ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions

Xiaohai Sun[1], Dominik Janzing[2] and Bernhard Schölkopf[1]

1- Max Planck Institute for Biological Cybernetics
72076 Tübingen - Germany

2- Universität Karlsruhe (TH) - Institute for Algorithms and Cognitive Systems
76128 Karlsruhe - Germany

Abstract. We propose a method to evaluate the complexity of probability measures from data that is based on a reproducing kernel Hilbert space seminorm of the logarithm of conditional probability densities. The motivation is to provide a tool for a causal inference method which assumes that conditional probabilities for effects given their causes are typically simpler and smoother than vice-versa. We present experiments with toy data where the quantitative results are consistent with our intuitive understanding of complexity and smoothness. Also in some examples with real-world data the probability measure corresponding to the true causal direction turned out to be less complex than those of the reversed order.

1 Introduction

First, let us sketch the basic idea of our causal inference rule. Given a joint distribution on n variables, all the conditional distributions that appear in the factorization of the joint measure $P(x_1,\ldots,x_n)=P(x_1)P(x_2|x_1)\ldots P(x_n|x_1,\ldots,x_{n-1})$ will typically be smoother 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. In some cases, this approach could be very useful, in particular where the conventional constraint-based inference rules (e.g. [1, 2]) fail. In particular, our inference rule can provide some hints about causal direction between only two observed variables[1] (see [3]) where constraint-based approaches are not capable. How to quantify smoothness and simplicity of a conditional distribution concerning causal asymmetry is herewith of vital importance. In this paper, we propose to measure the complexity of a distribution by a seminorm of the function which describes the logarithm of the probability distribution. The function is an element of a reproducing kernel Hilbert space (RKHS) and its seminorm can be computed by usual kernel methods. In contrast to common machine learning applications, the complexity measure here plays not only the role of a regularizer to avoid the overfitting of finite data. It is rather considered as an interesting quantity in its own right since it should provide hints on the causal direction. For this purpose, it is important to chose a definition of complexity which is well-behaved in some respects.

[1] It is to mention that we do not intend to treat the problem of confounding on assessing causality in the current paper and assume there are no hidden common causes in our setup.

2 Measuring complexity by Hilbert space seminorms

Let P be a measure[2] on some probability space $\mathcal{X}$, $\mathcal{H}$ a Hilbert space of real-valued functions on $\mathcal{X}$ containing the set of constant functions. We define the complexity measure on the space of distributions on $\mathcal{X}$ by $C(P) := \min\{\|\phi\|^2 \,|\, \phi \in \mathcal{H}$ with $P(x) = \exp(\phi(x) - \ln z_\phi)$, with the partition function $z_\phi := \sum_x \exp(\phi(x))$. Here $\|.\|$ denotes an arbitrary seminorm on $\mathcal{H}$ given by a positive semidefinite bilinear form $B : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfying $\|\mathbf{1}\|^2 = 1$, where $\mathbf{1}$ is the function having constant values 1. One can see that $C(P) = \|Q(\ln P)\|^2$, where Q denotes the projection perpendicular to the space of constant functions with respect to B. We have proved[3]:

Lemma 1 *Let $\mathcal{H}_1$, $\mathcal{H}_2$ be spaces of functions on $\mathcal{X}_1$, $\mathcal{X}_2$, respectively. Let C_1, C_2 be complexity measures on the probability distributions on $\mathcal{X}_1$, $\mathcal{X}_2$, respectively, defined by the corresponding seminorms in $\mathcal{H}_1$, $\mathcal{H}_2$. Let P be defined by a product of measures P_1, P_2, i.e., $P(x_1, x_2) = P_1(x_1)P_2(x_2)$ for all x_1, x_2. If a complexity measure C on the distributions on $\mathcal{X}$ is based on the seminorm of $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$ with $\|a \otimes b\| := \|a\| \otimes \|b\|$, the complexity of the product measure satisfies the additivity rule: $C(P) = C_1(P_1) + C_2(P_2)$.*

Now we introduce the complexity measure of *conditional* probabilities. Let $\mathcal{X}$, $\mathcal{Y}$ be the respective value sets of random variables X, Y, and $P_{X,Y}$ be a joint distribution on $\mathcal{X} \times \mathcal{Y}$. We define the complexity $C(P_{Y|X})$ of the corresponding conditional distribution $P_{Y|X}$ as

$$\min\left\{ \|\phi\|^2 \,\middle|\, \phi \in \mathcal{H} \text{ with } P_{Y|X}(y|x) = \exp(\phi(x,y) - \ln z_\phi(x)) \right\}, \tag{1}$$

with the partition function $z_\phi(x) := \sum_y \exp\big(\phi(x,y)\big)$. The complexity of a conditional distribution can also be given in a more explicit form: $C(P_{Y|X}) = \|(\mathbf{id} \otimes Q_2)(\ln P_{Y|X})\|^2$, where "**id**" denotes the identity map. We have shown:

Lemma 2 *Let X and Y be stochastically independent with respect to the joint measure P, i.e., $P_{Y|X}(y|x) = P_Y(y)$. Let C be a complexity measure based on the Hilbert space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ and C_2 be based on the seminorm of $\mathcal{H}_2$. Then we have $C(P_{Y|X}) = C_2(P_Y)$.*

This lemma is essential, if one intends to compare the complexity of marginal probabilities P_Y to conditional probabilities $P_{Y|X}$. The intention behind the comparison is the following inference principle: Having factorized a joint measure $P_{X,Y}$ into $P_{Y|X}P_X$ and $P_{X|Y}P_Y$ based on the both possible hypothetical causal orders, one calculates the sums of the complexity $C(P_{Y|X}) + C(P_X)$ and $C(P_{X|Y}) + C(P_Y)$. We attempt to consider these sums as the "total complexity" of the causal models $X \to Y$ and $Y \to X$, respectively and prefer

[2] Let us ignore issues of sampling for the moment and assume that a probability distribution P of X is given. For the sake of convenience, we assume that the value set $\mathcal{X}$ of X is finite.

[3] All the technical details and proofs will be provided in a forthcoming full paper.

the causal direction that corresponds to a smaller total complexity. For doing so, it is crucial that $C(P_Y)$ and $C(P_{Y|X})$ are comparable and 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 *Given the assumptions above, following inequalities hold:* $C(P_{X,Y}) \geq C(P_{Y|X})+C(P_X)+C(R)-2\sqrt{C(P_X)C(R)}$ *and* $C(P_{X,Y}) \leq C(P_{Y|X})+C(P_X)+C(R)+2\sqrt{C(P_X)C(R)}$, *where R is the following probability 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 $f := (\mathbf{id} \otimes Q_2)(\ln P_{Y|X})$.*

Note that in high dimensional spaces the angle between two vectors is typically close to 90^o. We have then $C(P_{X,Y}) \approx C(P_{Y|X})+C(P_X)+C(R)$. In other words: The complexity of the joint measure is typically the sum of the complexities of the conditional probabilities and the complexity of a measure defined by the partition function. The basic intuition behind our inference rule is that simple causal mechanism may generate conditionals $P_{Y|X}$ which are simple *up to a rather complex X-dependent normalization constant*, i.e., the partition function. Then the joint distribution can also be complex even when P_X is simple due to the additional complexity of the partition function. Moreover, also $P_{X|Y}$ may then be rather complex.

3 Implicit calculation of seminorms using kernels

We have shifted the problem of defining the complexity of distribution into the definition of seminorms. In this section, we rewrite our definition such that seminorms can be calculated in an implicit way by kernels. With the so-called "kernel trick" different seminorms can be chosen by simply replacing the kernel (see e.g. [4]). Let $k : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be a positive definite symmetric function and $\mathcal{X} \times \mathcal{Y}$ the probability space considered. Let $\mathcal{H}$ be the Hilbert space spanned by the functions $k((x,y),.)$ with the inner product $\langle k((x,y),.), k((x',y'),.)\rangle = k((x,y),(x',y'))$. Then we may represent the functions ϕ in Eq. (1) by $\phi(x,y) := \sum_{j=1}^n c_j k((x_j,y_j),(x,y)) = \left\langle \sum_{j=1}^n c_j k((x_j,y_j),.)\, ,\, k((x,y),.)\right\rangle$ with appropriate coefficients c_j and points (x_j,y_j). Since $(x,y) \mapsto \psi(x,y) = k\big((x,y),.\big) \in \mathcal{H}$ defines a non-linear map into the feature space $\mathcal{H}$ we may interpret the right term in the dot product of the equation above as the feature vector associated with the point (x,y) and the left term as the vector of sufficient statistics. We will in the following use kernels that are given by the sum of two distinct kernels k_1, k_2 $k := k_1 + k_2$, i.e., k defines a feature space that is a direct sum of the feature spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ defined by k_1 and k_2, respectively. Our seminorm will then be given by the norm of the first space: $\|.\|_{\mathcal{H}}^2 := \|.\|_{\mathcal{H}_1}^2$. The idea of using a seminorm is that the space $\mathcal{H}_2$ contains extremely simple functions (for instance polynomials of low degree) that should not contribute to the complexity measure at all. Let $P_{Y|X}$ be a conditional probability measure, given by

$P(y|x) = \exp\big(\sum_{j=1}^n c_j^{(1)} k_1((x_j,y_j),(x,y)) + \sum_{j=1}^n c_j^{(2)} k_2((x_j,y_j),(x,y)) - \ln z_{\mathbf{c}}(x)\big)$ with the appropriate partition function $z_{\mathbf{c}}(x)$. The complexity $C(P_{Y|X})$ is then defined by the minimum of $\sum_{j,j'=1}^n c_j^{(1)} c_{j'}^{(1)} k_1\big((x_j,y_j),(x_{j'},y_{j'})\big)$ 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 the equation above holds.

In order to define a seminorm that is multiplicative on tensor product vectors we choose k as the product $k_1((x_j,y_j),(x_{j'},y_{j'})) = k_X^{(1)}(x_j,x_{j'})\, k_Y^{(1)}(y_j,y_{j'})$. To achieve that the constant function $\mathbf{1}$ has seminorm 1 we proceed as follows. Define the matrix $K_X := k_X^{(1)}(x_j,x_{j'})$ and calculate its inverse K_X^{-1}. Let $c := (K_X^{-1})\mathbf{1}$. Now we have to renormalize such that $\langle c\,|K_X\, c\rangle = 1$, i.e., the sum of all entries of K_X^{-1} are 1. The same procedure is also to applied to $k_Y^{(1)}$. To calculate the complexity of a distribution from a finite data set we employ the regularized maximum likelihood estimation to fit the observed data points. In contrast to the usual methods we prefer exponential models to calculate the fit in our setting. We recall that, without regularizers, the method would read as follows. Given a family of conditional distributions by $P_\phi(y|x) = \exp\big(\langle\phi|\psi(x,y)\rangle - \ln z_\phi(x)\big)$ and N data points (x_i,y_i), the maximum likelihood estimation selects ϕ by

$$\max_\phi \left\{ \frac{1}{N} \sum_{i=1}^N \big(\langle\phi|\psi(x_i,y_i)\rangle - \ln z_\phi(x_i)\big) \right\}. \tag{2}$$

In order to avoid overfitting we add a regularizer

$$\max_\phi \left\{ \frac{1}{N} \sum_{i=1}^N \big(\langle\phi|\psi(x_i,y_i)\rangle - \ln z_\phi(x_i)\big) - \epsilon\|\phi\| \right\}. \tag{3}$$

We choose the norm itself and not its square (as opposed to our complexity measure) in agreement with the choice in [5]. The authors propose to use a value of ϵ that is proportional to $1/\sqrt{N}$. Note, as an aside, that the regularized maximum likelihood estimation for unconditional distributions can also be interpreted as maximizing the entropy of the distribution subject to expectation values of $\psi(X,Y)$ that coincide with the observed means of $\psi(X,Y)$ up to an error of ϵ (see [5]). For the experiments we use a sum of the Gaussian kernel $k_\sigma((x,y),(x',y')) = \exp(-\frac{\|(x,y)-(x',y')\|^2}{2\sigma^2})$ and a modified polynomial kernel $k_{a,b}((x,y),(x',y')) = (\frac{\langle x\cdot x'\rangle}{a_x} + b_x)(\frac{\langle y\cdot y'\rangle}{a_y} + b_y)^2$, where $(x,y),(x',y')$ are arbitrary value vectors of some vector-valued variables (X,Y). The additional scaling parameters a,b is used to ensure a numerically stable training. The idea behind this choice of kernels is the following: one checks easily that the second kernel induces a space of functions spanned by the monomials $1,x,xy,xy^2,y,y^2$. Such terms are considered as so elementary that they should not contribute to the complexity measure. In particular, we can then obtain Gauss distributions whose expectation value and variance changes linearly with the given variable X. An important property is furthermore that it induces a space which contains functions that tend to minus infinity. This is needed for the description of probability measures that vanish asymptotically in the infinity. The Gaussian kernel allows us to fit all local structures of the distribution.

Our experiments suggest that we have to learn appropriate values σ for the Gaussian kernel by optimizing Equation (3), otherwise we did not obtain reasonable fits. Clearly, we cannot directly compare the complexity values corresponding to kernels with different σ. However, we may define the complexity by the minimum over all seminorms squared within some given family of RKHSs. Denoting by $\mathcal{H}_i$ the Hilbert space as the one 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 seminorm in $\mathcal{H}_i$. In order to ensure the additivity with respect to product measures in product spaces for the redefined C we need to define a family of spaces by $\mathcal{H}_i^{(1)} \otimes \mathcal{H}_j^{(2)}$. Due to a combinatorial explosion such an optimization will only be feasible for a small set I and few tensor components. If we run the optimization procedure in Eq. (3) over all Hilbert spaces (i.e., all reasonable σ) the procedure will choose the vector ϕ from the Hilbert space that leads to the least norm among all those that lead to the same value in the non-regularized optimization of Eq. (2). We shall therefore consider the optimum of Eq. (3) over all kernels taken from a given family as an estimation of the minimal norm of the distribution over all considered Hilbert spaces.

4 Experiments with toy and real-world data

The following experiments should show the intuitive meaning of our complexity measure and that it seems to make sense for inferring the causal direction between two variables. Our idea is that stochastic dependence between "cause" and "effect" which is generated by "simple mechanism" should typically lead to "simple expressions" for $P(\text{effect}|\text{cause})$ but will not necessarily generate simple expressions for $P(\text{cause}|\text{effect})$. The latter is rather an abstract mathematical expression and does not directly describe the "physics" of the causal mechanism.

First, we sample 1000 data points from different distributions. P_1 is a standard Gaussian; P_2, P_3, P_4, P_5 are mixtures of 2 Gaussians; P_6, P_7, P_8 are mixtures of $3, 4, 5$ Gaussians. P_9 is a mixture of a Gaussian and a gamma distribution. P_{10} is a single gamma distribution, P_{11}, P_{12} are mixtures of $2, 3$ gammas. As expected (Fig. 1.), the complexity of a single Gaussian is 0. A single gamma distribution is also very smooth[4]. Here, the complexity value is increasing with the number of components. This holds even for the unimodal mixture P_2, P_{11}, P_{12}.

Since the causal inference problem was the motivation for our complexity measure, its performance with respect to real-world data is be the best criterion for judging whether it seems appropriate or not. We have 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. Statistical methods show that income and gender are indeed correlated. One can exclude that the personal income influences the gender, whereas the reverse causal direction makes sense. We found that the distribution of the income averaged over both genders is more complex than the distribution for both

[4] We observe slightly larger complexity values for a gamma distribution than for a Gaussian. We would like to leave open whether this property is desirable.

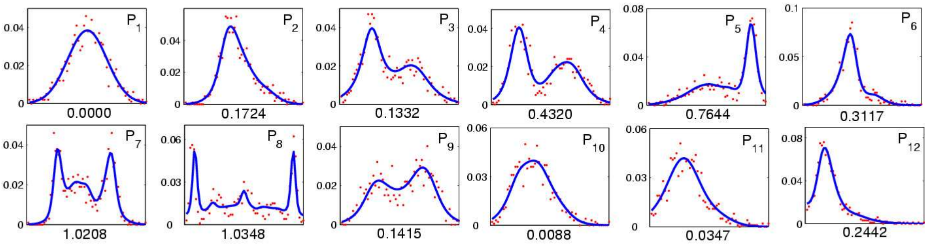

Fig. 1: 12 toy data sets sampled by distributions $P_1, ..., P_{12}$ (see text). The dots indicate the observed relative frequencies, the solid lines the estimated densities. The calculated complexity values are shown below each figure.

genders separately. Likewise, the conditional probability of the income given the gender is less complex than the marginal distribution. Having taken 10% from the total 13.803 data points randomly, we found the following complexity values: $C(P_{\text{Sex}}) = 0.0000$, $C(P_{\text{Income}|\text{Sex}}) = 0.4632$ and $C(P_{\text{Income}}) = 0.6725$, $C(P_{\text{Sex}|\text{Income}}) = 0.0000$, i.e., the sum of the complexities corresponding to the true causal direction is indeed smaller. 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 (“M-Status” takes two values: “never married” or “married, widowed, divorced or separated”). A 10% sampling shows the following results: $C(P_{\text{Age}}) = 0.0023$, $C(P_{\text{M-Status}|\text{Age}}) = 0.0012$ and $C(P_{\text{M-Status}}) = 0.0000$, $C(P_{\text{Age}|\text{M-Status}}) = 0.0164$. Our inference rule would then favor the causal hypothesis that the age should be a cause of marriage status of a person. It should be stressed that the relevance of the complexity in our applications lies rather in comparing the values of both hypothetical causal directions than in their absolute values. Further experiments with other real-life data show that our complexity measure works quite reliably in case of not too weak correlation between cause and effect. Nevertheless, we do not claim to have the right complexity measure with regard to causal reasoning, however RKHS-norms are a flexible way of constructing complexity measures having nice properties.

References

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

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

[3] X. Sun, D. Janzing, and B. Schölkopf. Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. In *Proc. of Ninth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, FL, 2006.

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

[5] Y. Altun and A. Smola. Unifying Divergence Minimization and Statistical Inference via Convex Duality. In *Proc. of the 19th Annual Conference on Learning Theory*, Pittsburgh, PA, 2006.