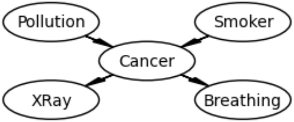
A Simple Unified Approach to Testing High-Dimensional Conditional Independencies for Categorical and Ordinal Data

Ankur Ankan Johannes Textor

Overview

Motivation: Example DAG / Causal Bayesian Network



An example of Directed Acyclic Graph (DAG) [1]

- Random variables are represented using nodes.
- Directed edges represent direct causal link between variables.
- Each variable is conditionally independent of all non-descendants given its parents. E.g.

  $XRay \perp Pollution | Cancer$

  $Breathing \perp Smoker | Cancer$

[1] K. B. Korb, A. E. Nicholson. Bayesian Artificial Intelligence

Motivation: Model Testing

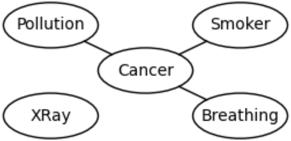
- In applied research, most of the DAGs are made by hand based on domain knowledge.
- Important to test whether the model is consistent with the data.
- Conditional Independence (CI) tests can be used to verify model structure.

```
                           x2 df    p.value
Brth _||_ Pllt | Cncr 4.7571803  2 0.09268115
Brth _||_ Smkr | Cncr 9.0058063  2 0.01107679
Brth _||_ XRay | Cncr 1.9104270  2 0.38472999
```

Example model testing output from R package *dagitty*

Motivation: Structure Learning

- CI implies that no direct causal link exists between the variables.
  $XRay \perp Smoker | Cancer \implies$ No edge b/w $XRay$ and $Smoker$
- Constraint-Based structure learning algorithms like PC and FCI use CI tests to systematically search for CIs in the dataset to determine model skeletons.



Structure Learning Example

(Conditional) Independence

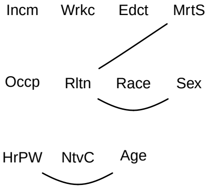
Independence

Two random variables X and Y are independent, $X \perp Y$ if and only if $P(X, Y) = P(X) \cdot P(Y)$.

Conditional Independence

Two random variables X and Y and are said to be conditionally independent given $\boldsymbol{Z}$, $X \perp Y | \boldsymbol{Z}$ if and only if for all z with $p(z) > 0$, $P(X, Y | Z = z) = P(X | Z = z) \cdot P(Y | Z = z)$

CI Testing is Difficult



Learned structure for US census income dataset using chi-square test

- Testing for CI is much harder compared to testing for non-conditional independence.
- Especially in case of high cardinality or high number of conditional variables.
- In the continuous case, no test can exist which is calibrated and has power over all distributions where CI is True. [2]
- Many different approaches and tests have been proposed.

[2] Shah, Rajen D., and Jonas Peters. "The hardness of conditional independence testing and the generalised covariance measure." The Annals of Statistics, 2020

Main classes of tests

- Stratification based tests
- Variable Importance based tests
- Residulaization based tests

Stratification Based Tests

- Most common type for discrete variables. E.g. chi-square, mutual information based test etc.
- Converts CI test into simple independence test by splitting the dataset.
  $D[X, Y, \boldsymbol{Z}] = \{D[X, Y, \boldsymbol{Z} = \boldsymbol{z_1}], D[X, Y, \boldsymbol{Z} = \boldsymbol{z_2}], \cdots\}$
- Runs test on each stratum and then combines the results.
- As the number of conditional variables is increased, exponentially less data is available in each stratum.
- Looses power when number of conditonal variables are increased.

Variable Importance Tests

- Based on comparing the probability models: $\hat{p}(x|y,z)$ and $\hat{p}(x|z)$. E.g. Stochastic Complexity-Based Conditional Independence Test (SCCI) [3].
- If the simpler model doesn't fit significantly worse, implies $X \perp Y|Z$.
- Can utilize any statistical model for which a reasonable goodness of fit exist.
- Inherently asymmetrical. The result of $X \perp Y|Z$ can be different from $Y \perp X|Z$.

[3] Marx, Alexander, and Jilles Vreeken. "Testing conditional independence on discrete data using stochastic complexity." PMLR, 2019

Residualization Based Tests

- Uses two estimators $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$ and checks for the multiplicative association between the residuals. E.g. Partial Correlation test, generalized covariance measure etc.
- Relies on the theorem from Daudin [1980] [4]that under CI, if the estimators have "valid" residuals such that $\mathbb{E}[R_{X|Z}] = \mathbb{E}[R_{Y|Z}] = 0$, then $\mathbb{E}[R_{X|Z}R_{Y|Z}] = 0$.
- Any estimator can be used as long as it has "valid" residuals.
- No residualization based test exists for categorical or ordinal variables.

[4]Daudin, J. J. "Partial association measures and an application to qualitative regression." Biometrika, 1980

Proposed Method

- Residualization based approach.
- Uses Li-Shepherd (LS) residuals [5].
- Any unbiased estimator can be used. We show empirical results using Logistic Regression (GLM) and Random Forest (RFT).

[5] C. Li and B. E. Shepherd. "A new residual for ordinal outcomes." Biometrika, 2012

LS-Residuals

Given an ordinal variable Y and an estimate $\hat{p}(y)$ of $p(y)$, LS-Residual for sample y_i is defined as:

$$R_{y_i} = \hat{p}(Y < y_i) - \hat{p}(Y > y_i)$$

For the binary case with $Y \in \{0, 1\}$:

$$R_{y_i} = y_i - \hat{p}(Y = 1)$$

For the conditional case for sample $(y|z)_i$,

$$R_{y_i|z_i} = \hat{p}(Y < y_i|Z = z_i) - \hat{p}(Y > y_i|Z = z_i)$$

Proposition

If $X \perp Y|Z$ and $\hat{p}(x|z)$ and $\hat{p}(y|z)$ are asymptotically unbiased estimators of $p(x|z)$ and $p(y|z)$ respectively, then $\mathrm{Cov}(R_{\boldsymbol{x}|\boldsymbol{z}}, R_{\boldsymbol{y}|\boldsymbol{z}}) = 0$ in large sample limit.

- For asymptotically unbaised estimators, LS-Residuals gives "valid" residuals: $E[R_{X|Z}] = E[R_{Y|Z}] = 0$.
- Under $X \perp Y|\boldsymbol{Z}$, valid residuals imply $\mathbb{E}[R_{X|Z} R_{Y|Z}] = 0$ [6].

[6] Daudin, J. J. "Partial association measures and an application to qualitative regression." Biometrika, 1980

Test Statistic: Both ordinal variables

$$Q_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \frac{(R_{\mathbf{x}} \cdot R_{\mathbf{y}})^2}{\boldsymbol{var}(R_{\mathbf{x}} R_{\mathbf{y}})}$$

If $X \perp\!\!\!\perp Y | Z$, then asymptotically $Q_1(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2(1)$.

- Train two estimators: $E_X = \mathbf{x} \sim \mathbf{z}$ and $E_Y = \mathbf{y} \sim \mathbf{z}$
- Make probability predictions for each data point: $\hat{p}(x|z)$ and $\hat{p}(y|z)$ using E_X and E_Y respectively.
- Compute the LS-Residuals for each data point: $R_{\mathbf{x}}$ and $R_{\mathbf{y}}$.
- Use $R_{\mathbf{x}}$ and $R_{\mathbf{y}}$ to compute Q_1.

Test Statistic: Both ordinal variables

$$Q_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \frac{(R_{\boldsymbol{x}} \cdot R_{\boldsymbol{y}})^2}{\boldsymbol{var}(R_{\boldsymbol{x}} R_{\boldsymbol{y}})}$$

If $X \perp Y | Z$, then asymptotically $Q_1(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2(1)$.

- From the first proposition, population mean $\mathbb{E}[R_X R_Y] = 0$
- From Central Limit Theorem, the standardized sample mean of $R_x R_y$, $\frac{1}{n} \frac{R_x \cdot R_y}{\sigma(R_x R_y)} \sim \mathcal{N}(0, \frac{1}{\sqrt{n}})$.
- Q_1 is chi-squared distributed with 1 degree of freedom (df).

Test Statistic: One ordinal and one categorical

$$Q_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}(d \times \hat{\Sigma}_d^{-1} \times d^T)$$

where $d = (R_{\mathbb{I}(\mathbf{x}=1)} \cdot R_{\mathbf{y}}, \ \ldots \ , R_{\mathbb{I}(\mathbf{x}=k-1)} \cdot R_{\mathbf{y}})$ and $\hat{\Sigma}_d$ is the covariance matrix.

If $X \perp\!\!\!\perp Y | Z$, then asymptotically $Q_2(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2(k-1)$.

- Dummy/one-hot encode the categorical variable.
- Similar to last case, train two estimators: $E_X = \mathbf{x} \sim \mathbf{z}$ and $E_Y = \mathbf{y} \sim \mathbf{z}$ and make probability predictions using them: $\hat{p}(x|z)$ and $\hat{p}(y|z)$.
- Compute the LS residuals for each dummy variable assuming them to be binary ($R_{\mathbf{x}}$) and the ordinal variable ($R_{\mathbf{y}}$).
- d is the product of residual from each dummy variable and the ordinal variable's residual.

Test Statistic: One ordinal and one categorical

$$Q_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}(d \times \hat{\Sigma}_d^{-1} \times d^T)$$

where $d = (R_{\mathbb{I}(\mathbf{x}=1)} \cdot R_{\mathbf{y}}, \ldots, R_{\mathbb{I}(\mathbf{x}=k-1)} \cdot R_{\mathbf{y}})$ and $\hat{\Sigma}_d$ is the covariance matrix.

If $X \perp Y | Z$, then asymptotically $Q_2(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2(k-1)$.

- Under CI, each component of d is asymptotically normal.
- Components of d are linearly correlated. Hence, d is a multivariate gaussian distributed.
- Q_2 is chi-squared distributed with $k-1$ df.

Test Statistic: Both categorical

$$Q_3(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}(d \times \hat{\Sigma}_d^{-1} \times d^T)$$

where

$$\begin{aligned} d \quad = \quad & (R_{\mathbb{I}(\mathbf{x}=1)} \cdot R_{\mathbb{I}(\mathbf{y}=1)}, \dots, R_{\mathbb{I}(\mathbf{x}=k-1)} R_{\mathbb{I}(\mathbf{y}=1)}, \dots, \\ & R_{\mathbb{I}(\mathbf{x}=1)} \cdot R_{\mathbb{I}(\mathbf{y}=r-1)}, \dots, R_{\mathbb{I}(\mathbf{x}=k-1)} R_{\mathbb{I}(\mathbf{y}=r-1)}) \end{aligned}$$

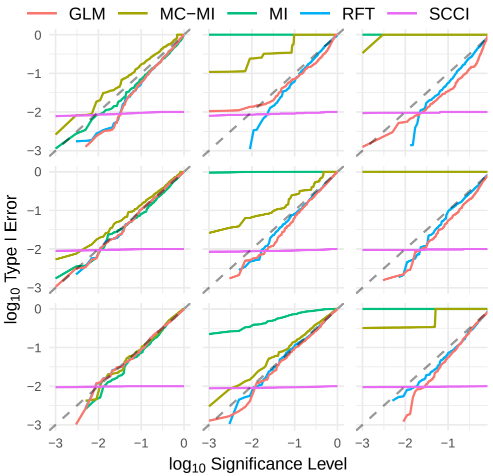
If $X \perp Y | Z$, then asymptotically $Q_3(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2((k-1)(r-1))$.

- Same as the last case, $Q_3(\boldsymbol{x}, \boldsymbol{y})$ is chi-squared distributed with $(k-1)(r-1)$ df.

Test Summary

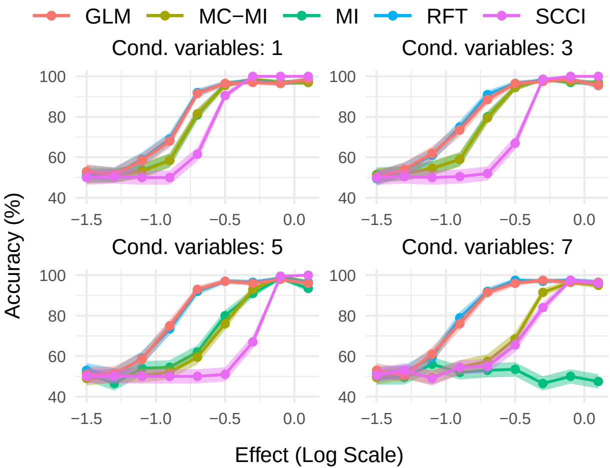
1. If $\mathbf{Z} = \emptyset$, do a non-conditional chi-squared test.
2. If either X or Y are non-binary categorical, dummy/one-hot encode them.
3. Train two estimators $E_x = \boldsymbol{x} \sim \boldsymbol{z}$ and $E_y = \boldsymbol{y} \sim \boldsymbol{z}$
4. Make probability predictions using these two estimators $\hat{p}(x) = E_x(\boldsymbol{z})$ and $\hat{p}(y) = E_y(\boldsymbol{z})$.
5. Use predictions and true values to compute LS-Residuals $R_{\boldsymbol{x}|\boldsymbol{z}}$ and $R_{\boldsymbol{y}|\boldsymbol{z}}$.
6. Compute the test statistic and df.

Empirical Analysis: Calibration



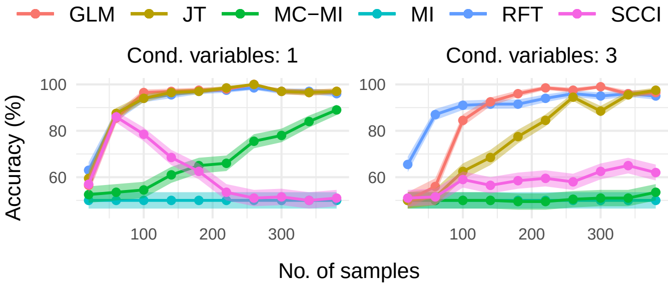
Type I error vs significance level for sample sizes (top to bottom): $[20, 40, 80]$ and number of conditional variables (left to right): $[1, 3, 5]$ on conditionally independent simulated binary datasets.

Empirical Analysis: Discrimination



Accuracy (shading: mean $\pm$ standard error, $N = 200$) of classifying simulated binary datasets (sample size: 1000) as conditionally dependent or independent.

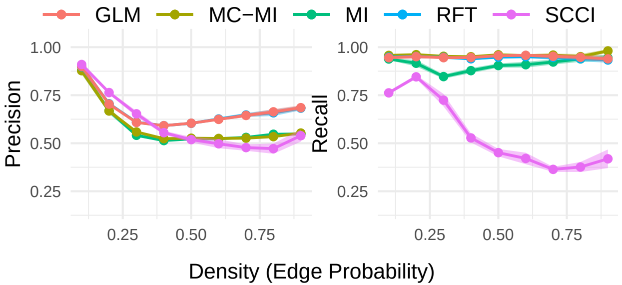
Empirical Analysis: Discrimination (Ordinal)



Accuracy (shading: mean $\pm$ standard error) of classifying simulated ordinal data (8 levels per variable) as conditionally dependent or independent.

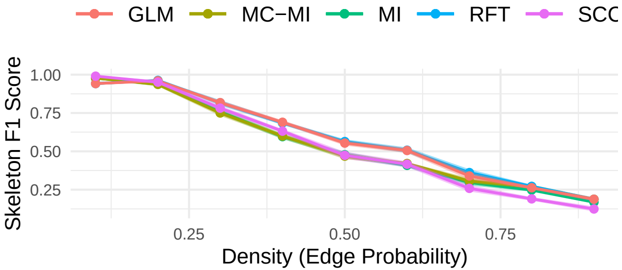
[7] JT = Jonckheere-Terpstra test

Applications: Model testing



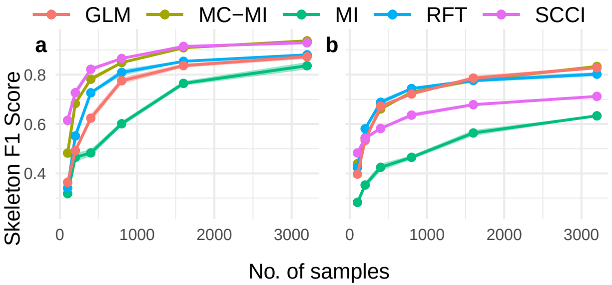
Precision and recall (shading: mean $\pm$ standard error) of testing implied CIs and equal number of randomly generated CIs in binary datasets (sample size: 1000) simulated from random DAGs on 20 variables.

Applications: Structure Learning



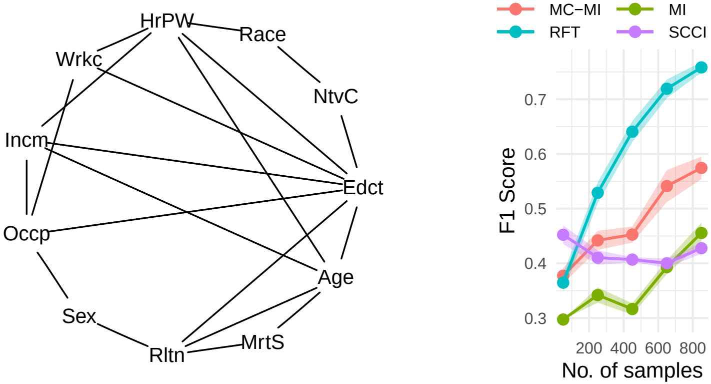
Structure learning on simulated data. F1-score (shading: mean $\pm$ standard error) of the learned model skeletons for randomly generated DAGs with 20 variables and varying edge probabilities.

Applications: Structure Learning



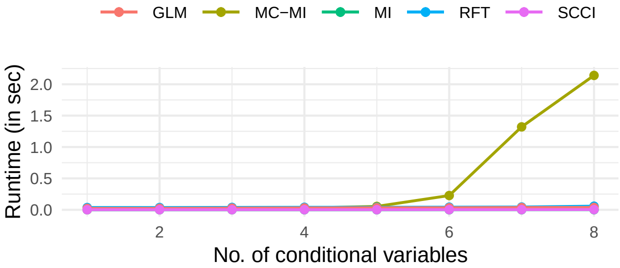
Structure learning on (a) “alarm”, and (b) “insurance” datasets. F1-score (shading: mean $\pm$ standard error, $N = 10$) of the learned model skeletons.

Applications: Structure Learning



Structure learning on US census income dataset. (a) Learnt skeleton using RFT. (b) F1-score (shading: mean $\pm$ standard error, $N = 10$) when comparing d-connected variable pairs from the CPDAG to correlated variable pairs in the dataset.

Runtime Analysis



Runtime (shading: mean $\pm$ standard error, $N = 100$) for CI tests with varying numbers of conditional variables and 1000 samples per dataset.

Conclusion/Future Work

- A residualization based CI test for categorical and ordinal variables.
- Properties: 1) Simple to implement; 2) Interpretable chi-square test statistic; 3) Symmetric by construction; 4) Computationally feasible
- Performs reasonably well for low number of conditional variable but performs better for high number of conditional variables.
- For structure learning, a hybrid approach can be used with other tests.
- Since Random Forests can work with combination of discrete and continuous variables, can possibly be extended to a single unified test.

Questions / Suggestions