Foundations of Causal Inference

Causal inference is usually concerned with exploring causal relations among random variables $X_1, \ldots, X_n$ after observing sufficiently many samples drawn from the joint probability distribution. The formal basis is the Causal Markov Condition stating that every variable is conditionally statistically independent of its non-descendants, given its parents with respect to the directed acyclic graph (DAG) that formalizes the causal relations. Apart from this local version of the Markov condition, there are other equivalent versions, e.g., the global one describing additional conditional independences that are implied by those appearing in the local one. Pearl and others have shown that the Markov condition can be justified by a functional causal model, where every node is a function of its parents and an unobserved noise term, with all noise terms being statistically independent.

However, causal conclusions in every-day life are usually not based on statistics. Instead, we even infer causal links among single objects: Significant similarities between two texts or pictures, for instance, indicate that one author or artist has copied from the other, or that both have been influenced by common third party material. There is no obvious way of interpreting such similarities as statistical dependences between random variables. We have therefore developed a causal inference scenario where the nodes of the causal DAG need not be random variables, but arbitrary mathematical objects $x_1, \ldots, x_n$ that formalize observations. We have argued that dependences between x_i and x_j can be defined by any information measure R as the difference $R(x_i) + R(x_j) - R(x_i, x_j)$, provided that the definition of R guarantees non-negativity of this expression. Postulating a stronger condition, namely submodularity, ensures that R defines a non-negative *conditional* dependence measure [16]. Using this measure, we can formulate the non-statistical analog of the different equivalent versions of the Markov condition mentioned above and shown that they are also equivalent.

To explore the conditions under which such an R-based Markov condition is related to causality, we have generalized the concept of a functional model to the non-statistical setting. The postulate that each X_j is deterministically given by its parents and the noise variable then translates into postulating that the mechanism generating x_j from its parents and its noise does not generate any additional R-information. This shows that the R-based Markov conditions link dependences to causality whenever R is appropriate for the class of causal mechanisms under consideration [16]. An interesting example is given by the Lempel-Ziv compression length of a string. It is appropriate for causal mechanisms like inserting, deleting, and combining substrings. An information measure that allows for even more complex causal mechanisms is given by Kolmogorov complexity. The corresponding "algorithmic Markov condition" is justified by an "algorithmic functional model", where every string can be computed from its parents by an appropriate program on a universal Turing machine [2]. This seems to be a weak restriction because the Church Turing principle implies that every mechanism in nature has a simulation on a Turing machine because it would otherwise be a computing device that is not Turing-computable.

We therefore consider the algorithmic Markov condition as a basis for justifying other inference methods. It includes the statistical Markov condition as a limiting case because the Shannon entropy of a random variable describes the asymptotic growth rate of the Kolmogorov complexity of a corresponding i.i.d. sample. Focusing on the growth rate, however, blurs the algorithmic information that is contained in the description of the probability distributions of the random variables. We have shown that the latter also contains valuable causal information [2]: the "principle of algorithmically independent conditionals (IC)" postulates that all the conditional probability distributions of every variable, given its direct causes, are algorithmically independent. In particular, the shortest description of $P(\text{cause}, \text{effect})$ is given by independent descriptions of $P(\text{cause})$ and $P(\text{effect}|\text{cause})$. This justifies, for instance, additive noise based causal discovery because an additive noise model in the wrong causal direction violates IC (provided that the probability distribution is sufficiently complex, which excludes cases like bivariate Gaussians) [3].

Novel Methods of Causal Inference and their impact for standard prediction tasks

In causal inference we are given a set of observations and estimate the underlying causal DAG (directed acyclic graph): each random variable is a vertex and parents are interpreted as direct causes. For approaching the problem, we have developed assumptions that make the causal graph identifiable from the joint distribution. These include versions of restricted Structural Equation Models and an approach based on Information Geometry.

In structural equation models (SEMs) each variable X_j is a function of a set of nodes $\mathbf{PA}_j$ and some noise variable N_j:

$$X_j = f_j(\mathbf{PA}_j, N_j)\,, \qquad j = 1, \ldots, p \qquad (0.1)$$

where the N_j are jointly independent, see Fig. 0.1.

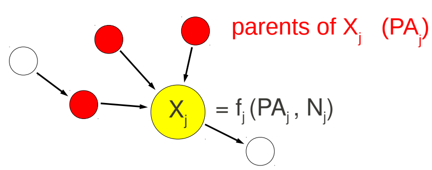

Figure 0.1: Each variable is a function of its parents and its noise

The corresponding graph is obtained by drawing directed arrows from each variable in $\mathbf{PA}_j$ to X_j (the $\mathbf{PA}_j$ become parents of X_j). In this form, SEMs are too general to be used for structure learning. For two variables X_1 and X_2, for example, any distribution can either be generated by $X_1 \to X_2$ or $X_2 \to X_1$. Traditional methods assume faithfulness and can therefore identify the Markov equivalence class of the graph; in particular, this does not help in the bivariate situation described above. As an alternative we propose *restricted* SEMs, in which some combinations of function and the distribution of noise and parents are excluded. If the structural equations satisfy an additive noise structure, that is, $X_j = f_j(\mathbf{PA}_j) + N_j$, then combinations of function, input and noise distribution only allow for $X_1 \to X_2$ and $X_1 \leftarrow X_2$ if the triple satisfies a specific differential equation [7]; the combination of linear function and Gaussian variables manifests one important exception. That is, in the generic case, the graph is identifiable from the joint distribution. A similar result holds if all variables are integer-valued [13] or if we interpret Equation (0.1) in $\mathbf{Z}/k\mathbf{Z}$ [4].

Theoretically, it is sufficient to consider the bivariate case: if a restricted functional model class allows for distinguishing between $X_1 \to X_2$ and $X_1 \leftarrow X_2$, this function class also allows for identifying a causal graph with p variables from the joint distribution [14]. Assuming that the data come from such a restricted functional model class, this result can be used in the following way: for each DAG we perform corresponding regressions and test the residuals for independence. According to the theory we should obtain independence for at most one DAG. Because the number of DAGs is growing hyper-exponentially in the number p of nodes, this approach becomes infeasible already for small values of p. [11] describes how one can estimate the graph while avoiding the enumeration of all possible DAGs. Since this method is based on the assumption of independent additive noise, one can make the independence of the residuals a criterion for regression.

In applications, we often find that the data are not i.i.d. but possess some time structure. It is therefore worthwhile to know that the identifiability results transfer to time series data, too, and constitute an improvement to the well-known Granger causality.

Additive noise models further allow for the detection of a hidden common cause. In the limit of small noise variance, one can distinguish between $X_1 \to X_2$, $X_1 \leftarrow X_2$ and $X_1 \leftarrow Z \to X_2$ with an unobserved variable Z [9].

Although the above methods inherently rely on noisy causal relations, statistical asymmetries between cause and effect can even appear for deterministic relations. We have considered the case where $Y = f(X)$ and $X = f^{-1}(Y)$, for some invertible function f, where the task is to tell which variable is the cause. Applying the general principle [2] that $P(X)$ and $P(Y|X)$ are algorithmically independent if X causes Y, we postulate that the shortest description of $P(X, Y)$ is given by separate descriptions of $P(X)$ and f. Description length in the sense of Kolmogorov complexity is uncomputable, but we can easily test the following kind of dependence: choosing $P(X)$ and f independently typically implies that $P(Y)$ tends to have high probability density in regions where f^{-1} has large Jacobian. This observation can be made precise within an information theoretic framework [6, 1] showing that applying nonlinear f to $P(X)$ decreases entropy and increases the relative entropy distance to Gaussians, provided that a certain independence between f and $P(X)$ is postulated which can be phrased as orthogonality in information space. The corresponding inference method is computationally simple and achieved positive results on real data.

While the aforementioned method requires non-linearity, there is a different approach called "trace method" for linear invertible relations between multi-dimensional variables that is related in spirit: if the covariance matrix of X and the structure matrix relating X and Y are chosen independently, directions with high covariance of Y tend to coincide with directions corresponding to small eigenvalues of A^{-1}, which can be checked by a simple formula relating traces of covariance matrices with traces of structure matrices. This way, cause and effect can be distinguished even in the Gaussian case [8] even in the regime where the dimension exceeds the number of data points [19]. The method turned out to be helpful even for noisy data.

To distinguish $X \rightarrow Y$ and $Y \rightarrow X$ and $X \leftarrow Z \rightarrow Y$ by observing $P(X, Y)$ only, is even more challenging. First steps in this direction have been made by additive noise models [5], the trace method [19], and a different method exploring the convex structure of the space of probability distributions [10].

Our recent work [15] discusses several implications of the above mentioned asymmetries between cause and effect for standard machine learning. Let us assume that Y is predicted from X.

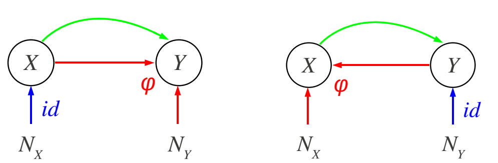

Figure 0.2: "Causal" and "anticausal" prediction scenario: predicting the effect from the cause or visa versa, respectively. The function φ describes the causal mechanism.

First, we have hypothesized that semi-supervised learning (SSL) does not help if X is the cause of Y as in Fig. 0.2 (left), whereas it often helps if Y is the cause (right). This is because additional x-values only tell us more about $P(X)$ – which is irrelevant in the case of causal prediction because the prediction requires information about the "unrelated" object $P(Y|X)$. Our meta-study analyzing results reported in the SSL-literature supports the hypothesis, see Fig. 0.3.

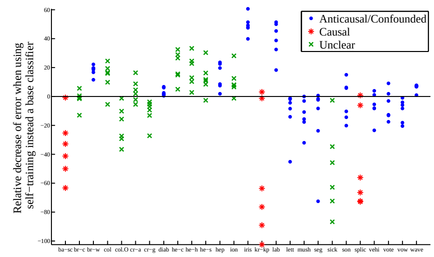

Figure 0.3: Comparison of performance of SSL for causal vs. anticausal prediction, as reported in the literature: all cases where SSL helped where anticausal, confounded, or examples where the causal structure was unclear.

The second implication refers to the case where the distribution changes from training data to test data. The fact that $P(\text{cause})$ provides no information about $P(\text{effect}|\text{cause})$ is closely related to the fact that both objects change independently across data sets. Therefore the covariate shift scenario where $P(X)$ changes and $P(Y|X)$ is constant is only justified if X is the cause. Otherwise it is more natural to assume that $P(Y)$ or $P(X|Y)$ have changed, both cases require non-trivial updates of the prediction rule $P(Y|X)$. We have discussed this for the case of additive-noise models.

In parallel to the above work, we have tried to improve traditional causal inference methods by employing non-parametric kernel independence tests that we develop. Based on the causal Markov condition and the faithfulness assumption, constraint-based methods for causal inference, such as the PC algorithm, (partially) recover the causal structure by exploiting the (conditional) independences that can be found in the data. It is natural to combine nonparametric tests with the PC algorithm for causal inference. An initial generalization of the PC algorithm using kernel tests was proposed by [17]; the KCI-test [18] subsequently yielded significant performance improvements. Because of the wide applicability of the KCI-test, such an approach produces interesting causal information on real data sets. Constraint-based methods can only recover the Markov equivalence classes of the causal structures, which are sets of graphs that impose the same independences and CIs and possibly have undetermined causal directions; if necessary, one can apply the causal inference approaches based on functional causal models (such as the post-nonlinear causal model and the one proposed in [12]) to find those undetermined causal directions.

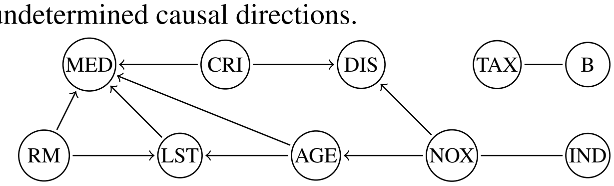

Figure 0.4: Output of KCI-test combined with the PC algorithm on the Boston housing data set.

Publications

Journal Articles

[1] D Janzing, J Mooij, K Zhang, J Lemeire, J Zscheischler, P Daniušis, B Steudel, and B Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 5 2012. 2

[2] D Janzing and B Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 10 2010. 1, 2

[3] D Janzing and B Steudel. Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17(2):189–212, 6 2010. 1

[4] J Peters, D Janzing, and B Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 12 2011. 2

[5] J Peters, J Morimoto, R Tedrake, and N Roy. Robot learning. *IEEE Robotics and Automation Magazine*, 16(3):19–20, 9 2009. 3

Articles in Conference Proceedings

[6] P Daniusis, D Janzing, J Mooij, J Zscheischler, B Steudel, K Zhang, and B Schölkopf. Inferring deterministic causal relations. In P Grünwald and P Spirtes, editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 143–150, Catalina Island, CA, USA, 7 2010. AUAI Press. Best Student Paper Award. 2

[7] PO Hoyer, D Janzing, JM Mooij, J Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *22nd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 689–696, Vancouver, BC, Canada, 6 2009. Curran. 2

[8] D Janzing, P Hoyer, and B Schölkopf. Telling cause from effect based on high-dimensional observations. In J Fürnkranz and T Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 479–486, Haifa, Israel, 6 2010. International Machine Learning Society. 3

[9] D Janzing, J Peters, JM Mooij, and B Schölkopf. Identifying confounders using additive noise models. In J Bilmes and AY Ng, editors, *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 249–257, Montréal, Canada, 6 2009. AUAI Press. 2

[10] D Janzing, E Sgouritsa, O Stegle, J Peters, and B Schölkopf. Detecting low-complexity unobserved causes. In FG Cozman and A Pfeffer, editors, *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 383–391, Barcelona, Spain, 7 2011. AUAI Press. 3

[11] JM Mooij, D Janzing, J Peters, and B Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In A Danyluk, L Bottou, and M Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 745–752, Montreal, Canada, 6 2009. ACM Press. 2

[12] JM Mooij, O Stegle, D Janzing, K Zhang, and B Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In J Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, and A Culotta, editors, *24th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1687–1695, Vancouver, BC, Canada, 2010. Curran. 3

[13] J Peters, D Janzing, and B Schölkopf. Identifying cause and effect on discrete data using additive noise models. In YW Teh and M Titterington, editors, *13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 597–604, Chia Laguna Resort, Italy, 5 2010. JMLR. 2

[14] J Peters, J Mooij, D Janzing, and B Schölkopf. Identifiability of causal graphs using functional models. In FG Cozman and A Pfeffer, editors, *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 589–598, Barcelona, Spain, 7 2011. AUAI Press. 2

[15] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262, Edinburgh, Scotland, 2012. Omnipress. 3

[16] B Steudel, D Janzing, and B Schölkopf. Causal Markov condition for submodular information measures. In AT Kalai and M Mohri, editors, *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 464–476, Haifa, Israel, 6 2010. OmniPress. 1

[17] RE Tillman, A Gretton, and P Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In Y Bengio, D Schuurmans, J Lafferty, C Williams, and A Culotta, editors, *23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1847–1855, Vancouver, BC, Canada, 2009. Curran. 3

[18] K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In FG Cozman and A Pfeffer, editors, *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813, Barcelona, Spain, 7 2011. AUAI Press. 3

[19] J Zscheischler, D Janzing, and K Zhang. Testing whether linear equations are causal: A free probability theory approach. In FG Cozman and A Pfeffer, editors, *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 839–847, Barcelona, Spain, 7 2011. AUAI Press. 3