Missing values and outliers

MIRI / MVA - Lab #1 Report

Authors: Cedric Bhihe <cedric.bhihe@gmail.com>
Santi Calvo <s.calvo93@gmail.com>

Delivery: before 2018.03.07 – 23:55

1. Impute missing values and save the result obtained.

The data set from Russet (1964) [1] consists of a heterogeneouos set of 9 continuous and categorical variables observed during the period 1945-1962, with a total of 4 missing values. The variables are observed in 47 countries. The distribution of missing values is best characterized by Fig. 1.

After inspection, there is no obvious pattern for the missing data's distribution and no readily observable correlation between the missing data and other variables' values. In the absence of a detailed statistical analysis to conclude with certainty on the matter, our working assumption is a missingness mechanism of type MCAR[i].

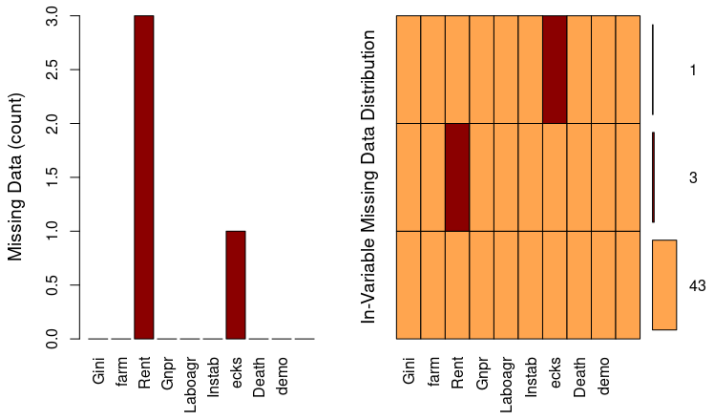


***Fig. 1**: Missing data count for each variable (left) and distribution of missing values among variables (right).*

The information in Fig. 1 is also available in tabular form, invoking either `` `md.pattern(<data_set>)` `` or `` `md.pairs(<data_set>)` `` in R. The latter method is useful when conducting a statistical analysis of the missingness mechanism: MCAR, MAR or MNAR.

No observation has more than 1 missing value as summarized in Table 1, where for each country col. "Index" exhibits the missing values' row numbers in the data set.

Index	Country	Nbr_missing	Var_missing
2	Australie	1	Rent
30	Nicaragua	1	Rent
31	Norvege	1	ecks
35	Péru	1	Rent

***Table 1**: Missing values distribution.*

Tables 2 summarizes imputed values according to the **kNN** and the **MICE**[ii] methods.

kNN results illustrate the fact that for small values of k, local "noise" (variability) may be captured when imputing a value from a small subset of nearest neighbors. Inversely for larger values of k, that local "noise" may be smeared out.

Country	Var_missing	k=1	k=3	k=5	k=7	MICE
Australie	Rent	22.3	34.7	22.3	20.4	14.6
Nicaragua	Rent	16.7	16.7	15.1	15.1	8.5
Norvege	ecks	0	0	0	0	4
Péru	Rent	15.1	15.1	16.7	15.1	2.4

***Table 2**:* Imputed values *for various numbers , k, of nearest neighbors, according to the Euclidian distance based kNN algorithm and to the MICE algorithm.*

The **Multivariate Imputation by Chained Equations** (MICE) [2] procedure was called with specific options for all numeric data variables in the data-set: predictive mean matching (pmm) imputation, simple random initial draw from the predictors' data range, a maximum number of iteration of 10, 8 data set imputation threads to be pooled, massive imputation, where each imputed variable uses all other variables as predictors.

Results for the kNN and MICE methods of imputation differ markedly.
They are saved in the directory `$(pwd)/Lab1/Report/` under `russet_imputed-values_knn-k5.txt` and `russet_imputed-values_mice-pmm.txt`.

---

i MCAR - Completely at random: missing values appear without any pattern. This is the most favorable situation, missing values just implies a reduction of the size, when no imputation is carried out.

ii MICE is a fully conditional specification (FCS) technique. It proceeds iteratively, starting from an incomplete data set, according to 3 steps: imputation, analysis and pooling. The term "chained equations" refers to the fact that MICE can be implemented as a concatenation of univariate value imputation procedures.

2. Outlier detection and analysis

The importance of detecting and ranking outliers in many high dimensional Data Mining applications is considerable [3]. For heterogeneous data (as in the present case) appropriate scoring functions are needed to produce a meaningful ranking, where parametric tests often fail to correctly represent the extent of outliers' deviation.

In the following results for the local outlier factor (LOF) approach, for the hierarchical-clustering-based outlier detection (HC) and for the Mahalanobis Classical Distance-based (MCD) detection are included. We used the data set resulting from our previous MICE imputation. All scores are rounded at least to the 3rd significant decimal.

Table 3 summarizes results for outlier ranking factors ordered, using hierarchical clustering (HC) with 2 different agglomeration methods. Results are remarkably different with three detected outliers (average) and only one (Ward.D).

Country	Cuba	USA	Bolivia	South-Vietnam
Outlyingness	0.957	0.957	0.911	0.911

Country	Cuba	Canada	USA	India
Outlyingness	0.941	0.692	0.692	0.600

***Table 3**: HC methods – outlyingness factors based on 2 agglomerative methods: (left) average and (right) Ward.D.*

Table 4 summarizes ranking for top most local outlier factors (LOF). The LOF ranking is sensitive to the number of nearest neighbors (k=3 or k=5) used to calculate local density around individual observations.

	USA	Canada	Cuba	South-Vietnam	Bolivia
LOF (k=3)	12.618	8.154	6.702	2.387	1.966
LOF (k=5)	5.174	2.612	7.784	2.702	1.944

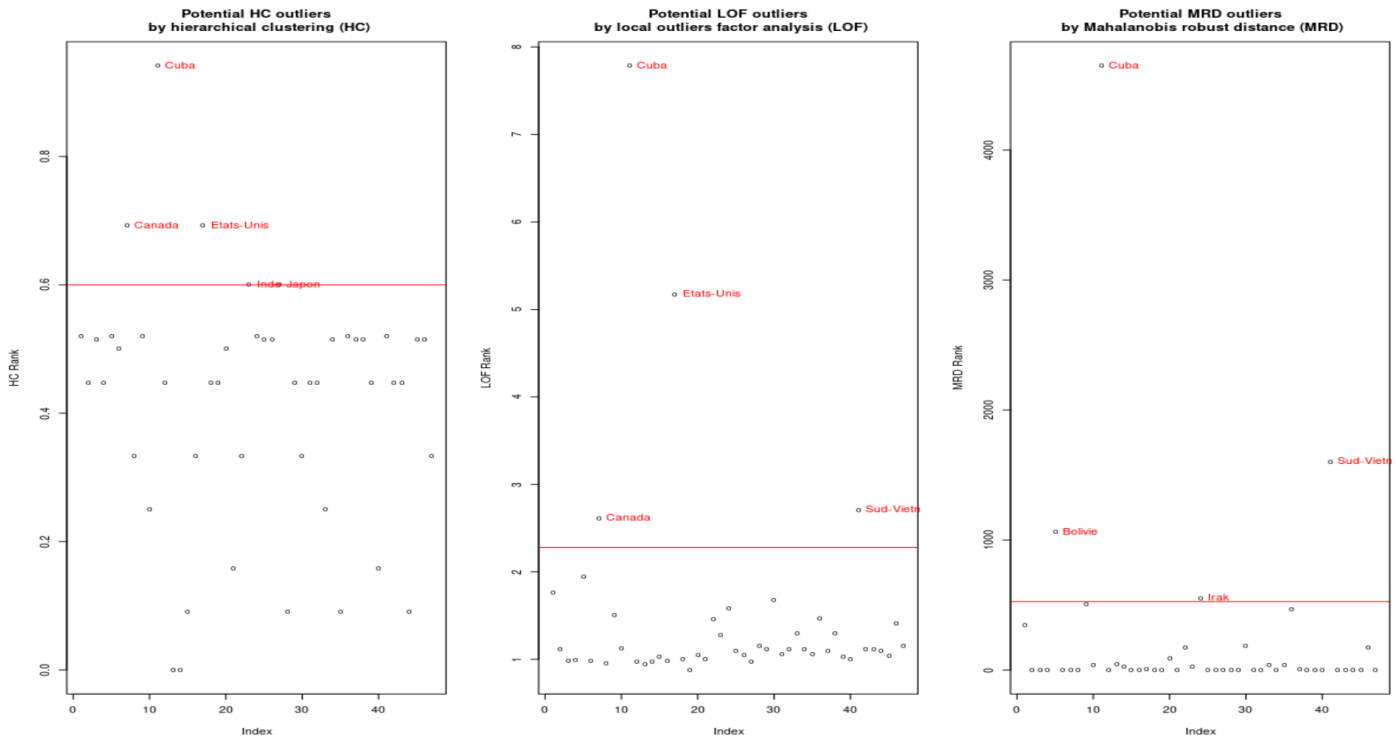
***Table 4**: LOF 5 top ranking for 3 and 5 neighbors respectively*

Table 5 summarizes MCD and MRD rankings using the Mahalanobis classical and robust distance respectively, calculated based on the "chemometrics" package in R. We used a quantile cutoff value of 0.975.

	Cuba	India	South-Vietnam	USA	Bolivia	Irak	Colombia
MCD	6.21	5.74	-	4.92	-	4.66	-
MRD	4646.84	-	1601.54	-	1060.35	548.98	504.69

***Table 5**: Mahalanobis central (MCD) and robust (MRD) distance based ranks for values greater than the cutoff and for the 5 top most potential outliers respectively.*

Finally we summarize result graphically according to the three ranking methods presented above (HC, LOF and MRD).



References

[1] B. M. Russet, “Inequality and Instability”, World politics, 21 (1964), 442-454.
[2] S. van Buuren and C. G. M. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *J. Stat. Soft.*, vol. 45, no. 3 (2011), 67 pages.
[3] E. Muller, I. Assent, U. Steinhausen, T. Seidl, “OutRank: ranking outliers in high dimensional data,”