Performing independence tests

Consider the following dataset consisting of three variables.

```
>>> import numpy as np, pandas as pd
>>>
>>> X = np.random.normal(loc=0, scale=1, size=1000)
>>> Y = 2 * X + np.random.normal(loc=0, scale=1, size=1000)
>>> Z = 3 * Y + np.random.normal(loc=0, scale=1, size=1000)
>>> data = pd.DataFrame(data=dict(X=X, Y=Y, Z=Z))
```

Assume that the provided causal graph is $X \to Y \to Z$. This graph implies that X should be independent of Z given Y. To test this condition, whether X is conditionally independent of Z given Y, using the kernel dependence measure, all you need to do is:

```
>>> import dowhy.gcm as gcm
>>>
>>> # Null hypothesis: x is independent of y given z
>>> p_value = gcm.independence_test(X, Z, conditioned_on=Y, method='kernel')
>>> p_value
0.48386151342564865
```

If we define a threshold of, for instance, 0.05, we would only reject independence if the p-value falls below this. Note that we can strictly speaking only reject independence, but cannot accept it if the p-value is above the threshold. However, in practice, one might still consider the result above as some evidence that the variables are conditionally independent. This is also what we would expect, given that we generated the data using the causal graph $X \to Y \to Z$, where Z is conditionally independent of X given Y.

Suppose that instead of the true graph, we had obtained the following graph for the dataset: $X \to Y \leftarrow Z$. According to the graph, Y is a collider and hence the graph implies that X and Z are independent. To test whether X is independent of Z (*without* conditioning on Y), we use the same function without the third argument.

Skip to main content

```
>>> # Null hypothesis: x is independent of y
>>> p_value = gcm.independence_test(X, Z, method='kernel')
>>> p_value
0.0
```

Again, we can define a threshold of 0.05, but this time the p-value is clearly below this threshold and we can clearly reject the null hypothesis of independence. That is, X and Z *are* dependent on each other. Therefore, the graph is incorrect and needs to modified before using it for any causal task. Again, this is what we would expect, since in our dataset, Z is dependent on Y and Y is dependent on X, but we don't condition on Y.

Previous
Refuting a Causal Graph

Next
Graph refutations