A new Mutual Information estimator for continuous censored variables

Ima Bernada*[1], Cécilia Samieri[1], and Gregory Nuel[2]

[1]Bordeaux population health – Université de Bordeaux, Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED), Institut National de la Santé et de la Recherche Médicale – France

[2]Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Universite, Centre National de la Recherche Scientifique, Université Paris Cité, Sorbonne Universite : UMR$_8$001, $CentreNationaldelaRechercheScientifique : UMR_8001, UniversitéParisCité : UMR_8001 - -France$

Résumé

Estimating dependency relationships between variables is an important issue in statistics. Mutual information (MI) is a measure of dependency based on entropy (1,2), and corresponds to the quantity of shared information between two variables. It is largely used in network sciences, as it presents very interesting properties: it can be used without making hypothesis on the underlying distribution of the variables and is applicable on both discrete and continuous variables; it allows capturing both linear and non-linear dependencies, and it equals zero if two variables are independent of each other.
MI estimation methods were primarily developed for exclusively discrete or exclusively continuous variables (3-5), or for mixed-data (datasets that can include both discrete and continuous variables). In practice, complex variables containing both discrete and continuous values (discrete-continuous variables) are often present in real datasets, for example in cohort studies. A common type of discrete-continuous variable is continuous censored. Left-censoring typically occurs with biological measures originating from analytical tools with a lower limit of detection; values which fall below that limit cannot be estimated and are missing at the left side of the distribution. A few methods have been developed to handle specifically discrete-continuous data (6-9), but their effectiveness on the specific case of censored data has not yet been demonstrated.

We propose a new MI estimation method for censored variables, in the form of a correction for other estimators. This new method is based on the decomposition of the MI formula in order to handle the censoring status of the variables on one side, and the simultaneously continuous values of the variables on the other. This second part can be estimated by any MI estimator adapted for continuous data.

We constructed different simulation scenarios of pairs of correlated censored log-normal variables, by varying the censoring, correlation, and size of sample. We evaluated our correction on a few existing estimators previously developed for mixed (10) or discrete-continuous data (6, 7), and on a simpler Gaussian approximation method. We compared the selected estimators, with and without the correction, on these different simulation scenarios.

*Intervenant

We found that MI estimation for all estimators tested is globally closer to the theoretical value when adding the correction. The correction, in the tested simulation cases, enables to reduce bias, and allows convergence towards the true MI value as the number of observations increases.

References:

(1) C. E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J., Oct. 1948.

(2) T. M. Cover and J. A. Thomas. Elements of Information Theory. John wiley edition, 2006.

(3) Y.-I. Moon et al.. Estimation of mutual information using kernel density estimators. Physical Review, Sept. 1995

(4) A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. Physical Review E, June 2004.

(5) C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. BMC Bioinformatics, 5(1):118, Aug. 2004.

(6) A. Marx et al. Estimating Conditional Mutual Information for Discrete-Continuous Mixtures using Multi-Dimensional Adaptive Histograms, Jan. 2021.

(7) O. C. Mesner and C. R. Shalizi. Conditional Mutual Information Estimation for Mixed Discrete and Continuous Variables with Nearest Neighbors, Dec. 2019.

(8) W. Gao, S. Kannan, S. Oh, and P. Viswanath. Estimating Mutual Information for Discrete-

Continuous Mixtures, Oct. 2018.

(9) A. Rahimzamani, H. Asnani, P. Viswanath, and S. Kannan. Estimators for Multivariate Infor-

mation Measures in General Probability Spaces, Oct. 2018.
(10) V. Cabeli et al. Learning clinical networks from medical records based on information estimates in mixed-type data. PLOS Computational Biology, May 2020.