Causal reasoning by evaluating the complexity of conditional densities with kernel methods

Xiaohai Sun*, Dominik Janzing, Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

Available online 26 January 2008

Abstract

We propose a method to quantify the complexity of conditional probability measures by a Hilbert space seminorm of the logarithm of its density. The concept of reproducing kernel Hilbert spaces (RKHSs) is a flexible tool to define such a seminorm by choosing an appropriate kernel. We present several examples with artificial data sets where our kernel-based complexity measure is consistent with our intuitive understanding of complexity of densities.

The intention behind the complexity measure is to provide a new approach to inferring causal directions. The idea is that the factorization of the joint probability measure $P(\text{effect}, \text{cause})$ into $P(\text{effect}|\text{cause})P(\text{cause})$ leads typically to "simpler" and "smoother" terms than the factorization into $P(\text{cause}|\text{effect})P(\text{effect})$. Since the conventional constraint-based approach of causal discovery is not able to determine the causal direction between only two variables, our inference principle can in particular be useful when combined with other existing methods.

We provide several simple examples with real-world data where the true causal directions indeed lead to simpler (conditional) densities.
© 2008 Elsevier B.V. All rights reserved.

Keywords: Learning causality; Complexity of densities; Reproducing kernel Hilbert space; Principle of plausible Markov kernels

1. Introduction

When visualizing the densities of some well-known distributions like Gaussian or gamma distributions, for instance, one would agree that the shape of the densities is rather smooth. In contrast, a mixture of two Gaussian functions, in particular when it is clearly bimodal, may be considered *less smooth*. The same holds for the mixture of two gamma distributions. Having observed a bimodal density after large sampling, one would prefer to interpret the observation as a mixture of two populations. It is implausible to assume that a density with such a shape stems from a *homogeneous statistical ensemble*. This shows that common sense gives us an intuitive idea about which densities should be considered natural and which ones demand an additional explanation as being a mixture of "more natural" and "smoother" distributions. Actually,

there is a broad variety of applications where the detection of mixtures is crucial for data analysis (see e.g. [4,5,7]). The definition of complexity of densities proposed in the current paper should, however, not merely be considered as a method for detecting mixtures of distributions. The main motivation behind it is to develop a tool for a new causal inference principle based on empirical data by quantifying smoothness of conditional densities.

Let us first sketch the basic idea of our causal inference rule. Given a joint probability measure P on n random variables $X_1, \ldots, X_n$, all the conditional measures (the so-called "Markov kernels") that appear in the factorization of the joint measure

$$P(x_1, \ldots, x_n) = P(x_1) P(x_2|x_1) \cdots P(x_n|x_1, \ldots, x_{n-1})$$

will typically be "smoother" or "simpler", if the order of the factorization $X_1, \ldots, X_n$ coincides with the causal order, in the sense that there is no pair (X_i, X_j) with $i < j$ such that X_j is a cause of X_i. We call this the principle of plausible Markov kernels. Throughout the paper, we assume that all joint probability measures are represented by densities.

*Corresponding author.

E-mail addresses: xiaohai.sun@tuebingen.mpg.de (X. Sun),
dominik.janzing@tuebingen.mpg.de (D. Janzing),
bernhard.schoelkopf@tuebingen.mpg.de (B. Schölkopf).

This inference principle can be very useful, especially where the conventional constraint-based causal learning algorithms (e.g. [10,16,11]) fail. For example, when only two dependent variables are measured, inferring the causal direction between them is impossible with the conventional approaches that are based on independence constraints. Our inference rule can provide some hints even in such cases about which causal direction should be preferred. We do not intend to treat the problem of confounders in the current paper and assume that there are no hidden common causes in our setting.

A first attempt to formalize the principle of plausible Markov kernels is to consider conditional probability measures most plausible that maximize the conditional entropy subject to *simple* constraints like the observed first and second statistical moments. This has been proposed in a conference paper [17]. For variables with $\mathbb{R}$ as range we obtain linear interactions, while for variables with other ranges we obtain conditional probabilities which are also smooth in an intuitive sense. Our approach is supported by positive results of experiments with real-world data on continuous and discrete range.

A related causal inference rule was described by Kano and Shimizu in [6]. They have observed that linear causal relations between non-Gaussian distributed random variables induce joint probability measures which would require non-linear cause–effect relations for causal hypotheses that differ from the true causal order. Accordingly, their inference principle [14,13] which is based on independent component analysis (ICA) selects causal hypotheses for which linear cause–effect relations are sufficient whenever such hypotheses are possible for a given probability measure. Unfortunately, the underlying idea is only justified for causal structures of *real-valued* variables since linear effects do not exist in the general case.

The main concern of this paper is to quantify complexity of (conditional) densities in order to detect the asymmetry between cause and effect. For this purpose, we propose to measure the complexity by a Hilbert space seminorm of the logarithm of the density. This function is an element of a reproducing kernel Hilbert space (RKHS) and its seminorm can therefore be computed by kernel methods. In contrast to common machine learning applications, the complexity measure in this paper plays not only the role of a regularizer, which is often used to avoid an overfitting of describing finite data points. Rather it should be considered as an interesting quantity in its own right since it provides hints on the causal direction. For this reason, it is essential to choose a definition of complexity which is well-behaved in some respects. This is described in the following section.

2. Defining complexity measure by Hilbert space seminorms

Before we introduce the complexity measure for *conditional* densities we define it for *unconditional* densities. Let us ignore for the moment the sampling issue and assume that the density P_X of some random variable X (possibly vectorial) is perfectly known. For the sake of convenience and in order to avoid some technical problems, we shall assume that the value set $\mathscr{X}$ of X is finite. We introduce a complexity measure on the space of densities on $\mathscr{X}$ as follows:

Definition 1 (*Complexity of marginals*). Let $\mathscr{X}$ be a probability space, X be a random variable on $\mathscr{X}$, and P_X a density on $\mathscr{X}$. Furthermore, let $\mathscr{H}$ be a Hilbert space of real-valued functions on $\mathscr{X}$ containing the set of constant functions. Then we define the complexity of P_X as

$$C(P_X):=\min\{\|\phi\|^2|\phi \in \mathscr{H} \text{ with } P_X(x) = \exp(\phi(x) - \ln z_\phi)\}$$

with the partition function

$$z_\phi := \sum_x \exp(\phi(x)).$$

Here $\|.\|$ denotes a seminorm on $\mathscr{H}$ given by

$$\|\phi\|:=\sqrt{B(\phi,\phi)},$$

where B denotes a positive definite (but not necessarily strictly positive) bilinear form $B : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$.

In the following we will use the following terminology: we call two vectors $v, w \in \mathscr{H}$ orthogonal if $B(v,w) = 0$. For a subspace V we define

$$V^\perp := \{w | B(w,v) = 0 \; \forall v \in V\}.$$

Since V and $V^\perp$ may have non-trivial intersection, we avoid the term "orthogonal complement". The term "orthogonal" will always refer to the bilinear form B unless something else is explicitly stated. An orthogonal projection R is said to be an projection onto $V^\perp$ if $R\mathscr{H} \subseteq V^\perp$ and $Rw = w - v$ for some $v \in V$ that minimizes $\|w - v\|$.

We have

$$C(P) = \|Q(\ln P)\|^2, \tag{1}$$

where Q denotes the projection onto $\mathbf{1}^\perp$. This is due to $\|\phi\| = \|Q(\phi - z_\phi \mathbf{1})\| = \|Q(\ln P)\|$.

We show the following lemma.

Lemma 1 (*Additivity*). Let $\mathscr{H}_1$ and $\mathscr{H}_2$ be spaces of functions on $\mathscr{X}_1$ and $\mathscr{X}_2$, respectively. Furthermore, let C_1 and C_2 be complexity measures on the densities on $\mathscr{X}_1$ and $\mathscr{X}_2$, respectively, defined by the corresponding seminorms in $\mathscr{H}_1$ and $\mathscr{H}_2$. Assume that a complexity measure C on the density on $\mathscr{X}$ is based on the seminorm of $\mathscr{H}:=\mathscr{H}_1 \otimes \mathscr{H}_2$ that satisfies the embedding property $\|a \otimes \mathbf{1}\| = \|a\| = \|\mathbf{1} \otimes a\|$, where $\mathbf{1}$ denotes the function taking the constant value 1. Then we have the following additivity rule: Let P be defined by a product of densities P_1 and P_2, i.e., $P(x_1, x_2) = P_1(x_1)P_2(x_2)$ for all x_1 and x_2. Then the complexity of the product measure satisfies $C(P) = C_1(P_1) + C_2(P_2)$.

Proof. Let Q, Q_1, Q_2 denote the projections onto the space of functions orthogonal to $\mathbf{1}$ for the spaces $\mathscr{H}, \mathscr{H}_1, \mathscr{H}_2$,

respectively. Then we have

$$\|Q(\ln P_1 \otimes \mathbf{1} + \mathbf{1} \otimes \ln P_2)\|^2$$
$$= \|Q_1(\ln P_1) \otimes \mathbf{1} + \mathbf{1} \otimes Q_2(\ln P_2)\|^2$$
$$= \|Q_1(\ln P_1)\|^2 + \|Q_2(\ln P_2)\|^2,$$

where the last equality is due to Pythagoras' theorem after taking into account that the vectors $Q_1(\ln P_1) \otimes \mathbf{1}$ and $\mathbf{1} \otimes Q_2(\ln P_2)$ are mutually orthogonal. $\quad\square$

Now we move to the definition of the complexity of *conditional* probabilities:

Definition 2 (*Complexity of conditionals*). Let $\mathscr{X}$ and $\mathscr{Y}$ be the respective value sets of random variables X and Y, and $P_{X,Y}$ be a joint density on $\mathscr{X} \times \mathscr{Y}$. Let $P_{Y|X}$ be the corresponding conditional density. We define the complexity of $P_{Y|X}$ as

$$C(P_{Y|X}) := \min\{\|\phi\|^2 | \phi \in \mathscr{H} \text{ with } P_{Y|X}(y|x)$$
$$= \exp(\phi(x,y) - \ln z_\phi(x))\}$$

with the partition function

$$z_\phi(x) := \sum_y \exp(\phi(x,y)).$$

Similarly to the reformulation of Definition 1 in Eq. (1), the definition of the complexity of a conditional density can also be given in a more explicit form:

$$C(P_{Y|X}) = \|(\mathbf{id} \otimes Q_2)(\ln P_{Y|X})\|^2, \tag{2}$$

where "**id**" denotes the identity map and Q_2 is as in the proof of Lemma 1. Under the assumptions of Definition 2, we have:

Lemma 2 (*Consistency*). *Let X and Y be stochastically independent with respect to the joint density P, i.e., $P_{Y|X} = P_Y$. Let C be a complexity measure based on a seminorm in $\mathscr{H} = \mathscr{H}_X \otimes \mathscr{H}_Y$ satisfying the embedding property in Lemma 1. Then we have $C(P_{Y|X}) = C_2(P_Y)$.*

Proof. Let ϕ be some function on $\mathscr{X} \times \mathscr{Y}$ such that $P_{Y|X}(y|x) = \exp(\phi(x,y) - \ln z_\phi(x)) = P_Y(y)$. We choose an arbitrary value y_0 and set $f(x) := \phi(x,y_0) - \ln P_Y(y_0)$ and $g(y) := \ln P_Y(y)$. Then we have $\phi(x,y) = f(x) + g(y)$. Thus

$$\|(\mathbf{id} \otimes Q_2)(\phi)\|^2 = \|(\mathbf{id} \otimes Q_2)(f \otimes \mathbf{1} + \mathbf{1} \otimes g)\|^2 = \|Q_2(g)\|^2.$$

Therefore, we conclude $C(P_{Y|X}) = C_2(P_Y)$. $\quad\square$

Lemma 2 is essential, if one intends to compare the complexity of marginal densities to that of conditional densities. The following causal inference principle stands behind such a comparison: having factorized a joint density $P_{X,Y}$ into $P_{Y|X}P_X$ and $P_{X|Y}P_Y$ based on both possible hypothetical causal orders, one calculates the sums of the complexities $C(P_{Y|X}) + C(P_X)$ and $C(P_{X|Y}) + C(P_Y)$ with respect to the different hypotheses. The intention is to consider the sums as the "total complexity" of the causal models $X \to Y$ and $X \leftarrow Y$, respectively, and to prefer the

causal direction that corresponds to the smaller total complexity. For doing so, it is crucial to make $C(P_Y)$ and $C(P_{Y|X})$ comparable. An essential property of the complexity measure is that we have

$$C(P_{Y|X}) + C(P_X) \neq C(P_{X|Y}) + C(P_Y)$$

in the generic case. The following lemma provides some deeper understanding why this is the case.

Lemma 3 (*Relation to complexity of partition function*). *Under the assumptions of Definition 2, the following inequalities hold*:

$$C(P_{X,Y}) \geqslant C(P_{Y|X}) + C(P_X) + C(R) - 2\sqrt{C(P_X)C(R)},$$

$$C(P_{X,Y}) \leqslant C(P_{Y|X}) + C(P_X) + C(R) + 2\sqrt{C(P_X)C(R)},$$

where R is the following measure on X: Set $R(x) := c \cdot z_f(x)$ with an appropriate normalization factor c and the partition function

$$z_f(x) = \sum_y \exp(f(x,y))$$

which is derived from $f := (\mathbf{id} \otimes Q_2)(\ln P_{Y|X})$.

Proof. Write

$$P(y|x) = \exp(f(x,y) - \ln z_f(x)),$$

where f satisfies by definition $(\mathbf{id} \otimes Q_2)(f) = f$. Furthermore, we set

$$P(x) = \exp(g(x) - \ln z)$$

with $Q_1(g) = g$ and normalization constant z. We observe that f is orthogonal to all functions that depend only on x since the latter have the form $h \otimes \mathbf{1}$ (where h is an arbitrary function). We have

$$\ln P_{X,Y} = \ln P_X + \ln P_{Y|X} = (-\ln z_f + g) \otimes \mathbf{1} + f - \ln z.$$

Due to the above remarks we have $f \perp (-\ln z_f + g) \otimes \mathbf{1}$. To compute the complexity of $P_{X,Y}$, we observe

$$C(P_{X,Y}) = \|Q(f + (-\ln z_f + g) \otimes \mathbf{1} + \ln z\,(\mathbf{1} \otimes \mathbf{1}))\|^2$$
$$= \|f + Q_1(-\ln z_f + g) \otimes \mathbf{1}\|^2.$$

Since the projected term is still orthogonal to f (note that it is a function that depends only on x) we have

$$C(P_{X,Y}) = \|f\|^2 + \|Q_1(-\ln z_f + g)\|^2$$
$$= \|f\|^2 + \|Q_1(\ln z_f) + g\|^2. \tag{3}$$

By elementary geometry we obtain

$$\|Q_1(-\ln z_f) + g\|^2 \geqslant \|Q_1(\ln z_f)\|^2 + \|g\|^2 - 2\|Q_1(\ln z_f)\|\,\|g\|,$$

$$\|Q_1(-\ln z_f) + g\|^2 \leqslant \|Q_1(\ln z_f)\|^2 + \|g\|^2 + 2\|Q_1(\ln z_f)\|\,\|g\|.$$

Having $C(R) = \|Q_1(\ln z_f)\|^2$, we finally conclude

$$C(P_{X,Y}) \geqslant C(P_{Y|X}) + C(P_X) + C(R) - 2\sqrt{C(P_X)C(R)},$$

$$C(P_{X,Y}) \leqslant C(P_{Y|X}) + C(P_X) + C(R) + 2\sqrt{C(P_X)C(R)}. \quad\square$$

Note that in high dimensional spaces the angle between two vectors is typically close to 90°. Therefore, it is likely that the vectors $Q_1(\ln z_f)$ and g in Eq. (3) satisfy $B(\ln z_f, g) \approx 0$. We then have

$$C(P_{X,Y}) \approx C(P_{Y|X}) + C(P_X) + C(R).$$

In other words, the complexity of the joint density is typically the sum of the complexities of the conditional densities and the complexity of a measure defined by the partition function. The basic idea behind our inference rule is that simple causal mechanism may generate conditional densities $P_{Y|X}$ which are simple *up to a rather complex X-dependent normalization constant*, i.e., the partition function. Note that the joint density could be complex even when P_X is simple due to the additional complexity of the partition function.

3. Calculation of seminorm using kernel methods

We have shifted the problem of defining the complexity of density into the definition of seminorms. We will rewrite our definition such that seminorms can be calculated in an implicit way. With the so-called "kernel trick" different seminorms can be chosen by simply replacing the kernel (see [12,2]).

Let $k_1, k_2 : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be positive definite symmetric kernels and $\mathcal{X} \times \mathcal{Y}$ the probability space under consideration. Let $\mathcal{H}_j$ for $j = 1, 2$ be the Hilbert spaces given by the completion of the spans of the functions $k_j((x, y), .)$ with the inner product

$$\langle k_j((x, y), .), k_j((x', y'), .) \rangle = k_j((x, y), (x', y')). \quad (4)$$

Hilbert spaces defined this way are usually referred to as RKHS. We assume that $\mathcal{H}_2$ is a subspace of $\mathcal{H}_1$. The vector ϕ in Definitions 1 and 2 can be approximated by

$$\begin{aligned}
\phi(x, y) &:= \sum_{j=1}^{n} c_j k((x_j, y_j), (x, y)) \\
&= \left\langle \sum_{j=1}^{n} c_j k((x_j, y_j), .), k((x, y), .) \right\rangle \quad (5)
\end{aligned}$$

with appropriate coefficients c_j and points (x_j, y_j).

We define our seminorm by

$$\|\phi\| := \|R(\phi)\|_{\mathcal{H}_1},$$

where R is the projector onto the subspace orthogonal to $\mathcal{H}_2$ with respect to the inner product in $\mathcal{H}_1$. The idea of using such a seminorm is that the space $\mathcal{H}_2$ contains simple functions (for instance polynomials of low degree) that should not contribute to the complexity measure at all.[1]

Let $P_{Y|X}$ be a conditional density, given by

$$\begin{aligned}
P_{Y|X}(y|x) = \exp\Bigg(&\sum_{j=1}^{n} c_j^{(1)} k_1((x_j, y_j), (x, y)) \\
&+ \sum_{j=1}^{n} c_j^{(2)} k_2((x_j, y_j), (x, y)) - \ln z_{\mathbf{c}}(x) \Bigg) \quad (6)
\end{aligned}$$

with the appropriate partition function $z_{\mathbf{c}}(x)$. The complexity $C(P_{Y|X})$ is then defined by the minimum of

$$\sum_{j,j'=1}^{n} c_j^{(1)} c_{j'}^{(1)} k_1((x_j, y_j), (x_{j'}, y_{j'})),$$

i.e., the square of the norm of the shortest component in $\mathcal{H}_1$, see Eq. (4), over all vectors $\mathbf{c} := (c_1^{(1)}, \ldots, c_n^{(1)}, c_1^{(2)}, \ldots, c_n^{(2)}) \in \mathbb{R}^{2n}$ for which Eq. (6) holds. The vector with coefficients $k_1((x_j, y_j), (x, y))$, $j = 1, \ldots, n$ and $k_2((x_j, y_j), (x, y))$, $j = 1, \ldots, n$ can be interpreted as the vector of sufficient statistics of an exponential model.

The framework introduced can also be considered as a method of density estimation with kernel methods. To make this method tractable in practice, there are some issues of implementation to be addressed. The choice of kernels k_1 and k_2 will be discussed in the next section. Given k_1 and k_2 described in the next section, our remarks above specified the choice of points (x_j, y_j) for $j = 1, \ldots, n$ in the range. Our experiments show that the seminorm is not sensitive against the choice of n, if n is not too small and the points (x_j, y_j) are somewhat evenly distributed over the whole range. The results of all our experiments in this paper are based on the choice of $n = 7$ for unconditional (one-dimensional) cases and $n = 49$ for conditional (two-dimensional) cases. The seven points for each dimension are chosen equidistantly in percentile over the whole observed range. For a binary range, $n = 2$.

To ensure that the embedding property $\|a \otimes \mathbf{1}\| = \|a\| = \|\mathbf{1} \otimes a\|$ is satisfied we proceed as follows. We choose the kernel k_1 as the product

$$k_1((x_j, y_j), (x_{j'}, y_{j'})) = k_X^{(1)}(x_j, x_{j'}) k_Y^{(2)}(y_j, y_{j'}).$$

Thus, the corresponding RKHSs have the form $\mathcal{H}_2 := \mathcal{H}_2^X \otimes \mathcal{H}_2^Y$ and $\mathcal{H}_1 := \mathcal{H}_1^X \otimes \mathcal{H}_1^Y$. We choose the kernels $k_X^{(2)}$ and $k_Y^{(2)}$ and the domains $\mathcal{X}, \mathcal{Y}$ such that $\mathcal{H}_2^X$ and $\mathcal{H}_2^Y$ contain the constant functions and normalize $k_X^{(1)}$ and $k_Y^{(1)}$ such that the constant functions $\mathbf{1}$ on $\mathcal{X}$ and $\mathcal{Y}$ satisfy $\|\mathbf{1}\|_{\mathcal{H}_1^X} = 1$ and $\|\mathbf{1}\|_{\mathcal{H}_1^Y} = 1$, respectively.

To this end, we define the matrix $K_X := k_X^{(1)}(x_j, x_{j'})$ and calculate its inverse K_X^{-1}. Let $c := (K_X^{-1})\mathbf{1}$ be the vector of coefficients of the constant function $\mathbf{1}$. This yields the normalization condition $\langle c | K_X c \rangle = 1$, i.e., the sum of all entries of K_X^{-1} are 1. The same procedure is also applied to $k_Y^{(1)}$. The seminorm of $a \otimes \mathbf{1}$ is given by the Hilbert space norm of its component in $(\mathcal{H}_2^X \otimes \mathcal{H}_2^Y)^\perp$. Let R_X and R_Y be the orthogonal projections onto $(\mathcal{H}_2^X)^\perp$ and $(\mathcal{H}_2^Y)^\perp$, respectively. Due to $R_Y(\mathbf{1}) = 0$ the relevant component of $a \otimes \mathbf{1}$ is given by $R_X(a) \otimes \mathbf{1}$. The Hilbert space norm of this

---

[1] This corresponds to the use of conditionally positive definite kernels in semiparametric models [15,19].

function is given by $\|R_X(a)\|_{\mathscr{H}_1^X}$ which coincides with the seminorm of a. Similar arguments apply to $\mathbf{1} \otimes a$.

4. Estimating densities from finite data with kernels

To calculate the complexity of a density we first use regularized maximum likelihood estimation to fit the observed data points using exponential models. A general framework for applying the kernel approach to exponential families can be found in [3]. Without regularizer, the method works as follows. Introducing the map $\psi : \mathscr{X} \times \mathscr{Y} \to \mathscr{H}_1$ with

$$\psi(x, y) := k_1((.,.),(x, y))$$

we define the family of conditional densities $P_\phi(y|x) = \exp(\langle \phi | \psi(x, y) \rangle - \ln z_\phi(x))$. For N observed data points (x_i, y_i), the maximum likelihood estimation selects ϕ by

$$\max_{\phi \in \mathscr{H}_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} (\langle \phi | \psi(x_i, y_i) \rangle - \ln z_\phi(x_i)) \right\}. \quad (7)$$

In order to avoid overfitting we include a regularizer

$$\max_{\phi \in \mathscr{H}_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} (\langle \phi | \psi(x_i, y_i) \rangle - \ln z_\phi(x_i)) - \varepsilon \|\phi\| \right\}. \quad (8)$$

The regularizer, the norm itself and not its square (as opposed to our complexity measure) is in agreement with the choice in [1]. The authors of [1] propose to use a value of ε that is proportional to $1/\sqrt{N}$. In our experiments, we chose $\varepsilon = 1/\sqrt{N}$. Note, as an aside, that the regularized maximum likelihood estimation for unconditional densities can also be interpreted as maximizing the entropy of the density subject to the expectations of $\psi(X, Y)$ coinciding with the observed means of $\psi(X, Y)$ up to an error of ε (see [1]).

For the sake of numerical stability, we normalize the observed data for X, Y, respectively. The data are linearly transformed such that the points ± 1 of the normalized data have the same percentiles as ± 3 of a standard normal distribution, respectively. Thus the normalized data points with continuous range will be located mostly in the interval $[-1, 1]$. A normalized binary variable then takes values ± 1. We choose a discretization of 0.1 to count the relative frequencies and calculate the sum in optimization. For the experiments described in the next section we use a sum of the Gaussian kernel

$$k_\sigma((x, y), (x', y')) = \exp\left(-\frac{\|(x, y) - (x', y')\|^2}{2\sigma^2} \right)$$

to define the space $\mathscr{H}_1$ and a polynomial kernel

$$k_{a,b,\tilde{a},\tilde{b}}((x, y), (x', y')) = \left(\frac{\langle x \cdot x' \rangle}{a} + b \right) \left(\frac{\langle y \cdot y' \rangle}{\tilde{a}} + \tilde{b} \right)^2$$

to define $\mathscr{H}_2$. The additional scaling parameters $a, b, \tilde{a}, \tilde{b}$ are used to ensure a numerically stable training. We choose $a, b, \tilde{a}, \tilde{b}$ so that the entries of $k_{a,b,\tilde{a},\tilde{b}}$ take the value between $[-1, 1]$. Since the normalized data have the value mostly

between -1 and 1, we choose $a = \tilde{a} = 2$ and $b = \tilde{b} = \frac{1}{2}$, if x, y are one-dimensional. The formulation of both kernels for the unconditional case is straightforward. Assuming that the range of random variables is compact, the space $\mathscr{H}_2$ (induced by a Gaussian kernel) contains the space $\mathscr{H}_1$ (induced by a polynomial kernel).

The idea behind the choice of kernels is the following: if x and y are one-dimensional, the second kernel induces a space of functions spanned by the monomials $1, x, xy, xy^2, y, y^2$. We consider these as sufficiently smooth such that they should not contribute to the complexity measure. In particular, we can then obtain Gaussian distributions whose expectations and variance changes linearly with the given variable X. The Gaussian kernel and the polynomial kernel induces, on the one hand, enough flexibility to fit various global and local structure of density. On the other hand, the density estimated this way is smooth. For a discussion of smoothing properties of Gaussian and polynomial kernels we refer to [8,15].

Our experience suggests that we have to learn appropriate values σ for the Gaussian kernel by optimizing Eq. (8), otherwise we could not obtain reasonable fits. Clearly, we cannot directly compare the complexity values corresponding to kernels with different values for σ. However, we may define the complexity by the minimum over all seminorms squared within some given family of RKHSs. Denoting by $\mathscr{H}_i$ the Hilbert space given by the kernel k_i we may define $C(P)$ by $C(P) := \inf_{i \in I}\{C_i(P)\}$, where C_i refers to the complexity measure defined by the seminorm in $\mathscr{H}_i$. In order to ensure additivity with respect to product measures in product spaces for the redefined C we need to define a family of spaces by $\mathscr{H}_i^{(1)} \otimes \mathscr{H}_j^{(2)}$ and optimize over all pairs (i, j). Due to a combinatorial explosion such an optimization will only be feasible for a small set I and few tensor components. In the experiments described in the next section we have therefore used the same σ for the Hilbert spaces for X and Y.

If we run the optimization procedure in Eq. (8) over all Hilbert spaces (i.e., all reasonable values σ) the procedure will choose the vector ϕ from the Hilbert space that leads to the smallest norm among all those that yield the same value in the non-regularized optimization given by Eq. (7). We shall therefore consider the optimum of Eq. (8) over all kernels taken from a given family as an estimation of the minimal norm of the density over all Hilbert spaces under consideration. Since the optimization problem with σ is no longer convex, one should choose the start value of σ properly. In our experiments we chose 200 equidistant starting values in the range $(0, \frac{2}{3})$. The value which leads to the maximum of Eq. (8) will then be taken as the start value of a subsequent optimization via gradient descent.

5. Experiments with simulated and real-world data

Some simulated experiments show the intuitive meaning of our complexity measure, while the real-world examples show that this complexity measure could be helpful for

inferring the causal direction between two variables, because we assume that stochastic dependences between cause and effect which are generated by a natural causal mechanism should typically lead to comparably *simple expressions* for P(cause) and P(effect|cause) but not necessarily generate simple expressions for P(effect) and P(cause|effect).

5.1. Unconditional densities

We have sampled 1000 data points from various distributions, as shown in Fig. 1. The underlying density P_1 follows a standard normal distribution; P_2, P_3, P_4, P_5 are various mixtures of 2 Gaussians; P_6, P_7, P_8 are mixtures of 3, 4, 5 Gaussians, respectively. P_9 is a mixture of a Gaussian and a gamma distribution. P_{10} follows a single gamma distribution and P_{11}, P_{12} are mixtures of 2, 3 gamma distributions, respectively. As expected, we see that the complexity of a single Gaussian is 0. A single gamma distribution has a very small complexity value. The measure increases as the number of components increases. This holds even for the unimodal mixture P_2, P_{11}, P_{12}. Moreover, we have examined the smoothness (complexity) of a real-world temperature data set[2] with 9162 entries. The estimated density (see Fig. 2) has a complexity of 0.0265, which suggests that the density of temperatures is more complex than a single normal or gamma distribution. We observe slightly larger complexity values for a gamma distribution than for a Gaussian. We do not want to speculate whether this property is desirable.

5.2. Conditional densities

If we define a density on a binary variable X and a continuous variable Y by

$$P(y) = 0.5P(y|X = x_1) + 0.5P(y | X = x_2),$$

where both conditionals $P(y|X = x_i)$ are Gaussian, the total complexity of the model $X \rightarrow Y$ is zero since the kernel k_2 induces such a density. Note that due to our choice of kernel the complexity of the density of a binary variable is always 0. We checked on randomly generated data with 1000 points whether this result is also obtained in finite sampling. We furthermore confirmed that the model $X \rightarrow Y$ was also preferred when the conditional $P(Y|X = x_2)$ was the gamma distribution and $P(Y|X = x_1)$ was a Gaussian. In a similar way, we defined joint densities on X and Y corresponding to the mixture models P_2, P_3, P_4, P_9 in Fig. 1 by using a binary variable X to indicate which one of the two pure ensembles is taken. The complexity values in Table 1 show that we indeed obtained the expected results.

Since the causal inference problem was the motivation for the construction of our complexity measure, its

performance with respect to some real-world data is the best criterion for judging whether it seems appropriate or not. To this end, we performed experiments with data sets from the Current Population Survey (CPS) 2001 on the relation between sex (binary variable) and income (continuous variable) in the US.[3] Statistical methods show that income and gender are indeed correlated. Common sense tells us that we can exclude that the personal income influences the gender, whereas the reverse causal direction makes sense. We found that the density of the income marginalized over both genders is more complex than the density for both genders separately.

First we intended to check to what extent the complexity measure recognizes mixtures as more complex. We found

$$C(P_{\text{Income|Sex="male"}}) < C(P_{\text{Income}}),$$

and the same for $P_{\text{Income|Sex="female"}}$. Note that left side of the inequality can also be considered as the complexity of an *unconditional* density since we assigned a specific value to the conditioning variable.

However, to check the performance of our causal inference principle we have to compute the total complexity of both hypothetical causal directions. Using one subsample of 10% of the data points from 13,803 entries, we found the following complexity values: $C(P_{\text{Sex}}) = 0.0000$, $C(P_{\text{Income|Sex}}) = 0.4632$, $C(P_{\text{Income}}) = 0.6725$, and $C(P_{\text{Sex|Income}}) = 0.0000$, i.e., the sum of the first two values (corresponding to the true causal direction) is indeed smaller that the sum of the last two.

Using the same data set, we consider another example where a continuous variable causally influences a binary variable. We examine the continuous variable "Age" and the binary variable marriage status (short "M-Status", it takes the two values: "never married" or "married, widowed, divorced or separated"). A 10% subsample leads to the following results: $C(P_{\text{Age}}) = 0.0023$, $C(P_{\text{M-Status|Age}}) = 0.0012$, $C(P_{\text{M-Status}}) = 0.0000$, $C(P_{\text{Age|M-Status}}) = 0.0164$. The sum of the first two values (corresponding to the true causal direction) is smaller than the sum of the last two. Our causal inference rule would then favor the causal hypothesis that the age should be a cause of marriage status of a person, not vice versa.

We repeated these experiments using different subsamples of 10% of the whole data set. All subsamples yielded the same result with regarding to both causal hypotheses. However, the complexity values were slightly different for different samples. Therefore, we should not overrate the meaning of the *absolute* value of the complexity measure. Its relevance consists rather in allowing us to *compare* complexity values for different causal directions.

The third example that we tested is a data set of handwritten numerals [9] containing PCA components of the pixel vectors for the symbols "0"–"9". We considered the symbols "0" and "1" and interpreted them as the

---

[2]Daily average temperatures from 1979 through 2004, Furtwangen, Germany.

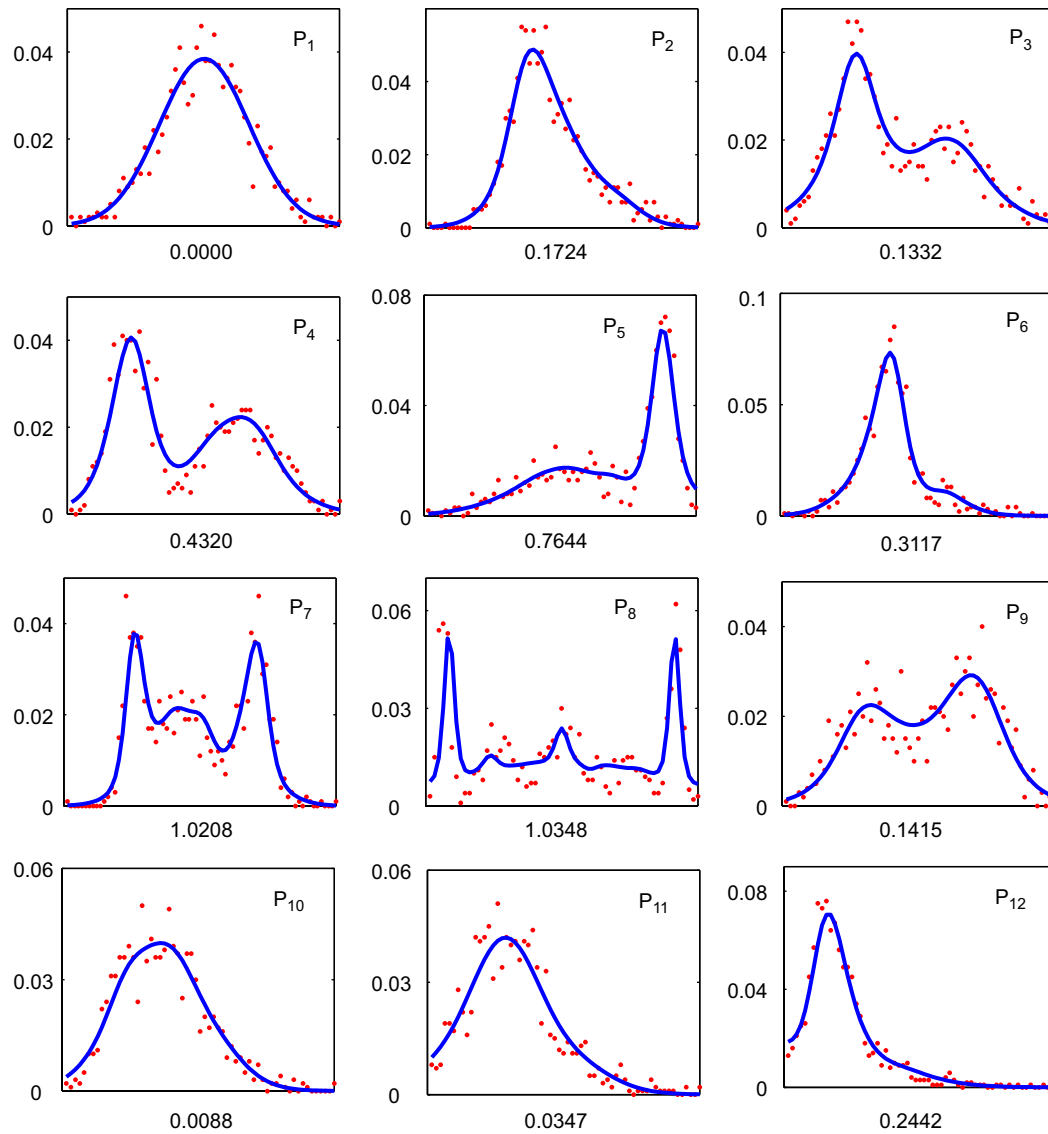[3]The data were transcribed by D. Freedman of UC Berkeley and are available online at http://www.stat.berkeley.edu/~census/.

Fig. 1. 12 toy data sets sampled by distributions $P_1, \ldots, P_{12}$ (see text). The dots indicate the observed relative frequencies, the solid lines the estimated densities. The calculated complexity values are shown below each plot.
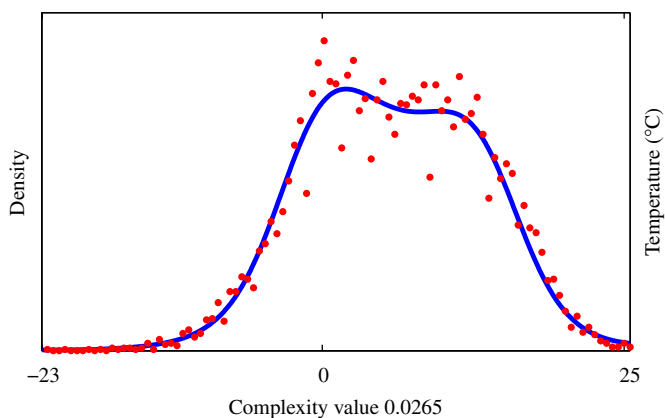

Fig. 2. Temperature data. The dots indicate the observed relative frequencies, the solid lines the estimated density. The calculated complexity value is shown below the plot.

Table 1
Complexity of conditional densities in binary mixture models

	P_2	P_3	P_4	P_9
$P_i(Y\|X = x_1)$	$\mathcal{N}(-1, 1)$	$\mathcal{N}(-2, 1)$	$\mathcal{N}(-3, 1)$	$\mathcal{N}(8, 1)$
$P_i(Y\|X = x_2)$	$\mathcal{N}(1, 4)$	$\mathcal{N}(2, 4)$	$\mathcal{N}(3, 4)$	$\mathcal{G}(9, 0.5)$
$C(P_X)$	0.0000	0.0000	0.0000	0.0000
$C(P_{Y\|X})$	0.0000	0.0000	0.0000	0.0004
$C(P_Y)$	0.1724	0.1332	0.4320	0.1415
$C(P_{X\|Y})$	0.0234	0.0000	0.0000	0.0000

values of a binary random variable X. For each symbol there are 200 instances. We chose a PCA coefficient as a continuous random variable Y. We assume that X is the cause of Y because the person first had the intention to

write the digit "1" or "0" and wrote it afterward. Hence the PCA coefficient Y is the *effect*.

We applied our inference rule to several coefficients. Their correlation with X attained, among others, the values $\rho = 0.8661$, -0.8079, 0.3233, 0.5674, 0.1086, -0.0601, -0.2547. For the cases with strong correlations we obtained results that were consistent with the ground truth, i.e., $C(P_X) + C(P_{Y|X}) < C(P_Y) + C(P_{X|Y})$. When the correlation coefficient was 0.3 or smaller, we have also observed several failures of the causal inference rule since $C(P_Y)$ and $C(P_{Y|X})$ are extremely small for these cases. This is because a density is hard to recognize as a mixture of two distributions if they are not sufficiently different.

6. Conclusions

We have presented a method to estimate the complexity of unconditional and conditional probability measures from finite samples. The complexity measure is based on an RKHS seminorm of the logarithm of the density. Experiments with real-world and toy data sets show that mixtures of two simple distributions like Gaussians and gamma distributions are recognized as more complex than the pure distributions. This confirmed that the complexity values are related to our intuitive understanding of smoothness.

Moreover, the total complexity of the true causal hypothesis $X \rightarrow Y$ given by the sum of the complexity of $P(X)$ and the complexity of $P(Y|X)$ was in most cases of our real-world experiments smaller than that of the erroneous model $X \leftarrow Y$. However, the relevance of the absolute value of complexity should not be overrated.

Every causal inference method is based on some kind of simplicity principle, which cannot be proved in principle. A final judgment on the performance of such methods requires a large number of examples from real-life data. Although the results obtained so far seem promising, we do not claim that our simplicity principle (which allows a space of functions spanned by certain simple monomials) is universally valid, since one can surely not expect that all real-world causal relationships exhibit the properties that we assumed. Different applications may require different complexity measures, but the kernel method seems to be quite flexible for designing appropriate measures by replacing the kernels k_1 and k_2.

The method presented here is computationally rather expensive. The optimization in Eq. (8) requires calculating the partition function. Due to the small size of our probability space (e.g. variables with binary range) this was nevertheless feasible. Evaluating conditionals with general continuous ranges or with more than two random variables seems (from the current perspective) to be feasible only after a coarse discretization. However, as we have shown, the tractable cases already lead to some interesting insights. They provide hints on the causal directions between only two variables (either discrete or continuous) where conventional constraint-based approaches as well as approaches using independent component analysis fail.

Due to the disadvantage that our method becomes computationally intractable for many variables we propose to apply conventional constraint-based approaches like the IC-algorithm (as well as a new algorithm using kernel-based independence tests [18]) to obtain partial information on the causal directions and apply methods of the type presented here to gain information on the causal directions that remained unspecified.

References

[1] Y. Altun, A. Smola, Unifying divergence minimization and statistical inference via convex duality, in: G. Lugosi, H. Simon (Eds.), Proceedings of the 19th Annual Conference on Learning Theory, Pittsburgh, PA, 2006, pp. 139–153.

[2] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Kluwer Academic Publishers, Norwell, MA, 2004.

[3] S. Canu, A. Smola, Kernel methods and the exponential family, Neurocomputing 69 (7–9) (2006) 714–720.

[4] S. Dasgupta, Learning mixtures of Gaussians, in: Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, 1999, pp. 634–644.

[5] J. Feldman, R. ODonnell, R. Servedio, Learning mixtures of product distributions over discrete domains, in: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, 2005, pp. 501–510.

[6] Y. Kano, S. Shimizu, Causal inference using nonnormality, in: Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion, Tokyo, Japan, 2003, pp. 261–270.

[7] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, NY, 2000.

[8] H.Q. Minh, P. Niyogi, Y. Yao, Mercer's theorem, feature maps, and smoothing, in: G. Lugosi, H. Simon (Eds.), Proceedings of the 19th Annual Conference on Learning Theory, Pittsburgh, PA, 2006, pp. 154–168.

[9] D.J. Newman, S. Hettich, C. Blake, C. Merz, UCI repository of machine learning databases ⟨http://www.ics.uci.edu/m̃learn/MLRepository.html⟩, 1998.

[10] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.

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

[12] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[13] S. Shimizu, P.O. Hoyer, A. Hyvärinen, A.J. Kerminen, A linear non-Gaussian acyclic model for causal discovery, J. Mach. Learn. Res. 7 (2006) 2003–2030.

[14] S. Shimizu, A. Hyärinen, Y. Kano, P.O. Hoyer, Discovery of non-Gaussian linear causal models using ICA, in: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, Edinburgh, Scotland, 2005, pp. 526–533.

[15] A. Smola, B. Schoelkopf, K.R. Müller, The connection between regularization operators and support vector kernels, Neural Networks 11 (1998) 637–649.

[16] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, Lecture Notes in Statistics, Springer, New York, NY, 1993.

[17] X. Sun, D. Janzing, B. Schölkopf, Causal inference by choosing graphs with most plausible Markov kernels, in: Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, 2006, pp. 1–11.

[18] X. Sun, D. Janzing, B. Schölkopf, K. Fukumizu, A kernel-based causal learning algorithm, in: Z. Ghahramani (Ed.), Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007, pp. 855–862.

[19] G. Wahba, Spline Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, PA, 1990.

Xiaohai Sun received a B.Sc. in economics from the University of International Business and Economics, Peking, China, in 1995, a Diploma in computer science from the University Karlsruhe (TH), Germany, in 2004. Since 2004, he is a Ph.D. student at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany.

Dominik Janzing received his Ph.D. in mathematics from the Eberhard-Karls-University of Tübingen, Germany, in 1998. From 1998 to 2006, he was a postdoc and senior scientist at the Computer Science Department of the University of Karlsruhe (TH). Since 2007, he is a senior scientist at the Max Planck Institute for Biological Cybernetics in Tübingen. His research interests include causal inference, physical aspects of causality and the physics of information (e.g. statistical physics and antum information).

Bernhard Schölkopf received an M.Sc. in mathematics from the University of London (1992), a Diploma in physics from the Eberhard-Karls-University, Tübingen (1994), and a doctorate in computer science from the Technical University, Berlin (1997). He has researched at the AT&T Bell Labs, at GMD FIRST, Berlin, at the Australian National University, Canberra, and at Microsoft Research, Cambridge (UK). He has taught at Humboldt University, Technical University, Berlin, and Eberhard-Karls-University, Tübingen. In July 2001, he was appointed as the scientific member of the Max Planck Society and director at the MPI for Biological Cybernetics. In October 2002, he was appointed as the Honorarprofessor for Machine Learning at the Technical University, Berlin. He won the University of London's Lionel Cooper Memorial Prize 1992 and the 1997 dissertation prize of the German Association for Computer Science (GI). He also received the prize for the best scientific project at the German National Research Center for Computer Science (GMD), 1998, and the J.K. Aggarwal Prize of the International Association for Pattern Recognition, 2006. He has been the program chair of COLT and NIPS and serves on the editorial boards of JMLR, IEEE PAMI, IJCV, and SIAM Imaging Sciences.