Environmental Modelling & Software 96 (2017) 283–290

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A simple nonparametric index of bivariate association for environmental data exploration

Varvara Vetrova [a], Earl Bardsley [b, *]

[a] School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[b] Faculty of Science and Engineering, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

ARTICLE INFO

Article history:
Received 29 April 2017
Accepted 5 July 2017

Keywords:
Randomization
Nearest neighbour
Bivariate association index
Nonparametric

ABSTRACT

A purely data-based index for detecting bivariate association is proposed for preliminary data exploration when seeking to model a dependent variable, associated with a possibly large number of independent variables. No particular form of association between the dependent and independent variables is assumed. The proposed bivariate association index is the value **p**, which is the probability that a scatter plot created by an *X*-randomization will generate a smaller mean nearest neighbour distance. The rationale is that randomizing an existing *X-Y* association will result in a scatter plot which will usually have a greater mean nearest neighbour distance. The process is then repeated for all other independent variables to give a specific **p** for each one. A subset of potentially informative independent variables is then obtained by noting all those with low **p** values, but just how small **p** should be is left to the user.

Software availability

Software is free MATLAB code written by Varvara Vetrova. Available at https://github.com/vetrovav/P-index.

1. Introduction

In seeking to construct an explanatory model for an environmental variable, it is common to have situations where there are a large number of available candidate independent variables which may or may not have causal association with the dependent variable. Also, there may be no prior reason to assume that any association should be linear.

A starting point toward model construction in such situations is to use a procedure for selecting a subset of predictor variables with statistical association with the dependent variable. A criterion is usually applied to reduce redundancy so there is minimal correlation between the selected variables. Examples of such studies in environmental applications include Quilty et al. (2016), Tran et al. (2015), Sharma and Mehrotra (2014), Hejazi and Cai (2009), and May et al. (2008). A methodology overview on variable selection in general is included in Huang and Zhu (2016).

There can be many forms of association between variables (not necessarily causal) and there is merit in preliminary data exploration on a per-variable basis to give some first insight into possible causal associations. For linear associations this can be achieved by computing Pearson correlation coefficients for each independent variable. However, it is less clear as to the choice of an appropriate bivariate index when there may be undefined nonlinear associations.

The statistical literature has an extensive history of general bivariate association measures, including both specific bivariate indices and bivariate indices as special cases of multivariate indices. Bivariate association measures to date include two-sample comparisons with respect to empirical distribution functions (Blum et al., 1961), empirical characteristic functions (Kankainen and Ushakov, 1998), information measure (Linfoot, 1957; Kraskov et al., 2004; Sugiyama, 2011; Reshef et al., 2011), general nonlinear functional relations (Delicado and Smrekar, 2009), rank-based methods (Kallenberg and Ledwina, 1999; Heller et al., 2013), bivariate copulas (Schweizer and Wolff, 1981), distance correlation (Szekely et al., 2007), and continuous analysis of variance (Wang et al., 2015). Murrell et al. (2016) introduced a generalized coefficient of variation and Karvanen (2005) gives a resampling test of independence for application to data from two stationary time series.

Many of these tests of bivariate association, though powerful,

* Corresponding author.
E-mail addresses: varvara.vetrova@canterbury.ac.nz (V. Vetrova), earl.bardsley@waikato.ac.nz (E. Bardsley).

http://dx.doi.org/10.1016/j.envsoft.2017.07.006
1364-8152/

are not intuitive and often require nonparametric estimation of intermediary assumed quantities such as joint or marginal probability distributions. This may require, for example, specification of a bandwidth parameter for kernel estimation. We propose here a simple and intuitive nearest-neighbour index to draw attention to possible bivariate associations when dealing with large multivariate data sets. The method is truly nonparametric in the sense that only the data values are used, there is no estimation of intermediary entities or parameters, and no assumption is made about the mechanism of data generation.

Nearest neighbours have been used previously as an association measure in the context of a mutual information estimation (Kraskov et al., 2004). However, we believe our approach to be the first purely data-based use of nearest neighbour distances as an index of bivariate association.

Section 2 defines the nearest neighbour index for bivariate association and Section 3 illustrates the nature of the index using synthetic examples. Section 4 applies the method for preliminary identification of geographical locations where atmospheric pressures may most influence westerly wind frequency on the west coast of the South Island of New Zealand.

2. Index of bivariate association

As noted by Murrell et al. (2016), randomization has the effect of inducing independence between variables. This means that for a given bivariate data set $\{X,Y\}$ there can be no association between the two variables if the X values have been rearranged in random order (which implies the associated Y values will also have been rearranged in random order).

We use the scatter plot mean nearest neighbour distance $\overline{D}$ as a familiar measure to form a data-based index of association, with randomization being the reference base for no association. That is, the association index **p** is defined as the probability that a scatter plot created from a randomization of X will have its mean nearest neighbour distance less than $\overline{D}$. If there is some form of association between X and Y then a randomization of X will tend to produce scatter plots with increased mean nearest neighbour distances, resulting in small values of **p**. In practice, **p** will be obtained to the required level of accuracy by carrying out a sufficiently large number of randomizations.

In summary, the proposed index of bivariate association **p** is calculated through the following steps:

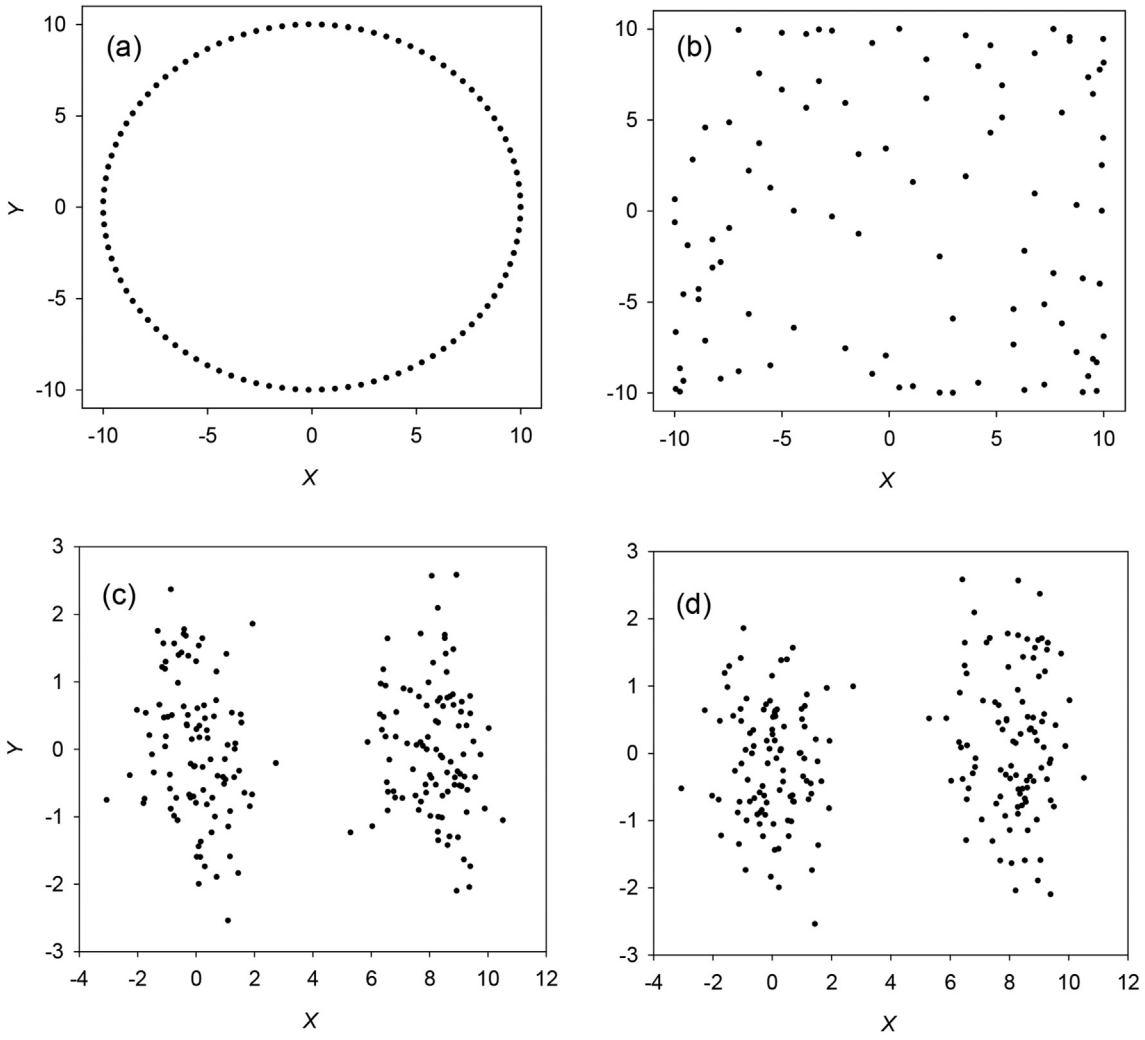

Fig. 1. Original scatter plots are shown in (a) and (c), with the effect of a single randomization shown in (b) and (d), respectively. X and Y are in arbitrary units.