Causal Inference

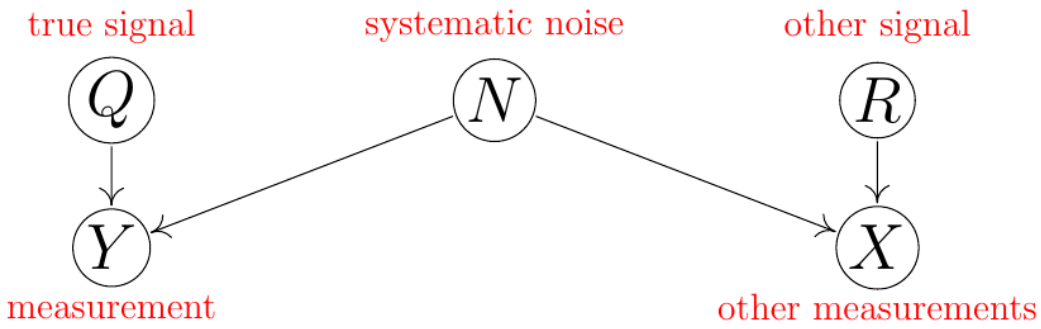

Figure 1.4: Visualization of a technique called "half-sibling regression" which corrects measurement errors. Measuring X allows us to remove parts of the systematic noise N from Y, which yields a better estimation of Q (the quantity of interest).

Causal Discovery In causal discovery, we try to learn a causal structure from data. This structure may be a directed acyclic graph (DAG) underlying a functional causal model (FCM), and possibly also the functions in the FCM. Without further assumptions, this goal is impossible to achieve: given only an observational distribution P, we can find an FCM generating P for any graph G s.t. P is Markovian w.r.t. G, i.e., the underlying graph is not identifiable. We investigate assumptions that make the graph identifiable from the observational distribution, and develop algorithms building on those assumptions. We now provide some examples, and refer to the literature for further papers that we cannot discuss below [353, 358, 401, 422, 518].

Example (1): Additive Noise Models. In additive noise models, the functional assignments used in the FCM are of the form $Z = f(X, Y) + N$. The subclass of linear functions and additive Gaussian noise does not lead to identifiability. This, however, constitutes an exceptional setting. If one assumes either (i) non-Gaussian noise, (ii) non-linear functions in the FCM [84, 624] or (iii) all noise variables to have the same variance, one can show that additive noise models are identifiable. Methods that are based on additive noise models perform above chance level not only on artifical data but also on the set of cause-effect pairs that we have collected over the last years [36]. A similar result holds if all variables are integer-valued [585] or if we interpret the additivity in $\mathbf{Z}/k\mathbf{Z}$ [200]. The concept of additive noise has been extended to time series, too [438], as well as to the identification of confounders [625].

Example (2): Information Geometric Causal Inference (IGCI). While the above methods inherently rely on noisy causal relations, statistical asymmetries between cause and effect also appear for deterministic relations. We have considered the case where $Y = f(X)$ and $X = f^{-1}(Y)$, for some invertible function f, the task being to tell which variable is the cause. The general assumption of independence of causal mechanisms [234] implies that $P(X)$ and $P(Y|X)$ should be independent if X causes Y. Choosing $P(X)$ and f independently 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 [134, 562]. Applying a non-linear 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.

A second approach, for linear invertible relations between multi-dimensional variables, 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 formula relating traces of covariance matrices with traces of the product of structure matrices with their transpose [518, 571].

Example (3): Invariant Prediction. In many situations, we are interested in the system's behavior under a change of environment. Here,

causal models become important because they are often invariant under those changes [470]. Following the assumption of independence of causal mechanisms, localized changes of some noises or mechanisms of a causal model will often leave other conditionals invariant, and thus a causal prediction (which uses only direct causes of the target variable as predictors) may remain valid even if we intervene on predictor variables or change the experimental setting. We can exploit this for causal discovery: given data from different experimental settings, we use invariance as a criterion to estimate the set of causal predictors for a given target variable. This method also leads to valid confidence intervals for causal relations [20].

Example (4): Hidden Confounding in Time Series. Assume we are given a multivariate time series $X_1, \ldots, X_L$ of measurements. In this project, our goal is to infer the causal structure underlying $X_1, \ldots, X_L$, in spite of a potential unobserved confounder $(Z_t)_{t\in\mathbb{Z}}$ (which other approaches such as Granger causality cannot handle). We assume a vector autoregressive causal model

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} := \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t.$$

Restricting the model class to non-Gaussian independent noise $(N_t)_{t\in\mathbb{Z}}$ makes B and C essentially identifiable [351]. We show that $D = 0$ is another sufficient restriction of the model class.

Example (5): Causal Strength. In real-world applications a measure of the *strength* of a causal influence is often required. This is a challenging question, even if the causal directed acyclic graph and the joint distribution are perfectly known. We have formulated a set of postulates that a measure of causal strength of an arrow (or of a set of arrows) in a causal network should satisfy [94]. We show that none of the measures in the literature satisfies all postulates, and describe examples where they therefore lead to poor results. This includes well-known approaches like Granger causality and transfer entropy for time series, and also measures for general graphs. The main problem is to quantify which part of the statistical dependences between two variables is due to a direct causal influence of one on the other, and which part is due to other causal paths. We propose an information-theoretic measure that satisfies all our postulates. It coincides with mutual information for a DAG with two nodes; for more complex DAGs it correctly removes the information propagated via alternative paths.

Causal Inference in Machine Learning We believe that causal knowledge is not only useful for predicting the effect of interventions, but that in some scenarios causal ideas can also improve the performance of classical machine learning methods. Again, we concentrate only on two examples and refer to some other papers [90, 470].

Example (6): Semi-supervised Learning. Our work [470] discusses several implications of the independence of cause and mechanism (as a special case of a more general independence principle) for standard machine learning. Let us assume that Y is predicted from X. We have argued that semi-supervised learning (SSL) does not help if X is the cause of Y ("causal learning"), whereas it often helps if Y is the cause of X ("anticausal learning"). This is because additional observations of X only tell us more about $P(X)$ – which is irrelevant in the case of causal prediction because the prediction requires information about the *independent* object $P(Y|X)$. Our meta-study analyzing results reported in the SSL-literature supports this hypothesis: all cases where SSL helped where anticausal, confounded, or examples where the causal structure was unclear. To elaborate on the link between causal direction and performance of SSL, we studied the toy problem of interpolating a monotonically increasing function for the case where the relation between X and Y is deterministic [24]. In such a scenario $P(X)$ can be shown to be beneficial for predicting Y from X whenever $P(X|Y)$ and $P(Y)$ satisfy a certain independence condition which coincides with the one postulated in our work on information-geometric causal inference [134].

Figure 1.4, finally, visualizes the idea of a recent method for correcting measurement errors called "half-sibling regression:" subtracting the conditional expectation of Y given X from Y can provide a better estimation of the quantity Q of interest than the noisy measurement Y, which has been used in [349] to process astronomical light curves for the detection of exoplanets.

More information: https://ei.is.tuebingen.mpg.de/project/causal-inference