Handling Missing data with Principal Component Analysis

Hoang Van Ha
University of Science - VNU-HCM
hvha@hcmus.edu.vn

Seminar on Applied Statistics

01 September 2021

Contents

Contents

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
    - Individuals may not answer the door/phone/email, or may respond only to certain questions

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
  - Study participants may fail to show up to some check-ups, some may drop out of the study

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
    - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
    - Study participants may fail to show up to some check-ups, some may drop out of the study
- Administrative registries: data were being collected for administrative purposes, but later we realize that they can be exploited for statistical analyses

Why do we get missing data?

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
  - Study participants may fail to show up to some check-ups, some may drop out of the study
- Administrative registries: data were being collected for administrative purposes, but later we realize that they can be exploited for statistical analyses
  - Certain variables might have missingness or even only be sporadically observed if their collection was not enforced.

Why do we get missing data?

Missing data can also occur by design:

Why do we get missing data?

Missing data can also occur by design:

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals

Why do we get missing data?

Missing data can also occur by design:

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals
- Survey sampling: we do not observe the characteristics for individuals who were not selected to be in the sample

Why do we get missing data?

Missing data can also occur by design:

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals
- Survey sampling: we do not observe the characteristics for individuals who were not selected to be in the sample
- Split-questionnaires: to reduce respondent burden, only subsets of questions are asked to individuals

Why do we get missing data?

Several problems can be framed as missing data problems:

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available

Why do we get missing data?

Several problems can be framed as missing data problems:

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available
    - Data: hospital data containing treatment information, mortality registry that measures survival. Missing data: "links" connecting records that refer to the same individuals

Why do we get missing data?

Several problems can be framed as missing data problems:

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available
    - Data: hospital data containing treatment information, mortality registry that measures survival. Missing data: "links" connecting records that refer to the same individuals
- Measurement error: we can only measure a noisy or surrogate version of what we want

Why do we get missing data?

Several problems can be framed as missing data problems:

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available
  - Data: hospital data containing treatment information, mortality registry that measures survival. Missing data: "links" connecting records that refer to the same individuals
- Measurement error: we can only measure a noisy or surrogate version of what we want
  - Data: 24-hour recall, self-reported measurement of daily fat intake. Missing data: true fat intake

Sometimes we make up missing data

Techniques for handling missing data can be useful for other problems:

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable

Sometimes we make up missing data

Techniques for handling missing data can be useful for other problems:

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"

Sometimes we make up missing data

Techniques for handling missing data can be useful for other problems:

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"
  - Data: friendship connections between people. Latent variable: "community membership" or "social space"

Sometimes we make up missing data

Techniques for handling missing data can be useful for other problems:

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"
  - Data: friendship connections between people. Latent variable: "community membership" or "social space"
- Causal inference: we only observe the outcome under the assigned treatment – what would the outcome be had the subject been assigned to another treatment?

Sometimes we make up missing data

Techniques for handling missing data can be useful for other problems:

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"
  - Data: friendship connections between people. Latent variable: "community membership" or "social space"
- Causal inference: we only observe the outcome under the assigned treatment – what would the outcome be had the subject been assigned to another treatment?
  - One can argue that "potential outcomes" under other treatments are made up missing data as their values never existed, although they could have existed

Data Example: Paris Hospitals

Traumabase (Paris Hospitals): 15000 patients/ 250 variables/ 11 hospitals.

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

....

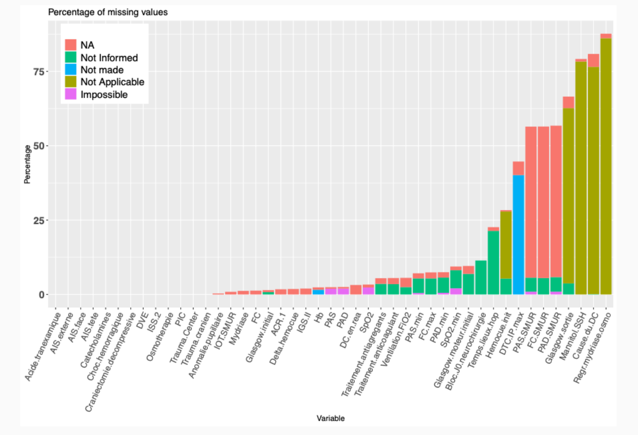
	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	....
1	97	35.6	<NA>	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NM	14.4	15	no	
7	100	36.6	NM	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NM	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

....

- Predict whether to start a blood transfusion, to administer fresh frozen plasma, etc.
- Study the effect of a treatment on survival.

Data Example

Traumabase (Paris Hospitals): 15000 patients/ 250 variables/ 11 hospitals.



- Missing: Not Recorded, Not Made, Note Applicable, etc.
- Multilevel data/ data integration: systematic missing variable in on hospital

Data Example: Ozone dataset

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	NA	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	17	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0622	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0642	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

http://www.airbreizh.asso.fr/

Impacts of the missingness

- Loss of non-relevant and/or non-explanatory information
    - Null impact
- Loss of relevant and/or explanatory information
    - Impact depending of proportion of missing values
    - Possible bias in the estimation of the precision and the accuracy

How to handle missing values?

1. Delete all missing values:

How to handle missing values?

1. Delete all missing values:
   - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!

How to handle missing values?

1. Delete all missing values:
    - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!
    - In general, this is a bad method, the default listwise my result in significant loss of information.

How to handle missing values?

1. Delete all missing values:
   - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!
   - In general, this is a bad method, the default listwise my result in significant loss of information.
2. Impute data with an imputation method:

How to handle missing values?

1. Delete all missing values:
   - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!
   - In general, this is a bad method, the default listwise my result in significant loss of information.
2. Impute data with an imputation method:
   - Replace all missing values by the mean/median/mode of the corresponding variable, or

How to handle missing values?

1. Delete all missing values:
   - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!
   - In general, this is a bad method, the default listwise my result in significant loss of information.
2. Impute data with an imputation method:
   - Replace all missing values by the mean/median/mode of the corresponding variable, or
   - imputation by Regression, Principal Component Analysis (PCA), Random Forests (RF), etc.

How to handle missing values?

1. Delete all missing values:
   - Easy, simple and maybe good enough with small amount of missing data. But defining "small" is problematic!!
   - In general, this is a bad method, the default listwise my result in significant loss of information.
2. Impute data with an imputation method:
   - Replace all missing values by the mean/median/mode of the corresponding variable, or
   - imputation by Regression, Principal Component Analysis (PCA), Random Forests (RF), etc.
3. Design method that handle missing values.

Contents

Missing value problematic

Dealing with missing values depends on:

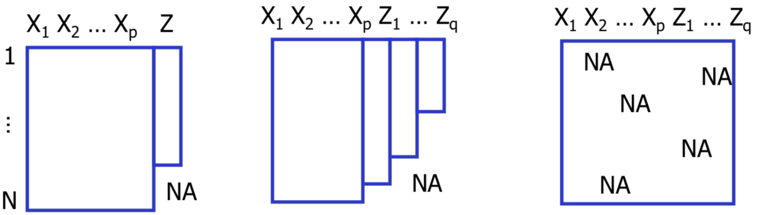
- the pattern of missing values
- the mechanism leading to missing values

Missing value problematic

Dealing with missing values depends on:

- the pattern of missing values
- the mechanism leading to missing values

Pattern of missing data:



Univariate Monotone General (non-monotone)

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

$$X = \begin{bmatrix} 6 & 7 & 8 & ?? \\ 0 & ?? & 11 & 15 \\ 1 & ?? & ?? & 9 \end{bmatrix}$$

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

$$X = \begin{bmatrix} 6 & 7 & 8 & ?? \\ 0 & ?? & 11 & 15 \\ 1 & ?? & ?? & 9 \end{bmatrix} \qquad M = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

$$X = \begin{bmatrix} 6 & 7 & 8 & ?? \\ 0 & ?? & 11 & 15 \\ 1 & ?? & ?? & 9 \end{bmatrix} \qquad M = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

There are 3 different types of missing data:

- **Missing Completely At Random (MCAR):** the missingness pattern M and the observation X are independent $\mathbb{P}(M|X;\phi) = \mathbb{P}(M;\phi)$, $\forall X$.

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \dots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

There are 3 different types of missing data:

- **Missing Completely At Random (MCAR):** the missingness pattern M and the observation X are independent $\mathbb{P}(M|X;\phi) = \mathbb{P}(M;\phi)$, $\forall X$.
- **Missing At Random (MAR):** $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi) = \mathbb{P}(M|X_{obs};\phi)$, $\forall X_{miss}$.

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

There are 3 different types of missing data:

- **Missing Completely At Random (MCAR):** the missingness pattern M and the observation X are independent $\mathbb{P}(M|X;\phi) = \mathbb{P}(M;\phi)$, $\forall X$.
- **Missing At Random (MAR):** $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi) = \mathbb{P}(M|X_{obs};\phi)$, $\forall X_{miss}$.
- **Missing Not At Random (MNAR):** other cases, $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{miss};\phi)$ or $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi)$.

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

There are 3 different types of missing data:

- **Missing Completely At Random (MCAR):** the missingness pattern M and the observation X are independent $\mathbb{P}(M|X;\phi) = \mathbb{P}(M;\phi)$, $\forall X$.
- **Missing At Random (MAR):** $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi) = \mathbb{P}(M|X_{obs};\phi)$, $\forall X_{miss}$.
- **Missing Not At Random (MNAR):** other cases, $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{miss};\phi)$ or $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi)$.

Missing data mechanisms

Let $X = (X_{obs}, X_{miss})$ a complete data model. Assume $X = (X_1, \ldots, X_p)$. Let $M = (M_{ik})$, $1 \leq i \leq p$, $1 \leq k \leq n$ where

$$M_{ik} = \begin{cases} 1 & \text{if } X_{ik} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

There are 3 different types of missing data:

- **Missing Completely At Random (MCAR):** the missingness pattern M and the observation X are independent $\mathbb{P}(M|X;\phi) = \mathbb{P}(M;\phi)$, $\forall\, X$.
- **Missing At Random (MAR):** $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi) = \mathbb{P}(M|X_{obs};\phi)$, $\forall\, X_{miss}$.
- **Missing Not At Random (MNAR):** other cases, $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{miss};\phi)$ or $\mathbb{P}(M|X;\phi) = \mathbb{P}(M|X_{obs}, X_{miss};\phi)$.

Most approaches for inference with missing data assume MAR.

Missing data mechanisms

- **MCAR:** the presence/absence of data is completely independent of observable variables and parameters of interest. In this case, the analysis performed on the data are unbiased.

Missing data mechanisms

- **MCAR:** the presence/absence of data is completely independent of observable variables and parameters of interest. In this case, the analysis performed on the data are unbiased.
- **MAR:** when missing data is not random but can be totally related to a variable where there is complete information. This kind of missing data can induce a bias in your analysis especially if it unbalances your data because of many missing values in a certain category.

Missing data mechanisms

- **MCAR:** the presence/absence of data is completely independent of observable variables and parameters of interest. In this case, the analysis performed on the data are unbiased.
- **MAR:** when missing data is not random but can be totally related to a variable where there is complete information. This kind of missing data can induce a bias in your analysis especially if it unbalances your data because of many missing values in a certain category.
- **MNAR:** the probability of being missing varies for reasons that are unknown to us. For example, in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case.

Missing data mechanisms: example

Age	Income (Inc)	M_{Age}	M_{Inc}
24	1500	1	1
19	NA	1	0
29	4200	1	1
68	NA	1	0

We want to explain the Income according to the Age. There are missing values in the Income.

Missing data mechanisms: example

Age	Income (Inc)	M_{Age}	M_{Inc}
24	1500	1	1
19	NA	1	0
29	4200	1	1
68	NA	1	0

We want to explain the Income according to the Age. There are missing values in the Income.

- If the observations are MCAR, the missingness of the Income does not depend on the Age and the Income.

Missing data mechanisms: example

Age	Income (Inc)	M_{Age}	M_{Inc}
24	1500	1	1
19	NA	1	0
29	4200	1	1
68	NA	1	0

We want to explain the Income according to the Age. There are missing values in the Income.

- If the observations are MCAR, the missingness of the Income does not depend on the Age and the Income.
- If the observations are MAR, the missingness of the Income does not depend on the Income. For example, it occurs if young and old people are less likely to give their incomes.

Missing data mechanisms: example

Age	Income (Inc)	M_{Age}	M_{Inc}
24	1500	1	1
19	NA	1	0
29	4200	1	1
68	NA	1	0

We want to explain the Income according to the Age. There are missing values in the Income.

- If the observations are MCAR, the missingness of the Income does not depend on the Age and the Income.
- If the observations are MAR, the missingness of the Income does not depend on the Income. For example, it occurs if young and old people are less likely to give their incomes.
- If the observations are MNAR, the missingness of the Income depends on the Income itself and may depend on the Age. A possible interpretation is that very rich or poor people are less likely to give their incomes.

Missing data mechanisims

- In general, missing data complicates inference

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<< MNAR$$

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<< MNAR$$

- How we determine MCAR/MAR/MNAR?

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<< MNAR$$

- How we determine MCAR/MAR/MNAR?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- How we determine MCAR/MAR/MNAR?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<< MNAR$$

- How we determine MCAR/MAR/MNAR?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it

Missing data mechanisims

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- How we determine MCAR/MAR/MNAR?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it
- Inference under MNAR is more realistic but more complicated. Most approaches for inference with missing data assume MAR.

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M | X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

If the data are **MAR** (or **MCAR**),

$$f(X_{obs}, M; \theta, \phi) = \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}; \phi) dX_{miss}$$

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M | X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

If the data are **MAR** (or **MCAR**),

$$f(X_{obs}, M; \theta, \phi) = \int f(X_{obs}, X_{miss}; \theta) f(M | X_{obs}; \phi) dX_{miss}$$

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M | X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

If the data are **MAR** (or **MCAR**),

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}; \theta) f(M | X_{obs}; \phi) dX_{miss} \\ &= f(M | X_{obs}; \phi) \int f(X_{obs}, X_{miss}; \theta) dX_{miss}. \end{aligned}$$

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

If the data are **MAR** (or **MCAR**),

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}; \phi) dX_{miss} \\ &= f(M|X_{obs}; \phi) \int f(X_{obs}, X_{miss}; \theta) dX_{miss}. \end{aligned}$$

Hence, $f(X_{obs}, M; \theta, \phi) = f(M|X_{obs}; \phi) f(X_{obs}; \theta)$.

Ignorable missing values

- In some cases, estimating θ form an incomplete data can be done in a simple way by *ignoring* the missing data mechanism.
- Assume that X has a density, parametrized by θ that we want to estimate, for example, if X is Gaussian, we have $\theta = (\boldsymbol{\mu}, \Sigma)$.

Maximum Likelihood estimation:

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}, M; \theta, \phi) dX_{miss} \\ &= \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}, X_{miss}; \phi) dX_{miss}. \end{aligned}$$

If the data are **MAR** (or **MCAR**),

$$\begin{aligned} f(X_{obs}, M; \theta, \phi) &= \int f(X_{obs}, X_{miss}; \theta) f(M|X_{obs}; \phi) dX_{miss} \\ &= f(M|X_{obs}; \phi) \int f(X_{obs}, X_{miss}; \theta) dX_{miss}. \end{aligned}$$

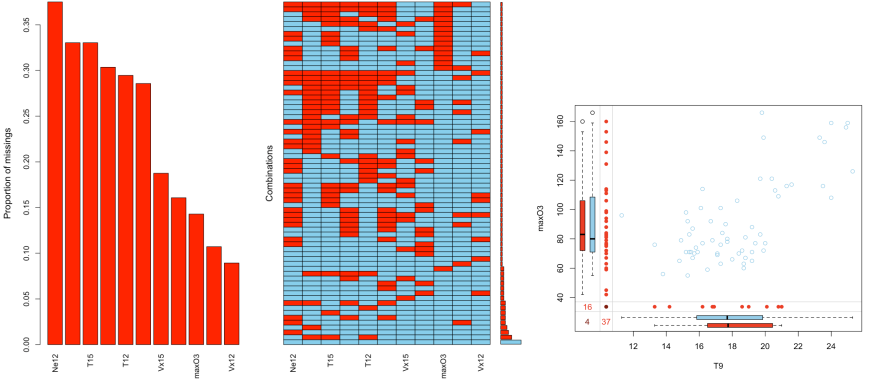
Hence, $f(X_{obs}, M; \theta, \phi) = f(M|X_{obs}; \phi) f(X_{obs}; \theta)$.

MAR is the minimal property to access the likelihood of missing data.

Contents

Visualization

- It is crucial to perform some descriptive statistics (how many missing? how many variables, individuals with missing?) and try to inspect and vizualize the pattern of missing entries and get hints on the mechanism that generated the missingness.



R package: VIM (M. Templ), naniar (N. Tierney), FactoMineR (Husson *et al.*).

Methods for handling missing values

1. Simple methods:

Methods for handling missing values

1. Simple methods:
   - A naïve method: Complete-Case Analysis

Methods for handling missing values

1. Simple methods:
   - A naïve method: Complete-Case Analysis
   - Single imputation

Methods for handling missing values

1. Simple methods:
   - A naïve method: Complete-Case Analysis
   - Single imputation
     - Mean Imputation

Methods for handling missing values

1. Simple methods:
    - A naïve method: Complete-Case Analysis
    - Single imputation
        - Mean Imputation
        - Regression Imputation

Methods for handling missing values

1. Simple methods:
    - A naïve method: Complete-Case Analysis
    - Single imputation
        - Mean Imputation
        - Regression Imputation
        - K-Nearest Neighbors (KNN)

Methods for handling missing values

1. Simple methods:
   - A naïve method: Complete-Case Analysis
   - Single imputation
     - Mean Imputation
     - Regression Imputation
     - K-Nearest Neighbors (KNN)
2. Imputation with EM algorithm and joint model with Gaussian distribution

Methods for handling missing values

1. Simple methods:
    - A naïve method: Complete-Case Analysis
    - Single imputation
        - Mean Imputation
        - Regression Imputation
        - K-Nearest Neighbors (KNN)
2. Imputation with EM algorithm and joint model with Gaussian distribution
3. Imputation with PCA

Contents

Complete-Case Analysis

Idea: discard observations with missingness, run intended analysis with remaining data.

Complete-Case Analysis

Idea: discard observations with missingness, run intended analysis with remaining data.

Gender	Age	Income	...
F	25	60,000	...
M	?	?	...
?	51	?	...
F	?	150,300	
...	...	...	...

Complete-Case Analysis

Idea: discard observations with missingness, run intended analysis with remaining data.

Gender	Age	Income	...
F	25	60,000	...
~~M~~	~~?~~	~~?~~	~~...~~
~~?~~	~~51~~	~~?~~	~~...~~
~~F~~	~~?~~	~~150,300~~	~~...~~
...	...	...	...

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!
- **Limitations:** Loss of information in incomplete cases has two aspects:

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!
- **Limitations:** Loss of information in incomplete cases has two aspects:
  - Increased variance of estimates

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!
- **Limitations:** Loss of information in incomplete cases has two aspects:
  - Increased variance of estimates
  - Bias when complete cases differ systematically from incomplete cases

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!
- **Limitations:** Loss of information in incomplete cases has two aspects:
  - Increased variance of estimates
  - Bias when complete cases differ systematically from incomplete cases
    - Restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question, this implies MCAR

Complete-Case Analysis

- Easy, simple and maybe good enough with small amount of missing data.
  - But defining "small" is problematic!!
- **Limitations:** Loss of information in incomplete cases has two aspects:
  - Increased variance of estimates
  - Bias when complete cases differ systematically from incomplete cases
    - Restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question, this implies MCAR
    - Example: suppose we are interested in estimating the median income of the some population. We send out an email asking a questionnaire to be completed, amongst which participants are asked to say how much they earn. But only a proportion of the target sample return the questionnaire, and so we have missing incomes for the remaining people. If those that returned an answer to the income question have systematically higher or lower incomes than those who did not return an answer, the median income of the complete cases will be biased.

Mean Imputation

Impute mean of observed values:

Age	*Income*
25	60,000
?	?
51	?
?	150,300
$\vdots$	$\vdots$

$\Longrightarrow$

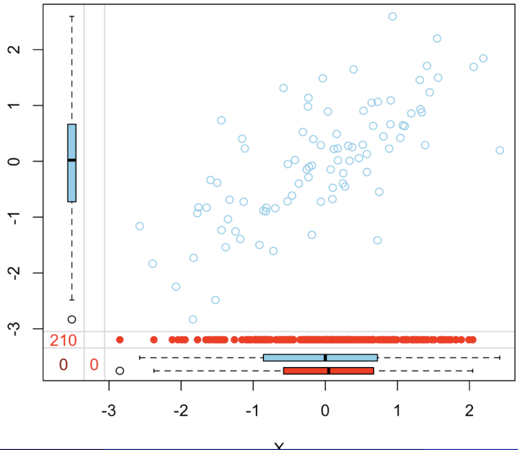
Age	*Income*
25	60,000
$\hat{\mu}^1_{Age}$	$\hat{\mu}^1_{Income}$
51	$\hat{\mu}^1_{Income}$
$\hat{\mu}^1_{Age}$	150,300
$\vdots$	$\vdots$

Mean Imputation

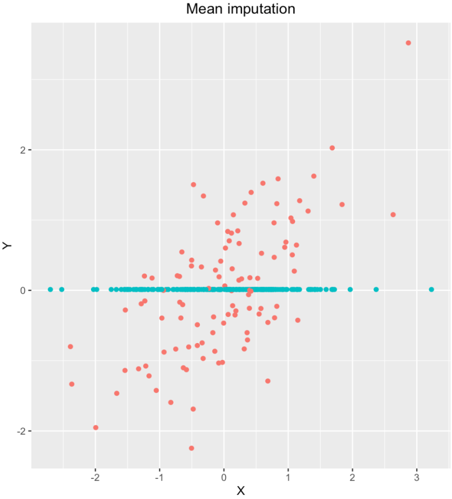
- Consider n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- 70% of missing entries completely at random on Y.
- Simulated data: $n = 300$, $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, $\rho_{XY} = 0.7$.

Mean Imputation

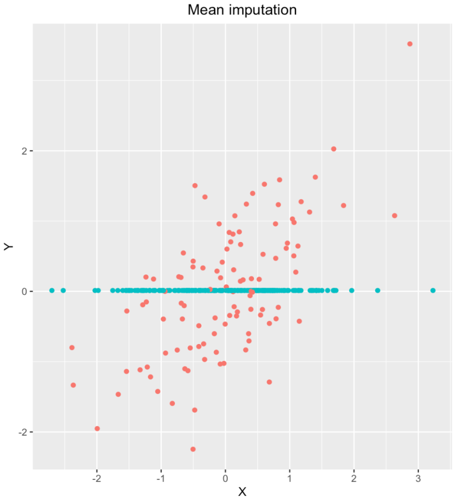
- Consider n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- 70% of missing entries completely at random on Y.
- Simulated data: $n = 300$, $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, $\rho_{XY} = 0.7$.



Mean Imputation



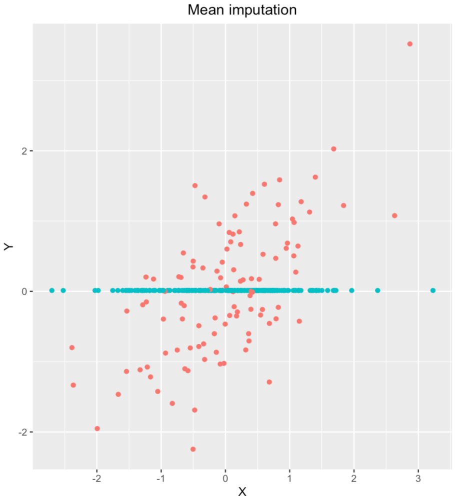
Mean Imputation



- preserve the mean of the imputed variable,

$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$$

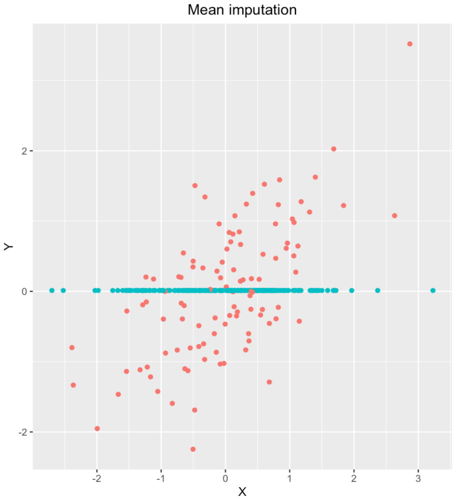
Mean Imputation



- preserve the mean of the imputed variable,
- reduces variance; standard errors of estimates from filled- in data are too small, since

$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$$

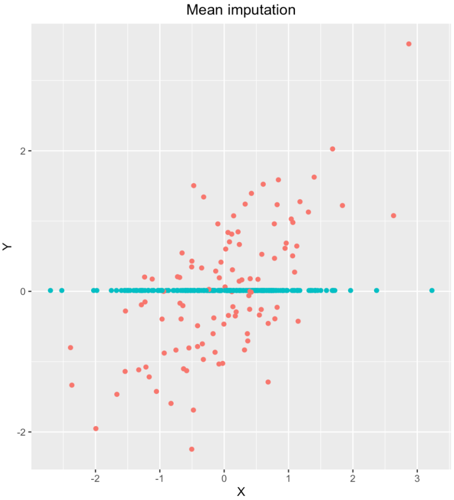
Mean Imputation



- preserve the mean of the imputed variable,
- reduces variance; standard errors of estimates from filled- in data are too small, since
  - standard deviations are underestimated

$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$$

Mean Imputation



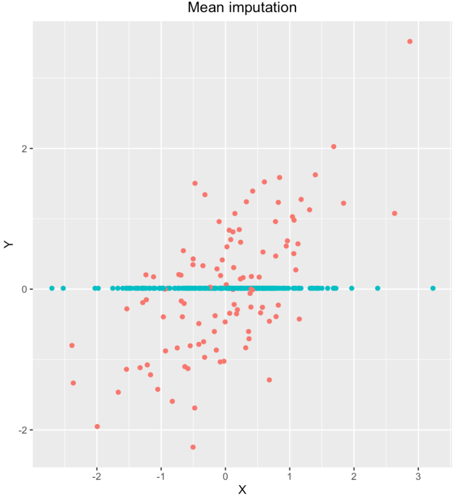
- preserve the mean of the imputed variable,
- reduces variance; standard errors of estimates from filled- in data are too small, since
  - standard deviations are underestimated
  - "Sample size" is overstated

$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$

$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$

$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$

Mean Imputation



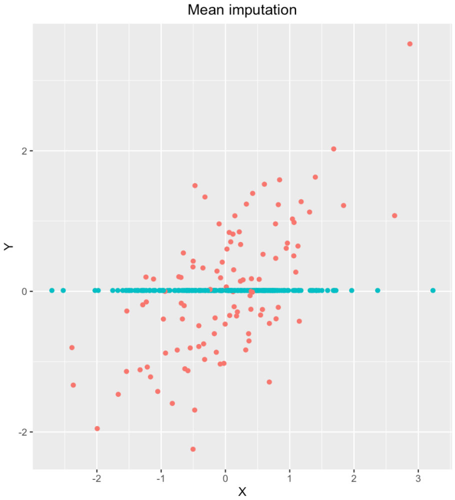
- preserve the mean of the imputed variable,
- reduces variance; standard errors of estimates from filled- in data are too small, since
  - standard deviations are underestimated
  - "Sample size" is overstated
- distorts the correlation with other variables,

$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$

$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$

$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$

Mean Imputation



- preserve the mean of the imputed variable,
- reduces variance; standard errors of estimates from filled- in data are too small, since
  - standard deviations are underestimated
  - "Sample size" is overstated
- distorts the correlation with other variables,
- deforms joint and marginal distributions.

$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.43$$

Regression Imputation

- Replace a missing value Y_i by a predicted value $\hat{Y}_i$ obtained by regression of Y on $X_1, X_2, \ldots, X_n$

Regression Imputation

- Replace a missing value Y_i by a predicted value $\hat{Y}_i$ obtained by regression of Y on $X_1, X_2, \ldots, X_n$
- Valide for means under MCAR

Regression Imputation

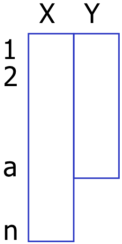
- Replace a missing value Y_i by a predicted value $\hat{Y}_i$ obtained by regression of Y on $X_1, X_2, \ldots, X_n$
- Valide for means under MCAR
- Underestimates true variance of estimators

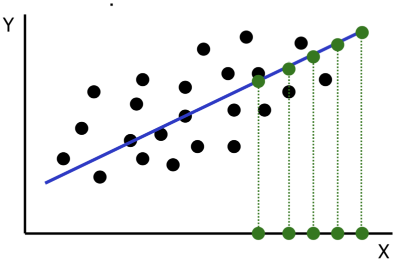
Regression Imputation

- Replace a missing value Y_i by a predicted value $\hat{Y}_i$ obtained by regression of Y on $X_1, X_2, \ldots, X_n$
- Valide for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Regression Imputation

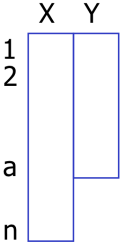
- Example: simple linear regression

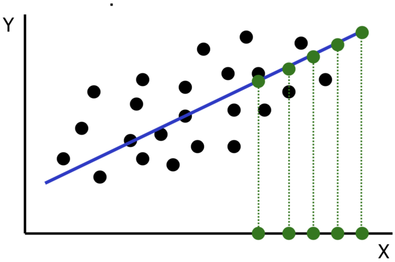




Regression Imputation

- Example: simple linear regression

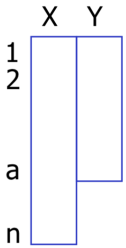


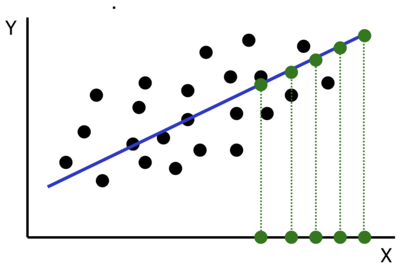


- Regression with the complete cases: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, a$

Regression Imputation

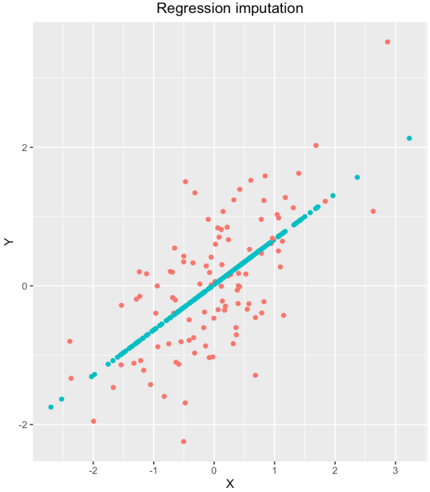
- Example: simple linear regression





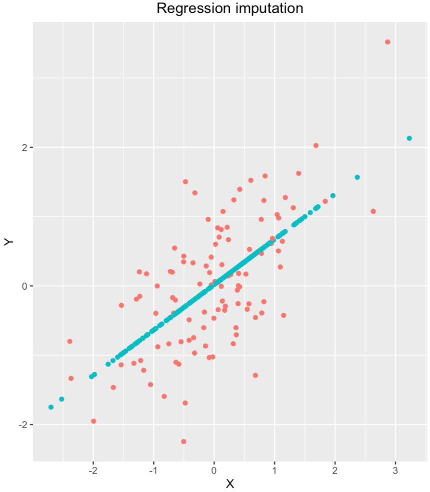
- Regression with the complete cases: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $i = 1, \ldots, a$
- Imputation by the prediction of the regression model:
  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = a + 1, \ldots, n$

Regression Imputation



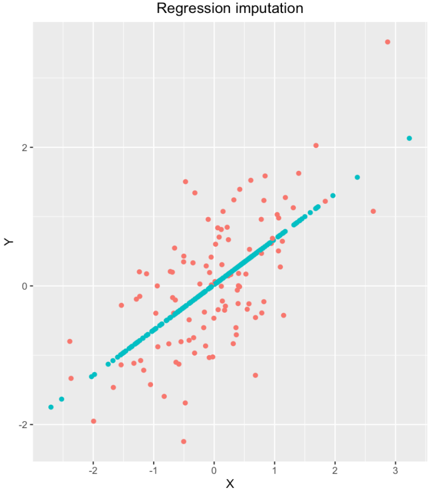
- Impute by regression take into account the relationship

Regression Imputation



- Impute by regression take into account the relationship
- variance underestimated and correlation overestimate

Regression Imputation

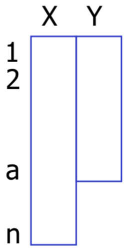


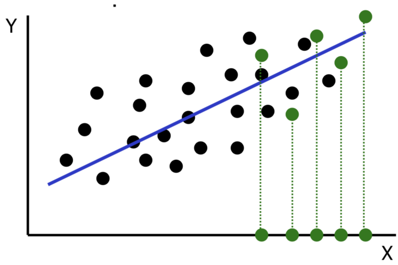
- Impute by regression take into account the relationship
- variance underestimated and correlation overestimate

$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.04$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.81$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.86$$

Stochastic Regression Imputation

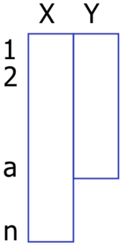
- Estimate the coefficients β_0, β_1 and the variance σ^2, then impute from the predictive $Y_i \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2)$

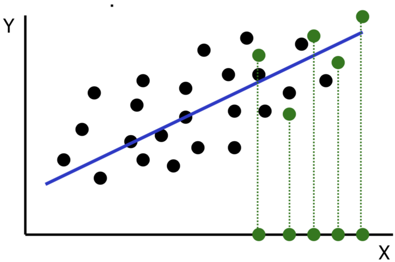




Stochastic Regression Imputation

- Estimate the coefficients β_0, β_1 and the variance σ^2, then impute from the predictive $Y_i \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2)$

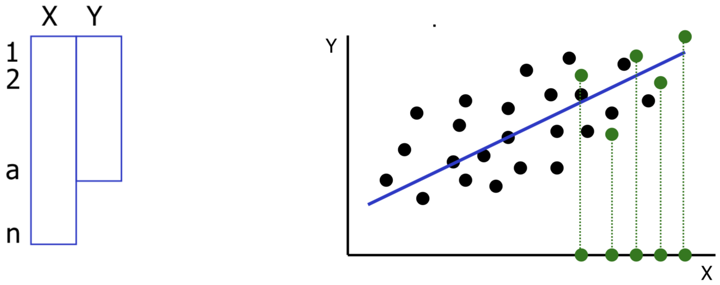




- Regression with the complete cases: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, a$ and $\hat{\sigma}^2$

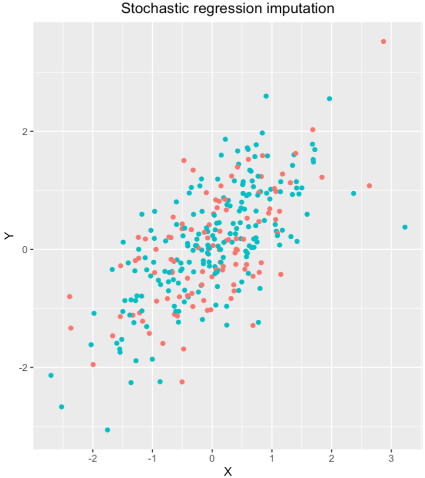
Stochastic Regression Imputation

- Estimate the coefficients β_0, β_1 and the variance σ^2, then impute from the predictive $Y_i \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2)$



- Regression with the complete cases: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $i = 1, \ldots, a$ and $\hat{\sigma}^2$
- Imputation by the prediction of the regression model:
  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i, i = a+1, \ldots, n$ with $\epsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2)$

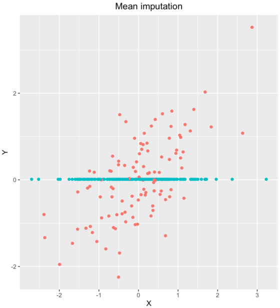
Regression Imputation



- Stochastic regression imputation preserve distribution

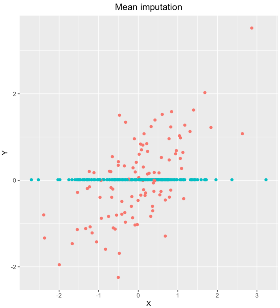
$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.98$$
$$\rho_{XY} = 0.7 \qquad \hat{\rho}_{XY} = 0.69$$

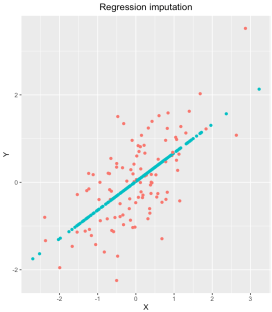
Single Imputation with means and regression: summary



$$\mu_Y = 0 \qquad \hat{\mu}_Y = 0.02$$
$$\sigma_Y = 1 \qquad \hat{\sigma}_Y = 0.58$$
$$\rho = 0.7 \qquad \hat{\rho} = 0.43$$

Single Imputation with means and regression: summary



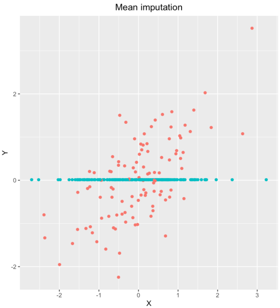


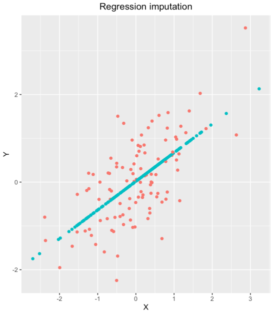
$\mu_Y = 0$ $\quad \hat{\mu}_Y = 0.02$ $\quad \hat{\mu}_Y = 0.04$

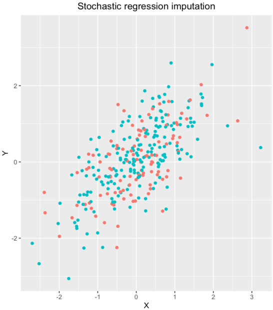
$\sigma_Y = 1$ $\quad \hat{\sigma}_Y = 0.58$ $\quad \hat{\sigma}_Y = 0.81$

$\rho = 0.7$ $\quad \hat{\rho} = 0.43$ $\quad \hat{\rho} = 0.86$

Single Imputation with means and regression: summary



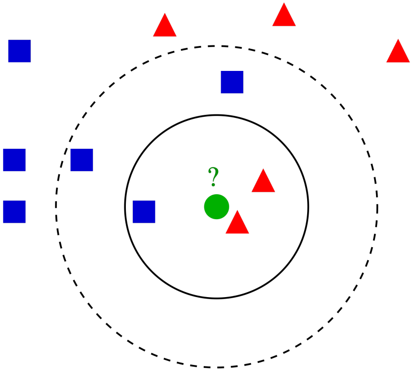




	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_Y = 0$	$\hat{\mu}_Y = 0.02$	$\hat{\mu}_Y = 0.04$	$\hat{\mu}_Y = 0.02$
$\sigma_Y = 1$	$\hat{\sigma}_Y = 0.58$	$\hat{\sigma}_Y = 0.81$	$\hat{\sigma}_Y = 0.98$
$\rho = 0.7$	$\hat{\rho} = 0.43$	$\hat{\rho} = 0.86$	$\hat{\rho} = 0.69$

Imputation with K-Nearest Neighbors

- **Idea**: The missing value is replaced by an observed value of an individual having similar characteristics.
  - similar characteristics $\Leftrightarrow$ "nearest neighbor"
    - determine an appropriate distance function on one or multiple auxiliary variables



K-Nearest Neighbors: Algorithm

- **Algorithm**:

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i*} is the variable with missing values.

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^*j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^*j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^*j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

- **Parameters calibration:**

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^* j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

- **Parameters calibration:**
  - the number of neighbors k

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^*j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

- **Parameters calibration:**
  - the number of neighbors k
  - the distance function (Euclidean/Mahalanobis distance for numeric variables; Hamming distance for categorical ones)

K-Nearest Neighbors: Algorithm

- **Algorithm**:
  1. Select an integer number k: $1 \leq k \leq n$.
  2. Calculate the distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$ where Y_{i^*} is the variable with missing values.
  3. Retain the k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ for which their distances are smallest.
  4. Replace the missing values by the mean of the k neighbors:

$$(y_{ij})_{miss} = y_{i^*j^*} = \frac{1}{k}\left(Y_{(i_1)} + \cdots + Y_{(i_k)}\right).$$

- **Parameters calibration:**
  - the number of neighbors k
  - the distance function (Euclidean/Mahalanobis distance for numeric variables; Hamming distance for categorical ones)
  - the aggregation method: we use arithmetic mean, median and mode for numeric variables and mode for categorical ones.

KNN: example

Consider the following dataset with the weight value of ID11 is missing:

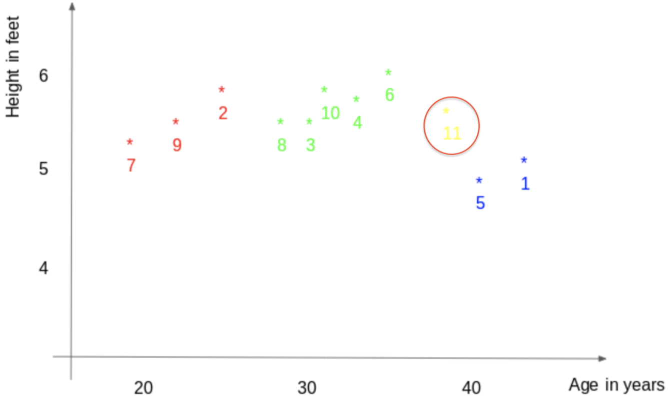
KNN: example

Consider the following dataset with the weight value of ID11 is missing:

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

KNN: example

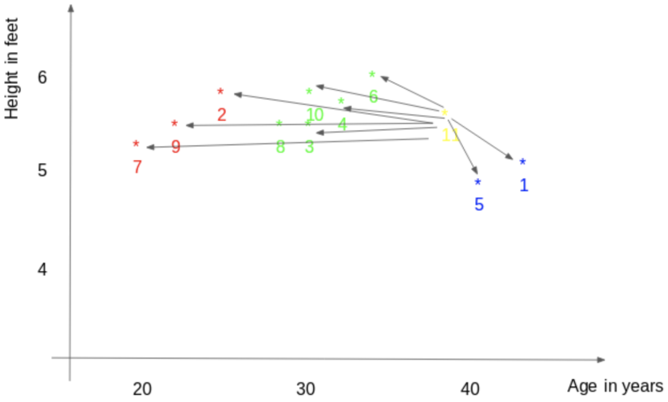
Consider the following dataset with the weight value of ID11 is missing:



KNN: example

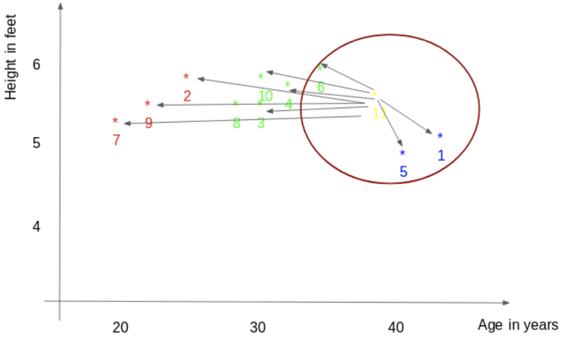
Step 1: Calculate the distance.

- Euclidean distance: $d(x, y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$ for two vectors $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$.



KNN: example

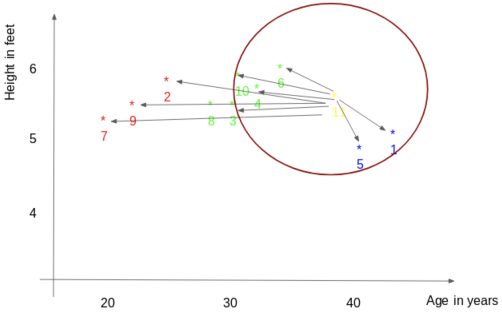
Step 2 & 3: Determine the k nearest neighbors (k closes points) based on the distance and compute the predicted value for $ID11$.
If we choose $k = 3$: $ID11 = (77 + 72 + 60)3 = 69.66$.



ID	Height	Age	Weight
1	5	45	77
5	4.8	40	72
6	5.8	36	60

KNN: example

Step 2 & 3: Determine the k nearest neighbors (k closes points) based on the distance and compute the predicted value for $ID11$.
If we choose $k = 5$: $ID11 = (77 + 59 + 72 + 60 + 58)5 = 65.2$.



ID	Height	Age	Weight
1	5	45	77
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
10	5.6	32	58

K-Nearest Neighbors: summary

- KNN is particularly useful for dealing with all kind of missing data. It can handle continuous, discrete, ordinal and categorical data

K-Nearest Neighbors: summary

- KNN is particularly useful for dealing with all kind of missing data. It can handle continuous, discrete, ordinal and categorical data
- The algorithm is easy to implement

K-Nearest Neighbors: summary

- KNN is particularly useful for dealing with all kind of missing data. It can handle continuous, discrete, ordinal and categorical data
- The algorithm is easy to implement
- time-consuming on larger datasets

K-Nearest Neighbors: summary

- KNN is particularly useful for dealing with all kind of missing data. It can handle continuous, discrete, ordinal and categorical data
- The algorithm is easy to implement
- time-consuming on larger datasets
- on high dimensional data, accuracy can be severely degraded

K-Nearest Neighbors: summary

- KNN is particularly useful for dealing with all kind of missing data. It can handle continuous, discrete, ordinal and categorical data
- The algorithm is easy to implement
- time-consuming on larger datasets
- on high dimensional data, accuracy can be severely degraded
- underestimation of variances

Contents

EM algorithm

- **Goal:** Estimate as well as possible the parameters and their variance despite missing values.

EM algorithm

- **Goal:** Estimate as well as possible the parameters and their variance despite missing values.
- Suppose that we are interested in estimating unknown parameters $\theta \in \mathbb{R}^d$ of a model, for example, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\theta = (\boldsymbol{\mu}, \Sigma)$.

EM algorithm

- **Goal:** Estimate as well as possible the parameters and their variance despite missing values.
- Suppose that we are interested in estimating unknown parameters $\theta \in \mathbb{R}^d$ of a model, for example, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\theta = (\boldsymbol{\mu}, \Sigma)$.
- Let $f(X; \theta)$ be the probability density function of $X = (X_{obs}, X_{miss})$.

EM algorithm

- **Goal:** Estimate as well as possible the parameters and their variance despite missing values.
- Suppose that we are interested in estimating unknown parameters $\theta \in \mathbb{R}^d$ of a model, for example, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\theta = (\boldsymbol{\mu}, \Sigma)$.
- Let $f(X; \theta)$ be the probability density function of $X = (X_{obs}, X_{miss})$.
- The EM algorithm aims at finding the estimate of θ that maximize the observed data log-likelihood

$$L_{obs}(\theta; X_{obs}) = \log f(X_{obs}; \theta) = \log \int f(X_{obs}, X_{miss}; \theta) dX_{miss}.$$

EM algorithm

- **Goal:** Estimate as well as possible the parameters and their variance despite missing values.
- Suppose that we are interested in estimating unknown parameters $\theta \in \mathbb{R}^d$ of a model, for example, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\theta = (\boldsymbol{\mu}, \Sigma)$.
- Let $f(X; \theta)$ be the probability density function of $X = (X_{obs}, X_{miss})$.
- The EM algorithm aims at finding the estimate of θ that maximize the observed data log-likelihood

$$L_{obs}(\theta; X_{obs}) = \log f(X_{obs}; \theta) = \log \int f(X_{obs}, X_{miss}; \theta) dX_{miss}.$$

- As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the MLE by iteratively maximizing the expected complete-data log-likelihood. Denote the complete-data log-likelihood as

$$L_{comp}(\theta; X) = \log f(X_{obs}, X_{miss}; \theta).$$

EM algorithm

Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate of θ at t-th iteration, then the next iteration of EM consists of two steps:

EM algorithm

Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate of θ at t-th iteration, then the next iteration of EM consists of two steps:

- **E step (conditional expectation):**

$$Q(\theta, \theta^{(t)}) = \mathbb{E}\left[L_{comp}(\theta; X) | X_{obs}; \theta^{(t)}\right] = \int L_{comp}(\theta; X) f(X_{miss} | X_{obs}; \theta^{(t)}) dX_{miss}.$$

EM algorithm

Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate of θ at t-th iteration, then the next iteration of EM consists of two steps:

- **E step (conditional expectation):**

$$Q(\theta, \theta^{(t)}) = \mathbb{E}\left[L_{comp}(\theta; X)|X_{obs}; \theta^{(t)}\right] = \int L_{comp}(\theta; X) f(X_{miss}|X_{obs}; \theta^{(t)}) dX_{miss}.$$

- **M step (maximization):** Determine $\theta^{(t+1)}$ by maximizing the function Q

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)}).$$

Convergence criterion: $|\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}| \leq \epsilon$.

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with EM.

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with *EM*.
- Then the conditional distribution of the missing data X_{miss} given X_{obs} can be derived using Schur complements. Denote by $\Sigma_{miss} \in \mathbb{R}^{m\times m}$, $\Sigma_{obs} \in \mathbb{R}^{r\times r}$ and $\Sigma_{miss,obs} \in \mathbb{R}^{m\times r}$, respectively, the covariance matrix of X_{miss}, X_{miss} and between X_{miss} and X_{obs}, then Σ given by:

$$\Sigma = \begin{pmatrix} \Sigma_{miss} & \Sigma_{miss,obs} \\ \Sigma_{miss,obs}^T & \Sigma_{obs} \end{pmatrix}.$$

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with *EM*.
- Then the conditional distribution of the missing data X_{miss} given X_{obs} can be derived using Schur complements. Denote by $\Sigma_{miss} \in \mathbb{R}^{m\times m}$, $\Sigma_{obs} \in \mathbb{R}^{r\times r}$ and $\Sigma_{miss,obs} \in \mathbb{R}^{m\times r}$, respectively, the covariance matrix of X_{miss}, X_{miss} and between X_{miss} and X_{obs}, then Σ given by:

$$\Sigma = \begin{pmatrix} \Sigma_{miss} & \Sigma_{miss,obs} \\ \Sigma_{miss,obs}^T & \Sigma_{obs} \end{pmatrix}.$$

- $X_{miss}|X_{obs}$ has a normal distribution with mean and covariance matrix given by:

$$\Sigma_{X_{miss}|X_{obs}} = \Sigma_{miss} - \Sigma_{miss,obs}\Sigma_{obs}^{-1}\Sigma_{miss,obs}^T,$$
$$\boldsymbol{\mu}_{X_{miss}|X_{obs}} = \mathbb{E}[X_{miss}] + \Sigma_{miss,obs}\Sigma_{obs}^{-1}(X_{obs} - \mathbb{E}[X_{obs}]).$$

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with *EM*.
- Then the conditional distribution of the missing data X_{miss} given X_{obs} can be derived using Schur complements. Denote by $\Sigma_{miss} \in \mathbb{R}^{m \times m}$, $\Sigma_{obs} \in \mathbb{R}^{r \times r}$ and $\Sigma_{miss,obs} \in \mathbb{R}^{m \times r}$, respectively, the covariance matrix of X_{miss}, X_{miss} and between X_{miss} and X_{obs}, then Σ given by:

$$\Sigma = \begin{pmatrix} \Sigma_{miss} & \Sigma_{miss,obs} \\ \Sigma_{miss,obs}^T & \Sigma_{obs} \end{pmatrix}.$$

- $X_{miss}|X_{obs}$ has a normal distribution with mean and covariance matrix given by:

$$\begin{aligned} \Sigma_{X_{miss}|X_{obs}} &= \Sigma_{miss} - \Sigma_{miss,obs}\Sigma_{obs}^{-1}\Sigma_{miss,obs}^T, \\ \boldsymbol{\mu}_{X_{miss}|X_{obs}} &= \mathbb{E}[X_{miss}] + \Sigma_{miss,obs}\Sigma_{obs}^{-1}(X_{obs} - \mathbb{E}[X_{obs}]). \end{aligned}$$

- Finally, impute the missing data using by drawing from the conditional distribution $X_{miss}|X_{obs}$.

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with *EM*.
- Then the conditional distribution of the missing data X_{miss} given X_{obs} can be derived using Schur complements. Denote by $\Sigma_{miss} \in \mathbb{R}^{m\times m}$, $\Sigma_{obs} \in \mathbb{R}^{r\times r}$ and $\Sigma_{miss,obs} \in \mathbb{R}^{m\times r}$, respectively, the covariance matrix of X_{miss}, X_{miss} and between X_{miss} and X_{obs}, then Σ given by:

$$\Sigma = \begin{pmatrix} \Sigma_{miss} & \Sigma_{miss,obs} \\ \Sigma^T_{miss,obs} & \Sigma_{obs} \end{pmatrix}.$$

- $X_{miss}|X_{obs}$ has a normal distribution with mean and covariance matrix given by:

$$\begin{aligned} \Sigma_{X_{miss}|X_{obs}} &= \Sigma_{miss} - \Sigma_{miss,obs}\Sigma^{-1}_{obs}\Sigma^T_{miss,obs}, \\ \boldsymbol{\mu}_{X_{miss}|X_{obs}} &= \mathbb{E}[X_{miss}] + \Sigma_{miss,obs}\Sigma^{-1}_{obs}(X_{obs} - \mathbb{E}[X_{obs}]). \end{aligned}$$

- Finally, impute the missing data using by drawing from the conditional distribution $X_{miss}|X_{obs}$.

Imputation with joint model with Gaussian distribution

Assumption: $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- First, estimate $\boldsymbol{\mu}$ and Σ from an incomplete data with *EM*.
- Then the conditional distribution of the missing data X_{miss} given X_{obs} can be derived using Schur complements. Denote by $\Sigma_{miss} \in \mathbb{R}^{m\times m}$, $\Sigma_{obs} \in \mathbb{R}^{r\times r}$ and $\Sigma_{miss,obs} \in \mathbb{R}^{m\times r}$, respectively, the covariance matrix of X_{miss}, X_{miss} and between X_{miss} and X_{obs}, then Σ given by:

$$\Sigma = \begin{pmatrix} \Sigma_{miss} & \Sigma_{miss,obs} \\ \Sigma^T_{miss,obs} & \Sigma_{obs} \end{pmatrix}.$$

- $X_{miss}|X_{obs}$ has a normal distribution with mean and covariance matrix given by:

$$\Sigma_{X_{miss}|X_{obs}} = \Sigma_{miss} - \Sigma_{miss,obs}\Sigma^{-1}_{obs}\Sigma^T_{miss,obs},$$
$$\boldsymbol{\mu}_{X_{miss}|X_{obs}} = \mathbb{E}[X_{miss}] + \Sigma_{miss,obs}\Sigma^{-1}_{obs}(X_{obs} - \mathbb{E}[X_{obs}]).$$

- Finally, impute the missing data using by drawing from the conditional distribution $X_{miss}|X_{obs}$.

Implementation in **R**: package norm

Contents

PCA: overview

PCA in the complete case boils down to finding a matrix of low rank S that gives:

- Best approximation of the data with projection.
- Best representation of the variability.

Figure: Camel or dromedary? (Source: J.P. Fénelon)

PCA: overview

PCA in the complete case boils down to finding a matrix of low rank S that gives:

- Best approximation of the data with projection.
- Best representation of the variability.

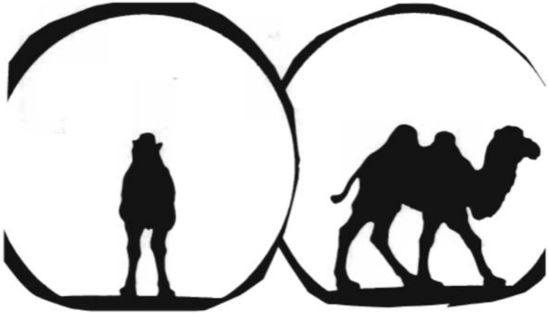
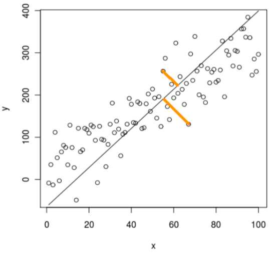
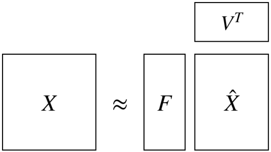


Figure: Camel or dromedary? (Source: J.P. Fénelon)

PCA reconstruction





- Minimizes distance between observations and their projections.
- Approximate the matrix $X_{n \times p}$ with a low rank matrix $S < p$ in the least square sense ($\| \cdot \|$ the Frobenius norm: $\|X\|_2^2 = tr(XX^T)$):

$$\operatorname{argmin}_Q \left\{ \|X_{n \times p} - Q_{n \times p}\|_2^2 : rank(Q) \leq S \right\}.$$

- The PCA solution (Eckart & Young, 1936) is the truncated singular value decomposition (SVD) of X at the order S:

$$\hat{X} = U_{n \times S} \Lambda_{S \times S}^{1/2} V_{S \times p}^T = F_{n \times S} V_{S \times p}^T.$$

$F = U\Lambda^{1/2}$: PC scores; V: principal axes - loadings.

PCA with missing values

- PCA complete: least squares criterion

$$\operatorname{argmin}_Q \left\{ \|X_{n\times p} - Q_{n\times p}\|_2^2 : rank(Q) \leq S \right\}.$$

[1] Josse, J.& Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), pp. 79-99.

PCA with missing values

- PCA complete: least squares criterion

$$\operatorname{argmin}_Q \left\{ \|X_{n\times p} - Q_{n\times p}\|_2^2 : rank(Q) \leq S \right\}.$$

- PCA with incomplete data: weighted least squares (WLS)

$$\operatorname{argmin}_Q \left\{ \|W_{n\times p} \odot (X_{n\times p} - Q_{n\times p})\|_2^2 : rank(Q) \leq S \right\},$$

where $W_{ij} = 0$ if X_{ij} is missing and $X_{ij} = 1$ otherwise. $\odot$ stands for the elementwise multiplication.

[1] Josse, J.& Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), pp. 79-99.

PCA with missing values

- PCA complete: least squares criterion

$$\operatorname{argmin}_Q \left\{ \|X_{n\times p} - Q_{n\times p}\|_2^2 : rank(Q) \leq S \right\}.$$

- PCA with incomplete data: weighted least squares (WLS)

$$\operatorname{argmin}_Q \left\{ \|W_{n\times p} \odot (X_{n\times p} - Q_{n\times p})\|_2^2 : rank(Q) \leq S \right\},$$

where $W_{ij} = 0$ if X_{ij} is missing and $X_{ij} = 1$ otherwise. $\odot$ stands for the elementwise multiplication.

[1] Josse, J.& Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), pp. 79-99.

PCA with missing values

- PCA complete: least squares criterion

$$\operatorname{argmin}_Q \left\{ \|X_{n\times p} - Q_{n\times p}\|_2^2 : rank(Q) \leq S \right\}.$$

- PCA with incomplete data: weighted least squares (WLS)

$$\operatorname{argmin}_Q \left\{ \|W_{n\times p} \odot (X_{n\times p} - Q_{n\times p})\|_2^2 : rank(Q) \leq S \right\},$$

where $W_{ij} = 0$ if X_{ij} is missing and $X_{ij} = 1$ otherwise. $\odot$ stands for the elementwise multiplication.

Algorithms: weighted alternating least squares (Gabriel and Zamir, 1979); iterative PCA (Kiers, 1997). See Josse and Husson[1], 2012 for more references.

[1] Josse, J.& Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), pp. 79-99.

Iterative PCA

Algorithms:

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**
   (a) perform the SVD on completed data to estimate $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$, S dimension kept.

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**
   (a) perform the SVD on completed data to estimate $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$, S dimension kept.
   (b) compute the fitted matrix $(\hat{X}_{ij}^{S})^{(t)} = U^{(t)} {\Lambda^{(t)}}^{1/2} {V^{(t)}}^{T}$ and define the new imputed data as $X^{(t)} = W \odot X + (\mathbf{1}_{n\times p} - W) \odot (\hat{X}_{ij}^{S})^{(t)}$, where $\mathbf{1}_{n\times p}$ is a matrix filled with ones. The observed values are the same and the missing ones are replaced by the fitted values.

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**
   (a) perform the SVD on completed data to estimate $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$, S dimension kept.
   (b) compute the fitted matrix $(\hat{X}_{ij}^{S})^{(t)} = U^{(t)} {\Lambda^{(t)}}^{1/2} {V^{(t)}}^{T}$ and define the new imputed data as $X^{(t)} = W \odot X + (\mathbf{1}_{n \times p} - W) \odot (\hat{X}_{ij}^{S})^{(t)}$, where $\mathbf{1}_{n \times p}$ is a matrix filled with ones. The observed values are the same and the missing ones are replaced by the fitted values.
3. Repeat steps (2a) and (2b) until the change in the imputed matrix smaller than a given threshold, for instance $\sum_{ij} \left(\hat{X}_{ij}^{(t)} - \hat{X}_{ij}^{(t-1)}\right)^2 \leq \epsilon$.

Iterative PCA

Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**
   (a) perform the SVD on completed data to estimate $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$, S dimension kept.
   (b) compute the fitted matrix $(\hat{X}_{ij}^{S})^{(t)} = U^{(t)} {\Lambda^{(t)}}^{1/2} {V^{(t)}}^{T}$ and define the new imputed data as $X^{(t)} = W \odot X + (\mathbf{1}_{n \times p} - W) \odot (\hat{X}_{ij}^{S})^{(t)}$, where $\mathbf{1}_{n \times p}$ is a matrix filled with ones. The observed values are the same and the missing ones are replaced by the fitted values.
3. Repeat steps (2a) and (2b) until the change in the imputed matrix smaller than a given threshold, for instance $\sum_{ij} \left(\hat{X}_{ij}^{(t)} - \hat{X}_{ij}^{(t-1)} \right)^2 \leq \epsilon$.

Iterative PCA

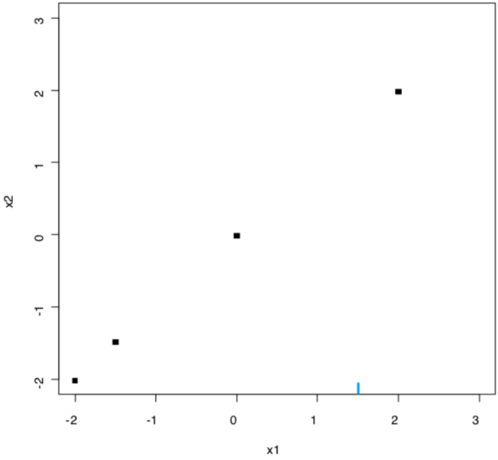
Algorithms:

1. **Initialization $t = 0$:** substitute missing values with initial values, $X^{(0)}$ (mean imputation).
2. **Step $t \geq 1$:**
   (a) perform the SVD on completed data to estimate $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$, S dimension kept.
   (b) compute the fitted matrix $(\hat{X}_{ij}^{S})^{(t)} = U^{(t)} {\Lambda^{(t)}}^{1/2} {V^{(t)}}^{T}$ and define the new imputed data as $X^{(t)} = W \odot X + (\mathbf{1}_{n\times p} - W) \odot (\hat{X}_{ij}^{S})^{(t)}$, where $\mathbf{1}_{n\times p}$ is a matrix filled with ones. The observed values are the same and the missing ones are replaced by the fitted values.
3. Repeat steps (2a) and (2b) until the change in the imputed matrix smaller than a given threshold, for instance $\sum_{ij} \left(\hat{X}_{ij}^{(t)} - \hat{X}_{ij}^{(t-1)}\right)^2 \leq \epsilon$.

Selection of the number of dimensions S: Cross-Validation.

Iterative PCA

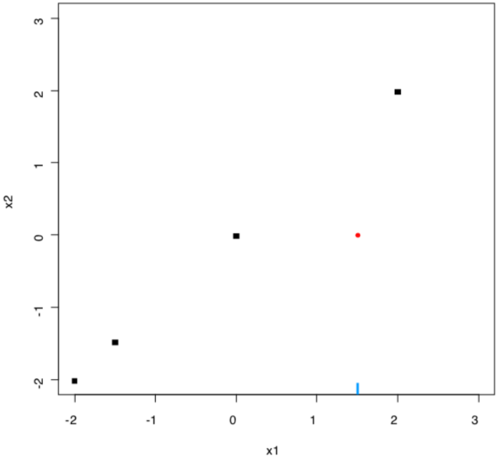
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98



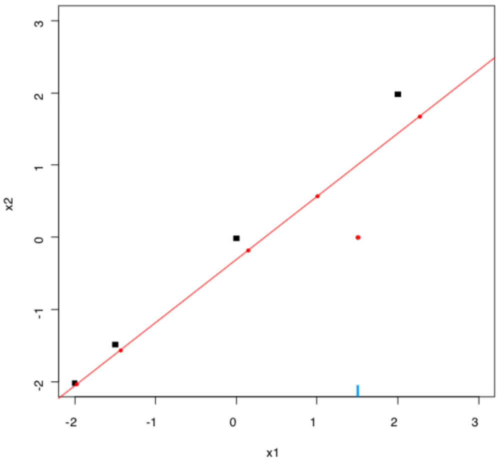
Initialization $t = 0$: $X^{(0)}$ (mean imputation)

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



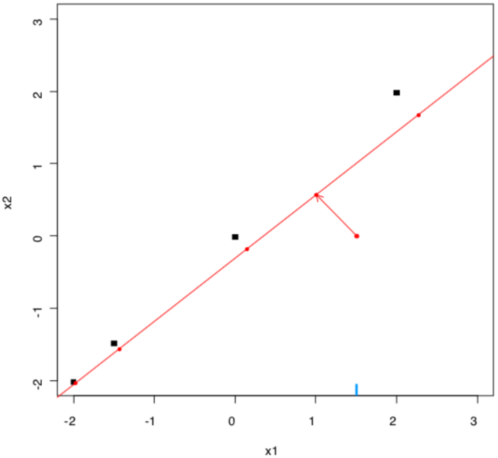
PCA on the completed dataset: $(U^{(t)}, \Lambda^{(t)}, V^{(t)})$

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



Missing values imputed with the fitted matrix $\hat{X}^{(t)} = U^{(t)} {\Lambda^{(t)}}^{1/2} {V^{(t)}}^{T}$

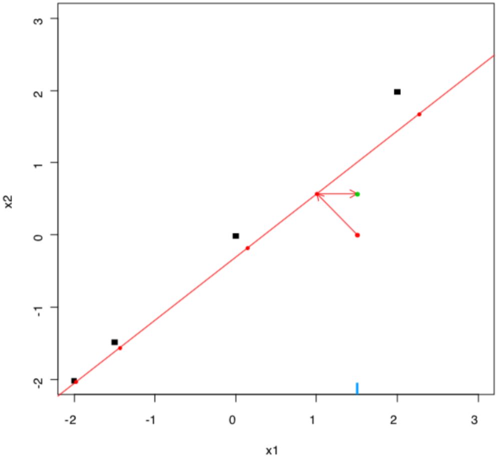
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\hat{x}1$	$\hat{x}2$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



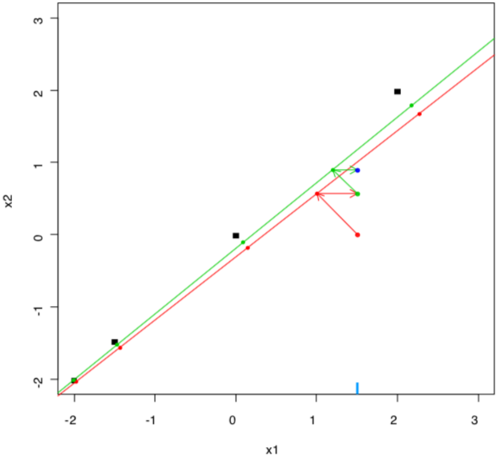
The new imputed dataset is $X^{(t)} = W \odot X + (\mathbf{1}_{n\times p} - W) \odot \hat{X}^{(t)}$

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.57
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.57
2.0	1.98



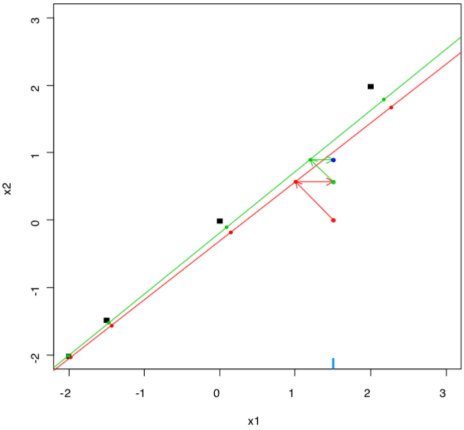
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	**0.90**
2.18	1.78

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.90**
2.0	1.98



PCA on the completed dataset: $(U^{(t+1)}, \Lambda^{(t+1)}, V^{(t+1)})$

Missing values imputed with the fitted matrix $\hat{X}^{(t+1)} = U^{(t+1)} {\Lambda^{(t+1)}}^{1/2} {V^{(t+1)}}^{T}$

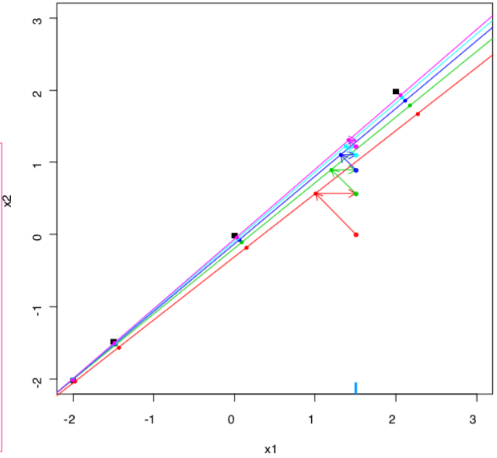
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Steps are repeated until convergence

Selection of S: Cross-Validation

- Review of of six methods of cross-validation in PCA: Bro *et al.*[2](2008).

[2] Bro, R., Kjeldahl, K., Smilde, A.K., Kiers. Cross-validation of component model: a critical look at current methods. Analytical and Bioanalytical Chemistry 390, 1241–1251.

Selection of S: Cross-Validation

- Review of of six methods of cross-validation in PCA: Bro *et al.*[2](2008).
- Criterion using the mean square error of prediction (MSEP):

$$MSEP(S) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(X_{ij} - (\hat{X}_{ij}^{S})^{-ij} \right)^2.$$

[2] Bro, R., Kjeldahl, K., Smilde, A.K., Kiers. Cross-validation of component model: a critical look at current methods. Analytical and Bioanalytical Chemistry 390, 1241–1251.

Selection of S: Cross-Validation

- Review of of six methods of cross-validation in PCA: Bro *et al.*[2](2008).
- Criterion using the mean square error of prediction (MSEP):

$$MSEP(S) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - (\hat{X}_{ij}^{S})^{-ij}\right)^2.$$

Estimate $(\hat{X}^{S})^{-ij}$ for fixed S and each $(i,j) \Rightarrow$ computational burden.

[2] Bro, R., Kjeldahl, K., Smilde, A.K., Kiers. Cross-validation of component model: a critical look at current methods. Analytical and Bioanalytical Chemistry 390, 1241–1251.

Selection of S: Cross-Validation

- Review of of six methods of cross-validation in PCA: Bro *et al.*[2](2008).
- Criterion using the mean square error of prediction (MSEP):

$$MSEP(S) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - (\hat{X}_{ij}^{S})^{-ij}\right)^2.$$

Estimate $(\hat{X}^S)^{-ij}$ for fixed S and each (i,j) $\Rightarrow$ computational burden.

- In a regression, denote a linear fitting method as $\hat{y} = Py$ with y a response vector and P a smoothing matrix. Craven & Whaba (1979) have shown

$$y_i - \hat{y}_i^{-i} = \frac{y_i - \hat{y}_i}{1 - P_{i,i}}.$$

[2] Bro, R., Kjeldahl, K., Smilde, A.K., Kiers. Cross-validation of component model: a critical look at current methods. Analytical and Bioanalytical Chemistry 390, 1241–1251.

Selection of S: Cross-Validation

- Review of of six methods of cross-validation in PCA: Bro *et al.*[2](2008).
- Criterion using the mean square error of prediction (MSEP):

$$MSEP(S) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - (\hat{X}_{ij}^{S})^{-ij}\right)^2.$$

Estimate $(\hat{X}^S)^{-ij}$ for fixed S and each $(i,j) \Rightarrow$ computational burden.

- In a regression, denote a linear fitting method as $\hat{y} = Py$ with y a response vector and P a smoothing matrix. Craven & Whaba (1979) have shown

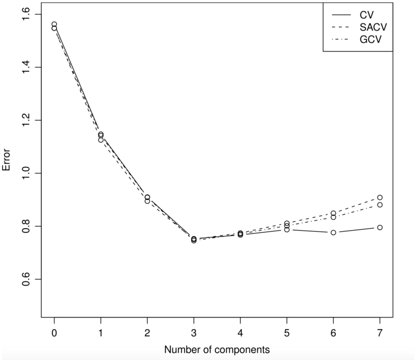
$$y_i - \hat{y}_i^{-i} = \frac{y_i - \hat{y}_i}{1 - P_{i,i}}.$$

- Write PCA as $\hat{X}^S = PX$, we get

$$X_{ij} - (\hat{X}_{ij}^S)^{-ij} \simeq \frac{X_{ij} - \hat{X}_{ij}^S}{1 - P_{ij,ij}}.$$

[2] Bro, R., Kjeldahl, K., Smilde, A.K., Kiers. Cross-validation of component model: a critical look at current methods. Analytical and Bioanalytical Chemistry 390, 1241–1251.

Cross-validation approximations



MSEP for cross-validation:

$$CV(S) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - (\hat{X}_{ij}^{S})^{-ij}\right)^2$$

Smoothing approximation of cross-validation (SACV) and Generalized CV (Josse & Husson[3], 2012):

$$SACV(S) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\frac{X_{ij} - \hat{X}_{ij}^{S}}{1 - P_{ij,ij}}\right)^2$$

$$GCV(S) = \frac{1}{np}\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \hat{X}_{ij}^{S}\right)^2}{(1 - tr(P)/np)^2}$$

R package: missMDA

[3] Josse, J. & Husson, F. Selecting the number of components in PCA using cross-validation approximations. Computational Statististics and Data Analysis.

Overfitting

Overfitting occurs when:

- Many parameters are estimated with respect to the number of observed values (the number of dimensions S and of missing values are important).
- Data are very noisy.

$\Rightarrow$ Trust to much the relationship between variables.

Overfitting

Overfitting occurs when:

- Many parameters are estimated with respect to the number of observed values (the number of dimensions S and of missing values are important).
- Data are very noisy.

$\Rightarrow$ Trust to much the relationship between variables.

Solution: Shrinkage methods.

Regularized iterative PCA[4]

The imputation step:

$$\hat{X}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$$

is replaced by a "shrunk" imputation step (Efron & Morris 1972):

$$\hat{X}_{ij}^{rPCA} = \sum_{s=1}^{S} \left(\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s} \right) \sqrt{\lambda_s} u_{is} v_{js} = \sum_{s=1}^{S} \left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}} \right) u_{is} v_{js},$$

with σ^2 estimated by

$$\hat{\sigma}^2 = \frac{n \sum_{s=S+1}^{p} \lambda_s}{np - p - nS - pS + S^2 + S}, \quad (X_{n \times p}, U_{n \times S}, V_{p \times S})$$

R package: missMDA (F. Husson, J. Josse).

[3] J Josse, J Pagès, and F Husson. Gestion des données manquantes en analyse en composantes principales. Journal de la Société Française de Statistique, 150:28–51, 2009.

Soft thresholding SVD

We replace the imputation step $\hat{X}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$ by a "shrunk" imputation step

$$\hat{X}_{ij}^{Soft} = \sum_{s=1}^{p} (\sqrt{\lambda_s} - \lambda)_+ u_{is} v_{js},$$

where $\hat{X}^{Soft}$ is the closed form solution to

$$\underset{Q}{\operatorname{argmin}} \left\{ \| W \odot (X - Q) \|_2^2 + \lambda \| Q \|_\star \right\},$$

where the nuclear norm $\|Q\|_\star$ is the sum of the singular values of Q.

Soft thresholding SVD

We replace the imputation step $\hat{X}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$ by a "shrunk" imputation step

$$\hat{X}_{ij}^{Soft} = \sum_{s=1}^{p} (\sqrt{\lambda_s} - \lambda)_+ u_{is} v_{js},$$

where $\hat{X}^{Soft}$ is the closed form solution to

$$\operatorname*{argmin}_{Q} \left\{ \| W \odot (X - Q) \|_2^2 + \lambda \| Q \|_\star \right\},$$

where the nuclear norm $\|Q\|_\star$ is the sum of the singular values of Q.

R package: softImpute (T. Hastie *et al.*, 2015, Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. JMLR)

Soft thresholding SVD

We replace the imputation step $\hat{X}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$ by a "shrunk" imputation step

$$\hat{X}_{ij}^{Soft} = \sum_{s=1}^{p} (\sqrt{\lambda_s} - \lambda)_+ u_{is} v_{js},$$

where $\hat{X}^{Soft}$ is the closed form solution to

$$\underset{Q}{\operatorname{argmin}} \left\{ \| W \odot (X - Q) \|_2^2 + \lambda \| Q \|_\star \right\},$$

where the nuclear norm $\|Q\|_\star$ is the sum of the singular values of Q.

R package: softImpute (T. Hastie *et al.*, 2015, Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. JMLR)

Selection of λ:

- Josse & Sardy, Adaptive shrinkage of singular values, Stat Comput, 2015.
- Josse & Wager, Bootstrap-Based Regularization for Low-Rank Matrix Estimation, JMLR, 2016.
- Gavish & Donoho, Optimal Shrinkage of Singular Values, 2016.

Implementation in R: package denoiseR (Josse, Wage, Sardy, denoiseR: A Package for Low Rank Matrix Estimation, 2018).

Single imputation with PCA: summary

- Impute large datasets of different dimensions with continuous, categorial variables:
  - reduce the dimensionality,
  - takes into account the similarities between individuals and the relationships between variables.
- Imputations with PCA are good for strong linear relationships.
- Tunning parameter: number of components S.

Contents

Why multiple imputation?

"Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing)." (Donald B. Rubin)

Single imputation cannot take into account the variability of the missing values prediction
$\Rightarrow$ underestimation of the variability
$\Rightarrow$ confidence intervals and tests that are not valid even if the imputation model is correct.

Solution: Multiple imputation.

Single imputation: confidence interval for mean

CI 95% for μ_Y:

$$\left[\bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\,;\, \bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\right]$$

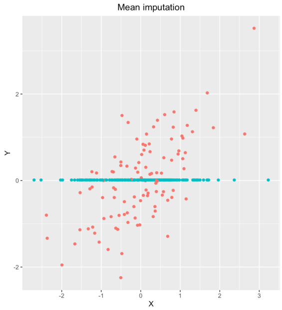
Compute the coverage of the confidence interval for μ_Y.

Single imputation: confidence interval for mean

CI 95% for μ_Y:

$$\left[\bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}} \,;\, \bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\right]$$

Compute the coverage of the confidence interval for μ_Y.



$\mu_Y = 0$ $\quad \hat{\mu}_Y = 0.02$

$\sigma_Y = 1$ $\quad \hat{\sigma}_Y = 0.58$

$\rho = 0.7$ $\quad \hat{\rho} = 0.43$

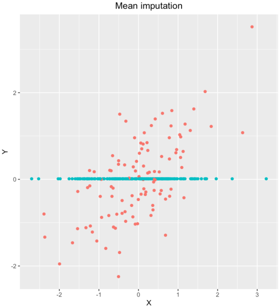
CI 95% $\quad$ 42.9%

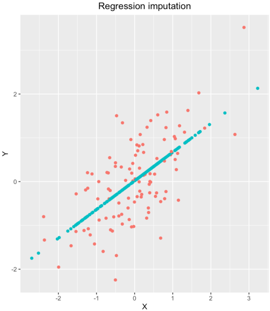
Single imputation: confidence interval for mean

CI 95% for μ_Y:

$$\left[\bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}} \,;\, \bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\right]$$

Compute the coverage of the confidence interval for μ_Y.





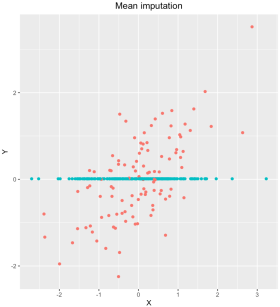
	Mean imputation	Regression imputation
$\mu_Y = 0$	$\hat{\mu}_Y = 0.02$	$\hat{\mu}_Y = 0.04$
$\sigma_Y = 1$	$\hat{\sigma}_Y = 0.58$	$\hat{\sigma}_Y = 0.81$
$\rho = 0.7$	$\hat{\rho} = 0.43$	$\hat{\rho} = 0.86$
CI 95%	42.9%	66.4%

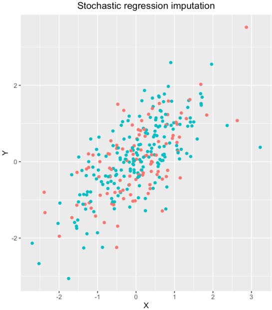
Single imputation: confidence interval for mean

CI 95% for μ_Y:

$$\left[\bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}} \; ; \; \bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\right]$$

Compute the coverage of the confidence interval for μ_Y.





	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_Y = 0$	$\hat{\mu}_Y = 0.02$	$\hat{\mu}_Y = 0.04$	$\hat{\mu}_Y = 0.02$
$\sigma_Y = 1$	$\hat{\sigma}_Y = 0.58$	$\hat{\sigma}_Y = 0.81$	$\hat{\sigma}_Y = 0.98$
$\rho = 0.7$	$\hat{\rho} = 0.43$	$\hat{\rho} = 0.86$	$\hat{\rho} = 0.69$
CI 95%	42.9%	66.4%	78.9%

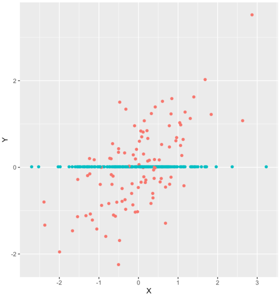
Single imputation: confidence interval for mean

CI 95% for μ_Y:

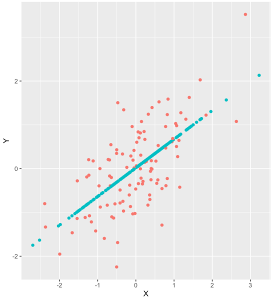
$$\left[\bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\ ;\ \bar{y} - t_{n-1}^{0.025} \times \frac{\hat{\sigma}_Y}{\sqrt{n}}\right]$$

Compute the coverage of the confidence interval for μ_Y.

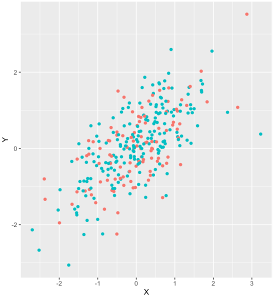
Mean imputation



Regression imputation



Stochastic regression imputation



	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_Y = 0$	$\hat{\mu}_Y = 0.02$	$\hat{\mu}_Y = 0.04$	$\hat{\mu}_Y = 0.02$
$\sigma_Y = 1$	$\hat{\sigma}_Y = 0.58$	$\hat{\sigma}_Y = 0.81$	$\hat{\sigma}_Y = 0.98$
$\rho = 0.7$	$\hat{\rho} = 0.43$	$\hat{\rho} = 0.86$	$\hat{\rho} = 0.69$
CI 95%	42.9%	66.4%	78.9%

$\Rightarrow$ Standard errors calculated from the imputed data are underestimated

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values
  - to preserve the important aspects of the distributions

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values
  - to preserve the important aspects of the distributions
  - to preserve the important relations between the variables

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values
  - to preserve the important aspects of the distributions
  - to preserve the important relations between the variables
- The goals is not

Multiple imputation

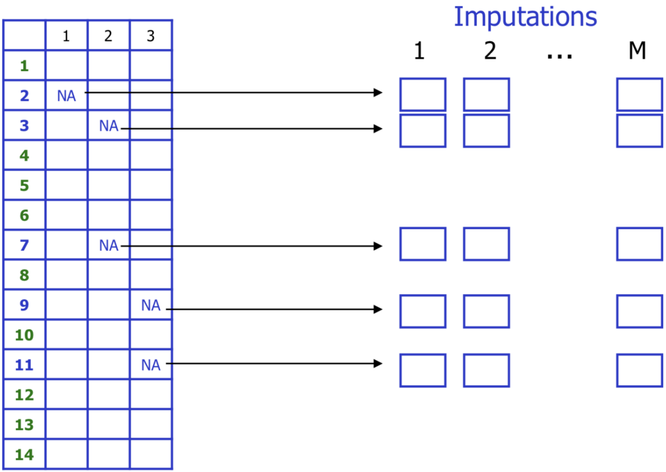
- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values
  - to preserve the important aspects of the distributions
  - to preserve the important relations between the variables
- The goals is not
  - to predict the missing values with the greatest precision

Multiple imputation

- Multiple imputation consists in creating several possible value of a missing value.
- The goals is
  - to reflect correctly the uncertainty of the missing values
  - to preserve the important aspects of the distributions
  - to preserve the important relations between the variables
- The goals is not
  - to predict the missing values with the greatest precision
  - to describe the data in the best possible way

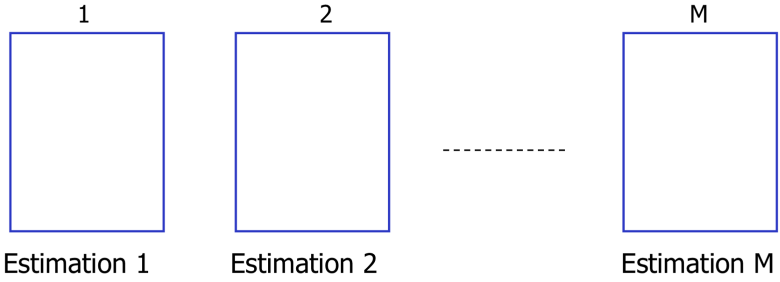
Multiple imputation: Step 1

- Generate $M > 1$ plausible for each missing value (*i.e.* M completed datasets).



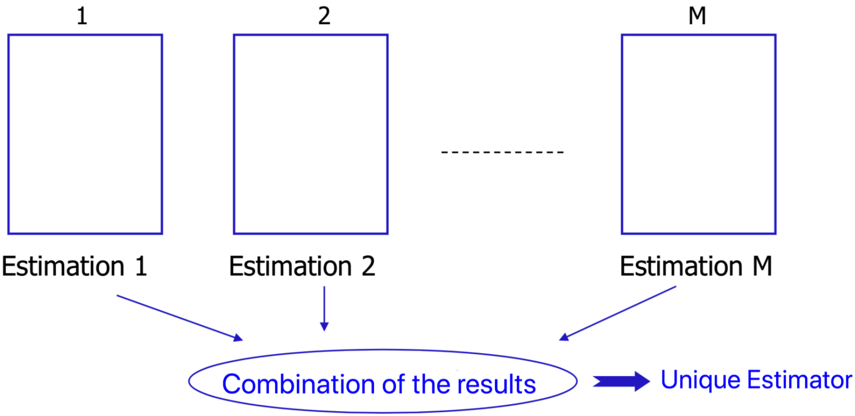
Multiple imputation: Step 2

- Perform independently the analysis on each imputed dataset: $\hat{\theta}_m$, $\widehat{\mathbb{V}ar}(\hat{\theta}_m)$.



Multiple imputation: Step 3

- Combine the results (Rubin's rules).



Multiple imputation: Step 3

- Combine the results (Rubin's rules).

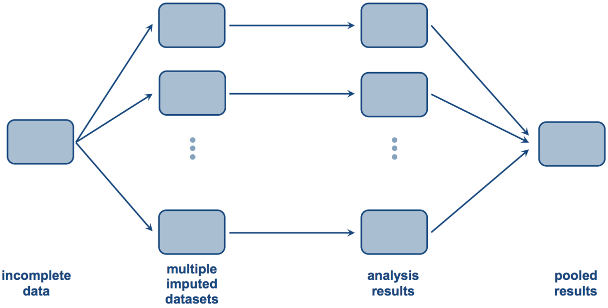
$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m,$$

$$\widehat{\mathbb{V}ar}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbb{V}ar}(\hat{\theta}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2 .$$

Variance $=$ Within $+$ Between imputation variance

$\Rightarrow$ variability of missing values taken into account.

Multiple imputation: Summary



Three steps:

1. **Imputation:** impute multiple times to get multiple completed datasets.
2. **Analysis:** analyse each of the datasets.
3. **Pooling:** combine results, taking into account additional uncertainty.

R packages for missing data imputation

- Imputations with Random Forests (RF): missForest (Daniel J. Stekhoven).
  ⇒ Good for non-linear relationships between continuous variables and when there are interactions.
- Imputations for Categorical data/Mixed/Multi-Blocks/MultiLevel: Multiple Correspondence Analysis (MCA)/ Regularized iterative MCA. R package: missMDA (Husson & Josse).
- Time Series imputation: imputeTS package (Steffen Moritz).

R packages for missing data imputation

- See Missing values taskview[5] (Julie Josse, Nicholas Tierney, Nathalie Vialaneix).
  - Single imputation:
    - *k-nearest neighbors:* DMwR, impute, VIM, wNNSel (for imputation in large dimensional datasets)
    - *regression based imputations:* VIM (linear regression based imputation in the function `regressionImp`), imputation
    - *Based on random forest:* missForest
    - *PCA/Singular Value Decomposition/matrix completion:* missMDA, softImpute
  - Multiple imputation: Amelia, mice, missMDA, miceMNAR
  - Specific application fields: visit Missing values taskview for further information.

[5] https://cran.r-project.org/web/views/MissingData.html

Contents

References

This talk relies mainly on the following sources:

Julie Josse. Course at École de recherche à Aussois 2018. http://juliejosse.com/

Marie Davidian. Course at NC State University, spring 2017. https://www4.stat.ncsu.edu/~davidian/st790/index.html

Stef van Buuren. Flexible Imputation of Missing Data, 2nd edition. Chapman & HallCRC, 2018. https://stefvanbuuren.name/fimd/

J. Josse and F. Husson. *Selecting the number of components in principal component analysis using cross-validation approximations*. Computational Statistics and Data Analysis, 56 (2012) 1869–1879.