Telling cause from effect based on high-dimensional observations

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Patrik O. Hoyer PATRIK.HOYER@HELSINKI.FI
University of Helsinki, Finland

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract

We describe a method for inferring linear causal relations among multi-dimensional variables. The idea is to use an asymmetry between the distributions of cause and effect that occurs if the covariance matrix of the cause and the structure matrix mapping the cause to the effect are independently chosen. The method applies to both stochastic and deterministic causal relations, provided that the dimensionality is sufficiently high (in some experiments, 5 was enough). It is applicable to Gaussian as well as non-Gaussian data.

1. Motivation

Inferring the causal relations that have generated statistical dependencies among a set of observed random variables is challenging if no controlled randomized studies can be made. Here, causal relations are represented as arrows connecting the variables, and the structure to be inferred is a directed acyclic graph (DAG). There are several well-known approaches to this task, of which perhaps the most established one is the independence-based approach (Pearl, 2000; Spirtes et al., 1993) based on the causal Markov condition and an assumption of faithfulness: The guiding principle is to accept only those causal DAGs that explain all of the observed dependencies in the data and furthermore explain *only* those dependencies, i.e. all inferred (marginal and conditional) *in*dependencies in the data are assumed to derive from the structure of the DAG.

Recently, several authors have proposed an entirely different route to causal discovery, based on ICA or (more generally) additive-noise models. These methods assume that the effects are given by some (possibly nonlinear) functions of the cause up to an additive noise that is statistically independent of the cause (Kano & Shimizu, 2003; Shimizu et al., 2006; Hoyer et al., 2009). A recent proposal generalizes this model class by further allowing non-linear transformations of the effect (Zhang & Hyvärinen, 2009).

These methods both have their relative advantages but also limitations. Approaches based solely on conditional independencies cannot distinguish between causally distinct models that impose the same set of indendences; in particular, they cannot infer whether $\mathbf{X}$ causes $\mathbf{Y}$ or $\mathbf{Y}$ causes $\mathbf{X}$ for just two observed variables (Mooij & Janzing, 2010) $\mathbf{X}$ and $\mathbf{Y}$. Methods based on additive noise models fail for linear relationships with Gaussian noise. Finally, neither of the above approaches can deal with deterministic relationships between the observed variables.

In the present paper, we describe a method based on a recently proposed third principle. The idea is to reject the causal hypothesis $\mathbf{X} \to \mathbf{Y}$ whenever there are some kind of dependences between $P(\mathbf{X})$ and $P(\mathbf{Y}|\mathbf{X})$ that suggest that $P(\mathbf{X})$ and $P(\mathbf{Y}|\mathbf{X})$ were not generated by "independent mechanisms" of nature. Janzing & Schölkopf (2008) show examples illustrating how an independent choice of $P(\mathbf{X})$ and $P(\mathbf{Y}|\mathbf{X})$ typically leads to joint distributions where $P(\mathbf{Y})$ and $P(\mathbf{X}|\mathbf{Y})$ satisfy non-generic dependences indicating that $\mathbf{Y} \to \mathbf{X}$ is not a plausible model. Based on an idea of Lemeire & Dirkx (2006), Janzing & Schölkopf (2008) express these dependences in terms of algorithmic informa-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

tion theory. Unfortunately, this leads to a criterion that cannot be used for practical purposes due to the uncomputability of Kolmogorov complexity. In this contribution we provide an easily computable criterion for detecting dependences between $P(\mathbf{X})$ and $P(\mathbf{Y}|\mathbf{X})$ for the case of two high-dimensional variables $\mathbf{X}$ and $\mathbf{Y}$ coupled by a linear causal mechanism. We show that the principle works even for multivariate Gaussian models, and also if the relation is deterministic, provided that the joint covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is sufficiently anisotropic.

Before proceeding to describe our method, we should mention connections to Bayesian approaches (Heckerman et al., 1999) to causal discovery. Such methods can, in principle and depending on the priors chosen, use any of the information relied upon by the abovementioned three approaches. However, to date most Bayesian causal discovery methods have focused on conditional independence information. Furthermore, the fact that deterministic relationships exist in real-world settings shows that priors that are densities on the parameters of the Bayesian networks, as is usually assumed, are problematic and the construction of good priors is difficult. Thus, rather than defining a prior explicitly, we will assume that it satisfies some symmetry constraints and show how this already leads to our inference rule (Theorem 1).

We start with two motivating examples. First, assume that $\mathbf{X}$ is a multivariate Gaussian variable with values in $\mathbb{R}^n$ and the isotropic covariance matrix $\Sigma_{\mathbf{XX}} = \mathbf{I}$. Let $\mathbf{Y}$ be another $\mathbb{R}^n$-valued variable that is deterministically influenced by $\mathbf{X}$ via the linear relation $\mathbf{Y} = A\mathbf{X}$ for some $n \times n$-matrix A. This induces the covariance matrix

$$\Sigma_{\mathbf{YY}} = A\Sigma_{\mathbf{XX}}A^T = AA^T.$$

The converse causal hypothesis $\mathbf{Y} \to \mathbf{X}$ becomes implausible because $P(\mathbf{Y})$ (which is determined by the covariance matrix AA^T) and $P(\mathbf{X}|\mathbf{Y})$ (which is described by $\mathbf{X} = A^{-1}\mathbf{Y}$) are related in a suspicious way: the mechanism from $\mathbf{Y}$ to $\mathbf{X}$ seems to be adjusted to the distribution $P(\mathbf{Y})$ because it exactly stretches the directions with small variance and shrinks the ones with large variance "in order" to get an *isotropic* output $P(\mathbf{X})$. This can, indeed, be the result of a "designed" mechanism, but it is unlikely to be obtained by a simple process in nature having no feedback loops.[1]

The "atypical" relation between $\Sigma_{\mathbf{YY}}$ and A^{-1} can also be phrased in terms of symmetries: $A^{-1}\Sigma_{\mathbf{YY}}A^{-T}$ (here we have used the short notation $A^{-T} := (A^{-1})^T$) is surprisingly isotropic compared to those matrices obtained by applying A^{-1} to $U\Sigma_{\mathbf{YY}}U^T$ for some *generic* orthogonal transformation $U \in O(n)$. We will show below that this remains true with high probability (in high dimensions) if we start with an *arbitrary* covariance matrix $\Sigma_{\mathbf{XX}}$ and apply a random linear transformation A chosen independently of $\Sigma_{\mathbf{XX}}$.

To understand in what sense independent choices of $\Sigma_{\mathbf{XX}}$ and A typically induce atypical relations between A^{-1} and $\Sigma_{\mathbf{YY}}$ we also discuss a second example where $\Sigma_{\mathbf{XX}}$ and A are simultaneously diagonal with c_j and a_j ($j = 1, \ldots, n$) as corresponding diagonal entries. Thus $\Sigma_{\mathbf{YY}}$ is also diagonal and its diagonal entries (which equal its eigenvalues) are $a_j^2 c_j$. We now assume that "nature has chosen" the values c_j independently from some distribution and the a_j independently from some other distribution. We can then interpret the values c_j as instances of n-fold sampling of the random variable c with expectation $\mathbb{E}(c)$ and the same for a_j. If we assume that a and c are independent, we have

$$\mathbb{E}(a^2 c) = \mathbb{E}(a^2)\mathbb{E}(c)\,. \tag{1}$$

Due to the law of large numbers, this equation will for large n approximately be satisfied by the empirical averages, i.e.,

$$\frac{1}{n}\sum_{j=1}^{n} a_j^2 c_j \approx \left(\frac{1}{n}\sum_{j=1}^{n} a_j^2\right)\left(\frac{1}{n}\sum_{j=1}^{n} c_j\right). \tag{2}$$

For the backward direction $\mathbf{Y} \to \mathbf{X}$ we observe that the diagonal entries $\tilde{c}_j = a_j^2 c_j$ of $\Sigma_{\mathbf{YY}}$ and the diagonal entries $\tilde{a}_j = a_j^{-1}$ of $\tilde{A} := A^{-1}$ have not been chosen independently because $\mathbb{E}(\tilde{a}^2\tilde{c}) = \mathbb{E}(a^{-2}a^2 c) = \mathbb{E}(c)$, whereas

$$\begin{aligned}\mathbb{E}(\tilde{a}^2)\mathbb{E}(\tilde{c}) &= \mathbb{E}(a^{-2})\mathbb{E}(a^2 c)\\ &= \mathbb{E}(a^{-2})\mathbb{E}(a^2)\mathbb{E}(c) > \mathbb{E}(c)\,.\end{aligned}$$

The last inequality holds because the random variables a^2 and a^{-2} are always negatively correlated (from the Cauchy-Schwarz inequality we obtain $\mathbb{E}(a^2)\mathbb{E}(a^{-2}) \geq 1$ with equality only for the trivial case where a is constant). We thus observe a systematic violation of (1) in the backward direction. The proof for non-diagonal matrices in Section 2 uses spectral theory, and is based upon the same intuition.

The paper is structured as follows. In Section 2, we show that the above mentioned atypical relations between covariance and structure matrices can be detected via a simple trace formula. In Section 3 we describe an algorithm that is based upon this result and

[1] The argument that complex processes like evolution may be able to develop such intelligent system design is certainly correct, but this is a problem for all approaches to causal inference from non-experimental data.

in Section 4 discuss experiments with simulated and real data. Section 5 proposes possible generalizations.

2. Identifiability results

Given a hypothetical linear causal model $\mathbf{Y} = A\mathbf{X}+\mathbf{E}$ (where $\mathbf{X}$ and $\mathbf{Y}$ are n- and m-dimensional, respectively) we want to check whether the pair $(\Sigma_{\mathbf{XX}}, A)$ satisfies some relation that typical pairs $(U\Sigma_{\mathbf{XX}}U^T, A)$ only satisfy with low probability if $U \in O(n)$ is randomly chosen. To this end, we introduce the renormalized trace

$$\tau_n(.) := \mathtt{tr}(.)/n$$

for dimension n and compare the values

$$\tau_m(A\Sigma_{\mathbf{XX}}A^T) \quad \text{and} \quad \tau_n(\Sigma_{\mathbf{XX}})\tau_m(AA^T)\,. \qquad (3)$$

One shows easily that the expectation of both values coincide if $\Sigma_{\mathbf{XX}}$ is randomly drawn from a distribution that is invariant under transformations

$$\Sigma_{\mathbf{XX}} \mapsto U\Sigma_{\mathbf{XX}}U^T.$$

This is because averaging the matrices $U\Sigma_{\mathbf{XX}}U^T$ over all $U \in O(n)$ projects onto $\tau_n(\Sigma_{\mathbf{XX}})\mathbf{I}$ since the average $U\Sigma_{\mathbf{XX}}U^T$ commutes with all matrices and is therefore a multiple of the identity. For our purposes, it is decisive that the typical case is close to this average, i.e., the two expressions in (3) almost coincide. To show this, we need the following result (Ledoux, 2001):

Lemma 1 (Lévy's Lemma)
Let $g : S_n \to \mathbb{R}$ be a Lipschitz continuous function on the n-dimensional sphere with

$$L := \max_{\gamma\neq\gamma'} \frac{|g(\gamma) - g(\gamma')|}{\|\gamma - \gamma'\|}\,.$$

If a point γ on S_n is randomly chosen according to an $O(n)$-invariant prior, it satisfies

$$|g(\gamma) - \bar{g}| \leq \epsilon$$

with probability at least $1 - \exp(-\kappa(n-1)\epsilon^2/L^2)$ for some constant κ, where $\bar{g}$ can be interpreted as the median or the average of $g(\gamma)$.

Given the above Lemma, we can prove the following Theorem:

Theorem 1 (multiplicativity of traces)
Let Σ be a symmetric, positive definite $n \times n$-matrix and A an arbitrary $m \times n$-matrix. Let U be randomly chosen from $O(n)$ according to the unique $O(n)$-invariant distribution (i.e. the Haar measure). Introducing the operator norm $\|B\| := \max_{\|x\|=1} \|Bx\|$, we have

$$|\tau_m(AU\Sigma U^TA^T) - \tau_n(\Sigma)\tau_m(AA^T)| \leq 2\epsilon\|\Sigma\|\|AA^T\|$$

with probability at least $q := 1 - \exp(-\kappa(n-1)\epsilon^2)$ for some constant κ (independent of $\Sigma, A, n, m, \epsilon$).

Proof: for an arbitrary orthonormal system $(\psi_j)_{j=1,\ldots,m}$ we have

$$\tau_m(AU\Sigma U^TA^T) = \frac{1}{m}\sum_{j=1}^{m}\langle\psi_j, AU\Sigma U^TA^T\psi_j\rangle\,.$$

We define the unit vectors $\gamma_j := U^TA^T\psi_j/\|A^T\psi_j\|$. Dropping the index j, we introduce the function $f(\gamma) := \langle\gamma, \Sigma\gamma\rangle$. For a randomly chosen $U \in O(n)$, γ is a randomly chosen unit vector according to a uniform prior on the n-dimensional sphere S_n.

The average of f is given by $\bar{f} = \tau_n(\Sigma)$. The Lipschitz constant is given by the operator norm of Σ, i.e., $L = 2\|\Sigma\|$. An arbitrarily chosen j satisfies

$$|\langle\gamma_j, \Sigma\gamma_j\rangle - \tau_n(\Sigma)| \leq 2\epsilon\|\Sigma\|$$

with probability $1 - \exp(-\kappa(n-1)\epsilon^2)$. This follows from Lemma 1 after replacing ϵ with ϵL. Hence

$$\begin{aligned}&|\langle\psi_j, AU\Sigma U^TA^T\psi_j\rangle - \tau_n(\Sigma)\langle\psi_j, AA^T\psi_j\rangle| \\ &\leq 2\epsilon\|\Sigma\|\|AA^T\|\,.\end{aligned}$$

Due to

$$\tau_m(AU\Sigma U^TA^T) = \frac{1}{m}\sum_{j=1}^{m}\langle\psi_j, AA^T\psi_j\rangle\langle\gamma_j, \Sigma\gamma_j\rangle\,,$$

we thus have

$$|\tau_m(AU\Sigma U^TA^T) - \tau_m(AA^T)\tau_n(\Sigma)| \quad \leq \quad 2\epsilon\|\Sigma\|\|AA^T\|\,.$$

□

It is convenient to introduce

$$\Delta(\Sigma, A) := \log\tau_m(A\Sigma A^T) - \log\tau_n(\Sigma) - \log\tau_m(AA^T)$$

as a scale-invariant measure for the strength of the violation of the equality of the expressions (3). We will also write $\Delta_{\mathbf{X}\to\mathbf{Y}}$ if Σ is the covariance matrix of $\mathbf{X}$ and A the structure matrix defining the linear model from $\mathbf{X}$ to $\mathbf{Y}$. Note that Δ vanishes for dimension one, our method is certainly not able to distinguish between cause and effect for just two one-dimensional variables.

We will assume that Δ can be used to detect the causal direction because we expect $\Delta \approx 0$ for the correct one (due to Theorem 1). Certainly, Δ can also be close to zero for the *wrong* direction, but our theory and experiments will suggest that this rarely happens.

First, we restrict the attention to deterministic models

$$\mathbf{Y} = A\mathbf{X}$$

and the case where $n \leq m$ and A has rank n. This ensures that the backward model is also deterministic, i.e.,

$$\mathbf{X} = A^{-1}\mathbf{Y}\,,$$

with $(.)^{-1}$ denoting the pseudo inverse. The following theorem shows that $\Delta_{\mathbf{X}\to\mathbf{Y}} = 0$ then implies $\Delta_{\mathbf{Y}\to\mathbf{X}} \leq 0$:

Theorem 2 (violation of trace formula)
Let n, m with $n \leq m$ denote the dimensions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. If $\mathbf{Y} = A\mathbf{X}$ and $\mathbf{X} = A^{-1}\mathbf{Y}$, the covariance matrices satisfy

$$\begin{aligned}&\Delta_{\mathbf{X}\to\mathbf{Y}} + \Delta_{\mathbf{Y}\to\mathbf{X}} \\ &= -\log\left(1 - \mathtt{Cov}(\mathbf{Z}, 1/\mathbf{Z})\right) + \log\frac{n}{m}\,,\end{aligned} \tag{4}$$

where $\mathbf{Z}$ is a real-valued random variable whose distribution is the empirical distribution of eigenvalues of AA^T, i.e., $\tau_m((AA^T)^k) = \mathbb{E}(\mathbf{Z}^k)$ for all $k \in \mathbb{Z}$.

Proof: We have

$$\begin{aligned}&\frac{\tau_n(\Sigma)}{\tau_m(A\Sigma A^T)\tau_n(A^{-1}A^{-T})} \\ &= \frac{1}{\tau_m(AA^T)\tau_n(A^{-1}A^{-T})}\,\frac{\tau_n(\Sigma)\tau_m(A^TA)}{\tau_m(A\Sigma A^T)}\,.\end{aligned} \tag{5}$$

Using

$$\begin{aligned}\tau_n(A^{-1}A^{-T}) &= \tau_n(A^{-T}A^{-1}) = \tau_n((AA^T)^{-1}) \\ &= \frac{m}{n}\tau_m((AA^T)^{-1})\end{aligned}$$

and taking the logarithm we obtain

$$\Delta_{\mathbf{Y}\to\mathbf{X}} = \log\frac{1}{\mathbb{E}(\mathbf{Z})\mathbb{E}(1/\mathbf{Z})} + \log\frac{n}{m} - \Delta(\Sigma, A)\,.$$

Then the statement follows from

$$\mathtt{Cov}(\mathbf{Z}, 1/\mathbf{Z}) = 1 - \mathbb{E}(\mathbf{Z})\mathbb{E}(1/\mathbf{Z})\,.$$

□

If $|\Delta_{\mathbf{X}\to\mathbf{Y}}|$ is smaller than the absolute value of the right hand side of eq. (4), we obtain a non-trivial lower bound on $|\Delta_{\mathbf{Y}\to\mathbf{X}}|$. To show that this bound need not be small in high dimensions, we consider the case $n = m$ and a sequence of $n \times n$ random matrices (A_n) whose eigenvalue distributions Z_n converge to some distribution on $\mathbb{R}$ with non-zero variance. $-\log\left(1 - \mathtt{Cov}(\mathbf{Z}, 1/\mathbf{Z})\right)$ then converges to some negative value.

We should, however, mention a problem that occurs for $m > n$ in the noise-less case discussed here: Since $\Sigma_{\mathbf{YY}}$ has only rank n, we could equally well replace A^{-1} with some other matrix $\hat{A}$ that coincides with A^{-1} on all of the observed y-values. For those matrices $\hat{A}$, the value Δ can get closer to zero because the term $\log n/m$ expresses the fact that the image of $\Sigma_{\mathbf{YY}}$ is orthogonal to the kernel of A^{-1}, which is already an atypical relation.

It turns out that the observed violation of the multiplicativity of traces can be interpreted in terms of relative entropy distances. To show this, we need the following result:

Lemma 2 (anisotropy and relative entropy)
Let Σ be the covariance matrix of a centralized non-degenerate multi-variate Gaussian distribution P_Σ in n dimensions. Let the anisotropy of Σ be defined by the relative entropy distance to the closest isotropic Gaussian

$$D(\Sigma) := \min_{Q \text{ isotropic}} D(P_\Sigma||Q)\,.$$

Then

$$D(\Sigma) = \frac{1}{2}\left(n\log\tau_n(\Sigma) - \log\mathtt{det}(\Sigma)\right)\,. \tag{6}$$

Proof: the relative entropy distance of two centralized Gaussians with covariance matrices Σ, Σ_0 in n dimensions is given by

$$D(P_\Sigma||P_{\Sigma_0}) = \frac{1}{2}\left(\log\left(\frac{\mathtt{det}\Sigma_0}{\mathtt{det}\Sigma}\right) + \mathtt{tr}(\Sigma_0^{-1}\Sigma) - n\right)\,.$$

Setting $\Sigma_0 = \lambda\mathbf{I}$, the distance is minimized for $\lambda = \tau_n(\Sigma)$, which yields eq. (6). □

Straightforward computations show:

Theorem 3 (multiplicativity and rel. entropy)
Let Σ and A be $n \times n$-matrices with Σ positive definite. Then

$$D(A\Sigma A^T) = D(\Sigma) + D(AA^T) + \frac{n}{2}\Delta(\Sigma, A)\,.$$

Hence, for independently chosen A and Σ, the anisotropy of the output covariance matrix $A\Sigma A^T$ is approximately given by the anisotropy of Σ plus the anisotropy of AA^T, which is the anisotropy of the output that A induces on an isotropic input. For the backward direction, the anisotropy is smaller than the typical value.

We now discuss an example with a stochastic relation between $\mathbf{X}$ and $\mathbf{Y}$. We first consider the general linear model

$$\mathbf{Y} = A\mathbf{X} + \mathbf{E}\,,$$

where A is an $m \times n$ matrix and $\mathbf{E}$ is a noise term (statistically independent of $\mathbf{X}$) with covariance matrix $\Sigma_{\mathbf{EE}}$. We obtain

$$\Sigma_{\mathbf{YY}} = A\Sigma_{\mathbf{XX}}A^T + \Sigma_{\mathbf{EE}}\,.$$

The corresponding backward model[2] reads

$$\mathbf{X} = \tilde{A}\mathbf{Y} + \tilde{\mathbf{E}}\,.$$

with

$$\tilde{A} := \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\,.$$

Now we focus on the special case where A is an orthogonal transformation and $\mathbf{E}$ is isotropic, i.e., $\Sigma_{\mathbf{EE}} = \lambda\mathbf{I}$ with $\lambda > 0$. Then $\Delta_{\mathbf{Y}\to\mathbf{X}}$ is *positive*:

Lemma 3 (noisy case)
Let $\mathbf{Y} = A\mathbf{X} + \mathbf{E}$ with $A \in O(n)$ and the covariance matrix of $\mathbf{E}$ be given by $\Sigma_{\mathbf{EE}} = \lambda\mathbf{I}$. Then we have

$$\Delta_{\mathbf{Y}\to\mathbf{X}} > 0\,.$$

Proof: We have $\Sigma_{\mathbf{YY}} = A\Sigma A^T + \lambda\mathbf{I}$ and $\Sigma_{\mathbf{YX}} = A\Sigma$, with $\Sigma := \Sigma_{\mathbf{XX}}$. Therefore,

$$\tilde{A} = \Sigma A^T(A\Sigma A^T + \lambda\mathbf{I})^{-1} = \Sigma(\Sigma + \lambda\mathbf{I})^{-1}A^T.$$

One checks easily that the orthogonal transformation A is irrelevant for the traces and we thus have

$$\begin{aligned}\Delta_{\mathbf{Y}\to\mathbf{X}} &= \log\frac{\tau(\Sigma^2(\Sigma+\lambda\mathbf{I})^{-1})}{\tau(\Sigma+\lambda\mathbf{I})\tau(\Sigma^2(\Sigma+\lambda\mathbf{I})^{-2})}\\ &= \log\frac{\mathbb{E}\left(\mathbf{Z}^2/(\mathbf{Z}+\lambda)\right)}{\mathbb{E}(\mathbf{Z}+\lambda)\mathbb{E}\left(\mathbf{Z}^2/(\mathbf{Z}+\lambda)^2\right)}\,,\end{aligned}$$

where $\mathbf{Z}$ is a random variable of which distribution reflects the distribution of eigenvalues of Σ. The function $z \mapsto z/(z+\lambda)$ is monotonously increasing for positive λ and z and thus also $z \mapsto z^2/(z+\lambda)^2$. Hence $\mathbf{Z}+\lambda$ and $\mathbf{Z}^2/(\mathbf{Z}+\lambda)^2$ are positively correlated, i.e.,

$$\begin{aligned}\mathbb{E}(\mathbf{Z}^2/(\mathbf{Z}+\lambda)) &= \mathbb{E}((\mathbf{Z}+\lambda)\mathbf{Z}^2/(\mathbf{Z}+\lambda)^2)\\ &> \mathbb{E}(\mathbf{Z}+\lambda)\mathbb{E}\left(\mathbf{Z}^2/(\mathbf{Z}+\lambda)^2\right)\,,\end{aligned}$$

for all distributions of $\mathbf{Z}$ with non-zero variance. Hence the logarithm is positive and thus $\Delta_{\mathbf{Y}\to\mathbf{X}} > 0$. □

Theorem 2 and Lemma 3 show that independent choices of A and $\Sigma_{\mathbf{XX}}$ can induce positive or negative values of Δ in the wrong direction. We therefore propose to prefer the causal direction for which Δ is closer to zero.

[2]For non-Gaussian $\mathbf{X}$ or $\mathbf{E}$, this induces a joint distribution $P(\mathbf{X},\mathbf{Y})$ that does not admit a linear backward model with an *independent* noise $\tilde{\mathbf{E}}$, we can then only obtain *uncorrelated* noise. We could in principle already use this fact for causal inference (Kano & Shimizu, 2003). However, our method also works for the Gaussian case or if the dimension is too high for testing higher-order statistical dependences reliably.

Algorithm 1 Identifying linear causal relations via traces

Input: $(x_1, y_1), \ldots, (x_k, y_k), \epsilon$
Compute the estimators $\Sigma_{\mathbf{XX}}$, $\Sigma_{\mathbf{XY}}$, $\Sigma_{\mathbf{YX}}$, $\Sigma_{\mathbf{YY}}$
Compute $A := \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}$
Compute $\tilde{A} := \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}$
if $|\Delta_{\mathbf{Y}\to\mathbf{X}}| > (1+\epsilon)|\Delta_{\mathbf{X}\to\mathbf{Y}}|$ **then**
  write "$\mathbf{X}$ is the cause"
else
  if $|\Delta_{\mathbf{X}\to\mathbf{Y}}| > (1+\epsilon)|\Delta_{\mathbf{Y}\to\mathbf{X}}|$ **then**
    write "$\mathbf{Y}$ is the cause"
  else
    write "cause cannot be identified"
  end if
end if

3. Inference algorithm

Motivated by the above theoretical results, we propose to infer the causal direction using Alg. 1.[3]

In light of the theoretical results, the following issues have to be clarified by experiments with simulated data:

1. Is the limit for dimension to infinity already justified for moderate dimensions?
2. Is the multiplicativity of traces sufficiently violated for noisy models? – Note that Lemma 3 only covers a special case.

Furthermore, the following issue has to be clarified by experiments with real data:

3. Is the behaviour of real causal structures qualitatively sufficiently close to our model with independent choices of A and $\Sigma_{\mathbf{XX}}$ according to an isotropic prior? How large must we choose the threshold ϵ in Alg. 1 to get reliable results?

4. Experiments

4.1. Simulated data

We have generated random models $\mathbf{Y} = A\mathbf{X} + \mathbf{E}$ as follows: We independently draw each element of the $m \times n$ structure matrix A from a standardized Gaussian distribution. This implies that the distribution of column vectors as well as the distribution of row vectors is isotropic. To generate a random covariance

[3]A code package implementing this algorithm (and reproducing all experiments reported in this paper), is available at: `http://www.kyb.tuebingen.mpg.de/causality/`

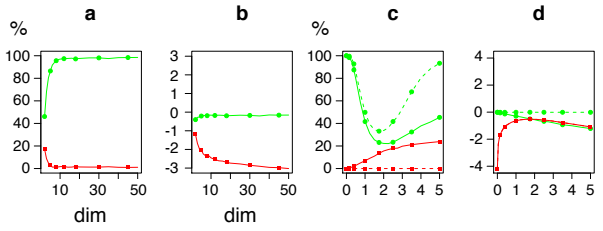

Figure 1. Simulation results. **(a)** Performance of the method as a function of the input dimensionality n, when the output dimensionality $m = n$ and the sample size is $N = 2n$. The green curve (circles) denotes the fraction of simulations on which the true causal direction was selected, while the red curve (squares) gives the fraction of wrong answers. **(b)** Mean values of Δ corresponding to the true direction (green) vs the wrong direction (red). **(c)** Performance as a function of noise level σ, for dimensionality $n = m = 10$ and sample size $N = 1000$. To compare, the dashed lines give the performance based on the exact covariance matrices rather than based on the samples. **(d)** Mean values of Δ corresponding to the true direction (green) vs the wrong direction (red). See main text for discussion.

matrix $\Sigma_{\mathbf{XX}}$, we similarly draw an $n \times n$ matrix B and set $\Sigma_{\mathbf{XX}} := BB^T$. Due to the invariance of our decision rule with respect to the scaling of A and $\Sigma_{\mathbf{XX}}$, the structure matrix and the covariance can have the same scale without loss of generality. The covariance $\Sigma_{\mathbf{EE}}$ of the noise is generated in the same way, although with an adjustable parameter σ governing the scaling of the noise with respect to the signal: $\sigma = 0$ yields the deterministic setting, while $\sigma = 1$ equates the power of the noise to that of the signal.

First, we demonstrate the performance of the method in the close-to deterministic setting ($\sigma = 0.05$) as a function of the dimensionality $n = m$ of the simulations, ranging from dimension 2 to 50. To show that the method is feasible even with a relatively small number of samples, we choose the number of samples N to scale with the dimension as $N = 2n$. (Note that we must have $N \geq \min(n, m)$ to obtain invertible estimates of the covariance matrices.) The resulting proportion of correct vs wrong decisions is given in Fig. 1a, with the corresponding values of Δ in Fig. 1b. As can be seen, even at as few as 5 dimensions and 10 samples, the method is able to reliably identify the direction of causality in these simulations.

To illustrate the degree to which identifiability is hampered by noise, the solid line in Fig. 1c gives the performance of the method for a fixed dimension ($n = m = 10$) and fixed sample size ($N = 1000$) as a function of the noise level σ. As can be seen, the performance drops markedly as σ is increased. As soon as there is significantly more noise than signal (say, $\sigma > 2$), the number of samples is not sufficient to reliably estimate the required covariance matrices and hence the direction of causality. This is clear from looking at the much better performance of the method when based on the exact, true covariance matrices, given by the dashed lines. In Fig. 1d we show the corresponding values of Δ, from which it is clear that the estimate based on the samples is quite biased for the forward direction.

The simulations point out at least one important issue for future work: the construction of unbiased estimators for the trace values or the Δ. The systematic deviation of the sample-based experiments from the covariance-matrix based experiments in Fig. 1c–d suggest that this could be a major improvement.

4.2. Handwritten digits

As experiments with real data with known ground truth, we have chosen 16×16 pixel images (so $n = 256$) of handwritten digits (LeCun et al., 1990). As the linear map A we have used both random local translation-invariant linear filters and also standard blurring of the images. (We added a small amount of noise to both original and processed images, to avoid problems with very close-to singular covariances.) The task is then: given a sample of pairs (x_j, y_j), each consisting of the picture x_j and its processed counterpart y_j, infer which of the set of pictures x or y are the originals ('causes'). By partitioning the image set by the digit class (0-9), and by testing a variety of random filters (and the standard blur), we obtained a number of test cases to run our algorithm on. Out of the total of 100 tested cases, the method was able to correctly identify the set of original images 94 times, with 4 unknowns (i.e. only two falsely classified cases).

4.3. Geographic position and precipitation

In the first experiment we took precipitation data from 4748 different locations in Germany[4]. The cause $\mathbf{X}$ was 3-dimensional and consisted of

$$\mathbf{X} = (\texttt{altitude}, \texttt{longitude}, \texttt{latitude}) .$$

The effect $\mathbf{Y}$ was 12-dimensional and consisted of the average precipitation per month:

$$\mathbf{Y} = (\texttt{prec. in Jan.}, \ldots, \texttt{prec. in Dec.}) .$$

Here we are faced with the following problem. The scales of $\mathbf{X}_1$ are uncomparable to those of $\mathbf{X}_2$ and $\mathbf{X}_3$

[4] http://www.dwd.de

since the quantities are measured in different units. We have therefore renormalized all $\mathbf{X}_j$ to unit variance. One may debate whether one should then also renormalize $\mathbf{Y}$ (even though all $\mathbf{Y}_j$ refer to the same unit) in order to have a method that treats $\mathbf{X}$ and $\mathbf{Y}$ in the same way. We have done two runs, one with only renormalizing $\mathbf{X}$ (top row of table below) and the second with renormalizing $\mathbf{X}$ and $\mathbf{Y}$ (bottom row of table below).

$$\begin{array}{rclcrcl} \Delta_{\mathbf{X}\to\mathbf{Y}} & = & -0.240; & & \Delta_{\mathbf{Y}\to\mathbf{X}} & = & -2.25 \\ \Delta_{\mathbf{X}\to\mathbf{Y}} & = & -0.278; & & \Delta_{\mathbf{Y}\to\mathbf{X}} & = & -2.11 \end{array}$$

The results qualitatively coincide for both versions and our method prefers the correct causal direction $\mathbf{X} \to \mathbf{Y}$. We will henceforth renormalize whenever a vector contains quantities of different units, even if some of the units coincide.

We also tried the same experiment with 3-dimensional precipitation vector $\mathbf{Y}$ containing only months j to $j+2$ for all $j = 1, \ldots, 10$. In all these 10 experiments we also obtained correct results.

4.4. Weather and air pollution

In another experiment we used data on the relation between weather conditions and air pollution in Chemnitz, Germany, measured at 1440 days.[5] We defined a 3-dimensional vector of weather conditions

$$\mathbf{X} = (\sin(\phi_{wind}), \cos(\phi_{wind}), T)\,,$$

where ϕ_{wind} is the direction of the wind and T is the air temperature. We then define a 6-dimensional vector "air pollution"

$$\mathbf{Y} = (\texttt{ozone}, \texttt{sulfur dioxid}, \texttt{dust}, \texttt{CO}, \texttt{NO}_2, \texttt{NO}_\texttt{x})\,.$$

Clearly, the wind can blow the pollutants away from or to the location of measurement, depending on its direction. Moreover, the production of ozone is known to be a problem of days with strong sun radiation. We therefore assume the ground truth to be $\mathbf{X} \to \mathbf{Y}$. Here, we must obviously renormalize $\mathbf{X}$ and $\mathbf{Y}$. We obtained the result

$$\Delta_{\mathbf{X}\to\mathbf{Y}} = -0.0504; \quad \Delta_{\mathbf{Y}\to\mathbf{X}} = -1.44\,,$$

and thus prefer the correct causal direction. To exclude a systematic preference of the lower dimension as the cause we have also defined 3-dimensional vectors $(\mathbf{Y}_j, \mathbf{Y}_{j+1}, \mathbf{Y}_{j+2})$ and obtained the correct result for all $j = 1, \ldots, 4$. The dataset also contained a variable wind strength, but it was discretized into 3 values. We have therefore dropped it because our method actually refers to continuous variables only. However, it is noteworthy that our method actually assumes that the data points are independently drawn from $P(\mathbf{X}, \mathbf{Y})$, rather than being part of a time series as it is the case here and in the following experiment.

[5] http://www.mathe.tu-freiberg.de/Stoyan/Umweltdaten/chemnitz.txt

4.5. Stock returns

We also ran an experiment with the relations between the daily stock returns in different countries from the Yahoo finance database at 2394 days (from January 04, 2000 to March 10, 2009). We defined

$$\mathbf{X} := (\texttt{SH}, \texttt{HSI}, \texttt{TWI}, \texttt{N225})$$

where the abbreviations stand for the stock returns of Shanghai (China), Hang Seng (Hong Kong), Taiwan, Nikkei (Japan). Moreover, we combine the European indices

$$\mathbf{Y} := (\texttt{FTSE}, \texttt{DAX}, \texttt{CAC})\,,$$

corresponding to England, Germany, and France, respectively. Due to the different time zones, the European returns at the same day refer to a later time than the Asian ones. We therefore assume that the ground truth is $\mathbf{X} \to \mathbf{Y}$. The results (without renormalizing covariance matrices) are

$$\Delta_{\mathbf{X}\to\mathbf{Y}} = 0.408; \quad \Delta_{\mathbf{Y}\to\mathbf{X}} = 0.363\,.$$

Here we infer the wrong direction but the difference between the values is small. When applying a conservative decision rule (via setting the threshold appropriately) we would not make a decision here.

To get vectors of higher dimensions we combined the returns of Europe and the USA (DJ and Nas) because both refer to a later time than the Asian ones.

$$\mathbf{Y} := (\texttt{FTSE}, \texttt{DAX}, \texttt{CAC}, \texttt{DJ}, \texttt{Nas})\,.$$

We obtain

$$\Delta_{\mathbf{X}\to\mathbf{Y}} = 0.312; \quad \Delta_{\mathbf{Y}\to\mathbf{X}} = 0.147\,,$$

which also infers the wrong direction (with higher significance than before). The problem might be the following. The assumption that "the market chooses" $\Sigma_{\mathbf{XX}}$ and A independently is questionable since $\Sigma_{\mathbf{XX}}$ is the result of complex market phenomena that also depend on what happened the day before (which was also determined by A). To further explore the robustness of our method with respect to violating the model assumptions must be left to the future.

Since the threshold ϵ in Alg. 1 is hard to interpret one may prefer a decision rule that is closer to standard

statistical hypothesis testing: generate a large number of random orthogonal transformations U, apply them to $\mathbf{X}$ (while keeping A) and consider the distribution of all these values $\Delta_{U\mathbf{X}\to\mathbf{Y}}$. The observed value then defines a p-value and we can infer directions by just comparing p-values. We have done preliminary studies for some of the above examples and obtained qualitatively the same results.

5. Outlook: generalizations of the method

In this section, we want to place our theoretical results in a more general context and state:

Postulate 1 (typicality for group orbits)
Let $\mathbf{X}$ *and* $\mathbf{Y}$ *be random variables with joint distribution* $P(\mathbf{X},\mathbf{Y})$ *and* G *be a group of transformations on the range of* $\mathbf{X}$*. Each* $g \in G$ *defines a modified distribution of* $\mathbf{Y}$ *via* $P^{(g)}(\mathbf{Y}) := \sum_x P(\mathbf{Y}|g(x))P(x)$*. Let* $K(.)$ *be some real-valued function on the probability distributions of* $\mathbf{Y}$*. The causal hypothesis* $\mathbf{X}\to\mathbf{Y}$ *is unlikely if* $K(P(\mathbf{Y}))$ *is smaller or greater than the big majority of all distributions* $(P^{(g)}(\mathbf{Y}))_{g\in G}$*.*

Our prior knowledge about the structure of the data set determines the appropriate choice of G. The idea is that G expresses a set of transformations that generate input distributions $P_g(\mathbf{X})$ that we consider equally likely. We have chosen $K(P(\mathbf{Y})) := \tau(\Sigma_{\mathbf{YY}})$ and $G = O(n)$, but the permutation of components of $\mathbf{X}$ also defines an interesting transformation group. For time series, the translation group would be the most natural choice.

Interpreting this approach in a Bayesian way, we thus use symmetry properties of priors without the need to explicitly define the priors themselves.

6. Discussion

Our experiments with simulated data suggest that the method performs quite well already for moderate dimensions provided that the noise level is not too high. Certainly, the model of drawing $\Sigma_{\mathbf{XX}}$ according to a distribution that is invariant under $\Sigma_{\mathbf{XX}} \mapsto U\Sigma_{\mathbf{XX}}U^T$ may be inappropriate for many practical applications. Even though the example with diagonal matrices in Section 1 shows that the statement $\Delta \approx 0$ holds for a much broader class of models, it remains to show that most cause-effect pairs in the real world indeed satisfy $\Delta \approx 0$ for the true causal direction. Our studies with empirical data are still preliminary in this respect. It is possible that the method presented here only shows a future direction for the development of more mature causal inference algorithms.

It would be interesting to know whether our method could be a sanity check for complex causal networks: Consider, e.g., a causal DAG G connecting $2n$ real-valued variables. If $\mathbf{X}_1,\ldots,\mathbf{X}_{2n}$ is an ordering that is consistent with G, we define $\mathbf{Y} := (\mathbf{X}_1,\ldots,\mathbf{X}_n)$ and $\mathbf{W} := (\mathbf{X}_{n+1},\ldots,\mathbf{X}_{2n})$ and test whether $\Delta_{\mathbf{Y}\to\mathbf{W}} \approx 0$.

References

Heckerman, D., Meek, C., and Cooper, G. A Bayesian approach to causal discovery. In Glymour, C. and Cooper, G. (eds.), *Computation, Causation, and Discovery*, pp. 141–165, Cambridge, MA, 1999. MIT Press.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *NIPS 2008*.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. To appear in *IEEE Transactions on Information Theory*.

Kano, Y. and Shimizu, S. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pp. 261–270, Tokyo, Japan, 2003.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pp. 396–404. Morgan Kaufman, 1990.

Ledoux, M. *The concentration of measure phenomenon*. American Mathematical Society, 2001.

Lemeire, J. and Dirkx, E. Causal models as minimal descriptions of multivariate systems. `http://parallel.vub.ac.be/~jan/`, 2006.

Mooij, J. and Janzing, D. Distinguishing between cause and effect. *JMLR, W&CP*, 6:147–146, 2010.

Pearl, J. *Causality*. Cambridge University Press, 2000.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 7:2003–2030, 2006.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. Springer, New York, 1993.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. *UAI 2009*.