Probablistic Association Discovery using Functions of Observation Graph

Hesen Peng
Independent
136 102nd Ave SE Apt 326
Bellevue, Washington, USA 98004
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

Probablistic association discovery aims to identify the association between random vectors, regardless of number of variables involved or linear/nonlinear functional forms. Application to high-dimensional data analysis has generated rising interest in probablistic association discovery.

We developed a framework for probablistic association discovery on the Euclidean space using functions on the observation graph. We first discuss the property of mean observation distance and its novel application to association discovery. Then we generalize the advantageous property to a group of functions on the observation graph. The group of functions encapsulates major existing methods in association discovery, like mutual information and Mira score, and can be expanded to more complicated forms. We conducted numerical comparison that shows differentiated testing power of related methods under multiple scenarios.

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 propose and generalize new methods to discover probablistic association 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 of probablistic association between X and Y based on the n pairs of samples. The discussion in this paper will focus on the probablistic association between continuous random variables defined in the Euclidean space. However, general theory developed in this paper can be applied in other spaces as well.

Classical association statistics like Pearson's correlation coefficient assume functional forms (for example, piecewise linear, monotonicity) between X and Y, which are judged as *correlated* if

$$Corr(X, Y) \neq 0$$

Probablistic association statistic, as the name suggests, perceives associations from the level of probablistic dependence. That is, X and Y are judged as independent if and only if their joint probability density function can be factored

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

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

We have noticed that multiple methods on probablistic association discovery are linked to functions on the observation distance 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 probablistic association statistic to date[4, 6, 9]. To estimate mutual information, the joint entropy can be approximated using log-transformed K-nearest neighbour distance averaged for each observation [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 linear functions of pairwise distances between sample elements calculated with X and Y dimensions, respectively. Given fixed marginal distribution for X

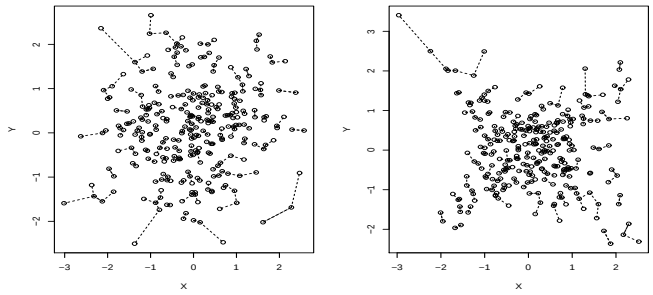

Figure 1: Random samples generated from independent bivariate normal distribution (left), and mixture bivariate normal distribution with ± 0.8 covariates (right). The dashed lines connects two observations if they are nearest neighbours.

and Y, large Brownian distance covariate suggests the existence of probablistic association.

In an independent research, the authors 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. Given fixed marginal distribution for X and Y, small Mira score suggests the existence of probablistic association.

We would like to discuss the property of functions applied to the distances of the observation graph. The discussion of the section below is inspired by comparing mean observation distance between dependent and independent random vectors of the same marginal distribution. Our intution suggest that when two random vectors are probablistically associated, their observation distances tend to be smaller than their independent counterparts.

Take Figure 1 for example, both plots show 300 samples from two bivariate random vectors. Observations on the left panel are sampled from independent bivariate normal distributions. Observations on the right panel are sampled from mixture bivariate normal distribution with ± 0.8 covariates. Coincidentally, both distributions have standard normal mariginal distribution and zero coerrelation coefficient. However, the two samples differ on a group of metrics defined on the observation distances. Table 1 shows the mean distance, mean nearest neighbour distance, and mean log-nearest neighbour distance are all smaller for the dependent case compared with independent case. We have repeated the simulation multiple times and the same trend is observed in large probability.

The above observation is no coincidence. In this article we will generalize the functions on the observation graph, discuss their properties in identifying probablistic associations. More specifically, we will contribute:

1. Point out that non-trial monotonic functions of the observation graph would be capable of discovering universal association.

2. Present numerical comparison between existing methods.

Metric	Left (Ind)	Right (Mix. Normal)
mean distance	1.81	1.70
mean NN	0.14	0.12
mean log(NN)	-2.24	-2.43

Table 1: Comparison between the independent bivariate normal distribution and mixture normal distribution in Figure 1. Statistics used in the comparison include mean observation distance, mean nearest neighbour distance, mean log-nearest neighbour distance.

For illustration purpose we applied the proposed methods to the association identication in image analysis and text mining. The results are very interesting.

2. THEORY

We are interested in testing the existence of probablistic association between two random vectors (X, Y), given n pairs of observations. Consider another pair of random vectors $(\hat{X}, \hat{Y})$, where $\hat{X}$ follows independent and identical distribution (*i.i.d.*) as X and $\hat{Y}$ follows *i.i.d.* as Y. The only difference is that $\hat{X}$ and $\hat{Y}$ are mutually independent. As mentioned above, we would like to compare the sample observation distance from (X, Y) against that from $(\hat{X}, \hat{Y})$.

Theorem 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*

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

For the sake of space, proofs of the theoretical results in this section are presented in Appendix A. Theorem 1 above confirms our earlier intuition: when two random vectors are probablistically associated, their observations tend to be closer compared with their independent counterparts.

Based on Theorem 1, we can use permutation tests to investigate the existence of probabilistic association given n pairs of *i.i.d.* random samples. The mean sample distance is compared against a null distribution to generate p-value. The null distribution is generated by permuting the relative index of observations from X and Y. Denote distances between two random observations as d_{ij}, where i and j are the indeces among the n observations. To our advantage, we have the following property:

Corollary 1. *For a given observation i, define its mean peer distance as Equation 2. Also define the mean observation distance for n observations as Equation 3:*

$$\bar{d}_i = \frac{1}{n-1} \sum_{i \neq j} d_{ij} \quad \forall i \tag{2}$$

$$\bar{d} = \frac{1}{n} \sum \bar{d}_i \tag{3}$$

Under the null hypothesis that random vectors X and Y are independent, the mean observation distance $\bar{d}$ follows asymptotic normal distribution with $n \to \infty$.

Corollary 1 is easily proved using central limit theorem. Based on Corollary 1, we can approximate the null distribu-

tion of mean distance using normal distribution, which dramatically alleviate computational burden given large sample size.

Finally, functional transformations might be applied to the observation distances for each observation. We have

Corollary 2. *Under regulatory conditions, denote a monotonically increasing function $g(\cdot)$ applied to peer distances for each observations.*

$$\begin{aligned} \tilde{d}_i &= g(d_{i1}, \ldots, d_{in}) \\ \tilde{d} &= \frac{1}{n}\sum \tilde{d}_i \end{aligned} \tag{4}$$

The g-transformed mean observation distances still enjoy all properties above, more specifically:

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

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

From here we can see that when g is the minimum function, Mira score, calculated as the mean nearest neighbour observation distance, is a special case of $\tilde{d}$. When g is generating the mean of the smallest K elements, $\tilde{d}$ is equivalent to mean K-nearest neighbour edge sum. When g is generating the mean of log-transformed smallest K elements, $\tilde{d}$ is equivalent to linear transformation of entropy estimates, which translates to association discovery using mutual information.

3. ASSOCIATION DETECTION

Following theoretical results from the section above, we propose permutation test of probablistic association using mean observation distance and its transformations. Given n pairs of *i.i.d.* observations from random vectors (X, Y), the test statistic, mean g-transformed observation distance, is calculated using Equation **??**. Null distribution of the test statistic is generated with the following procedure:

1. Permute relative indeces of samples $\{X_i, Y_i\}_{i=1}^n$ and calculate mean g-transformed observation distance after permutation.
2. Repeat the above step R times and record all mean g-transformed observation distances, denoted as $\{\bar{d}_i\}$.
3. Calculate mean and and standard deviation of $\{\hat{d}_i\}$, denoted as $(\bar{\mu}, \bar{\sigma})$.
4. Approximate the null distribution using normal distribution with mean and standard deviation equaling to $(\bar{\mu}, \bar{\sigma})$.
5. Compare $\bar{d}$ with the approximated null distribution, and generate p-value of the test.

When g is identity function, test statistic above equals to the mean observation distance.

3.1 Numerical Comparison

The permutation and normal approximation procedure above exemplifies the general probablistic association testing process that most existing methods use, as summarized in Table 2. All existing methods of probablistic association identification would utilize observation distances and generate test statistics from it. The power of the above methods are compared using simulation under scenarios below:

Name	Statistic	Perm-Test
This paper	Mean distance	Yes
Mira score [5]	Mean NN distance	Yes
Mutual Info [2, 3, 1]	Mean log-NN distance	Yes
Brownian Cov [7, 8]	Distance covariate	Yes

Table 2: Summary of existing methods on probablistic association discovery. Perm-test indicates if the method utilizes permutation test to generate null distribution for hypothesis testing. (NN is short for Nearest Neighbour)

- **Linear association**: X and Y are 5-dimensional random vectors. (X, Y) follow multivariate normal distribution with zero mean, unit variance and $Cov(X_i, Y_j) = 0.1$ for all (i, j).
- **Variance association**: X follows 5-dimensional normal distribution with zero mean, zero covariance, and unit variance. Y follows 5-dimensional normal distribution with zero mean, zero covariance, and $|X|$ variance. In this case Y maintains zero mean regardless of X.
- **Triple association**: Marginally, X, Y, Z are 5-dimensional random vectors following normal distribution with zero mean, zero covariance, and unit variance. However, jointly we have

  $$\text{sign}(Z) = \text{sign}(XY)$$

  In this case (X, Y, Z) are pairwise independent but jointly dependent. We would like to test the association between X and (Y, Z) given n pairs of random samples.

For each of the above scenarios, we generated n pairs of random samples. We tested the existence of association using methods listed in Table 2. The sample size n ranged from 40 to 1000 with steps of 20. For each scenario/method/sample size tuple, we generated 1000 simulations. The p-value for each simulation is recorded. And finally the power for each method under each scenario and sample size combination is calculated as the percentage of tests with p-values smaller than 0.05.

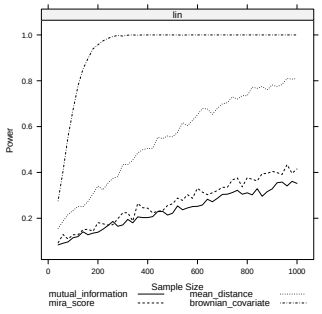

Figure 2: Power comparison for the linear association scenario.

Power comparison shows differentiated method performance in different scenarios. In the linear association scenario, the

Brownian Covariate outperformed all other methods (Figure 2). The superior performance comes possibly from the tight connection between pearson correlation and Brownian Covariate under multivariate normal distributions as illustrated in [7].

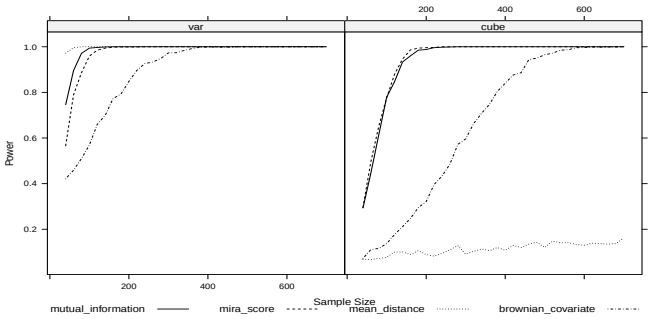

Figure 3: Power comparison for the nonlinear association scenario: variance association (left) and triple association (right).

On the other hand, mutual information and Mira score both outperforman Brownian Covariate in the two nonlinear scenarios (Figure 3). Mira score is estimated as average nearest neighour observation graph edge length. Mutual information is estimated as average log-transformed nearest neighbour edge length. Mira score and mutual information share similar power possibly due to the similar forms of estimation. Observations in the variance association and triple association case are "cornered" into half of the distribution space. Thus association metric concerned about the nearest neighbour distance can quickly detect the change of distribution space. In contrary, Brownian Covariate takes into account the distributions of all observations. Even when observations space is squeezed into half of the independent one because of nonlinear associations, long distances between observations still exists. This might have slowed down the detection of nonlinear association by Brownian Covariate.

4. APPLICATIONS

5. DISCUSSION

In this paper we have discussed the general theory of association discovery using functions on the observation graph. Statistics of similar form to Equation 4 are capable of detecting associations between continuous random vectors using permutation test of association. However, we would like to point out that Equation 4 is not the only way to test probablistic association. For example, Brownian distance covariate (dCov) utilizes the covariates between observation distance calculated using either random vectors under consideration for statistic. This is different from Equation 4. We are confident that there are way more methods to test probablistic associations to be discovered on the theoretical front.

In Section 2 we have generalized the estimation of mean observation distance, Mira score, and mutual information estimate into the same framework of functions on the observation graph. Simulation study in Section 3 shows that the three statistics have differentiated performance in terms

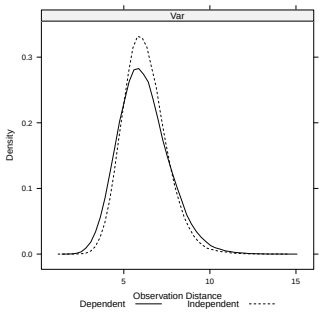

Figure 4: Density plot of observation distances for variance association (solid line) between two random vectors with $p = 10$ and their independent counterparts (dashed line).

of test power under different scenarios. In hindsight, we realized that: testing of probablistic association using observation distance under framework of Equation **??** actually rests on the testing observation distance distributions. More specifically, when random vectors X and Y under consideration are associated, their distribution of observation distance should be different from their independent counterpart $\hat{X}$ and $\hat{Y}$.

For illustrations, consider the following situation, 500 random *i.i.d.* random samples are generated from variance association distribution with $p = 10$. For comparison, 500 random samples are generated from independent random variable pairs of the same marginal distribution. Observation distances are calculated for each case, respectively. Density plot of observation distances for each case are plotted in Figure 4. We observe that the distribution of observation distance for samples from variance association distribution are visually different from the null distribution. In our upcoming work, we are proposing an omnibus probablistic association test based on the observation distance distribution. We would expect this test to be even more flexible compared with the methods compared in this paper.

APPENDIX

A. PROOF OF THEOREMS

Theorem 1. We would like to make the following definitions before proceeding:

- (X', Y'): follows *i.i.d.* as (X, Y)
- $(\hat{X}, \hat{Y})$: $\hat{X}$ follows *i.i.d.* as X; $\hat{Y}$ follows *i.i.d.* as Y. But $\hat{X}$ and $\hat{Y}$ are independent. $(\hat{X}, \hat{Y})$ is independent of all previously defined random variables.
- $(\hat{X}', \hat{Y}')$: follows *i.i.d.* as $(\hat{X}, \hat{Y})$ and is independent of all the above random variables.

Based on the abvoe definition, the distance between independent random pairs of observations from (X, Y) and its independent counterpart $(\hat{X}, \hat{Y})$ can be expressed as:

$$\begin{aligned} E(d_{XY}) &= E(|(X, Y) - (X', Y')|_{p+q}) \\ E(d_{\hat{X}\hat{Y}}) &= E(|(\hat{X}, \hat{Y}) - (\hat{X}', \hat{Y}')|_{p+q}) \end{aligned}$$

Further more, based on Equation 2.5 of [8], we have:

$$E(|(X,Y)-(X',Y')|_{p+q}) = \int_{R^{p+q}} \frac{1-|f_{X,Y}(s,t)|^2}{C_{p+q}|(t,s)|_{p+q}^{1+p+q}} d(s,t)$$

$$E(|(\hat{X},\hat{Y})-(\hat{X}',\hat{Y}')|_{p+q}) = \int_{R^{p+q}} \frac{1-|f_{\hat{X},\hat{Y}}(s,t)|^2}{C_{p+q}|(t,s)|_{p+q}^{1+p+q}} d(s,t)$$

where $f_{X,Y}(s,t)$ is the characteristics functions of (X,Y); $f_{\hat{X},\hat{Y}}(s,t)$ is the characteristics functions of $(\hat{X},\hat{Y})$; C_{p+q} is a constant defined in Equation 2.3 of [8].

Meanwhile, the characteristics function of $(\hat{X},\hat{Y})$ can be factored because of their independence

$$f_{\hat{X},\hat{Y}}(s,t) = f_{\hat{X}}(s) f_{\hat{Y}}(t)$$

Summarizing the above, we have

$$\begin{aligned} E(d_{XY} - d_{\hat{X}\hat{Y}}) &= \int_{R^{p+q}} \frac{|f_X(s)|^2|f_Y(t)|^2 - |f_{XY}(s,t)|^2}{C_{p+q}|(t,s)|_{p+q}^{1+p+q}} d(t,s) \\ &\leq 0 \end{aligned}$$

according to Cauchy-Schwarz inequality. □

B. 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.