Available online at www.sciencedirect.com
ScienceDirect
Acta Mathematica Scientia 2018,**38B**(1):57–72
Acta Mathematica Scientia 数学物理学报
http://actams.wipm.ac.cn

ROBUST DEPENDENCE MEASURE FOR DETECTING ASSOCIATIONS IN LARGE DATA SET*

Hangjin JIANG (蒋杭进)[1,2] *Qiongli WU* (吴琼莉)[2†]

1. University of Chinese Academy of Sciences, Beijing 100049, China

2. Key Laboratory of Magnetic Resonance in Biological Systems, Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, Wuhan 430071, China

E-mail: jianghangjin10@mails.ucas.ac.cn; wuqiongli@wipm.ac.cn

Abstract In this paper, we proposed a new statistical dependency measure for two random vectors based on copula, called copula dependency coefficient (CDC). The CDC is proved to be robust to outliers and easy to be implemented. Especially, it is powerful and applicable to high-dimensional problems. All these properties make CDC practically important in related applications. Both experimental and application results show that CDC is a good robust dependence measure for association detecting.

Key words CDC; dependence measure; EDC; association; large dataset; robust

2010 MR Subject Classification 62B10; 62P99

1 Introduction

Measuring statistical dependence between random variables is a fundamental problem in a wide range of applications in science and engineering [5] such as information theory, machine learning, functional magnetic resonance imaging (fMRI) data analysis, image registration. The most well-known dependence measure is Pearson correlation coefficient, which is popular in many applications, especially in fMRI data analysis. It is the most powerful measurement for the linear relationship, however, its performance is not adequate in non-linear cases, which motivates us to develop new dependence measure.

The emergence of large data set in modern scientific research problems brings new challenges to the development of non-linear measurements (there are some theoretical constraints on dependence measure [15, 17]).

(a) The high requirement on the computing efficiency of the dependence measure becomes more and more urgent. Taking biological application for example, if it takes about 1 minute to get the dependence measure between two observations with large sample size, it will be quite a long time to get that of 2000 or more pairs of observations of the same size.

*Received January 26, 2017. Supported by the National Natural Science Foundation of China (31600290).

†Corresponding Author: Qiongli WU.

(b) The performance of the measurement should be powerful for all functional types. In practice, no one knows exactly what's the underlying relationship behind the data, which implies that the measurement should be sensitive to each kind of functional types.

(c) The robustness of the measurement should be guaranteed. The data we have is always noisy/contaminated, and the robustness is the prerequisition of handling with outliers in the data to make sure the result is reliable.

(d) It should be able to handle with high-dimensional variables. Because in many cases, we want to know the dependence between two random vectors, instead of between two variables.

These requirements are difficult to be fulfilled simultaneously, which is one of the reasons why so many dependence measures have been proposed in the literatures. The development of non-linear dependence measure goes deeper and deeper, although confronting with these difficulties. Maximum correlation coefficient (MCC) is the earliest measure that tries to fulfil these requirements and the properties needed for a dependence measure [15]. ACE (alternative conditional expectation, [1]) gives a quite good estimation of MCC in additive models, and Papadatos et al. [14] gave a simple method for obtaining it under some special situations. The measures proposed recently such as maximal information coefficient (MIC) [16], HHG [7], distance correlation (dCor, [20]) get some improvements on some special cases [19], however, there is still a long way to go. For example, MIC is not available for high-dimensional variables and time-consuming for large sample problems. dCor is not robust; one single large enough outlier can arbitrarily ruin the estimator. HHG is time-consuming when the sample size is large. Detailed comparison of these dependence measures can be found in [11]. Jiang et al. [11] classified the popular dependence measures into five classes, and found that ACE, RDC [12], MI (mutual information), HHG [7], HSIC (Hilbert-Schmidt independence criterion, [6]) with Laplace kernel is the most powerful one among dependence measures of its own kind respectively.

There are two types of the underlying model (functional relationship) between the predictors $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ and the response Y. The first is additive model, say, it can be assumed that $Y = \sum_{i=1}^{p} \phi_j(X_j) + \varepsilon$, where ε is the noise term. The MCC is the best method for dealing with such kind of model, since it obtains first the consistent estimator of ϕ_j, $\hat{\phi}_j$, and then compute $\rho(Y, \hat{Y})$, the correlation between Y and its estimator, where $\hat{Y} = \sum_{i=1}^{p} \hat{\phi}_j(X_j)$. For the second type, the non-additive model, methods such as dCor, HHG, all have its own shortcomings as showed in [11]. Generally, we do not know which type of model behind the data, which motivates us to propose in next section a new dependence measure, Copula dependence coefficient (CDC) based on copula and MCC, to deal with both cases. Defined as the MCC of two distribution-transformed random variables, say, $F_{\mathbf{X}}(\mathbf{X})$ and $F_{\mathbf{Y}}(\mathbf{Y})$, CDC has nice properties of MCC, where $F_{\mathbf{X}}(\mathbf{x})$ and $F_{\mathbf{Y}}(\mathbf{y})$ is the distribution function of $\mathbf{X}$ and $\mathbf{Y}$ respectively. The distribution function $F_{\mathbf{X}}(\mathbf{x})$ and $F_{\mathbf{Y}}(\mathbf{y})$ are usually unknown, so we replace them by their estimators to obtain the estimation of CDC, EDC.

The rest of this paper is organized as following. Section 2 focuses on the definition of CDC, and its estimation. Section 3 gives experimental results and a real data application. Section 4 concludes the paper.