of observations $\mathbf{X}^k$ issued from K data sources

Output: Causal directions between $\mathbf{X}^i$ and $\mathbf{X}^j$ for all $i, j = \{1, \dots, K\}$

STEP 1: Perform PCA for all data sources $\mathbf{X}^k$ for $k \in \{1, \dots, K\}$

STEP 2: Project $\mathbf{X}^k$ on principal components and get

$$\bar{\mathbf{X}}^k \text{ for } k \in \{1, \dots, K\}$$

STEP 3: Use deep restricted Boltzmann machines to approximate the probabilities

$$P(\bar{\mathbf{X}}^j_{\cdot m}) \text{ and } P(\bar{\mathbf{X}}^i_{\cdot l} | \bar{\mathbf{X}}^j_{\cdot m}), \tag{20}$$
$$P(\bar{\mathbf{X}}^i_{\cdot l}) \text{ and } P(\bar{\mathbf{X}}^j_{\cdot m} | \bar{\mathbf{X}}^i_{\cdot l}), \text{ for all } l, m, i, j \tag{20}$$

STEP 4: Infer causal directions
Using one of the criteria eq. (17) – eq. (20)

With model parameter w, the energy of state $(\mathbf{v}, \mathbf{h})$ is given as follows:

$$\mathrm{E}(\mathbf{v}, \mathbf{h}, w) = -\mathbf{v}^T w \mathbf{h}, \tag{21}$$

$$p(\mathbf{v}, w) = \frac{1}{Z(w)} \sum_{\mathbf{h}} \exp(-\mathrm{E}[\mathbf{v}, \mathbf{h}, w]), \tag{22}$$

$$Z(w) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-\mathrm{E}[\mathbf{v}, \mathbf{h}, w]). \tag{23}$$

The conditional distributions over observed and hidden variables are defined as:

$$p\big(h_j = 1 | \mathbf{v}, \mathbf{h}_{-j}\big) = \sigma\left(\sum_{i=1}^{D} w_{ij} v_i\right), \tag{24}$$

$$p\big(v_i = 1 | \mathbf{h}, \mathbf{v}_{-i}\big) = \sigma\left(\sum_{j=1}^{P} w_{ij} h_j\right), \tag{25}$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{26}$$

where σ is the logistic (sigmoid) function. The first derivative is used in the optimisation procedure, and it takes the following form:

$$\Delta w = \alpha\big(\mathrm{E}_{Pdata}\big[\mathbf{v}\mathbf{h}^T\big] - \mathrm{E}_{Pmodel}\big[\mathbf{v}\mathbf{h}^T\big]\big), \tag{27}$$

where α is the learning rate. The first term of the gradient is the expectation with respect to the completed data distribution, and the second term is the expectation with respect to the distribution defined by the model.

In case of a DRBM which contains two hidden layers, the energy of a state can be computed as follows:

$$\mathrm{E}\big[\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, w\big] = -\mathbf{v}^T w^1 \mathbf{h}^1 - \mathbf{h}^1 w^2 \mathbf{h}^2, \tag{28}$$

where $w = \{w^1, w^2\}$ are the parameters (weights associated with the layers) of the model, and

$$p(\mathbf{v}, w) = \frac{1}{Z(w)} \sum_{\mathbf{h}^1, \mathbf{h}^2} \exp\left(-\mathrm{E}\left[\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, w\right]\right). \tag{29}$$

The conditional distributions over the hidden and observed layers are defined:

$$p\left(h_j^1 = 1 | \mathbf{v}, \mathbf{h}^2\right) = \sigma\left(\sum_i w_{ij}^1 v_i + \sum_m w_{jm}^2 h_j^2\right), \tag{30}$$

$$p\left(h_m^2 = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_i w_{im}^2 h_i^1\right), \tag{31}$$

$$p\left(v_i = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_j w_{ij}^1 h_j^2\right). \tag{32}$$

Pre-training. Pre-training is a step of the learning procedure to initialise the weights with some reasonable values. The greedy layerwise pre-training [43] is a standard method. It consists in learning stacked restricted Boltzmann machines in an unsupervised manner. It was demonstrated [24,43] that the pre-training procedure fastens the supervised training. To initialise the weights, we set:

$$p\left(h_j^1 = 1 | \mathbf{v}\right) = \sigma\left(\sum_i w_{ij}^1 v_i + \sum_i w_{ij}^1 v_i\right), \tag{33}$$

$$p\left(v_i = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_j w_{ij}^1 h_j\right), \tag{34}$$

$$p\left(h_j^1 = 1 | \mathbf{h}^2\right) = \sigma\left(\sum_m w_{jm}^2 h_m^2 + \sum_m w_{jm}^2 h_m^2\right), \tag{35}$$

$$p\left(h_m^2 = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_j w_{jm}^2 h_j^1\right), \tag{36}$$

where the input is doubled to get rid of the double-counting problem when we perform top-down and bottom-up inferences. Putting together the Eqs. (33) – (36), we get:

$$p\left(h_j^1 = 1 | \mathbf{v}, \mathbf{h}^2\right) = \sigma\left(\sum_i w_{ij}^1 v_i + \sum_m w_{jm}^2 h_m^2\right). \tag{37}$$

Training. Let

$$F(\mathbf{v}) = -\log \sum_{\mathbf{h}^1, \mathbf{h}^2} \exp\left(-E\left[\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2\right]\right), \tag{38}$$

and

$$\frac{-\partial \log p(\mathbf{v}, w)}{\partial w} = \frac{\partial F(\mathbf{v})}{\partial w} - \sum_{\tilde{v}} p(\tilde{v}) \frac{\partial F(\tilde{v})}{\partial w}. \tag{39}$$

During the training of the DRBM, the first term of Eq. (39) is supposed to increase the probability of training data (the so-called positive phase). The second term of the equation decreases the probability of generated samples (negative phase). The expectation over all possible configurations which is the second term of the gradient, can be computationally intractable in many real applications. A sampling method such as MCMC can be applied to approximate it. Another way to compute the approximation, is to use Annealed Importance Sampling (AIS) with variational inference [24,25]. The approximation takes the following form:

$$\mathrm{E}_p\left[\frac{\partial F(\mathbf{v})}{\partial w}\right] = \frac{1}{|\mathcal{V}|} \sum_{\tilde{v} \in \mathcal{V}} \frac{\partial F(\tilde{v})}{\partial w}, \tag{40}$$

where $\tilde{v} \in \mathcal{V}$ are generated samples.

Prediction. In a supervised learning scenario, the classification is done as follows [44]:

$$p(y|\mathbf{v}) = \sum_{\mathbf{h}} p(y|\mathbf{h}) p(\mathbf{h}|\mathbf{v}) = \mathrm{E}_{p(\mathbf{h}|\mathbf{v})} p(y|\mathbf{h}). \tag{41}$$

6. Experiments

In this section, we show the results of our experiments on a standard benchmark which includes multivariate cause-effect pairs, and on an original biomedical data set gathered and maintained at the Pitié-Salpêtrière hospital, Paris, France.

6.1. Benchmark of multivariate cause-Effect pairs

The causal inference benchmark which is publicly available from http://webdav.tuebingen.mpg.de/cause-effect (version 1.0) contains 100 pairs of causes and their effects from different scientific domains. The ground truth is given. For our experiments, we extract the cause-effect pairs which are multivariate problems, namely, pairs 52–55, and 71.

- Pair 52. X and Y are both four-dimensional variables for day 51 and 50 of year 2000. The measurements are temperature, pressure, sea level pressure, and relative humidity. The ground truth tells that day 50 influences day 51.
- Pair 53. The task is to verify whether wind speed, global radiation, and temperature cause changes in ozone concentration.
- Pair 54. The data concerns city-cycle fuel consumption. The X group includes displacement, horsepower, and weight of a car, and Y variable contains information about mpg and acceleration. The ground truth tells than X causes Y.
- Pair 55. The problem is to reveal causality between ozone values and temperature at 16 different places. The ground truth states that the temperature has an impact on the ozone.
- Pair 71. The task concerns an acute inflammation of urinary bladder. From the clinical literature it is known that such symptoms as temperature of patient, occurence of nausea, lumbar pain, urine pushing, micturition pains, and burning of urethra predict the state of a patient which can be either inflammation of urinary bladder, or nephritis.

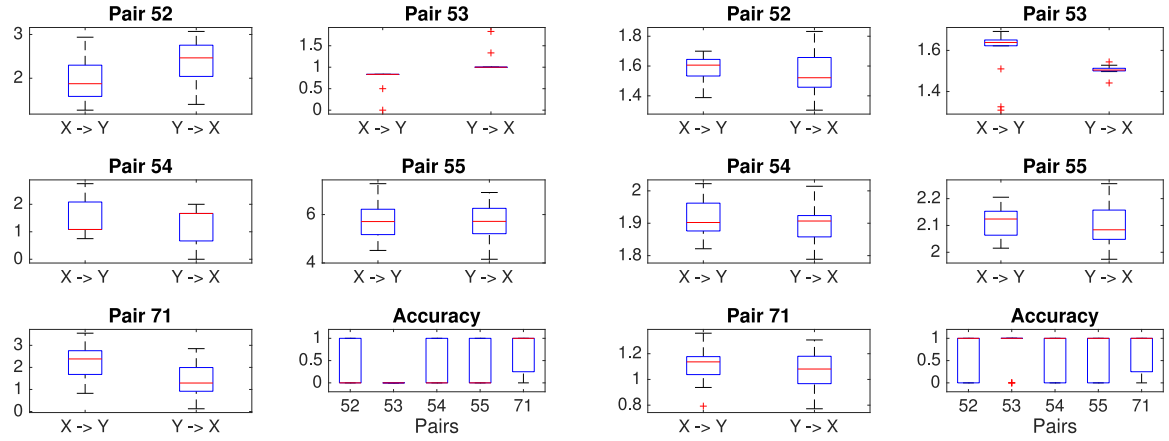
Here, to simplify the analysis of the obtained results, we read the data in such a way that the ground truth is $X \to Y$ for all pairs. So, ideally, we expect that for all considered cause-effect pairs the distribution $X \to Y$ contains bigger values then $Y \to X$.

First, we perform a PCA on the original X and Y data, and consider their projections for the further causal analysis. The data are discretized using the equal frequency algorithm into 5 bins. The choice of deep architecture is an engineering problem. We fixed the number of hidden layers and the number of latent variables in them by 10-fold-cross validation. The best empirical accuracy on the validation set was obtained using 3 hidden layers, each containing 5 units. In our experiments, 25 epochs were enough to converge to an optimal solution.

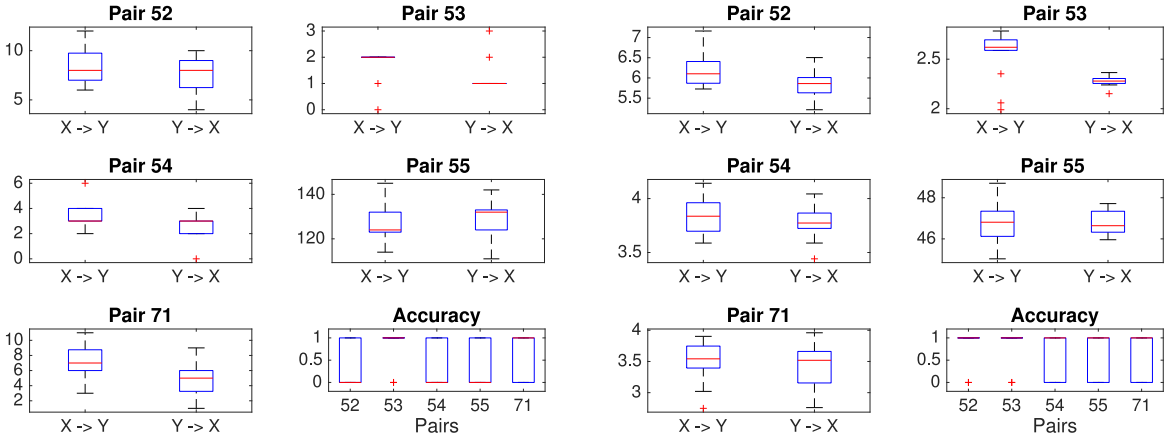
We test the criteria Eq. (17) – Eq. (20), and boxplot the corresponding distributions. Fig. 1 shows the results for the four tested criteria. To compare our approach to the state-of-the-art, we consider the performance of the MIIC approach, and apply it also to the projections on the principal components with Eq. (17) – Eq. (18). The criteria Eq. (19) – Eq. (20) can not be tested with the MIIC, since the causal directions are not quantified. The output of the MIIC algorithm is a direction $\to$ or $\leftarrow$ between two egdes. In case where the MIIC did not infer any direction between two edges, the edge is undirected. The results obtained with the MIIC are illustrated on Fig. 2, and we see that for some pairs the values are equal to 0 what means that the MIIC did not provide any direction.

The question whether the prediction is stable or not is very important. In the experiments, we split the data into 10 parts, and in each run we use 9 parts to infer the causal directions, and we repeat the experiment 10 times.

To answer the important question about the accuracy and the number of bins, we tested various numbers of bins, and we also tested the logistic regression to estimate the probability distributions. Fig. 3 shows

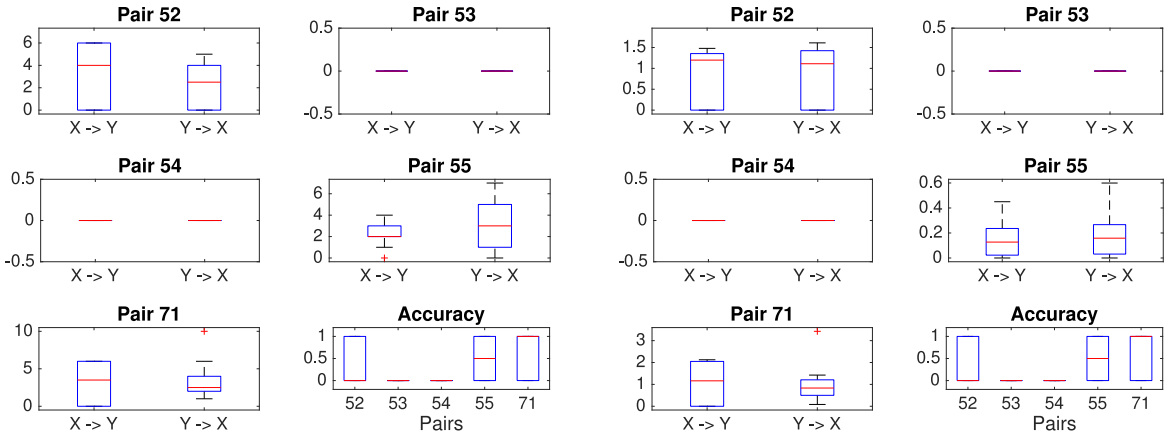


(a) Criterion eq. (17) (b) Criterion eq. (18)



(c) Criterion eq. (19) (d) Criterion eq. (20)

Fig. 1. Multivariate cause-effect pairs. Performance of the proposed criteria for causal inference: obtained distributions and accuracy.



(a) Criterion eq. (17) (b) Criterion eq. (18)

Fig. 2. Performance of the MIIC algorithm on the multivariate cause-effect pairs.

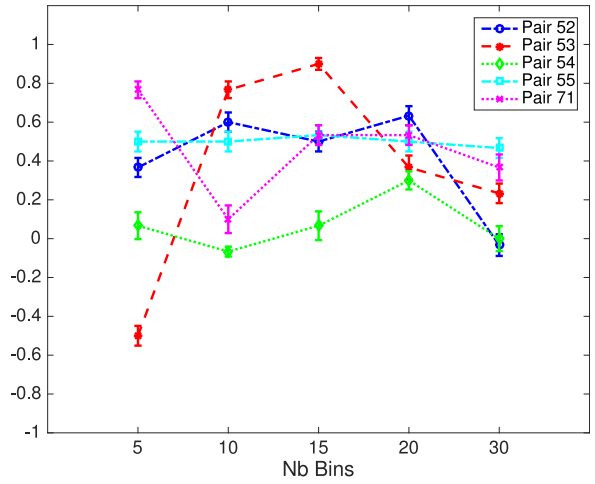


Fig. 3. The difference between the predictive accuracies where the probabilities are estimated by the DRBM and by the logistic regression for five multi-variate pairs of the cause-effect data set.

the difference between the predictive accuracies of causal directions using the DRBM and the logistic regression for five multi-variate pairs of the cause-effect data set. If the difference is positive, the DRBM estimates the probabilities more accurately than the logistic regression, and it leads to a more accurate prediction of causal directions. If the values are negative, these are the cases where the logistic regression performs better.

We observe that the criteria Eq. (18) and Eq. (20) which take into consideration the absolute difference and the order of the principal components, are quite accurate, and outperform others.

6.2. Microobes data

The MicroObes corpus [5] contains heterogeneous biomedical data of obese patients. Clinicians of the NutriOmics team, Pitié-Salpêtrière hospital, Paris, France, examined 49 patients. Parameters which can be called environmental, include alimentary patterns reflecting nourishing habits of subjects, and also information about their physical activity. The host data contain measurements of glucose homeostasis markers, blood lipids, inflammatory markers and adipokines, body composition, kidney function, and subcutaneous adipose tissue (AT) markers. 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. The challenge is to reveal causal relations between the groups of variables.

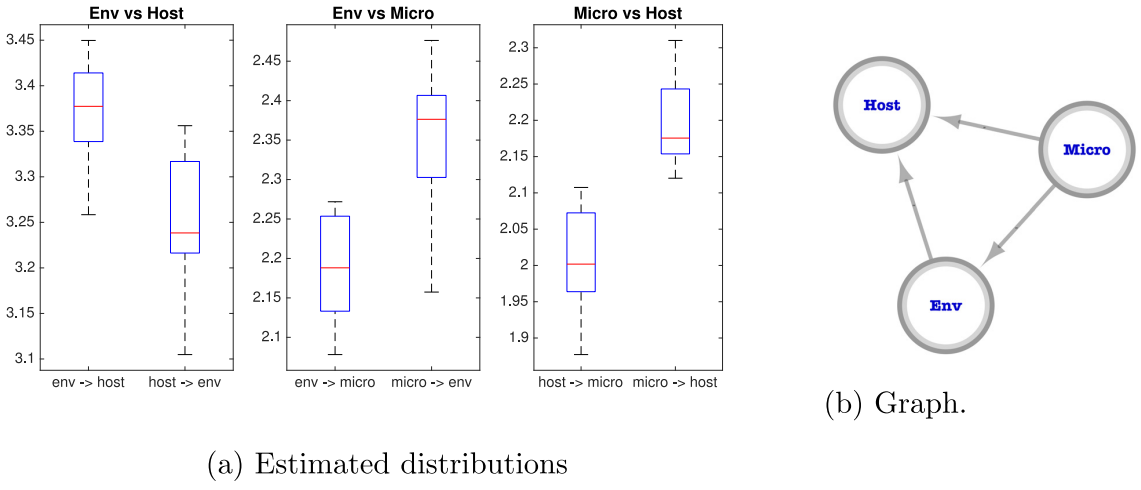
The parameters of the MicroObes data can be naturally divided into several groups. After discussions with the clinicians of the Pitié-Salpêtrière hospital, we decided to consider two scenarios:

1. Find causal relations between 3 groups (environment, host, and bacteria)
2. Find causal directions between 10 groups (glucose homeostasis markers, blood lipids, inflammatory markers and adipokines, body composition, kidney function, subcutaneous AT markers, food groups, nutrients, physical activity, and gut flora bacteria).

The experiments on the cause-effect pairs benchmark confirmed that the criterion Eq. (20) is the best one in terms of accuracy of causal directions. We have chosen it for the MicroObes data exploration. The DRBM have a random element, since the weights are initialised randomly, and the result on the same data can vary. We run the experiments 10 times to make a decision which is robust.

Fig. 4 on the left, demonstrates the distributions, similar to what we produced for the cause-effect pairs, for the scenario with three groups. We can construct a directed graph, and Fig. 4 on the right visualizes the causal relations between the three heterogeneous groups learnt purely from the data.

We run the same experiments for the second scenario which includes 10 groups, and we visualize the obtained graph as Fig. 5. We observe that food has an important impact on the level of blood lipids, on the composition of the human gut, on the markers of the adipose tissue, on body composition, on physical activity, on glucose level, on the inflammation markers, and on the kidney function. The width of an edge is equal to an average taken over 10 runs of the absolute values $|D_{X|Y} - D_{Y|X}|$. Such a heuristic can be a reasonable indicator of significance of a causal direction. We kept all edges whose strength is bigger than 0.1. To provide more details on the edge statistics, we show the values for all edges in Table 1 in Appendix. Note that the node *Nutr* stands for nutrition and includes nutritional values, calculated by clinicians from the food questionnaires, and, therefore, the causal direction *Food* → *Nutr* was expected. The hypothesis that the gut flora (*Bacteria*) can have an important impact on the inflammatory status is verified in the data.



(a) Estimated distributions

(b) Graph.

Fig. 4. Causal inference on the MicroObes data with three groups of heterogeneous data sources.

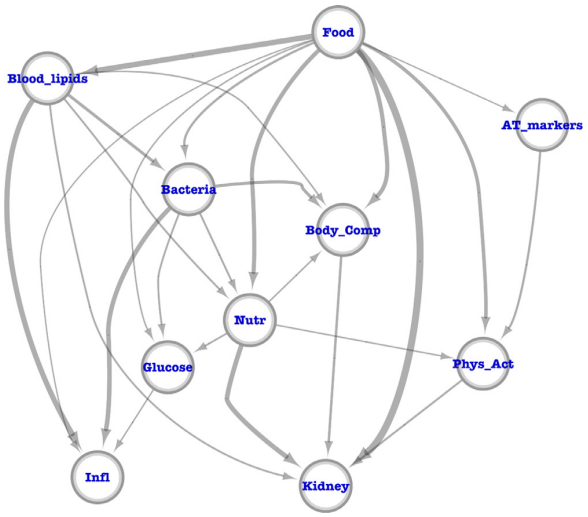


Fig. 5. Resulting causal graph for the MicroObes data with 10 groups of heterogeneous observations.

7. Conclusions

We proposed a principled approach to causal discovery from purely observational non-temporal blocks of heterogeneous data. In this work, we developed an improved version of the causal inference algorithm with distance correlation method, by incorporating Principal Component Analysis and Deep Restricted Boltzmann machines. Our approach enables causal inference in wider, namely in high-dimensional, application scenarios. We showed by numerical experiments that the approach is computationally efficient. Note that its implementation is simple. The method does not rely on any hyper-parameters which are needed to be adjusted.

The proposed causal inference algorithm was compared to the state-of-the-art methods. The results of the experiments on the multivariate cause-effect pairs (discrete or discretized) show that our algorithm reaches the best performance in terms of empirical performance. We have also demonstrated that the proposed approach is efficient for real applications, and can help clinicians and researchers doing fundamental biological research to verify complex mechanistic hypotheses.

Currently we are planning to extend the proposed method for confounding variables. An important research avenue is to find an optimal discretization method, in particular, an adaptive discretization, since empirical performance is highly data dependent. A promising research direction are information theoretic causal inference methods based on Kolmogorov complexity. They use the minimum description length assuming that the probability distribution is simpler in causal direction.

Code

Code for the proposed method is available at the following link: https://integromics.fr/~nsokolovska/code.html.

Acknowledgements

This work was supported by the European Union's Seventh Framework Program under grant agreement HEALTH-F4-2012-305312 (MetaCardis project). This work was also supported by Agence Nationale de la Recherche (ANR MICRO-Obes).

Appendix

Table 1
Edge Statistics for the MicroObes Data.

Source	Target	Strength	Source	Target	Strength
Bacteria	AT_markers	0.068523	Blood_lipids	AT_markers	0.065421
AT_markers	Body_Comp	0.087794	Food	AT_markers	0.10483
Glucose	AT_markers	0.070102	Infl	AT_markers	0.073882
AT_markers	Kidney	0.089233	Nutr	AT_markers	0.099515
AT_markers	Phys_Act	0.13261	Blood_lipids	Bacteria	0.098889
Bacteria	Body_Comp	0.22556	Food	Bacteria	0.070223
Bacteria	Glucose	0.16729	Bacteria	Infl	0.10617
Bacteria	Kidney	0.044709	Bacteria	Nutr	0.074261
Bacteria	Phys_Act	0.060303	Blood_lipids	Bacteria	0.15581
Blood_lipids	Body_Comp	0.13514	Food	Blood_lipids	0.22781
Blood_lipids	Glucose	0.071551	Blood_lipids	Infl	0.29245
Blood_lipids	Kidney	0.11803	Blood_lipids	Nutr	0.075615
Blood_lipids	Phys_Act	0.087583	Bacteria	Body_Comp	0.15825
Blood_lipids	Body_Comp	0.11272	Food	Body_Comp	0.19478
Body_Comp	Glucose	0.086518	Body_Comp	Infl	0.046299
Body_Comp	Kidney	0.16786	Nutr	Body_Comp	0.11126
Phys_Act	Body_Comp	0.070197	Food	Bacteria	0.13606
Food	Blood_lipids	0.23331	Food	Body_Comp	0.19441
Food	Glucose	0.083089	Food	Infl	0.10574
Food	Kidney	0.27788	Food	Nutr	0.16854
Food	Phys_Act	0.13919	Bacteria	Glucose	0.121
Blood_lipids	Glucose	0.066414	Body_Comp	Glucose	0.096817
Food	Glucose	0.11026	Glucose	Infl	0.063315
Kidney	Glucose	0.05642	Nutr	Glucose	0.13963
Phys_Act	Glucose	0.075443	Bacteria	Infl	0.20479
Blood_lipids	Infl	0.22789	Body_Comp	Infl	0.0505
Food	Infl	0.090273	Glucose	Infl	0.10894
Kidney	Infl	0.093364	Nutr	Infl	0.079415
Infl	Phys_Act	0.068262	Bacteria	Kidney	0.070516
Blood_lipids	Kidney	0.12113	Body_Comp	Kidney	0.13921
Food	Kidney	0.28229	Kidney	Glucose	0.053427
Kidney	Infl	0.090114	Nutr	Kidney	0.15269
Phys_Act	Kidney	0.12751	Bacteria	Nutr	0.1242
Blood_lipids	Nutr	0.12295	Nutr	Body_Comp	0.096798
Food	Nutr	0.19157	Nutr	Glucose	0.11799
Nutr	Infl	0.070178	Nutr	Kidney	0.21011
Nutr	Phys_Act	0.11041	Phys_Act	Bacteria	0.050546
Blood_lipids	Phys_Act	0.067077	Body_Comp	Phys_Act	0.071918
Food	Phys_Act	0.17273	Phys_Act	Glucose	0.075065
Phys_Act	Infl	0.066928	Phys_Act	Kidney	0.12567
Nutr	Phys_Act	0.075765			

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2018.11.016.

References

[1] Z. Hao, H. Zhang, R. Cai, W. Wen, Z. Li, Causal discovery on high dimensional data, Appl. Intell. 32 (3) (2015) 594–607.
[2] D. Janzing, P.O. Hoyer, B. Schölkopf, Telling cause from effect based on high-dimensional observations, ICML, 2010.
[3] J. Zscheischler, D. Janzing, K. Zhang, Testing whether linear equations are causal: a free probability theory approach, UAI, 2011.
[4] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference, The MIT Press, 2017.
[5] A. Cotillard, S.P. Kennedy, L.C. Kong, E. Prifti, N. Pons, E.L. Chatelier, M. Almeida, B. Quinquis, F. Levenez, N. Galleron, S. Gougis, S. Rizkalla, J.-M. Batto, P. Renault, A.M. consortium, J. Doré, J.-D. Zucker, K. Clément, S.D. Ehrlich, Dietary intervention impact on gut microbial gene richness, Nature 500 (2013) 585–588.
[6] R. Martìn, S. Miquel, P. Langella, L. Bermúdez-Humarán, The role of metagenomics in understanding the human microbiome in health and disease, Virulence (2014).
[7] National Research Council Committee on Metagenomics, The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet, National Academies Press (US), 2007.
[8] W.-H. Lee, H.-M. Chen, S.-F. Yang, C. Liang, C.-Y. Peng, F.-M. Lin, L.-L. Tsai, B.-C. Wu, C.-H. Hsin, C.-Y. Chuang, T. Yang, T.-L. Yang, S.-Y. Ho, W.-L. Chen, K.-C. Ueng, H.-D. Huang, C.-N. Huang, Y.-J. Jong, Bacterial alterations in salivary microbiota and their association in oral cancer, Sci. Rep. (2018).

[9] M. Schirmer, E.A. Franzosa, J. Lloyd-Price, L.J. McIver, R. Schwager, T.W. Poon, A.N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, M. Prasad, J. Sauk, B. Stevens, R.G. Wilson, J. Braun, L.A. Denson, S. Kugathasan, D.P.B. McGovern, H. Vlamakis, R.J. Xavier, C. Huttenhower, Dynamics of metatranscription in the inflammatory bowel disease gut microbiome, Nat. Microbiol. (2018).
[10] X. Cui, L. Ye, J. Li, L. Jin, W. Wang, S. Li, M. Bao, S. Wu, L. Li, B. Geng, X. Zhou, J. Zhang, J. Cai, Metagenomic and metabolomic analyses unveil dysbiosis of gut microbiota in chronic heart failure patients, Sci. Rep. (2018).
[11] R. Shaw, K.A. Lamia, D. Vasquez, S.H. Koo, N. Bardeesy, R.A. Depinho, M. Montminy, L. Cantley, The kinase lkb1 mediates glucose homeostasis in liver and therapeutic effects of metformin, Science 310 (2005) 1642–1646.
[12] A. Madiraju, D.M. Erion, Y. Rahimi, X. Zhang, D. Braddock, R. Algright, B.J. Prigaro, J.L. Wood, S. Bhanot, M.J. MacDonald, M.J. Jurczak, J.P. Camporez, H.Y. Lee, G.W. Cline, V.T. Samuel, R.G. Kibbey, G.I. Shulman, Metformin suppresses gluconeogenesis by inhibiting mitochondrial glycerophosphate dehydrogenase, Nature 510 (2014) 542–546.
[13] L. McCreight, C.J. Bailey, E. Pearson, Metformin and the gastrointestinal tract, Diabetologia 59 (2016) 426–435.
[14] D. Janzing, B. Schölkopf, Causal inference using the algorithmic markov condition, IEEE Trans. Inf. Theory 56 (2010) 5168–5194.
[15] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Streudel, B. Schölkopf, Information-geometric approach to inferring causal directions, Artif. Intell. 182-183 (2012) 1–31.
[16] E. Sgouritsa, D. Janzing, P. Hennig, B. Schölkopf, Inference of cause and effect with unsupervised inverse regression, AISTATS, 2015.
[17] J. Pearl, Causality: Models, Reasoning and Inference, 2nd edition, Cambridge University Press, Oxford, 2009.
[18] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and search, Springer, New York, NY, 2000.
[19] S. Affeldt, L. Verny, H. Isambert, 3off2: a network reconstruction algorithm based on 2-point and 3-point information statistics, BMC Bioinform. 17 (S-2) (2016) 12.
[20] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, Massachusetts, 2016.
[21] Y.L. Cun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[22] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[23] Y. Li, F.-X.Wu, A. Ngom, A review on machine learning principles for multi-view biological data integration, Brief. Bioinform. (2016).
[24] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, AISTATS, 2009.
[25] R. Salakhutdinov, H. Larochelle, Efficient learning of deep Boltzmann machines, AISTATS, 2010.
[26] A. Sankaran, G. Goswami, M. Vatsa, R. Singh, A. Majumdar, Class sparsity signature based restricted boltzmann machine, Pattern Recognit. 61 (2017) 674–685.
[27] J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks, JMLR 17 (2016).
[28] C. Heinze-Deml, M.H. Maathuis, N. Meinshausen, Causal structure learning, 2017. arXiv: 1706.09141.
[29] P. Hoyer, D. Janzing, J. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models., NIPS, 2009.
[30] J. Peters, J. Mooij, D. Janzing, B. Schölkopf, Causal discovery with continuous additive noise models, JMLR 1 (15) (2014) 2009–2053.
[31] P. Bühlmann, J. Peters, J. Ernest, Cam: causal additive models, high-dimensional order search and penalized regression., Ann. Stat. 42 (2014) 2526–2556.
[32] K. Zhang, A. Hyvärinen, On the identifiability of the post-nonlinear causal models, UAI, 2009.
[33] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, J. Mooij, Semi-supervised Learning in Causal and Anticausal Settings, in: Festschrift in Honor of Vladimir Vapnik, Springer-Verlag, 2013, pp. 129–141.
[34] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, B. Schölkopf, Cause-effect inference by comparing regression errors, AISTATS, 2018.
[35] F. Liu, L.-W. Chan, Causal inference on multidimensional data using free probability theory, IEEE Trans. Neural Netw. Learn. Syst. (2017).
[36] K. Budhathoki, J. Vreeken, Causal inference by compression, ICDM, 2016.
[37] A. Marx, J. Vreeken, Telling cause from effect using MDL-based local and global regression, ICDM, 2017.
[38] L. Verny, N. Sella, S. Affeldt, P. Singh, H. Isambert, Learning causal networks with latent variables from multivariate information in genomic data, PLoS Comput. Biol. 13 (10) (2017).
[39] F. Liu, L. Chan, Causal inference on discrete data via estimating distance correlations, Neural Comput. 28 (2016).
[40] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Annal. Stat. 35 (6) (2007) 2769–2794.
[41] A. Foroushani, R. Agrahari, R. Docking, L. Chang, G. Duns, M. Hudoba, A. Karsan, H. Zare, Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: an introduction to the pigengene package and its applications, BMC Med. Genomics 10 (16) (2017).
[42] I. Markovsky, Low Rank Approximation: Algorithms, Implementation, Applications, Springer, Cham, 2011.
[43] G. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[44] T.D. Nguyen, D. Phung, V. Huynh, T. Lee, Supervised restricted Boltzmann machines, UAI, 2017.

Revealing causality between heterogeneous data sources with deep restricted Boltzmann machines

N. Sokolovska, K. Clément, J.-D. Zucker

Information Fusion 2019

Résumé

Une question importante en microbiologie est de savoir si le traitement provoque des modifications de la flore intestinale et s'il affecte également le métabolisme. La reconstruction de relations causales purement à partir de données d'observation non temporelles est difficile. Nous abordons le problème de l'inférence causale dans un cas multivarié, où la distribution jointe de ces variables est observée. Nous nous attaquons ici au problème de l'inférence causale entre groupes d'observations hétérogènes non temporelles obtenues à partir de sources multiples. Les méthodes d'inférence causale de l'état de l'art pour les données continues souffrent d'une grande complexité de calcul. Certaines approches modernes ne conviennent pas aux données catégorielles, tandis que d'autres nécessitent d'estimer et de fixer plusieurs hyper-paramètres.

Dans cette contribution, nous introduisons une nouvelle méthode d'inférence causale qui est basée sur l'hypothèse largement utilisée que si X cause Y, alors $P(X)$ et $P(Y|X)$ sont indépendantes. Nous proposons d'explorer une approche semi-supervisée dans laquelle $P(Y|X)$ et $P(X)$ sont estimés à partir de données étiquetées et non étiquetées, alors que la probabilité marginale est estimée potentiellement à partir de données plus grandes mais non étiquetées.

Nous validons la méthode proposée sur les paires cause à effet standard. Nous illustrons par des expériences avec des tâches de la reconstruction des réseaux biologiques que l'approche proposée est très compétitive en termes de temps de calcul et de précision par rapport a des méhodes existantes.

Pour télécharger le papier publié:

https://www.sciencedirect.com/science/article/abs/pii/S1566253517307509