Recovering Joint from two Biased Distributions

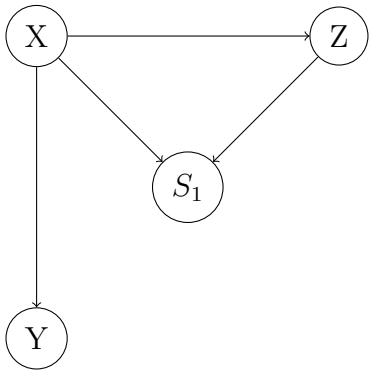
Canyon Foot

December 1, 2019

1

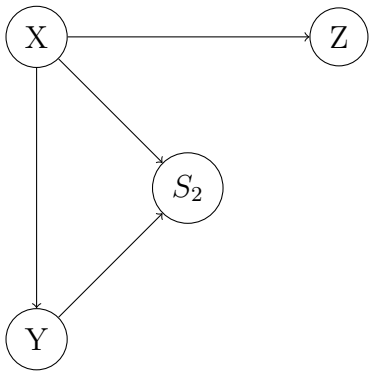
Here I give a simple example of two causal graphs measuring the same variables but with different selection mechanisms. In this case, the full joint distribution is recoverable only if we have access to the distributions for both graphs. We assume external information $P(x)$.

G_1



As always, we have access to $P(x, y, z|S_1 = 1)$, and since $Y \perp\!\!\!\perp S|X$, we also have $P(y|x)$. Additionally, since we have assumed access to $P(x)$, we could obtain $P(x, y)$ if we wanted. Notice the because of their is an edge for Z to S, we cannot obtain $P(z)$.

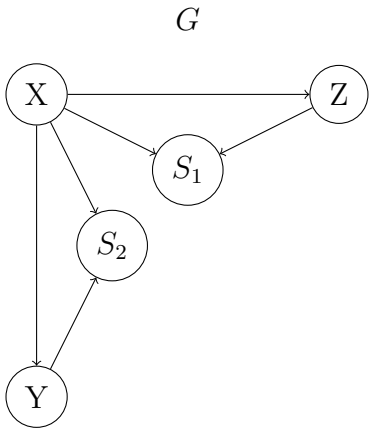
G_2



The story here is similar. We can get $P(x, y, z|S_2 = 1)$, $P(z|x)$, and $P(x, z)$ (using the external data). Then, using the chain rule for probability,

$$\begin{aligned} P(X, Y, Z) &= P(Y|X, Z)P(X, Z) \\ &= P(Y|X)P(Z|X)P(X) \end{aligned}$$

Since we have all of these quantities we can find $P(X, Y, Z)$ without issue. However, to do so we need to use $P(Z|X)$ and $P(Y|X)$, which means we need both biased distributions. To really investigate what happens when multiple selection biased distributions are present, we will want to express the selection mechanisms within a single graph. To do so, we simply 'superimpose' the graphs. So, for our current example, we get a new graph:



A few things of note. We do not allow the selection nodes S_1 and S_2 to have children, and therefore any path passing through a selection node (i.e. $V \to S_i \leftarrow W$) will include a collider at S_1. This means that the definition of path blocking gives that any such path is blocked, and therefore any d-separation statement that holds in one of the original graphs will hold in the superimposed graph. This fact allows us to apply the first and second theorems from Barenboim 2014 to get the following propositions.

Proposition 1.1. *A conditional distribution $P(y|\mathbf{x})$ can be recovered from a if there is a selection node S_i such that $Y \perp\!\!\!\perp S_i|\mathbf{X}$.*

And similarly,

Proposition 1.2. *When external information $P(\mathbf{x}, \mathbf{c})$ is available, the conditional distribution $P(y|x)$ is recoverable if there is a selection node S_i such that $Y \perp\!\!\!\perp S_i|(\mathbf{X}, \mathbf{C})$.*

These propositions are little more than sanity checks since they state that any distribution recoverable from one of the biased distribution alone is recoverable when all are accessible. What is more interesting is establishing the criteria under which distributions not recoverable from any of the biased distributions in isolation are recoverable when the collection of distributions are accessible.