Robustness under missing data

BAUM Carole Department of Mathematics, University of Liège, Belgium carole.baum@uliege.be

Joint work with VAN MESSEM Arnout

Many **imputation techniques** have been introduced to handle missing data. However, a large scale comparison of these techniques seems to be missing and it is relatively unknown how the different imputation techniques perform when outliers are present in the data. Therefore, our goals are to:

❶ **compare their performance** under different settings; ❷ **investigate their robustness** when outliers are added.

Experimental setup

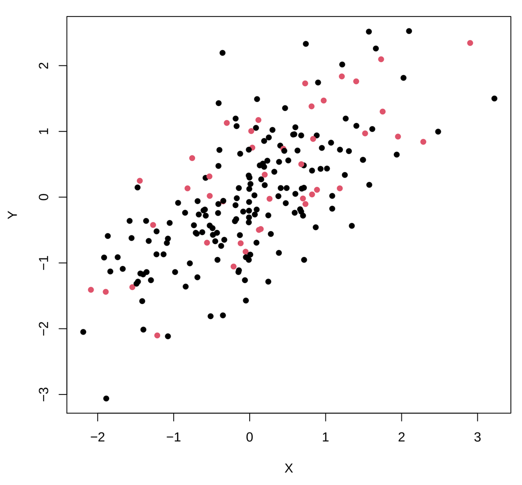
Data creation

- According to a Gaussian distribution
- Independent or dependent variables
- In multiple dimensions (2/3/20)

Introducing missing values

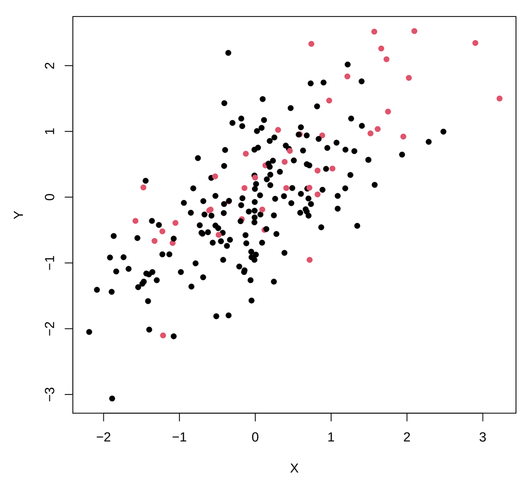
Missing values are created using multivariate amputation [1].

MCAR



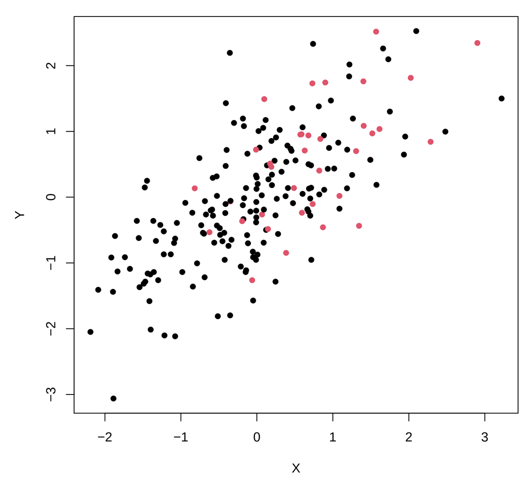
(Missing completely at random)

MAR



(Missing at random)

MNAR

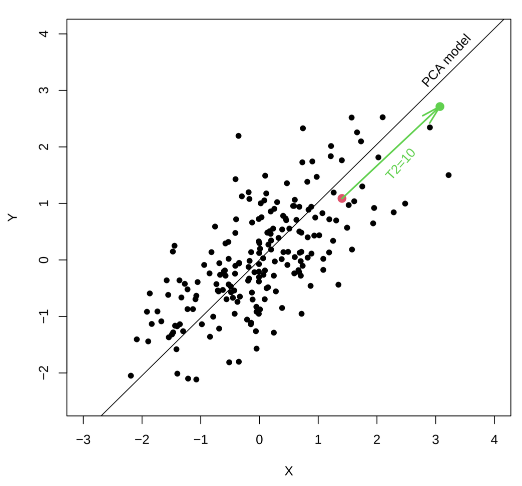


(Missing not at random)

Introducing outliers

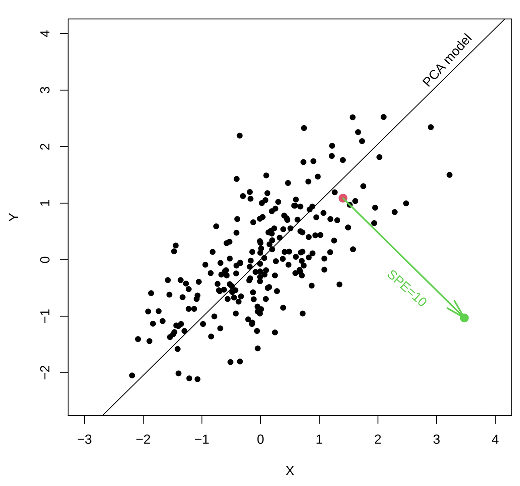
Outliers are generated by modifying either the Square Prediction Error (SPE), the Hotteling's T^2 statistic, or both [2].

Type 1



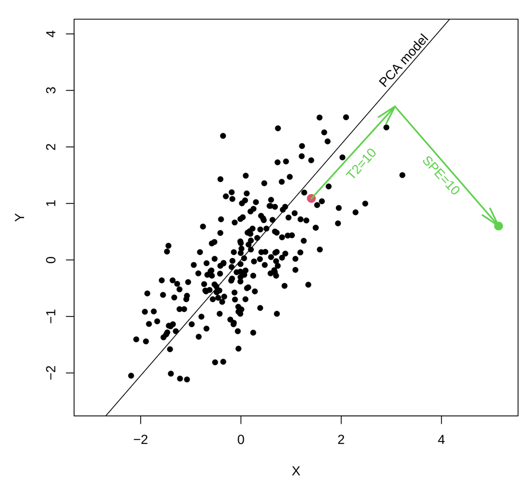
Good leverage - Outside the range

Type 2



Bad leverage - Inside the range

Type 3



Bad leverage - Outside the range

Imputation techniques

- Mean
- Expectation maximisation (EM)
- Linear regression (LR)
- Robust linear regression (RLR)
- Random forest (RF)
- K nearest neighbours (KNN)

Evaluation

Simulation 1: Without outliers

- Computational time
- Root mean squared error

$$RMSE = \sqrt{\frac{1}{n}\sum_{(i,j)\in I}(x_{i,j}-\hat{x}_{i,j})^2}$$

I is the set of indices of the missing values, $n = \#I$ is the number of missing values, $x_{i,j}$ is the original value in the complete data set at position (i,j) and $\hat{x}_{i,j}$ is the imputed value at position (i,j)

- Impact on statistics in 2/3D (mean, median, standard deviation, IQR, correlation)

Simulation 2: With outliers

- Difference in RMSE between the imputation with and without outliers

Simulation procedure

❶ Generate a complete data set → `data1`

❷ Ampute the complete data set by deleting a selection of values

❸ Impute the missing values to obtain a full data set → `data2`

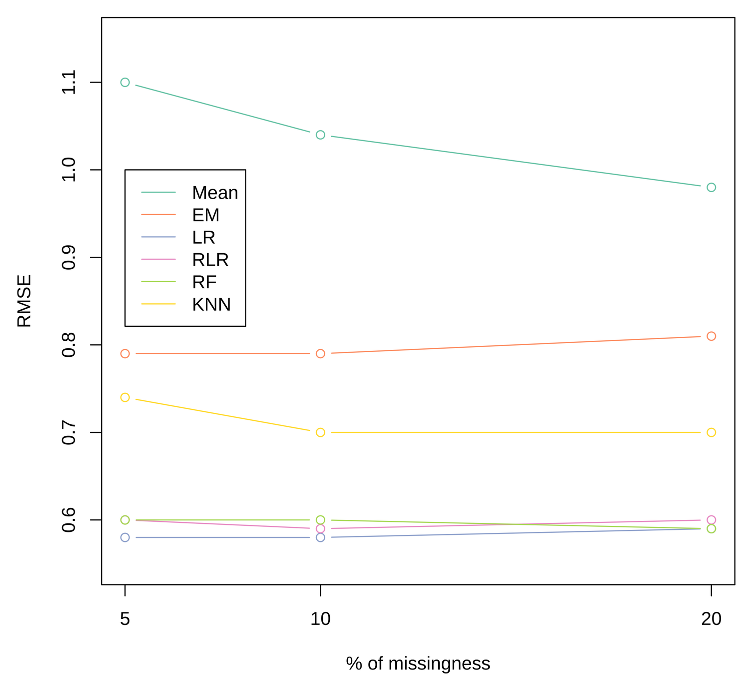
❹ *Introduce outliers in* `data1` *and ampute this data set*

❺ *Impute the missing values to obtain a full data set* → `data3`

❻ Two comparisons: `data1` ↔ `data2` and `data1` ↔ `data3`

Simulation 1: Without outliers

Results



Gaussian dependent variables, 20D, MNAR

Conclusions in 2D

- Decrease in central tendency
- Decrease in dispersion
- Increase in correlation for linear methods (LR, RLR)

Conclusions in 20D

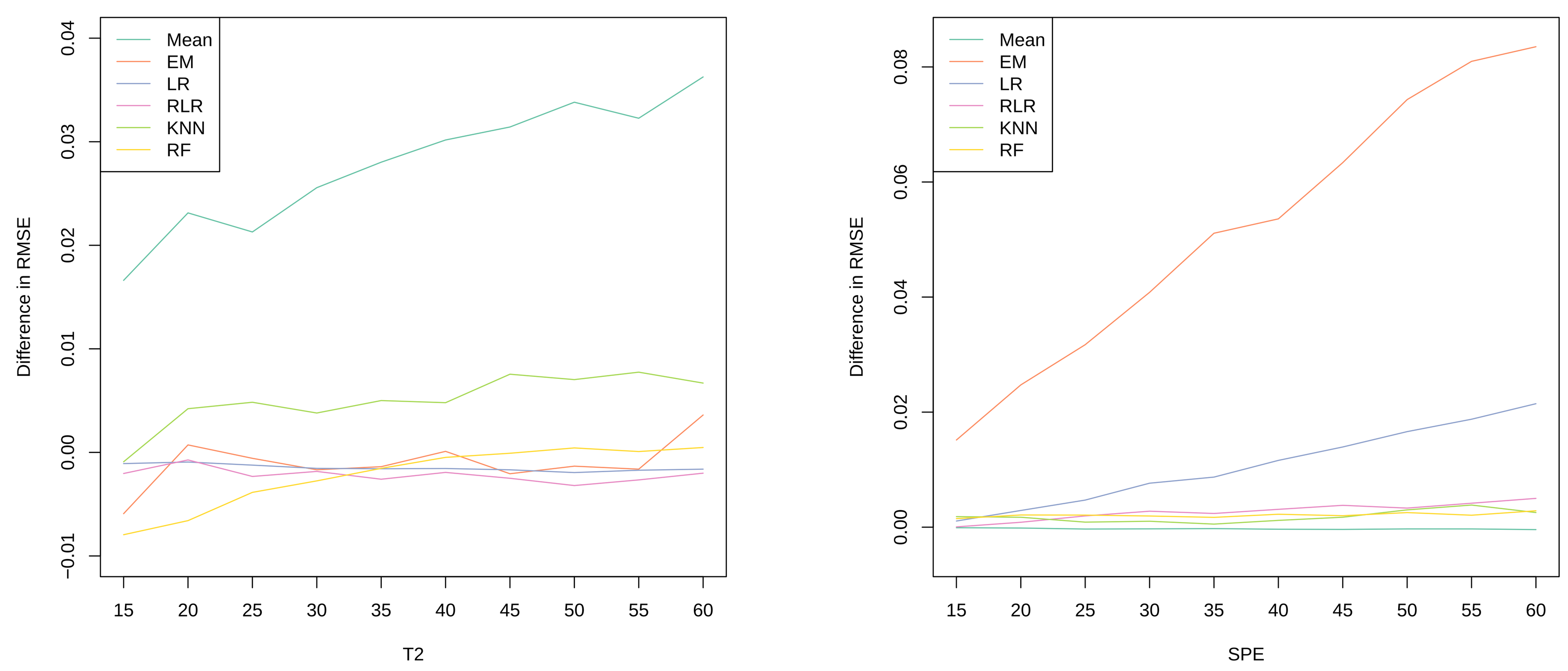
- Smallest RMSE: LR, RLR, RF
- Fastest methods: Mean, EM
- Slowest method: RLR

⋆ Similar results hold for other dimensions and all 3 missingness patterns

⋆ Independent variables: mean method also has small RMSE

Simulation 2: With outliers

Results



Gaussian dependent variables, 20D, MNAR

Conclusions

- Most robust methods: RLR, RF and KNN
- Robustness does not change when the % of missing values/outliers varies

⋆ Similar results hold for other dimensions and all 3 missingness patterns

Ongoing work

- Investigate the evolution of the mean vector and the covariance matrix after imputation with outliers (in 20D)
- Create linear or PCA models on both imputed data sets, i.e., with or without outliers, and compare the coefficients
- Compare imputations with and without outliers through binary classification

References:

[1] Schouten, Rianne Margaretha and Lugtig, Peter and Vink, Gerko (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909-2930.

[2] González-Cebrián, Alba and Arteaga, Francisco and Folch-Fortuny, Abel and Ferrer, Alberto (2021). How to simulate outliers with the desired properties. *Chemometrics and Intelligent Laboratory Systems*, 212:104301