Multi-Dimensional Causal Discovery

Ulrich Schaechtle and **Kostas Stathis**
Dept. of Computer Science,
Royal Holloway,
University of London, UK.
{u.schaechtle, kostas.stathis}@rhul.ac.uk

Stefano Bromuri
Dept. of Business Information Systems,
University of Applied Sciences,
Western Switzerland, CH.
stefano.bromuri@hevs.ch

Abstract

We propose a method for learning causal relations within high-dimensional tensor data as they are typically recorded in non-experimental databases. The method allows the simultaneous inclusion of numerous dimensions within the data analysis such as samples, time and domain variables construed as tensors. In such tensor data we exploit and integrate non-Gaussian models and tensor analytic algorithms in a novel way. We prove that we can determine simple causal relations independently of how complex the dimensionality of the data is. We rely on a statistical decomposition that flattens higher-dimensional data tensors into matrices. This decomposition preserves the causal information and is therefore suitable for structure learning of causal graphical models, where a causal relation can be generalised beyond dimension, for example, over all time points. Related methods either focus on a set of samples for instantaneous effects or look at one sample for effects at certain time points. We evaluate the resulting algorithm and discuss its performance both with synthetic and real-world data.

1 Introduction

Causal discovery seeks to develop algorithms that learn the structure of causal relations from observation. Such algorithms are increasingly gaining importance since, as argued in [Tenenbaum *et al.*, 2011], producing rich causal models computationally can be key to creating human-like artificial intelligence. Diverse problems in domains such as aeronautical engineering, social sciences and bio-medical databases [Spirtes *et al.*, 2010] have acted as motivating applications that have gained new insights by applying suitable algorithms to large amounts of non-experimental data, typically collected via the internet or recorded in databases.

We are motivated by a class of causal discovery problems characterised by the need to analyse large and non-experimental datasets where the observed data of interest are recorded as continuous-valued variables. For instance, consider the application of causal discovery algorithms to large bio-medical databases recording data of diabetic patients. In such databases we may wish to find causal relations between variables such as the administration of medications like insulin dosage and the effects it has on diabetes management, for example, the patients' glucose level [Kafalı *et al.*, 2013]. We are particularly concerned with datasets that are multi-dimensional, for example, consider insulin dosage and glucose level measurements for different patients over time. Insulin dosage and glucose level are variables. Patients, variables for a patient, and time are dimensions.

A convenient way to represent cause-and-effect relations between variables is as directed edges between nodes in a graph. Such a graph, understood as a Bayesian Network [Pearl, 1988], allows us to factorise probability distributions of variables by defining one conditional distribution for each node given its causes. A more expressive representation of cause-and-effect relations is based on generative models that associate functions to variables, with the concomitant advantages of testability of results, checking equivalence classes between models and identifiability of causal effects [Pearl, 2000; Spirtes *et al.*, 2000].

We are taking a generative modelling approach to discover cause-and-effect relations for multi-dimensional data. Our starting point is the generative model LiNGAM (Linear Non-Gaussian Additive Model) [Shimizu *et al.*, 2006] as it relies on independent component analysis (ICA) [Hyvaerinen and Oja, 2000], which in turn is known to be easily extensible to data with multiple-dimensions. Multi-dimensional extensions for time series data using LiNGAM exist [Hyvaerinen *et al.*, 2008; 2010; Kawahara *et al.*, 2011]. However, for datasets such as the one for diabetic patients mentioned before, it is unclear how to answer even simple questions like *Does insulin dosage have an effect on glucose values?* There is at least one main problem that we see, namely, how to directly include more than one patient into the model learning process, as these approaches fit one model to a time series at a time. What we need here is an algorithm that abstracts-away from selected dimensions of the data while preserving the causal information that we seek to discover.

In this paper we present *Multi-dimensional Causal Discovery* (MCD), a method that discovers simple acyclic graphs of cause-and-effect relations between variables independently of the dimensions of the data. MCD integrates LiNGAM with tensor analytic techniques [Kolda and Bader, 2009] to support causal discovery in multi-dimensional data that was not possible before. MCD relies on a statistical decomposition that flattens higher dimensional data tensors into matrices and preserves the causal information. MCD is therefore suitable for structure learning of causal graphical models, where a causal relation can be generalised beyond dimension, for example, over all points in time. However, we can include not only time as a dimension, but also space or viewpoints, still resulting in plain knowledge representation of cause-and-effect. This is crucial for an intuitive understanding of time series in the context of causal graphs. As part of our contribution we also prove that we can determine simple causal relations independently of the dimensionality of the data. We specify an algorithm for MCD and we evaluate the algorithm's performance.

The paper is structured as follows. The background on LiNGAM and the (multi) linear transformations relevant to understand the work are reported in section 2. Then in section 3 we introduce MCD for causal discovery in multi-linear data. The algorithm is extensively evaluated on synthetic data as well as on real-world data in section 4. Apart from synthetic data, here, we show the benefits of our method for time series data analysis within application domains such as medicine, meteorology and climate studies. Related work is discussed in section 5, where we put our approach into the scientific context of the multi-dimensional causal discovery. We summarise the advantages of MCD in section 6, where we also outline our plans for future work.

2 Background

In the following, we will denote with x a random variable, with $\mathbf{x}$ a column vector, with $\mathbf{X}$ a matrix and with $\mathcal{X}$ a tensor.

2.1 LiNGAM

LiNGAM is a method for finding the instantaneous causal structure of non-experimental matrix data [Shimizu *et al.*, 2006]. By instantaneous we mean that causality is not time-dependent and is modelled in terms of functional equations [Pearl, 2000]. The underlying functional equation for LiNGAM is formulated as [Shimizu *et al.*, 2005]:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i + c_i \tag{1}$$

with

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{2}$$

put together as

$$\mathbf{x} = \mathbf{A}\mathbf{e} \tag{3}$$

where

$$\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}. \tag{4}$$

The observed variables are arranged in a causal order denoted by $k(j) < k(i)$. The value assigned to each variable x_i is a linear function of the values already assigned to the earlier variables x_j, plus a noise term e_i, and an optional constant term c_i that we will ommit from now on. $\mathbf{x}$ is a column vector of length m. $\mathbf{e}$ is an error vector. We assume non-Gaussian error distributions. This has an empirical advantage since a change in variance of some Gaussian noise distribution (e.g.: over time) will induce non-Gaussian noise [Hyvaerinen *et al.*, 2010]. The non-Gaussianity assumption yields also a practical advantage resulting in one unambiguous graph instead of a set of possible graphs.

We know that a directed acyclic graph (DAG) can be expressed as a strictly lower triangular matrix [Bollen, 1989]. Shimizu and colleagues define strictly lower triangular as lower triangular with all zeros at the diagonal [2006]. Here, LiNGAM exploits that we can permute an unknown matrix into a strictly lower triangular matrix, given enough entries in this matrix are zero. The matrix which is permuted in LiNGAM is the result of an Independent Component Analysis (ICA) with additional processing.

Linear transformations in data, such as ICA, can reduce a problem into simpler, underlying components suitable for analysis. Regarding assumptions such as linearity and noise we can have the full spectrum of possible models. For our purposes, we are looking into methods that can be suitably integrated in causal discovery algorithms. The basis of each of these methods is a latent variable model. Here, we assume the n-dimensional data to be "generated" by $l \leq m$ latent (or hidden) variables. We can compute the data matrix by linearly mixing these components [Bishop, 2006; Hyvaerinen and Oja, 2000]:

$$x_j = a_{j1} y_1 + ... + a_{jm} y_m \tag{5}$$

for all j with $j \in [1, m]$. Similar for the complete sample:

$$\mathbf{x} = \mathbf{A}\mathbf{y}. \tag{6}$$

The above equation corresponds to (3), where $\mathbf{e} = \mathbf{y}$. We can extend this idea to the entire dataset $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{A}\mathbf{Y}. \tag{7}$$

$\mathbf{X}$ is an $m \times n$ data matrix where m is the number of variables n is the number of cases or samples .

ICA builds on the sole assumption that the data is generated by a set of statistically independent components. Assuming such independence, we can transform the axis system determined by the m variables so that we can detect m independent components. We can use this in LiNGAM because the order of the independent components in ICA cannot be determined. As explained in [Hyvaerinen and Oja, 2000], the reason for this indeterministic nature is that, since both $\mathbf{y}$ and $\mathbf{A}$ are unknown, one can freely change the order of the terms in the sum (5) and call any of the independent components the first one. Formally, a permutation matrix $\mathbf{P}$ and its inverse can be substituted in the model to give:

$$\mathbf{x} = \mathbf{B}\mathbf{P}^{-1}\mathbf{P}\mathbf{y}. \tag{8}$$

The elements of $\mathbf{P}\mathbf{y}$ are the original independent variables $\mathbf{y}$, but in another order. The matrix $\mathbf{B}\mathbf{P}^{-1}$ is just a new unknown mixing matrix, to be solved by the ICA algorithm.

We now describe the LiNGAM algorithm [Shimizu *et al.*, 2005]:

1. Given an $m \times n$ data matrix $\mathbf{X}$ $(m < n)$ where each column contains one sample vector $\mathbf{x}$, first subtract the mean from each row of $\mathbf{x}$, then apply an ICA algorithm to obtain a decomposition $\mathbf{X} = \mathbf{A}\mathbf{S}$ where $\mathbf{S}$ has the

same size as $\mathbf{X}$ and contains in its rows the independent components. From now on we will work with (9):

$$\mathbf{W} = \mathbf{A}^{-1}. \tag{9}$$

2. Find the permutation of rows of $\mathbf{W}$ which yields a matrix $\tilde{\mathbf{W}}$ without any zeros on the main diagonal. In practice, small estimation errors will cause all elements of $\mathbf{W}$ to be non-zero, and hence the permutation is sought which minimises (10):

$$\sum_i \frac{1}{|\tilde{W}_{ii}|}. \tag{10}$$

3. Divide each row of $\tilde{\mathbf{W}}$ by its corresponding diagonal element, to yield a new matrix $\tilde{\mathbf{W}}'$ with all ones on the diagonal.
4. Compute an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$ using

$$\hat{\mathbf{B}} = \mathbf{I} - \tilde{\mathbf{W}}'. \tag{11}$$

5. To find a causal order, find the permutation matrix $\mathbf{P}$ (applied equally to both rows and columns) of $\hat{\mathbf{B}}$ yielding

$$\tilde{\mathbf{B}} = \mathbf{P}\hat{\mathbf{B}}\mathbf{P}^T, \tag{12}$$

which is as close as possible to strictly lower triangular. This can be measured for instance using $\sum_{i \le j} \tilde{B}_{ij}^2$.

We describe next the background of how to represent multi-dimensional data.

2.2 Tensor Analysis

A tensor is a multi-way array or multi-dimensional array.

Definition [Cichocki *et al.*, 2009] Let $I_1, I_2, \ldots, I_K \in K$ denote upper bounds. A tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2, \ldots, I_1, I_K}$ of order K is a K-dimensional array where elements $y_{i1,i2,\ldots,i_k}$ are indexed by $i_k \in \{1, 2, \ldots, Ik\}$ for k with $1 \le k \le K$.

Tensor analysis is applied in datasets with a high number of dimensions, other than the conventional matrix data (see Figure 1). An example where tensor analysis can be applicable is time series in medical data. Here, we have a number of patients n, a number of treatment variables m such as medication and symptoms of a disease, and t discrete points in time at which the treatment data for different patients have been collected. This makes one $n \times m$ data matrix for each point in time t or a tensor of the dimension $n \times m \times t$.

Definition [Kolda and Bader, 2009] The order of a tensor is the number of its dimensions, also known as ways or modes.

We also need to define the n-dimensional tensor product.

Definition [Cichocki *et al.*, 2009] The mode-n tensor matrix product $\mathcal{X} = \mathcal{G} \times_n A$ of a tensor $\mathcal{G} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_N}$ and a matrix $\mathbf{A} \in \mathbb{R}^{I_n \times J_n}$ is a tensor $\mathcal{Y} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_{n-1} \times I_n \times J_n + 1 \times \ldots \times J_N}$, with elements

$$x_{j_1, j_2, \ldots, j_{i-1}, i_n, j_{n+1}, \ldots, j_N} = \sum_{j_n = 1}^{J_n} \mathcal{G}_{j_1, j_2, \ldots, j_N} a_{i_n j_n}. \tag{13}$$

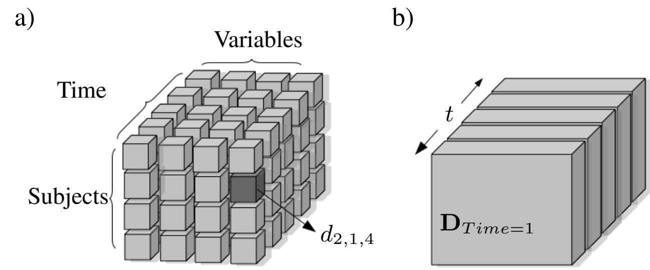

Figure 1: a) A three-dimensional tensor: with subjects (dimension 1), time (dimension 2), and variables (dimension 3) yielding a cube (instead of a matrix). b) We can frontally slice the data - each slice represents a snapshot of the variables for a fixed point in time.

Tensor decomposition can be described as a multi-linear extension of PCA (Principal Component Analysis), for a K-dimensional tensor:

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U_1} \times_2 \mathbf{U_1} \times_3 \ldots \times_k \mathbf{U_k}. \tag{14}$$

$\mathcal{G}$ is the core-tensor, that is the multi-dimensional extension of the latent variables $\mathbf{Y}$, with $\mathbf{U_1} \ldots \mathbf{U_k}$ being orthonormal.

Unlike for matrix decomposition, there is no trivial solution for computing the tensor decomposition. We use alternating least square (ALS) methods (e.g.: Tucker-ALS [Kolda and Bader, 2009]), since efficient implementations are available [Bader *et al.*, 2012]. The decomposition can be optimised in terms of the components of a factorisation for every dimension iteratively [De Lathauwer *et al.*, 2000].

In practical applications, it is useful to work with a projection of the actual decomposition. By projection, we mean a mapping of the information of an arbitrary tensor onto a second order tensor (a matrix). Such a mapping can be achieved using a linear operator represented as a matrix, which from now on we will refer to as projection matrix. ALS optimisation works on a projected decomposition as well. However, the resulting projection allows us to apply well-known methods for matrix data on very complex datasets.

K-dimensional Independent Component Analysis. We can determine a K-dimensional extension [Vasilescu and Terzopoulos, 2005] for ICA. We can decompose a tensor $\mathcal{X}$ as the k-dimensional product of k matrices A_k and a core tensor $\mathcal{G}$:

$$\mathcal{X} = \mathcal{G} \times_\mathbf{1} \mathbf{A_1} \times_\mathbf{2} \mathbf{A_2} \times_\mathbf{3} \ldots \times_\mathbf{k} \mathbf{A_k}. \tag{15}$$

To extend ICA for higher-dimensional problem domains, we first need to relate ICA with PCA.

$$\begin{aligned} \mathbf{X} &= \mathbf{U\Sigma V^T} \\ &= (\mathbf{UH^{-1}})(\mathbf{H\Sigma V^T}) \\ &= \mathbf{AY}. \end{aligned} \tag{16}$$

Here we compute PCA with SVD (Singular Value Decomposition) as $\mathbf{U\Sigma V^T}$. For the K-dimensional ICA, we can make use of (16) and define the following sub-problem:

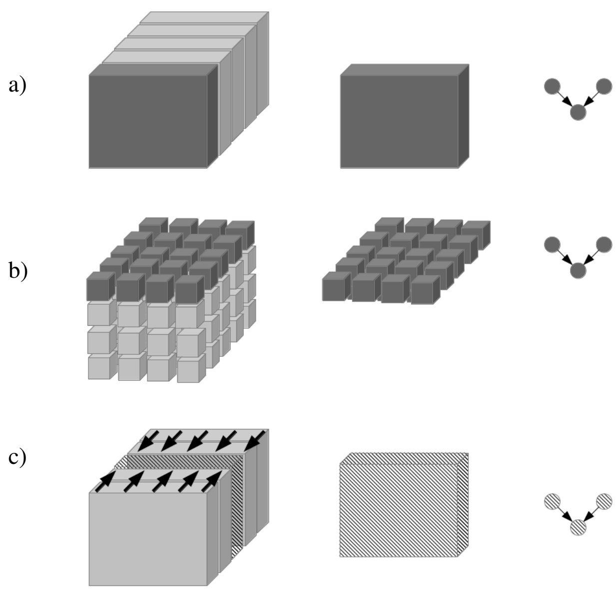

Figure 2: Causal analysis of time series data in tensor form. As before, the dimensions are subjects (dimension 1), time (dimension 2), and variables (dimension 3). (a) LiNGAM does not take into account temporal dynamics; it only inspects a snapshot, ignoring temporal correlations. (b) LiNGAM extensions investigate several points in time for one case. (c) MCD flattens the data whilst preserving the causal information and then applies linear causal discovery.

$$\begin{aligned} \mathbf{X_{(k)}} &= \mathbf{U_k \Sigma_k V_k^T} \\ &= (\mathbf{U_k H_k^{-1}})(\mathbf{H_k \Sigma_k V_k^T}) \\ &= \mathbf{A_k Y_k}. \end{aligned} \quad (17)$$

This sub-problem is due to [Vasilescu and Terzopoulos, 2005], who argue that the core tensor $\mathcal{G}$ enables us to compute the coefficient vectors via a tensor decomposition using a K-dimensional SVD algorithm.

3 Multi-dimensional Causal Discovery (MCD)

The main idea of our work is to integrate LiNGAM with K-dimensional tensors in order to efficiently discover causal dependencies in multi-dimensional settings, such as time series data. The intuition behind MCD is that we want to flatten the data, that is decompose the data and project the decomposition on a matrix (see Figure 2(c)).

Definition Meta-dimensionality reduction is the process of reducing the order of a tensor via optimising the equation $\mathcal{X} = \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1} \times_2 ... \times_k \mathbf{U_k}$ so that we can compute $\mathbf{X} = \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1} \times_2 ... \times_k \mathbf{U_{k-1}}$.

After a meta-dimensionality reduction step, we can apply LiNGAM directly and, as a result, reduce the temporal complexity of causal inferences (compare Figures 3(a) and 3(b)). We interpret the output graph of the algorithm as an indicator of cause-and-effect that is significant according to the tensor analysis for a sufficient subset of all the tensor slices. However, before applying LiNGAM, we need to ensure that the information on causal dependencies is preserved when meta-dimensionality reduction has been applied.

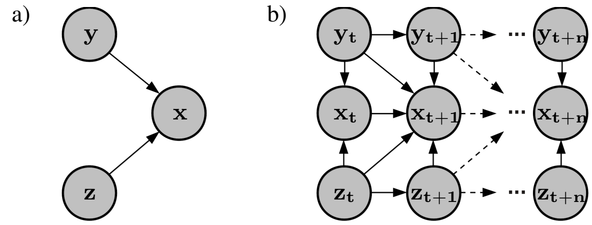

Figure 3: a) The causal graph determined by MCD b) A causal graph determined by extensions of LiNGAM for time series.

Theorem 3.1. *Let $\mathcal{X} = \mathcal{G} \times_1 \mathbf{S_1} \times_2 \mathbf{S_2} ... \times_k \mathbf{S_k}$ be any decomposition of a data tensor $\mathcal{X}$ that can be computed using SVD. Let $\mathbf{X}$ be the projection (mapping) of $\mathcal{X}$ where we remove one or more tensor dimensions for meta-dimensionality reduction. The independent components of tensor data subspaces which are to be permuted are independent of previous projections in the meta-dimensionality reduction process.*

Proof. The decomposition is computed independently for all dimensions (see (17)) using SVD. Furthermore, we know that the tensor matrix product is associative:

$$\mathcal{D} = (\mathcal{E} \times_1 \mathbf{A} \times_2 \mathbf{B}) \times_3 \mathbf{C} = \mathcal{E} \times_1 (\mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}). \quad (18)$$

Therefore, as long as it is computed with SVD we can establish a relation between Tucker-Decomposition and Multi-Linear ICA. Vasilescu and Terzopoulos define this relation in the following way [2005]:

$$\begin{aligned} \mathcal{X} &= \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1} \times_2 ... \times_k \mathbf{U_k} \\ &= \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1 H_1^{-1} H_1} \times_2 ... \times_k \mathbf{U_k H_k^{-1} H_k} \\ &= (\mathcal{Y}_{Tucker} \times_1 \mathbf{H_1} ... \times_k \mathbf{H_k}) \times_1 \mathbf{A_1} \times_2 ... \times_k \mathbf{A_k} \\ &= \mathcal{Y} \times_1 \mathbf{A_1} \times_2 ... \times_k \mathbf{A_k} \end{aligned} \quad (19)$$

where

$$\mathcal{Y} = \mathcal{Y}_{Tucker} \times_1 \mathbf{H_1} ... \times_k \mathbf{H_k}. \quad (20)$$

Now, when we look at an arbitrary projection of a data tensor $\mathcal{X}$ to a data matrix $\mathbf{X}$, we first need to generalise (19) for projections. This can simply be done by replacing the inverse with the Moore-Penrose pseudo-inverse (denoted with †), which is defined as:

$$\mathbf{H}^\dagger = (\mathbf{H^T H})^{-1} \mathbf{H^T} \quad (21)$$

where

$$\mathbf{H H^\dagger H} = \mathbf{H} \quad (22)$$

We can then rewrite (19) to the following:

$$\begin{aligned} \mathcal{X} &= \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1} \times_2 ... \times_k \mathbf{U_k} \\ &= \mathcal{Y}_{Tucker} \times_1 \mathbf{U_1 H_1^\dagger H_1} \times_2 ... \times_k \mathbf{U_k H_k^\dagger H_k} \\ &= (\mathcal{Y}_{Tucker} \times_1 \mathbf{H_1} ... \times_k \mathbf{H_k}) \times_1 \mathbf{A_1} \times_2 ... \times_k \mathbf{A_k} \\ &= \mathcal{Y} \times_1 \mathbf{A_1} \times_2 ... \times_k \mathbf{A_k} \end{aligned} \quad (23)$$

Finally, we look at the projection that we compute using Tucker-Decomposition. Here we project our tensor into a matrix by reducing the order of a K-dimension tensor. If one thinks of $\mathbf{H_k}$ in terms of a projection matrix, this can be expressed as:

$$\mathcal{X} = (\mathcal{Y}_{Tucker} \times_k \mathbf{H_k}) \times_1 \mathbf{U_1} \times_2 ... \times_{k-1} \mathbf{U_{k-1}} \times_k \mathbf{U_k}\mathbf{H_k}^{\dagger}. \tag{24}$$

We see that compared to (23) we find projections that allow the meta-dimensionality reduction in two places: at $(\mathcal{Y}_{Tucker} \times_k \mathbf{H_k})$ and at $\mathbf{U_k}\mathbf{H_k}^{\dagger}$. Accordingly, we define the projection of the decomposition to the matrix data $\mathbf{X}$ as:

$$\mathbf{X} = (\mathcal{Y}_{Tucker} \times_k \mathbf{H_k}) \times_1 \mathbf{U_1} \times_2 ... \times_{k-1} \mathbf{U_{k-1}} \tag{25}$$

Here, we see that $U_1 ... U_{k-1}$ are not affected by the projection. Therefore, by computing the $H_1 ... H_{K-1}$ we can still find the best result for ICA on our flattened tensor data, if we apply it on the projection. This means, that if we want to apply multi-linear ICA with the aim on having statistically independent components in meta-dimensionality reduced space, we can resort to Tucker-Decomposition for reducing the order of the tensor - until we finally compute the interesting part that is used by ICA for LiNGAM. Thus, the theorem is proven.

On these grounds, we can see that LiNGAM works, without making additional assumptions on the temporal relations between variables, if it is combined with any decomposition that can be expressed in terms of SVD. The algorithm is summarised in Algorithm 1, the variable-mode contains the causes and effects that we try to find.

Algorithm 1 Multi-dimensional Causal Discovery

procedure MCD($\mathcal{X}$, sample-mode, variable-mode)
    $k_{sm} \leftarrow$ sample-mode
    $k_{vm} \leftarrow$ variable-mode
    $\mathcal{Y}_{projection}, \mathbf{U_{k_{sm}}}, \mathbf{U_{k_{vm}}} \leftarrow$ Tucker-ALS($\mathcal{X}$)
    $\mathbf{X} \leftarrow \mathcal{Y}_{projection} \times_{k_{sm}} \mathbf{U_{k_{sm}}} \times_{k_{vm}} \mathbf{U_{k_{vm}}}$
    $\mathbf{B} \leftarrow$ LiNGAM($\mathbf{X}$)
end procedure

It is worth noting how MCD exploits Gaussianity: unlike plain LiNGAM, we do not forbid Gaussian noise totally. In the tensor-analytically reduced dimensions, we can have Gaussian noise, this does not influence MCD, as long as we have non-Gaussian noise in the non-reduced dimensions to identify the direction of cause-and-effect. Having Gaussian noise in one dimension and non-Gaussian noise in the other is an empirical necessity if time is involved [Hyvaerinen *et al.*, 2010].

4 Evaluation

4.1 Experiments with Synthetic Data

We have simulated a 3-dimensional tensor with dimensions cases, variables and time. 5000 cases with 5 variables and 50 points in time have been produced. For each case, we have created the time series with a special case of a structural autoregressive model:

$$\mathbf{X_t} = \mathbf{c} + \sum_{\mathbf{p=0}}^{\mathbf{P}} \phi_{\mathbf{p}} \mathbf{X_{t-p}} + \epsilon_t \tag{26}$$

with

$$\phi_{\mathbf{p}} = \mathbf{B} \tag{27}$$

and

$$\epsilon_{\mathbf{t}} \sim \mathcal{N},\ \mathbf{c} \sim \mathcal{SN} \tag{28}$$

where $\mathcal{SN}$ is a sub or super-Gaussian distribution. In contrast to the classical autoregressive model we start indexing at $\mathbf{p} = 0$. This allows us to include instantaneous and time-lagged effects. Furthermore, we allowed each of the nodes (variables) to have either one or two incoming edges at random. In that manner we created three different kinds of datasets, one with time-lag $\mathbf{p} = 1$, one with time-lag $\mathbf{p} = 2$ and one with time-lag $\mathbf{p} = 3$ to test the MCD algorithm with. Each kind we created 500 times, so that we could test the algorithm on a number of different datasets. We found the output of the algorithm to be correct in 73.00 % of all the datasets with time-lag $\mathbf{p} = 1$, 69.20 % with $\mathbf{p} = 2$ and 68.20 % with time-lag $\mathbf{p} = 3$. The decrease in accuracy can be explained by the increasing complexity of the time series function that comes with increasing $\mathbf{p}$. The algorithm's output was determined to be incorrect if there was any type of structural error in the graph, that is false positive or false negative findings. Due to this very conservative measure, we could achieve very high precision (ca. 99 %) and recall (ca. 96 %) when investigating the total number of correct classifications, that is whether there is a cause-effect-relation between one variable and another (i.e. $a \rightarrow b$ true or false). For pruning the edges in the LiNGAM part of the algorithm, we used a simple resampling method (described in [Shimizu *et al.*, 2006]).

4.2 Application to Real-world Data

To show how MCD works on real-world problems, we applied it to three different real-world datasets. Where possible, we also applied an implementation of multi-trial version of Granger Causality (MTGC) [Seth, 2010] to compare MCD's results to something known to the community. Also, from all related methods, Granger Causality is the only method where there is an extension available for multiple realisations of the same process [Ding *et al.*, 2006]. However, the multiple realisations are interpreted in terms of repetitive trials with a single subject or case. This suggests dependence due to repetition instead of the desired independence of cases. For example, if we look at a number of subjects and their medical treatment over time, we expect the subjects to be independent from each other.

First of all, we applied MCD and MTGC to a dataset on Diabetes [Frank and Asuncion, 2010]. Here, the known ground truth was Insulin $\rightarrow$ Glucose. Glucose curves and insulin dose were analysed for 69 patients - the number of points in time differed from patient to patient, thus we had to cut them all to similar size. MCD successfully found the causal ground truth, MTGC did not and resulted in a cyclic graph.

Secondly, we investigated a dataset with two variables, 72 points in time, 16 different places. The two variables were ozone and radiation with the assumed ground truth that radiation has an causal effect on ozone.[1] Again, MCD found the causal ground truth and MTGC did not and resulted in a cyclic graph.

Finally, we tested the algorithm on meteorological data[2]. 10,226 samples have been taken for how the weather conditions of one day cause the weather conditions of the second day. The variables that were measured were mean daily air temperature, mean daily pressure at surface, mean daily sea level pressure and mean daily relative humidity. Ground truth was that the conditions on day t affect the conditions on day $t+1$ which was found by MCD. We did not apply MTGC here because of its conceptual dependency to the time-dimension.

5 Related Work

The most well-known example of causality for time series is Granger Causality [Granger, 1969]. Granger affiliates his definition of causality with the time-dimension. Statistical tests regarding predictive power, when including a variable, detect an effect of this variable. Granger Causality cannot incorporate instantaneous effects, which is often cited as a drawback [Peters *et al.*, 2012]. MCD complements Granger Causality in this. Likewise, this is the case for transfer entropy (TE) [Schreiber, 2000]: proven equivalent to Granger Causality for the case of Gaussian noise [Barnett *et al.*, 2009], TE is bound to the notion of time. TE cannot detect instantaneous effects because potential asymmetries in the underlying information theory models are only due to different individual entropies and not due to information flow or causality.

Entner and Hoyer make use of similarities between causal relations over time to extend the Fast Causal Inference (FCI) algorithm [Spirtes, 2001] for time series [2010]. In contrast to MCD, FCI supports the modelling of latent confounding variables and it does not exploit the non-Gaussian noise assumptions.

The closest approaches to MCD are the approaches connecting LiNGAM to models of the Autoregressive-moving-average model (ARMA) class. For example, a link was established between LiNGAM and structural vector autoregression [Hyvaerinen *et al.*, 2008; 2010] in the context of non-Gaussian noise. The authors focus on an ICA interpretation of the autoregressive residuals. This was generalised for the entire ARMA class [Kawahara *et al.*, 2011]. These methods can be seen as a LiNGAM-based generalisation of Granger Causality, since they can take into account time-lagged and instantaneous effects. Similarly, the Time Series Models with Independent Noise, which can be used in a multi-variate, linear, non-linear setting, with or without instantaneous interactions [Peters *et al.*, 2012].

The main difference between our approach and these LiNGAM extensions (and the other related work discussed earlier in this section) is the possibility to directly include a number of dimensions in the analysis using the MCD algorithm. Previous research takes into account single time series, but does not allow abstracting away modes to produce simple and clear cause-and-effect relations. Here, it is unclear how to analyse multiple cases of multi-variate time series for causality. Only for Granger Causality, there are methods available for a direct comparison of performance.

[1]causal ground truth was given, data taken from `https://webdav.tuebingen.mpg.de/cause-effect/`

[2]same source as above

6 Conclusions

In this paper we have proposed MCD, a method for learning causal relations within high-dimensional data, such as multi-variate time series, as they are typically recorded in non-experimental databases. The contribution of the work is the implementation of an algorithm that integrates linear non-Gaussian additive models (LiNGAM) with tensor analytic techniques and opens up new ways of understanding causal discovery in multi-dimensional data that was previously impossible. We have shown how the algorithm relies on a statistical decomposition that flattens higher dimensional data tensors into matrices. This decomposition preserves the causal information and is therefore suitable to be included in the structure learning process of causal graphical models, where a causal relation can be generalised beyond dimension, for example, over all points in time. Related methods either focus on a set of samples for instantaneous effects or look at one sample for effects at certain points in time. We have also evaluated the resulting algorithm and discussed its performance both with synthetic and real-world data.

The practical value of MCD analysis needs to be determined by applying it to more real-world data sets and comparing it to other causal inference methods for non-experimental data. The real-world data analysed here are rather simple as they contain relations between two variables only. It has been quite difficult to find multi-dimensional time series where the underlying causality is clear. Here it would be useful to see how we can include discrete variables into the MCD analysis, because in most cases of non-experimental datasets we can find discrete-valued and continuous-valued variables.

Also, in the current method, the tensor analytic process of flattening the data relies on the variance of the linear interaction between the decomposed subspaces. A more direct integration of this aspect into the LiNGAM discovery process would be desirable. We aim to address this issue in future research too.

Finally, we plan to compare our approach to algorithms with other assumptions such as non-linearity and Gaussian error. The heteroscedastic nature of time series data could give rise to a formal integration of the interplay of the Gaussian and non-Gaussian noise assumption, that is how the non-Gaussian assumption's usefulness is "triggered" by the time-dimension. This may bring further light into the interplay between instantaneous and time-lagged causal effects.

Acknowledgements

We thank the anonymous referees for their valuable comments and Alberto Paccanaro for helpful discussions. The work was partially supported by the EU FP7 Project COMMODITY$_{12}$ (`www.commodity12.eu`).

References

[Bader *et al.*, 2012] B. W. Bader, T. G. Kolda, et al. Matlab tensor toolbox version 2.5. Available online, January 2012.

[Barnett *et al.*, 2009] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103(23):238701, 2009.

[Bishop, 2006] C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[Bollen, 1989] K. A. Bollen. Structural equations with latent variables. 1989.

[Cichocki *et al.*, 2009] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[De Lathauwer *et al.*, 2000] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[Ding *et al.*, 2006] M. Ding, Y. Chen, and S. L. Bressler. 17 granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis*, page 437, 2006.

[Entner and Hoyer, 2010] D. Entner and P. O. Hoyer. On causal discovery from time series data using FCI. *Probabilistic Graphical Models*, 2010.

[Frank and Asuncion, 2010] A. Frank and A. Asuncion. UCI machine learning repository. Available online, 2010.

[Granger, 1969] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[Hyvaerinen and Oja, 2000] A. Hyvaerinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[Hyvaerinen *et al.*, 2008] A. Hyvaerinen, S. Shimizu, and P. O. Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 424–431, 2008.

[Hyvaerinen *et al.*, 2010] A. Hyvaerinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *The Journal of Machine Learning Research*, 11:1709–1731, 2010.

[Kafalı *et al.*, 2013] Ö. Kafalı, S. Bromuri, M. Sindlar, T. van der Weide, E. Aguilar Pelaez, U. Schaechtle, B. Alves, D. Zufferey, E. Rodriguez-Villegas, M. I. Schumacher, and K. Stathis. Commodity$_{12}$: A smart e-health environment for diabetes management. *Journal of Ambient Intelligence and Smart Environments, IOS Press (To appear)*, 2013.

[Kawahara *et al.*, 2011] Y. Kawahara, S. Shimizu, and T. Washio. Analyzing relationships among arma processes based on non-gaussianity of external influences. *Neurocomputing*, 74(12):2212–2221, 2011.

[Kolda and Bader, 2009] T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[Pearl, 1988] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[Pearl, 2000] J. Pearl. *Causality: models, reasoning, and inference*, volume 47. Cambridge University Press, 2000.

[Peters *et al.*, 2012] J. Peters, D. Janzing, and B. Schoelkopf. Causal inference on time series using structural equation models. *arXiv preprint arXiv:1207.5136*, 2012.

[Schreiber, 2000] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[Seth, 2010] A. K. Seth. A matlab toolbox for granger causal connectivity analysis. *Journal of Neuroscience Methods*, 186(2):262, 2010.

[Shimizu *et al.*, 2005] S. Shimizu, A. Hyvaerinen, Y. Kano, and P. O. Hoyer. Discovery of non-gaussian linear causal models using ICA. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 526–533, 2005.

[Shimizu *et al.*, 2006] S. Shimizu, P. O. Hoyer, A. Hyvaerinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

[Spirtes *et al.*, 2000] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. 2000.

[Spirtes *et al.*, 2010] P. Spirtes, C. N. Glymour, R. Scheines, and R. Tillman. Automated search for causal relations: Theory and practice. Technical report, Department of Philosophy, Carnegie Mellon University, 2010.

[Spirtes, 2001] P. Spirtes. An anytime algorithm for causal inference. In *Proceedings of AISTATS*, pages 213–231. Citeseer, 2001.

[Tenenbaum *et al.*, 2011] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

[Vasilescu and Terzopoulos, 2005] M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 547–553. IEEE, 2005.