# Generalized Distand Association

Hesen Peng Tianwei Yu

Draft: May 25, 2015

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 $\{(X_i, Y_i)\}_{i=1}^n$. [1] 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 [1]) Denote the distance between two independent random samples from $(X, Y)$ as $d_{XY}$, and the distance between two independent random samples from $(\hat{X}, \hat{Y})$ as $d_{\hat{X}\hat{Y}}$. Then we have*

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

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:*

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

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

*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:

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

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

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 continuous and monotonically increasing on every dimension of input. Define*

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

*Then we have:*

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

# References

[1] Hesen Peng, Junjie Ma, Yun Bai, Jianwei Lu, and Tianwei Yu. Media: Mean distance association and its applications in nonlinear gene set analysis. *PLoS ONE*, 10(4):e0124620, 04 2015.