A proposal of a method for distinguishing MAR type missing data from MCAR type

Jerzy Korzeniewski
University of Lodz, Lodz, Poland – jurkor@wp.pl

Abstract

Correct assessment of the mechanism responsible for creating missing data is very vital for empirical data sets analysis. If the missing data have random character MCAR (Missing Completely At Random) the analysis is easier because the missing data can be ignored or replaced. The scope of methods dealing with distinguishing the MCAR type from the MAR (Missing At Random) is very limited. Actually, only one test with a rather limited applicability was proposed. In this article we present a proposal of a new method which might be of some help in this task in the case of data sets with a cluster structure. The idea of the method consists in using a measure of the strength of correlation between two variables based on the linear correlation between pairs of objects distances corresponding to the two variables. The measure can be computed in a number of instances e.g. taking into account missing values and treating them as a new value or, omitting missing values. If the strengths of correlation in both instances differ significantly it is a reason for considering the missing values to of the MAR type. The new method is first checked on some binary data sets with a known cluster structure and then on some data sets from the Irvine Data Sets Repository. The results sometimes return very certain answers.

Keywords: class structure, missing data, variable correlation, testing of MCAR type.

1. Introduction

In the literature one can find a couple of types of missing data. The most welcome one is the MCAR type by which no dependence on anything (neither other attributes or observations) or any kind of rule for their appearance is understood. The MAR type is less welcome because there is no dependence on the missing values but the appearance on missing values may depend on the values of other attributes. A good example of the MAR missing data on individuals income are misses for people with higher education but not for those with high income. Another type of missingness is the MNAR (Missing Not At Random) type in which the misses appear for some predefined values of the attribute on which they are to appear. In the example aforementioned that would mean missing values for those with high income. The appearance of missing data has a range of bad consequences like making the population or sample size smaller, making the estimators variances larger, causing heavier bias, distorting variables distributions etc. One should know the type of the missingness, on that the ways in which the misses are treated depends (deletion, weighing, imputation). The most convenient ones are the missing data of the MCAR type. However, there are very few methods which allow to check this type of missingness. Little (1988) proposed a test for comparing two means of quantitative attributes but only for those normally distributed. Some authors mention the possibility of using the chi-square test of independence but clear and successful applications are very hard to find. The applicability of both methods is very limited.
In this article a method using distance based correlation between attributes is proposed. This correlation measure might be useful in investigating data sets possessing a cluster structure, therefore the new method is investigated on this kind of data sets. In section two there is a statement of the new method. Section three contains examples of some applications of the new method to data sets from the Irvine Data Set Repository possessing cluster structure.

2. Method formulation

When there is a cluster structure in the data set, one can use a distance based correlation measure in order to measure the strength of correlation between two (or more) attributes. The idea consists in comparing two sets of corresponding distances between sampled pairs of objects, one set consists of distances on one attribute, the other of distances on the other attribute. If the data set has *n* objects, the measure can be formulated in the following way (*Korzeniewski*, 2012).

Definiton 1 The coefficient of distance based correlation between sets A, B of attributes is given by the formula

$$DBC(A,B,l) = \frac{\frac{1}{l}\sum_{i=1}^{l} d_i^A d_i^B - \bar{d}^A \bar{d}^B}{s^A s^B} \qquad (1)$$

where $1 \le l \le n$ denotes the number of pairs of objects drawn dependently from the set pf all pairs of objects; d_i^A, d_i^B denote distances for the *i*-th pair computed on the attributes from sets, respectively, *A*, *B*; $\bar{d}^A$, $\bar{d}^B$, s^A, s^B denote, respectively, arithmetic means and standard deviations computed on all l distances.

Table 1. Binary data sets. Rows represent clusters and columns represent attributes.

Number of clusters	Number of attributes		
	4	6	8
4	1 0 0 1 1 1 1 0 0 0 1 1 0 1 0 1	1 0 0 1 1 0 1 1 1 0 0 0 0 0 1 1 0 0 0 1 0 1 0 1	1 0 0 1 1 0 1 0 1 1 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 1 0
6	1 0 0 1 1 1 1 1 1 0 1 0 0 1 0 1 0 0 0 1 0 1 1 0	1 0 0 0 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 1 1 0	1 0 0 0 1 1 0 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 0 0 1
8	1 0 1 1 1 0 0 0 1 1 1 0 1 1 0 1 0 1 0 1 0 1 0 0	1 0 0 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 0 1 0 0 1 1 0 0 1	1 0 0 1 1 1 0 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 1

Source: Brusco (2004)

The coefficient of distance based correlation (*DBC*) depends on the number l of the pairs drawn, but this number will be fixed and equal to 30 throughout the investigation. Additionally, the computations are repeated 500 times and averaged. In the remainder of this article, the sets A and B will be one element sets, therefore the coefficient will be called the coefficient for a pair of attributes.

Applications of *DBC* to the investigation of attribute correlation is especially useful for data sets with a cluster structure due to the fact that two objects from the same cluster are, usually, not so far apart as two objects from different clusters. Following this reasoning one can try to use *DBC* to check if the missing data are of the MCAR type. The investigation of attribute correlation is the only way out if we want to tell the MCAR type from the MAR type. The *DBC* may help to establish the pairs of attributes which may be prone to generate missing data of the MAR type. We proceed in the following way.

We compare the strength of correlation computed on the set of objects without data misses with the strength of correlation computed on the set of objects with the data misses treated as new variant. The

applicability of *DBC* grows with growing strength of measurement scale as the weaker the scale, the less precise distance measurement. Therefore, in order to check this supposition, we will check how big the differences of the values of *DBC* in both cases will be in an experiment on artificial data sets made up by binary attributes only i.e. the ones with the weakest possible scale.
Each set consisted of 2000 objects generated according to the parameters set up given in table 1. The numbers of objects in clusters were equal. The missing data were introduced in the following way. One attribute was fixed (always attribute number 1) and on some of the remaining attributes (one or two) we introduce data misses (denoted with 2) in two modes. In the first mode the misses are random with respect to objects. In the second mode (nonrandom mode) the misses appear when there is 1 on attribute number 1. The missing data were introduced in the following manners:

- with frequencies equal to 5%, 10%, 20%;
- on one (last) attribute and on two (last) attributes.

The strength of correlation was measured by *DBC* given by formula (1) with the Sokal-Michener distance measure. The values of *DBC* were computed for each pair of attributes in two variants: 1) only on objects with data misses on a fixed attribute, 2) on all objects. If the set of all possible 30-element subsets is treated as population, then 500 such subsets drawn, constitute a random sample drawn from the population. Then we test the following null hypothesis against its alternative

$$H_0: \ DBC\{u,v\} \text{ on missing data} = DBC\{u,v\} \text{ on all objects}$$
$$H_1: \ DBC\{u,v\} \text{ on missing data} \neq DBC\{u,v\} \text{ on all objects}$$

by means of the statistic:

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}} \qquad (2)$$

where: $\bar{x}_1$ - sample arithmetic mean of *DBC(u,v)* , on all objects (missing data is treated as a new variant if one of the two attributes *u, v* is the attribute with missing data); $\bar{x}_2$ - sample arithmetic mean of *DBC(u,v)* , on objects for which there is missing data on a fixed attribute.
The results of this experiment were as follows. In the random variant, the differences between *DBC* were statistically insignificant for all pairs of variables. In the nonrandom variant, the differences were big for some pairs of attributes e.g. 0.2 differences for 10% of missing data and 0.4 for 20% of missing data. Such differences turn out to be statistically significant because the absolute values of statistic (2) were greater than 4.
These results prove that even for very weak measurement scales big differences in the values of *DBC* might be of some help in investigating the MCAR type of missing data.

3. Method evaluation on real world data sets

In this section the new method will be applied to some data sets from the Irvine Data Sets Repository meeting the following conditions: existing cluster structure, existing missing data. The sets examined are big enough with respect to the number of elements that each time the test statistic (2) will be used with 500 draws of 30-element subsets of pairs of objects.
Data set ***Votes.*** Objects: USA congressmen, number of objects: 435, number of attributes: 16, number of clusters: 2. Attribute characteristic: 16 binary attributes (*yes* or *no*) being the results of voting on 16 bills. Missing data characteristic: attribute 16 – 24% misses, attribute 2 – 11% misses, attribute 12 – 7% misses, attribute 15 – 6,7% misses.
The numbers presented in table 2 let one have some suspicions on which attributes missing data are likely to be considered of the MAR type and not MCAR type.

Table 2. Biggest differences in *DBC* values for the *Votes* data set.

First attribute	Second attribute	*DBC* on all objects	*DBC* on misses on attribute 16	Statistic value
1	3	.144	.258	-10.1
1	4	.165	.312	-13.2
1	7	.119	.221	-9.2
1	12	.150	.241	-8.1
1	13	.117	.222	-9.3
3	4	.527	.685	-16.9
3	13	.527	.337	-8.1
4	8	.473	.557	-8.8
4	13	.326	.451	-11.9
5	7	.443	.534	-8.9
5	9	.571	.403	16.2

Source: own computations.

Data set ***Adult.*** Objects: adult individuals, number of objects: 32162, number of attributes: 14, number of clusters: 2. Attribute characteristic: 16 continuous attributes and 8 nominal variables. Missing data characteristic: attribute 2 – 6% misses, attribute 7 – 11% misses, attribute 14 – 2% misses.

Table 3. Biggest differences in *DBC* values for the *Adult* data set.

First attribute	Second attribute	*DBC* on all objects	*DBC* on misses on attribute 2	Statistic value
3	9	.067	.319	-12.6
3	14	.085	-0.041	11.3
4	5	.518	.636	-15.4
5	14	.267	.038	18.7
6	10	.168	.262	-8.6

Source: own computations.

Data set ***Hepatitis.*** Objects: patients, number of objects: 155, number of attributes: 19, number of clusters: 2. Attribute characteristic: 6 continuous attributes and 13 binary attributes. Missing data characteristic: attribute 7 – 7% misses, attribute 8 – 6% misses, attributes 9, 10, 11, 12 – about 3% misses.

Table 4. Biggest differences in *DBC* values for the *Hepatitis* data set.

First attribute	Second attribute	*DBC* on all objects	*DBC* on misses on attribute 8	Statistic value
1	3	.006	.186	-18.7
1	5	.007	.134	-12.8
1	6	.032	.274	-25.4
1	16	.033	.288	18.7
1	18	-0.039	.322	-38.6
3	5	.014	.159	-15.2
3	15	.011	.206	-20.2
3	17	.036	.621	-69.9
5	6	.254	.653	-46.2
6	7	.366	.167	22.8
7	11	0.171	.020	15.1

Source: own computations.

For the *Hepatitis* data set similar differences appear on many attributes, e.g. for the data missing on attribute 10 nearly all *DBC* have high values and nearly all change substantially.

Data set ***Watertreatment.*** Objects: results of water quality measurements, number of objects: 527, number of attributes: 38, number of clusters: 2. Attribute characteristic: all variables are continuous. Missing data characteristic: more than a dozen attributes with the fraction of missing data of about a couple percent.

Table 5. Biggest differences in *DBC* values for the *Watertreatment* data set.

First attribute	Second attribute	*DBC* on all objects	*DBC* on misses on attribute 11	Statistic value
1	4	.000	.128	-11.07
1	5	.037	.168	-11.57
1	16	.085	.234	-12.7
1	21	-0.033	.125	-13.8
1	27	-0.021	.126	-12.3

Source: own computations.

In the case of this data set one can also try to compare coefficients of linear correlation. For example, on missing data on attribute 11 for some pairs of attributes the values of these coefficients is changed considerably.

Data set ***Audiology.*** Objects: patients, number of objects: 200, number of attributes: 69, number of clusters: 22. Attribute characteristic: 8 categorical ordered attributes and 61 binary attributes. Missing data characteristic: attribute 6 – 51% misses and four attributes with missing data fraction of about 3-5%. One attribute was excluded from the analysis because it had 97% of missing data.

Table 6. Biggest differences in *DBC* values for the *Audiology* dataset.

First attribute	Second attribute	*DBC* on all objects	*DBC* on misses on attribute 6	Statistic value
1	14	.020	.198	-11.7
1	19	-0.011	.156	-22.5
1	28	.002	.128	-8.5
1	47	.004	.211	-19.4

Source: own computations.

4. Conclusions

The research carried out allows to draw the following conclusions. The methods constructed so far by other authors usually are not suitable to investiagate the type of missingness for the real world data sets. The distance based correlation is flexible, it can be used for every measurement scale, however weak or strong. In the case of some data sets one can compare two values of correlation coefficients computed with and without taking into account missing values. It seems that big differences between the two values prove some indication for stating that the MCAR type of missingness is not true. If one manages to find an attribute which causes considerable changes in the strength of correlation for some pairs of other variables, then such result is an indication of the MAR type of missingness.

References

Brusco M., 2004, *Clustering Binary Data in the Presence of Masking Variables*, Psychological Methods vol. 9, p. 510-521.

Heitjan, D.F., Basu, S., 1996, *Distinguishing `Missing at Random' and `Missing Completely at Random*', American Statistician, 50, p. 207-213.

Kim, J., Curry, J., 1977, *The treatment of missing data in multivariate analysis*, Sociological, Methods & Research, 6, p. 215-240.

Korzeniewski J., 2012, *Metody selekcji zmiennych w analizie skupień. Nowe procedury.* Wydawnictwo Uniwersytetu Łódzkiego.

Little, R.J., 1988, *A test of missing completely at random for multivariate data with missing Values*, Journal of the American Statistical Association, 83, p. 1198-1202.