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

Learning causality by identifying common effects with kernel-based dependence measures

Xiaohai Sun[1] and Dominik Janzing[2]

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 describe a method for causal inference that measures the strength of statistical dependence by the Hilbert-Schmidt norm of kernel-based conditional cross-covariance operators. We consider the increase of the dependence of two variables X and Y by conditioning on a third variable Z as a hint for Z being a common effect of X and Y. Based on this assumption, we collect "votes" for hypothetical causal directions and orient the edges according to the majority vote. For most of our experiments with artificial and real-world data our method has outperformed the conventional constraint-based inductive causation (IC) algorithm.

1 Introduction

A major aim of many studies in the social, behavioral, and biological sciences is the identification of cause-effect relationships among variables or events. With the seminal work of Pearl and Spirtes et al. [1, 2] it became clear that under reasonable assumptions, it is possible to derive causal information from purely observational data. Their well-known approach for automatically generating causal hypotheses, formalized by a directed acyclic graph (DAG), is based on the Markov condition and the faithfulness assumption: Among all graphs that contain enough causal arrows to explain *all* conditional statistical dependences, one prefers those structures which allow *only* these conditional dependences. A notable algorithm based on these principles is the inductive causation (IC) algorithm. A refined version of IC is the PC algorithm[1] (after its authors Spirtes and Glymore [3]). Roughly speaking, the IC algorithm consists of three steps:

Step 1 Connect vertices $X-Y$ if and only if no set of variables (excluding X, Y) S_{XY} can be found with $X \perp\!\!\!\perp Y \,|\, S_{XY}$, i.e. X, Y are independent with respect to the conditional probability measure given all variables in S_{XY}.

Step 2 For each substructure $X - Z - Y$, where X and Y are non-adjacent, orient the edges to $X \rightarrow Z \leftarrow Y$ (a so-called v-structure), if $Z \notin S_{XY}$.

Step 3 Orient as many of undirected edges as possible subject to the condition: It should create neither a new v-structure nor a directed cycle.

However, if too few or no conditional independent relations are observed, IC would have little chance to orient the edges in step 2. Another disadvantage

[1] The PC algorithm is implemented, for instance, in the TETRAD software, available at www.phil.cmu.edu/projects/tetrad.

of IC is that the categorical (maybe erroneously) decisions for independence in step 1 will affect all the future algorithm behavior. In addition, testing independence is a challenging task in its own right. The PC algorithm, a refinement of IC, use the partial correlations for continuous variables under the assumption of multivariate normal distribution and χ^2 tests for categorical variables. This paper tries to elaborate on these problems. We argue that the *kernel-based statistical independence* measure seems to be helpful in learning causality. In particular, taking the *strength* of dependence into account, our method is robust to over-determination of statistical dependence and thus also applicable to datasets with all-over dependent networks. This requires an appropriate measure for the strength of unconditional dependence and conditional dependence. For this purpose, we extend a dependence measure proposed by Gretton et al. [4] which is based on the Hilbert-Schmidt (HS) norm of cross-covariance operators to measure *conditional* dependence.

2 Measuring statistical dependence with kernels

The idea of measuring dependence by reproducing kernel Hilbert spaces (RKHS) [5] is that statistical dependence can always be detected by correlations after data are mapped into an appropriate feature space which is only implicitly given by a kernel. Fukumizu, Bach and Jordan [6, 7] presented a similar approach for kernelized dimension reduction and independent component analysis.

2.1 Cross-Covariance Operator and Independence

First, we introduce cross-covariance operators [8] expressing correlations in the feature space and show its relation to independence of variables. Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces, and let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X}), (\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ be RKHSs of functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with measurable positive definite kernels $k_\mathcal{X}, k_\mathcal{Y}$. We consider a random vector $(X, Y)\colon \Omega \to \mathcal{X} \times \mathcal{Y}$ such that the expectations $\mathrm{E}_X\left[k_\mathcal{X}(X,X)\right], \mathrm{E}_Y\left[k_\mathcal{Y}(Y,Y)\right]$ are finite. Fukumizu et al.[7] have shown that there exists a unique operator Σ_{YX} from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ such that

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{E}_{XY}\left[f(X)g(Y)\right] - \mathrm{E}_X\left[f(X)\right]\mathrm{E}_Y\left[g(Y)\right] = \mathrm{Cov}\left[f(X), g(Y)\right]$$

holds for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. This is called the cross-covariance operator. It is known [8] that Σ_{YX} has a representation of the form $\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$, where $V_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is a unique bounded operator such that $\|V_{YX}\| \leq 1$. Bach et al. [6] have shown that $\Sigma_{YX} = 0 \Leftrightarrow X \perp\!\!\!\perp Y$ holds if Gaussian kernels[2] are used. In analogy to the conditional covariance operator in [7], we define the conditional *cross*-covariance operator as follows. Let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$, $(\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ and $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ be RKHSs on measurable spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, and (X, Y, Z) a random vector on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The conditional cross-covariance operator is defined by

$$\Sigma_{YX|Z} := \Sigma_{YX} - \Sigma_{YY}^{1/2} V_{YZ} V_{ZX} \Sigma_{XX}^{1/2} .$$

[2] For $\mathcal{X} \subset \mathrm{I\!R}^m$ with some $m \in \mathrm{I\!N}$, the so-called Gaussian radial basis function (RBF) kernel $k_\sigma(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ with parameter $\sigma \in \mathrm{I\!R}^+$.

where V_{YZ} and V_{ZX} are the bounded operators derived from Σ_{YZ} and Σ_{ZX}. It can be shown that $\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathrm{E}_Z[\mathrm{Cov}[f(X), g(Y)|Z]]$ for any $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. If we "blow up" the variables X and Y by setting $\ddot{X} := (X, Z)$ and $\ddot{Y} := (Y, Z)$ we can capture every conditional dependence using the cross-covariance operator in the sense that $\Sigma_{\ddot{Y}\ddot{X}|Z} = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z$, if Gaussian kernels are used. Furthermore, if $(X, Y) \perp\!\!\!\perp Z$, we have $\Sigma_{\ddot{Y}\ddot{X}|Z} = \Sigma_{YX} \otimes T_Z$, where T_Z is defined by $\langle h_2, T_Z h_1 \rangle := \mathrm{E}[h_1(Z)h_2(Z)]$ for arbitrary $h_1, h_2 \in \mathcal{H}_{\mathcal{Z}}$. We rescale the measure with β_Z.

Definition 1 *The strength of the marginal and conditional dependence can be respectively defined by* $\mathbb{H}_{YX} := \|\Sigma_{YX}\|_{\mathrm{HS}}^2$ *and* $\mathbb{H}_{YX|Z} := \beta_Z \left\|\Sigma_{\ddot{Y}\ddot{X}|Z}\right\|_{\mathrm{HS}}^2$ *with* $\beta_Z := 1/\|T_Z\|_{\mathrm{HS}}^2$.

By means of rescaling in this way, the measure of conditional dependence equals that of unconditional dependence, if the conditional variable Z is independent of X and Y.

Theorem 1 *We have* $(X, Y) \perp\!\!\!\perp Z \Longrightarrow \mathbb{H}_{YX|Z} = \mathbb{H}_{YX}$. *Moreover, if Gaussian kernels are used,* $\mathbb{H}_{YX} = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$ *and* $\mathbb{H}_{YX|Z} = 0 \Longleftrightarrow X \perp\!\!\!\perp Y \mid Z$.

For notational convenience, we will henceforth drop the dots on X and Y for the indices that appear in the context of *conditional* cross-covariance operators.

2.2 Empirical estimation of Hilbert-Schmidt dependence measures

In this section, we will consider the estimation of $\mathbb{H}_{YX}$ and $\mathbb{H}_{YX|Z}$ after finite sampling. It has been shown in [4] that

$$\widehat{\mathbb{H}}_{YX}^{(n)} := \frac{1}{(n-1)^2} \mathrm{Tr}\left(\widehat{K}_{\mathcal{Y}} \widehat{K}_{\mathcal{X}}\right) .$$

is a consistent estimator for $\mathbb{H}_{YX}$. Here $\widehat{K}$ is the centralized Gram matrix [9]. Fukumizu et al. [10] showed that the estimator of the cross-covariance operator guarantees to converge in HS norm at rate $n^{-1/2}$.

In some analogy to the construction of an estimator for $\Sigma_{XX|Z}$ given in [11] we have constructed a consistent estimator on $\mathbb{H}_{YX|Z}$ by

$$\begin{aligned} \widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)} \quad := \quad & \frac{\hat{\beta}_Z^{(n)}}{(n-1)^2} \mathrm{Tr}\Big(\widehat{K}_{\mathcal{Y}} \widehat{K}_{\mathcal{X}} - 2\widehat{K}_{\mathcal{Y}} \widehat{K}_{\mathcal{Z}} (\widehat{K}_{\mathcal{Z}} + \epsilon I)^{-2} \widehat{K}_{\mathcal{Z}} \widehat{K}_{\mathcal{X}} \\ & \qquad + \widehat{K}_{\mathcal{Y}} \widehat{K}_{\mathcal{Z}} (\widehat{K}_{\mathcal{Z}} + \epsilon I)^{-2} \widehat{K}_{\mathcal{Z}} \widehat{K}_{\mathcal{X}} \widehat{K}_{\mathcal{Z}} (\widehat{K}_{\mathcal{Z}} + \epsilon I)^{-2} \widehat{K}_{\mathcal{Z}}\Big) . \end{aligned}$$

Here the estimator $\beta_Z^{(n)}$ is given by $n(n-1)/\sum_{i \neq j} k_{\mathcal{Z}}(z_i, z_j)^2$ and $\epsilon > 0$ a regularization constant that enables inversion[3]. If ϵ converges to zero more slowly than $n^{-1/2}$ one can show that this estimator converges to $\mathbb{H}_{YX|Z}$.

[3] The regularizer is therefore required as the observed data are finite, whereas the feature space could be infinite dimensional. The regularization may be understood as a smoothness assumption on the eigenfunctions of $H_{\mathcal{Z}}$. Our experiments showed that the empirical measures are insensitive to ϵ, if it is chosen sufficiently small.

3 Causal learning algorithm using dependence measures

One could certainly use the above conditional dependence measure for the IC algorithm. However, it seems as if sometimes so much dependence is detected that an orientation of edges thereafter is impossible. We propose therefore the following heuristic rule: Conditioning on a common effect has the tendency to generate dependence between the causes. This is at least true if the dependence between the causes is small without conditioning. If the causes X, Y are already strongly dependent, conditioning on Z can, of course, decrease the dependence. Nevertheless we assume that it will typically decrease the dependence less than conditioning on a common *cause* would do.

Based on this intuition, we introduce a voting-like procedure for orientation of the edges: for any triple X–Z–Y (X and Y not necessarily non-adjacent) one gets a vote for Z being a common effect of X and Y, if and only if $\mathbb{H}_{YX|Z} > \lambda\, \mathbb{H}_{YX}$, with an appropriate $\lambda > 0$. Counting these votes we may direct most (not always all) edges according to the majority vote. We choose λ_1 very large in the first run. Then we set $\lambda_2 := \max\{1, \frac{\mathbb{H}_{ZX|Y}}{\mathbb{H}_{ZX}}, \frac{\mathbb{H}_{ZY|X}}{\mathbb{H}_{ZY}}\}$ and $\lambda_3 := \max\{\frac{\mathbb{H}_{ZX|Y}}{\mathbb{H}_{ZX}}, \frac{\mathbb{H}_{ZY|X}}{\mathbb{H}_{ZY}}\}$ in the second and third run. In summary, we sketch our kernel-based causal learning (KCL) algorithm as follows:

Step 1 Connect vertices $X - Y$ if and only if no set of variables (excluding X and Y) S_{xy} can be found with $\mathbb{H}_{YX|S_{xy}} < \epsilon_0$ (ϵ_0 very small).

Step 2 Direct edges as follows: (a) Check for all substructures $X - Z - Y$ (X and Y not necessarily non-adjacent) whether Z is a candidate for being a common effect of X and Y on the basis of λ_1. If this is the case the orientations $X \rightarrow Z$ and $Z \leftarrow Y$ both obtain a vote. Direct all edges which obtained at least one vote (for either of both directions) according to the majority principle. If the result is balanced, leave the edge undirected. (b) The same procedure with λ_2. (c) The same procedure with λ_3.

Step 3 As IC in Section 1.

4 Experiments

This section describes some experiments with real-world data. We have chosen datasets where some statements about the true causal structure is obvious. The first and last dataset are listed as examples for TETRAD on its project webpage.

The first dataset describes a test of food products for palatability. The experiment involved the effects on palatability of a coarse versus fine screen (large "pieces" versus small "pieces") and of a low versus high concentration of a liquid component. The dataset consists of 16 cases and three variables, i.e., SCORE: total palatability score for 50 consumers: General Foods employed a 7-point scale from −3 (terrible) to +3 (excellent) with 0 representing "average"; LIQUID: liquid level (0: "low", 1: "high"); and SCREEN: screen type (0: "coarse", 1: "fine"). The result of KCL (Fig. 1, right) is consistent with our knowledge that SCORE is the common effect of the independent causes SCREEN and LIQUID, whereas the PC algorithm (Fig. 1, left) detects merely an undirected edge between Screen and Score. Note that this is a shortcoming of the independence tests used by the standard PC. That KCL performs better is due to the kernel-based independence measures having kept the type II errors (deciding independence when there is dependence) to a lower level than the conventional approach.

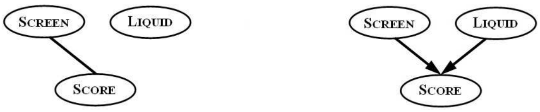

Fig. 1: Results obtained by PC (left) and KCL (right) for taste score data.

In the next experiment, the influence on sintered bodies of the variation of supplemental powder content was investigated. The supplemental powder was added to a powder mixture in different ratios (0% – 70%), from which sintered ceramic parts were fabricated. The ceramic parts were sintered at four different temperature levels: 1300°C, 1350°C, 1400°C and 1450°C. Using an optical scanning device, the surface roughness of these parts is characterized by the roughness average R_a as well as roughness depths R_z^{ISO} and R_z^{DIN}, depending on ISO or DIN standards. The dataset contains 80 measurements. We know that the supplemental Powder Content and sintering Temperature influence the Surface Roughness of sintered parts and not vice versa. In our experiments, we used different definitions for the variable Surface Roughness: R_a, R_z^{ISO}, R_z^{DIN}, (R_a, R_z^{ISO}), (R_a, R_z^{DIN}), $(R_z^{\text{ISO}}, R_z^{\text{DIN}})$, and $(R_a, R_z^{\text{ISO}}, R_z^{\text{DIN}})$. In all 7 cases KCL identified Surface Roughness as the common effect. This is an advantage of KCL against PC, since the former can be extended to multidimensional domains in a straightforward way. The result[4] of PC (Fig. 2, left) is less specific than KCL (Fig. 2, right).

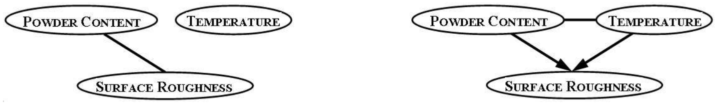

Fig. 2: Results obtained by PC (left) and KCL (right) for ceramic surface data.

As cheese ages, various chemical processes take place that determine the taste of the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese and a subjective measure of taste for each sample. Overall Taste scores were obtained by combining the scores from several tasters. The variables Acetic, H2S and Lactic

[4] We interpreted variable Surface Roughness in case of PC as a one-dimensional variable: R_a, R_z^{ISO} or R_z^{DIN}. All the three interpretations yielded the same result as (Fig. 2, left).

represent the concentrations of acetic asid, hydrogen sulfide, and lactic acid. The result of KCL (Fig. 3, right) is also more specific than PC (Fig. 3, left). The detected causal knowledge that Taste is only an effect and not a cause of any other variable is in agreement with the ground truth. Due to our lack of chemical understanding, we do not speculate on the plausibility of the influences among the various chemicals, i.e., Acetic, H2S and Lactic.

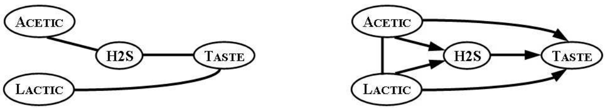

Fig. 3: Results obtained by PC (left) and KCL (right) for cheese data.

Acknowledgments

The authors gratefully acknowledge S. Norajitra, M. Beck, R. Heldele and B. Schumacher of Institute for Materials Research Ⅲ, Research Center Karlsruhe, Germany, for providing the ceramic surface data. Thanks are due to K. Fukumizu and A. Gretton for helpful comments on kernel-based dependence measures.

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] P. Spirtes and C. Glymour. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, 9(1):67–72, 1991.

[4] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Proc. Algorithmic Learning Theory*, pages 63–77, Berlin, 2005. Springer-Verlag.

[5] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[6] F. Bach and M. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[7] K. Fukumizu, F. Bach, and M. Jordan. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[8] C.R. Baker. Joint Measures and Cross-Covariance Operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[9] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neurocomputing*, 10:1299–1319, 1998.

[10] K. Fukumizu, F. Bach, and A. Gretton. Statistical Convergence of Kernel CCA. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Proc. of 18th Conf. Advances in Neural Information Processing Systems*, pages 387–394, Cambridge, MA, 2005. MIT Press.

[11] K. Fukumizu, F. Bach, and M. Jordan. Kernel Dimension Reduction in Regression. Technical Report 715, Department of Statistics, University of California, Berkeley, 2006.