Causal inference by choosing graphs with most plausible Markov kernels

Xiaohai Sun* Dominik Janzing† Bernhard Schölkopf‡

November 17, 2005

Abstract

We propose a new inference rule for estimating causal structure that underlies the observed statistical dependencies among n random variables. Our method is based on comparing the conditional distributions of variables given their direct causes (the so-called "Markov kernels") for all hypothetical causal directions and choosing the most plausible one. We consider those Markov kernels most plausible, which maximize the (conditional) entropies constrained by their observed first moment (expectation) and second moments (variance and covariance with its direct causes) based on their given domain.

In this paper, we discuss our inference rule for causal relationships between two variables in detail, apply it to a real-world temperature data set with known causality and show that our method provides a correct result for the example.

1 Introduction

Causal inference plays a significant role in many areas of science, finance and industry. But how can causal knowledge be discovered automatically from non-experimental data? Given correlations among observed random variables there is in principle no method to identify causal relationships between the variables uniquely. Nevertheless there are some interesting inference rules [6, 9] that provide at least some hints on a causal relationship. The formal basis of these approaches are graphical models [5], where the random variables are the nodes of a directed acyclic graph (DAG) and an arrow from variable X to Y indicates that there is a direct causal influence from X to Y. The definition of "direct causal effect" from X to Y refers to a hypothetical intervention where all variables in the model except from X and Y are adjusted to fixed values and one observes whether the distribution of Y changes while X is adjusted to different values. As clarified by Pearl in full detail [6], the change of the distribution of Y in such an intervention cannot be derived from the joint distribution of all variables without defining a causal graph. The relation indicating whether there is a causal effect from X to Y is inherently asymmetric, because if X causes Y then intervening to change the value of X can change the disbribution of Y but intervening to change the value of Y cannot change the distribution of X, whereas statistical dependency defines a symmetric relation.

In Pearl's approach, the Markov condition is the essential axiom that unifies a causal structure and statistical dependencies among the variables. This assumption is based originally on the idea of the philosophers Reichenbach [7] and Salmon [8]. The Markov condition can be stated simply: Conditional on all its direct causes, a variable X is independent of every other variable except its effects. The intuition behind the causal Markov assumption is quite plausible: Ignoring a variable's

*xiaohai.sun@tuebingen.mpg.de, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

†janzing@ira.uka.de, Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

‡bernhard.schoelkopf@tuebingen.mpg.de, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

effects, all the relevant probabilistic information about a variable that can be obtained from a system is contained only in its direct causes. When one gives DAGs a causal interpretation, it then becomes necessary to argue that the Markov condition is in fact the correct connection between causal structure and probabilistic independence. A causal inference rule formulated by Pearl [6], Spirtes, Glymour and Scheines [9] is based on the principle to choose among all "possible" DAGs (in the sense that they satisfy the Markov condition) a causal graph that explains exactly (if possible) these conditional independencies that are entailed by the Markov condition or as many of them as possible.

However, any causal inference based on the Markov condition needs a threshold value for the decision of the statistical independency, which is chosen somehow arbitrarily. Moreover, there are often many distinct causal structures where the rules above do not allow to prefer one of them to the others. In an extreme case that there are no (conditional) independent relations among the observed variables, a causal inference based on the Markov condition is inapplicable, because the only possible causal structures are the complete acyclic graphs and there is no simplicity criterion that allows to prefer one of the $n!$ complete acyclic graphs on n variables to the others. In particular, one cannot determine the causal direction between two variables X, Y if only these two are observed, because both hypothetical causal directions ($X \to Y$ and $Y \to X$) can in principle generate all joint distributions. Our proposal is to try to capture the asymmetry of causality by the shape of conditional distributions on a hypothetical "true" causal graph. The hope is, roughly speaking, that the conditional distribution of an effect given all its causes is typically a "smoother" distribution than the distribution of the effect itself, and a cause itself has typically a much "smoother" distribution than any conditional distribution of the cause given some of its effects. Therefore, our approach will allow us to obtain some hints about causality despite the absence of conditional independencies. In particular, one could get some ideas to determine the causal direction already in case of only two observed variables.

2 Markov kernels of causal directions

We begin by introducing the concept of Markov kernels corresponding to a hypothetical causal direction. A Markov kernel formalizes the distribution of an effect given all its direct causes with respect to a given hypothetical causal graph $\mathcal{G}$. We characterize the joint distribution on n random variables $(X_1, \cdots, X_n)$ by all values

$$P(x_1, \cdots, x_n)$$

where $(x_1, \cdots, x_n)$ run over all possible values of $(X_1, \cdots, X_n)$ and interpret them as probabilities or probability densities according to whether it is a discrete or continuous variable. In general, the possible values of every variable in all our discussion might be either continuous or discrete. For the sake of simplicity and general computability, we assume in the following that the domain of each variable is discrete and finite, since data in the real world are mostly given with finite accuracy on a finite domain. For a continuous variable, the only change required is a suitable discretization with a proper scale where appropriate. Due to the causal Markov condition the joint measure can be factorized into

$$P(x_1, \cdots, x_n) = \prod_{j=1}^{n} P(x_j|pa_j)$$

where pa_j is a tuple of values of all k_j parents of X_j in $\mathcal{G}$. We call the conditional probabilities in the product the Markov kernels of P with respect to $\mathcal{G}$.

Actually, we only need to focus on complete acyclic causal graphs $\mathcal{K}$'s which are defined by an ordering of the nodes and drawing arrows from each node to all its successors. One can easily identify any causal graph $\mathcal{G}$ as an embedded subgraph in a suitable $\mathcal{K}$ by checking for each node X_j the set of its parents in $\mathcal{K}$ which can be dropped without changing the Markov kernels $P(x_j|pa_j)$ and consequently the joint distribution P. We call an ordering of variables a true causal direction if the corresponding complete graph $\mathcal{K}$ contains the true graph $\mathcal{G}$ as a subgraph. As a causal ordering consists of causal connections between a variable and all its parents, we will restrict our attention

in the following to identifying true causal directions between only two variables, which in general might be one- or multidimensional. Our method of estimating causation between two variables is based on the plausibility of the shape of the Markov kernels corresponding to a hypothetical causal direction.

3 Criteria for plausible Markov kernels

We allow ourselves to assume some "smoothness" conditions on Markov kernels and to make sense of the shapes of some plausible Markov kernels. We employ here the constrained maximum entropy approach to define smooth Markov kernels. Constrained entropy maximization is a widely used method for estimating a probability distribution. Collins, Downson and Wragg provided in [2, 3] a mathematical framework of the maximum Shannon entropy approach to assign a probability distribution on the basis of a limited number of moments. Although the intention is rather different in our setting, we refer to these articles for the mathematical framework. Furthermore, as the first and second moment can be estimated quite well from few data points, we determine the most plausible Markov kernel as follows: Given the (joint) distribution of all its parents $P(Pa_j)$, the most plausible Markov kernel of a variable X_j is the conditional probability that maximizes its conditional entropy constrained on the expectation and variance of X_j as well as the cross-covariance of X_j with all its parents Pa_j.

3.1 Plausible Markov kernel of causes

We consider the solution of the following optimization as the most plausible Markov kernel $P(X)$ of a vectorial cause variable $X = (X^{(1)}, \cdots, X^{(d_x)})$ with a domain $\mathcal{S}^x \subseteq \mathbb{R}^{d_x}$ in the causal direction $X \to Y$.

optimize	$\max_{P(X)} -\sum_x P(x) \cdot ln\,(P(x))$	(Entropy of $P(X)$)
subject to	$P(x) \geq 0 \quad \forall x \in \mathcal{S}^x$	(Non-negativity)
	$\sum_x P(x) = 1$	(Normalization)
	$\sum_x x \cdot P(x) = \mu$	(1st moment)
	$\sum_x x^{(i)} \cdot x^{(j)} \cdot P(x) = \gamma_{ij} \quad \forall i,j = 1, \cdots, d_x$	(2nd moment)

Here $\mathcal{S}^x$ denotes the domain of X, μ the first moment vector and $\gamma \equiv (\gamma_{ij})$ the second moment matrix of X. These values are estimated on the basis of the observed data. In a one-dimensional case ($d_x = 1$), μ is just the expectation and γ the second moment of X.

3.2 Plausible Markov kernel of effects

To determine a plausible Markov kernel $P(Y|X)$ for an effect variable $Y = (Y^{(1)}, \cdots, Y^{(d_y)})$ in the causal direction $X \to Y$ with $\mathcal{S}^y \subseteq \mathbb{R}^{d_y}$, we maximize the entropy of the conditional distribution of Y given X constrained by the expectation vector, the within-block covariance of Y as well as the cross-covariance of Y with its direct cause X. We consider the solution of the following optimization as the most plausible Markov kernel for Y.

optimize	$\max_{P(Y\|X)} -\sum_x \sum_y P(x) \cdot P(y\|x) \cdot ln(P(y\|x))$	(Entropy of $P(Y\|X)$)
subject to	$P(y\|x) \geq 0 \quad \forall (x,y) \in \mathcal{S}^x \times \mathcal{S}^y$	(Non-negativity)
	$\sum_y P(y\|x) = 1 \quad \forall x \in \mathcal{S}^x$	(Normalization)
	$\sum_x \sum_y y \cdot P(x) \cdot P(y\|x) = \mu$	(1st moment)
	$\sum_x \sum_y y^{(i)} \cdot y^{(j)} \cdot P(x) \cdot P(y\|x) = \gamma_{ij}$	(2nd moment)
	$\sum_x \sum_y y^{(k)} \cdot x^{(l)} \cdot P(x) \cdot P(y\|x) = \eta_{kl}$	(2nd mixed moment)
	$\forall i,j,k = 1, \cdots, d_y$ and $l = 1, \cdots, d_x$	

In this context the (joint) distribution of the cause variable $P(X)$ is given. $\mathcal{S}^y$ denotes the domain of Y, $\mu \in \mathbb{R}^{d_y}$ the first moment vector of Y, $\gamma \equiv (\gamma_{ij}) \in \mathbb{R}^{d_y \times d_y}$ the second moment matrix (within-block covariance) of Y and $\eta \equiv (\eta_{kl}) \in \mathbb{R}^{d_y \times d_x}$ the second mixed moment matrix (cross-covariance) of X and Y. Actually, in the optimization we need not take account of the non-negativity condition explicitly , because the logarithms in the objective function already implies this condition. The same applies also for the optimization problem in Section 3.1.

3.3 Analytic solutions for plausible Markov kernels

It is known that the optimization problems described in Subsections 3.1 and 3.2 are strictly convex [1], which ensures the existence of a unique optimal solution for them. In case of the continuous limit, these optimization problems can be formulated analogously. The only change required is a substitution of integration for summation when appropriate. In some special cases we can even find a closed-form solution. An example is the plausible Markov kernels for a causation between a binary variable $X = \{-1, +1\}$ and a one-dimensional real-valued variable Y.

For one hypothetical causal direction $X \to Y$, the plausible Markov kernels take the form of a Bernoulli distribution for the discrete cause X and Gaussian distributions with different expectations but the same variance for the continuous effect Y.

$$\begin{aligned} \mathcal{Q}(x_{-1}) = p \qquad & \text{and} \qquad \mathcal{Q}(x_{+1}) = 1 - p \\ \mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}\left(\mu_{-1}, \sigma^2\right) \qquad & \text{and} \qquad \mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}\left(\mu_{+1}, \sigma^2\right). \end{aligned}$$

For the other hypothetical causal direction $Y \to X$, the plausible Markov kernels are in form of a Gaussian distribution for a continuous cause Y and a family of hyperbolic tangent functions for a discrete effect X.

$$\begin{aligned} \mathcal{R}(Y) \qquad & \propto \qquad \mathcal{N}\left(\mu, \sigma_0^2\right) \\ \mathcal{R}(x_{-1}|Y) = \frac{1}{2} - \frac{1}{2} tanh(\lambda y + \nu) \qquad & \text{and} \qquad \mathcal{R}(x_{+1}|Y) = \frac{1}{2} + \frac{1}{2} tanh(\lambda y + \nu). \end{aligned}$$

The derivations are available in Appendix A. Note that for the causal direction $Y \to X$, the cause variable Y shows a unimodal Gaussian distribution, whereas for the other direction $X \to Y$ the plausible Markov kernels lead to a bimodal mixture Gaussian distribution for effect variable Y as its marginal distribution. That means, if we observe a variable with a mixture Gaussian distribution, it is more plausible to consider it as effect, because to regard the reverse as "true" causation, one must accept an unusual or contrived distribution as a natural or plausible Markov kernel. Therefore, the plausibility of Markov kernels might help us to guess the "true" causal direction.

However, due to the existence of awkward normalizing constants it is typically non-trivial to present an analytic solution for plausible Markov kernels in a closed form of some smooth function families. For example, the computation of $P(X|Y)$ requires for each given value y of Y a constraint that the probability or density of X should sum or integrate to 1, which will be awkward, if the set of possible values of Y becomes very large or infinite. Fortunately, if we admit a suitable discretization on a given continuous domain, the plausible Markov kernels can be always determined for a finite value set numerically.

4 Estimating causal direction based on maximum likelihood

Beginning with statistical information from data and all possible hypothetical complete causal graphs, we now turn to the subject of causal inference. Each hypothetical complete causal graph defines a unique causal ordering. For every hypothetical causal ordering $X_1, \cdots, X_n$, we compute a set of plausible Markov kernels $P(X_j|Pa_j)$ $(j = 1, \cdots, n)$ by maximizing the conditional entropy $\mathcal{H}(X_j|Pa_j)$ subject to a known joint distribution of all parents (direct causes) Pa_j and the cross-covariance of the effect X_j with all its direct causes as well as the first and second moment of

the effect X_j. Although such an approach is applied often to estimate distributions from data, a plausible Markov kernel should be regarded rather as a function with a causal interpretation, which comes from prior knowledge or assumption. It characterizes the likely impact of any intervention on a hypothetical causal graph. The intuition behind such entropy maximization is to complete the linear part of the effect of Pa_j on X_j by fixing it with the maximal uncertainty, since we consider, like the authors of [4], a linear causal relationship as a simplest form of causality.

We obtain the joint distribution of variables with hypothetical causal ordering $X_1, \ldots, X_n$ by, first, constrained maximization of the entropy $\mathcal{H}(X_1)$, next of the conditional entropy $\mathcal{H}(X_2|X_1)$, followed by $\mathcal{H}(X_3|X_2, X_1)$ and so on. The sum of all conditional entropies is the joint entropy and the constraints on expectations, variances and covariances coincide for all different causal orderings. But due to the order of maximizing it can happen that we obtain different joint distributions

$$P_{ordering_{X_1,\cdots,X_n}} = P(X_n|X_{n-1}, \cdots X_1) \cdots P(X_3|X_2, X_1) \cdot P(X_2|X_1) \cdot P(X_1)$$

with the same constraints. Here the order of maximizing conditional entropies matters. Having calculated the joint distributions from plausible Markov kernels based on all possible causal orderings, we apply the maximum likelihood approach to decide on these different orderings. We choose an ordering (causal graph) as "true", if its derived plausible Markov kernels lead to a joint distribution that has the maximum log-likelihood score by given observed data. The log-likelihood score tells us how strong the data support the hypothesis of causal ordering in the context of plausible Markov kernels.

In particular, for estimating the causal direction between only two observed variables X, Y we start out with both hypothetical causal orderings and calculate the plausible Markov kernels $\{\mathcal{Q}(X), \mathcal{Q}(Y|X)\}$ corresponding to $X \to Y$ (causal ordering X, Y) and the plausible Markov kernels $\{\mathcal{R}(Y), \mathcal{R}(X|Y)\}$ corresponding to $Y \to X$ (causal ordering Y, X). For the hypothetical causal direction $X \to Y$ we obtain a joint distribution

$$\mathcal{Q}_{X\to Y} = \mathcal{Q}(Y|X) \cdot \mathcal{Q}(X)$$

and we get for the other hypothetical causal direction $Y \to X$

$$\mathcal{R}_{Y\to X} = \mathcal{R}(X|Y) \cdot \mathcal{R}(Y).$$

Note that generally

$$\mathcal{Q}_{X\to Y} \neq \mathcal{R}_{Y\to X}$$

in a causal context of plausible Markov kernels. For example, one can show that the inequality holds in the case described in Section 3.3 whensoever correlations between variables are observed (see Appendix A for some more detail). By given data, we calculate the log-likelihood scores for $\mathcal{Q}$ (based on $X \to Y$) and $\mathcal{R}$ (based on $Y \to X$) respectively and choose the causal direction with larger log-likelihood as "true". This way, we hope to pick up not the symmetric dependency but the asymmetric causality between X and Y.

5 Real-world temperature data example

To test the effectiveness of our method, we examined the causation between dates of the year (*Date*) and daily average temperatures (*Temperature*) as an example. Common sense tells us that the seasonal cycle is a cause of temperature variation (Figure 1), not vice versa.

Figure 1: Causation from *Date* to *Temperature*

A real data set of daily average temperatures in Furtwangen (Black Forest, Germany) from Jan. 1, 1979 to Jan. 31, 2004 with 9162 entries was analyzed. Due to the cyclic property of dates of the year, we assign the unit circle, a proper subset of $\mathbb{R}^2$, to the domain of the variable *Date* (X, Y) with $\mathcal{S}^{Date} = \{(x, y)|x^2 + y^2 = 1\}$. This value set can be parameterized, for example, by $x = cos\left(\frac{2\pi}{366}k\right)$ and $y = sin\left(\frac{2\pi}{366}k\right)$ with $k = 1, \cdots, 366$ (maximum days per year).

Consequently, the first moment of *Date* is a two dimensional vector and states the expectations in X and Y. The second mixed moment of *Date* is also a two dimensional vector, which defines cross-covariance between (X, Y) and *Temperature*. The second moment of *Date* is a symmetric matrix, which fixes the within-block covariance of (X, Y). Table 1 summarizes all the statistical features from the data which we need for our optimization described in Sections 3.1 and 3.2.

	Date	Temperature
	(X, Y)	(degree Celsius)
Value set	$\{(x, y)\|x^2 + y^2 = 1\} \subseteq \mathbb{R}^2$	$[-23, 25] \subseteq \mathbb{R}$
1st moment	$(0.0022, -0.0009)$	5.7053
2nd moment	$\begin{pmatrix} 0.5019 & 0 \\ 0 & 0.4981 \end{pmatrix}$	84.6079
2nd mixed moment	$(-3.9702, -1.4548)$	

Table 1: Temperature data of Furtwangen

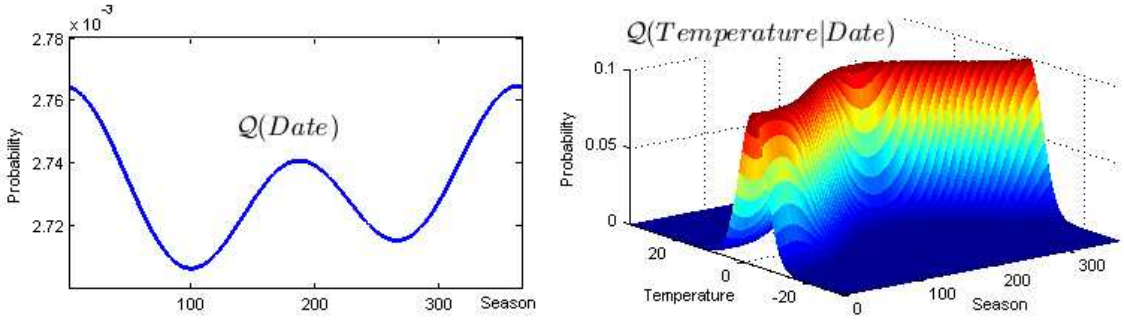

Figure 2: Plausible Markov kernels $\mathcal{Q}(Date)$ and $\mathcal{Q}(Temperature|Date)$ of causation $Date \rightarrow Temperature$

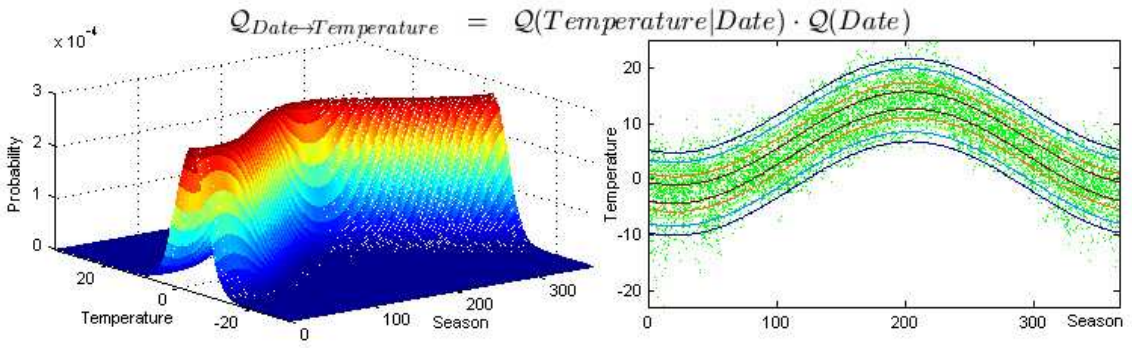

Figure 3: Hypothetical joint distribution $\mathcal{Q}_{Date \rightarrow Temperature}$ based on causation $Date \rightarrow Temperature$

With these constraints we computed the plausible Markov kernels for both hypothetical causal directions. Note that in all figures the variable *Date* is parameterized in k. Because of the non-

uniform sampling of *Date* (there are often only 365 days in year and in the real data set there is one year more observed for the days in January), the plausible Markov kernel of the cause *Date* in $Date \to Temperature$ differs slightly from the usually expected uniform distribution (Figure 2, left). For the effect variable *Temperature* in $Date \to Temperature$, the plausible Markov kernel (Figure 2, right) has a conditional expectation in a sinus form, which traces back to the cyclic property of the cause *Date*, and a Gaussian-shaped function for every given value of *Date*, which is due to the method of entropy maximization constrained by first and second moments.

In case of the other hypothetical causal direction $Temperature \to Date$, the cause variable *Temperature* has a Gaussian distribution (Figure 4, left), as a result of constrained entropy maximization. For the effect variable *Date* in $Temperature \to Date$, we obtain a bizarre shape for its most plausible Markov kernel (Figure 4, right).

Then we calculated the joint distributions from these plausible Markov kernels based on both hypothetical causal directions.

$$\begin{aligned} \mathcal{Q}_{Date \to Temperature} &= \mathcal{Q}(Temperature|Date) \cdot \mathcal{Q}(Date) \\ \mathcal{R}_{Temperature \to Date} &= \mathcal{R}(Date|Temperature) \cdot \mathcal{R}(Temperature). \end{aligned}$$

Figure 3 (left) visualizes the resulting joint distribution $\mathcal{Q}_{Date \to Temperature}$ and Figure 5 (left) visualizes $\mathcal{R}_{Temperature \to Date}$. Our computation is based on a discretization of one day for the variable *Date* and one degree for the variable *Temperature*. Figures 3 (right) and 5 (right) display both joint distributions as contours of equal with the same observed data points, respectivly. We note that $\mathcal{Q}$ and $\mathcal{R}$ have different numbers of modes, which we found to be invariant to changes of location and scale of discretization.

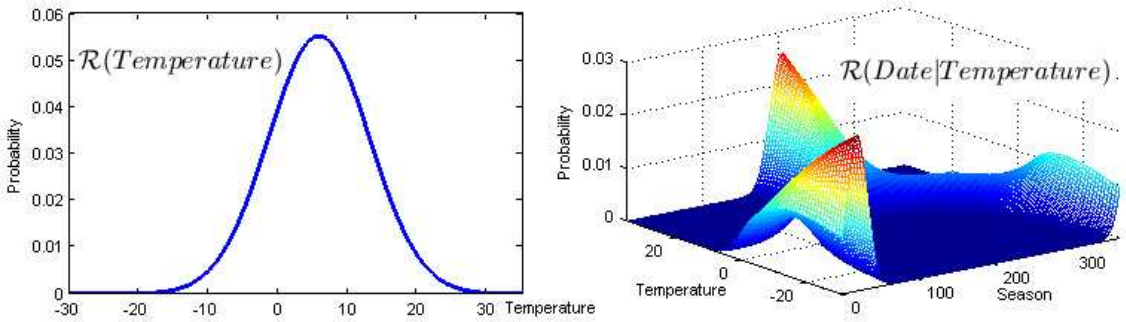

Figure 4: Plausible Markov kernels $\mathcal{R}(Temperature)$ and $\mathcal{R}(Date|Temperature)$ of hypothetical causation $Temperature \to Date$

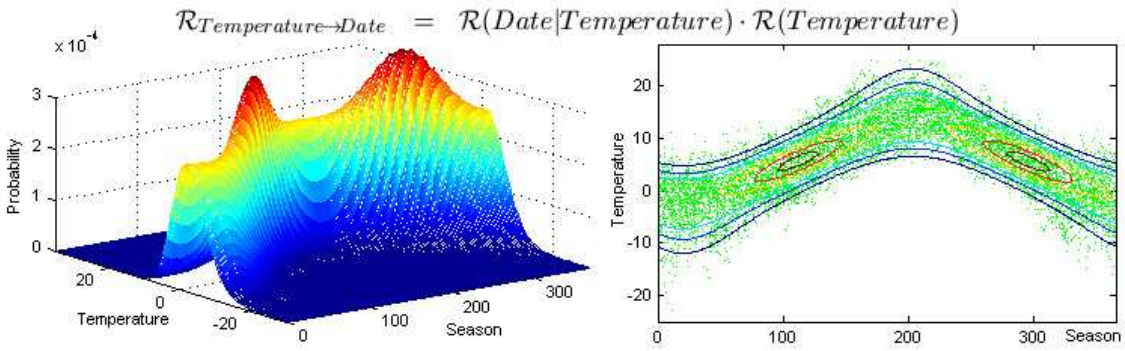

Figure 5: Hypothetical joint distribution $\mathcal{R}_{Temperature \to Date}$ based on hypothetical causation $Temperature \to Date$

We use the log-likelihood based on $\mathcal{Q}_{Date \to Temperature}$ and $\mathcal{R}_{Temperature \to Date}$ to quantify how strong the data support a hypothetical causal direction in the context of plausible Markov kernels. Our calculation shows that given data the "true" causal direction $Date \to Temperature$ achieves a log-likelihood score of -7.9844×10^4, whereas the other direction gets a lower log-likelihood score of -8.0027×10^4. Our method yields herewith the correct result.

6 Conclusion

Between only two observed variables X and Y (either discrete or continuous, either one- or multidimensional), both possible hypothetical causal directions $X \to Y$ and $Y \to X$ come into consideration. However, it is well known that they are equivalent under the Markov condition assumption and thus indistinguishable solely on the basis of probabilistic independence. In this paper, we developed a new method of causal inference using plausibility of Markov kernels. Through some additional "smoothness" conditions on the shape of plausible Markov kernels we found a possible way to capture the asymmetry of causality and showed a novel approach to estimate the causal direction between X and Y. The encouraging result[1] of real-world temperature data raise a slight hope for making causal inferences from purely observational (non-interventional) data among equivalent causal structures with respect to the causal Markov condition. Our further work is to generalize our method by exploring diverse criteria for most plausible Markov kernels concerning true causality in nature.

7 Acknowledgement

We thank B. Janzing of the Meteorological Station Furtwangen (Germany) for providing the temperature data. D. Janzing acknowledges financial support by the DFG project STE 1041/1.

A Appendix

Here we derive the plausible Markov kernels of the causation between a binary variable X with $\mathcal{S}^x = \{-1, +1\}$ and a real-valued variable Y with $\mathcal{S}^y = \mathbb{R}$. For the sake of simplicity, we denote $x_{\pm 1}$ for the cases $X = \pm 1$. Assuming a hypothetical causal direction $X \to Y$, the plausible Markov kernel $\mathcal{Q}(X)$ is determined just through the constraint of its first moment μ^x. Note that the second moment of X is a constant 1. It applies

$$\begin{aligned} \mathcal{Q}(x_{+1}) &= \frac{1}{2}(1 + \mu^x) =: q \\ \mathcal{Q}(x_{-1}) &= \frac{1}{2}(1 - \mu^x) = 1 - q. \end{aligned}$$

To determine the plausible Markov kernel $\mathcal{Q}(Y|X)$ we maximize the entropy function

$$\mathcal{H}(Y|X) = q \cdot \mathcal{H}(Y|x_{+1}) + (1 - q) \cdot \mathcal{H}(Y|x_{-1}) \tag{1}$$

subject to the constraints

$$q \cdot E_{+1} + (1 - q) \cdot E_{-1} = \mu^y \tag{2}$$

$$q \cdot E_{+1} - (1 - q) \cdot E_{-1} = \eta^{xy} \tag{3}$$

$$q \cdot (E_{+1})^2 + (1 - q) \cdot (E_{-1})^2 + q \cdot Var_{+1} + (1 - q) \cdot Var_{-1} = \gamma^y \tag{4}$$

[1] We have so far applied our method to two real-world problems. One of them is shown in the present paper, the other one was a medical one. In both cases, the results were positive. Unfortunately we cannot present the results of the second study in the present paper.

Here μ^y is the first moment of Y, η^{xy} the second mixed moment of X and Y, γ^y the second moment of Y. These values are known. $E_{\pm 1}$ denote the expectations of the conditional variable $(Y|x_{\pm 1})$ and $Var_{\pm 1}$ the variances of $(Y|x_{\pm 1})$, respectively. These values are still to be determined. However, $E_{\pm 1}$ can be determined by equations (2) and (3) uniquely.

$$\begin{aligned} E_{+1} &= \frac{\mu^y + \eta^{xy}}{1+\mu^x} \\ E_{-1} &= \frac{\mu^y - \eta^{xy}}{1-\mu^x}. \end{aligned}$$

Therefore, it remains actually only one constraint to be satisfied:

$$q \cdot Var_{+1} + (1-q) \cdot Var_{-1} =: \sigma^2 \tag{5}$$

where

$$\sigma^2 = \eta^y - \left(q \cdot (E_{+1})^2 + (1-q) \cdot (E_{-1})^2 \right) = \eta^y - \frac{(\mu^y + \eta^{xy})^2}{2(1+\mu^x)} - \frac{(\mu^y - \eta^{xy})^2}{2(1-\mu^x)}.$$

Here σ^2 can be calculated directly from all known values. The maximization of the function (1) with satisfying the constraint (5) has obviously a unique solution that $\mathcal{Q}(Y|x_{+1})$ and $\mathcal{Q}(Y|x_{-1})$ are both Gaussian distributed:

$$\mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}(E_{+1}, Var_{+1}) \quad \text{and} \quad \mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}(E_{-1}, Var_{-1}).$$

Otherwise it would be inconsistent with the well known fact that a normal distribution maximizes the entropy by given expectation and variance. The maximal entropy of $\mathcal{Q}(Y|X)$ of equation (1) in such case can be formulated as follows:

$$\mathcal{H}(Y|X) = \frac{1}{2} ln(2\pi e) + \frac{q}{2} ln(Var_{+1}) + \frac{1-q}{2} ln(Var_{-1}) \tag{6}$$

since the entropies of both Gaussian distributions are $\frac{1}{2} ln(2\pi e Var_{+1})$ and $\frac{1}{2} ln(2\pi e Var_{-1})$ respectively. Substitute (5) into (6), to achieve the maximum the first-order derivative must vanish and the second-order derivative should be negative. We obtain

$$Var_{+1} = Var_{-1} = \sigma^2$$

which means $\mathcal{H}(Y|X)$ achieves its maximum if and only if

$$\mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}\left(\mu_{-1}, \sigma^2\right) \quad \text{and} \quad \mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}\left(\mu_{+1}, \sigma^2\right).$$

The Markov kernels $\mathcal{R}(Y)$ and $\mathcal{R}(X|Y)$ for the other causal direction $X \to Y$ can also be determined analytically. Firstly, it is known that for fixed first (μ^y) and second moment (γ^y) bell-shaped Gaussian distribution $\mathcal{N}(\mu^y, \gamma^y - (\mu^y)^2)$ maximizes the differential entropy of the real-valued variable Y. To determine $\mathcal{R}(X|Y)$ we maximize the entropy function

$$\mathcal{H}(X|Y) = -\int \left(\mathcal{R}(x_{+1}|y) \cdot ln\left(\mathcal{R}(x_{+1}|y)\right) + \mathcal{R}(x_{-1}|y) \cdot ln(\mathcal{R}(x_{-1}|y)) \right) \cdot \mathcal{R}(y) dy$$

subject to the constraints

$$\mathcal{R}(x_{+1}|y) + \mathcal{R}(x_{-1}|y) = 1 \quad \forall y \in \mathbb{R} \tag{7}$$

$$\int \left(\mathcal{R}(x_{+1}|y) - \mathcal{R}(x_{-1}|y) \right) \cdot \mathcal{R}(y) dy = \mu^x \tag{8}$$

$$\int y \cdot \left(\mathcal{R}(x_{+1}|y) - \mathcal{R}(x_{-1}|y) \right) \cdot \mathcal{R}(y) dy = \eta^{xy} \tag{9}$$

$$\int \left(\mathcal{R}(x_{+1}|y) + \mathcal{R}(x_{-1}|y) \right) \cdot \mathcal{R}(y) dy = \gamma^x \equiv 1 \tag{10}$$

Here μ^x and γ^x is the known first and second moment of X. The equation (10) holds trivially. Through the substitution of (7) in (8) and (9) only the following two constraints are left:

$$\int (2\mathcal{R}(x_{+1}|y) - 1) \cdot \mathcal{R}(y)dy = \mu^x \tag{11}$$

$$\int y \cdot (2\mathcal{R}(x_{+1}|y) - 1) \cdot \mathcal{R}(y)dy = \eta^{xy} \tag{12}$$

By introducing two positive Lagrange multipliers λ and ν the solution of $\mathcal{R}(X|Y)$ must be of the form

$$\begin{aligned} \mathcal{R}(x_{-1}|y) &= \frac{e^{-(\lambda y+\nu)}}{e^{\lambda y+\nu} + e^{-(\lambda y+\nu)}} = \frac{1}{2} - \frac{1}{2}tanh(\lambda y + \nu) \\ \mathcal{R}(x_{+1}|y) &= \frac{e^{\lambda y+\nu}}{e^{\lambda y+\nu} + e^{-(\lambda y+\nu)}} = \frac{1}{2} + \frac{1}{2}tanh(\lambda y + \nu). \end{aligned}$$

Together with (11) and (12) the unknowns λ and μ should satisfy the following equations system

$$\begin{aligned} \int tanh(\lambda y + \nu) \cdot \mathcal{R}(y)dy &= \mu^x \\ \int y \cdot tanh(\lambda y + \nu) \cdot \mathcal{R}(y)dy &= \eta^{xy} \end{aligned}$$

where $\mathcal{R}(y) \propto \mathcal{N}(\mu^y, \gamma^y - (\mu^y)^2)$. Solving this nonlinear equations system, we will be able to determine λ and μ, thus $\mathcal{R}(X|Y)$ for every given μ^x and η^{xy}.

In summary, we obtain a closed-form solution for the causation between a binary and a real-valued variable. For one causal direction $X \to Y$, we have plausible Markov kernels in a form of

$$\begin{aligned} \mathcal{Q}(x_{-1}) = \frac{1}{2}(1 - \mu^x) \qquad &\text{and} \qquad \mathcal{Q}(x_{+1}) = \frac{1}{2}(1 + \mu^x) \\ \mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}\left(\mu_{-1}, \sigma^2\right) \qquad &\text{and} \qquad \mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}\left(\mu_{+1}, \sigma^2\right) \end{aligned}$$

where

$$\mu_{-1} = \frac{\mu^y - \eta^{xy}}{1 - \mu^x}, \quad \mu_{+1} = \frac{\mu^y + \eta^{xy}}{1 + \mu^x} \quad \text{and} \quad \sigma^2 = \gamma^y - \frac{(\mu^y + \eta^{xy})^2}{2(1 + \mu^x)} - \frac{(\mu^y - \eta^{xy})^2}{2(1 - \mu^x)}.$$

For the other causal direction $Y \to X$, the plausible Markov kernels have a form of

$$\begin{aligned} \mathcal{R}(Y) \quad &\propto \quad \mathcal{N}\left(\mu^y, \gamma^y - (\mu^y)^2\right) \\ \mathcal{R}(x_{-1}|y) = \frac{1}{2} - \frac{1}{2}tanh(\lambda y + \nu) \qquad &\text{and} \qquad \mathcal{R}(x_{+1}|y) = \frac{1}{2} + \frac{1}{2}tanh(\lambda y + \nu). \end{aligned}$$

Having computed these plausible Markov kernels, the corresponding joint distributions

$$\mathcal{Q}_{X\to Y} = \mathcal{Q}(Y|X) \cdot \mathcal{Q}(X) \quad \text{(with respect to causation } X \to Y\text{)}$$

$$\mathcal{R}_{Y\to X} = \mathcal{R}(X|Y) \cdot \mathcal{R}(Y) \quad \text{(with respect to causation } Y \to X\text{)}$$

can be calculated. The question is whether $\mathcal{Q}_{X\to Y}$ could equal $\mathcal{R}_{Y\to X}$ under certain conditions, because if the equation

$$\mathcal{Q}_{X\to Y} = \mathcal{R}_{Y\to X}$$

applies, causal directions ($X \to Y$ and $Y \to X$) cannot be distinguished from each other anymore, based on our "principle of plausible Markov kernels". However, one checks that whenever there exists correlation between X and Y, our method with most plausible Markov kernels leads always to different joint distributions. This is because the marginal distribution of Y based on the causal direction $X \to Y$ is a convex sum of two Gaussian distributions which have different expectation values for non-vanishing correlation between X and Y. This distribution cannot coincide with the marginal distribution of Y based the causal direction $Y \to X$ since the latter is unimodal Gaussian distributed.

References

[1] S. Boyd. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[2] R. Collins and A. Wragg. Maximum entropy histograms. *Jounal of Physics A: Mathematical and General*, 10(9):1441–1464, 1977.

[3] D.C. Dowson and A. Wragg. Maximum-entropy distributions having prescribed first and second moments. *IEEE Transactions on Information Theory*, 19:5:689–693, 1973.

[4] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proc. International Symposium on Science of modeling -The 30th Anniversary of the Information Criterion (AIC)-*, pages 261–270, Tokyo, Japan, 2003.

[5] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, Oxford, 1996.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[7] H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

[8] W. Salmon. Probabilistic causality. *Pacific Philosophical Quarterly*, 61:50–74, 1980.

[9] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search (Lecture Notes in Statistics)*. Springer-Verlag, New York, 1993.