International Journal of Business Analytics and Intelligence
7 (2) 2019, 01-03
http://publishingindia.com/ijbai/

Submitted: 05 August, 2019
Accepted: 11 September, 2019

Analytically Yours:

Measures of Association Among Several Variables

Arnab Kumar Laha*

In the current times, humongous amount of data is being collected from a variety of sources and are being put to use for various business applications. Such data is currently being referred to as Big Data. Not only are the size of these datasets huge but often they also have information about a large number of variables which may be dependent on one another. This leads us to the question of understanding the extent of association among these variables. When dealing with only two variables, in elementary statistics courses, one discusses a measure of association called the (Pearson product-moment) correlation coefficient (ρ). When two random variables X and Y have a joint distribution then $\rho_{X,Y}$ is computed as follows: $\rho_{X,Y} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$ where, $Cov(X,Y) = E((X - E(X))(Y - E(Y))$ is the covariance between X and Y, $SD(X)$ and $SD(Y)$ are the standard deviations of the random variables X and Y respectively. This simple measure of association takes value between –1 and 1 where the extreme values $\rho_{X,Y} = 1$ or $\rho_{X,Y} = -1$ indicates perfect linear dependence between the two variables i.e., $Y = \alpha + \beta X$ where α, β are two constants. When $\rho_{X,Y} = 0$ we say that the variables X and Y are uncorrelated i.e. they are not linearly related. It may be noted that X and Y may be non-linearly related even if $\rho_{X,Y} = 0$. To see this, consider the case that X is symmetrically distributed about 0 as a result of which $E(X) = E(X^3) = 0$. Let $Y = X^2$. A simple computation shows that $Cov(X,Y) = E(X^3) - E(X)E(X^2) = 0$ implying $\rho_{X,Y} = 0$. When the random variables X and Y are independent we have $\rho_{X,Y} = 0$ but the converse is not true as can be seen from the example given above. The sample correlation coefficient is denoted by $\rho_{X,Y}$.

It is therefore natural to seek measures of association akin to the correlation coefficient when more than two variables are involved. Suppose $\boldsymbol{X} = (X_1,\ldots, X_p)'$ is a $p \times 1$ random vector and $\boldsymbol{Y} = (Y_1,\ldots,Y_q)'$ is a $q \times 1$ random vector and we are interested in measuring the association between X and Y. In this context, Hotelling (1936) introduced the concept of Canonical Correlation. Consider the linear combination $\boldsymbol{Z}_X = \boldsymbol{u}'\boldsymbol{X} = u_1X_1 + \cdots + u_pX_p$ of the variables $X_1,\ldots,X_p$ and the linear combination $Z_Y = \boldsymbol{v}'\boldsymbol{Y} = v_1Y_1 + \cdots + v_qY_q$ of the variables $Y_1,\ldots, Y_q$. To obtain the first canonical correlation Hotelling suggests choosing the vectors u and v so that $\rho Z_X, Z_Y$ is maximised while ensuring variance of Z_X and Z_Y are both equal to 1. The maximum value $\rho Z_X, Z_Y$ so obtained is called the first (or maximum) canonical correlation (ρ_1) between the random vectors X and Y and the linear combinations $Z_X^{(1)} = u'_1X$ and $Z_Y^{(1)} = v'_1X$ for which the maximum value of $\rho Z_X, Z_Y$ is obtained are called the first canonical variates. It is possible to obtain upto $t = \min(p, q)$ canonical correlations $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_t \geq 0$ by repeatedly following a procedure similar to above with the additional constraint that the canonical variates to be obtained in the current step are uncorrelated with those that had been obtained in the earlier steps. The details are omitted here and the interested reader may look at Chapter 10 of Eaton (2007). A measure of affine dependence based on the canonical correlations is $\phi = \sum_{i=1}^{t}\left(1 - \rho_i^2\right)$. When $\phi = 0$ and $p \leq q$ it can be shown that X and Y are related through an affine transformation i.e. $X_{p\times 1} = C_{p\times q}Y_{q\times 1} + d_{p\times 1}$ for some $p \times q$ matrix C and $p \times 1$ vector d. Since $0 \leq \phi \leq 1$, one may consider $1 - \frac{\phi}{t}$ which satisfies $0 \leq \phi \leq 1$ with $\phi = 1$ giving

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

the case of affine dependence between the two random vectors.

How does one estimate the canonical correlations from given data? Suppose (x_i, y_i) for $1 \le i \le n$ be the given dataset of n observations. It may be noted that $x_i = (x_{i1},\ldots, x_{ip})'$ and $y_i = (y_{i1},\ldots, y_{iq})'$. We define the following:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_{XX} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})',$$

$$S_{YY} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})' \text{ and}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})' = S'_{XY}$$

Let the k-th sample canonical correlation be denoted by ρ_k. Then ρ_k^2 is the k-th eigenvalue of $S_{XX}^{-1/2} S_{XY} S_{YY}^{-1} S_{YX} S_{XX}^{-1/2}$ (or equivalently that of $S_Y Y^{-1/2} S_Y X S_X X^{-1} S_X Y S_Y Y^{-1/2}$). Another coefficient that has fairly wide usage for measuring association between two sets of variables is the RV coefficient which is a consistent estimator of the coefficient ρ_V which is discussed below. Let us denote by Σ_{XY} the $p \times q$ covariance matrix between the random vectors X and Y i.e. the (i, j)-th entry in this matrix is the $\mathrm{Cov}(X_i, Y_j)$. We analogously define $q \times p$ matrix Σ_{YX}, $p \times p$ matrix Σ_{XX} and the $q \times q$ matrix Σ_{YY}. Escoufier (1973) defined the coefficient $\rho V(\boldsymbol{X}, \boldsymbol{Y}) = \dfrac{tr(\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YX})}{\sqrt{tr(\boldsymbol{\Sigma}_{XX}^2)\, tr(\boldsymbol{\Sigma}_{YY}^2)}}$ where for any square matrix A, $tr(A)$ denotes the trace of the matrix A (i.e. the sum of the diagonal elements of the matrix A). It is easy to see that when $p = q = 1$, $\rho V = \rho_{X,Y}^2$. Further, $0 \le \rho V(X, Y) \le 1$ and $\rho V(X, Y) = 0$ if and only if $\Sigma_{YX} = 0$. Moreover, $\rho V(X, aBX + c) = 1$ i.e. when X and Y are related through an affine transformation then $\rho V(X, Y) = 1$.

The RV-coefficient is defined as $RV(X,Y) = \dfrac{tr(S_{XY}S_{YX})}{\sqrt{tr(S_{XX}^2)\, tr(S_{YY}^2)}}$. This expression of the RV-coefficient can be expressed in several alternative ways. We note one of these here. Let Δ_X be a $n \times n$ matrix whose (i, j)-th element (d_{ij}^X) is the Euclidean distance between x_i and x_j i.e. $d_{ij}^X = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$ and Δ_Y be another $n \times n$ matrix whose (i, j)-th element (d_{ij}^Y) is the Euclidean distance between y_i and y_j i.e. $d_{ij}^Y = \sqrt{\sum_{k=1}^{q}(y_{ik} - y_{jk})^2}$. Further let $C_{n\times n} = I_n - \frac{1}{n} 1_n 1_n'$ where I_n is the $n \times n$ identity matrix and 1_n is the $n \times 1$ vector of all 1's. The RV-coefficient can now be expressed as:

$$RV(X,Y) = \frac{\langle C\Delta_X^2 C, C\Delta_Y^2 C\rangle}{\|C\Delta_X^2 C\|\,\|C\Delta_Y^2 C\|}$$

where for any two $n \times n$ real matrices A, B the Hilbert-Schmidt inner product $\langle A, B\rangle$ is defined as $\langle A, B\rangle = \mathrm{tr}(A'B)$ and

$$\|A\| = \sqrt{\langle A, A\rangle}$$

The third coefficient that we discuss in this article is the dCor coefficient which is based on a measure of dependence between random vectors introduced by Szekely, Rizzo and Bakirov (2007) called the distance covariance or dCov. A very important property of dCov is that $dCov = 0$ if and only if there is independence between the random vectors X and Y. The dCov coefficient is defined as the distance between the joint and the product of the marginal characteristic functions of the random vectors in a weighted L^2 sense that ensures that zero-independence property. The dCov coefficient is then scaled to obtain an association measure $R(X, Y)$ whose sample estimate is the dCor coefficient which is defined as:

$$dCor^2(X,Y) = \frac{\langle C\Delta_X C, C\Delta_Y C\rangle}{\|C\Delta_X C\|\,\|C\Delta_Y C\|}$$

Table 1: Systolic and Diastolic Blood Pressure Readings Taken in the Left Arm (SysLH and DiasLH) and the Right Arm (SysRH and DiasRH) Measured (Nearly) Simultaneously for a Person on 19 Days

LH-RH Blood Pressure			
SysLH	*DiasLH*	*SysRH*	*DiasRH*
123	85	125	83
112	76	112	76
107	75	115	75
108	72	111	72
110	79	115	79
119	85	125	79
112	78	124	78
117	77	113	79
117	83	110	81
112	80	118	77
106	75	113	72
105	74	112	77
114	81	114	82
107	77	112	72
103	73	119	76
107	71	107	73
112	77	118	84
119	81	129	84

It can be shown that when $p = q = 1$, then $dCor(X, Y) \leq |r_{X,Y}|$. Further $0 \leq dCor(X, Y) \leq 1$ and $dCor(X, aBX + c) = 1$ i.e. when X and Y are related through an affine transformation then $dCor(X,Y) = 1$

In Table 1 above we give Systolic and Diastolic blood pressure readings of an individual measured (nearly) simultaneously on the left arm and right arm of the same individual on 19 days. It is expected that the readings in the left arm would be associated with that in the right arm. Table 2 below gives the R-code and output for computing the three measures discussed above. It can be seen that the measures give quite different values for the same dataset. The two canonical correlations are QUOTE $\rho_1 = 0.824$ and QUOTE $\rho_2 = 0.199$ yielding the measure of affine dependence QUOTE $\phi = 0.359$. The RV-coefficient is 0.999 while the dCor coefficient is 0.777.

Table 2: R-code and Output for Computing Canonical Correlation, RV-coefficient and dCor.

```
#Canonical Correlations
cc=cancor(LH,RH)
cc$cor
[1] 0.8239106 0.1994156

#RV-coefficient
library(MatrixCorrelation)
RV(LH,RH)
[1] 0.9987494

#dCor coefficient
library(energy)
dcor(LH,RH)
[1] 0.7774218
```

The RV-coefficient is 0.999 while the dCor coefficient is 0.777.

The interested reader looking for further information about measures of multivariate association may refer to the survey article by Josse and Holmes (2016).

References

Eaton, M. L. (2007). *Multivariate statistics: A vector space approach, 53*, IMS Lecture Notes Monograph Series.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics, 29,* 751–760.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28,* 321–377.

Josse, J., & Holmes, S. (2016). Measuring multivariate association and beyond. *Statistics Surveys, 10,* 132-167.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics, 35*(6), 2769–2794.