Universal Association Discovery based on Functions of Observation Graph, the good and the ugly

Hesen Peng
Microsoft
6330 NE Radford Dr Apt 3727
Seattle, Washington, USA 98115
hesen.peng@gmail.com

Tianwei Yu
Dept of Biostatistics and Bioinformatics
Emory University
1518 Clifton Rd NE 3F
Atlanta Georgia, USA 30322
tianwei.yu@emory.edu

ABSTRACT

Here we start the Mira paper.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory, Happy

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

In this paper we would like to discuss emerging methods on the discovery of universal probablistic assocation between random vectors. Consider two random vectors X and Y and n pairs of independent and indentically distributed (i.i.d.) random samples $\{X_i, Y_i\}_{i=1}^n$. We would like to draw inference for the existence between X and Y based on the n pairs of samples. Classical association statistics like Pearson's correlation coefficient assume functional forms (linear, monotonicity) between X and Y, which are judged as *uncorrelated* if

$$Corr(X, Y) = 0$$

Universal association statistic perceive associations from the level of probablistic dependence. That is, X and Y are judged as independent if and only if

$$F(X, Y) = F(X)F(Y) \tag{1}$$

where $F(\cdot)$ is the probability density function for the random vector under consideration. Probabblistic association as captured by universal association statistics encapsulates a larger group of associations than traditional correlation coefficient. For example, universal association would consider nonlinear interactions involving multiple variables.

We have noticed that multitude of methods on universal association discovery link to distance functions on the observation graph. The distance graph consists of nodes representing each observation (X_i, Y_i) in the $p+q$ Euclidean space. Here p and q are the dimensions of X and Y, respectively. Edges of the observation graph would connect two nodes (observations) if specific criteria is satisfied.

For example, mutual information and its derivatives have been the most popular universal association statistic to date[4, 6, 9]. To estimate mutual information, the joint entropy can be approximated using K-nearest neighbour [2, 3, 1]. Recent breakthrough on distance covariate [7, 8] sheds light on universal association discovery with its simplicity of form and theoretical flexibility. Brownian distance [7] covariate proposed `dCov` as

$$V_N^2 = \frac{1}{n^2} \sum_{k,l=1}^{n} D_{kl}^X D_{kl}^Y$$

where D_{kl}^X and D_{kl}^Y are simple linear functions of pairwise distances between sample elements calculated on X and Y dimensions, respectively. In an independent research, the author proposed Mira score [5] as

$$M = \sum_{k,l=1}^{n} D_{kl}^{(X,Y)} w_{kl}$$

where $D_{kl}^{(X,Y)}$ is the distance between sample elements calculated using both X and Y dimensions, and $w_{kl} = 1$ when the involved elements are nearest neighbors, $w_{kl} = 0$ otherwise.

In this article we will contribute:

1. Point out that non-trial functions of the observation graph would be capable of discoverying universal association.
2. Present numerical comparison between existing methods.

2. FUNCTIONS ON THE OBSERVATION GRAPH

3. REFERENCES

[1] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy

estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3):277–297, 2005.

[2] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[3] N. Leonenko, L. Pronzato, and V. Savani. A class of rÃl'nyi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, 10 2008.

[4] C. Nardini, L. Wang, H. Peng, L. Benini, and M. Kuo. Mm-correction: Meta-analysis-based multiple hypotheses correction in omic studies. In A. Fred, J. Filipe, and H. Gamboa, editors, *Biomedical Engineering Systems and Technologies*, volume 25 of *Communications in Computer and Information Science*, pages 242–255. Springer Berlin Heidelberg, 2009.

[5] H. Peng. High-dimensional universal dependence discovery. *Emory University Laney Graduate School Dissertation*, 2012.

[6] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[7] G. J. SzÃl'kely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 12 2009.

[8] G. J. SzÃl'kely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 12 2007.

[9] F. Tostevin and P. R. Ten Wolde. Mutual information between input and output trajectories of biochemical networks. *Physical review letters*, 102(21):218101, 2009.