

# Missing data analysis

---

**University College London, 2015**

# Contents

---

1. Introduction
2. Missing-data mechanisms
3. Missing-data methods that discard data
4. Simple approaches that retain all the data
5. RIBG
6. Conclusion

# Introduction

- Databases are often corrupted by missing values
- Most data mining algorithms cannot be immediately applied to incomplete data
- The simplest method to deal with missing data is data reduction which deletes the instances with missing values. However it will lead to great information loss.

# Why are data missing

- Random error
  - Someone forgot to write down a number, to fill in a questionnaire item, etc.
- Systematic bias
  - Certain types of people didn't want or couldn't or preferred not to answer certain types of questions

## Basic notions

- Let  $D$  denote an incomplete dataset with  $r$  variables  $D = \{A_1, A_2, \dots, A_r\}$  and  $n$  instances. For each variable  $A_j = \{A_j^{obs}, A_j^{mis}\}$ .

The entire dataset consists also of two components:

$$D = \{D^{obs}, D^{mis}\}$$

Let's introduce a response indicator matrix

$$R_{ij} = \begin{cases} 0 & \text{if } v_{ij} \text{ is missing} \\ 1 & \text{if } v_{ij} \text{ is observed} \end{cases}$$

## Types of missing data mechanisms (Rubin)

- *Missing Completely At Random (MCAR)*

If  $\Pr(R|D^{\text{mis}}, D^{\text{obs}}) = \Pr(R)$ . It implies that the missingness is unrelated to both missing and observed values in the dataset.

- *Missing At Random (MAR)*

If  $\Pr(R|D^{\text{mis}}, D^{\text{obs}}) = \Pr(R|D^{\text{obs}})$ . It means that the missingness depends only on observed values.

- *Not Missing At Random (NMAR)*

If  $\Pr(R|D^{\text{mis}}, D^{\text{obs}})$  is not equal to  $\Pr(R|D^{\text{obs}})$  and depends on  $D^{\text{mis}}$ .

# Missing-data methods that discard data

- Complete-case analysis
  - excluding all units for which the outcome or any of the inputs are missing

Problems with this approach:

- if the units with missing values differ systematically from the completely observed cases, this could bias the complete-case analysis.
- if many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a sample analysis.

# Missing-data methods that discard data

- Available-case analysis
  - study of different aspects of a problem with different subsets of the data.

Example: in the 2001 Social Indicators Survey, all 1501 respondents stated their education level, but 16% refused to state their earnings. This allow summarizing the distribution of education levels using all the responses and the distribution of earnings using 84% of respondents who answered the question.

## Problems with this approach:

- different analyses will be based on different subsets of the data and may not be consistent with each other
- if non-respondents differ systematically from the respondents, this will bias the available-case summaries.

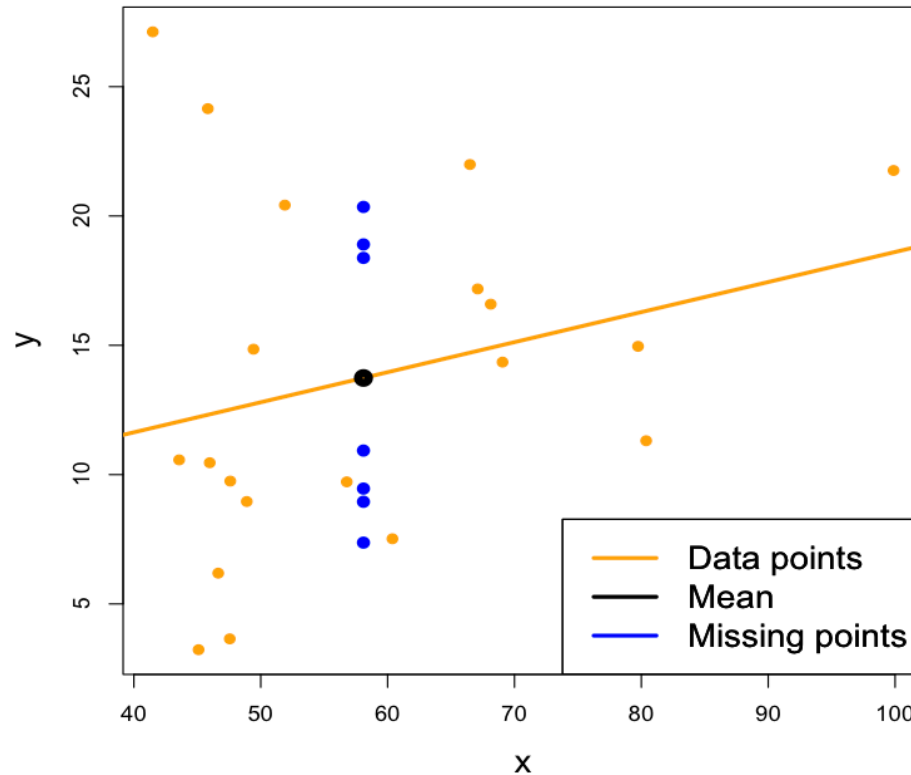
# Approaches that retain the data

- Mean substitution
  - replacing the missing values by the mean of all observed values at the same variable

## Problems with this approach:

- if the units with missing values differ systematically from the completely observed cases, this could bias the complete-case analysis.
- if many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a sample analysis.

# Mean substitution



- Regression line always pass through the mean of X and the mean of Y
- Missing values of X can be placed at the mean of X without affecting the slope of the line

# Mean substitution

## Advantages:

- All subjects have data for all values

## Disadvantages

- False impression of N
- Variance decreases
- What if data are missing for a reason?

# Approaches that retain the data

- Hot deck imputation
  - replacing missing values with values from a “similar” responding unit. Usually used in data from surveys. Involves replacing missing values of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to characteristics observed by both cases.

## Types of HTD:

- random hot deck methods (donor is selected randomly from a set of potential donors)
- deterministic hot deck methods (single donor is identified and values are imputed from that case, “nearest” in some sense)

## Other imputation methods

- *Regression imputation.* It uses regression models (different forms of them) to predict missing values.

### Package “VIM”

- *EM imputation.* It uses the iterative procedure of Expectation-Maximization algorithm to calculate the sufficient statistics. Missing values will be produced in the process.

# Amelia

Expectation-Maximization Bootstrap-based algorithm (EMB)

It assumes that the complete data are multivariate normal

Advantages:

- fast
- can deal with time-series data
- never crashes (according to official description)

AMELIA II: A Program for Missing Data

James Honaker, Gary King, and Matthew Blackwell

Version 1.7.3  
November 14, 2014

## Approaches that retain the data

- *Multiple imputation.* First proposed by Rubin way to handle missing data. It produces  $m$  complete datasets and then each of them is analyzed by complete-data method. At last the results derived from these  $m$  datasets are combined.

### Package ‘mi’

April 16, 2015

**Type** Package

**Title** Missing Data Imputation and Model Checking

**Version** 1.0

**Date** 2015-04-16

# Multiple imputation

Basic steps:

1. Make a model that predict every missing data item (linear or logistic regression, non-linear models, etc.)
2. Use the above models to create a “complete” dataset.
3. Each time a “complete” dataset is created, do an analysis of it, keeping the mean and SE of each parameter of interest.
4. Repeat this between 2 and tens of thousands of time
5. To form final inferences, for each repetition, average across means, and sum the within and between variances for each parameter.

R package: “mi”

## Machine learning-based imputation

- *Machine-learning-based approach.* Decision tree approach, clustering procedures, k-nearest neighbors approach and other can be used to fill in the missing data.

Example: function “impute.knn” from package “impute”

```
data(mtcars); mtcars<-as.matrix(mtcars[,c(1,3:7)]); mtcars_imp<- mtcars; mis_level<- 0.3
x1<- sample(1:length(mtcars[,1]), round(length(mtcars[,1])*mis_level), replace=F)
x2<- sample(1:length(mtcars[,1]), round(length(mtcars[,1])*mis_level), replace=F)
mtcars_imp[x1, 2]<- NA; mtcars_imp[x2, 5]<- NA
knn_res=rep(0,length(mtcars[,1])) #k-nearest neighbours
for (i in 1:length(mtcars[,1]))
  {knn<- impute.knn(mtcars_imp,k=i)
   knn_res[i]=sqrt(sum((mtcars[x1,2]-knn$data[x1,2])^2, (mtcars[x2,5]-knn$data[x2,5])^2))
  /sum(length(x1), length(x2))}
am=amelia(mtcars_imp, k=5) #Amelia
amelia_imp=(am$imputations$imp1+am$imputations$imp2+am$imputations$imp3+am$imputations$imp4+am$imputations$imp5)/5
amelia_res=sqrt(sum((mtcars[x1,2]-amelia_imp[x1,2])^2, (mtcars[x2,5]-amelia_imp[x2,5])^2)) /sum(length(x1),
length(x2))
mult_imp=mi(missing_data.frame(mtcars_imp), n.chains=5) #Multiple Imputation
mi_imp=(complete(mult_imp)[[1]][,1:6]+complete(mult_imp)[[2]][,1:6]+complete(mult_imp)[[3]][,1:6]+complete(mult_imp)[[4]][,1:6]+complete(mult_imp)[[5]][,1:6])/5
mi_res=sqrt(sum((mtcars[x1,2]-mi_imp[x1,2])^2, (mtcars[x2,5]-mi_imp[x2,5])^2)) /sum(length(x1), length(x2))
imp1=regressionImp(dis~mpg+hp+drat+qsec, data=mtcars_imp) #Regression
imp2=regressionImp(wt~mpg+hp+drat+qsec, data=mtcars_imp)
reg_imp=cbind(mtcars_imp[,1],imp1$dis, mtcars_imp[,3:4],imp2$wt,mtcars_imp[,6])
reg_res=sqrt(sum((mtcars[x1,2]-reg_imp[x1,2])^2, (mtcars[x2,5]-reg_imp[x2,5])^2)) /sum(length(x1), length(x2))
knn_res; amelia_res; mi_res; reg_res
```

## GMDH algorithm

- Group Method of Data Handling is an inductive method that constructs a hierarchical (multi-layered) network structure to identify complex input-output functional relationship from data.
- The process of GMDH is based on sorting-out of gradually complicated models and selection of the best solution by external criterion.

## RIBG (robust imputation based on GMDH) algorithm

- The main idea of RIBG is using the mechanism GMDH to impute missing data even when data contain noise.

- Let's consider an incomplete dataset

$$D = \{A_1, A_2, \dots, A_r\}$$

- First RIBG will fill in the original dataset by simple mean imputation to get an initial complete dataset.
- Then the GMDH mechanism will be used to predict and update these initial estimated missing values with an iterative process.

## RIBG criterion

- The criterion is introduced which integrates the systematic regularity criterion (SR) and minimum bias criterion (MB):

$$RM = SR + MB =$$

$$= \left\{ \left( \sum_{i \in B} (y_i - \hat{y}_i^C)^2 + \sum_{i \in C} (y_i - \hat{y}_i^B)^2 \right) \right\} + \sum_{i \in B \cup C} (\hat{y}_i^B - \hat{y}_i^C)^2$$

$B, C$  - two disjoint subsets,  $B \cup C = D$

$\hat{y}_i^B, \hat{y}_i^C$  - estimated outputs of the model

## Algorithm RIBG

### Input:

$D$ — $n \times r$  incomplete dataset

### Output:

$\hat{D}$ — $n \times r$  complete dataset

- Step 1: Divide dataset  $D$  into two disjoint subsets:  $D = B \cup C$ , where  $B$  is the training set and  $C$  is the validation set;
- Step 2: For each variable in  $D$ , replace missing elements by mean (if the variable is a numeric variable) or mode (if the variable is nominal) of the observed elements to get an initial complete dataset;

- Step 3: Select variable  $A_j$  that needs to be imputed as output variable ( $y = A_j$ ) and all the remaining variables as input variables ( $x = \{A_s | s = 1, \dots, r, s \neq j\}$ ) to enter the first layer of the GMDH network;
- Step 4: Combine input variables in pairs  $(x_i, x_j)$ ,  $1 \leq i, j \leq m$  and generate model candidates from each combination using the following quadratic polynomial:

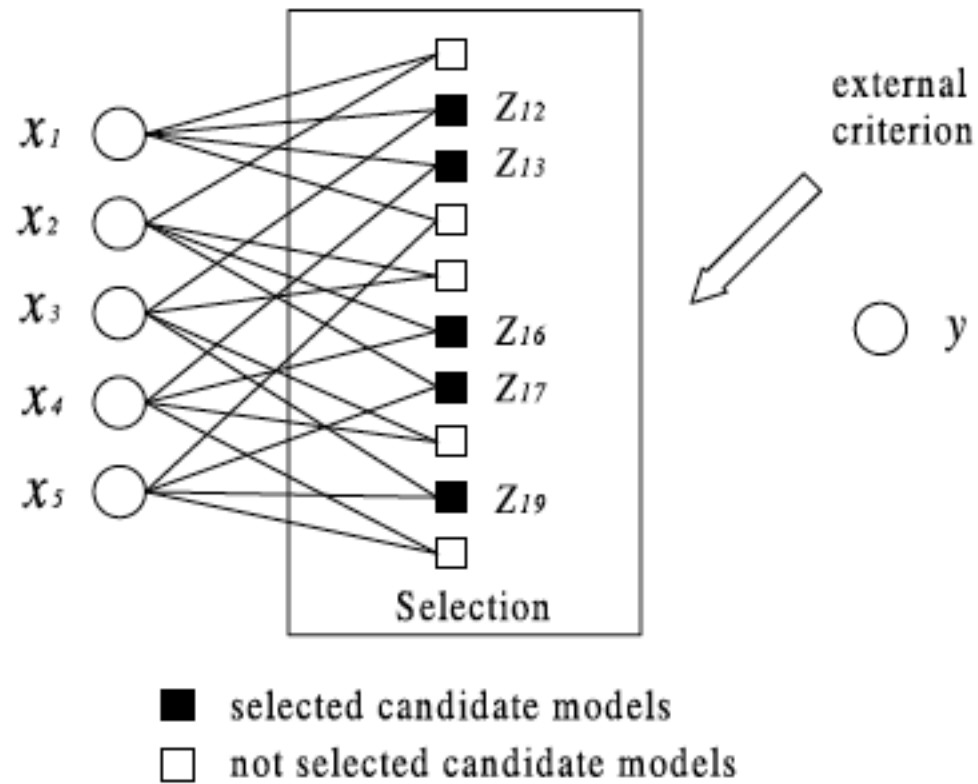
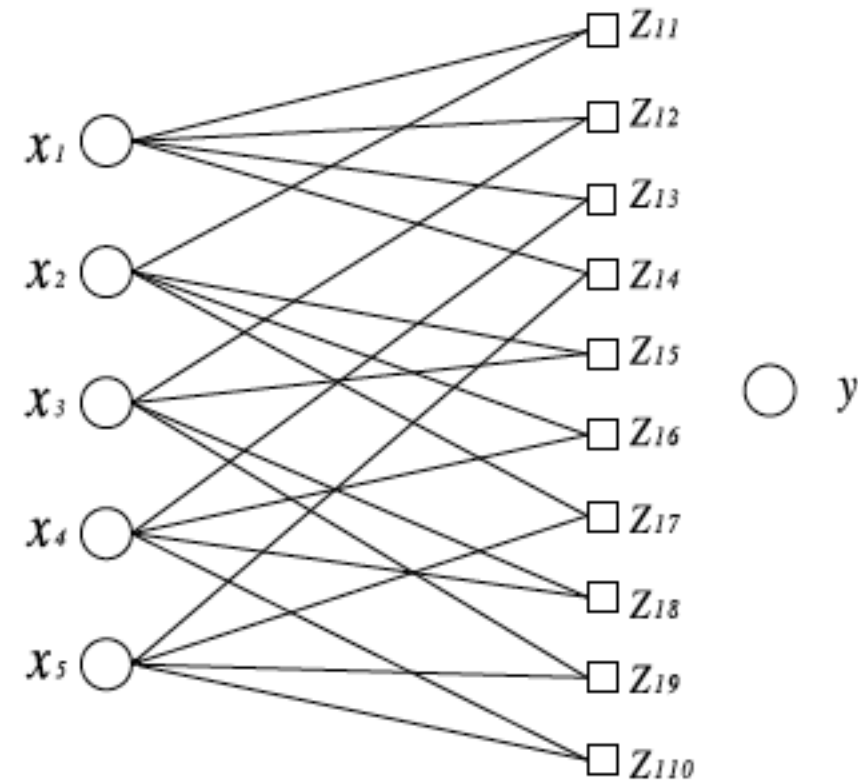
$$y = c_0 + c_1x_i + c_2x_j + c_3x_i \cdot x_j + c_4x_i^2 + c_5x_j^2 \quad (3)$$

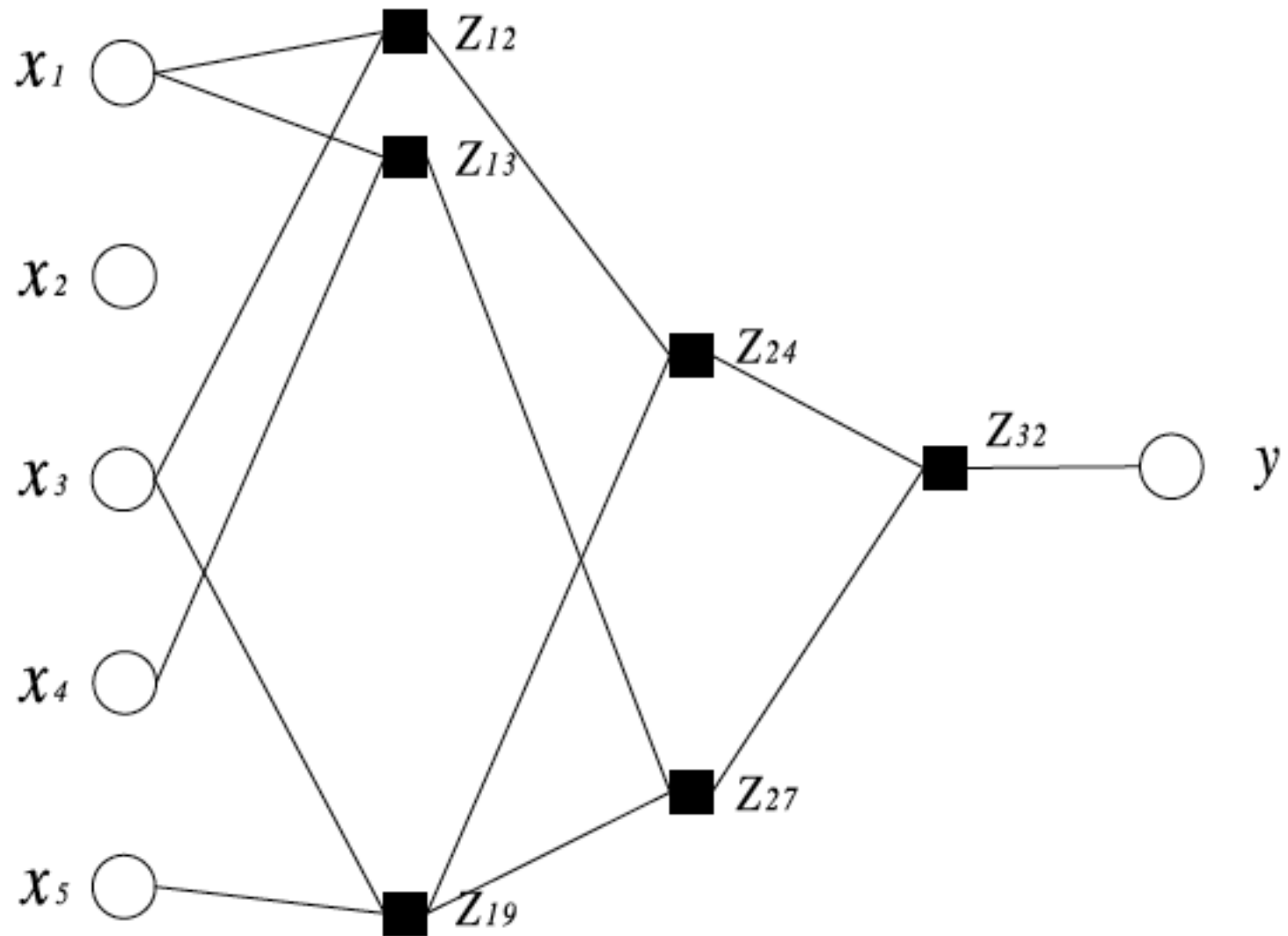
where  $c_0, c_1, \dots, c_5$  are parameters to be estimated by the ordinary least square (OLS) method. For example, if we use  $x_1, x_2, \dots, x_5$  as the input variables to estimate the output variable  $y$ , then 10 model candidates are produced in Fig. 1 and input variables  $x_1$  and  $x_2$  will be combined to produce the model candidate  $Z_{11}$  as follows:

$$Z_{11} = c_0 + c_1x_1 + c_2x_2 + c_3x_1 \cdot x_2 + c_4x_1^2 + c_5x_2^2 \quad (4)$$

- Step 5: Evaluate the external criterion of each model using the combined criterion  $RM$ . Record the minimum of the external criterion  $R_i$  from the current layer. Select  $F_i$  best models with lower criterion values and their outputs  $z_{ti}$  are employed as new input variables for the second layer of the GMDH network ( $x_{it} = z_{it}, t = 1, \dots, F_i$ ). For instance, model candidates which have lower external criterion values  $z_{1t}$  ( $t = 2, 3, 6, 7, 9$ ) are selected and used as input variables for the second layer in Fig. 2;
- Step 6: Repeat Steps 4–5 to produce model candidates of the second layer, the third layer, ... until the lowest value of external criterion at the current layer  $R_i$  is greater than that in previous layer. The model with the minimum external criterion at the  $i-1$  layer is then selected as the final optimal complex model. Figure 3 gives an example that the optimal complex model  $z_{32}$  is obtained at the third layer;

- Step 7: Use the corresponding estimates of the optimal complex model to update the missing elements  $A_j^{mis}$  of variable  $A_j$ ;
- Step 8: Repeat Steps 3–7 until the change of missing element estimates  $\hat{A}_j^{mis}$  becomes smaller than a threshold or maximum number of iterations is reached. Then assign current values of  $A_j$  to the corresponding elements in  $\hat{D}$ ;
- Step 9: Follow Steps 3–8 to update the missing values of remaining variables.





# Simulations

Data sets:

- Housing (economics)
- Breast (medical science)
- Bupa, Cmc, Iris (life sciences)
- Glass2, Ionosphere, Wine (physics)

## Missingness and noise

Levels of missing rate: 5%, 10%, 20%

Levels of noise ( $\delta$ ): 0%, 10%, 20%

Every value at each variable had a ( $\delta$ ) chance to be changed to any other random value

## Methods to compare

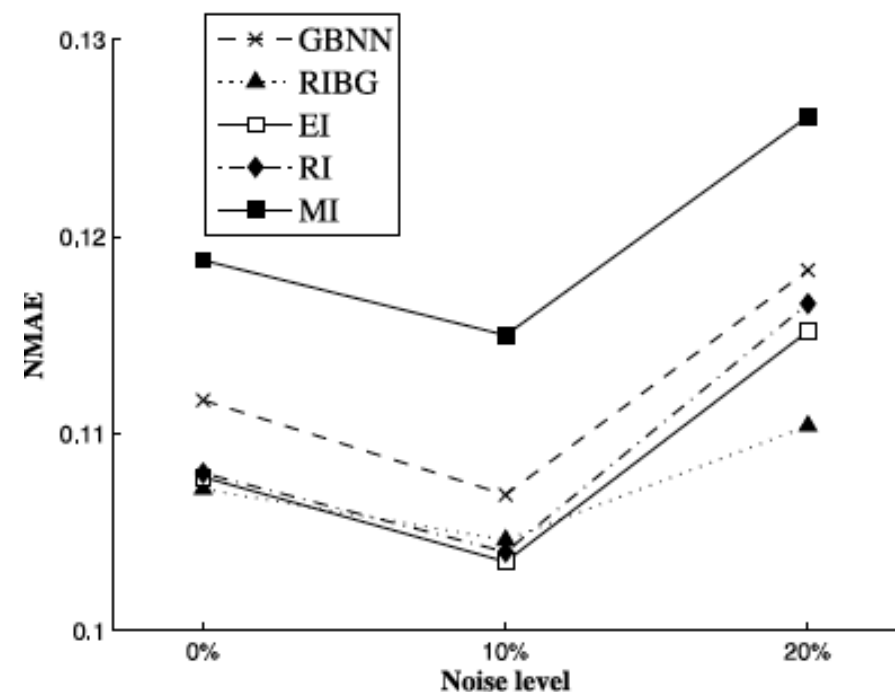
- Regression imputation
- EM imputation
- GBNN imputation (based on knn method)
- Multiple imputation

## Performance measure

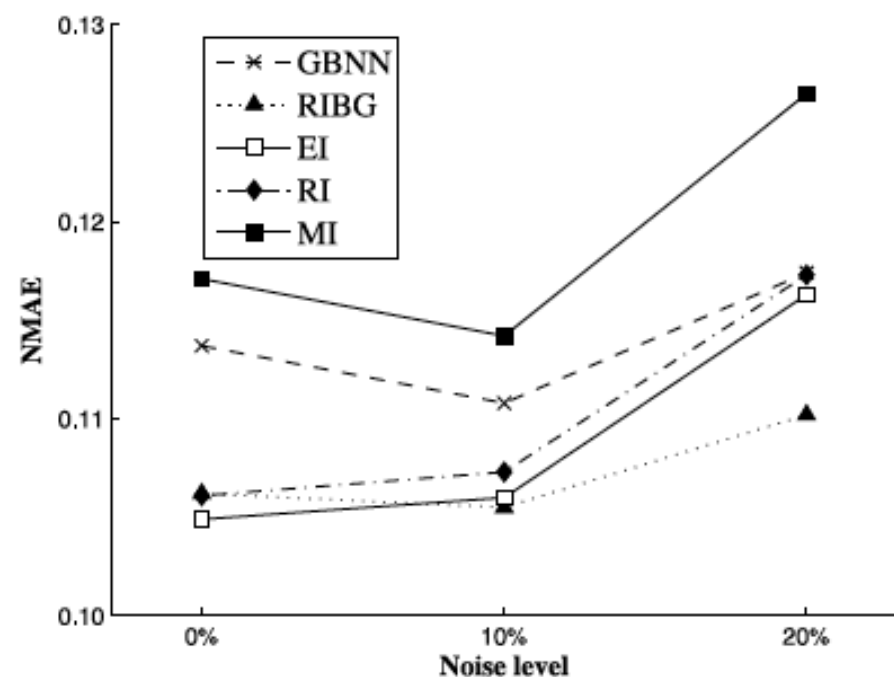
$$NMAE_j = \begin{cases} \frac{1}{n_{mis}^j} \sum_{i=1}^{n_{mis}^j} \left( \frac{\hat{v}_{ij} - v_{ij}}{v_j^{\max} - v_j^{\min}} \right) & \text{if variable is numerical} \\ 1 - \frac{n_j^{cor}}{n_j^{mis}} & \text{if variable is nominal} \end{cases}$$

$n_j^{mis}$  - number of missing values;  $v_{ij}, \hat{v}_{ij}$  - true and imputed values;  $v_j^{\max}, v_j^{\min}$  - maximum and minimum for this variable;

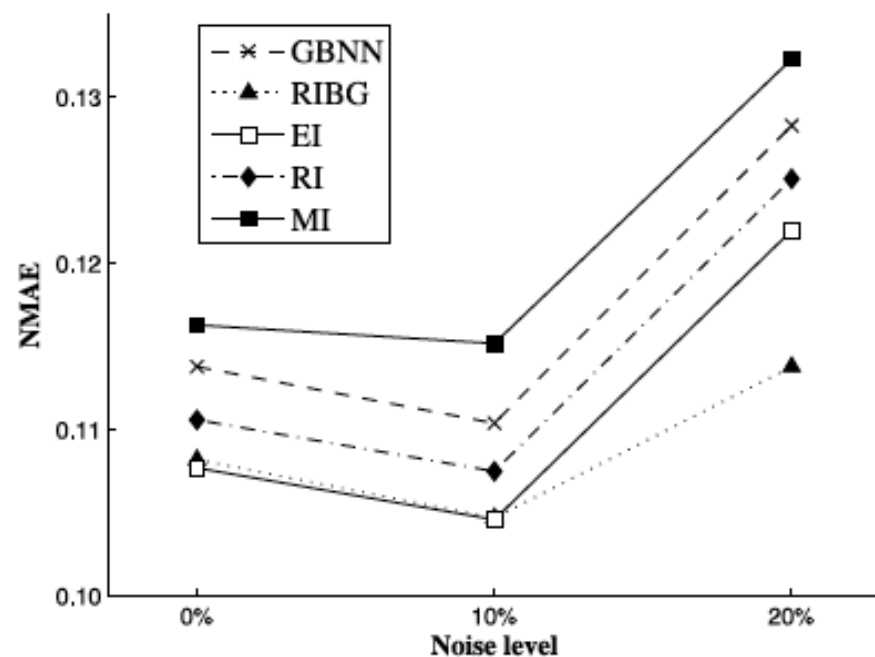
$n_j^{cor}$  - number of correctly predicted nominal values



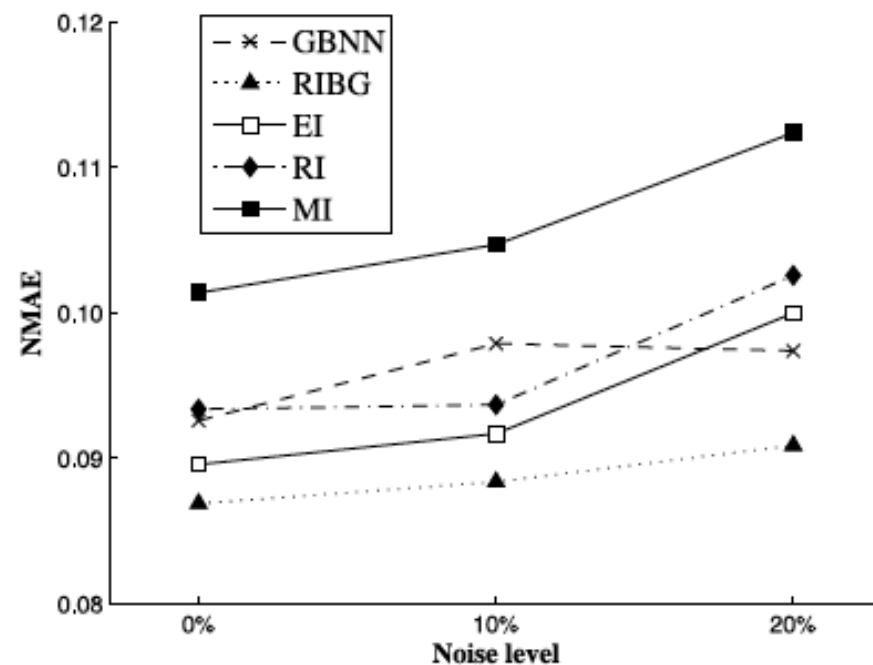
(a) Mechanism=MCAR, Missing rate=5%



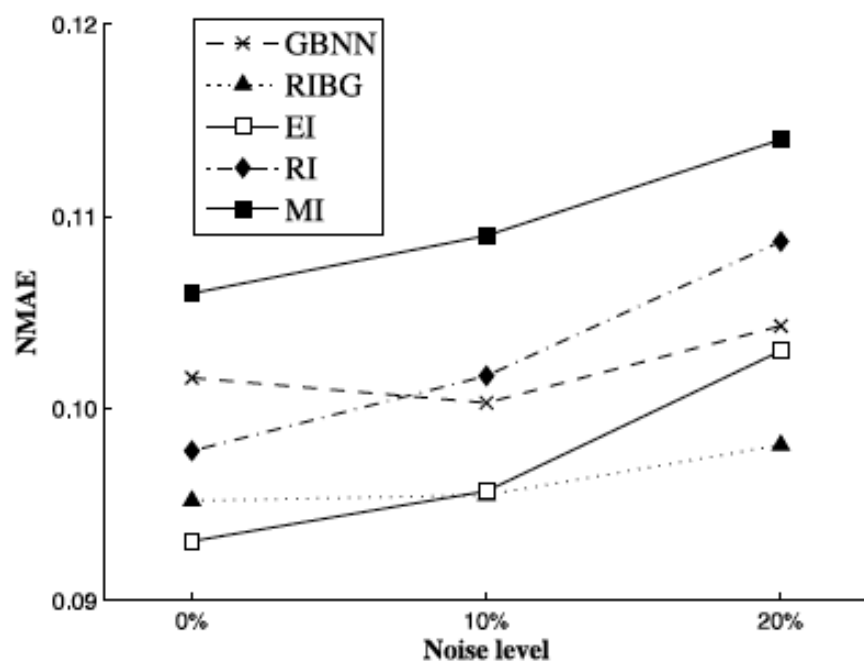
(b) Mechanism=MCAR, Missing rate=10%



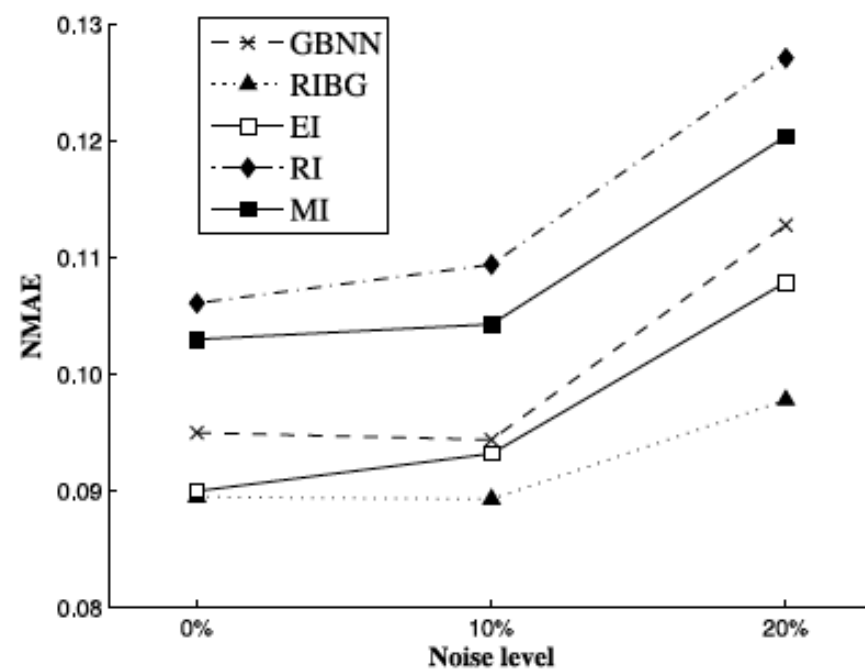
(c) Mechanism=MCAR, Missing rate=20%



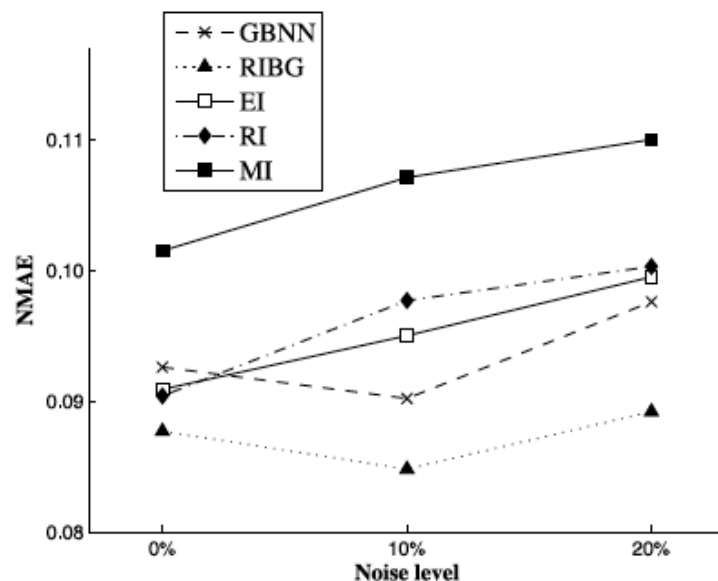
(d) Mechanism=MAR, Missing rate=5%



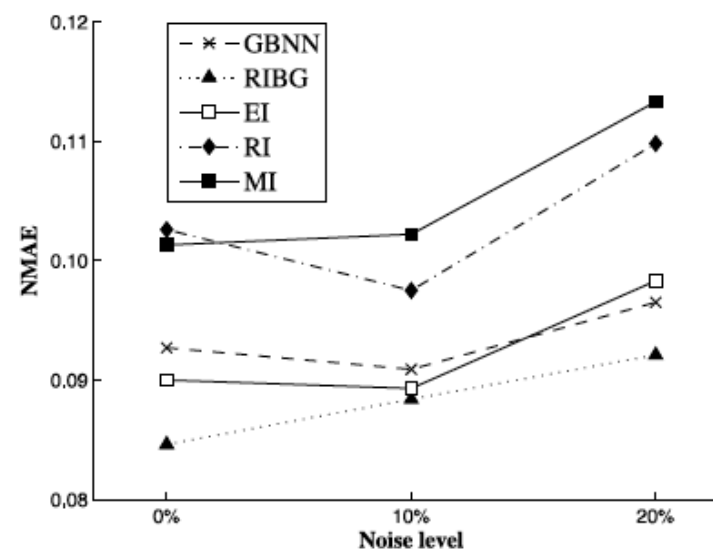
(e) Mechanism=MAR, Missing rate=10%



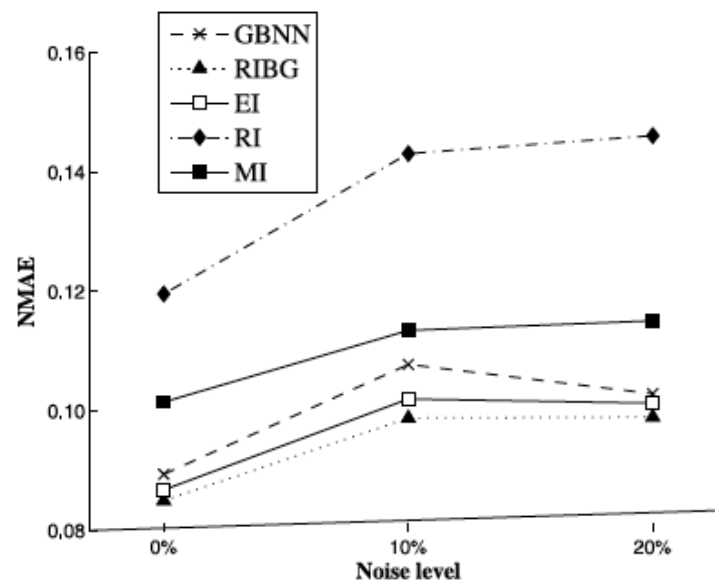
(f) Mechanism=MAR, Missing rate=20%



(g) Mechanism=NMAR, Missing rate=5%



(h) Mechanism=NMAR, Missing rate=10%



(i) Mechanism=NMAR, Missing rate=20%

# Literature

1. Andridge R.R., Little R.J.A. A review of Hot Deck Imputation for Survey Non-response. International statistical Review. 78, 2010, 40-64 pp.
2. Honaker J., King G., Blackwell M. Amelia II: A program for missing data, 2014.
3. Zhu B., He C., Liatsis P. A robust missing value imputation method for noisy data. Applied Intelligence. 36, 1, 2012, 61-74 pp.
4. Packages “HotDeckImputation”, “Amelia”, “mi”

# Questions

---