Research Journal of Applied Sciences, Engineering and Technology 4(19): 3580-3583, 2012
ISSN: 2040-7467

Submitted: January 27, 2012 Accepted: March 06, 2012 Published: October 01, 2012

Measuring Association Between Two Variables: A New Approach

Masoud Yarmohammadi
Department of Statistics, Payame Noor University, 19395-4697 Tehran I.R. of Iran

Abstract: In this study we introduce a new non parametric method for measuring the level of dependence between two variables. The proposed method does not require any assumptions about the normality or linearity. The results show that the proposed method works well for any kind of relationship between two data sets.

Keywords: Association, embedding, decomposition, trajectory matrix

INTRODUCTION

Correlation analysis is an essential tool in many different statistical methods such as regression analysis, time series analysis, multivariate analysis. There are various measures of association with different strength. The sample correlation coefficient, for example, commonly used for estimating the linear correlation is the Pearson's product moment correlation coefficient. Spearman's rank correlation and Kendall's coefficient also are well known non-parametric correlation coefficients. These coefficients achieve their maximum values not only for linear dependence, but for any strictly monotonic relationship. However, for non-monotonic relationships (e.g., $Y = X^2$ with positive and negative values of X), these measures may fail to detect the dependence. Generally speaking, none of the existing well-known measures of association are appropriate when the range of possible relationships between X and Y is wide. Here, we introduce a new approach to overcome these shortcomings. The idea of this approach arises from Singular Spectrum Analysis (SSA), which is a relatively new and powerful technique for time series analysis. One of the advantages of the proposed approach in this paper is that it does not require any assumptions about the normality or linearity. Moreovere, it works well for any kind of relationship, either linear and nonlinear, between two data sets.

Singular spectrum analysis SSA: In recent years SSA as a powerful technique of time series analysis has been developed and applied to many practical problems (Hassani, 2007; Hassani, 2009; Hassani *et al*., 2009a-d; Ghodsi *et al*., 2009; Hassani and Thomako, 2010; Mahmoudvand and Zokaei, 2011, 2012). A thorough description of the theoretical and practical foundations of the SSA technique (with several examples) can be found in Golyandina *et al*. (2001).

It should be noted that despite the fact that a lot of probabilistic and statistical elements are employed in the SSA-based technique but the technique does not make any statistical assumption concerning either signal or noise while performing the analysis and investigating the properties of the algorithms (Hassani, 2007). This matter can be considered as one of the advantages of the technique against other classical methods which usually rely on some restricted assumptions.

The SSA technique consists of two complementary stages: decomposition and reconstruction and both of which include two separate steps. The original time series is decomposed into a number of additive time series, each of which can be easily identified as being part of the modulated signal, or as being part of the random noise. This is followed by a reconstruction of the original series. Here, we mainly follow Hassani (2007).

Consider the real-valued non-zero time series $Y_T = (y_1,...,y_T)$ of sufficient length T. Let K = T-L+1, where L ($L \leq T/2$) X is some integer called the window length. Define the matrix:

$$X = (x_{ij})_{i,j=1}^{L,K} = [X_1 \ldots, X_K] \quad (1)$$

where, $X_j = (y_j,...,y_{L+j-1})^T$. We then consider x as multivariate data with L characteristics and K = T – L + 1 observations. The columns $X_j = (y_j,...,y_{L+j-1})^T$ of, considered as vectors, lie in an L -dimensional space R^L. Define the matrix xx^T. Singular Value Decomposition (SVD) of xx^T provides us with the collections of L eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L \geq 0$ and the corresponding eigenvectors U_1, ..., U_L, where U_i is the normalized eigenvector corresponding to the eigenvalue λ_i (i = 1, ..., L).

A group of r (with $1 \leq r < L$) eigenvectors determine an r-dimensional hyperplane in the L-dimensional space R^L of vectors X_j. If we choose the first r eigenvectors $U_1,\ldots,U_r$, then the squared L_2-distance between this projection and X is equal to $\sum_{j=r+1}^{L} \lambda_j$. According to the Basic SSA algorithm, the L-dimensional data is projected onto this r-dimensional subspace and the subsequent averaging over the diagonals allows us to obtain an approximation to the original series.

MAIN RESULTS

Let $(x_1, y_1), ..., (x_N, y_N)$ is an i.i.d sample of a bivariate distribution of the random variables (X,Y) and the aim is to find a correlation between X and Y, if there is any one. In the following we give an steps-by-steps algorithm to show the way of estimating association between two variables X and Y.

Step 1: Sorting: Sort sample $(x_1, ..., x_N)$ into ascending order and let $(R_1, ..., R_N)$ are the corresponding rank. Then, consider the new sample $\left(x_{R_1}, y_{R_1}\right),...,\left(x_{R_N}, y_{R_N}\right)$.

Step 2: Embedding: Let L is a integer number, $2 \le L \le N-1$ and k = N-L+1. Then, the result of this step is the block Hankel trajectory matrix:

$$H = [X\ Y] \quad (2)$$

where, X and Y are the trajectory matrices corresponding to samples $(x_1,...,x_N)$ and $(y_1,...,y_N)$ with the following definition:

$$X = \begin{bmatrix} x_{R_1} & \cdots & x_{R_k} \\ \vdots & \cdots & \vdots \\ x_{R_L} & \cdots & x_{R_N} \end{bmatrix} Y = \begin{bmatrix} y_{R_1} & \cdots & y_{R_N} \\ \vdots & \cdots & \vdots \\ y_{R_L} & \cdots & y_{R_N} \end{bmatrix}$$

Step 3: SVD: In this step we perform SVD of H. Denote by $\lambda_1, ..., \lambda_L$ the eigenvalues of HH^T arranged in the decreasing order $(\lambda_1 \ge ... \lambda_L \ge 0)$

Step 4: Estimation: Considering the r largest eigenvalues of XX^T, we define the following quantity as a measure of association by order (L,r) and denote it by $Dep^L_r(X,Y)$:

$$Dep_r^L(X,Y) = \frac{\sum_{j=1}^{r}\lambda_j}{\sum_{j=1}^{L}\lambda_j} = \frac{\sum_{j=1}^{r}\lambda_j}{tr(HH^T)} = \frac{\sum_{j=1}^{r}\lambda_j}{tr(XX^T)+tr(YY^T)} \quad (3)$$

Measure $Dep^L_r(X, Y)$ shows both severity and simplicity of association between X and Y. Larger values of $Dep^L_r(X, Y)$ shows more association. Moreover, larger value of $Dep^L_r(X, Y)$ for small value of r shows more simplicity. For example, the linear association is achieved by r = 2, whereas r = 3 should be considered for the quadratic association.

Properties of $Dep^L_r(X, Y)$: According to the definition of $Dep^L_r(X,Y)$, we have the following properties:

- For all possible values of L, r we have $0 < Dep^L_r(X,Y) \le 1$
- For all values of L we have $Dep^L_r(X, Y) = 1$
- If Rank (H) = $r_0 \le L$ then $Dep^L_{r_0}(X,Y) = 1$
- For $r_1 < r_2 \ge L$ we have $Dep^L_{r_1}(X,y) \le Dep^L_{r_2}(X,Y)$.
- For all nonzero constants c we have: $Dep^L_r(cX,cY) = Dep^L_r(X,Y)$
- Let $I = \{i_1,\ldots,i_n\}$ be an arbitrary permutation of the index $\{1,\ldots,N\}$ and $X^{(I)} = \left(x_{i_1},...,x_{i_N}\right)$ and $Y^{(I)} = \left(y_{i_1},...,y_{i_N}\right)$ then $Dep^L_r(X^{(I)}, Y^{(I)}) = Dep^L_r(X, Y)$

Proof of these properties are straightforward, therefore we do not report them here. Note that the considered condition in the third case is a necessary condition as it is not difficult to find examples such that $y_i = g(x_i)$ for all values of x_i. The equality does not hold for some points. Furthermore, may not be a a full rank matrix. For instance, let $y_i = \beta_0+\beta_1x_i$ for i = 1,...,N-1 and $y_N = \sigma(\beta_0+\beta_1x_N)$, where, σ is an arbitrary real number. If $x_i = i$ for i = 1,...,N it is easy to show that:

$$Rank(H) = \begin{cases} 2 \ if \ \sigma = 1 \\ 3 \ if \ \sigma \ne 1 \end{cases} \quad (4)$$

However, the above example indicates that $Dep^L_r(X,Y)$ is robust with respect to the presence of outlier. This can be considered as another advantage of the proposed method.

Choosing L and r: It worth mentioning that the amount of $Dep^L_r(X, Y)$ depends on the value of L. Thus, choosing an improper value of L maybe mislead us to find a proper value for dependence. For instance according to the properties of $Dep^L_r(X, Y)$, choosing small values of L, increase the contributions of eigenvalues and show the improper value of dependence. The larger values of L are better than the smaller values. Furthermore, larger values of L provides better separability between subcomponents. There is no general rule about the value of r. However, if there is predefined model then there are some criteria. For example, let $X = (x_1, \ldots, x_N)$ be a sequence with fixed interval between x_i's. Then, the proper values of L and r are provided in the following table.

Note that the values of r in Table 1 are Rank(H). It is easy to expand the value of r for the other complicated models. For example, consider a polynomial model with order p:

$$y_i = \beta_0 + \beta_1x_i + ... + \beta_px_i^p$$

In this case we need to consider r = p+1. Moreover, as we mentioned in the previous section, $Dep^L_r(X, Y)$ is a non-decreasing functi on with respect to r. However, this

Table 1: The value of r or several models

Model	$\beta_0+\beta_1x$	$\beta_0+\beta_1x+\beta_2x^2$	$\exp(\beta_0+\beta_1x)$	$\sin(\alpha x)$	$\log(x)$
r	2	3	3	4	6

Table 2: Association measures for example 1

	σ		
Measure	5	15	25
$Dep^{L}_{r}(X,Y)$	0.994	0.946	0.896
R_p	0.986	0.879	0.767
R_{sp}	0.987	0.878	0.774
R_k	0.912	0.694	0.585

Table 3: Association measures for example 2

	σ		
Measure	15	75	150
$Dep^{L}_{r}(X, Y)$	0.985	0.763	0.547
R_p	0.095	0.084	0.063
R_{sp}	0.073	0.069	0.057
R_k	0.050	0.047	0.039

Table 4: Association measures for example 3

	σ		
Measure	0.10	0.50	1.00
$Dep^{L}_{r}(X,Y)$	1.000	0.999	0.998
R_p	-0.109	-0.091	-0.072
R_{sp}	-0.108	-0.092	-0.072
R_k	-0.072	-0.062	-0.048

does not mean that the larger values of r are better than the smaller ones.

Numerical examples: In this section, we provide simulation results for linear and non-linear models with additive white noise with variance σ^2. For all data generating processes, we have simulated 1000 samples of length N by means of statistical software R. Then, standard simulation procedures are used to obtain the correlation coefficient estimates . Moreover, we consider a range of values for σ, from small to the large, to examine the robustness of the proposed approach in the condition where there are a high percentage of contamination in data. furthermore, we have considered Pearson (R_p), Speramn (R_{sp}) and Kendall (R_k) correlation coefficients as our benchmark. Although, there are several other correlation coefficients that work better than R_p, R_{sp} and R_k in some situations, but as the aim of this paper is just introducing a new measure of association, thus we do not compare the proposed method with other competitive methods here.

Example 1: Let $y_t = 1+2t+\varepsilon_t$, where, t = 1,…,50 and ε_t is a Gaussian error term with mean zero and variance σ^2. We have computed $Dep^{L}_{r}(X, Y)$ for L = 25 and r = 2. Table 2 shows the results for several values of σ. As appears from the table, R_p, R_{sp} and R_k show a small association (particularly for higher values of σ), whereas $Dep^{L}_{r}(X, Y)$ shows a relatively high association, even for the large values of σ .

Example 2: Let $y_t = x_t^2 + \varepsilon_t$, where $x_t = t-10$,t = 1,…,40 and ε_t is a Gaussian error term with mean zero and variance σ^2. We have computed $Dep^{L}_{r}(X, Y)$ for L = 20 and r = 3. Table 3 shows the results for several values of σ. As appears from this table, R_p, R_{sp} and R_k show a very small association, whereas $Dep^{L}_{r}(X, Y)$ shows a relatively high association, even for the large values of σ.

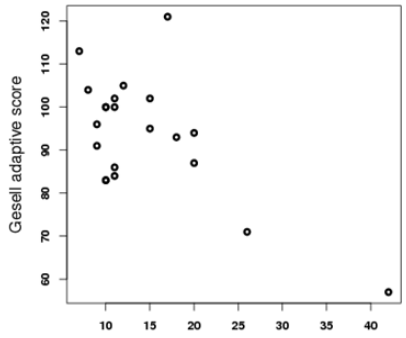

Fig. 1: Scatter plot of gesell adaptive score versus age at first word

Example 3: Let $y_t = \sin\left(\frac{2\pi t}{12}\right) + \varepsilon_t$, where, t = 1,…,40 and ε_t is a Gaussian error term with mean zero and variance σ^2. Table 4 shows the results for several values of σ. Here, we also observe similar results.

Example 4: The data set we consider here is a two-dimensional data set which has been widely used from different prospectives (Rousseeuw, 1987). The explanatory variable is age (in months) at which a child utters its first word, and the response variable is its Gesell adaptive score. Data are depicted in Fig. 1. For this data set we have $Dep_2^{11}(X, Y) = 0.986$, $R_p = -0.640$, $R_{sp} = -0.317$ and $R_k = -0.230$, confirming the superiority of the proposed approach in this study for a real data set.

CONCLUSION

In this study, we proposed a new measure of association based on the idea of subspace methods (more precisely SSA). The proposed approach has several desirable theoretical properties and does not require any assumptions about the normality or linearity. Moreover, it works well for any kind of relationship between two sets of data and therefore can be used for the situation where form of the relationship is unknown. The empirical results, simulation results and real data set, indicate that the performances of the new approach is promising with compare with other association measures in most of cases, even for the the situation where there are high percentages of contamination in data.

REFERENCES

Ghodsi, M., H. Hassani, S. Sanei and Y. Hicks, 2009. The use of noise information for detection of temporomandibular disorder. J. Biomed. Signal Proc. Contr., 4(2): 79-85.

Golyandina, N., V. Nekrutkin and A. Zhigljavsky, 2001. Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC, New York-London.

Hassani, H., 2007. Singular spectrum analysis: Methodology and comparison. J. Data Sci., 5(2): 239-257.

Hassani, H., 2009. Singular spectrum analysis based on the minimum variance estimator. Nonlinear Anal-Real, 11(3): 2065-2077.

Hassani, H., A. Dionisio and M. Ghodsi, 2009a. The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets. Nonlinear Anal-Real, 11(1): 492-502.

Hassani, H., S. Heravi and A. Zhigljavsky, 2009b. Forecasting european industrial production with singular spectrum analysis. Int. J. Forecasting, 25(1): 103-118.

Hassani, H., M. Zokaei, D. von. Rosen, S. Amiri and M. Ghodsi, 2009c. Does noise reduction matter for curve fitting in growth curve models? Comput. Meth. Prog. Bio., 96(3): 173-181.

Hassani, H. and A. Zhigljavsky, 2009d. Singular spectrum analysis: Methodology and application to economics data. J. Syst. Sci. Complex., 22(3): 372-394.

Hassani, H. and D. Thomakos, 2010. A review on singular spectrum analysis for economic and financial time series. Stat. Interface, 3(3): 377-397.

Mahmoudvand, R. and M, Zokaei, 2011. A filter based correlation coefficient by using singular spectrum analysis. Proceeding of the 31th Interantional Syposium on Forecasting. Prague, Czeck, June, 26-29.

Mahmoudvand, R. and M. Zokaei, 2012. On the singular values of the hankel matrix with application in the singular spectrum analysis. Chil. J. Stat. Forthcomming.

Rousseeuw, P.J. and A.M. Leroy, 1987. Robust Regression and Outlier Detection. John Wiley and Sons Inc., New York.