Generalized Distand Association

Hesen Peng   Tianwei Yu

Draft: September 26, 2015

Abstract

1 Motivation

Given two random vectors X and Y, we are interested in testing their probablistic association given n pairs of independent and identically distributed random samples $\{(X_i, Y_i)\}_{i=1}^n$. Peng *et al.* (2015) proposed Mean Distance Association (MeDiA), a set of probablistic association statistics as functions of observation distances. The theoretical foundation of MeDiA relies upon the result below:

Theorem 1. *(from Peng* et al. *(2015)) Denote the distance between two independent random samples from (X, Y) as d_{XY}, and the distance between two dependent random samples from $(\hat{X}, \hat{Y})$ as $d_{\hat{X}\hat{Y}}$. Then we have*

$$E(d_{XY}) \geq E(d_{\hat{X}\hat{Y}})$$

In this paper, we would like to expand the theory above to general functions on the observation graph. The generalized mean distance would encompass a number of existing methods, like mutual information. Besides, the generalized mean distance naturally leads to the construction of several other probabilistic association statistics.

Theorem 2. *(univariate g-transformation) Using the same notation, denote a monotonically increasing continuously differentiable function $g(\cdot)$. Denote the g-transformed distance as*

$$\begin{aligned}\tilde{d}_{XY} &= g(d_{XY})\\ \tilde{d}_{\hat{X}\hat{Y}} &= g(d_{\hat{X}\hat{Y}})\end{aligned}$$

Then we have:

$$E(\tilde{d}_{XY}) \geq E(\tilde{d}_{\hat{X}\hat{Y}})$$

and the average of the transformed distances $\tilde{d}$ follow asymptotic normal distribution.

Following Theorem 2, we can see that distance based mutual information statistic $MI = \sum \log(d_{ij})$ actually falls into the generalized mean distance family.

However, it would be helpful to realize that Theorem 2 has not yet encompass functions on the observation graph that give different weigths depending on the value. We would make this up with the results below:

Theorem 3. *(Multivariate f-transformation) Using the same notation as above, n-variate function f is monotonically increasing on every dimension of input. Define*

$$\begin{aligned} \bar{d}_{XY} &= f(d_{i1}^{XY}, \ldots, d_{in}^{XY}) \\ \bar{d}_{\hat{X}\hat{Y}} &= f(d_{i1}^{\hat{X}\hat{Y}}, \ldots, d_{in}^{\hat{X}\hat{Y}}) \end{aligned}$$

Then we have:

$$E(\bar{d}_{XY}) \geq E(\bar{d}_{\hat{X}\hat{Y}})$$

Theorem 3 shows that k-nearest neighbour edge sum as defined in Mira score, and k-nearest neighbour log edge sum as defined in Mutual Information, also falls into this category and can be used to identify random vector associations.

2 Test of Probabilistic Association

2.1 Numerical Comparison

3 Applications

4 Discussions

A Appendix

A.1 Proof of Theorem 2

Proof. The proof follows directly from delta method. More specifically, for given $d_{\hat{X}\hat{Y}}$ and d_{XY}, there exists d'_{XY}, such that:

$$\tilde{d}_{\hat{X}\hat{Y}} = g(d_{\hat{X}\hat{Y}})$$

$$
\begin{aligned}
&= g\left[(d_{\hat{X}\hat{Y}} - d_{XY}) + d_{XY}\right] \\
&= g(d_{XY}) + g'(d'_{XY})(d_{\hat{X}\hat{Y}} - d_{XY}) \\
&= \tilde{d}_{XY} + g'(d'_{XY})(d_{\hat{X}\hat{Y}} - d_{XY})
\end{aligned}
$$

Following Theorem 1, taking expectation on both sides, and realizing that $g'(\cdot) \geq 0$, we conclude the proof. $\square$

Proof of Theorem 3

Proof. When f is monotonically increasing and continuously differentiable, the proof to Theorem 3 is straight forward and similar to the proof to Theorem 2 usin delta method.

In addition, when f is monotonically increasing but not continously differentiable, there exists a sequence of monotonically increasing and continuously differentiable functions $\{f_i(\cdot)\}_{i=1}^{+\infty}$, such that

$$\lim_{i\to\infty} ||f_i - f||_{d_{XY}} \to 0 \tag{1}$$

Define

$$
\begin{aligned}
\bar{d}^i_{XY} &= f_i(d^{XY}_{i1}, \ldots, d^{XY}_{in}) \\
\bar{d}^i_{\hat{X}\hat{Y}} &= f_i(d^{\hat{X}\hat{Y}}_{i1}, \ldots, d^{\hat{X}\hat{Y}}_{in})
\end{aligned}
$$

Then for each i, we have:

$$E(\bar{d}^i_{XY}) \geq E(\bar{d}^i_{\hat{X}\hat{Y}})$$

Summing the results above, we have

$$E(\bar{d}_{XY}) \geq E(\bar{d}_{\hat{X}\hat{Y}}) \tag{2}$$

$\square$

References

Peng, Hesen, Ma, Junjie, Bai, Yun, Lu, Jianwei, & Yu, Tianwei. 2015. Media: Mean distance association and its applications in nonlinear gene set analysis. *Plos one*, **10**(4), e0124620.