SUPPLEMENT TO "QUANTIFYING CAUSAL INFLUENCES"

By Dominik Janzing*,
David Balduzzi*†, Moritz Grosse-Wentrup*, and Bernhard Schölkopf*

*Max Planck Institute for Intelligent Systems
†ETH Zürich

S.1. Generating an iid copy via random permutations. We now describe in which sense random permutations simulate an i.i.d. copy X'. First observe:

Lemma 1. *Let X be a discrete random variable with probability mass function $P(x)$. Given an i.i.d. sample $x_1, \ldots, x_m$. Let $\pi \in S_m$ be a random permutation (drawn uniformly from the symmetric group S_m). Let $(x, x') \mapsto \hat{P}_m(x, x')$ denote the empirical distribution of the sample $(x_1, x_{\pi(1)}), \ldots (x_m, x_{\pi(m)})$. Then $\hat{P}_m(x, x')$ converges for $m \to \infty$ almost surely to $P(x)P(x')$, where X' is an i.i.d. copy of X.*

Proof. It is sufficient to show that

$$\frac{1}{m} \sum_{j=1}^{m} f(X_j, X_{\pi(j)}) \to \mathbb{E}[f(X, X')] , \tag{S.1}$$

almost surely[1] for every function f. We only need to consider the case that π has no fix points $\pi(j) = j$ because this occurs only with probability $1/m$. It is clear that for every such π

$$\mathbb{E}[f(X_j, X_{\pi(j)})] = \mathbb{E}[f(X, X')] ,$$

(where X, X' are distributed according to $P(X)P(X')$) because X_j ad $X_{\pi(j)}$ are independent. Thus, the expectation of the left hand side of (S.1) coincides with the right hand side for all m. It therefore remains to show that the left hand side converges to its expectation almost surely. If $(X_1, X_{\pi(1)}), \ldots, (X_m, X_{\pi(m)})$ were a sequence of independent pairs, this would be the usual law of large numbers. Here, the dependence is still weak enough to ensure convergence. To show this, assume without loss of generality that $Z_j := f(X_j, X_{\pi(j)})$ has zero mean. Then one easily checks that

$$\sum_{j=1}^{m} \sum_{i=1}^{m} \mathbb{E}(Z_i Z_j) \in O(m) ,$$

which is a sufficient condition for $\sum_{j=1}^{m} Z_j/m$ converging to zero almost surely ([1], Theorem 2.1).

■

For our application, we need a slightly stronger version that ensures that the permuted sample is also independent of the other parents of X_l:

[1] Note that the left hand side formally is a function of the $m + 1$ random variables $X_1, \ldots, X_m, \pi$.

LEMMA 2. *Let X, W be two random variables with joint density $P(x,w)$. Given an iid sample (x_j, w_j) with $j = 1,\ldots,m$. Let $\pi \in S_m$ be a random permutation. Then the empirical distribution of the sample $(x_j, w_j, x_{\pi(j)})$ converges for $m \to \infty$ almost surely to $P(x,w)P(x')$ where X' is an i.i.d. copy of X.*

PROOF. Using vector valued random variables in Lemma 1, we obtain $P(X,W)P(X',W')$ by jointly permuting x, w. Then the statement follows by marginalizing over W'. ∎

S.2. Another option to define causal strength. We now discuss a slightly different measure of causal strength that should be mentioned because it can easily be confused with ours. Although it has many nice properties and it is quite intuitive, it fails satisfying Postulate 3. To define the strength of the arrow $X \to Y$ we consider X and its parents PA_Y and define a modified joint distribution on PA_Y by

$$P'(X, PA_Y^X) := P(X)P(PA_Y^X)P(Y|PA_Y^X)\,.$$

In words: we remove the dependencies between X and the other parents PA_Y^X of Y. Then we could define causal strength by the conditional mutual information

$$I_{P'}(X;Y\,|PA_Y^X)$$

with respect to the modified distribution. The modification can be thought of describing the post-interventional distribution where X is set to the values x according to the observed marginal distribution $P(X)$. To show that Postulate 3 is violated, we consider the case where two dependent variables X and Z influence Y. Let X consist of $k+1$ bits, Y be k bits and Z be just one bit. Call the first bit of X the control bit and the remaining k bits of X the message. Define $P(Y|X,Z)$ such that the message bits are copied to Y whenever both the control bit of X and the variable Z are set to 1. Otherwise, they are uniformly distributed on $\{0,1\}^k$. To specify $P(X,Z)$, we first define a distribution on $P(X_1,Z)$ by

$$P(x_1,z) = \begin{cases} 1/2 & \text{for } x_1 = z \\ 0 & \text{otherwise} \end{cases}$$

Then we set

$$P(X,Z) := P(X_1,Z)P(X_2,\ldots,X_k)\,,$$

where $P(X_2,\ldots,X_k)$ is the uniform distribution on $\{0,1\}^k$. It easy to see that

$$I(X;Y\,|Z) = k/2$$

because the k message bits are copied whenever $Z = X_1 = 1$, which happens with probability $1/2$. However, modifying P to P' breaks the coupling between X_1 and Z, and $X_1 = Z = 1$ only happens with probability $1/4$. Thus, the message bits are only copied with probability $1/4$ and therefore

$$I_{P'}(X;Y\,|Z) = I_{P'}(X_1;Y\,|Z) + I_{P'}(X_2,\ldots,X_{k+1};Y\,|X_1,Z) = 0 + k/4 < I(X;Y\,|Z)\,,$$

while Postulate 3 requires

$$I_{P'}(X;Y\,|Z) \geq I(X;Y\,|Z)\,.$$

S.3. The problem of defining total influence. If X influences Y via directed paths other than a direct arrow, we may want to measure the total influence of X on Y. Clearly, we cannot quantify total influence by quantifying the impact of removing all the arrows on the directed paths connecting X and Y. To see this, consider the causal chain $X \to Z \to Y$ and assume that X and Z are strongly coupled (e.g. by a copy operation), but Z and Y are weakly coupled (e.g. a very noisy copy operation). Then, removing both arrows $X \to Z$ and $Z \to Y$ has a large impact on $P(X, Z, Y)$ although Y obtains almost no signal from X. In this simple case, total influence may be defined by replacing the path $\to Z \to$ with a single arrow and computing the direct influence in $X \to Y$ after marginalizing $P(X, Z, Y)$ to $P(X, Y)$. However, we do not see how to construct a general rule for shrinking total influence to a direct arrow. To describe the problem, consider DAG d) in Figure 1 in the main article. Since X influences Y only directly, we may want to consider $\mathfrak{C}_{X \to Y} = I(X; Y \,|Z)$ also as the strength of the total influence. On the other hand, for case b) in Figure 2, we would tend to consider $I(X; Y)$ as the total influence. This is because shrinking the DAG to a direct influence would represent the net effect of both influences (the direct one $X \to Y$, and the indirect one $X \to Z \to Y$) into the single arrow $X \to Y$. The distribution $P(X, Y)$ is simply given by marginalizing $P(X, Z, Y)$. On the other hand, DAG d) in Figure 1 can be considered as a special instance of the DAG in Figure 2 b) for the limit of weakening the influence of X on Z. Inserting the "virtual edge" $X \to Z$ (i.e., an arrow with zero strength) in Figure 1d) then changes the total influence from $I(X; Y \,|Z)$ to $I(X; Y)$. If, for instance $Y = X \oplus Z$ and X and Z are unbiased coins, $I(X; Y) = 0$ but $I(X; Y \,|Z) = 1$.

References.

[1] B. Ninness. Strong law of large numbers under weak assumptions with applications. *IEEE Transactions on Automatic Control*, 45(11):2117–2122, 2000.

Dominik Janzing
Spemannstr. 38
72076 Tübingen
Germany
E-mail: dominik.janzing@tuebingen.mpg.de

David Balduzzi
CAB F 63.1
Universitätstrasse 6
8092 Zurich
Switzerland
E-mail: david.balduzzi@inf.ethz.ch

Moritz Grosse-Wentrup
Spemannstr. 38
72076 Tübingen
Germany
E-mail: moritz.grosse-wentrup@tuebingen.mpg.de

Bernhard Schölkopf
Spemannstr. 38
72076 Tübingen
Germany
E-mail: bernhard.schoelkopf@tuebingen.mpg.de