Detecting Novel Associations in Large Astrophysical Data Sets

Elizabeth Martínez-Gómez[*] Mercedes T. Richards[†]
Donald St. P. Richards[‡]

Abstract
The distance correlation as a measure of dependence between collections of random variables was introduced by Székely, Rizzo, and Bakirov (2007) and Székely and Rizzo (2009). Unlike the classical Pearson correlation coefficient, the distance correlation is zero only in the case of independence. Moreover, the distance correlation applies to random vectors of any dimension, rather than to two-dimensional variables only, and it is now known to be capable of detecting nonlinear associations that are not detectable by the Pearson correlation coefficient. We apply the distance correlation to analyze high-dimensional, large-sample astrophysical databases on galaxy clusters, and we identify new associations and correlations between numerous astrophysical variables. For certain pairs of variables, we find that it is also possible to estimate the corresponding Pearson correlation coefficients from the distance correlation measures, with high accuracy. Indeed, the distance correlation has a clear tendency to resolve some high-dimensional data into highly concentrated "horseshoe" graphs, which make it easier to identify patterns in the data. For comparison we also compute the Maximal Information Coefficient (MIC score) and we conclude that Distance Correlation is more general and more powerful than the Pearson and MIC measures of dependence.

Key Words: Association, large astrophysical data sets, distance correlation, mutual information, Pearson correlation coefficient

1. Introduction

The Pearson correlation coefficient is the classical measure of (mainly linear) dependence between two variables (Pearson 1895). Because of its deficiency in detecting nonlinear relationships and the coefficient can easily be zero for dependent variables, Székely, Rizzo and Bakirov (2007) introduced a new measure named *Distance correlation* and recently Reshef et al. (2011) proposed a new measure of association between variables based on Shannon's mutual information (Shannon and Weaver 1949), the Maximal Information Coefficient (MIC).

1.1 Measures to detect associations between variables

To understand the measures of association between variables it is necessary to refer to the concept of statistical independence. Events (or measurements) are termed probabilistically independent if information about some does not change the probabilities of the others. By convention, any measure of association between two variables must be zero if the variables are independent. Those are also known as *measures of dependence*. There are other requirements of a good measure of dependence, including symmetry (Rényi 1959), and since Galton's correlation coefficient (1886) the statisticians have been defining suitable measures, including rank

[*] Department of Statistics, Instituto Tecnológico Autónomo de México, Río Hondo 1, Col. Tizapán San Ángel, 01080, México D. F., México.

[†] Department of Astronomy & Astrophysics, Pennsylvania State University, University Park, PA 16802, U.S.A.

[‡] Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

correlation (Spearman 1904, Kendall 1938), maximal linear correlation after transforming both variables (Hirschfeld 1935), the distance correlation (Székely, Rizzo and Bakirov 2007), and the curve-based methods reviewed in (Delicado 2009).

The following conditions form a set for a symmetric, nonparametric measure of dependence $\delta(X, Y)$ for two continuously distributed random variables X and Y (Schweizer and Wolff 1981):

1. $\delta(X, Y)$ is defined for any X and Y, neither of them being constant with probability 1
2. $\delta(X, Y) = \delta(Y, X)$
3. $0 \leq \delta(X, Y) \leq 1$
4. $\delta(X, Y) = 0$ if X and X and Y are independent
5. $\delta(X, Y) = 1$ if and only if each of X, Y is a strictly monotone function of the other
6. If f and g are strictly monotone on Range X and Range Y, respectively, then $\delta(f(X), g(Y)) = \delta(X, Y)$
7. If the joint distribution of X and Y is bivariate normal, with correlation coefficient r, then $\delta(X, Y)$ is a strictly increasing function ϕ of $|r|$
8. If (X, Y) and (X_n, Y_n), $n = 1, 2, \ldots$, are pairs of random variables with joint distributions J and J_n, respectively, and if the sequence J_n converges weakly to J, then $\lim_{n\to\infty} \delta(X_n, Y_n) = \delta(X, Y)$

Rényi's (1959) original postulates differ from the above in that: (1) They were not restricted to continuously distributed random variables; (2) Condition 5 was "$\delta(X, Y) = 1$ if either $X = f(Y)$ or $Y = g(X)$ for some Borel–measurable functions f and g"; (3) Condition 6 was "If f and g are Borel–measurable, one–one mappings of the real line into itself then $\delta(f(X), g(Y)) = \delta(X, Y)$"; (4) In condition 7, $\delta(X, Y)$ was required to be equal to $|r|$; (5) Condition 8 was not included.

Rényi (1959) defined the maximal correlation ρ between a pair of two random variables (X, Y) as

$$\sup\left\{\frac{\mathrm{Cov}(f(X), g(Y))}{\sqrt{V(f(X))V(g(Y))}}\right\}; V(f(X)) > 0, V(g(Y)) > 0 \tag{1}$$

where the supremum is taken over all functions of X and Y with finite second moments. The random variables X and Y are independent if and only if $\rho = 0$.

An explicit evaluation of the Rényi maximal correlation is not available for a general random variable (X, Y) except in very special cases. For a bivariate normal distribution with correlation r, the Rényi maximal correlation is $|r|$, testifying to the fact that $r = 0$ implies independence.

A general function for $f(X)$ of X is a linear function $\mathbf{a}^{'}\mathbf{U}$ of $\mathbf{U}$ for some vector $\mathbf{a}$. Similarly for $g(Y)$ of Y, that is, $\mathbf{b}^{'}\mathbf{V}$ of $\mathbf{V}$ for some vector $\mathbf{b}$. It can be shown that the maximal Rényi correlation in this case is given by $\rho = \sqrt{\mu_1}$ where $\sqrt{\mu_1}$ is the canonical correlation between the two random vectors, $\mathbf{U}$ and $\mathbf{V}$ (Sethuraman 1990).

Some of the most frequently used measures of dependence are: (1) Correlation coefficient which satisfies Postulates 2, 3, and 7; (2) Correlation ratios which satisfies Postulate 5; and (3) Maximal Correlation which has all the properties.

Actually modern data sets, contain hundreds of thousands or even millions of variable pairs from which we need to examine all possible associations and examine the most important.

1.2 Maximal Information Coefficient (MIC) and Maximal Information–based Nonparametric Exploration (MINE)

Reshef et al. (2011) have introduced a novel statistic to measure dependence which has two heuristic properties: *generality* (with sufficient sample size the statistic captures a wide range of associations, including specific function types as the exponential, periodic, linear, or any other functional relationship), and *equitability* (the statistic gives similar scores to equally noisy relationships.)

This statistic has been called **Maximal Information Coefficient (MIC)** and gives rise to a larger family of statistics, referred as Maximal Information–based Nonparametric Exploration (MINE).

The Maximal Information–based Nonparametric Exploration (MINE) is a class of statistics for identifying and classifying relationships between variables. It encompasses several measures to identify and characterize the type of associations in a data set. MINE also includes: MAS (Maximum Asymmetry Score) which captures departure from monotonicity; MEV (Maximum Edge Value) which captures closeness to being a function, and MCN (Minimum Cell Number) which captures complexity of the association.

To familiarize with the theory behind MIC we need to recall Shannon's entropy definition (Shannon and Weaver 1949) and mutual information.

1.2.1 Entropy and Mutual Information

The entropy is a measure of uncertainty, that is, a measure of the amount of information required on the average to describe the random variable. Thus, the higher the entropy, the more uncertain one is about a random variable. This concept was introduced by Shannon (1949) and it is a straightforward adaptation of the Gibbs entropy formula. The Shannon entropy $H(X)$ of a discrete random variable X with possible values $(x_1, \ldots, x_n)$ and probability mass function $p(X)$ is defined as

$$H_b(X) = -\sum_{X \in \mathcal{H}} p(x) \log_b p(x) \tag{2}$$

where b is referred to the base of the logarithm. In information theory it is common to assume $b = 2$ and hence the entropy is expressed in bits. Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities. We shall denote expectation by $\mathbb{E}$. Thus if $X \sim p(x)$, then the expected value of the random variable $g(X)$ is written

$$\mathbb{E}_p g(X) = \mathbb{E} g(X) = \sum_{X \in \mathcal{H}} g(x) p(x) \tag{3}$$

The entropy of X can also be interpreted as the expected value of $-\log p(X)$, where X is drawn according to probability-mass function $p(x)$. Thus,

$$H(X) = \mathbb{E}_p \log \frac{1}{p(X)} \tag{4}$$

This definition of entropy (eq. 4) is related to the definition of entropy in thermodynamics and has the following properties: $H(X) \geq 0$, and $H_b(X) = H_a(X)(\log_b a)$. We can extend the definition of entropy to a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ as

$$\begin{aligned} H(X,Y) &= -\sum_{X\in\mathcal{X}}\sum_{Y\in\mathcal{Y}} p(x,y)\log p(x,y) \qquad (5)\\ &= -\mathbb{E}\log p(X,Y) \end{aligned}$$

We also need to define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable. If $(X, Y) \sim p(x, y)$, then the conditional entropy $H(Y|X)$ is

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x)\\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)\\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \qquad (6)\\ &= -\mathbb{E}_{p(x,y)}\log p(Y|X) \end{aligned}$$

The *Mutual Information* is a measure of the amount of information that one random variable contains about another random variable. In other words, it is the reduction in the uncertainty of one random variable due to the knowledge of the other. This quantity turns out to be a new measure of dependence and was first proposed by Linfoot (1957).

Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution of the marginals $p(x)$ $p(y)$,

$$\begin{aligned} I(X;Y) &= -\sum_{X\in\mathcal{X}}\sum_{Y\in\mathcal{Y}} p(x,y)\log p(x,y)p(x)p(y) \qquad (7)\\ &= \mathbb{E}_{p(x,y)}\frac{p(X,Y)}{p(X)p(Y)} \end{aligned}$$

The entropy and mutual information are related through the expression,

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\\ &= \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)}\\ &= -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(x,y)\log p(x|y) \qquad (8)\\ &= -\sum_{x,y} p(x)\log p(x) - \left[-\sum_{x,y} p(x,y)\log p(x|y)\right]\\ &= H(X) - H(X|Y) \end{aligned}$$

Thus the mutual information $I(X; Y)$ is the reduction in the uncertainty of X due to the knowledge of Y. By symmetry of eq.8, it follows that $I(X; Y) = H(Y) - H(Y|X)$. Thus X says as much about Y as Y says about X. More details about information theory can be found in Cover and Thomas (1991).

1.2.2 Maximal Information Coefficient (MIC)

We consider a partition of the data set D into a grid G of ordered pairs $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, such that there are x bins (of variable size) covering x and y bins covering y (of variable size). The probability mass function of a particular grid cell is proportional to the number of data points falling inside that cell and so, for a given (x, y), there will be a maximal mutual information. We can construct a characteristic matrix $\mathbf{M}(D)$ whose elements obtained as

$$\mathbf{M}(D)_{x,y} = \frac{\max(I(X;Y))}{\log \min\{x, y\}} \tag{9}$$

are the highest normalized mutual information achieved by any of the corresponding grids. The MIC is defined to be the maximum value in $\mathbf{M}$, such that, $xy < C$:

$$MIC(D) = \max_{xy<C}\{\mathbf{M}(D)_{x,y}\}, \tag{10}$$

where C is a function of the sample size and represents the maximal grid size considered. Reshef et al. (2011) found empirically a satisfactory limit for C, that is, $C(n) = n^{0.6}$

The behaviour of the MIC statistic is that it tends to 1 for all never–constant noiseless functional relationships and to 0 for statistically independent variables.

Its statistical significance can be determined from comparison of a real value against a set of values from α^{-1} surrogate dats sets where α is the probability of false rejection. Since MIC is a rank–order statistic, the uncorrected p-value depends only on the score and on the sample size of the relationship under consideration. Pre–computed uncorrected p-values are available for different sample sizes at MINE's website `http://www.exploredata.net/Downloads/P-Value-Tables`.

1.3 Distance Covariance (dCov) and Distance Correlation (dCor)

Let be X and Y two random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, where p and q are positive integers. The characteristic functions of X and Y are denoted by f_X and f_Y, respectively, and the joint characteristic function of X and Y is denoted by $f_{X,Y}$. Distance covariance ($\mathcal{V}$) can be applied to measure the distance $\|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|^2$ between the joint characteristic function and the product of the marginal characteristic functions, and to test the hypothesis of independence $H_0 : f_{X,Y} = f_X f_Y$ against $H_1 : f_{X,Y} \neq f_X f_Y$. Thus, the distance covariance between two random vectors with finite first moments is the nonnegative number $\mathcal{V}(X, Y)$ defined by

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds \end{aligned} \tag{11}$$

Similarly, distance variance (dVar) is defined as the square root of

$$\mathcal{V}^2(X,X) = \mathcal{V}^2(X)\|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2 \tag{12}$$

It is clear that $\mathcal{V}(X,Y) \geq 0$ and $\mathcal{V}(X,Y) = 0$ if and only if X and Y are independent.

Distance correlation is a new measure of dependence between random vectors introduced by Székely, Rizzo and Bakirov (2007). For all distributions with finite first moments, distance correlation $\mathcal{R}$ generalizes the idea of correlation in the sense that $\mathcal{R}(X,Y)$ is defined for X and Y in arbitrary dimension, and $\mathcal{R}(X,Y) = 0$ characterizes independence of X and Y.

Distance correlation satisfies $0 \leq \mathcal{R} \leq 1$, and $\mathcal{R} = 0$ only if X and Y are independent.

The distance dependence statistics are defined as follows. For an observed random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, 2, \ldots, n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, define

$$a_{kl} = \|X_k - X_l\|_p, \quad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl} \tag{13}$$

$$\bar{a_{\cdot\cdot}} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \quad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot} \tag{14}$$

Similarly,

$$b_{kl} = \|Y_k - Y_l\|_q, \quad \bar{b}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} b_{kl}, \quad \bar{b}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} b_{kl} \tag{15}$$

$$\bar{b_{\cdot\cdot}} = \frac{1}{n^2}\sum_{k,l=1}^{n} b_{kl}, \quad B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot} \tag{16}$$

for $k, l = 1, 2, \ldots, n$, where the subscript "$\cdot$" denotes that the mean is computed for the index that it replaces.

The empirical (that is, obtained from the data) distance covariance $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} \tag{17}$$

Similarly, $\mathcal{V}_n(\mathbf{X})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2 \tag{18}$$

The empirical distance correlation $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ of two random variables is the square-root of

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}} & ,\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0 & ,\mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases} \tag{19}$$

The asymptotic distribution of $n\mathcal{V}_n^2$ is a quadratic form of centered Gaussian random variables, with coefficients that depend on the distributions of X and Y. If we use that statistic to test independence, we shall implement it as a permutation test.

2. Application of the Distance Correlation (dCor) and Maximal Information Coefficient (MIC) to an astronomical large data set

We choose the well–known catalog named COMBO–17 for our computations about the relationships among variables. In Figure 1 we show one of the COMBO-17 fields to illustrate the type of objects we are interested in.

2.1 Description of the COMBO–17 data set

The COMBO–17 project ("Classifying Objects by Medium–Band Observations in 17 Filters") was mainly carried out to study the evolution of galaxies and their associated dark matter haloes at $z \lesssim 1$ as well as the evolution of quasars at $1 \lesssim z \lesssim 5$. In order to obtain large samples of objects (~ 50000 galaxies and $\lesssim 1000$ quasars) with precise photometric redshifts, four fields with a total area of $1\square^\circ$ were observed with a 17–band filter set covering the range of $\lambda_{obs} \sim 350 - 930$nm. In practice, such a filter set provides a redshift accuracy of $\sigma_{z,gal} \approx 0.03$, $\sigma_{z,QSO} \lesssim 0.1$, smoothing the true redshift distribution of the sample only slightly and allowing the derivation of luminosity functions.

All objects in the catalogue are found in the Chandra Deep Field South, based upon images obtained in 2003 with the Wide Field Imager on the ground–based 2.2–m MPG/ESO telescope located at the European Southern Observatory (ESO) on La Silla, Chile. This camera covers an area of more than $0.5^\circ \times 0.5^\circ$, which is larger than the field initially observed from space by the Great Observatories Origins Deep Survey (GOODS).

The foremost data analysis goal of the COMBO–17 approach is to convert the photometric observations into a very low–resolution spectrum that allows simultaneously a reliable spectral classification of stars, galaxies of different types and QSOs as well as an accurate redshift (or SED) estimation for the latter two. The full survey catalogue contains 63501 astronomical objects with classifications and redshifts on $1.5\square^\circ$ of area. It also includes restframe luminosities in Johnson, SDSS, and Bessell passbands and estimated errors.

The catalogue can be used to analyze aspects of galaxy evolution (Wolf et al. 2003a; 2003b, Bell et al. 2004), the evolution of faint AGN from redshift 5 to 1 (Wolf et al. 2003b), weak lensing studies (see for example Gray et al. 2002, Kleinheinrich et al. 2003).

The data set is available at the COMBO website (`http://www.mpia.de/COMBO/combo_index.html`). It lists identifiers, positions, magnitudes, morphologies, classification and redshift information. A detailed description of the column entries in both published FITS and ASCII versions of the catalogue and further explanations are given in (Wolf et al. 2004) and also at (`http://www.mpia.de/COMBO/cat_legend.html`).

2.2 Application of the MINE and dCor statistics

From the set of variables listed in Table 3 of Wolf's original paper (Wolf et al. 2004) we select 33 variables (5 contain general information, 3 correspond to the classification results, 3 are the total restframe luminosities and 22 are observed seeing–adaptive aperture fluxes in runs D, E, and F). We are not considering the estimated errors in the variables, we only include positive values for the fluxes and we neglect any missing information in the set of variables. In Table 1 we summarize the variables and their meanings chosen for our analysis.

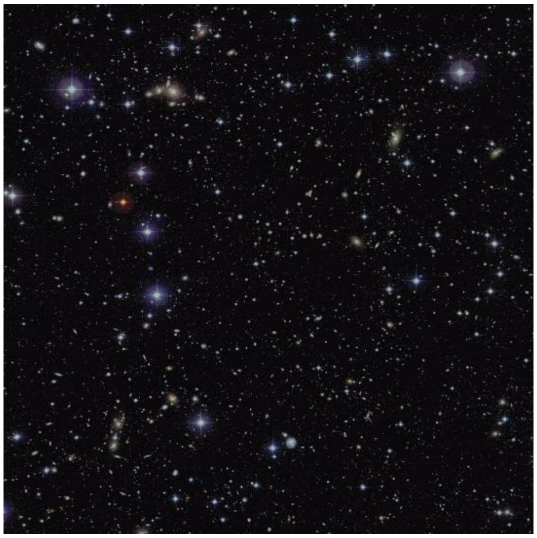

Figure 1: Cluster field Abell 901/902 (Taken from COMBO–17 website).

Table 1: Set of variables and their description considered in our analysis based on the COMBO17 catalog.

General information about the object	
Variable	Meaning
Rmag	total R-band magnitude
mu_max	central surface brightness
MajAxis	major axis
MinAxis	minor axis
PA	position angle
Classification results	
Variable	Meaning
MC_z	mean redshift in distribution
MC_z2	alternative redshift if distribution is bimodal
MC_z_ml	peak redshift in distribution
dl	luminosity distance of MC_z
Total object restframe luminosities	
Variable	Meaning
BjMag	$M_{abs,gal}$ in Johnson B ($z \approx [0, 1.1]$)
rsMag	$M_{abs,gal}$ in SDSS r ($z \approx [0, 0.5]$)
S280Mag	$M_{abs,gal}$ in 280/40 ($z \approx [0.25, 1.3]$)
Observed seeing-adaptive aperture fluxes	
Variable	Meaning
W420F_E	photon flux in filter 420 in run E
W462F_E	photon flux in filter 462 in run E
W485F_D	photon flux in filter 485 in run D
W518F_E	photon flux in filter 518 in run E
W571F_D	photon flux in filter 571 in run D
W571F_E	photon flux in filter 571 in run E
W604F_E	photon flux in filter 604 in run E
W646F_D	photon flux in filter 646 in run D
W696F_E	photon flux in filter 696 in run E

W753F_E	photon flux in filter 753 in run E
W815F_E	photon flux in filter 815 in run E
W856F_D	photon flux in filter 856 in run D
W914F_D	photon flux in filter 914 in run D
W914F_E	photon flux in filter 914 in run E
UF_F	photon flux in filter U in run F
BF_D	photon flux in filter B in run D
BF_F	photon flux in filter B in run F
VF_D	photon flux in filter V in run D
RF_D	photon flux in filter R in run D
RF_E	photon flux in filter R in run E
RF_F	photon flux in filter R in run F

The final study includes 15,352 galaxies over a redshift range from 0 to 2. The data were subdivided by galaxy type (1 to 4) based on their magnitudes as defined by Wolf et al. (2003a). The complete set of galaxies is summarized in Table 2.

Table 2: Galaxy Analysis Scheme

Redshift Color-Magnitude	$0 \leq z < 0.5$	$0.5 \leq z < 1$	$1 \leq z < 2$	Total
$B - r > 1.25$, $m_{280} - B \geq 1.1$	38	50	16	104
$B - r > 1.25$, $m_{280} - B < 1.1$	45	19	4	68
$0.95 < B - r \leq 1.25$	328	277	109	714
$B - r \leq 0.95$	3254	9284	1928	14466
Total	3665	9630	2057	15352

We apply MINE statistics to each group of galaxies according to their (1) redshifts and (2) galaxy type (Figure 2). For the set of 33 variables, there are $\binom{33}{2} = 528$ possible pairs to compute. The MINE application, which computes MIC and other statistics from the MINE family, can be downloaded from the website `http://www.exploredata.net/` for use in both Java and `R` statistical language. The input parameters for MINE are summarized in Table 3. For each computed pair, we obtain the MIC (Maximal Information Coefficient) that represents the strength of a relationship.

We compute dCor for each pair of variables using the `energy` package in `R` (Rizzo and Székely 2013).

3. Results and Discussion

The results of this first statistical study of the COMBO-17 galaxies are displayed in Figures 3–5 for the four galaxy types and three redshift groups shown in Table 2. Figure 2 shows the galaxy types based on their $m_{280} - B$ and $B - r$ colors. The magnitude ranges were derived from the galaxy-type cutoffs seen in Figure 2 of Wolf's paper (2003a); these ranges are associated with the Kinney et al. (1996) galaxy classification template of galaxy types.

The application of MINE statistics to the 528 pairs of variables based on the list of 33 variables in Table 1 revealed horseshoe patterns in the data. These patterns

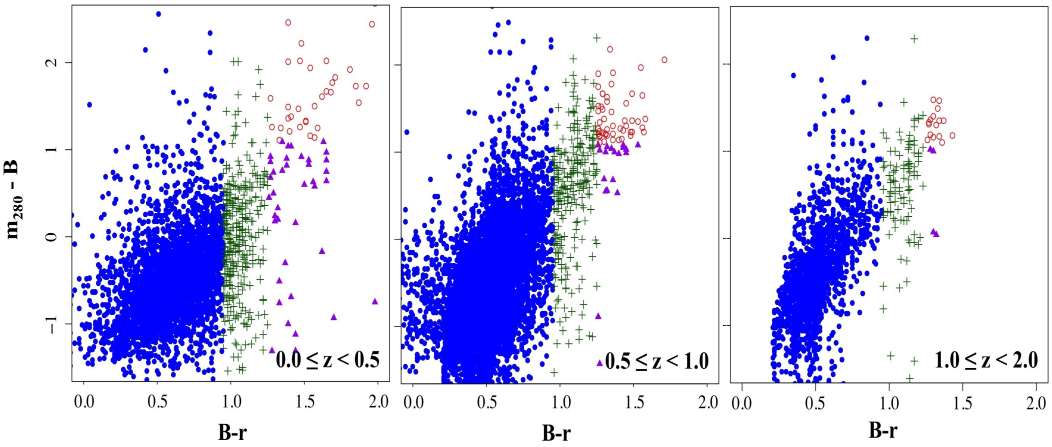

Figure 2: Galaxy types based on their $m_{280} - B$ and $B - r$ colors for $0 \leq z < 0.5$ (left), $0.5 \leq z < 1$ (middle),and $1 \leq z < 2$ (right): Type 1 (red), Type 2 (purple), Type 3 (dark green), and Type 4 (blue).

can be seen in Figures 3–5, which show the Pearson correlation coefficient versus the MIC score for galaxies of all types, over three redshift ranges from $z = 0$ to 2. In these graphs, a low MIC score corresponds to a weak relationship between a given pair of variables, while a high MIC score corresponds to a strong relationship between variables. The horseshoe patterns are due to the negative values that the Pearson correlation coefficient can take which indicates the direction of such an association, while MIC scores just provides the strength of it. In that sense, Pearson provides more information about a (linear) relationship.

Even though the figure clearly shows that the horseshoe pattern persists across the galaxy types, it exhibits more variability (dispersion) as the sample size increases. Thus, the accuracy to the exact value of the Pearson correlation coefficient and to the MIC score is better for large sample sizes (Type 4 galaxies with redshifts $0 < z < 2$).

Moreover, the MIC scores show lower values about 0.17 to 0.20 for pairs comparing observational variables like the position angle, the minor and major axes with some photon fluxes. In fact, those pairs also have a low Pearson correlation coefficient (in some cases it is negative). That probably means statistical independence between some fluxes and the localization of the galaxy.

Although the distance correlation measure can be applied to variables of any dimension (Székely et al. 2007), we compared the Pearson coefficient with the distance correlation coefficient for the same set of 528 pairs of variables as done for the MIC score. The results displayed in the low panels of Figures 3–5 reveal the much more distinctive relation between the Pearson and Distance Correlation measures. In other words, the distance correlation provides a more accurate measure of dependence (because of less variability) than the MIC score for this very large sample of 15,352 galaxies. Here again, a low distance correlation coefficient suggests near-independence between a given pair of variables, while a high coefficient represents a strong relationship between variables.

Finally, we find an interesting relationship between Distance Correlation measures and MIC scores for the COMBO-17 dataset. Figure 6 shows the case for galaxies with low redshift.

Although Reshef et al. (2011) proposed MIC scores as an alternative to classical

Table 3: Input parameters for MINE statistics

Mandatory	Description	Default value
infile	File (csv format) containing the data	–
style	Tells MINE which variable pairs to analyze	allPairs adjacentPairs masterVariable onePair pairsBetween
Optional		M
cv	A floating point number indicating which % of the records need to have data in them for both variables	0
exp	The exponent in $C(n) = n^\beta$	0.6
c	Determines by what factor clumps may outnumber columns when the algorithm starts to partition	15
notify	Number of variable pairs to analyze before printing a status message	100
gc	Number of pairs to analyze before forcing a Java garbage collection	Integer.MAX_VALUE

correlation measures, some criticisms of the MIC approach have been made recently by Gorfine et al. (2012), Simon and Tibshirani (2012), and Kinney and Atwal (2013). They noted that MIC scores sometimes are less powerful than the Pearson correlation coefficient for the linear case; when sample sizes are small (such as 50); for functional relationships at identical noise levels; and also that the power of the MIC procedure varies dramatically between the various relationships, i.e., MIC tends to have a strong preference for certain types of functions.

4. Conclusions

The rate of scientific discovery in astronomy is tied to the amount of data available, which has grown enormously due to modern detectors and computational resources. The new challenge is to analyze these huge amounts of data to find significant relationships – linear, non-linear, functional, structural – between pairs, triplets, and groups of properties or variables, that characterize an astronomical object.

In recent years, a number of approaches to extract and identify such relationships or associations have been derived, for example, Székely et al. (2007), Ball and Brunner (2010). We have focused in this paper on two novel techniques: Distance Correlation (dCor) and Maximal Information Coefficient (MIC). We applied both techniques to a well-known survey for galaxy clusters, called the Chandra Deep Field South COMBO-17 database. This database consists of observations in 17 filters of photon fluxes drawn from different astrophysical sources: stars, galaxies, and quasars. We only considered galaxies in the database, and these were classified by redshift and galaxy type, and subsequently we studied bivariate associations between the underlying astrophysical variables.

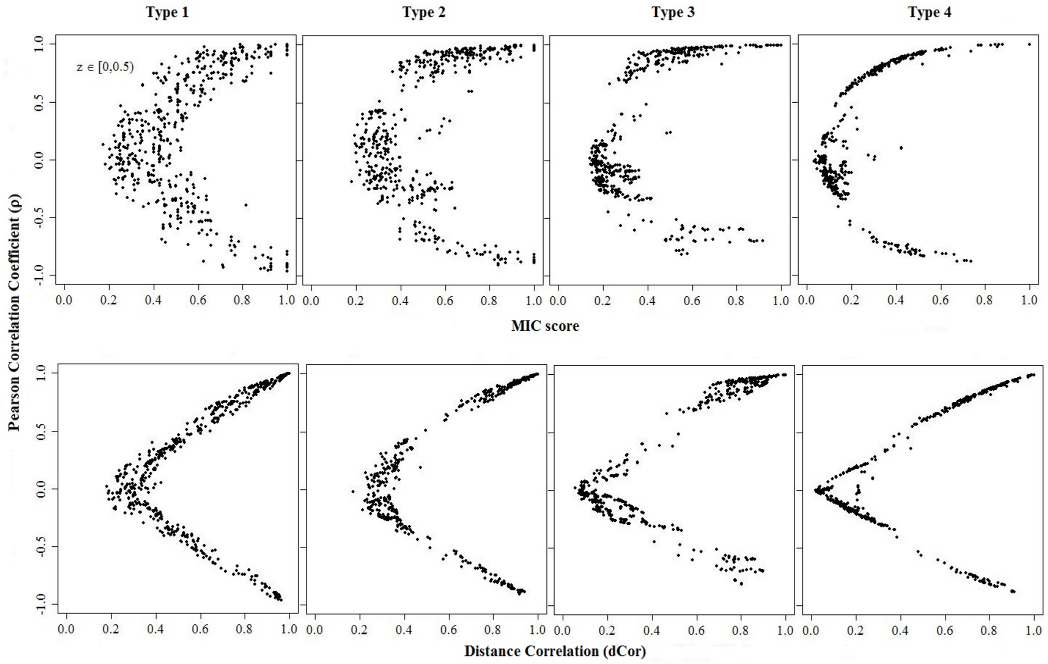

Figure 3: Comparison between Pearson correlation coefficient (ρ) versus MIC score and dCor for galaxies with redshifts $0 \leq z < 0.5$ and the color types described in Table 2.

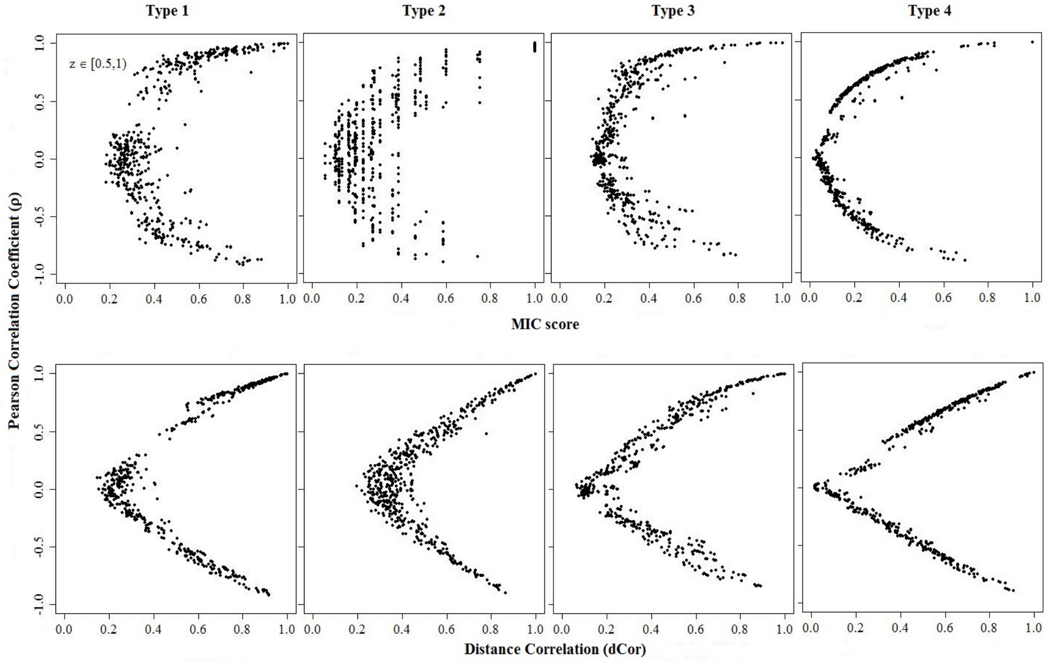

Figure 4: Comparison between Pearson correlation coefficient (ρ) versus MIC score and dCor for galaxies with redshifts $0.5 \leq z < 1$ and the color types described in Table 2.

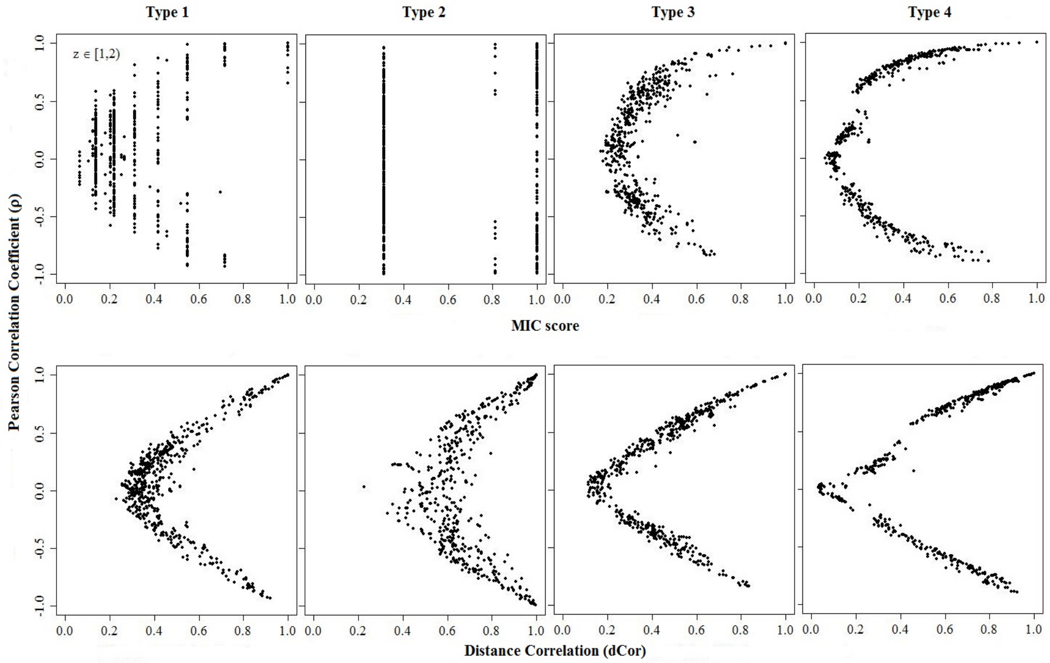

Figure 5: Comparison between Pearson correlation coefficient (ρ) versus MIC score and dCor for galaxies with redshifts $1 \leq z < 2$ and the color types described in Table 2.

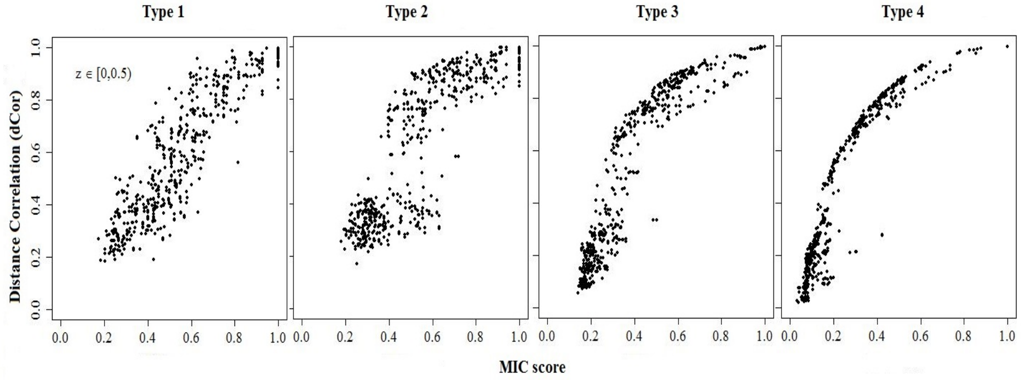

Figure 6: Distance correlation (dCor) versus MIC score for galaxies with redshifts $0 \leq z < 0.5$ and the color types described in Table 2.

To illustrate some of our results, we provide in the upper rows of Figures 3-5 graphs of Pearson correlations (ρ) *vs.* MIC scores for four galaxy types and three redshift ranges. In the lower rows of Figures 3-5, we provide plots of Pearson correlations (ρ) *vs.* distance correlations (dCor) for the same types of galaxies. The galaxies themselves are classified according to their redshift, z. Galaxies in Type 1 tend to be of spiral shape; Type 3 galaxies tend to be of elliptical shape; and Type 2 galaxies tend to be in-between spiral and elliptical in shape. We have found that distance correlation provides a more accurate measure of dependence between variables, and this has helped us to confirm some obvious astrophysical relationships.

REFERENCES

Ball, N. M., and Brunner, R. J. (2010), "Data mining and machine learning in astronomy," *International Journal Modern Physics D*, 19, 1049–1106.

Bell, E. F., Wolf, C., Meisenheimer, K., Rix, H.-W., Borch, A., Dye, S., Kleinheinrich, M., Wisotzki, L., and McIntosh, D. H. (2004), "Nearly 5000 distant early-type galaxies in COMBO-17: A red sequence and its evolution since $z \sim 1$," *Astrophysical Journal*, 608, 752–767.

Cover, T. M., and Thomas, J. A. (1991), *Entropy, Relative Entropy and Mutual Information in Elements of Information Theory*: Wiley, New York.

Delicado, P., and Smrekar, M. (2009), "Measuring non-linear dependence for two random variables distributed along a curve," *Statistical Computing*, 19, 255–269.

Galton, F. (1886), "Regression towards mediocrity in hereditary stature," *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

Heller, R., Heller, Y., and Gorfine, M. (2013), "A consistent multivariate test of association based on ranks of distances," *Biometrika*, 100, 503–510.

Gray, M. E., Taylor, A. N., Meisenheimer, K., Dye, S., Wolf, C., and Thommes, E. (2002), "Probing the distribution of dark matter in the A901/902 supercluster with weak lensing," *Astrophysical Journal*, 568, 141–164.

Gorfine, M., (2012), "Comment on Detecting novel associations in large data sets," `http://iew3.technion.ac.il/gorfinm/files/science6.pdf`.

Hirschfeld, H. O.(1935), "A connection between correlation and contingency," *Proceedings of the Cambridge Philosophical Society*, 31, 520–524.

Kendall, M. G. (1938), "A new measure of rank correlation," *Biometrika*, 30, 81–93.

Kinney, A. L., Calzetti, D., Bohlin, R. C., McQuade, K., Storchi-Bergmann, T., and Schmitt, H. R. (1996), "Template UV to near-IR Spectra of Starforming Galaxies and their Application to K-Corrections." *Astrophysical Journal*, 467, 38–60.

Kinney, J. B. and Atwal, G. S. (2013), "Equitability, mutual information, and the maximal information coefficient," arXiv, 1301.7745v1.

Kleinheinrich, M., Schneider, P., Erben, T., Schirmer, M., Rix, H.-W., Meisenheimer, K. and Wolf, C. (2003), "Galaxy-galaxy lensing in the COMBO-17 survey," In: *Astronomische Nachrichten*, Supplementary Issue 2, Vol. 324, Short Contributions of the Annual Scientific Meeting of the Astronomische Gesellschaft in Berlin, September 23-28, 2002, p.37.

Linfoot, E. H. (1957), "An informational measure of correlation," *Information and Control* 1, 85–89.

Pearson, K. (1895), "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, 58, 240–242.

Rényi, A. (1959), "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungaricae*, 10, 441–451.

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

Rizzo, M. L., and Székely, G. J. (2013), "E-statistics (energy statistics) *R*-package," `http://cran.us.r-project.org/web/packages/energy/index.html`

Schweizer, B. and Wolff, E. F. (1981), "On nonparametric measures of dependence for random variables," *Annals of Statistics*, 9, 879–885.

Sethuraman, J. (1990), "The asymptotic distribution of the Rényi maximal correlation," *Communications in Statistics – Theory & Methods*, 19, 4291–4298.

Shannon, C. E. and Weaver, W. (1949), *The Mathematical Theory of Communication*: University of Illinois Press, Urbana, Illinois.

Simon, N. and Tibshirani, R. (2012), Comment on "Detecting novel associations in large data sets" by Reshef, et al., Science, Dec. 16, 2011. http://www-stat.stanford.edu/tibs/reshef/comment.pdf.

Spearman, C. (1904), "The proof and measurement of association between two things," *American Journal of Psychology*, 15, 72–101

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

Székely, G. J., and Rizzo, M. (2009), "Brownian distance covariance," *Annals of Applied Statistics*, 3, 1236–1265.

Wolf, C., Meisenheimer, K., Rix, H.-W., et al. (2003a), "The COMBO-17 survey: Evolution of the galaxy luminosity function from 25000 galaxies with $0.2 < z < 1.2$," *Astronomy and Astrophysics*, 401, 73–98.

Wolf, C., Wisotzki, L., Borch, A., et al. (2003b), "The evolution of faint AGN between $z \simeq 1$ and $z \simeq 5$ from the COMBO-17 survey," *Astronomy and Astrophysics*, 408, 499–514.

Wolf, C., Meisenheimer, K., Kleinheinrich, M., Borch, A., Dye, S., Gray, M., Wisotzki, L., Bell, E. F., Rix, H.-W., Cimatti, A., Hasinger, G., Szokoly, G. (2004), "A catalogue of the Chandra Deep Field South with multi-colour classification and photometric redshifts from COMBO-17," *Astronomy and Astrophysics*, 421, 913–936.