A Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test

JFRB 2023 : 11èmes Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes
1-2 juin 2023 Nantes, LS2N, site de Polytech’Nantes (France)

Lei ZAN∗†, Anouar Meynaoui∗, Charles K. Assaad†, Emilie Devijver∗, and Eric Gaussier∗

* Univ. Grenoble Alpes, CNRS, INP, LIG
† EasyVista
Outline

1. Challenges & Objectives
2. CMI Estimator for Mixed Data and Associated Test
3. Limitations and future work
Mixed data occur frequently in many applications, such as health, marketing, medical, and finance. (Ahmad and Dey, 2007; Hennig and Liao, 2013; Morlini and Zani, 2010)

\[ e.g. \]

| Index | message_dispatcher_bolt | metric_bolt | check_message_bolt |
|-------|-------------------------|-------------|-------------------|
| 1     | 0.56                    | 0.51        | Normal            |
| 2     | 0.60                    | 0.53        | Warning           |
| 3     | 0.87                    | 0.52        | Critical          |
| 4     | 1.06                    | 0.51        | Normal            |
| 5     | 0.58                    | 0.54        | Normal            |
| ...   | ...                     | ...         | ...               |
Measuring the (in)dependence between random variables from data when the underlying joint distribution is unknown plays a key role in several settings:

1. Causal discovery (Spirtes et al., 2000)
2. Graphical model inference (Whittaker, 2009)
3. Feature selection (Vinh, Chan, and Bailey, 2014)

Objectives: Estimating and testing conditional independence via Conditional Mutual Information (CMI), from observable mixed data.

Conditional Mutual Information (CMI) has good properties:

\[ I(X, Y|Z) = 0 \Rightarrow X \perp Y|Z \]
\[ I(X, Y|Z) \neq 0 \Rightarrow X \not\perp Y|Z \]
Measuring the (in)dependence between random variables from data when the underlying joint distribution is unknown plays a key role in several settings:

1. Causal discovery (Spirtes et al., 2000)
2. Graphical model inference (Whittaker, 2009)
3. Feature selection (Vinh, Chan, and Bailey, 2014)

Objectives: Estimating and testing conditional independence via Conditional Mutual Information (CMI), from observable mixed data.

Conditional Mutual Information (CMI) has good properties:

\[ I(X, Y|Z) = 0 \Rightarrow X \perp Y|Z \]
\[ I(X, Y|Z) \neq 0 \Rightarrow X \not\perp Y|Z \]
Measuring the (in)dependence between random variables from data when the underlying joint distribution is unknown plays a key role in several settings:

1. Causal discovery (Spirtes et al., 2000)
2. Graphical model inference (Whittaker, 2009)
3. Feature selection (Vinh, Chan, and Bailey, 2014)

Objectives: Estimating and testing conditional independence via Conditional Mutual Information (CMI), from observable mixed data.

Conditional Mutual Information (CMI) has good properties:

\[ I(X, Y|Z) = 0 \Rightarrow X \perp Y|Z \]
\[ I(X, Y|Z) \neq 0 \Rightarrow X \not\perp Y|Z \]
Outline

1 Challenges & Objectives

2 CMI Estimator for Mixed Data and Associated Test

3 Limitations and future work
Consider 3 mixed random vectors $X^{t, \ell}, Y^{t, \ell}$ and $Z^{t, \ell}$. $X^{t, \ell}$ (resp. $Y^{t, \ell}, Z^{t, \ell}$) can be denoted as $(X^t, X^\ell)$, where

- $X^t$ contains all quantitative dimensions of $X^{t, \ell}$
- $X^\ell$ contains all qualitative dimensions of $X^{t, \ell}$

The Conditional Mutual Information $I(X^{t, \ell}; Y^{t, \ell} | Z^{t, \ell})$ is defined as:

$$I(X^{t, \ell}; Y^{t, \ell} | Z^{t, \ell}) = H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell) + H(X^t, Z^t | X^t, Z^t)$$

$$+ H(Y^t, Z^t | Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t | X^\ell, Y^\ell, Z^\ell) - H(Z^t | Z^\ell)$$

$\implies I(X^{t, \ell}; Y^{t, \ell} | Z^{t, \ell})$ is a combination of

- Entropy of qualitative dimensions
- Entropy of quantitative dimensions conditioning on qualitative dimensions
Consider 3 mixed random vectors $X^{t,\ell}, Y^{t,\ell}$ and $Z^{t,\ell}$. $X^{t,\ell}$ (resp. $Y^{t,\ell}, Z^{t,\ell}$) can be denoted as $(X^t, X^\ell)$, where

- $X^t$ contains all quantitative dimensions of $X^{t,\ell}$
- $X^\ell$ contains all qualitative dimensions of $X^{t,\ell}$

The Conditional Mutual Information $I(X^{t,\ell}; Y^{t,\ell} | Z^{t,\ell})$ is defined as:

$$I(X^{t,\ell}; Y^{t,\ell} | Z^{t,\ell}) = H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell) + H(X^t, Z^t | X^t, Z^t)$$
$$+ H(Y^t, Z^t | Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t | X^\ell, Y^\ell, Z^\ell) - H(Z^t | Z^\ell)$$

$\implies I(X^{t,\ell}; Y^{t,\ell} | Z^{t,\ell})$ is a combination of
- Entropy of qualitative dimensions
- Entropy of quantitative dimensions conditioning on qualitative dimensions
Consider 3 mixed random vectors $X^t, Y^t$ and $Z^t$. $X^t$ (resp. $Y^t, Z^t$) can be denoted as $(X^t, X^\ell)$, where
- $X^t$ contains all quantitative dimensions of $X^t, \ell$
- $X^\ell$ contains all qualitative dimensions of $X^t, \ell$

The Conditional Mutual Information $I(X^t, \ell; Y^t, \ell \mid Z^t, \ell)$ is defined as:

$$
I(X^t, \ell; Y^t, \ell \mid Z^t, \ell) = H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell) + H(X^t, Z^t \mid X^t, Z^t)
+ H(Y^t, Z^t \mid Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t \mid X^\ell, Y^\ell, Z^\ell) - H(Z^t \mid Z^\ell)
$$

$\implies I(X^t, \ell; Y^t, \ell \mid Z^t, \ell)$ is a combination of
- Entropy of qualitative dimensions
- Entropy of quantitative dimensions conditioning on qualitative dimensions
Consider 3 mixed random vectors $X^t, Y^t$, and $Z^t$. $X^t$ (resp. $Y^t$, $Z^t$) can be denoted as $(X^t, X^\ell)$, where
- $X^t$ contains all quantitative dimensions of $X^t, \ell$
- $X^\ell$ contains all qualitative dimensions of $X^t, \ell$

The Conditional Mutual Information $I(X^t, \ell; Y^t, \ell | Z^t, \ell)$ is defined as:

$$I(X^t, \ell; Y^t, \ell | Z^t, \ell) = H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell) + H(X^t, Z^t | X^t, Z^t)$$

$$+ H(Y^t, Z^t | Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t | X^\ell, Y^\ell, Z^\ell) - H(Z^t | Z^\ell)$$

$\implies I(X^t, \ell; Y^t, \ell | Z^t, \ell)$ is a combination of
- Entropy of qualitative dimensions
- Entropy of quantitative dimensions conditioning on qualitative dimensions
Entropy of qualitative dimensions can be estimated as:

\[
\hat{H}(X^\ell, Z^\ell) = - \sum_{x^\ell \in \Omega(X^\ell), z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \log \left( \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \right)
\]

Entropy of quantitative dimensions conditioning on qualitative dimensions can be estimated as:

\[
\hat{H}(X^t, Z^t | X^\ell, Z^\ell) = \sum_{x^\ell \in \Omega(X^\ell), z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell)
\]

- \(\hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell)\) is estimated using histogramme.
- \(\hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell)\) is estimated using the nearest neighbors estimator (Kozachenko and Leonenko, 1987):

\[
\hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell) = \psi(n_{xz}) - \psi(k_{xz}) + \log (v_{dxz}) + \frac{d_{xz}}{n_{xz}} \sum_{i=1}^{n_{xz}} \log \xi_{xz}(i)
\]
Hybrid conditional mutual information estimation for mixed data (CMIh)

Entropy of qualitative dimensions can be estimated as:

\[
\hat{H}(X^\ell, Z^\ell) = - \sum_{x^\ell \in \Omega(X^\ell)} \sum_{z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) \log \left( \hat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) \right)
\]

Entropy of quantitative dimensions conditioning on qualitative dimensions can be estimated as:

\[
\hat{H}(X^t, Z^t | X^\ell, Z^\ell) = \sum_{x^\ell \in \Omega(X^\ell)} \sum_{z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) \hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell)
\]

\[\hat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell)\] is estimated using histogram.

\[\hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell)\] is estimated using the nearest neighbors estimator (Kozachenko and Leonenko, 1987):

\[
\hat{H}(X^t, Z^t | X^\ell = x^\ell, Z^\ell = z^\ell) = \psi(n_{xz}) - \psi(k_{xz}) + \log \left( v_{dxz} \right) + \frac{d_{xz}}{n_{xz}} \sum_{i=1}^{n_{xz}} \log \xi_{xz}(i)
\]
Hybrid conditional mutual information estimation for mixed data (CMIh)

- Entropy of qualitative dimensions can be estimated as:

\[
\hat{H}(X^\ell, Z^\ell) = - \sum_{x^\ell \in \Omega(X^\ell), z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \log(\hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell))
\]

- Entropy of quantitative dimensions conditioning on qualitative dimensions can be estimated as:

\[
\hat{H}(X^t, Z^t|X^\ell, Z^\ell) = \sum_{x^\ell \in \Omega(X^\ell), z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell)
\]

- \(\hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell)\) is estimated using histogram.
- \(\hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell)\) is estimated using the nearest neighbors estimator (Kozachenko and Leonenko, 1987):

\[
\hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) = \psi(n_{xz}) - \psi(k_{xz}) + \log(v_{dxz}) + \frac{d_{xz}}{n_{xz}} \sum_{i=1}^{n_{xz}} \log(\xi_{xz}(i))
\]
Hybrid conditional mutual information estimation for mixed data (CMIh)

Entropy of qualitative dimensions can be estimated as:

$$\hat{H}(X^\ell, Z^\ell) = - \sum_{x^\ell \in \Omega(X^\ell)} \sum_{z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \ast \log \left( \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \right)$$

Entropy of quantitative dimensions conditioning on qualitative dimensions can be estimated as:

$$\hat{H}(X^t, Z^t|X^\ell, Z^\ell) = \sum_{x^\ell \in \Omega(X^\ell)} \sum_{z^\ell \in \Omega(Z^\ell)} \hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell) \hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell)$$

- $\hat{P}_{X^\ell, Z^\ell}(x^\ell, z^\ell)$ is estimated using histogram.
- $\hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell)$ is estimated using the nearest neighbors estimator (Kozachenko and Leonenko, 1987):

$$\hat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) = \psi(n_{xz}) - \psi(k_{xz}) + \log \left( v_{dxz} \right) + \frac{d_{xz}}{n_{xz}} \sum_{i=1}^{n_{xz}} \log \xi_{xz}(i)$$
Hybrid conditional mutual information estimation for mixed data (CMIh)

**Experiments of estimator**

- **Configuration of experiments:**

| Scenarios                              | X                        | Y                        | Z                        |
|----------------------------------------|--------------------------|--------------------------|--------------------------|
| Dependence quantitative                | quantitative, Gaussian   | quantitative, Gaussian   | qualitative, Poisson     |
| Dependence mixed                       | qualitative, uniform     | quantitative, uniform    | qualitative, Poisson     |
| Dependence mixed imbalanced            | qualitative, exponential | quantitative, exponential | qualitative, Poisson     |
| Conditional dependence quantitative    | quantitative, Gaussian   | quantitative, Gaussian   | qualitative, Poisson     |
| Conditional dependence mixed           | qualitative, uniform     | quantitative, uniform    | qualitative, Poisson     |
| Conditional dependence mixed imbalanced| qualitative, exponential | quantitative, exponential | qualitative, Poisson     |
| Conditional independence quantitative  | quantitative, Gaussian   | quantitative, Gaussian   | qualitative, Poisson     |
| Conditional independence mixed         | quantitative, uniform     | quantitative, uniform    | qualitative, Poisson     |
| Conditional independence mixed imbalanced | qualitative, exponential | quantitative, exponential | qualitative, Poisson     |
Hybrid conditional mutual information estimation for mixed data (CMIh)

Experiments of estimator

- MSE (on a log-scale) of each method with sample size $n \in \{500, 600, \ldots, 2000\}$ over 100 repetitions.

![MSE plots for various conditions](image)

Conclusions:

- FP performs well in the purely quantitative case;
- MS and RAVK have similar performance in most cases and MS has a main drawback as it gives the value close to 0 in some particular cases;
- LH and CMIh, overall, are more robust than the other ones, but LH is more computation-costly than CMIh.
Hypothesis Null & Hypothesis Alternative:

- $H_0 : X \perp Y \mid Z$
- $H_1 : X \not\perp Y \mid Z$

The main concept of the local permutation test (LocT) can be described as follows for three one-dimensional random variables, namely $X$, $Y$, and $Z$:

1. Estimate the conditional mutual information of the original data as $\hat{I}(X, Y \mid Z)$,
2. Shuffle the value of $X$ within its neighbours that have a similar $Z$ value, resulting in $X_{\pi}$. This permutation ensures that $X_{\pi} \perp Y \mid Z$,
3. Repeat Step 2 $B$ times, and estimate $\hat{I}_i(X_{\pi}, Y \mid Z)$ for each permutation $i \in \{1, \ldots, B\}$,
4. Calculate p-value by using $\hat{I}(X, Y \mid Z)$ and $\{\hat{I}_i(X_{\pi}, Y \mid Z)\}_{i \in \{1, \ldots, B\}}$.

* Intuitive explanations:

- If $X \perp Y \mid Z$, in most cases, $\hat{I}_i(X_{\pi}, Y \mid Z) \approx \hat{I}(X, Y \mid Z)$, where $i \in \{1, \ldots, B\}$.
- If $X \not\perp Y \mid Z$, in most cases, $\hat{I}_i(X_{\pi}, Y \mid Z) \geq \hat{I}(X, Y \mid Z)$, where $i \in \{1, \ldots, B\}$.

Extend local permutation test (LocT) to mixed data, by defining the nearest neighbours should have the same qualitative values in $Z$ and denote it as (LocAT).
**Experiments of independent test**

Here, we propose to analyze 3 structures that are classical:

- **Chain:** $X \rightarrow Z \rightarrow Y$
- **Fork:** $X < Z \rightarrow Y$
- **Collider:** $X \rightarrow Z < Y$

For each structure, we consider the following configurations of experiments:

- **tlt:** $X$ and $Z$ are quantitative, $Y$ is qualitative;
- **ttt:** $X$, $Y$, and $Z$ are quantitative;
- **llt:** $X$ and $Y$ are qualitative, $Z$ is quantitative;
- **ttl:** $X$ is quantitative, $Y$ and $Z$ are qualitative;
- **tll:** $X$ and $Y$ are quantitative, $Z$ is qualitative;
- **lll:** $X$, $Y$, and $Z$ are qualitative.

We use acceptance rate over 10 repetitions of two thresholds (0.01 and 0.05) to show the results:

- The acceptance rate closer to 1 under different threshold the better.
- The number of sampling point is 500.
### Experiments of independent test

|                | CMih-LocT 0.01 | CMih-LocAT 0.01 | CMih-GloT 0.01 | MS-LocT 0.01 | MS-LocAT 0.01 | MS-GloT 0.01 |
|----------------|----------------|-----------------|----------------|--------------|---------------|--------------|
| **Chain**      |                |                 |                |              |               |              |
| \( t\ell t \)  | 1 1            | 1 1             | 1 1            | 1 1          | 1 1           | 1 1          |
| \( ttt \)      | 1 1            | 1 1             | 0 0            | 1 1          | 1 1           | 1 0.9        |
| \( \ell \ell t \) | 1 0.9         | 1 0.9           | 1 0.8          | 1 1          | 1 1           | 1 1          |
| \( t\ell \ell \) | 1 1            | 1 1             | 1 1            | 1 1          | 1 1           | 1 1          |
| \( ttt \ell \) | 0 0            | 0.8 0.4         | 0 0            | 0 0          | 0 0           | 0.5 0.3      |
| \( \ell \ell \ell \) | 1 0.9          | 1 0.9           | 1 1            | 1 1          | 1 1           | 1 1          |
| **Fork**       |                |                 |                |              |               |              |
| \( t\ell t \)  | 0.9 0.9        | 0.9 0.9         | 0 0            | 1 1          | 1 1           | 1 1          |
| \( ttt \)      | 1 1            | 1 1             | 0 0            | 1 1          | 1 1           | 1 1          |
| \( \ell \ell t \) | 1 1            | 1 1             | 1 1            | 1 1          | 1 1           | 1 1          |
| \( t\ell \ell \) | 1 1            | 1 0.9           | 1 1            | 1 1          | 1 1           | 1 1          |
| \( ttt \ell \) | 0 0            | 0.9 0.8         | 0 0            | 0 0          | 0 0           | 0.8 0.5      |
| \( \ell \ell \ell \) | 1 1            | 1 1             | 1 1            | 1 1          | 1 1           | 1 1          |
| **Collider**   |                |                 |                |              |               |              |
| \( t\ell t \)  | 1 1            | 1 1             | 1 1            | 1 1          | 0 0           | 0 0          |
| \( ttt \)      | 1 1            | 1 1             | 0.8 0.9        | 1 1          | 1 1           | 1 1          |
| \( \ell \ell t \) | 1 1            | 1 1             | 1 1            | 0 0          | 0 0           | 0 0          |
| \( t\ell \ell \) | 0 0            | 0.4 0.7         | 0 0            | 0 0          | 0 0           | 0 0          |
| \( ttt \ell \) | 0.6 1          | 1 1             | 0.2 0.4        | 0 0          | 0 0           | 0 0          |
| \( \ell \ell \ell \) | 1 1            | 1 1             | 1 1            | 1 1          | 1 1           | 0.4 0.9      |

### Conclusions:
- CMih with the test LocAT allows one to correctly identify the true (in)dependence relation on all configurations of all structures;
- For other combinations, these is at least one case where it can not work.
Outline

1. Challenges & Objectives

2. CMI Estimator for Mixed Data and Associated Test

3. Limitations and future work
The proposed test may suffer from the problem indicated in \(^1\).
Check the performance of the method on more sophisticated structures.

\(^1\) Shah, Rajen & Peters, Jonas. (2018). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. Annals of Statistics. 48. 10.1214/19-AOS1857.
Limitations and future work

References

Ahmad, Amir and Lipika Dey (2007). “A k-mean clustering algorithm for mixed numeric and categorical data”. In: Data & Knowledge Engineering 63.2, pp. 503–527.

Hennig, Christian and Tim F Liao (2013). “How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification”. In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 62.3, pp. 309–369.

Kozachenko, Lyudmyla F and Nikolai N Leonenko (1987). “Sample estimate of the entropy of a random vector”. In: Problemy Peredachi Informatsii 23.2, pp. 9–16.

Morlini, Isabella and Sergio Zani (2010). “Comparing approaches for clustering mixed mode data: An application in marketing research”. In: Data Analysis and Classification: Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Societ`a Italiana di Statistica. Springer, pp. 49–57.

Spirtes, Peter et al. (2000). Causation, prediction, and search. MIT press.

Vinh, Nguyen, Jeffrey Chan, and James Bailey (2014). “Reconsidering mutual information based feature selection: A statistical significance view”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 28. 1.

Whittaker, Joe (2009). Graphical models in applied multivariate statistics. Wiley Publishing.
Thank you!