SCM and z-identification

In this note we explore situations in which it is impractical to directly manipulate the causal or exposure variable of interest, say X, and instead attempt to identify the causal effect of interest by auxiliary experiments manipulating another variable, say Z. We refer to this as z-identification.[1]

z-identification is nonparametrically achievable whenever permitted by the rules of do-calculus (Pearl, 1995) and X is only employed passively (observation, no *do*-operators on X). The rules of do-calculus are below.

do-calculus

Let G be the DAG associated with a causal model and let $\Pr()$ be the probability distribution induced by the model. For any dis-joint set of variables X, Y, Z, and W the following rules apply.

Rule 1 (insertion/deletion of observations):
$\Pr(y \mid do(x), z, w) = \Pr(y \mid do(x), w)$ if $(Y \perp Z \mid X, W)_{G_{\overline{X}}}$ where $\perp$ refers to stochastic independence or d-separation in the graph.

Rule 2 (action/observation exchange):
$\Pr(y \mid do(x), z, w) = \Pr(y \mid do(x), do(z), w)$ if $(Y \perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$.

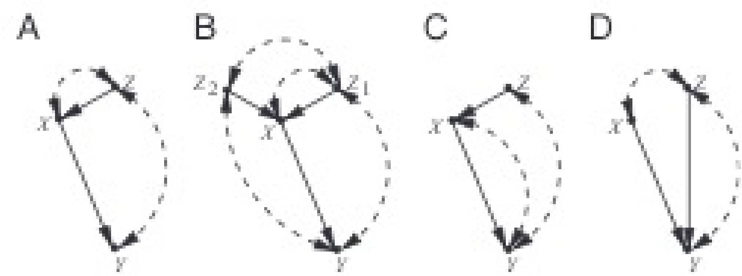
Rule 3 (insertion/deletion of actions):
$\Pr(y \mid do(x), w) = \Pr(y \mid do(x), do(z), w)$ if $(Y \perp Z \mid X, W)_{G_{\overline{X\,Z(W)}}}$
where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$.

z-identification

The quantity of interest is the causal effect of X on Y, $\Pr(Y = y \mid do(X = x))$. z-identification here refers to nonparametric identification of the effect by manipulation of other variables, say Z, rather than X. **z-identification** is feasible if and only if X intercepts all directed paths from Z to Y and $\Pr(y \mid do(x))$ is identified in $G_{\overline{Z}}$.

Consider the DAGs below.

[1] This note draws from Bareinboim and Pearl (2012), "Causal inference by surrogate experiments: z-identifiability," *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* and Bareinboim and Pearl (2016), "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Science.*



In DAG A, X intercepts the only directed path from Z to Y and there is no confounding back-door into X connected to Y in the subgraph $G_{\overline{Z}}$; hence, z-identification is feasible. Specifically, do-calculus rule 3, insertion of action Z, is satisfied. Therefore,

$$\Pr\left(y \mid do\left(x\right)\right) = \Pr\left(y \mid do\left(x\right), do\left(z\right)\right)$$

Further, rule 2 applies to the subgraph $G_{\overline{Z}\underline{X}}$ and

$$\Pr\left(y \mid do\left(x\right), do\left(z\right)\right) = \Pr\left(y \mid \ x \ , do\left(z\right)\right)$$

Action is replaced by observation of X. This latter quantity can also be written

$$\Pr\left(y \mid do\left(x\right)\right) = \Pr\left(y, x \mid do\left(z\right)\right) / \Pr\left(x \mid do\left(z\right)\right)$$

In either case, all *do*-terms or manipulations only involve Z so the causal effect is estimable from available data.

In DAG B, X intercepts the directed path from Z_1 to Y and Z_2 blocks the confounding back-door path into X connected to Y in the subgraph $G_{\overline{Z_1}}$; hence, z-identification is feasible.

$$\Pr\left(y \mid do\left(x\right)\right) = \Pr\left(y \mid do\left(x\right), do\left(z_1\right)\right)$$

$$= \sum_{z_2} \Pr\left(y \mid do\left(x\right), do\left(z_1\right), z_2\right) \Pr\left(z_2 \mid do\left(x\right), do\left(z_1\right)\right)$$

$$= \sum_{z_2} \Pr\left(y \mid x, do\left(z_1\right), z_2\right) \Pr\left(z_2 \mid do\left(x\right), do\left(z_1\right)\right)$$

$$\Pr\left(y \mid do\left(x\right)\right) = \sum_{z_2} \Pr\left(y \mid x, do\left(z_1\right), z_2\right) \Pr\left(z_2\right)$$

The first line utilizes rule 3 to insert $do\left(z_1\right)$ via subgraph $G_{\overline{XZ_1}}$. The second line utilizes Bayes chain rule to insert Z_2. The third line employs rule 2 to exchange observation with action on X in the first term. The fourth line applies rule 3 to delete X and Z_1 from the second term.

DAGs C and D are **not** z-identified. While X blocks the directed path from Z to Y in DAG C, the confounding bow between X and Y prevents identifying $\Pr(y \mid do(x))$ in subgraph $G_{\overline{Z}}$ so manipulation through Z fails to identify the causal effect of X on Y.

The failure of z-identification in DAG D is the reverse of that in DAG C. While $\Pr(y \mid do(x))$ in subgraph $G_{\overline{Z}}$ is identified, the path from Z to Y is unblocked by X. Consequently, do-calculus (in particular, rules 2 and 3) cannot be employed to satisfy z-identification.