A missing values tour in R

Wei Jiang
Ecole Polytechnique

IM UWr @ Wroclaw

December 20, 2019

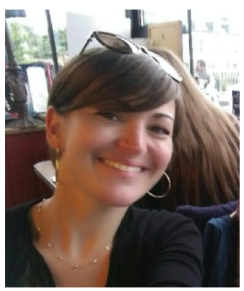
Self presentation

- PhD student in Statistics @ CMAP, Ecole Polytechnique
- Research interest: modeling with missing values
  e.g. parameter estimation / model selection for regression
- Fields of application: health data – Paris hospital
- R packages: misaem, ABSLOPE
- Ongoing work with M. Bogdan:
  false discovery control with missing covariates

Self presentation

- PhD student in Statistics @ CMAP, Ecole Polytechnique
- Research interest: modeling with missing values
  e.g. parameter estimation / model selection for regression
- Fields of application: health data – Paris hospital
- R packages: misaem, ABSLOPE
- Ongoing work with M. Bogdan:
  false discovery control with missing covariates

Julie Josse



Marc Lavielle

Tobias Gauss

Sophie Hamada

Outline

10am - 10:50am :

- Overview of missing values problems
- Imputation methods
- Multiple imputation

11:05am - 12am :

- Expectation-Maximization algorithm
- Lab

https://github.com/StatsIMUWr/MissingDataWorkshop

Outline

Missing values

When we attempt to explore data as a source of knowledge, **missing values** lies in the process of obtaining, recording, and preparing the data.

- Unanswered questions in a survey
- loss of data
- machines that fail

"We should be suspicious of any dataset (large or small) which appears perfect." – David J. Hand

$\Rightarrow$ Still an issue in the "big data" area

Paris Hospitals - TraumaBase dataset

20 000 severely traumatised patients + 250 measurements

	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	NA	14.4	15	no	
7	100	36.6	NA	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NA	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	
.....							

Paris Hospitals - TraumaBase dataset

20 000 severely traumatised patients + 250 measurements

	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	NA	14.4	15	no	
7	100	36.6	NA	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NA	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	
.....							

$\Rightarrow$ Predict the Glasgow score, whether to start a blood transfusion, etc...

$\Rightarrow$ Linear regression/ Logistic regression /Random Forests with missing covariates

Missing values problematic

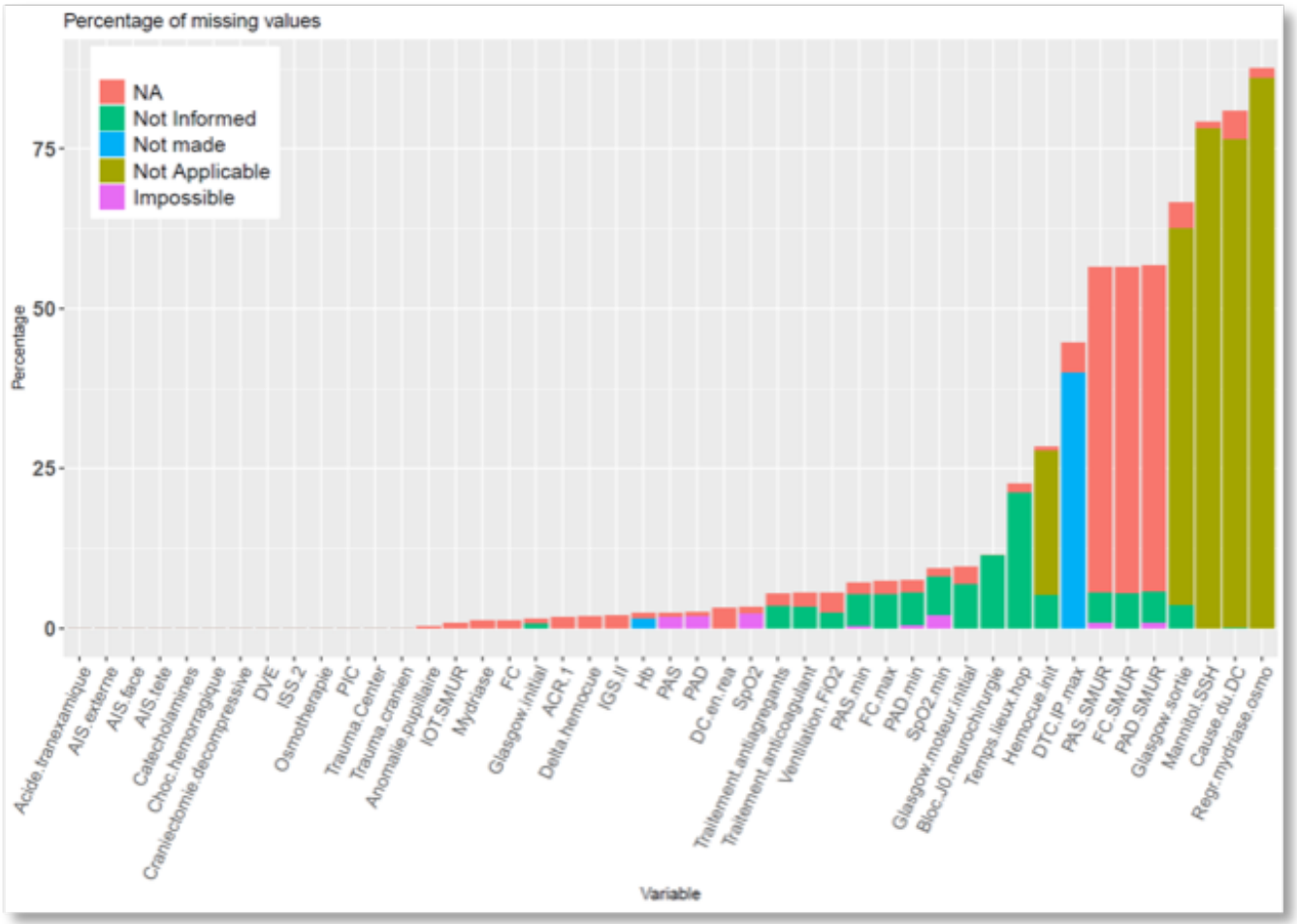
List-wise deletion (default `lm` function in R)

$\Rightarrow$ loss of information

Missing values problematic

List-wise deletion (default `lm` function in R)

$\Rightarrow$ loss of information



$\Rightarrow$ less than 10% remained

Mean imputation

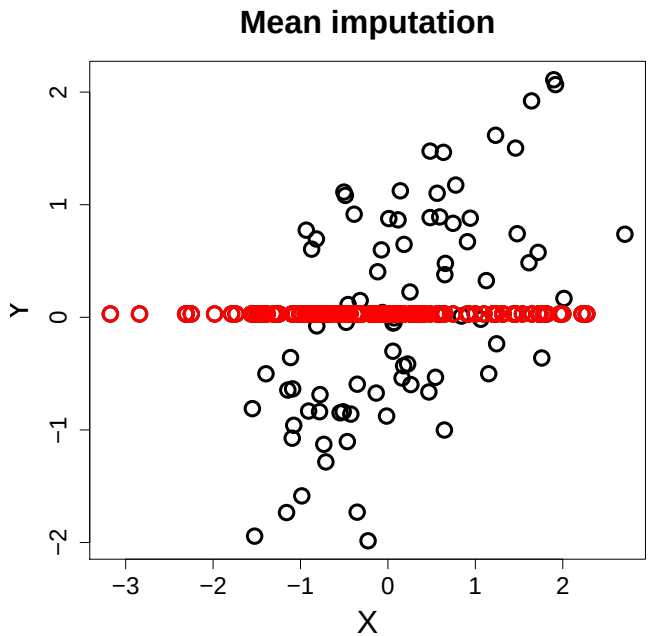
- $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$ *i.i.d.*
- 70% missing entries on y randomly

Date completion by the mean of observed values in y
$\Rightarrow$ Estimate parameters:

Mean imputation

- $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$ *i.i.d.*
- 70% missing entries on y randomly

Date completion by the mean of observed values in y
$\Rightarrow$ Estimate parameters:



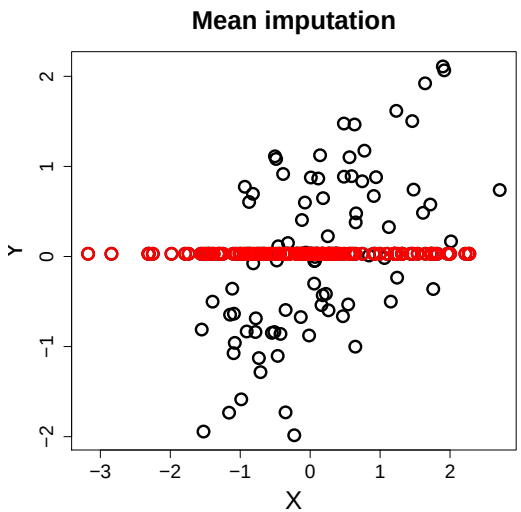
$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

$\hat{\mu}_y = 0.01$
$\hat{\sigma}_y = 0.5$
$\hat{\rho} = 0.30$

$\Rightarrow$ Biased estimates

Imputation by linear regression

- Mean imputation

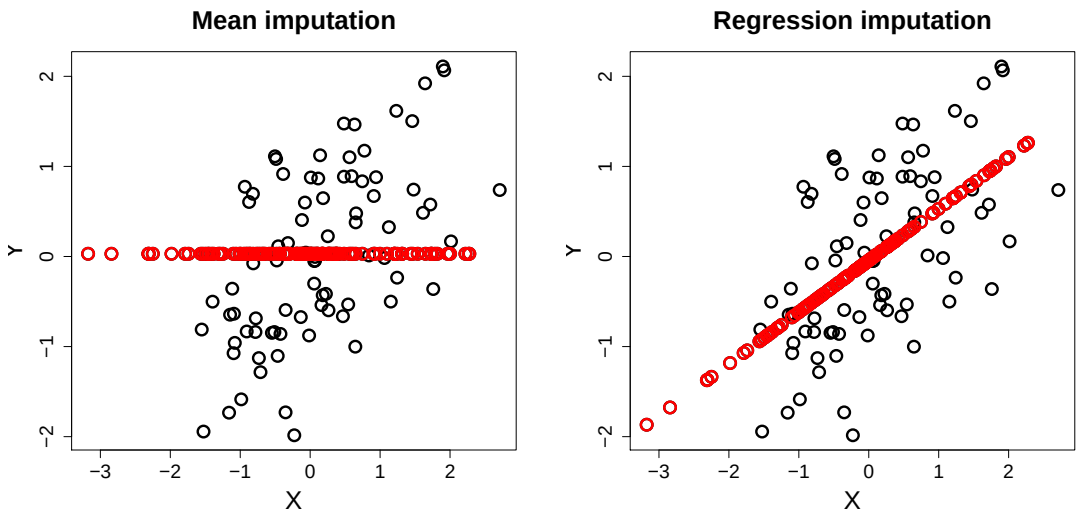


$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

0.01
0.5
0.30

Imputation by linear regression

- Mean imputation
- Impute by regression: impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  $\Rightarrow$ variance underestimated and correlation overestimated.



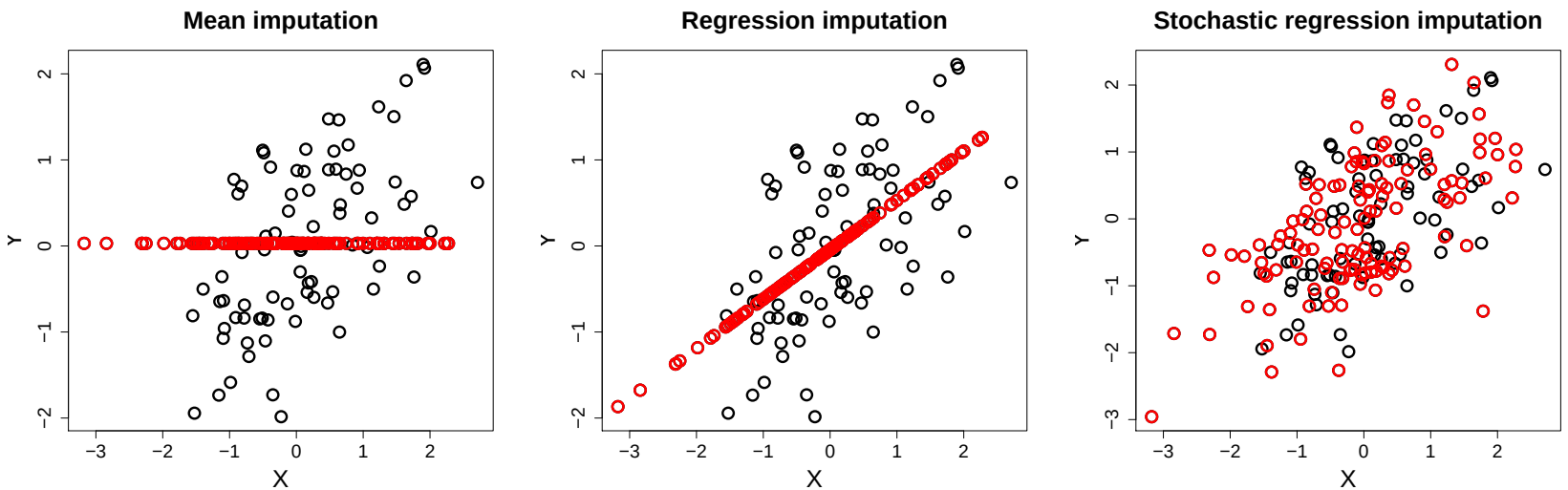
$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

0.01
0.5
0.30

0.01
0.72
0.78

Imputation by linear regression

- Mean imputation
- Impute by regression: impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  $\Rightarrow$ variance underestimated and correlation overestimated.
- Impute by stochastic regression: impute $\hat{y}_i \sim \mathcal{N}\left(x_i\hat{\beta}, \hat{\sigma}^2\right)$
  $\Rightarrow$ preserve distribution



$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

How about real dataset?

Dealing with missing values depends on:

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

$\Rightarrow$ Explore dataset

R packages: `VIM`, `naniar` (Matthias Templ, Nick Tierney)

`FactoMineR` (YouTube)

Ozone data set

112 daily records of meteorological variables (wind speed, temperature, rainfall, etc.) and ozone concentration recorded in Rennes

	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/

Aim: complete ozone

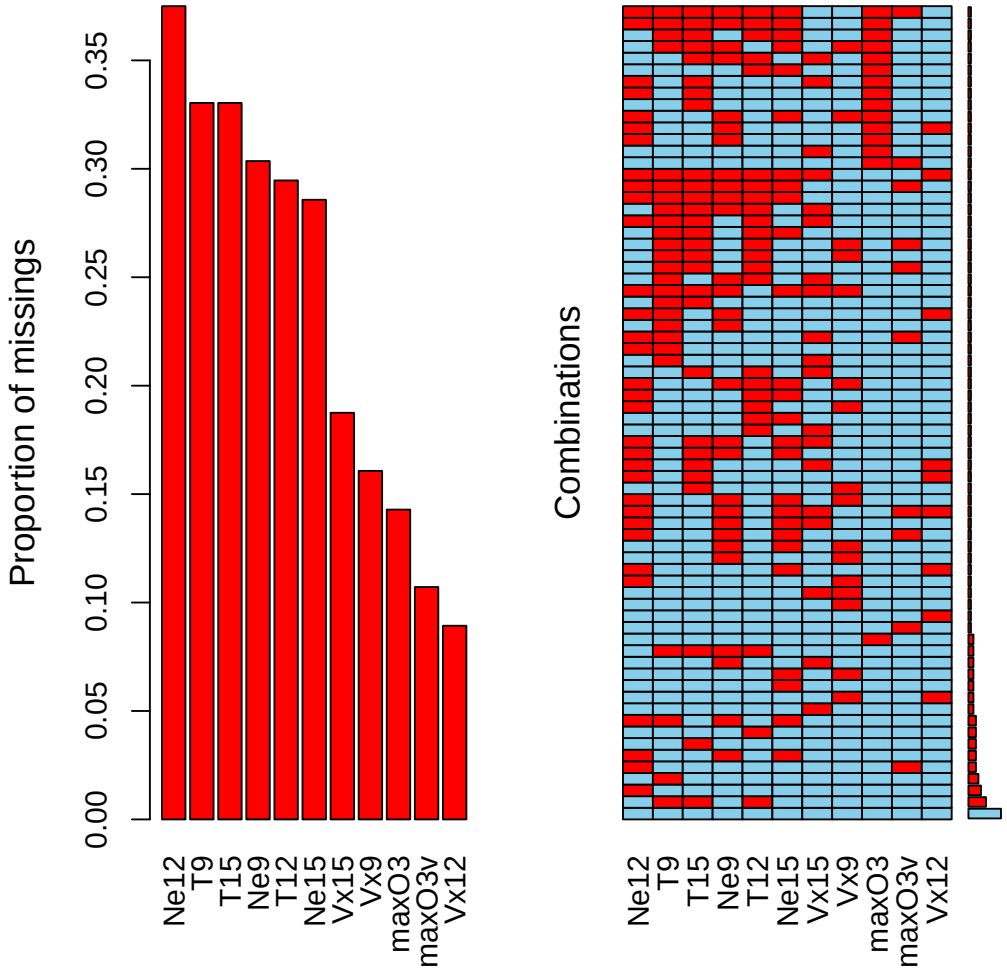
Count missing values

```
> library(missMDA)
> WindDirection <- ozo[,12]
> don <- ozo[,1:11]
> library(VIM)
> res <- summary(aggr(don, sortVar = TRUE))$combinations
> res[rev(order(res[, 2])),]
```

```
Variables sorted by
number of missings:
Variable      Count
    Ne12 0.37500000
      T9 0.33035714
     T15 0.33035714
     Ne9 0.30357143
     T12 0.29464286
    Ne15 0.28571429
    Vx15 0.18750000
     Vx9 0.16071429
   maxO3 0.14285714
  maxO3v 0.10714286
    Vx12 0.08928571
```

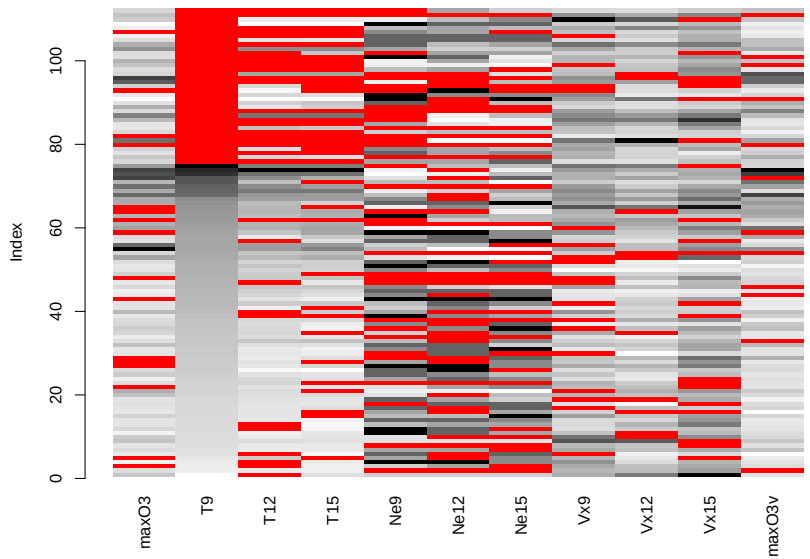
Combinations	Count	Percent
0:0:0:0:0:0:0:0:0:0:0	13	11.6071429
0:1:1:1:0:0:0:0:0:0:0	7	6.2500000
0:0:0:0:0:1:0:0:0:0:0	5	4.4642857
0:1:0:0:0:0:0:0:0:0:0	4	3.5714286
0:1:0:0:1:1:1:0:0:0:0	3	2.6785714
0:0:1:0:0:0:0:0:0:0:0	3	2.6785714
0:0:0:1:0:0:0:0:0:0:0	3	2.6785714
0:0:0:0:1:1:1:0:0:0:0	3	2.6785714
0:0:0:0:0:1:0:0:0:0:1	3	2.6785714
0:1:1:1:1:0:0:0:0:0:0	2	1.7857143
0:0:0:0:1:0:0:0:0:1:0	2	1.7857143
0:0:0:0:0:0:1:1:0:0:0	2	1.7857143
0:0:0:0:0:0:1:0:0:0:0	2	1.7857143
.....	.	...

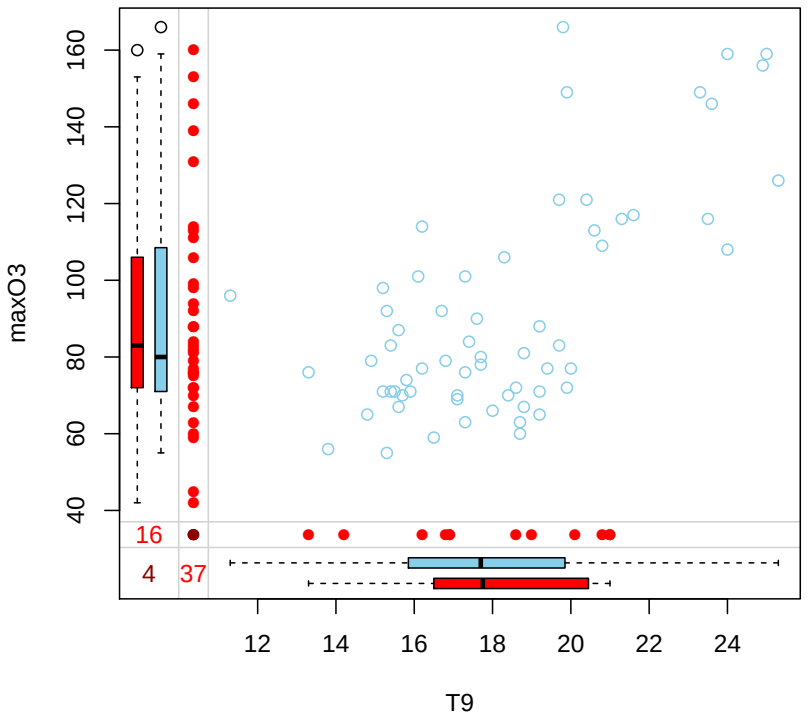
Pattern visualization




```
 #library(VIM)
> aggr(don, sortVar = TRUE)
```

Visualization






```
# library(VIM)
> matrixplot(don, sortby = 2)
> marginplot(don[ ,c("T9", "maxO3")])
```

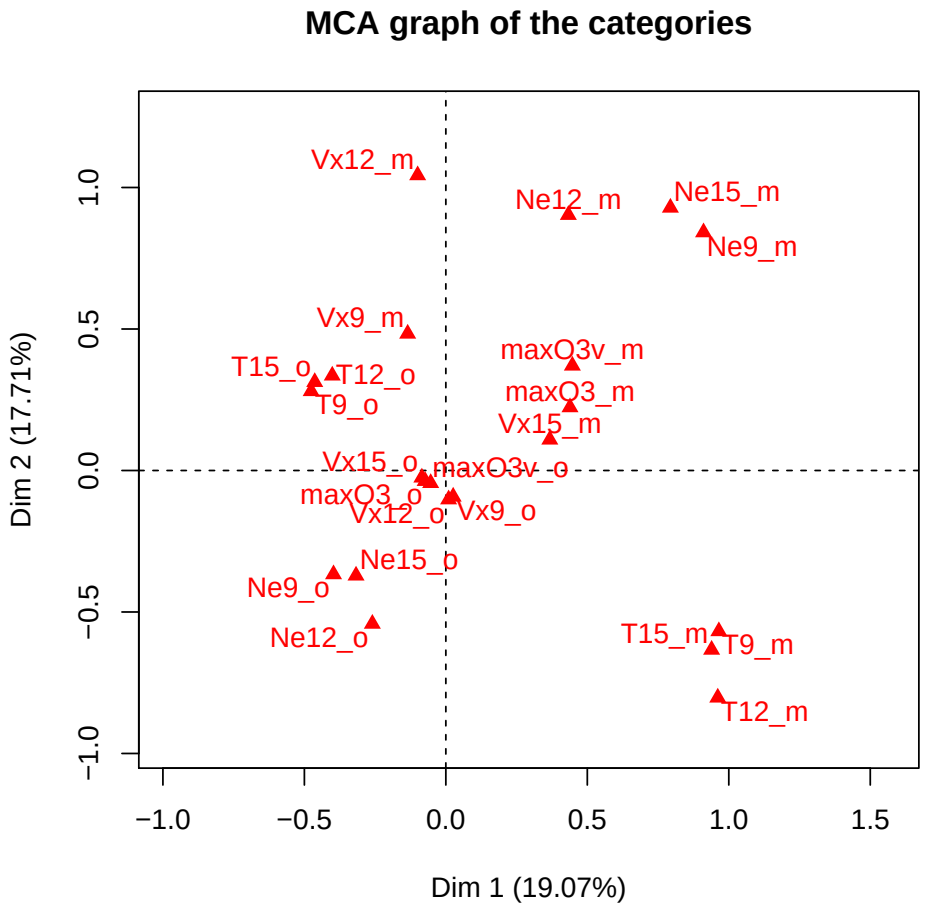
Visualization with Multiple Correspondence Analysis

$\Rightarrow$ Create the missingness matrix

```
> mis.ind <- matrix("o", nrow = nrow(don), ncol = ncol(don))
> mis.ind[is.na(don)] = "m"
> dimnames(mis.ind) = dimnames(don)
> mis.ind
```

```
         maxO3 T9  T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
20010601 "o"   "o" "o" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010602 "o"   "m" "m" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010603 "o"   "o" "o" "o" "o" "m"  "m"  "o" "m"  "o"  "o"
20010604 "o"   "o" "o" "m" "o" "o"  "o"  "m" "o"  "o"  "o"
20010605 "o"   "m" "o" "o" "m" "m"  "m"  "o" "o"  "o"  "o"
20010606 "o"   "o" "o" "o" "o" "m"  "o"  "o" "o"  "o"  "o"
20010607 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
20010610 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
```

Visualization with Multiple Correspondence Analysis




```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA, invis = "ind", title = "MCA graph of the categories")
```

Missing data mechanism

Pattern of missingness: M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

Missing data mechanism

Pattern of missingness: M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

- Missing completely at random (MCAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M)$$

Missing data mechanism

Pattern of missingness: M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

- Missing completely at random (MCAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M)$$

- Missing at random (MAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M|x_{\mathrm{obs}})$$

Missing data mechanism

Pattern of missingness: M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

- Missing completely at random (MCAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M)$$

- Missing at random (MAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M|x_{\mathrm{obs}})$$

- Missing not at random (MNAR):

$$\mathrm{p}(M|x_{\mathrm{obs}}, x_{\mathrm{mis}}) = \mathrm{p}(M|x_{\mathrm{mis}})$$

Example: age and income.

Outline

Imputation methods

- PCA or MCA
  R package: `missMDA`
- *k*-nearest neighbor
  R packages: `VIM`, `yaImpute`, `impute`
- random forest
  R package: `missForest`
- chained equation (conditional distribution)
  R packages: `mice`

⇒ R-miss-tastic (Josse et al.): Methods and references for managing missing data
⇒ Flexible imputation of missing data (Stef van Buuren)

PCA (for complete data)

Principal component analysis:
Find the subspace that best represents the data

Figure: Camel or dromedary?

$\Rightarrow$ Best approximation with projection

PCA (for complete data)

Principal component analysis:
Find the subspace that best represents the data

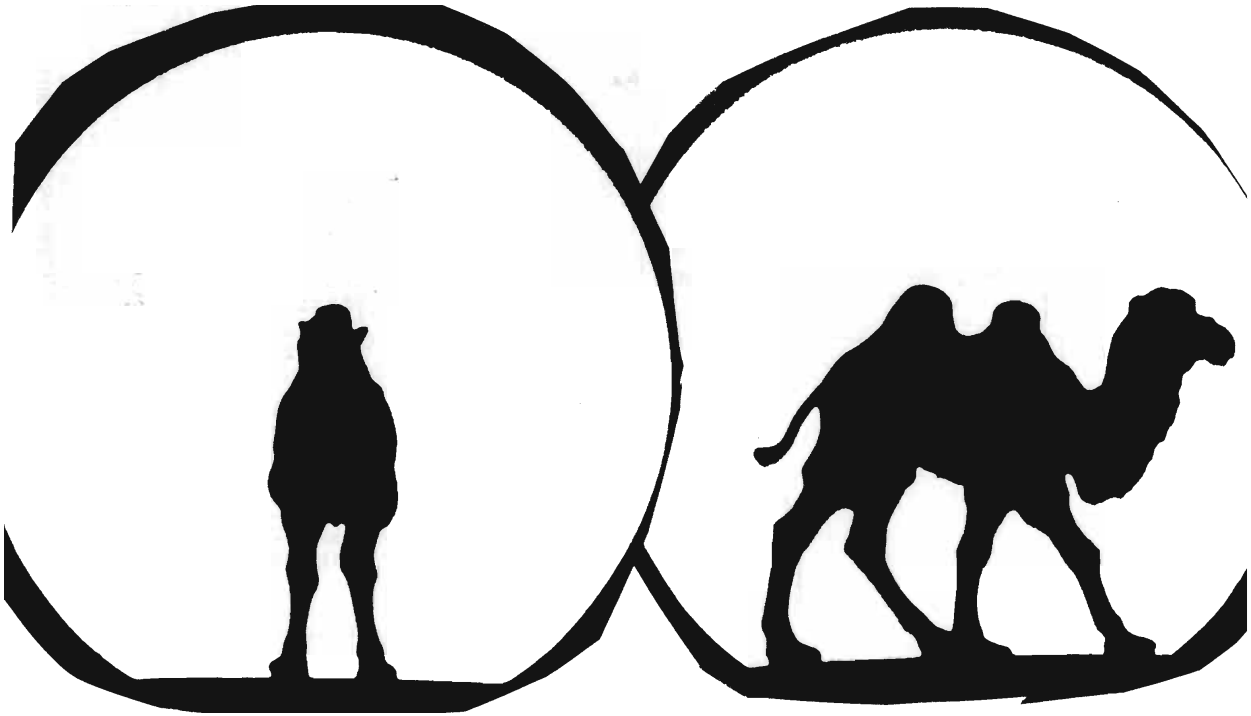


Figure: Camel or dromedary? source J.P. Fénelon

$\Rightarrow$ Best approximation with projection
$\Rightarrow$ Best representation of the variability

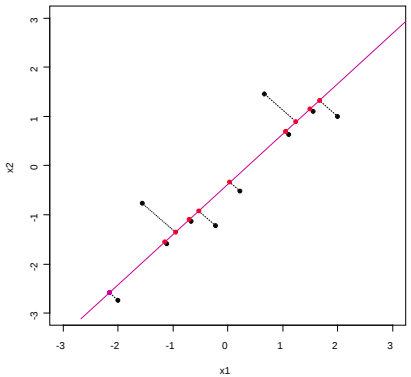
PCA reconstruction

X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33



$$X \approx F \; V'\hat{\mu}$$

$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq S \right\}$$

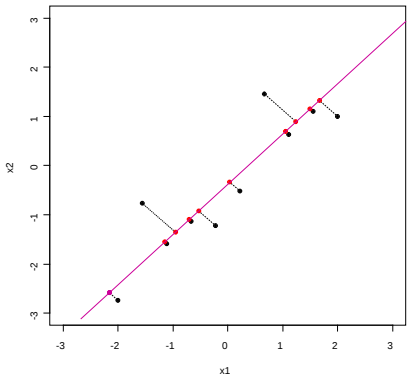
PCA reconstruction

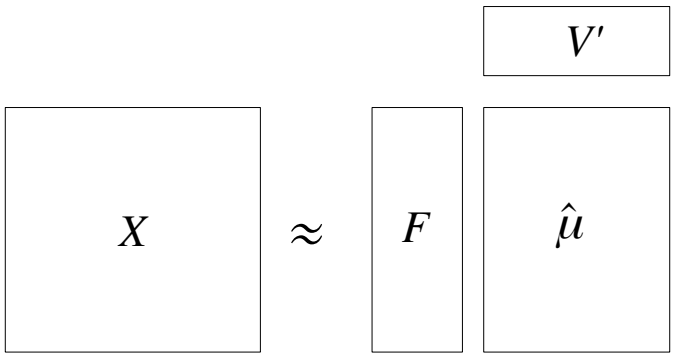
X

-2.00	-2.74
NA	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	NA
0.22	-0.52
0.67	1.46
NA	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \text{tr}(AA^\top)$:

$$\text{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

SVD X: $\hat{\mu}^{\text{PCA}} = U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S}$ $F = U\Lambda^{\frac{1}{2}}$ PC - scores

$= F_{n\times S}V'_{p\times S}$ V principal axes - loadings

Missing values in PCA

$\Rightarrow$ PCA: least squares

$$\text{argmin}_{\mu} \left\{ \|X_{n\times p} - \mu_{n\times p}\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

$\Rightarrow$ PCA with missing values: weighted least squares

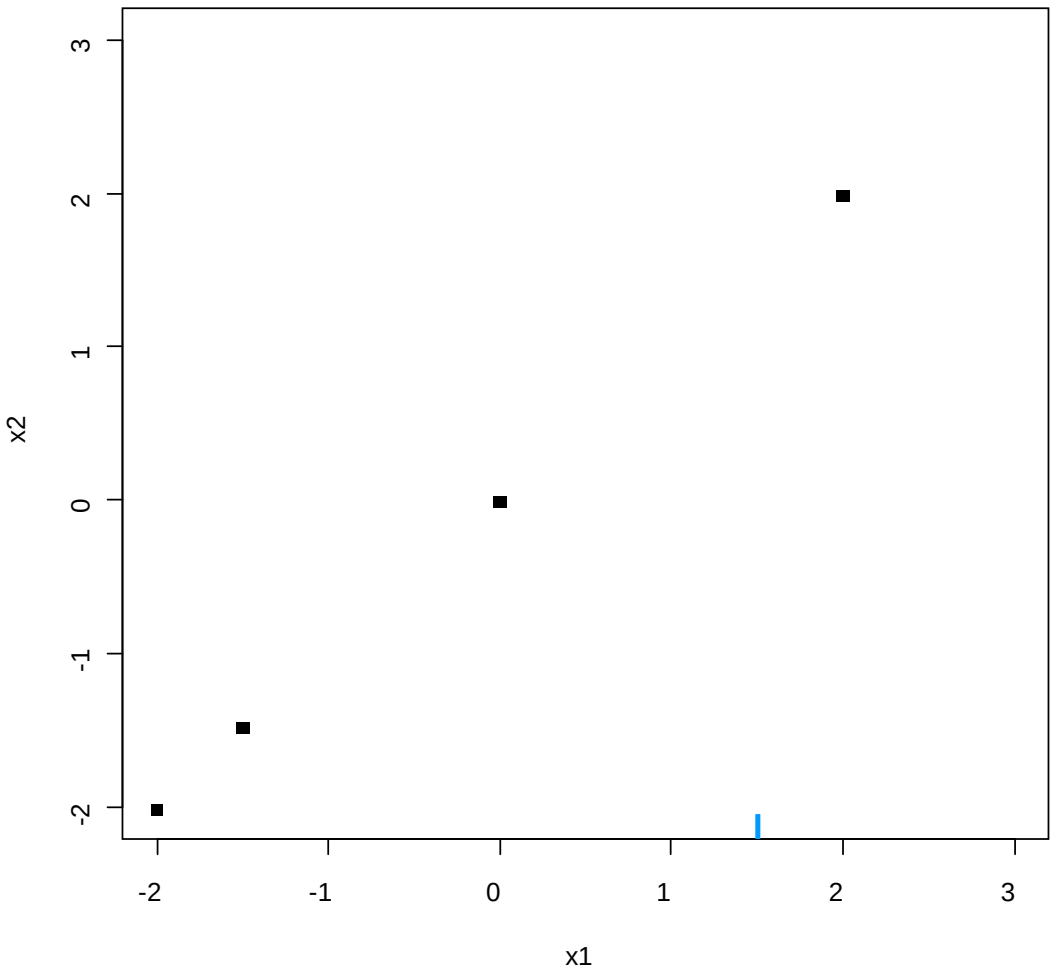
$$\text{argmin}_{\mu} \left\{ \|W_{n\times p} * (X - \mu)\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise; $*$ elementwise multiplication

Many algorithms: weighted alternating least squares (Gabriel & Zamir, 1979); iterative PCA (Kiers, 1997)

Iterative PCA

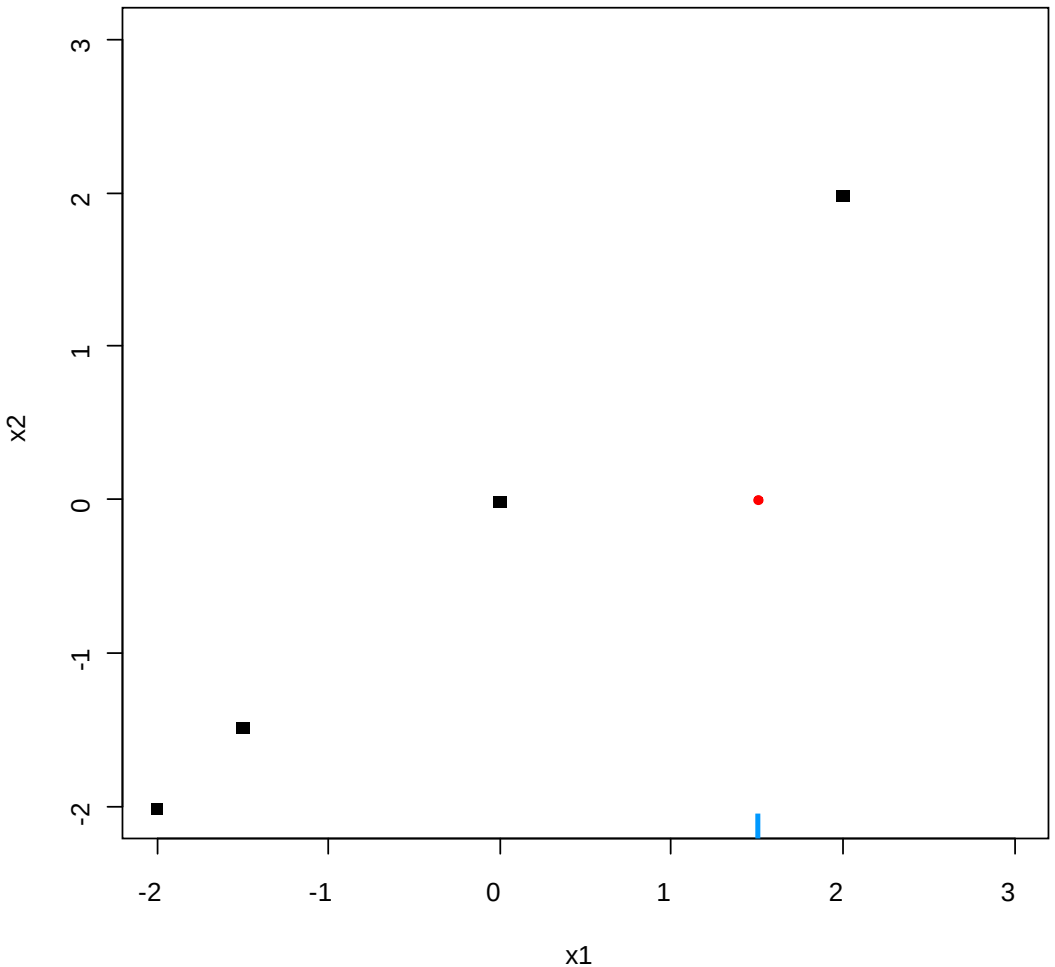
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



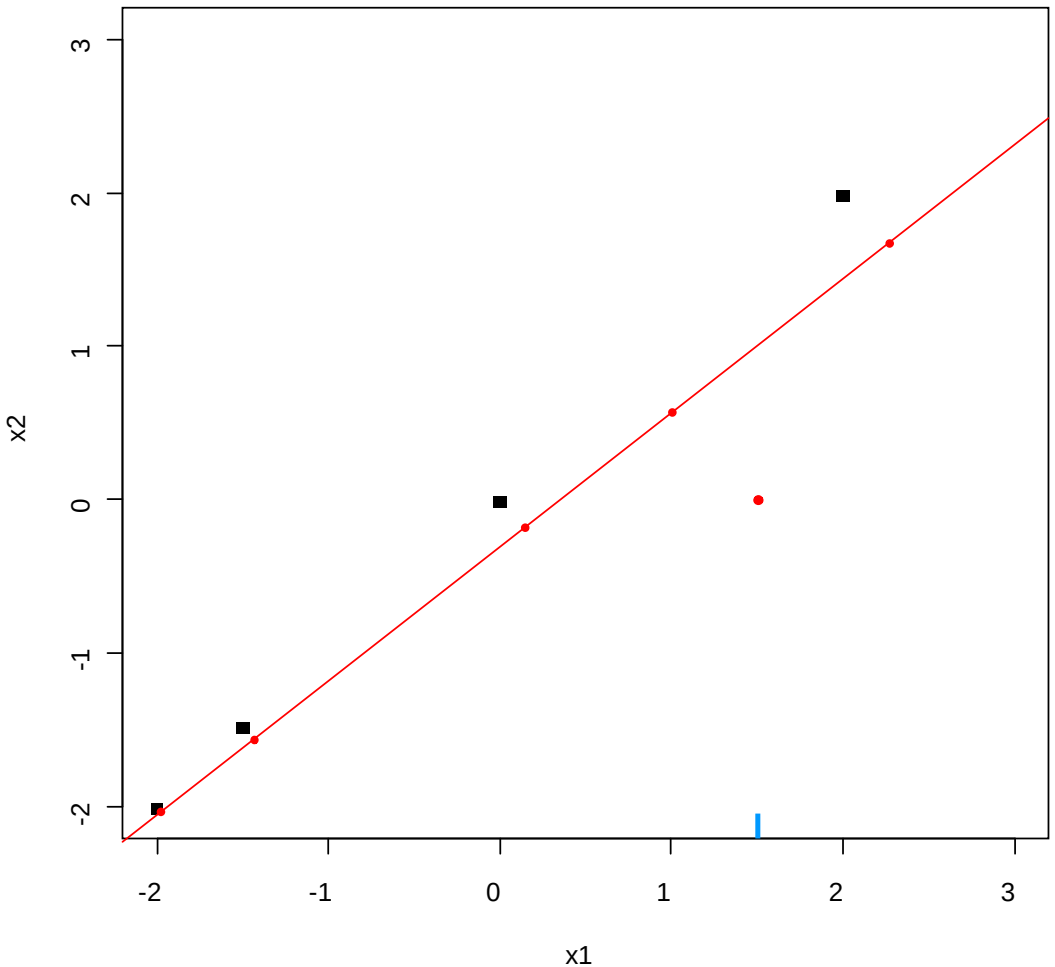
Initialization $\ell = 0$: X^0 (mean imputation)

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



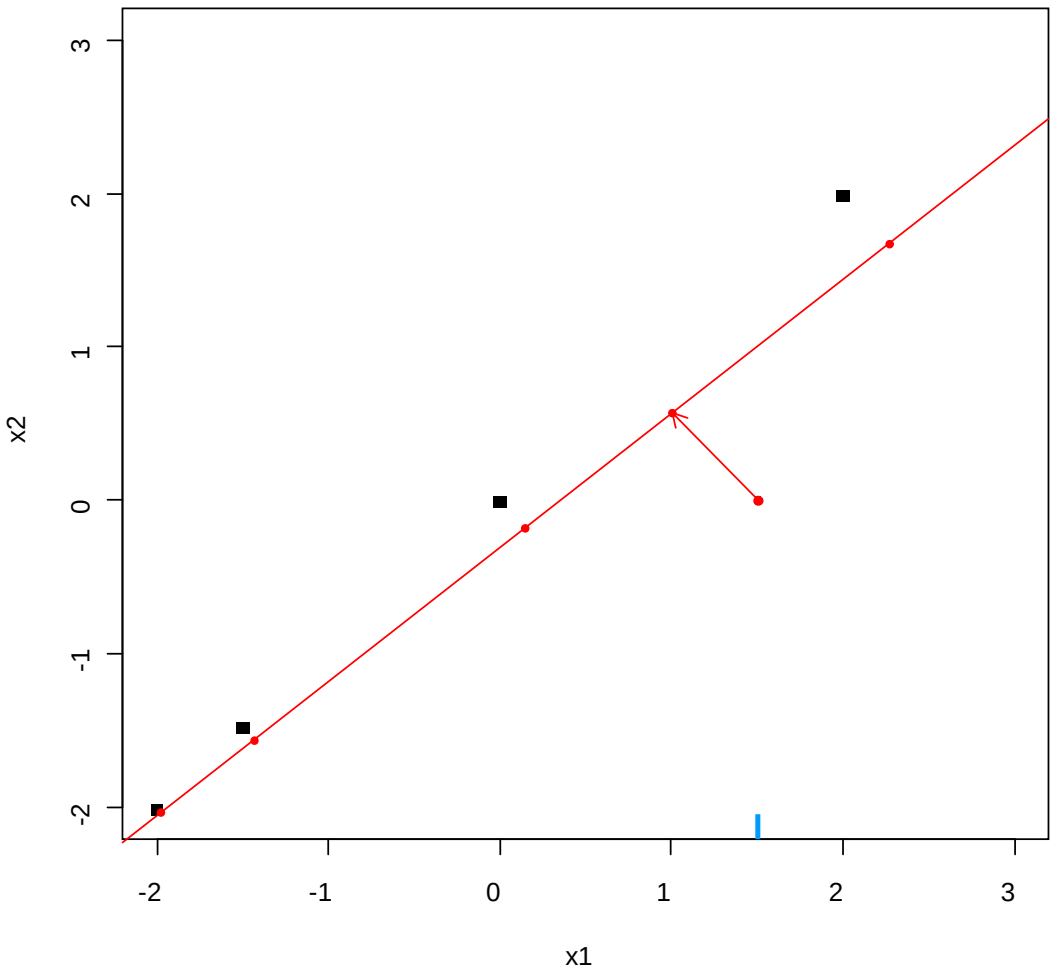
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

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



Missing values imputed with the fitted matrix $\hat{\mu}^{\ell} = U^{\ell}\Lambda^{1/2^{\ell}}V^{\ell\prime}$

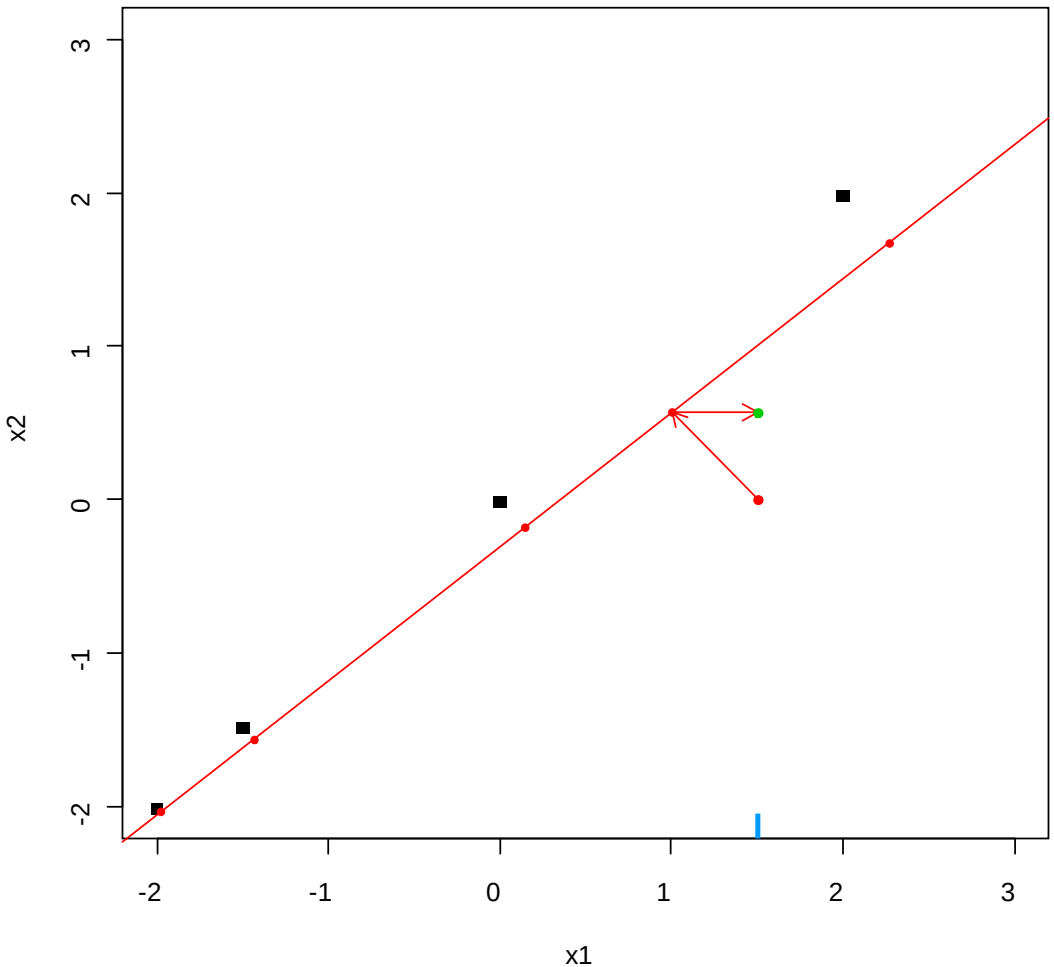
Iterative PCA

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



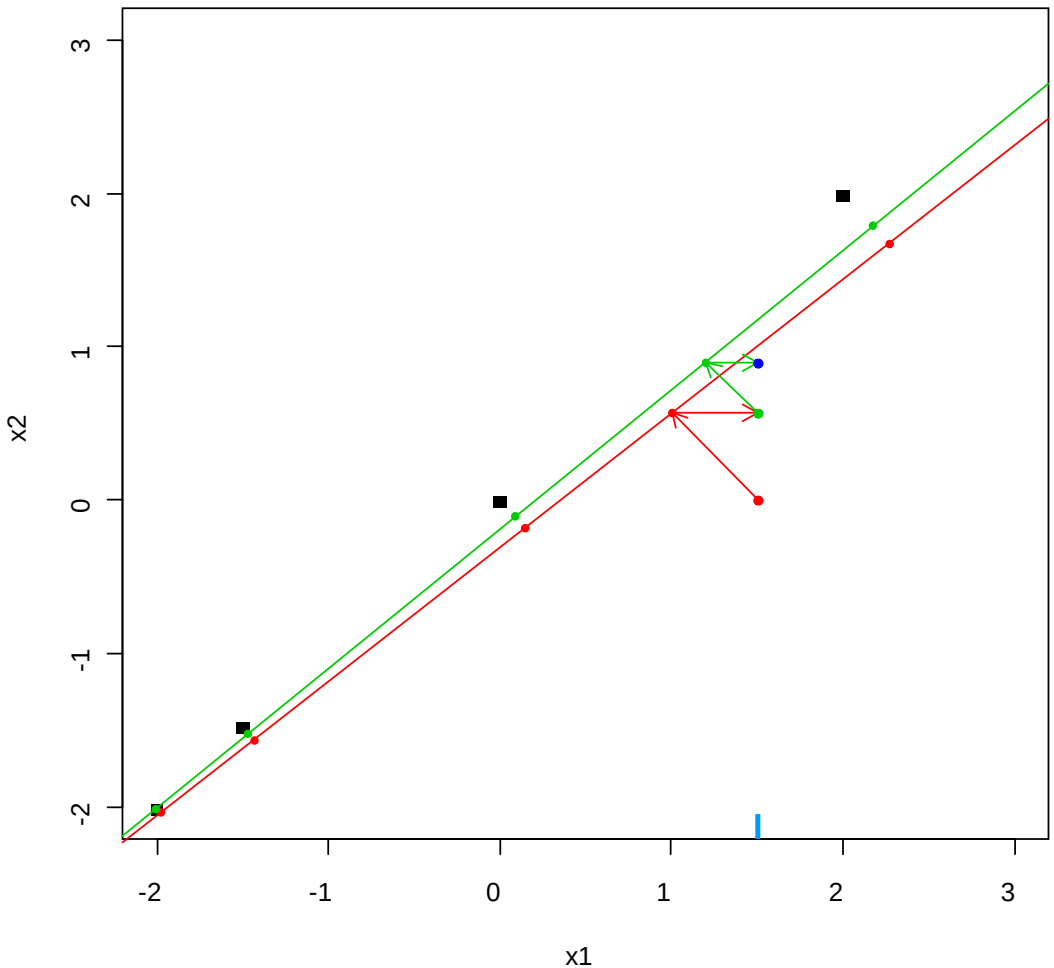
The new imputed dataset is $\hat{X}^{\ell} = W * X + (\mathbf{1} - W) * \hat{\mu}^{\ell}$

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



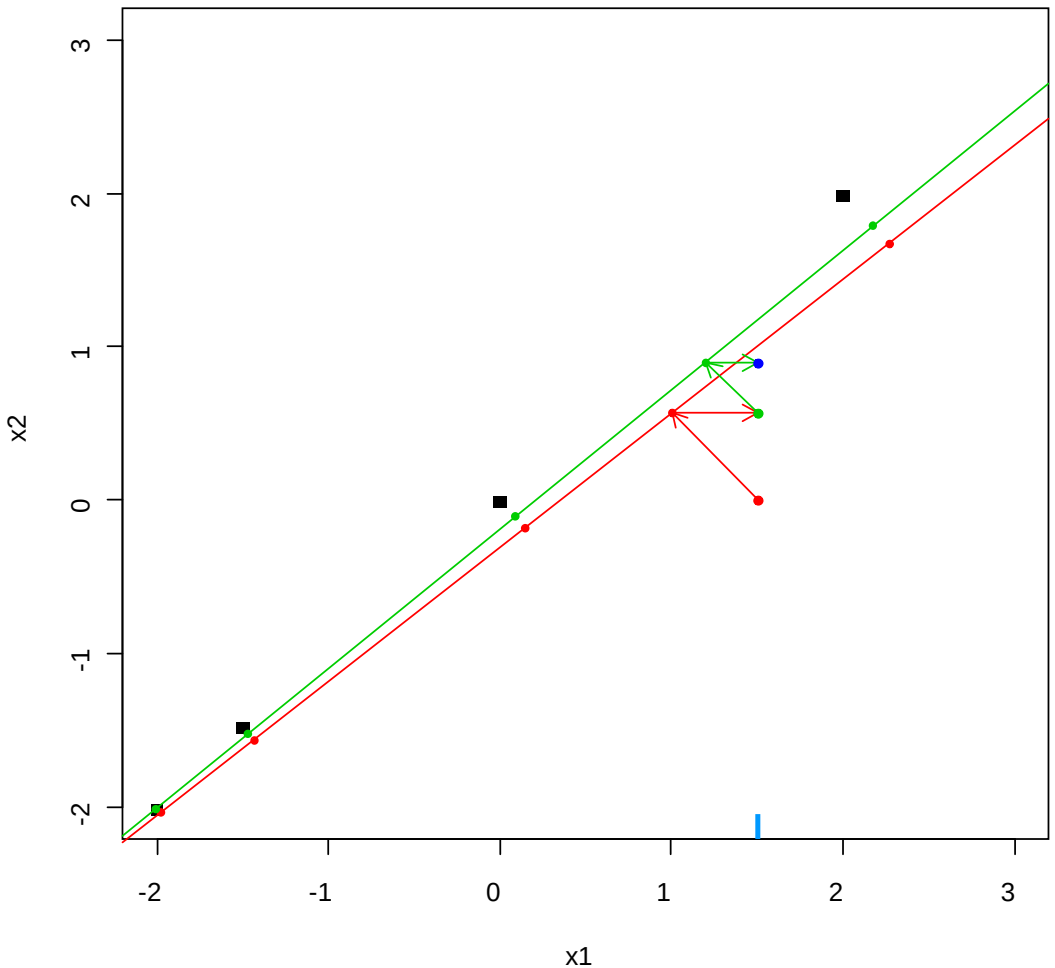
Iterative PCA

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



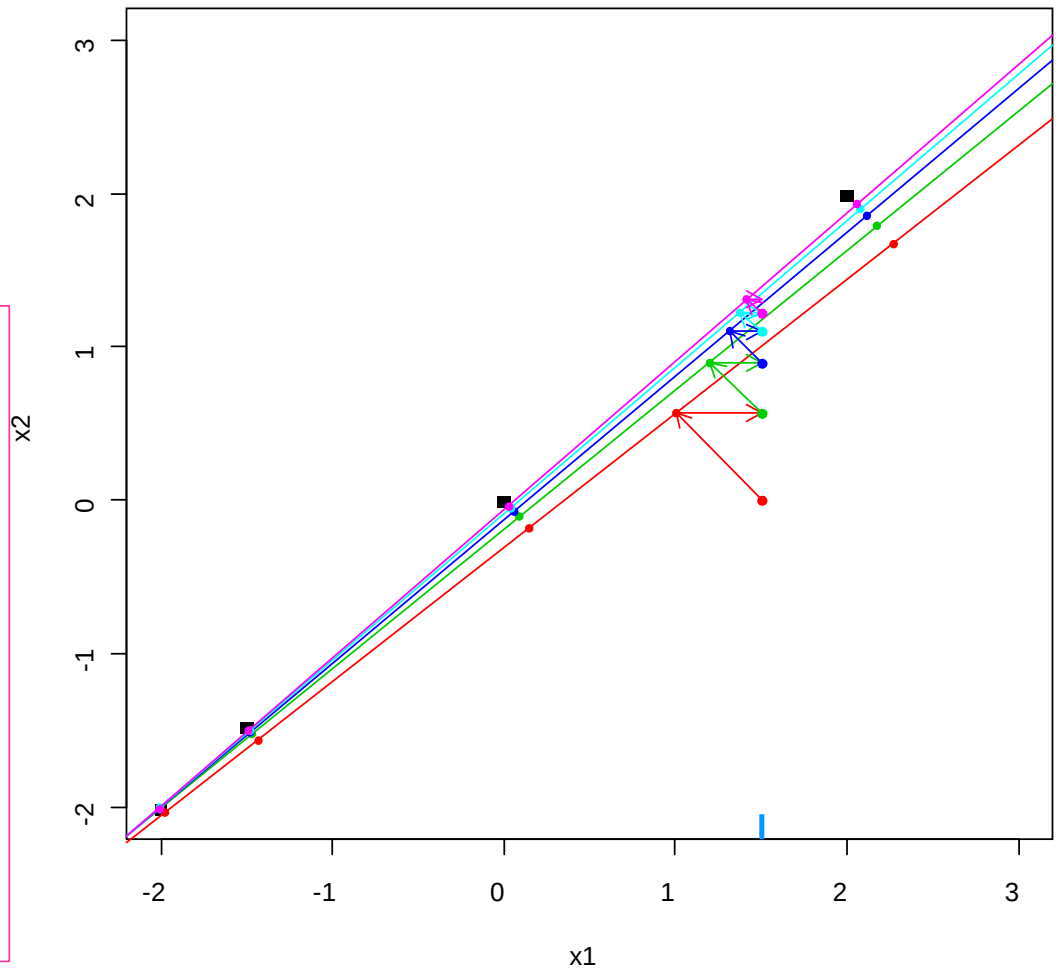
Iterative PCA

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

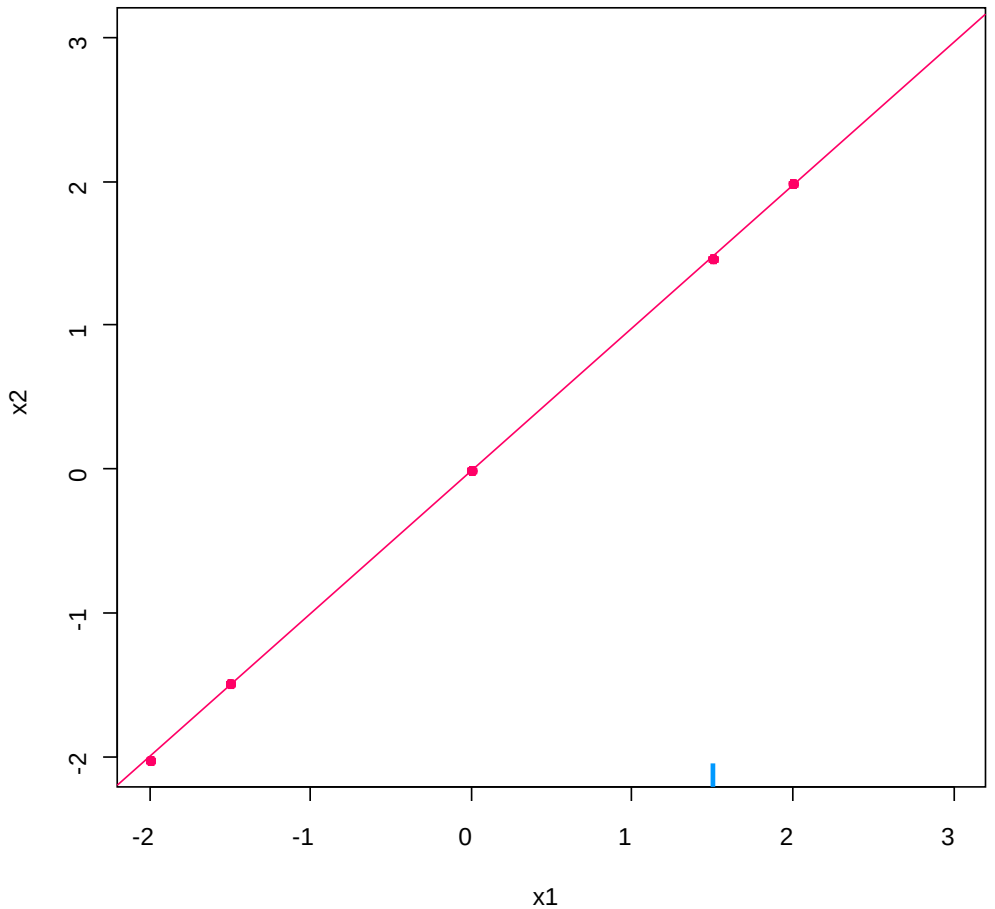
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Steps are repeated until convergence

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	**1.46**
2.0	1.98

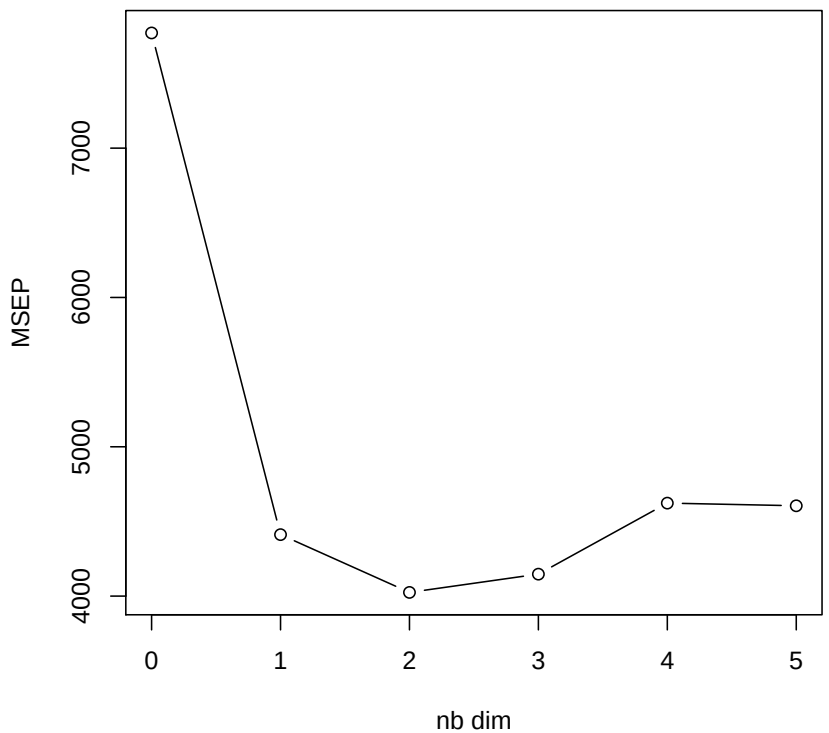
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$

Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$

Imputation with Principal Component Analysis in practice

$\Rightarrow$ Step 1: Estimation of the number of dimensions (Cross Validation)

```
> library(missMDA)
> nb <- estim_ncpPCA(don, method.cv = "Kfold")
> nb$ncp    #2
> plot(0:5, nb$criterion, xlab = "nb dim", ylab ="MSEP")
```



Imputation with PCA in practice

$\Rightarrow$ Step 2: Imputation of the missing values

```
> res.comp <- imputePCA(don, ncp = 2)
> res.comp$completeObs[1:3, ]
     maxO3    T9   T12   T15 Ne9 Ne12 Ne15   Vx9  Vx12  Vx15 maxO3v
0601    87 15.60 18.50 20.47   4 4.00 8.00  0.69 -1.71 -0.69     84
0602    82 18.51 20.88 21.81   5 5.00 7.00 -4.33 -4.00 -3.00     87
0603    92 15.30 17.60 19.50   2 3.98 3.81  2.95  1.97  0.52     82
```

Properties of imputation with PCA

- Very good quality of imputation. Using similarities between individuals and relationship between variables.
  Popular in machine learning with recommandation systems (Netflix: 99% missing).
- Model makes sense: Data = structure of low rank S + noise (Udell & Townsend, 2017)

Properties of imputation with PCA

- Very good quality of imputation. Using similarities between individuals and relationship between variables.
  Popular in machine learning with recommandation systems (Netflix: 99% missing).
- Model makes sense: Data = structure of low rank S + noise (Udell & Townsend, 2017)

Q: How about random forest imputation? Any Difference?

Iterative Random Forests imputation

1. Initial imputation: mean imputation - random category
   Sort the variables according to the amount of missing values
2. Fit a RF X_j^{obs} on variables X_{-j}^{obs} and then predict X_j^{miss}
3. Cycling through variables
4. Repeat step 2 and 3 until convergence

- number of trees: 100
- number of variables randomly selected at each node $\sqrt{p}$
- number of iterations: 4-5

Implemented in the R package `missForest`
(Daniel J. Stekhoven, Peter Buhlmann, 2011)

Missing At Random

	Feat1	Feat2	Feat3	Feat4	Feat5...
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	NA	NA	8	8
Frank	8	NA	NA	8	8
Bertrand	9	NA	NA	9	9
Alex	9	NA	NA	9	9
Yohann	10	NA	NA	10	10
Jean	10	NA	NA	10	10

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1.0	1.00	1	1
C2	1	1.0	1.00	1	1
C3	2	2.0	2.00	2	2
C4	2	2.0	2.00	2	2
C5	3	3.0	3.00	3	3
C6	3	3.0	3.00	3	3
C7	4	4.0	4.00	4	4
C8	4	4.0	4.00	4	4
C9	5	5.0	5.00	5	5
C10	5	5.0	5.00	5	5
C11	6	6.0	6.00	6	6
C12	6	6.0	6.00	6	6
C13	7	7.0	7.00	7	7
C14	7	7.0	7.00	7	7
Igor	8	6.87	6.87	8	8
Frank	8	6.87	6.87	8	8
Bertrand	9	6.87	6.87	9	9
Alex	9	6.87	6.87	9	9
Yohann	10	6.87	6.87	10	10
Jean	10	6.87	6.87	10	10

$\Rightarrow$ with Random Forests

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	8	8	8	8
Frank	8	8	8	8	8
Bertrand	9	9	9	9	9
Alex	9	9	9	9	9
Yohann	10	10	10	10	10
Jean	10	10	10	10	10

$\Rightarrow$ with PCA

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1.0	1.00	1	1
C2	1	1.0	1.00	1	1
C3	2	2.0	2.00	2	2
C4	2	2.0	2.00	2	2
C5	3	3.0	3.00	3	3
C6	3	3.0	3.00	3	3
C7	4	4.0	4.00	4	4
C8	4	4.0	4.00	4	4
C9	5	5.0	5.00	5	5
C10	5	5.0	5.00	5	5
C11	6	6.0	6.00	6	6
C12	6	6.0	6.00	6	6
C13	7	7.0	7.00	7	7
C14	7	7.0	7.00	7	7
Igor	8	6.87	6.87	8	8
Frank	8	6.87	6.87	8	8
Bertrand	9	6.87	6.87	9	9
Alex	9	6.87	6.87	9	9
Yohann	10	6.87	6.87	10	10
Jean	10	6.87	6.87	10	10

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	8	8	8	8
Frank	8	8	8	8	8
Bertrand	9	9	9	9	9
Alex	9	9	9	9	9
Yohann	10	10	10	10	10
Jean	10	10	10	10	10

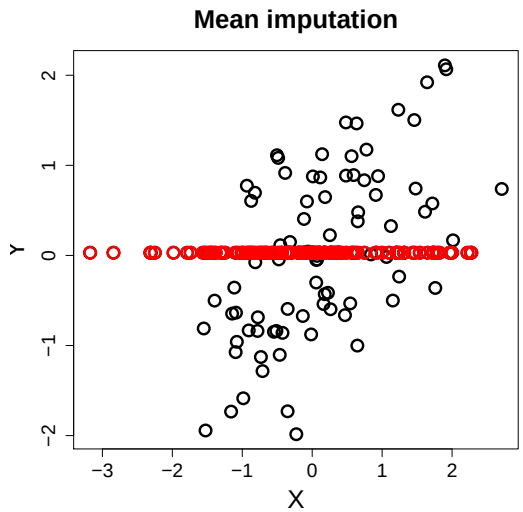
$\Rightarrow$ with Random Forests $\Rightarrow$ with PCA

$\Rightarrow$ forests: Non-linear relationship & variable interactions

Outline

Is single imputation a safe method?

Evaluate the variability of single imputation:



$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	

Confidence interval for a mean

Let $Y = (Y_1, \ldots, Y_n)'$ i.i.d. $\sim \mathcal{N}(\mu_y, \sigma_y^2)$

A confidence interval for μ:

- when variance known

$$\mathbb{P}\left(\bar{Y} - \frac{\sigma_y}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{Y} + \frac{\sigma_y}{\sqrt{n}} z_{1-\alpha/2}\right) = 1 - \alpha$$

- when variance unknown

$$\mathbb{P}\left(\bar{y} - \frac{\widehat{\sigma}_y}{\sqrt{n}} t_{1-\alpha/2}(n-1)\ ,\ \bar{y} + \frac{\hat{\sigma}_y}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right) = 1 - \alpha$$

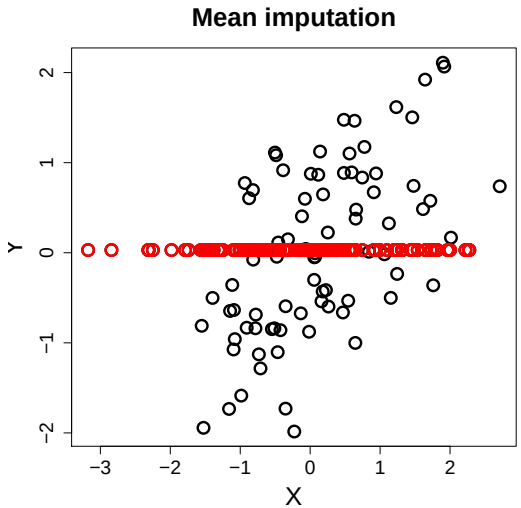
e.g. $\alpha = 0.05 \Rightarrow 95\%$ confidence

Simulation

1. Generate bivariate Gaussian data ($\mu_y = 0, \sigma_y = 1, \rho = 0.6$)
2. Put missing values on y
3. Impute missing entries
4. Compute the confidence interval of $\mu_y \Rightarrow$
   count if the true value $\mu_y = 0$ is in the confidence interval
5. Repeat the steps 10000 times

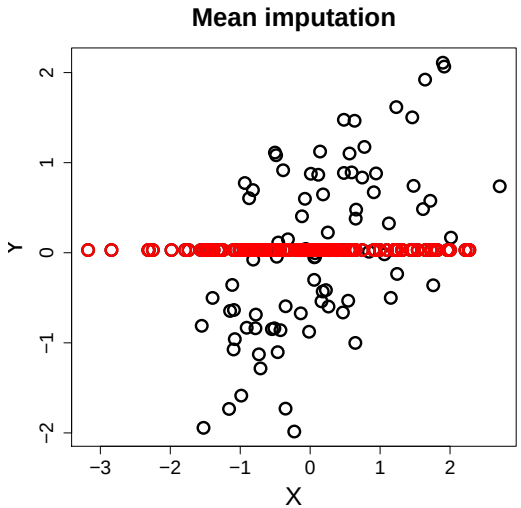
$\Rightarrow$ coverage of confidence interval $= 95\%$?

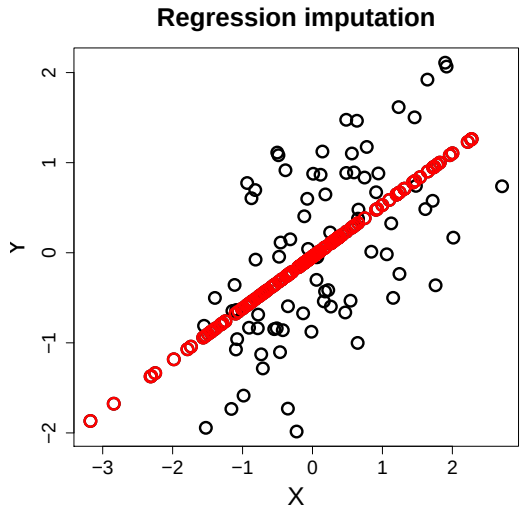
Single imputation methods



$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	39.4

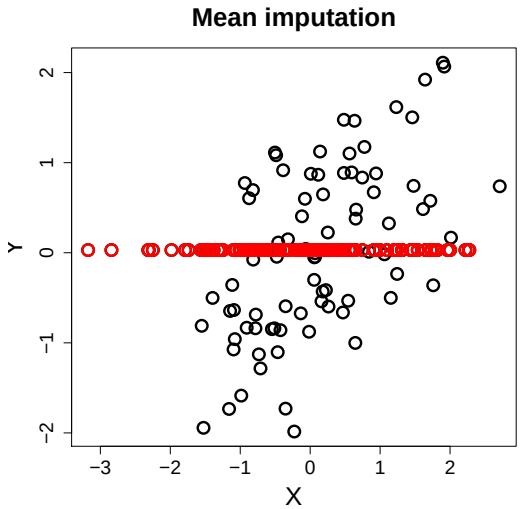
Single imputation methods

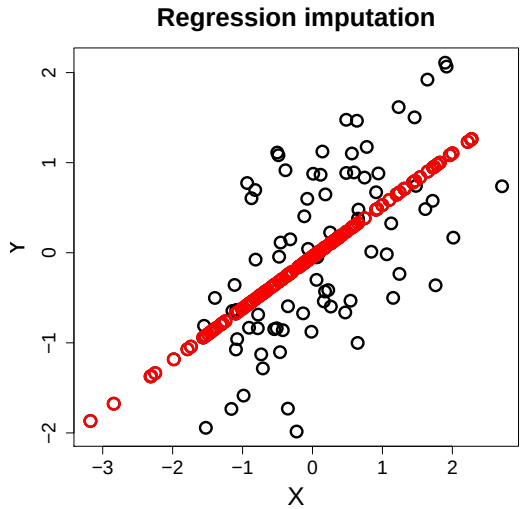


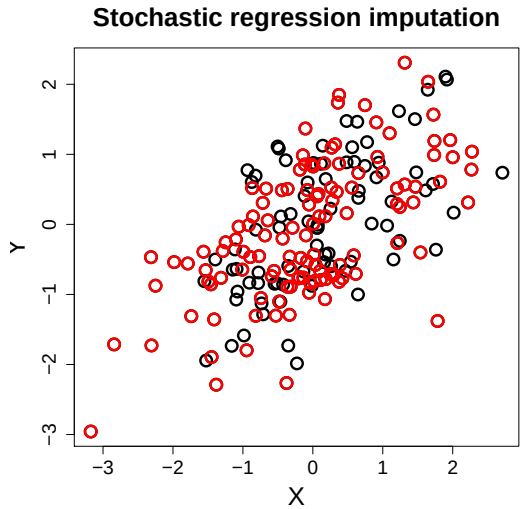


$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.5	0.72
$\rho = 0.6$	0.30	0.78
$CI\mu_y 95\%$	39.4	61.6

Single imputation methods

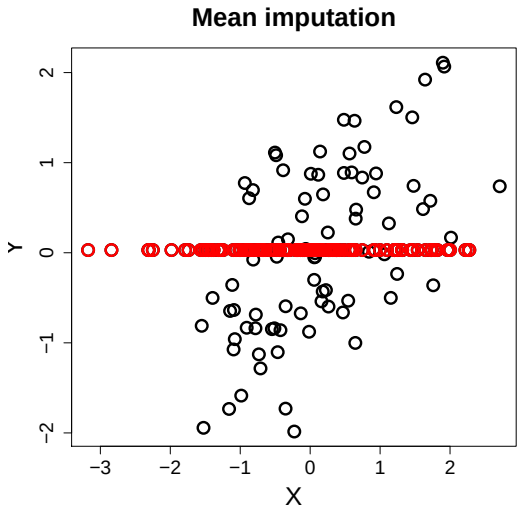


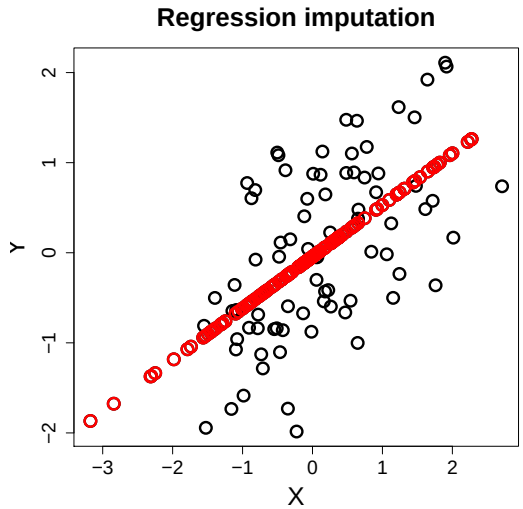


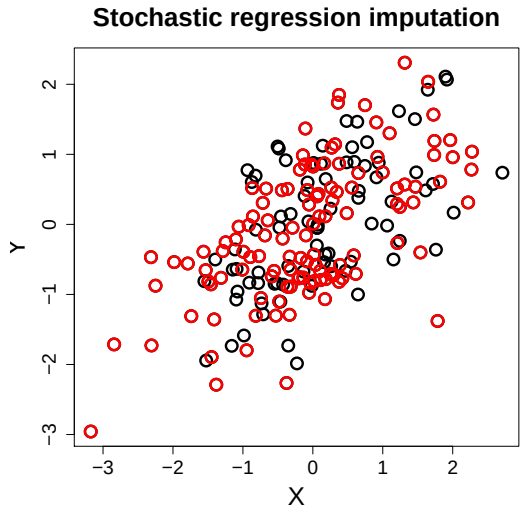


$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

Single imputation methods







$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

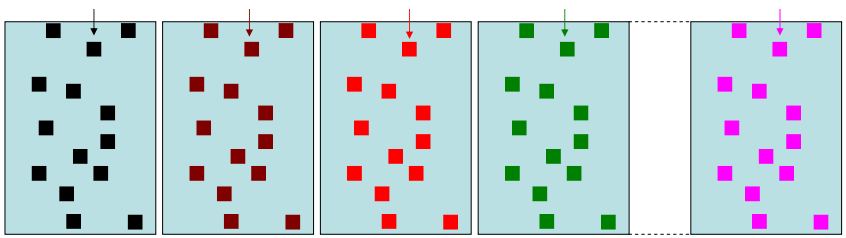
$\Rightarrow$ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated underestimated

$\Rightarrow$ A single value cannot reflect the uncertainty of the predictions.

Multiple imputation (Rubin, 1987)

Aim: provide the variability of estimation (taken into account the variability due to missing values)

1. Generating M imputed data sets

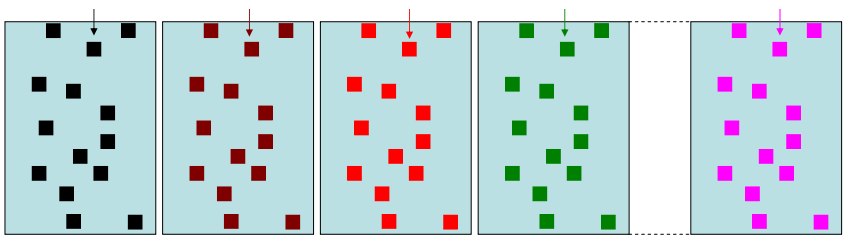


2. Performing the analysis on each imputed data set

Multiple imputation (Rubin, 1987)

Aim: provide the variability of estimation (taken into account the variability due to missing values)

1. Generating M imputed data sets



2. Performing the analysis on each imputed data set
3. Combining:

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

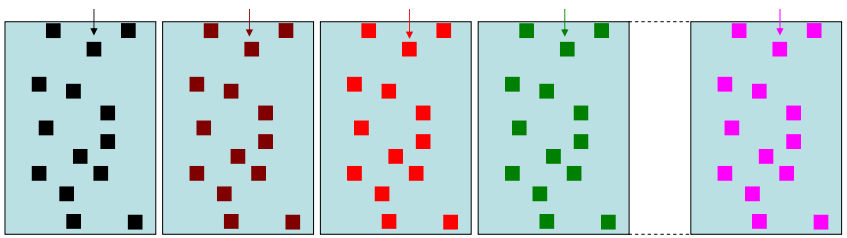
$$T = \frac{1}{M} \sum_m Var\left(\hat{\beta}_m\right) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_m \left(\hat{\beta}_m - \hat{\beta}\right)^2$$

variance = within + among imputed datasets

Multiple imputation (Rubin, 1987)

Aim: provide the variability of estimation (taken into account the variability due to missing values)

1. Generating M imputed data sets



2. Performing the analysis on each imputed data set
3. Combining:

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

$$T = \frac{1}{M} \sum_m Var\left(\hat{\beta}_m\right) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_m \left(\hat{\beta}_m - \hat{\beta}\right)^2$$

variance = within + among imputed datasets

packages: `mice`, `Amelia`, `missMDA`

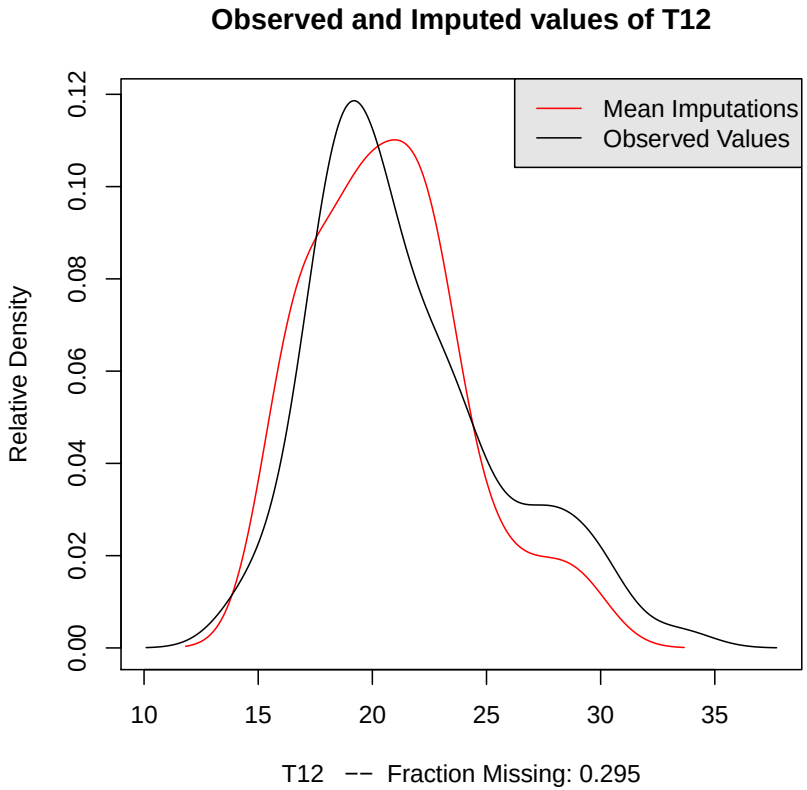
Multiple imputation in practice

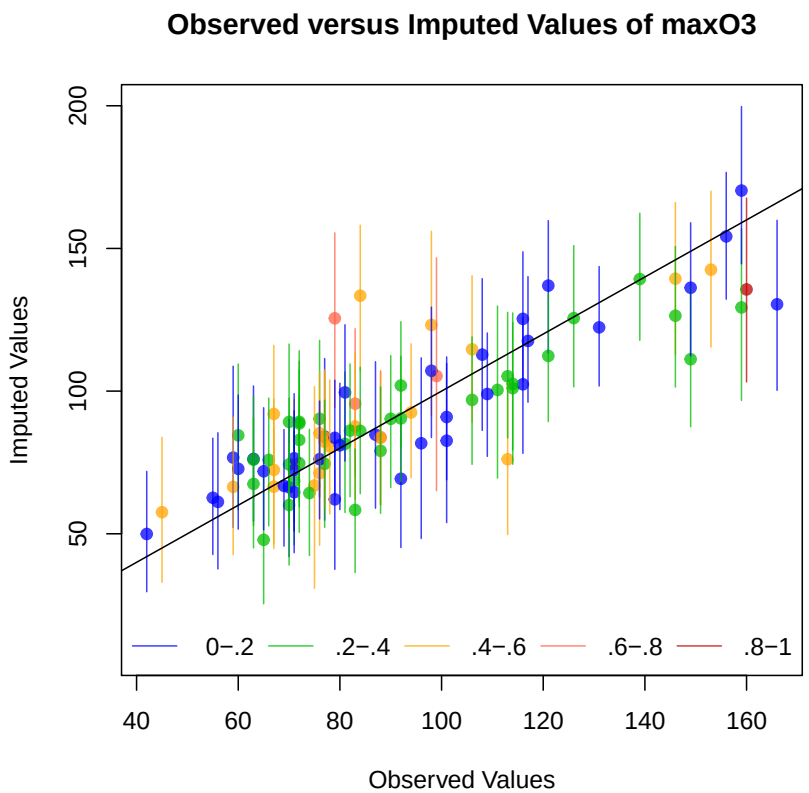
$\Rightarrow$ Step 1: Generate M imputed data sets

```
> library(Amelia)
> res.amelia <- amelia(don, m = 100)

> library(mice)
> res.mice <- mice(don, m = 100, defaultMethod = "norm.boot")

> library(missMDA)
>  res.MIPCA <- MIPCA(don, ncp = 2, nboot  = 100)
> res.MIPCA$res.MI
```

Multiple imputation in practice

⇒ Step 2: visualization



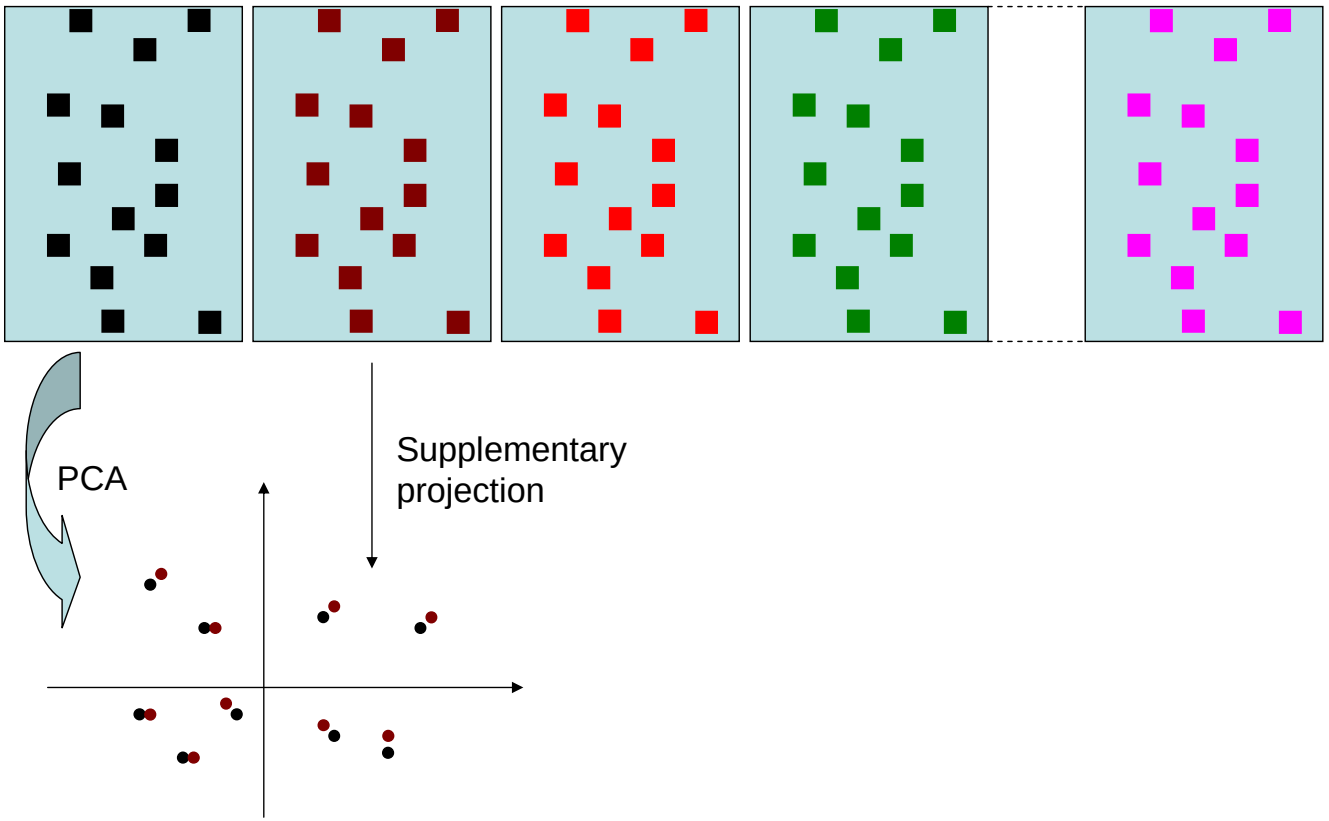



```
# library(Amelia)
> res.amelia <- amelia(don, m = 100)
> compare.density(res.amelia, var = "T12")
> overimpute(res.amelia, var = "maxO3")

# library(missMDA)
res.over<-Overimpute(res.MIPCA)
```

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



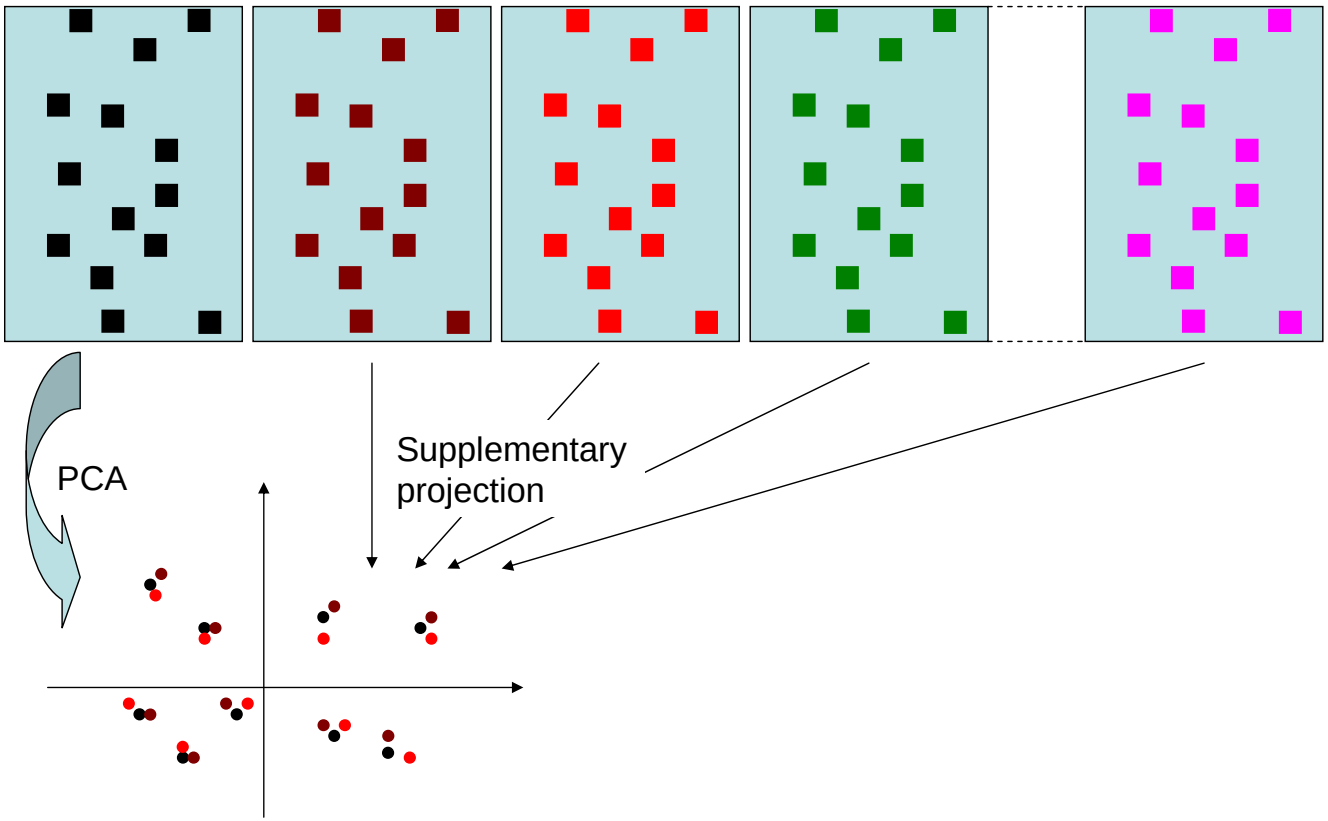
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



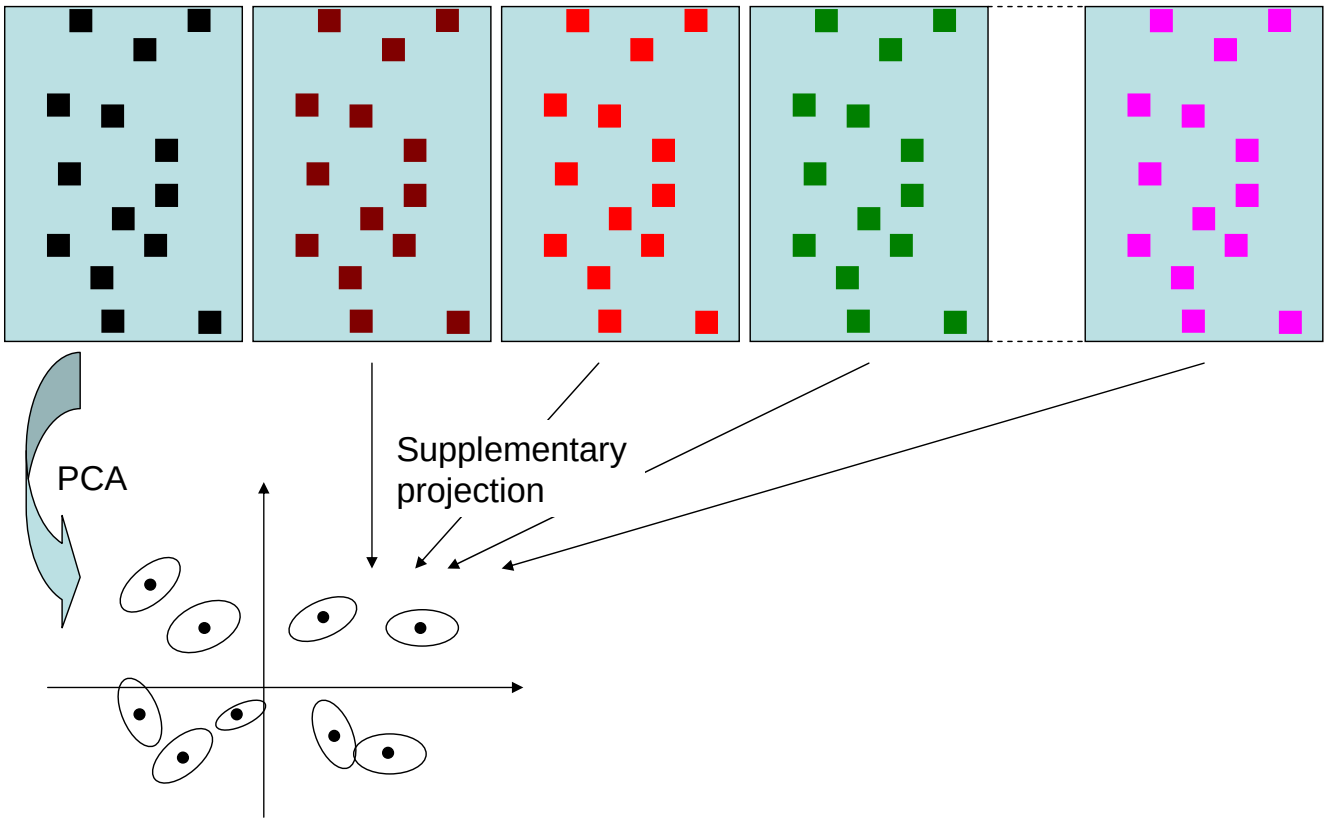
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

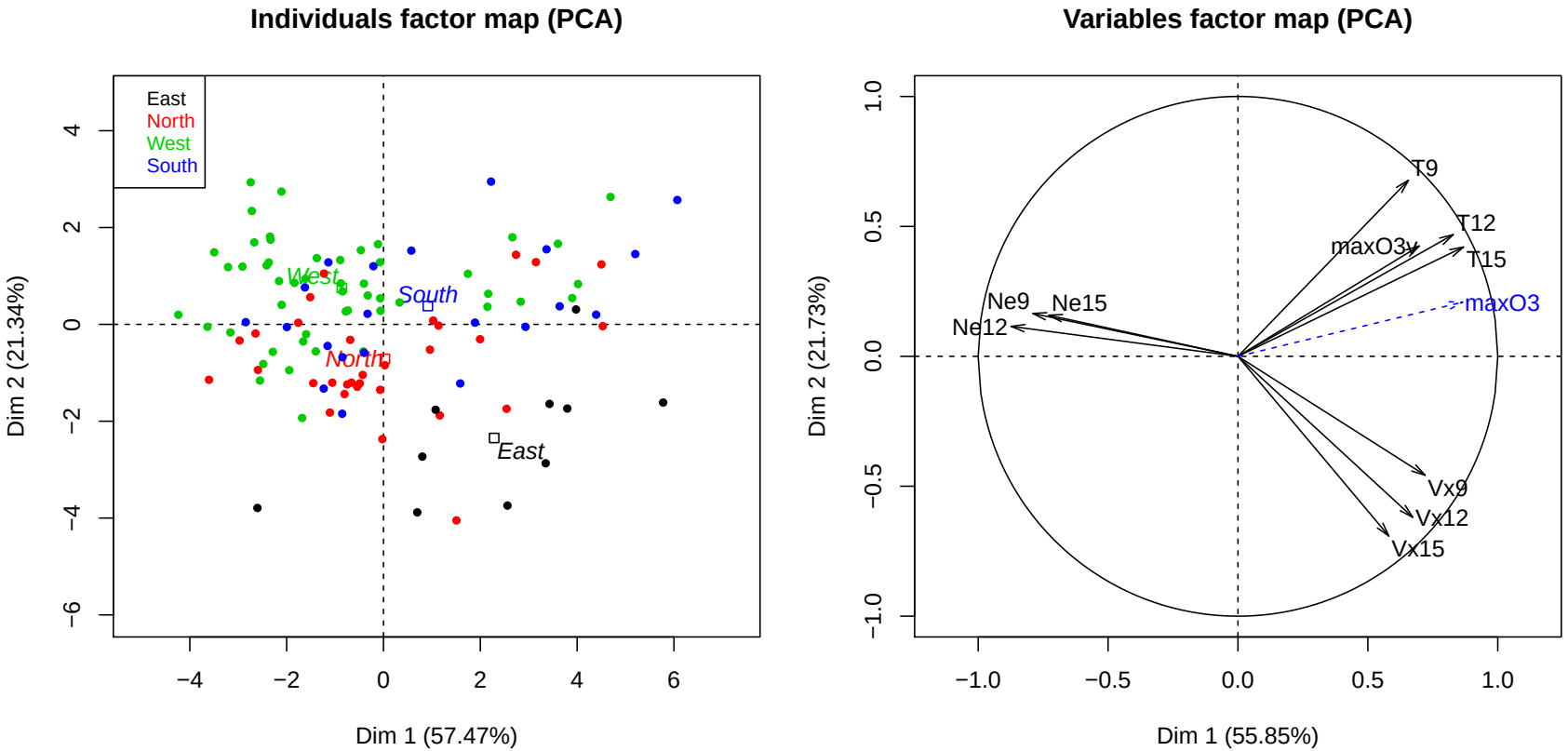
⇒ Individuals position (and variables) with other predictions



Regularized iterative PCA

⇒ reference configuration

PCA representation

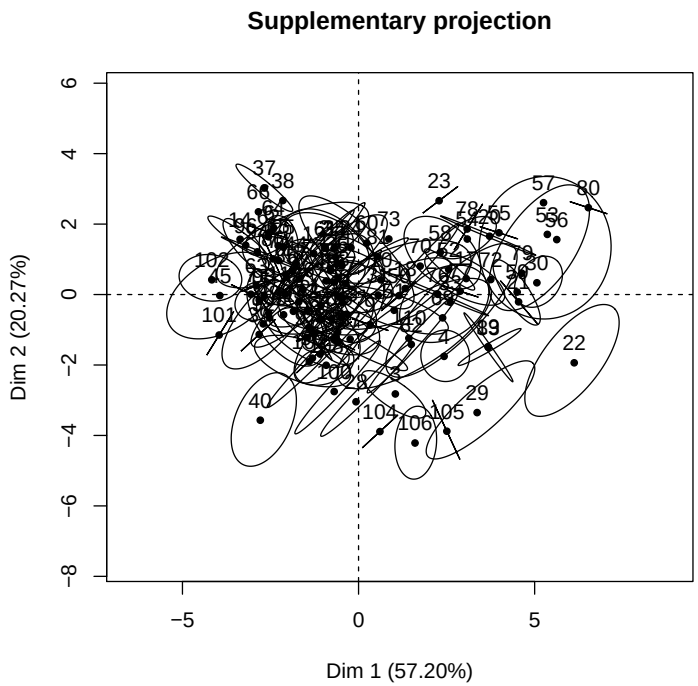


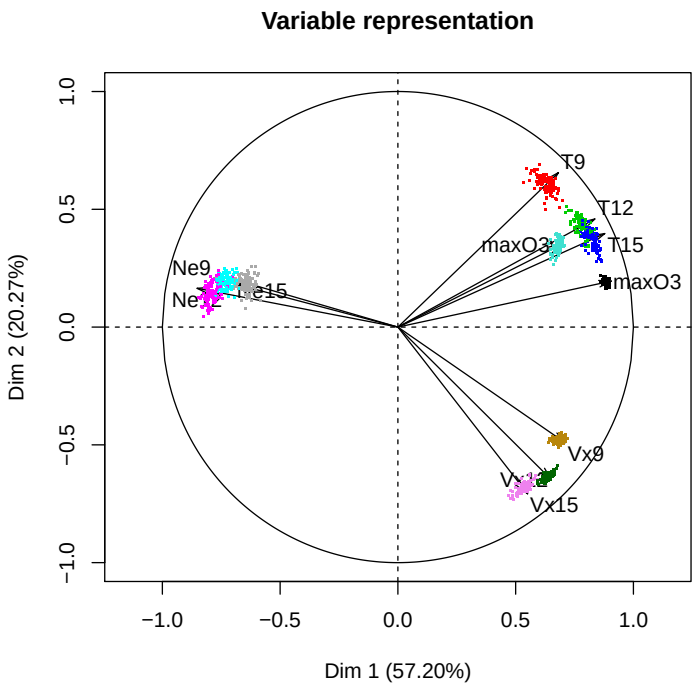

```
> imp <- cbind.data.frame(res.comp$completeObs, ozo[, 12])
> res.pca <- PCA(imp,quanti.sup = 1, quali.sup = 12)
> plot(res.pca, hab =12, lab = "quali"); plot(res.pca, choix = "var")
> res.pca$ind$coord #scores (principal components)
```

Multiple imputation in practice

⇒ Step 2: visualization

```
> res.MIPCA <- MIPCA(don, ncp = 2)
> plot(res.MIPCA, choice = "ind.supp"); plot(res.MIPCA, choice = "var")
```





Multiple imputation in practice

$\Rightarrow$ Step 3. Regression on each table and pool the results

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

$$T = \frac{1}{M}\sum_m \widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_m\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

```
> library(mice)
> res.mice <- mice(don, m = 100)
> imp.micerf <- mice(don, m = 100, defaultMethod = "rf")
> lm.mice.out <- with(res.mice, lm(maxO3 ~ T9+T12+T15+Ne9+...+Vx15+maxO3v))
> pool.mice <- pool(lm.mice.out)
> summary(pool.mice)
```

```
               est    se     t    df Pr(>|t|)  lo 95 hi 95 nmis  fmi lambda
(Intercept) 19.31 16.30  1.18 50.48     0.24 -13.43 52.05   NA 0.46   0.44
T9          -0.88  2.25 -0.39 26.43     0.70  -5.50  3.75   37 0.71   0.69
T12          3.29  2.38  1.38 27.54     0.18  -1.59  8.18   33 0.70   0.68
....
Vx15         0.23  1.33  0.17 39.00     0.87  -2.47  2.93   21 0.57   0.55
maxO3v       0.36  0.10  3.65 46.03     0.00   0.16  0.56   12 0.50   0.48
```

Outline

Estimation without imputation

Modify the estimation process to deal with missing values.

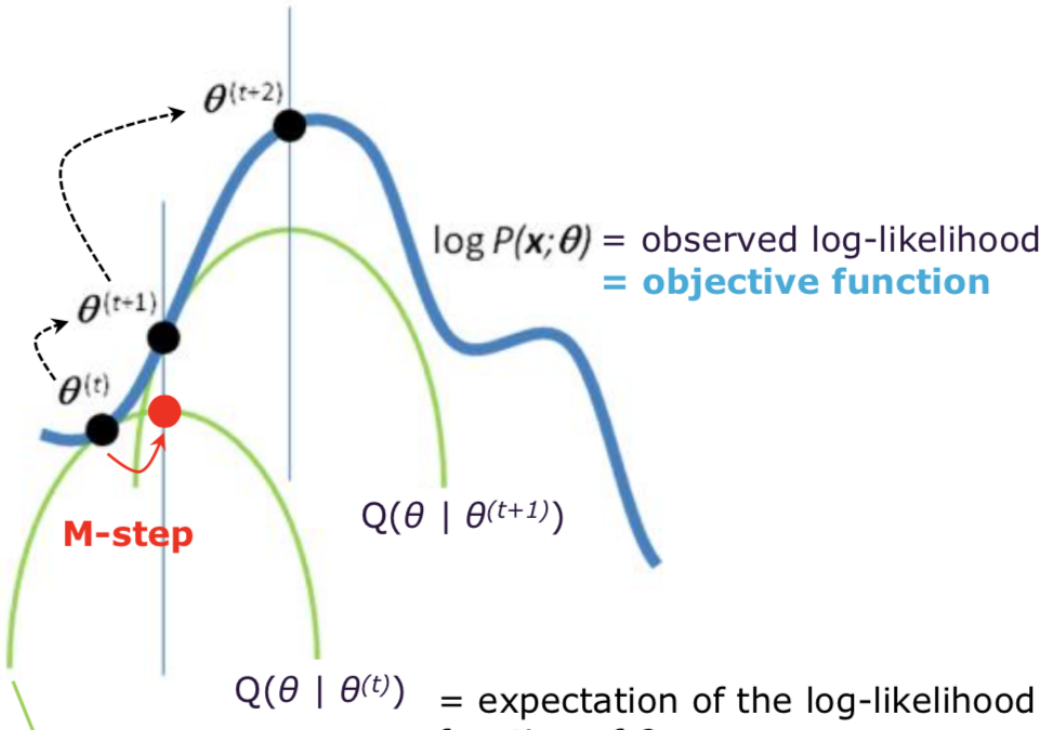
Maximum observed likelihood: $\operatorname{argmax} \ell(\theta; x_{\text{obs}}) = \int \ell(\theta; x) dx_{\text{mis}}$.

Expectation-Maximization algorithm:

- **E-step:** Evaluate the quantity

$$Q_k(\theta) = \mathbb{E}[\ell(\theta; x)|x_{\text{obs}}; \theta_{k-1}] = \int \ell(\theta; x) \mathrm{p}(x_{\text{mis}}|x_{\text{obs}}; \theta_{k-1}) dx_{\text{mis}}.$$

- **M-step:** $\theta_k = \operatorname{argmax}_\theta Q_k(\theta)$.



EM algorithm with missing data

Example: Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$, point estimates with EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

$\Rightarrow$ Natural model selection procedure!
$\Rightarrow$ One specific algorithm for each statistical method.
$\Rightarrow$ Not many implementations even for simple models.

package `misaem`: Logistic regression with missing covariates

To conclude

Take home message:

- "***The idea of imputation is both seductive and dangerous***. *It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.*" (Dempster and Rubin, 1983)
- Single imputation aims to complete a dataset as best as possible (prediction)
- Multiple imputation aims to perform other statistical methods after and to estimate parameters and their variability taking into account the missing values uncertainty
- EM algorithm estimates parameters without imputation but requires computational cost.

Some references

Schafer (1997),

Joseph L. Schafer
Analysis of incomplete data

Little & Rubin (1987, 2002)

Roderick Little

Donald Rubin
Statistical analysis with missing values

Suggested reading: chap 25 of Gelman & Hill (2006)

Andrew Gelman

Jennifer L. Hill
Data Analysis Using Regression and Multilevel/Hierarchical Models

R-miss-tastic: https://rmisstastic.netlify.com/
A resource website on missing values - Methods for managing missing data