A New Method for Estimation of Missing Data Based on Sampling Methods for Data Mining

Rima Houari, Ahcène Bounceur, Tahar Kechadi, Reinhardt Euler

A New Method for Estimation of Missing Data Based on Sampling Methods for Data Mining

Rima Houari[1], Ahcène Bounceur[2] , Tahar Kechadi[3], Reinhardt Euler[2]

[1] University of Abderrahmane Mira Bejaia
[2] Lab-STICC Laboratory - European University of Britanny - University of Brest
[3] University College Dublin, Ireland

Abstract

Today we collect large amounts of data and we receive more than we can handle, the accumulated data are often raw and far from being of good quality they contain Missing Values and noise.
The presence of Missing Values in data are major disadvantages for most Datamining algorithms. Intuitively, the pertinent information is embedded in many attributes and its extraction is only possible if the original data are cleaned and pre-treated.
In this paper we propose a new technique for preprocessing data that aims to estimate the Missing Values, in order to obtain representative Samples of good quality, and also to assure that the information extracted is more safe and reliable.

Key-words : Datamining, Copulas, Missing Value, Multidimensional Sampling, Sampling.

1 Introduction and previous work

During the process of knowledge extraction in databases, companies are trying to understand how to extract the values of all the data they collect. Consequently the presence of Missing Data devaluates the power of Data Mining algorithms causes a major problem on the way to achieve knowledge. Phase specific treatment of some data is often necessary to remove or complete them. Especially during the extraction of knowledge, these incomplete data are mostly removed. This sometimes leads to the elimination of more half of the base; the information extracted is no more representative and not reliable.

Many techniques for processing Missing Data have been developed [17][11][2]. According to [10] and [20], there are three possible strategies for dealing with Missing Data; the first technique use the deletion procedures [19] [16] [11]. These methods allow to have a complete database, therefore this method sacrifices a large amount of data [13], which is a major weakness in Data Mining. The following technique relies on the use of alternative procedures (substitution), that are intended to built to a comprehensive basis by finding an appropriate way to replace Missing Values. Among these methods we can mention: the method

of mean imputation [12], the regression method [13][16][24], and the method of imputation by the k-nearest neighbor [3] [5] [27][28]. Generally, these methods are not adapted to the characteristics of Data Mining when processing large databases or large percentage of missing values.

The latter technique is to estimate certain parameters of data distribution containing Missing Values, such as the method of maximum likelihood [4] [16] and the expectation maximization[1][29][14][24], these techniques are very costly in computation time, they require more specification of a data generation model. This task involves making a certain number of hypotheses, which is always difficult for they solutions because they are not always feasible.

The great advantage of the presented method is to create a complete database. However, it is not beneficial for Datamining unless the replaced data on the large databases with a great percentage of Missing Values are very close to the original data and they do not alter the relationship between the variables. For this reason, we propose a new approach based on the theory of Copulas which involves estimating Missing Values in a manner to better reflect the uncertainty of data when the most significant knowledge is to be extracted.

The paper is organized as follows: basic concepts are presented in Section 2, Section 3 contains a description of the proposed method, and experimental results are given in Section 4, Section 5 concludes the paper.

2 Basic concepts

In this Section, we will introduce some basic concepts that will be used in the rest of the paper. These concepts include the inverse transform Sampling to generate random samples from a probability distribution given its cumulative distribution function (CDF). A range of family of Copulas will also be presented.

2.1 Inverse transform sampling

A classical approach for generating samples from a one dimensional CDF is the inverse transform sampling method. Given a continuous random variable X with a CDF $F(x) = P[X \leq x]$, the transform $U = F(X)$ results in a random variable U that is uniform in [0,1]. Moreover, if U has a uniform distribution in [0,1] and X is defined as $X = F^{-1}(U)$, where $F^{-1}(u) = inf\{x, F(x) \geq u\}$ and inf denotes the greatest lower bound of a set of real numbers, then the CDF of X is F(X). In order to generate an arbitrary large sample of X, we start with a uniformly distributed sample of U that can easily be generated using a standard pseudo-random generator. For each sampled value of U, we can calculate a value of X using the inverse CDF given by $X = F^{-1}(U)$. Figure 1 illustrates this method for the case of a Gaussian random variable.

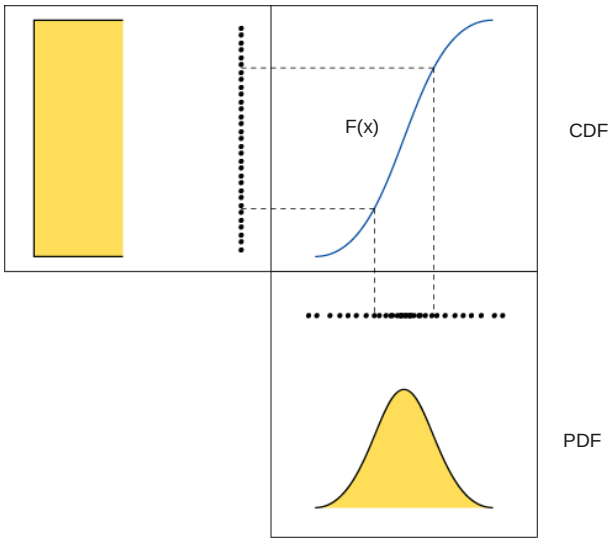

Fig. 1. *The inverse method to generate a sample from a Gaussian distribution.*

2.2 Definition and Illustration of Copulas

Datamining seeks to identify the knowledge of massive data and the importance of the dependence structure between the variables is essential. By adopting Copulas, Datamining can take advantage of this theory to construct multivariate probability distribution without constraint to specific types of marginal distributions of attributes, in order to predict Missing Values in large databases.

2.3 Definition

Formally, a Copula [23] is a joint distribution function whose margins are uniform on $[0, 1]$.
C $[0, 1]^m \mapsto [0, 1]$ is a Copula if $U_1...U_m$ are random variables which as uniformly distributed in [0,1] , such that [20]

$$C(u_1..., u_m) = p[U_1 \leq u_1, ..., U_m \leq U_m] \tag{1}$$

The most important theorem in the theory of Copulas is given by Sklar [22]
Let F be a distribution function with marginal distribution functions $F_1...F_m$. Then there exists a Copula such that $\forall(X_1....X_m) \in R^m$

$$F(x_1..., x_m) = C[F_1(x_1), ..., F_n(x_m)] \tag{2}$$

To illustrate the method of calculating a Copula from any sample, we consider the bivariate example in the Figure below

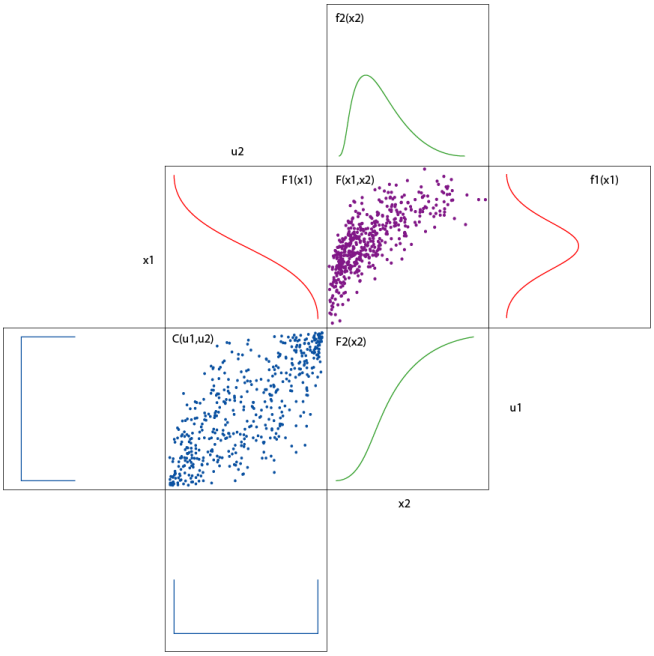

Fig. 2. *Generation of a sample of a Gaussian copula from a sample of a bivariate distribution that has as marginal distributions a Gamma and a Gaussian.*

The above example illustrates that Gaussian Copulas can model the dependency between random variables that do not necessarily follow a Gaussian distribution (Figure 2) we consider two random variables X_1 and X_2 and we assume that $f_1(x_1)$ is a Gaussian distribution and $f_2(x_2)$ is a Gamma distribution, follow as shown in the top right corner of Figure 2. Using the transformation $U_i = F(x_i)$, we obtain the corresponding sample in the space (u_1, u_2).The result of this transformation is the bivariate sample from the Gaussian Copula.

Empirical Copula

To evaluate the suitability of a chosen Copula with the estimated parameter, we can use the empirical Copula structure to model the observed dependence. Formally, the calculation of the empirical Copula is given by the following equation

$$(c_{ij}) = \frac{1}{n}(\sum_{k=1}^{n} I_{(v_{kj} \leq v_{ij})}) \tag{3}$$

wher n is the number of observations; v_{kj} is the value of the k^{th} row and j^{th} column; v_{ij} is the value of the i^{th} row and j^{th} column.

2.4 Family of Copulas

There is a wide range of family of Copulas

Gaussian Copula

Let $f_i(x_i)$ be standard Gaussian distributions, i.e. $X_i \sim N(0,1)$, and let Σ be the correlation matrix. The resulting Copula $C(u_1, u_2, ..., u_n)$ is called a Gaussian Copula. The density associated with $C(u_1, u_2, ..., u_n)$ is obtained using the following equation

$$C(\Phi(x_1)..., \Phi(x_m)) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp[\frac{-1}{2} X^t(\Sigma^{-1} - I)X] \tag{4}$$

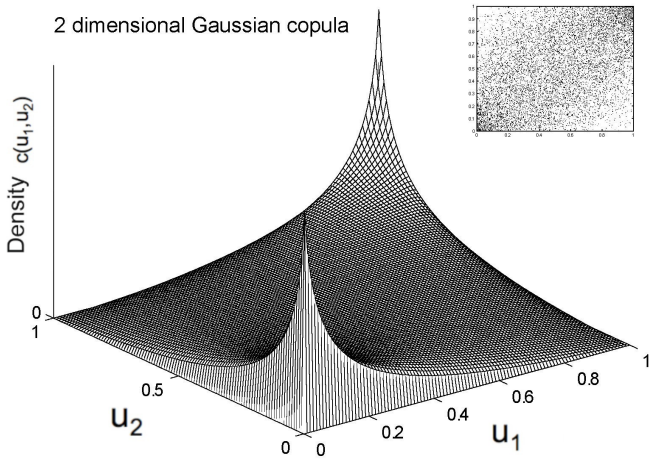

Fig. 3. *2-dimensional Gaussian Copula density resulting from a sample of a bivariate standard Gaussian distribution.*

where $X = (x_1, ..., x_n)$ and $\Phi(x)$ denotes the CDF of the standard Gaussian distribution. Similarly by using $U_i = \Phi(x_i)$ we can write

$$C(u_1..., u_m) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp[\frac{-1}{2}\xi^t(\Sigma^{-1} - I)\xi] \tag{5}$$

where $\xi = (\Phi^{-1}(u_1)....\Phi^{-1}(u_m)^T$.

Student Copula

The Student Copula is extracted from the multivariate t distribution which is given by the following : $\forall(u_1..., u_m) \in [0, 1]^m$

$$C(u_1..., u_m) = \frac{(f_{(v,\Sigma)}(t_v^{(-1)}u_1, ...t_v^{(-1)}u_1))}{(\prod_{i=1}^{(n)}(f_{(v)}t_v^{(u_i)}))} \tag{6}$$

where $t_v^{(-1)}$ is the inverse of the t distribution centered and reduced to univariate degrees of freedom.
$f_{(v,\Sigma)}$ is the probability density function of the Student distribution which is centered and reduced.
Σ is the correlation matrix and $f_{(v)}$ is the density unirange of the Student distribution, centered and reduced (Σ =1).

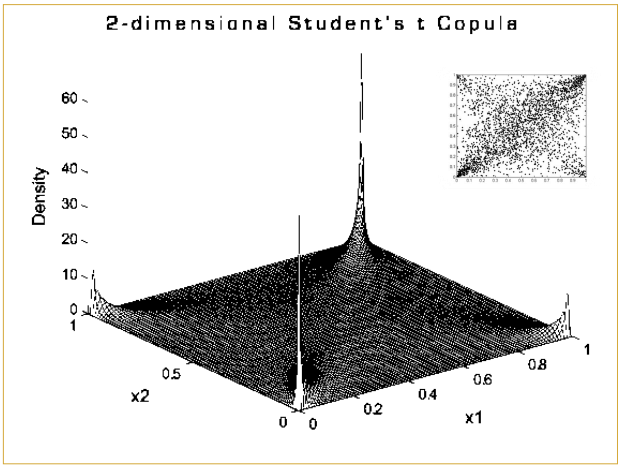

Fig. 4. *2-dimensional Student Copula density resulting from a sample of a bivariate standard Student distribution.*

Archimedean Copulas

The Archimedean Copulas are defined as follows

$$C(u_1...,u_n) = \begin{cases} \varphi^{-1}(\varphi(u_1 + ... + u_n) \text{ if } \sum_{\varphi(u_i)\leq\varphi(0)}, \\ 0 \qquad\qquad\qquad\qquad\quad \text{else} \end{cases}$$

φ is called the generating function that checks the Copula
$\varphi(1) = 0$, $\varphi(\acute{u}) < 0$ and $\acute{u}\varphi(\acute{u}) > 0$, $0 \leq u < 1$

In this table we have summarized some examples of Archimedean Copulas.

Table 1. Examples of Archimedean Copulas

Copules	$\varphi(u)$	C(u)
$\prod$	-ln u	$\prod_{i=1}^{d} u_i$
Gumbel	$(-lnu)^{\theta}, \theta \geq 1$	$exp\{-[\Sigma_{i=1}^{d}(-lnu)^{\theta}]^{1/\theta}\}$
Frank	$\frac{-lnexp((-\theta u)-1}{exp(-\theta)-1}$	$-\frac{1}{\theta}ln(1 + \frac{\prod_{i=1}^{d} exp((-\theta u_i)-1}{(exp(-\theta)-1)^{d-1}})$
Clayton	$u^{-\theta} - 1, \theta > 0$	$(\Sigma_{i=1}^{d} u^{-\theta} - d + 1)^{\frac{-1}{\theta}}$

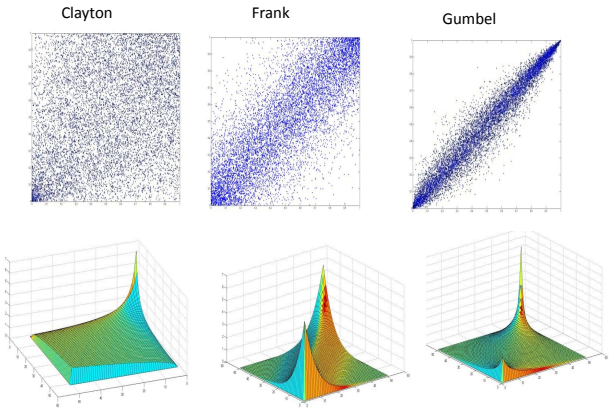

Fig. 5. *2-dimensional Archimedean Copulas : Clayton, Gumbel, Frank,*

3 Proposed approach

To illustrate our approach, we give an overview in the following flowchart

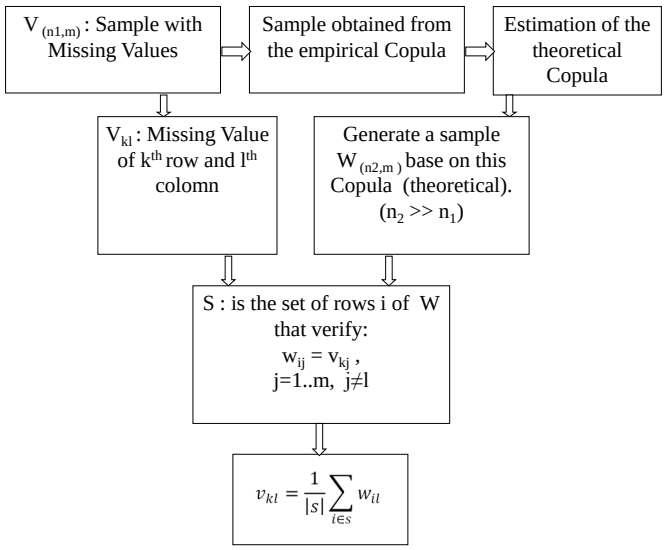

Fig. 6. *general outline of the proposed approach*

First, we calculate the empirical Copula from the sample containing Missing Values using formula (3) to better observe the dependence between the variables of these data.

According to the marginal distributions from the observed and approved empirical Copula, we can determine the theoretical Copula adjusted by the family of Copulas presented in Section 2, in order to generate the theoretical sample of millions of points, having the same distribution as the empirical sample, by computing the inverse CDF of each empirical marginal of the sample.

To estimate a Missing Value, we will determine a subset of rows in the theoretical Sample whose its variables are the same as the i^{th} row and k^{th} column of this Missing Value searched. At this time, we will calculate the average values found in order to acquire the Missing Values. Then, we calculate the standard deviation to estimate the accuracy of the mean and derive a confidence interval.

4 Results and discussion

Data source
To evaluate the effectiveness of our solution, we have developed a large-scale experiment on a based UCI machine learning repository server. This is a version of database generator waveform of 21 column and 33367 rows .

In the following we will give all the results obtained.
In our case study, we noticed that most of the Copula obtained have the form of scatter plot shown in Figure 7, For this we choose a bivariate Copula among 21 and we will illustrate all the results from a this Copula.

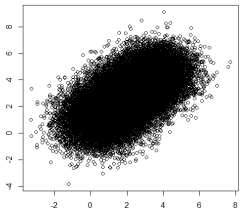

Fig. 7. The original sample with Missing Values for variables X^5, X^6

Generating an empirical sample
Given that the statistical model of the joint distribution is not known, we can calculate the empirical Copula of this sample to see if it follows a known Copula. Figure 8 shows the estimated Copula. This Copula is obtained by converting each point of the original sample by the cumulative distribution function of each marginal. The resulting Copula has an elliptical form just as a Gaussian Copula. To verify this formally, we used the fit test for Gaussian Copulas .

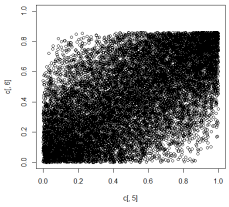

Fig. 8. The empirical sample from the empirical Copula of variables X^5, X^6

Generation of a larger theoretical sample
Above, we have shown that we have a Gaussian Copula. To generate a large sample from this Copula and with the same parameters, we can calculate the

inverse marginal CDFs to obtain a sample of large size with the same statistical model as the empirical sample. To calculate the inverse CDFs, we observed the marginal distributions of variables X^5, X^6. These distributions are Gaussian. They were validated by fit test classic such as univariate Kolmogorov-Smirnov. Figure 9 shows the theoretical sample obtained with a million points.

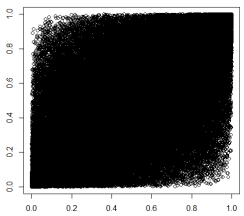

Fig. 9. The theoretical sample from the theoretical Gaussian Copula

The new sample obtained

To estimate a Missing Value, we have determined the subset of rows in the theoretical Sample (Figure 9) whose its variables are the same of the Missing Value searched. The average of these values has the value estimated.

We have repeated the same steps for all the Missing values and we have calculated standard deviation. We noticed that the new Sample obtained is Gaussian which is shown in the top left corner of Figure 10.

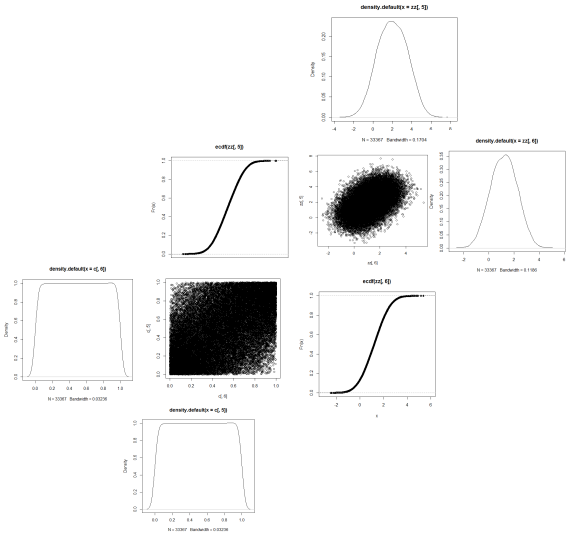

Fig. 10. The new sample obtained

To test the performance of our method, we calculated the error rate for the percentage of Missing Values, and we also applied the technique of imputation of the mean and that of regression for the same database in order to better see the advantage of our approach.

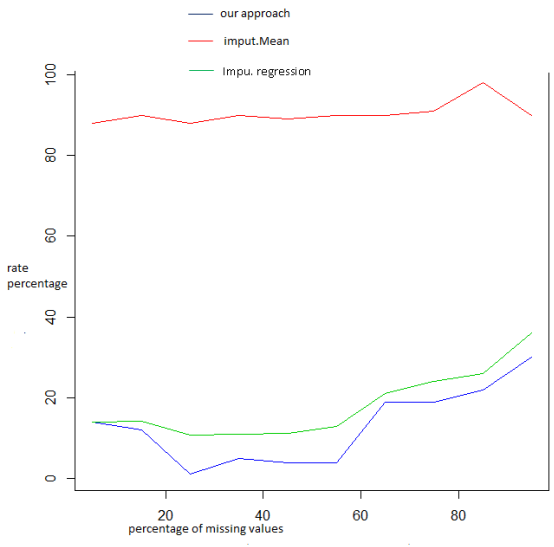

Fig. 11. *Performance of substitution techniques according to missing values*

Discussion

The comparison of the different graphs in Figure 11, shows the correspondance with the results obtained with the waveform database for the same evaluation criteria.

The increase in missing values by 5 % to 95 % caused a decrease in the minimum accuracy of 88 % for the mean imputation method, 14 % by the regression technique and a 1% by our method, for against the maximum error is 98 % for the imputation method, and 36 % by the regression technique and the average and 30 % our method.

Degradation of the error is always evident when increasing missing values. However our strategy (blue curve) based on Copula is much better than the mean imputation (red curve) and also superior to regression technique (green curve). The difference is most sensitive for all values.

Our approach should be a practical solution to estimate missing values for a very large database because it overcomes the Missing Values using Copulas

with small errors even with a very large percentage of missing values which is contrary to the mean imputation if is much less conclusive and may be better on small databases which is not the case in the data Mining.

5 CONCLUSION AND FUTURE WORK

Incomplete and noisy data, are a major obstacle in the pre-treatment process of KDD, those that lead to knowledge extraction from low quality and consequently the KDD process slows and the results that it provided are not reliable.

Within this context, we propose in this paper a new approach based Sampling techniques that essentially seeks to predict Missing Values, which constitute a major problem, because the information available is incomplete, less reliable and knowledge extracted from these data is not representative.
An experimental study on a large scale has provided very good results, those that show the effectiveness of our method.
Future work will focus on the application at this same theory to eliminate redundant data to reduce the size of large amounts of data.

References

1. A. Dempster, N. Laird et D. Rubin : Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1) 1-38, 1977.
2. Allison P. D , Multiple Imputation for Missing Data : A Cautionary Tale. Sociological Methods Research, 28(3), 301-309, 2000.
3. Chen J, Shao J, Nearest neighbor imputation for survey data, Journal of Official Statistics, Vol.16, No.2, pp.113-131,2000.
4. DeSarbo, W.S, Green, P.E, Carroll, J.D, Missing data in product-concept testing. Decision Sciences 17,163-185,1986.
5. Engels JM, Diehr P, Imputation of missing longitudinal data: a comparison of methods, Journal of Clinical Epidemiology, Volume 56 ,Issue 10, Pages 968-976 Statistical Association, 83, 1198-1202,2003.
6. G. Saporta, Probabilités, analyse des données et statistique, Editions Technip, Paris, 2006.
7. J.W, Some simple procedures for handling missing data in multivariate analysis. Psychometrika 41,409-415,1976.
8. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, edition Elsevier, new York,2006
9. Joe H, Multivariate Models and Dependence Concepts, Monographs on Statistics and Applied Probability,73, Chapmann et Hall, London,1997
10. Kline R.B, Principles and Practice of Structural Equation Modelling, Guilford Press, New York, 1989.
11. Kaufman, C.J, The application of logical imputation to household measurement, Journal of the Market Research Society 30, 453-466, 1988.

12. Kim, J.O, Curry, The treatment of missing data in multivariate analysis, Sociological Methods and Research 6, 215-241, 1977.
13. Little, R.J.A, and Rubin, D.B, Statistical Analysis with Missing Data,2002, New York: John Wiley and Sons, Inc , pp. 11-13.
14. Laird N.M, Missing data in longitudinal studies. Statistics in Medicine 7,1988, 305-315.
15. Lee, S.Y., Chiu, Y.M, Analysis of multivariate polychoric correlation models with incomplete data. British Journal of Mathematical and Statistical Psychology 43,1990, 145-154.
16. M. Hu, S.M. Salvucci et M.P. Cohen: Evaluation of some popular imputation algorithms, in Section on Survey Research Methods, pages 309-313, 2000. American Statistical Association.
17. Mélanie Glasson Cicognani , André Berchtold ,Imputation des données manquantes :Comparaison de différentes approches, J. Statist, Plann. Inference,,inria -00494698, version 1, 2010
18. Malhotra, N.K.,Analyzing marketing research data with incomplete information on the dependent variable. Journal of Marketing Research 24,1987, 74-84.
19. P.Deheuvels , La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance, Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5^{me} série.
20. Q. Song et M. Shepperd, A new imputation method for small software project data sets, Journal of Systems and Software, 80(1)51-62, 2007.
21. R.B.Nielsen, an introduction to copulas, second edition,springer,2005
22. L. Ruschendorf, On the distributional transform, Sklar's theorem, and the empirical copula process, J. Statist. Plann. Inference, 139(11); 39213927; 2009.
23. Roth, P.L, Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-560,1994.
24. Ruud, P.A,Extensions of estimation methods using the EM algorithm. Journal of Econometrics 49,305-341,1991.
25. Sinharay,Stern, H.S.Russell, The use of multiple imputation for the analysis of missing data, Psychological Methods, 6 (4), 317-329, 2001.
26. V. Barnett and T. Lewis. Outliers in statistical data. John Wiley et Sons, 1994.
27. Shichao Zhang, Parimputation From imputation and null-imputation to partially imputation, IEEE Intelligent Informatics Bulletin, Vol 9(1), 2008, 32-38.
28. Zhang S.C and al, Missing Value Imputation Based on Data Clustering. Transactions on Computational Science Journal, LNCS 4750, pp 128-138.
29. Z. Ghahramani et M.I. Jordan, Supervised learning from incomplete data via an EM approach, In J.D. Cowan, G. Tesauro et J. Alspector, éditeurs : Advances in Neural Information Processing Systems 6, pages 120-127, Morgan Kaufman, 1994.