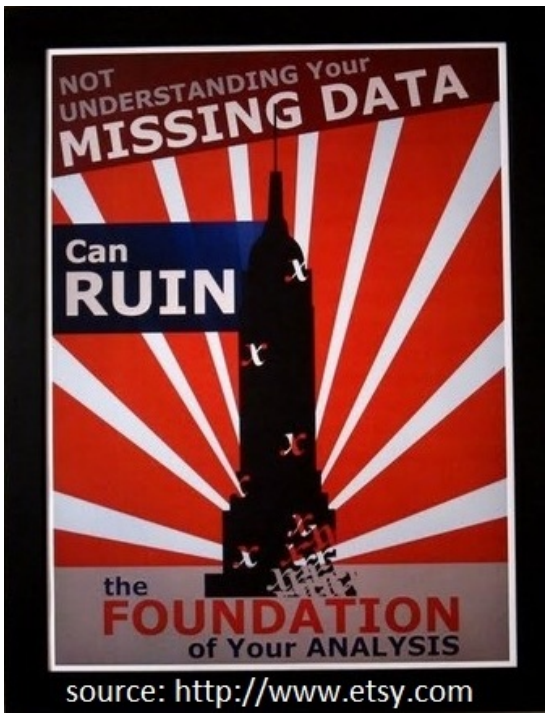
Dealing with missing values

Julie Josse



Meetup Rladies, Paris, 20 April 2017

Research activities

- Dimensionality reduction methods to visualize complex data (PCA based): multi-sources data, textual data, arrays
- Missing values - matrix completion
- Low rank estimation, selection of regularization parameters
- Fields of application: bio-sciences (agronomy, sensory analysis), health data (hospital APHP)
- R community: book R for Statistics, R foundation, R taskforce for women, R packages and JSS papers:

  FactoMineR explore continuous, categorical, multiple contingency tables (correspondence analysis), combine clustering and PC, ..
  MissMDA for single and multiple imputation, PCA with missing
  denoiseR to denoise data

Missing values

are everywhere: unanswered questions in a survey, lost data, damaged plants, machines that fail...

The best thing to do with missing values is not to have any" Gertrude Mary Cox.

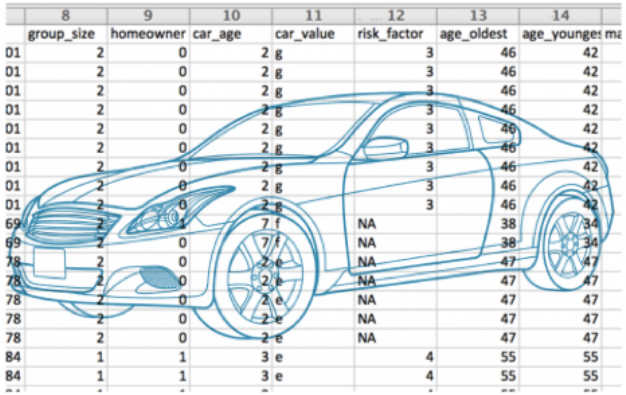
$\Rightarrow$ Still an issue in the "big data" area

Data integration: data from different sources

Kaggle

You can use any type of data to create a Kaggle competition, for example:

Numerical
Text
Media
Multiple formats



	8	9	10	11	12	13	14
	group_size	homeowner	car_age	car_value	risk_factor	age_oldest	age_younges ma
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
01	2	0	2	g	3	46	42
69	2	0	7	f	NA	38	34
69	2	0	7	f	NA	38	34
78	2	0	2	e	NA	47	47
78	2	0	2	e	NA	47	47
78	2	0	2	e	NA	47	47
78	2	0	2	e	NA	47	47
78	2	0	2	e	NA	47	47
84	1	1	3	e	4	55	55
84	1	1	3	e	4	55	55

Allstate ran a competition to predict a customer's purchase based on a limited amount of shopping history data.

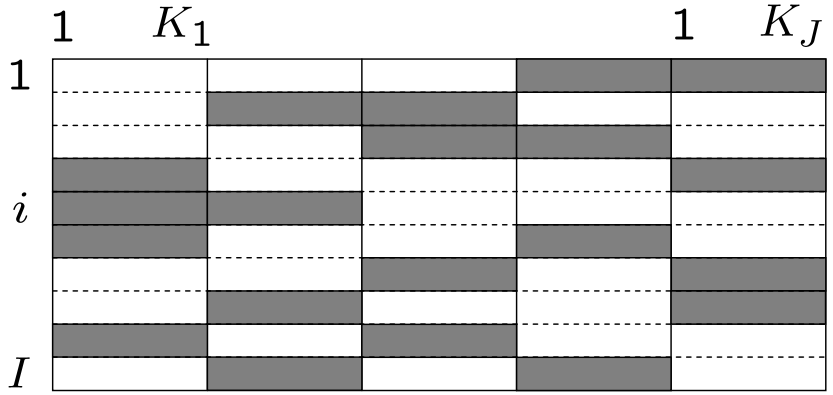
Jobs • 1,429 teams

Airbnb New User Bookings

Wed 25 Nov 2015 — Thu 11 Feb 2016 (40 hours to go)

Predict in which country a new user will make his first booking:
age: 42.4 %
date first booking: 6.7 %
first affiliate tracked: 2.2 %
gender: 46 %

Multi-blocks data set



L'OREAL: 100 000 women in different countries - 300 questions

- Self-assessment questionnaire: life style, skin and hair characteristics, care and consumer habits
- Clinical assessments by a dermatologist: facial skin complexion, wrinkles, scalp dryness, greasiness
- Hair assessments by a hair dresser: abundance, volume, breakage, curliness
- Skin and Hair photographs and measurements: sebum quantity, etc.

Ozone data set

	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/

```
ozo <- read.table("http://juliejosse.com/wp-content/uploads/2016
/06/ozoneNA.csv",header=TRUE,sep=",",row.names=1)
```

Recommended methods

$\Rightarrow$ Modify the method, the estimation process to deal with missing values. Maximum likelihood: EM algorithm to obtain point estimates (+ other algorithms for their variability)

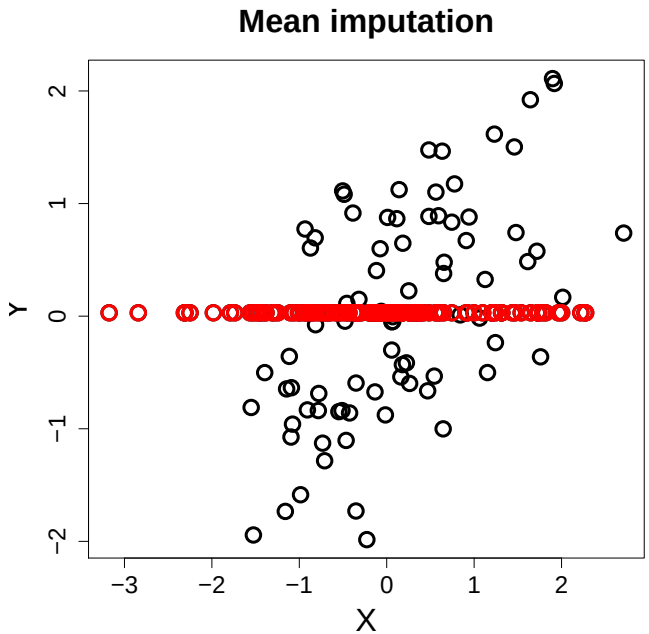
One specific algorithm for each statistical method...

$\Rightarrow$ Imputation (multiple) to get a completed data set on which you can perform any statistical method

Dealing with missing values

$\Rightarrow$ A simple way: delete rows. Be careful to delete subpopulations!

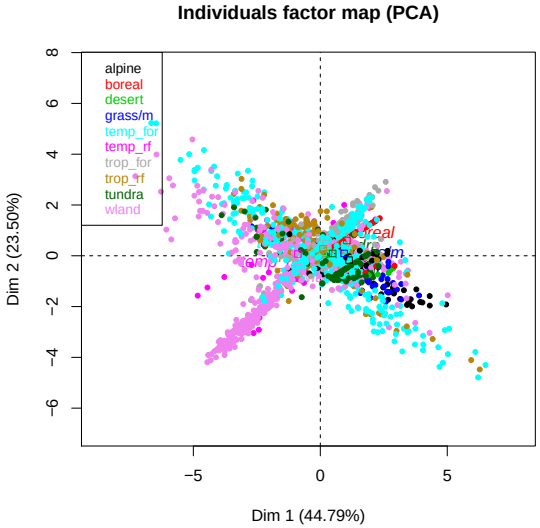
$\Rightarrow$ Imputation to get a completed data set

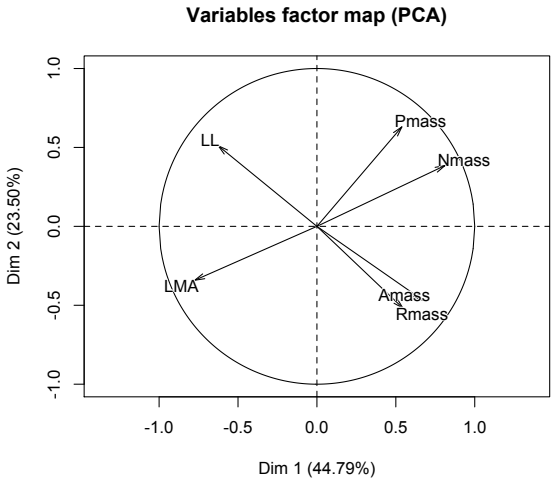


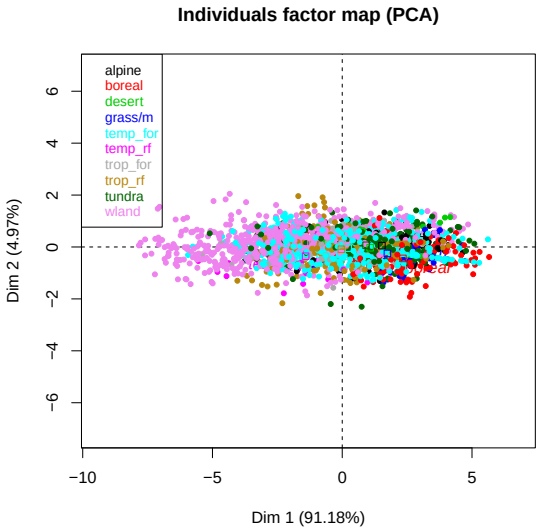
$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

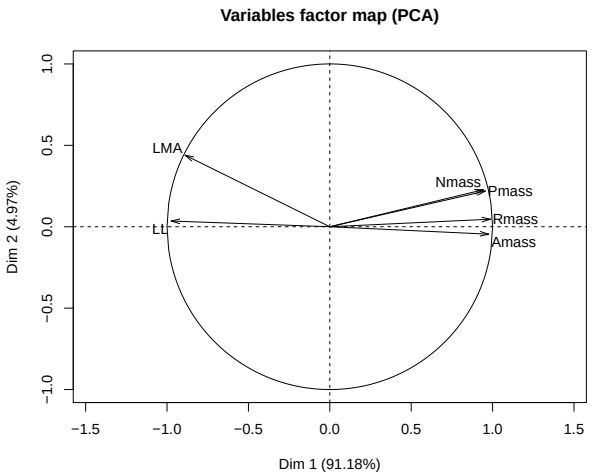
$\hat{\mu}_y = 0.01$
$\hat{\sigma}_y = 0.5$
$\hat{\rho} = 0.30$

Dealing with missing values







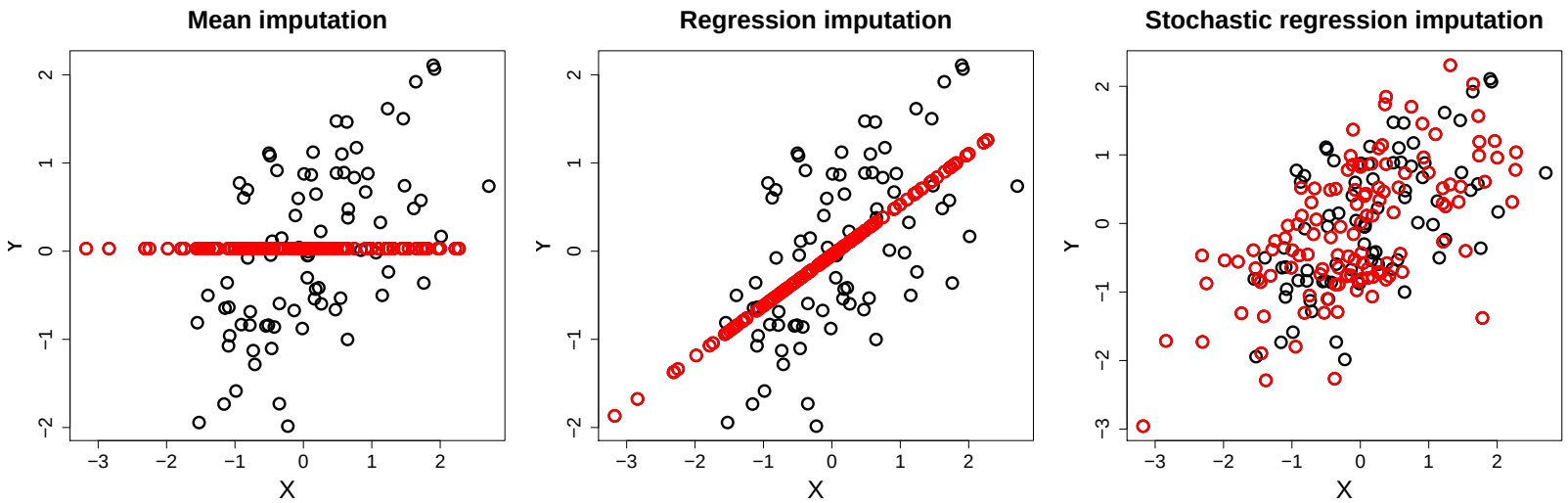


Wright IJ, et al. (2004). The worldwide leaf economics spectrum. *Nature*,
69 000 species - LMA (leaf mass per area), LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), Rmass (dark respiration rate)

Dealing with missing values

$\Rightarrow$ A simple way: delete rows. Be careful to delete subpopulations!

$\Rightarrow$ Imputation to get a completed data set

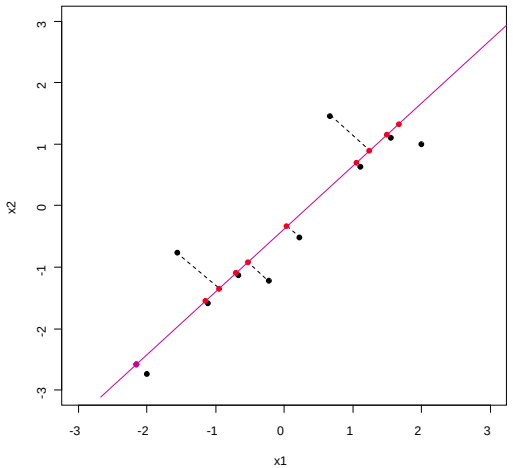


$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

PCA reconstruction

X

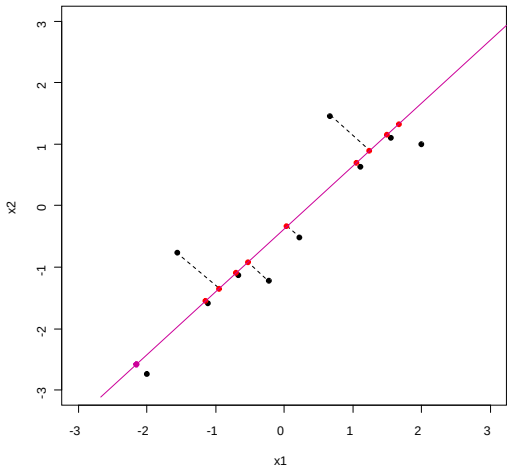
-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00



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



-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

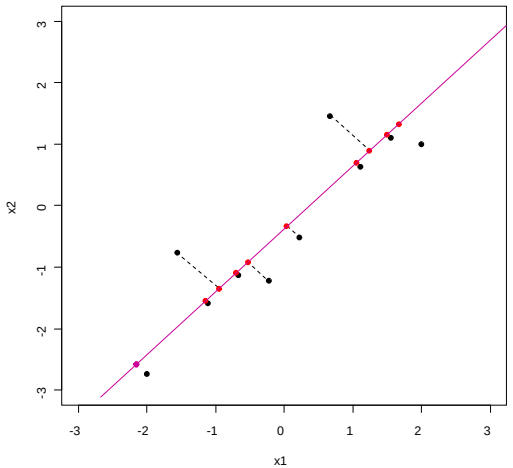
$\hat{X}$

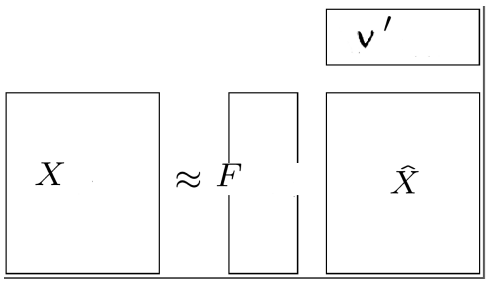
$\Rightarrow$ Approx of X with a low rank matrix $S < p$: $\|X_{n\times p} - \hat{X}_{n\times p}\|^2$

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





$\hat{X}$

-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

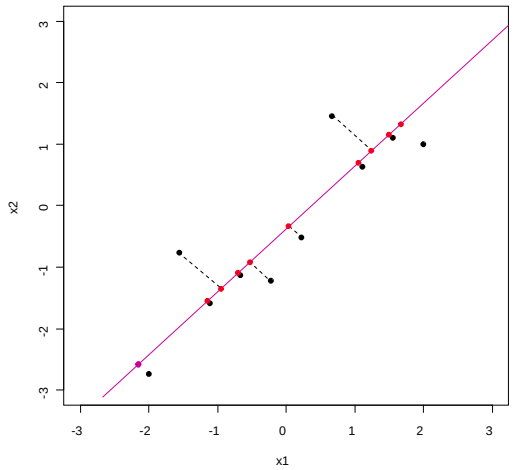
$\Rightarrow$ Approx of X with a low rank matrix $S < p$: $\|X_{n\times p} - \hat{X}_{n\times p}\|^2$

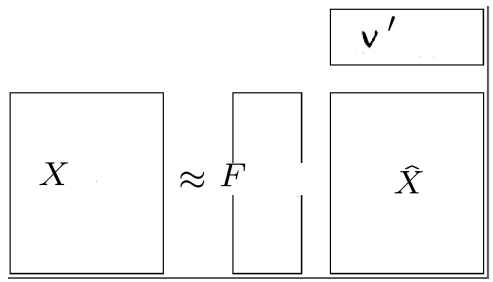
$$\text{SVD: } \hat{X}^{\text{PCA}} = U_{n\times s}\Lambda^{\frac{1}{2}}_{s\times s}V'_{p\times s} = F_{n\times s}V'_{p\times s}$$

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





$\hat{X}$

-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

$\Rightarrow$ Approx of X with a low rank matrix $S < p$: $\|X_{n\times p} - \hat{X}_{n\times p}\|^2$

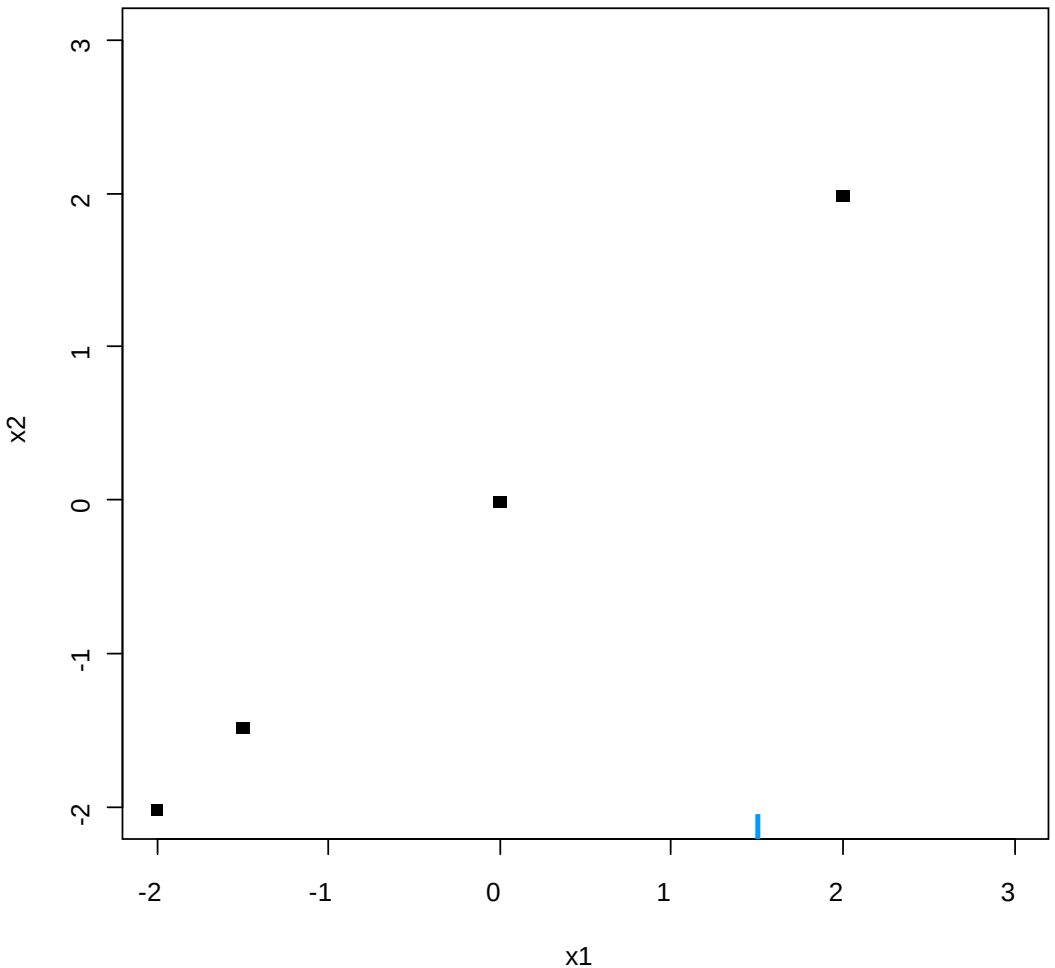
$$\text{SVD: } \hat{X}^{\text{PCA}} = U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S} = F_{n\times S}V'_{p\times S}$$

$\Rightarrow$ Missing: weighted least squares $w_{ij} = 0$ if x_{ij} is miss/ $w_{ij} = 1$

$$\|W_{n\times p} \odot (X_{n\times p} - U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S})\|^2$$

Iterative PCA

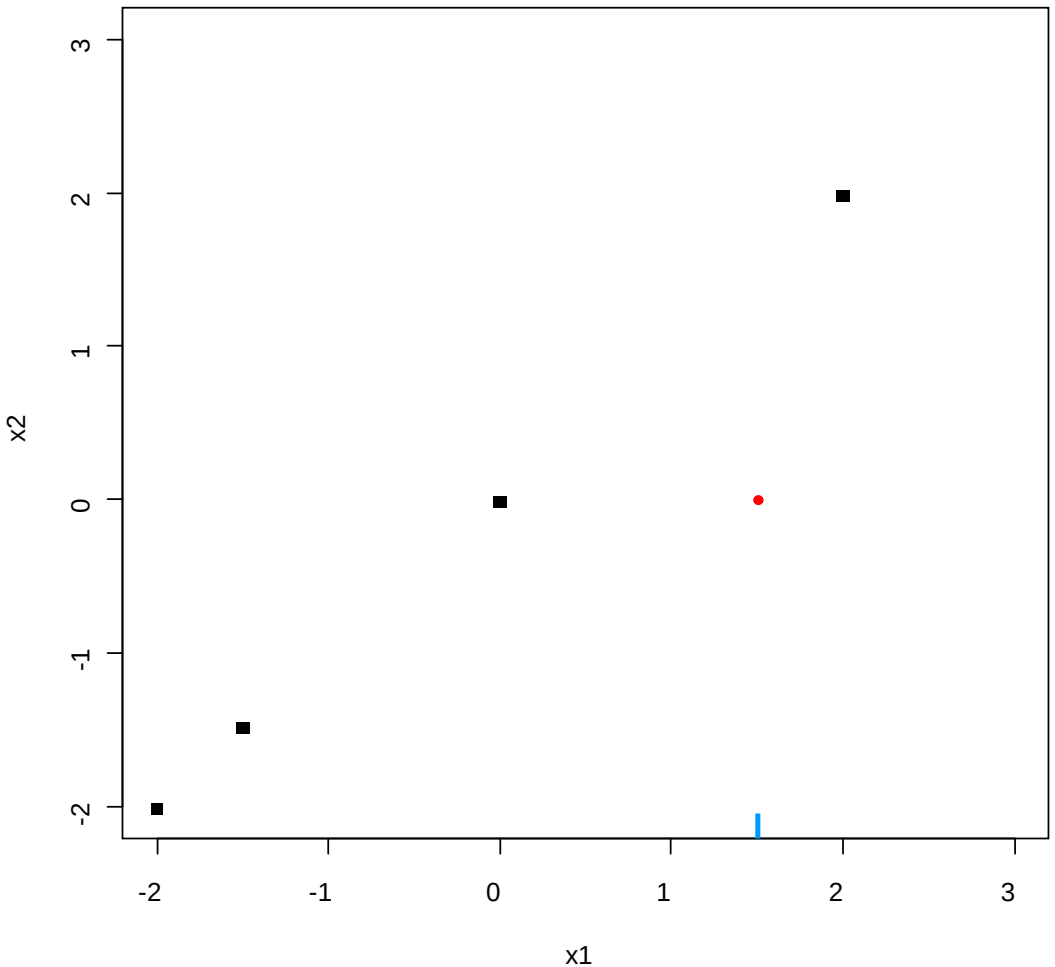
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



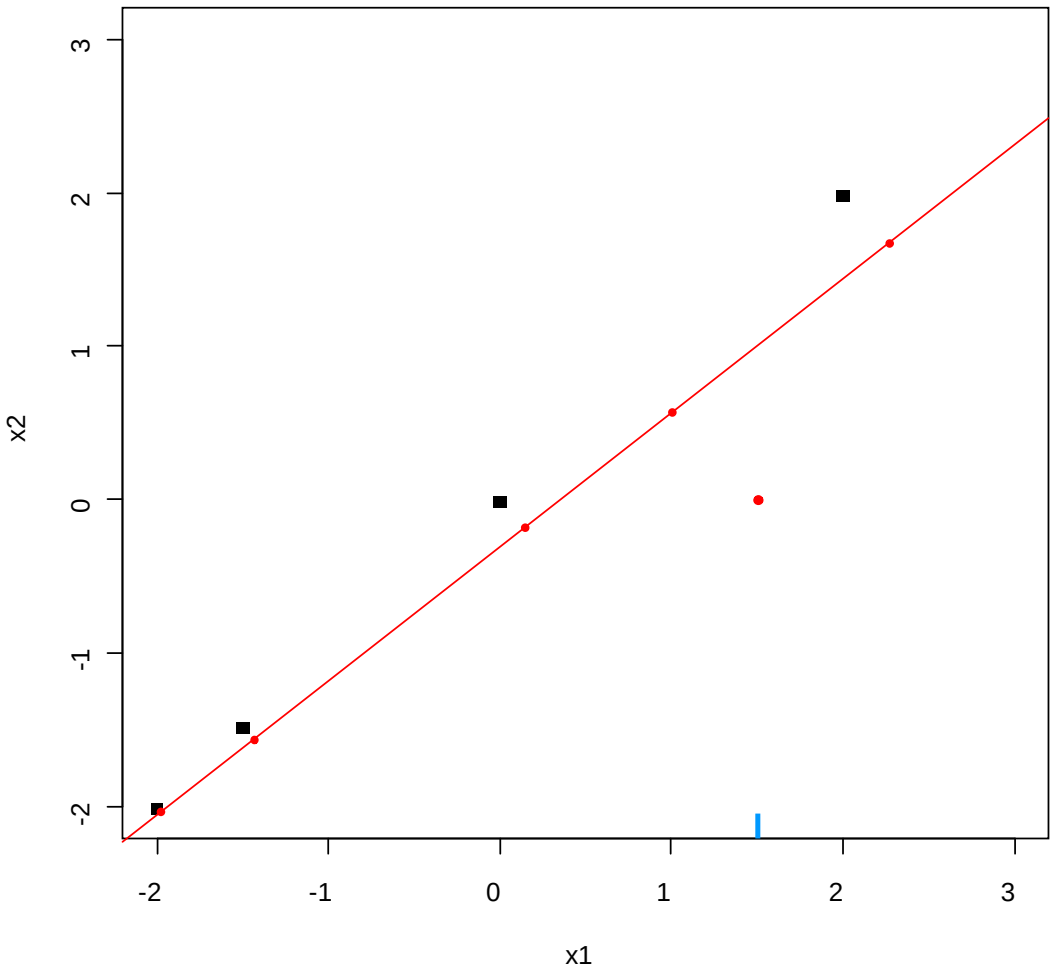
Initialization $\ell = 0$: X^0 (mean imputation)

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



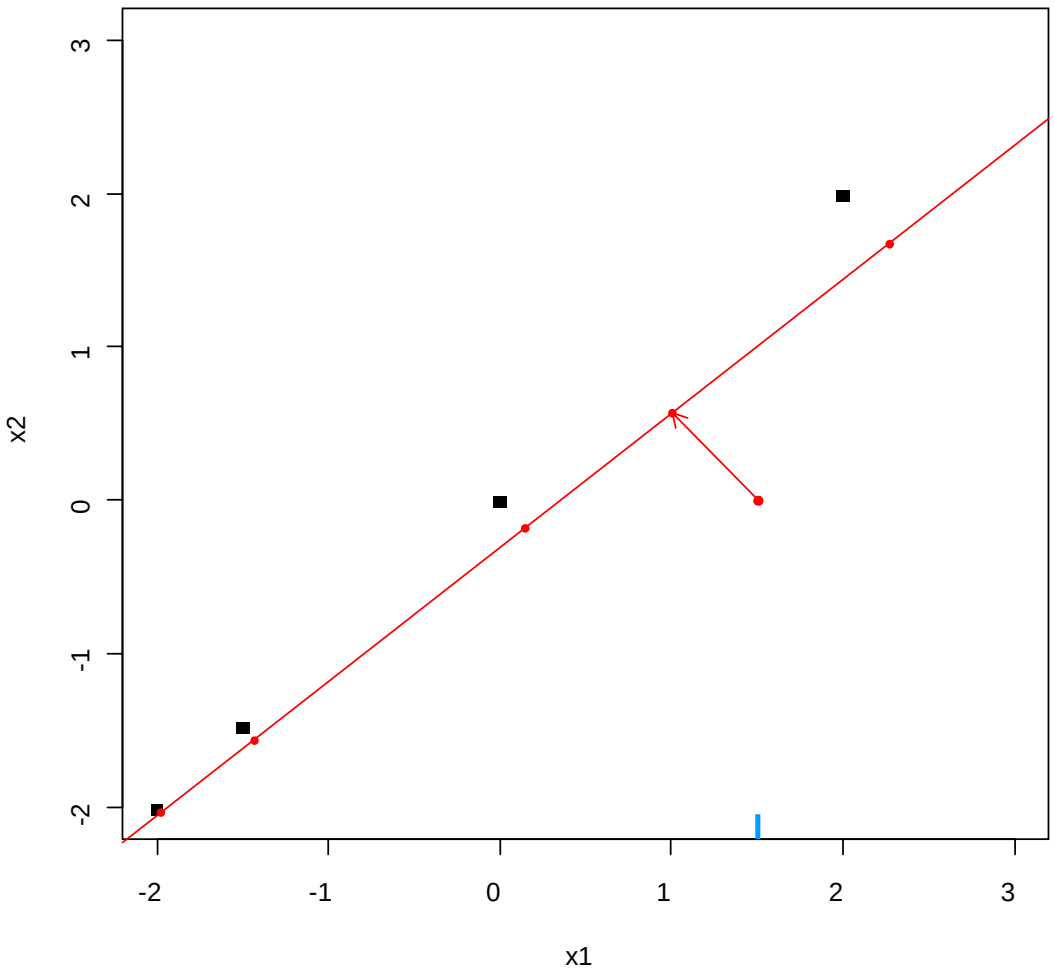
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

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

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



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

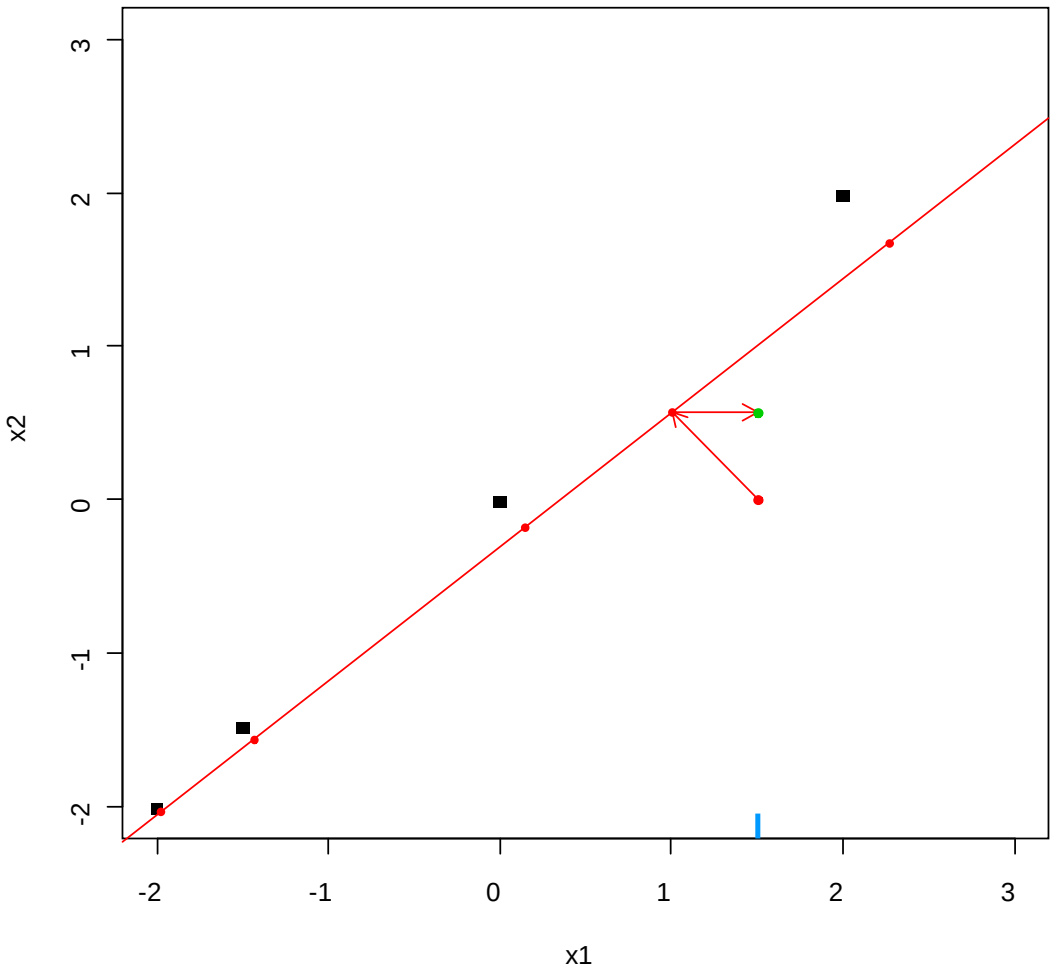
Iterative PCA

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



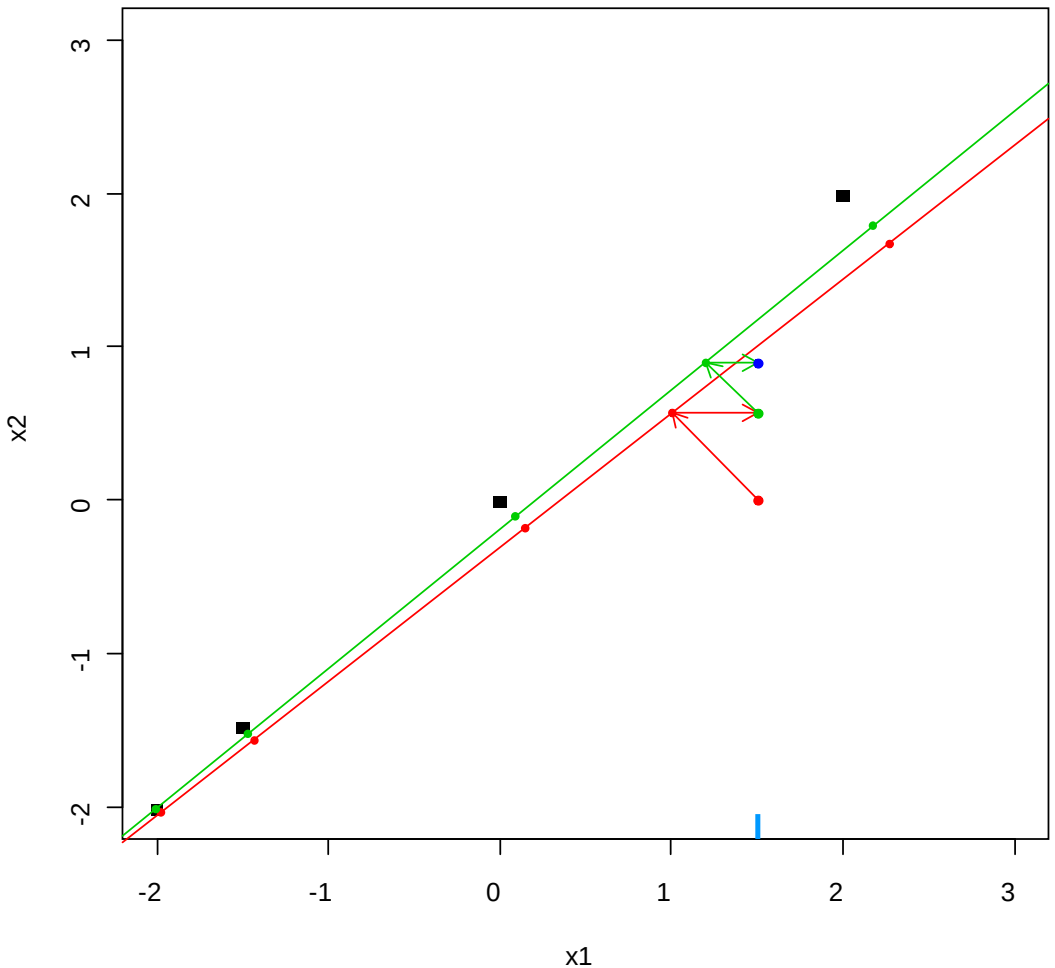
The new imputed dataset is $\hat{X}^{\ell} = W * X + (\mathbf{1} - W) * \hat{X}^{\ell}$

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



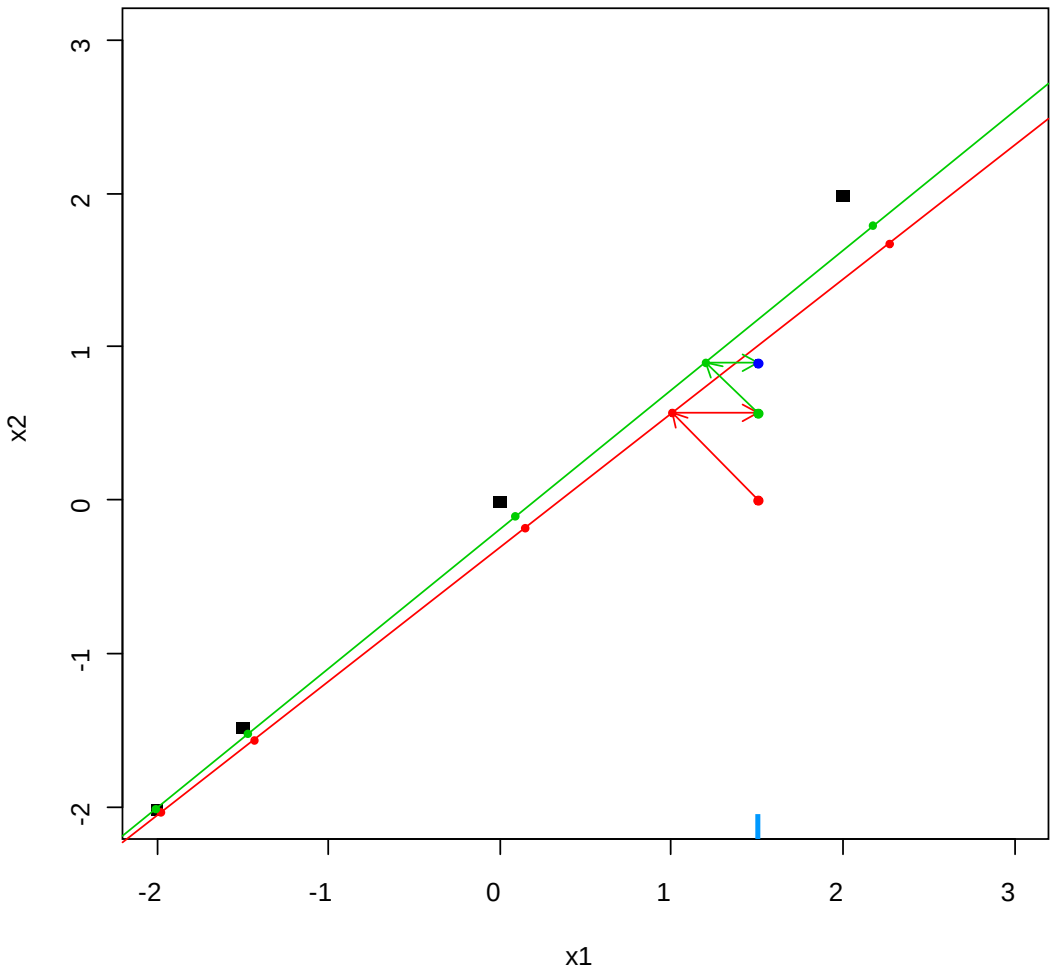
Iterative PCA

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

$\hat{x}1$	$\hat{x}2$
-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



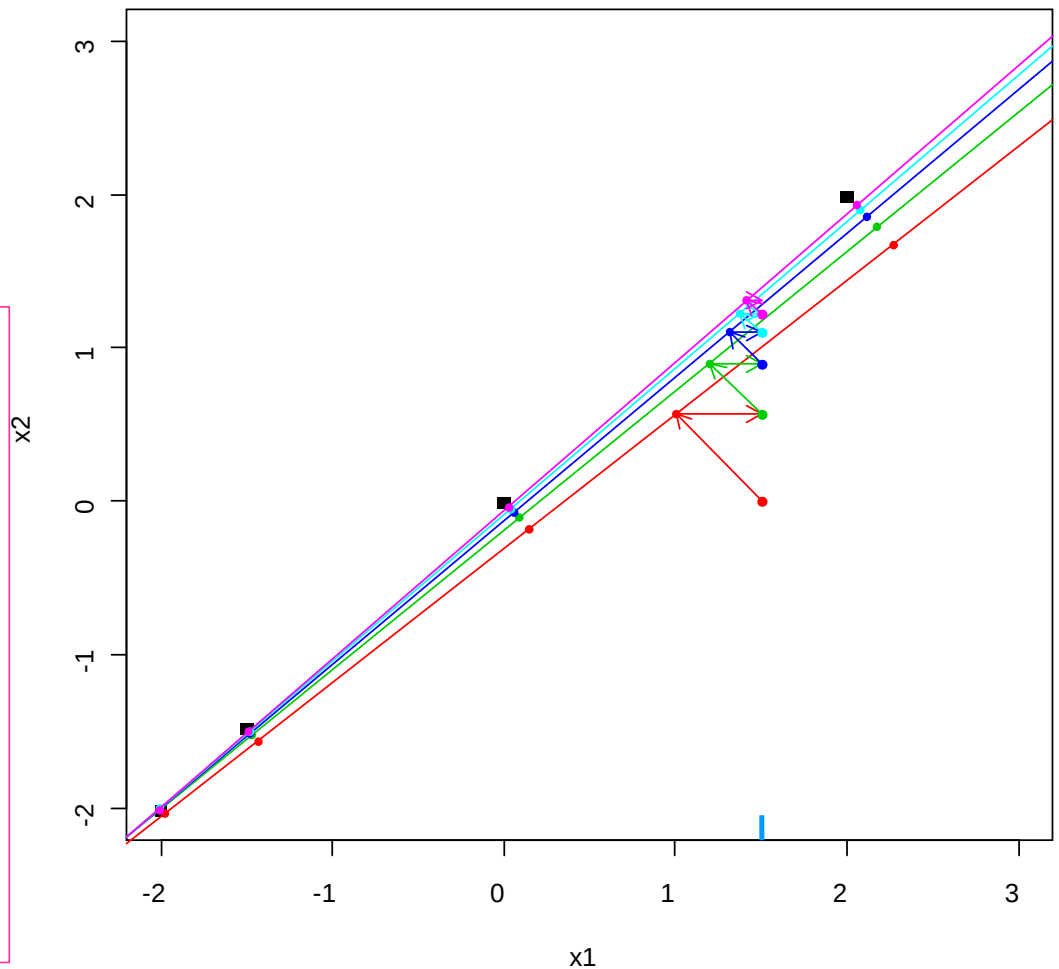
Iterative PCA

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

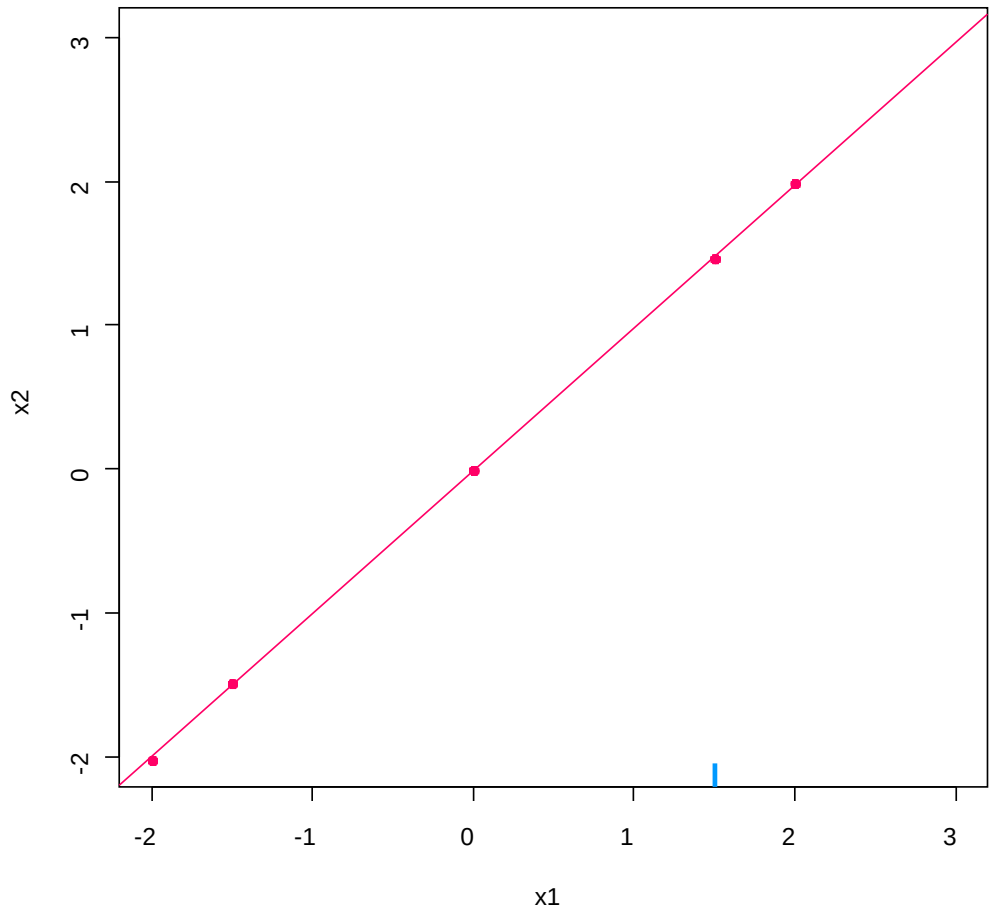
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



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 $\hat{X}^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$

Iterative PCA

- initialization $\ell = 0$: X^0 (mean imputation)
- step ℓ:
  - PCA on completed data $\hat{x}_{ij}^{\ell} = \sum_{s=1}^{S} \sqrt{\lambda_s}^{\ell} U_{is}^{\ell} V_{js}^{\ell}$
  - missing values imputed with $\hat{X}$
    new imputed data is $X^{\ell} = W \odot X + (1 - W) \odot \hat{X}^{\ell}$
- steps of estimation and imputation are repeated

$\Rightarrow$ Very good quality of imputation. (Netflix: 99% missing).
$\Rightarrow$ Model: Data = structure of rank S + noise makes sense.

Iterative PCA

- initialization $\ell = 0$: X^0 (mean imputation)
- step ℓ:
  - PCA on completed data $\hat{x}_{ij}^{\ell} = \sum_{s=1}^{p} (\sqrt{\lambda_s} - \lambda)_+^{\ell} U_{is}^{\ell} V_{js}^{\ell}$
  - missing values imputed with $\hat{X}$
    new imputed data is $X^{\ell} = W \odot X + (1 - W) \odot \hat{X}^{\ell}$
- steps of estimation and imputation are repeated

$$\text{argmin}_{\mu} \left\{ \| W \odot (X - \mu) \|_2^2 + \lambda \| \mu \|_* \right\}$$

$\Rightarrow$ Different transformations - ways of selecting λ

$\Rightarrow$ Very good quality of imputation. (Netflix: 99% missing).
$\Rightarrow$ Model: Data = structure of rank S + noise makes sense.

Iterative Random Forests imputation

1. Initial imputation: mean imputation
2. Fit a RF X_1^{obs} on X_{-1} and then predict X_1^{miss}
   Fit a RF X_2^{obs} on X_{-2} and then predict X_2^{miss}
   ...
   cycling through variables
3. Repeat until convergence

$\Rightarrow$ Conditional modeling based on RF

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

$\Rightarrow$ Good for complex relationships between variables

Random Forests versus PCA

	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

$\Rightarrow$ MAR missing values

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

Ozone data

	O3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	O3v
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

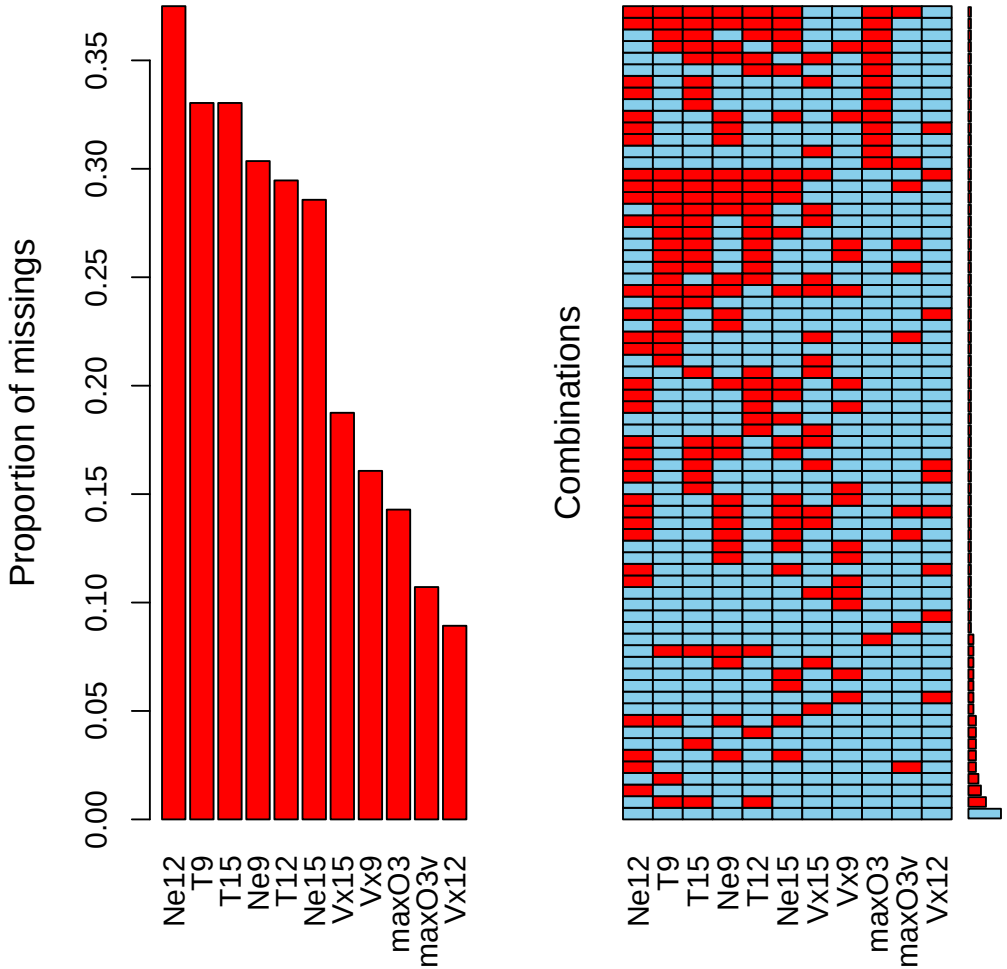
Count missing values

```
> 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
```

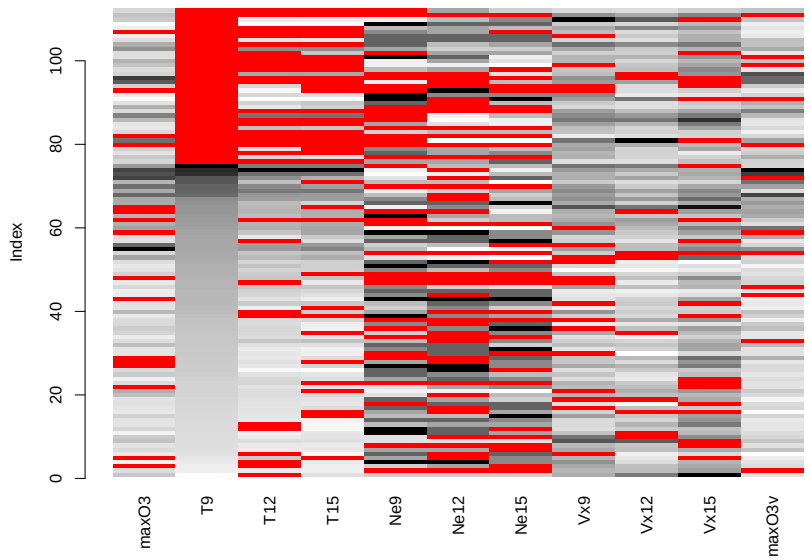
Combinations	Count	Percent
0:0:0:0:0:0:0:0:0:0:0	13	11.6071429
0:1:1:1:0:0:0:0:0:0:0	7	6.2500000
0:0:0:0:0:1:0:0:0:0:0	5	4.4642857
0:1:0:0:0:0:0:0:0:0:0	4	3.5714286
0:1:0:0:1:1:1:0:0:0:0	3	2.6785714
0:0:1:0:0:0:0:0:0:0:0	3	2.6785714
0:0:0:1:0:0:0:0:0:0:0	3	2.6785714
0:0:0:0:1:1:1:0:0:0:0	3	2.6785714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

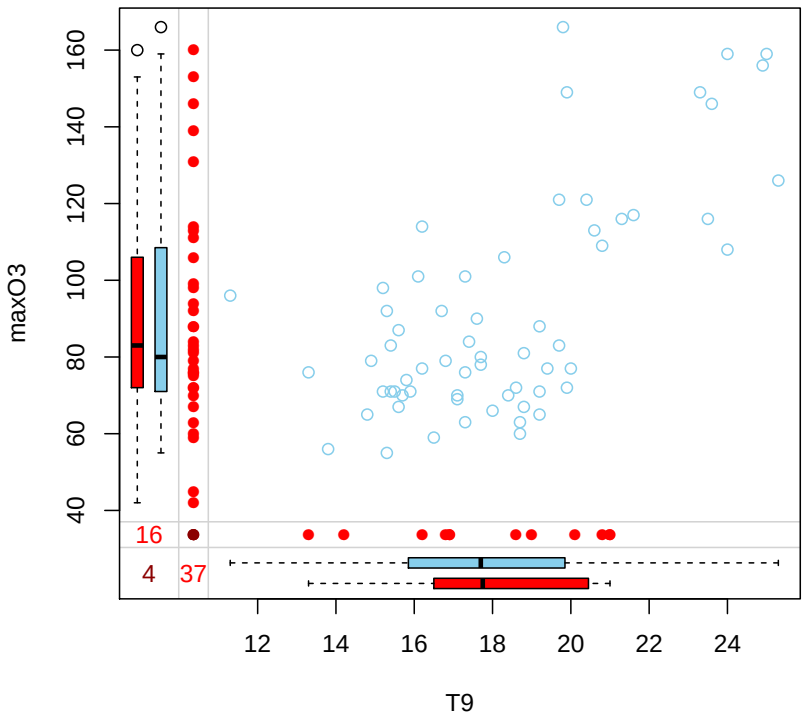
Pattern visualization




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

Visualization






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

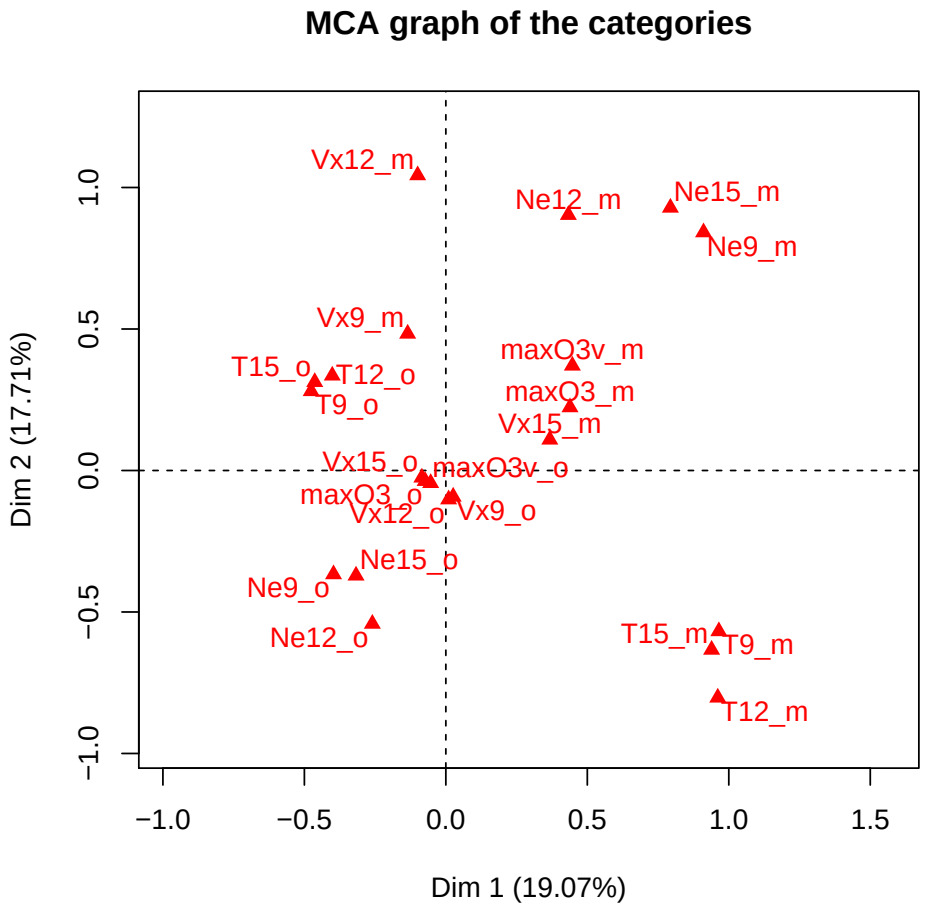
Visualization with Multiple Correspondence Analysis

$\Rightarrow$ 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

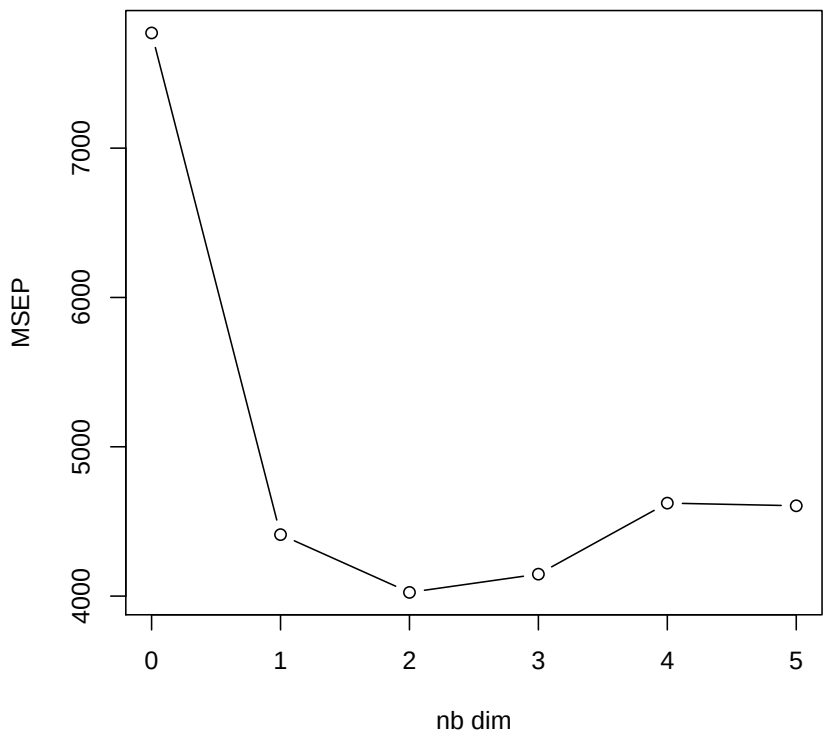



```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA, invis = "ind", title = "MCA graph of the categories")
```

Imputation with PCA in practice

⇒ Step 1: Estimation of the number of dimensions
(Cross Validation, Bro, 2008; GCV, Josse & Husson, 2011)

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

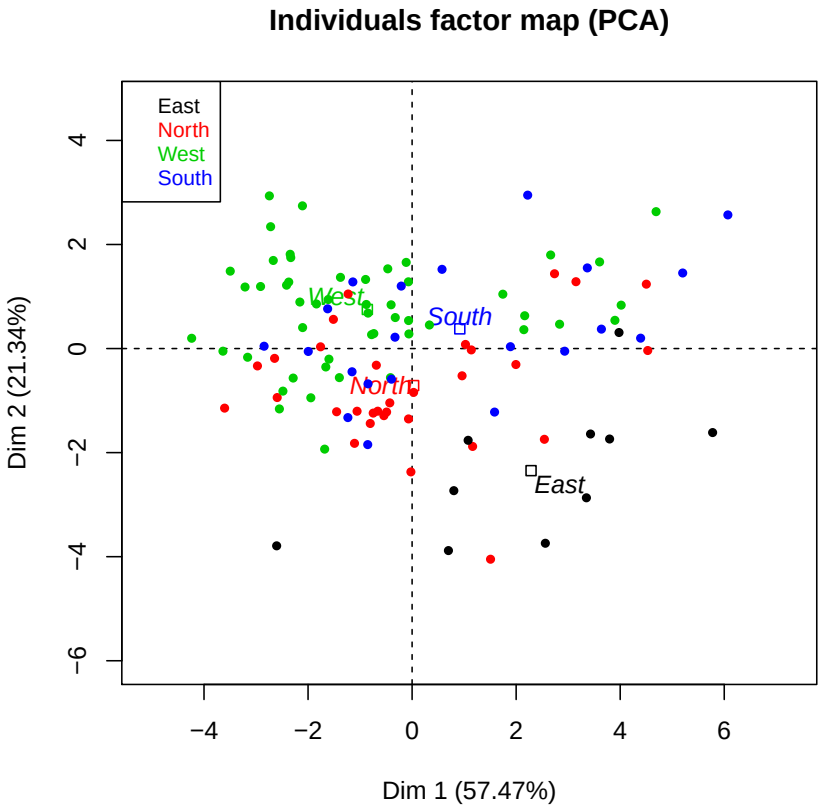


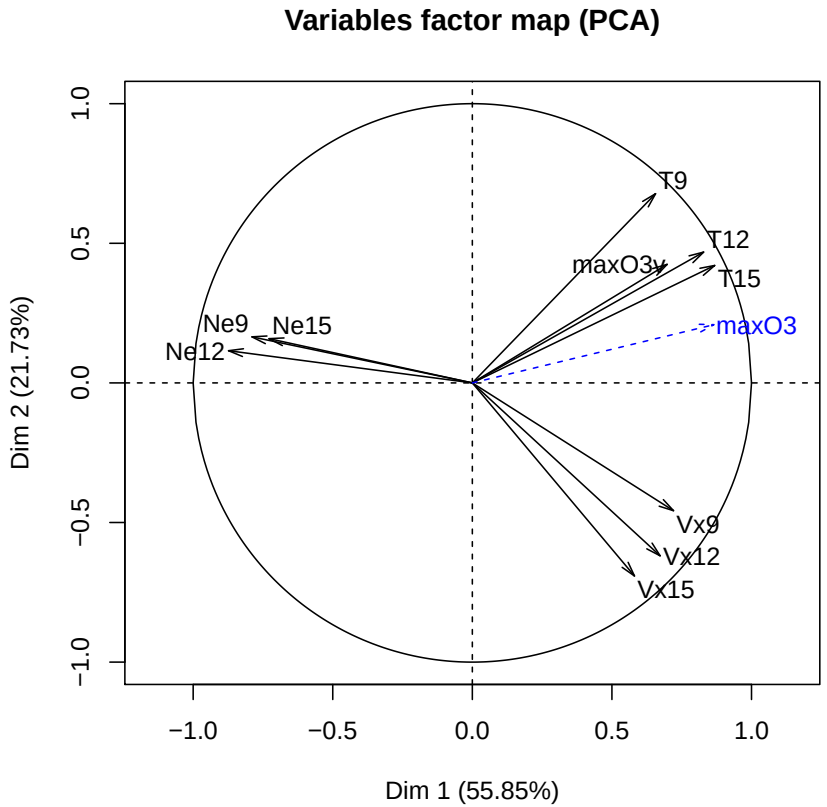
Imputation with PCA in practice

⇒ 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
```

Cherry on the cake: PCA on incomplete data!

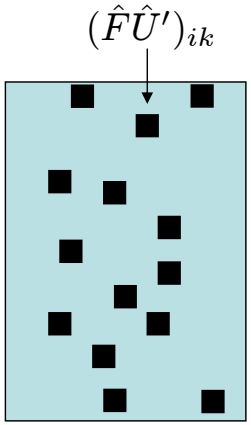





```
> imp <- cbind.data.frame(res.comp$completeObs, ozo[, 12])
> res.pca <- PCA(imp, quanti.sup = 1, quali.sup = 12)
> plot(res.pca, hab = 12, lab = "quali"); plot(res.pca, choix = "var")
> res.pca$ind$coord #scores (principal components)
```

Multiple imputation

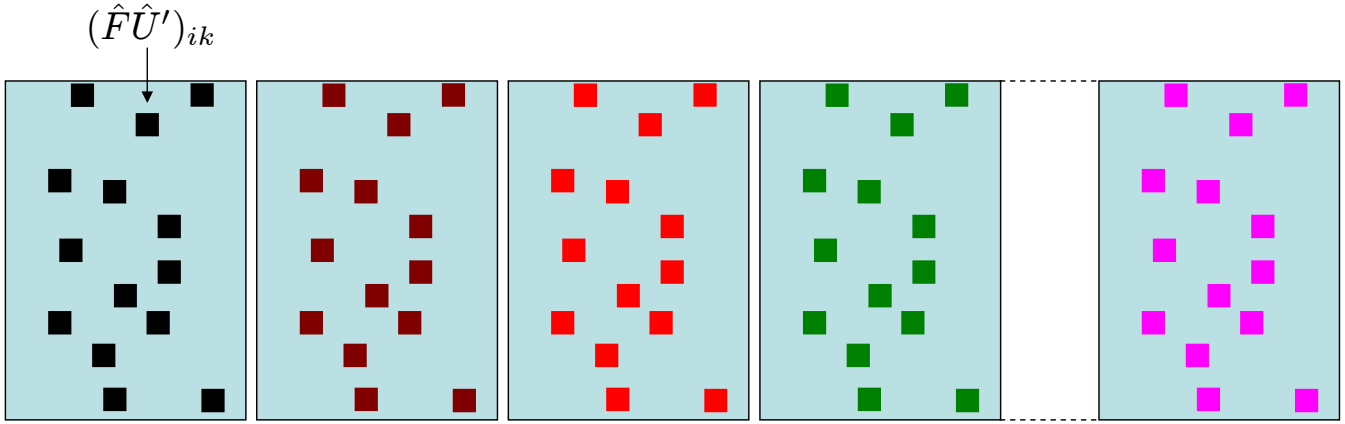
$\Rightarrow$ iterative PCA: single imputation



$\Rightarrow$ A unique value can not reflect the uncertainty of prediction

Multiple imputation

$\Rightarrow$ iterative PCA: single imputation



$\Rightarrow$ A unique value can not reflect the uncertainty of prediction

$\Rightarrow$ Multiple imputation (Rubin, 1987): different plausibles values for each missing entry (bootstrap/ bayesian approaches)

Ex: one cell: predict 18.5. Empirical interval [14.8 ; 22.18]

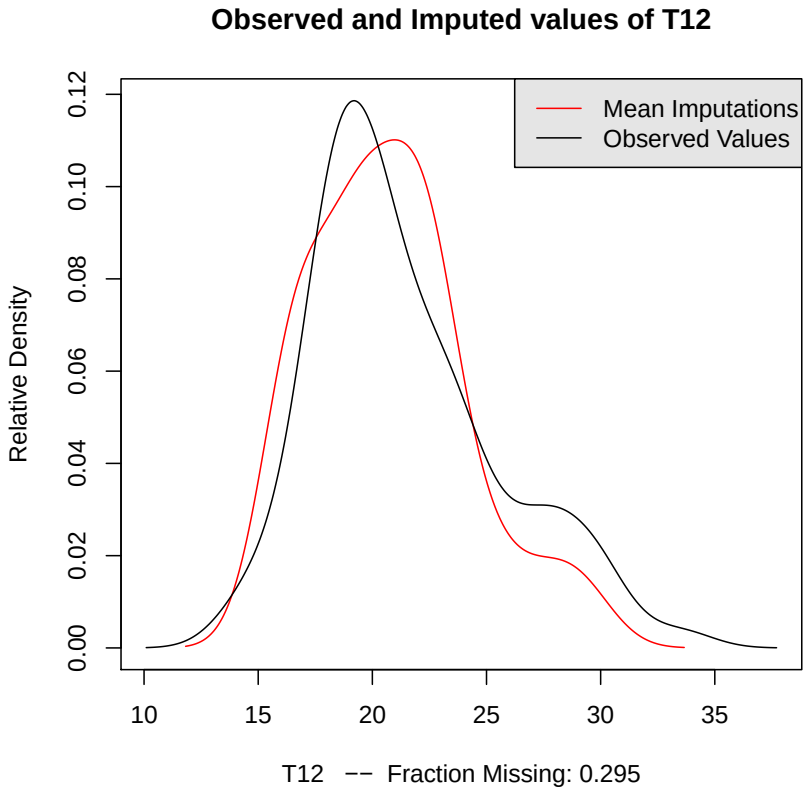
Multiple imputation in practice

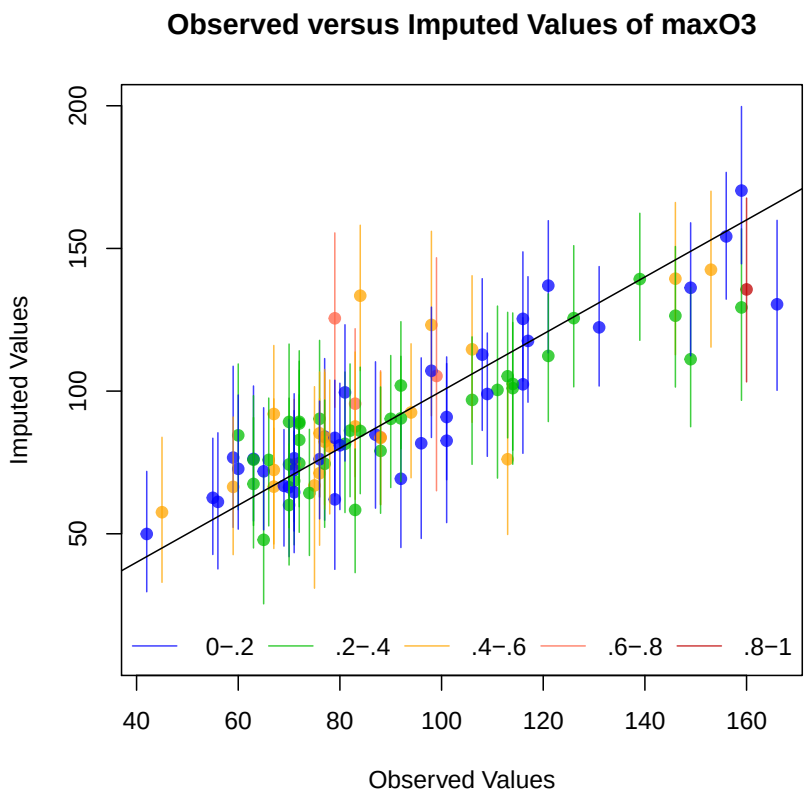
$\Rightarrow$ Step 1: Generate M imputed data sets

```
> 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



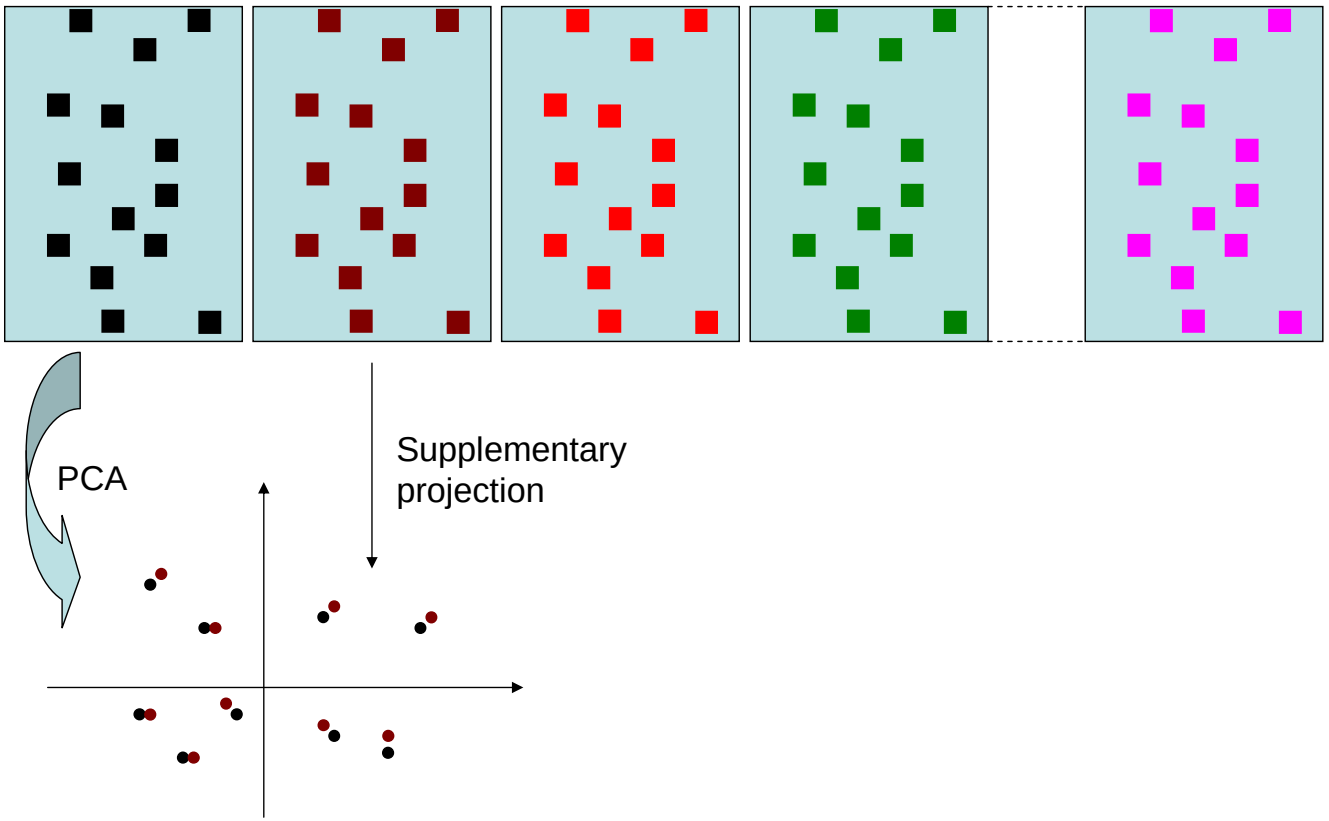



```
# library(Amelia)
> res.amelia <- amelia(don, m = 100)
> compare.density(res.amelia, var = "T12")
> overimpute(res.amelia, var = "maxO3")

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

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



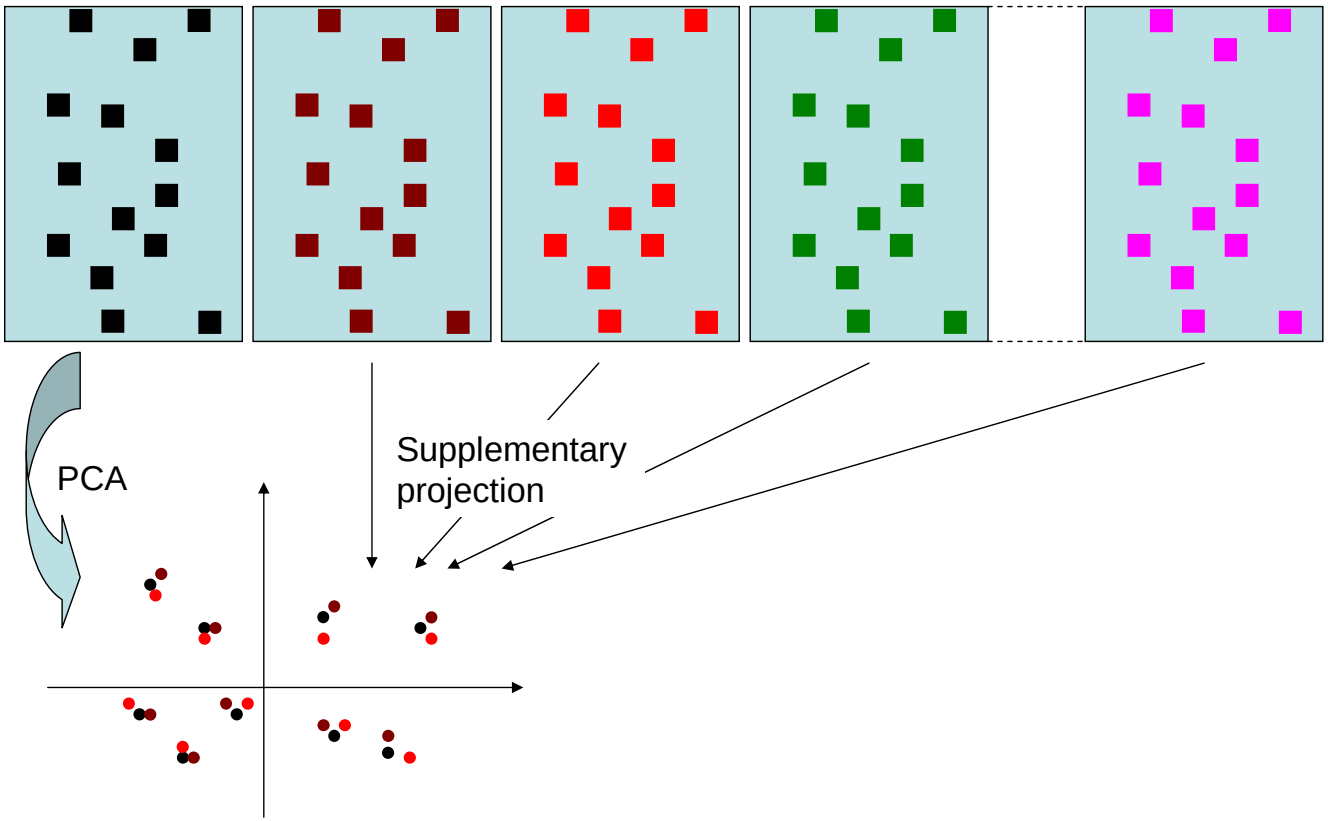
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



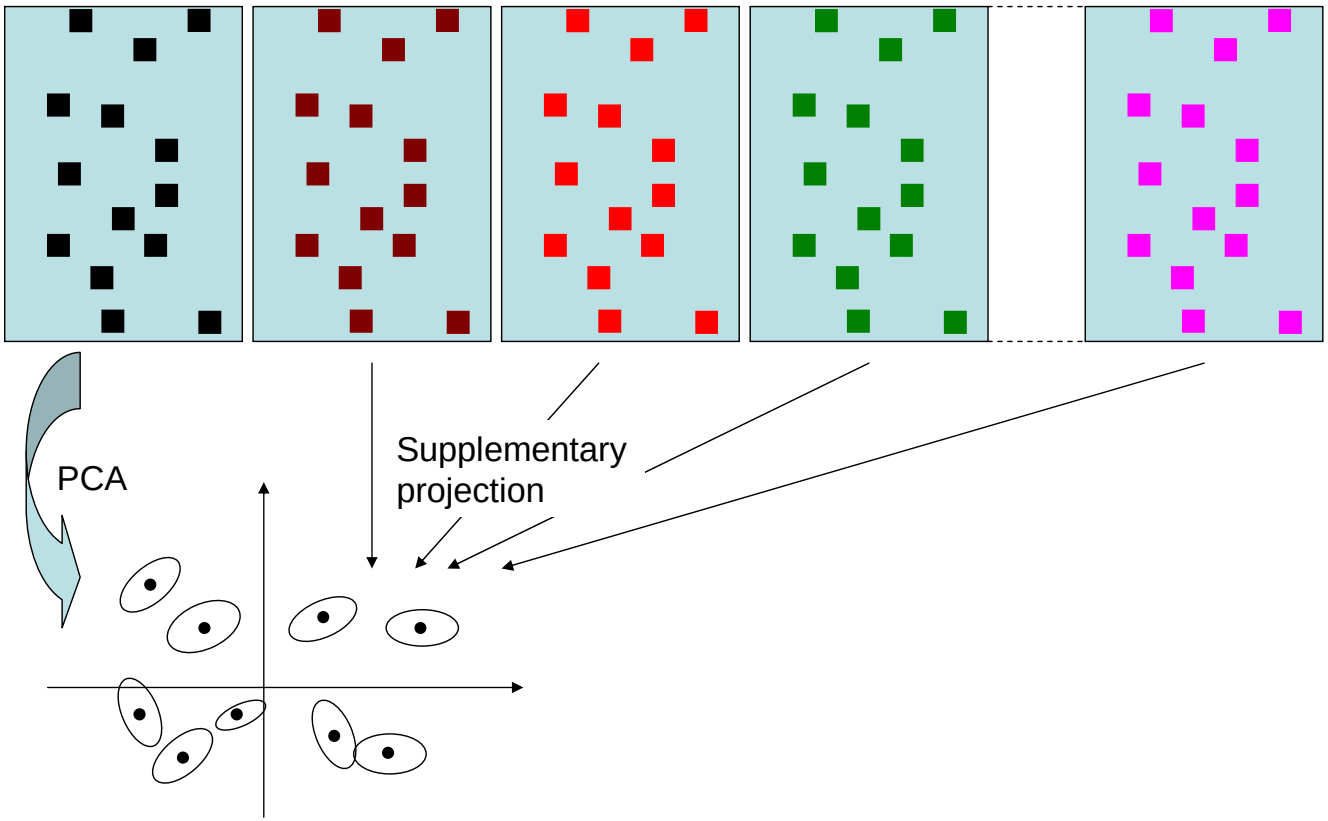
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

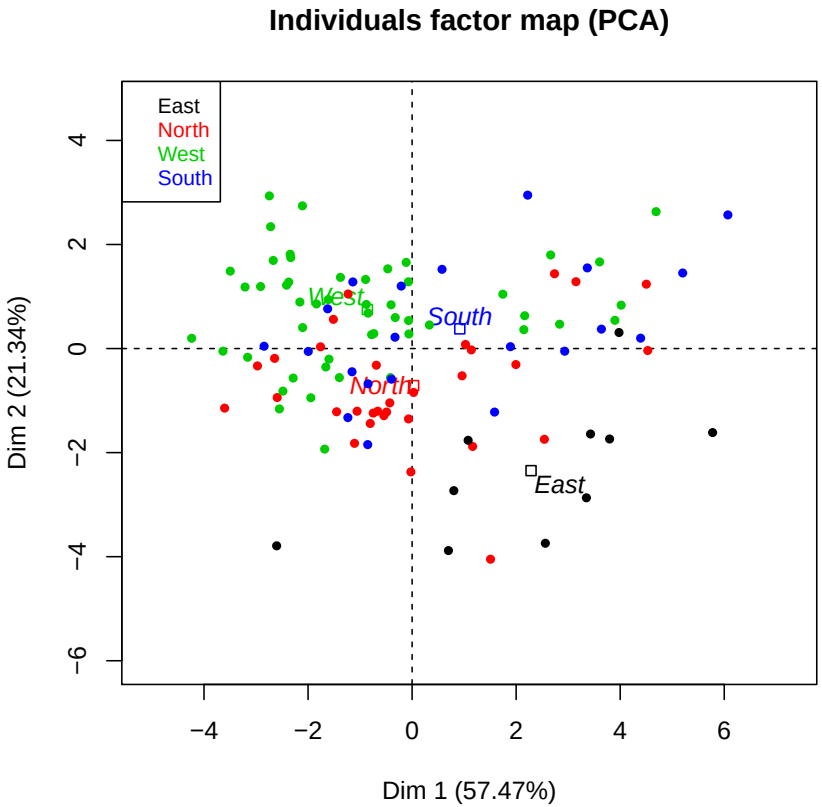
⇒ Individuals position (and variables) with other predictions

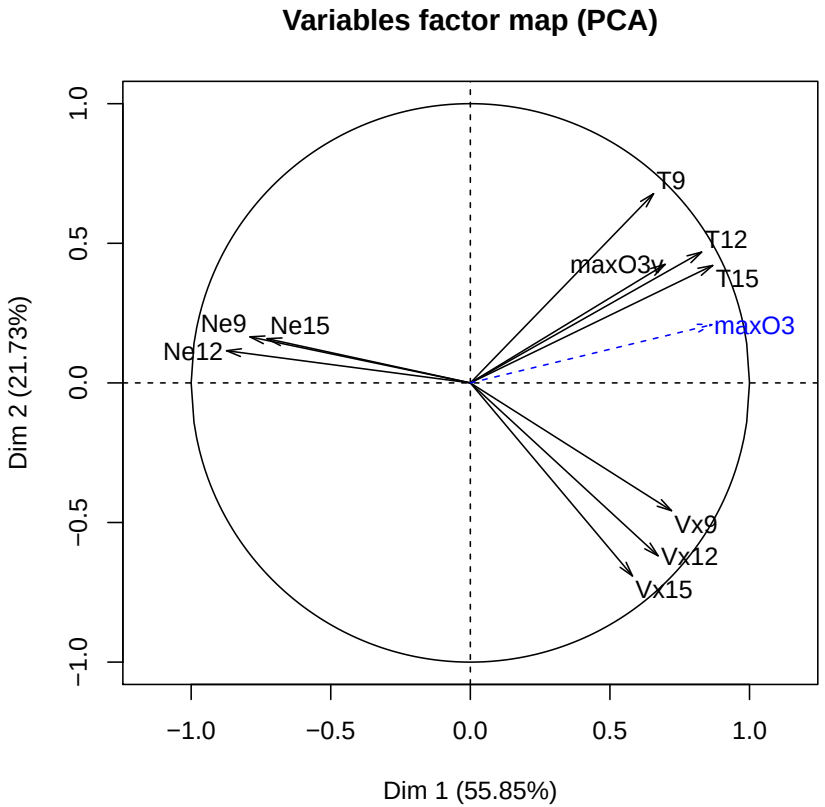


Regularized iterative PCA

⇒ reference configuration

PCA representation



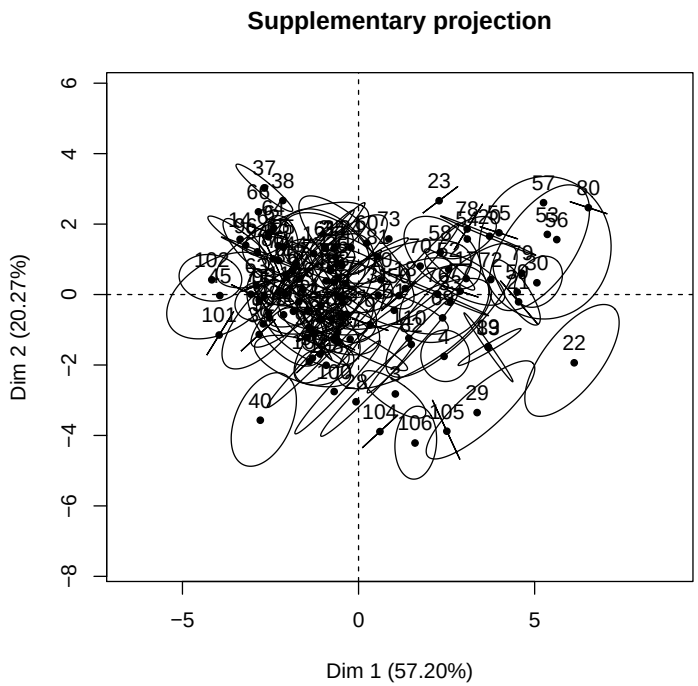


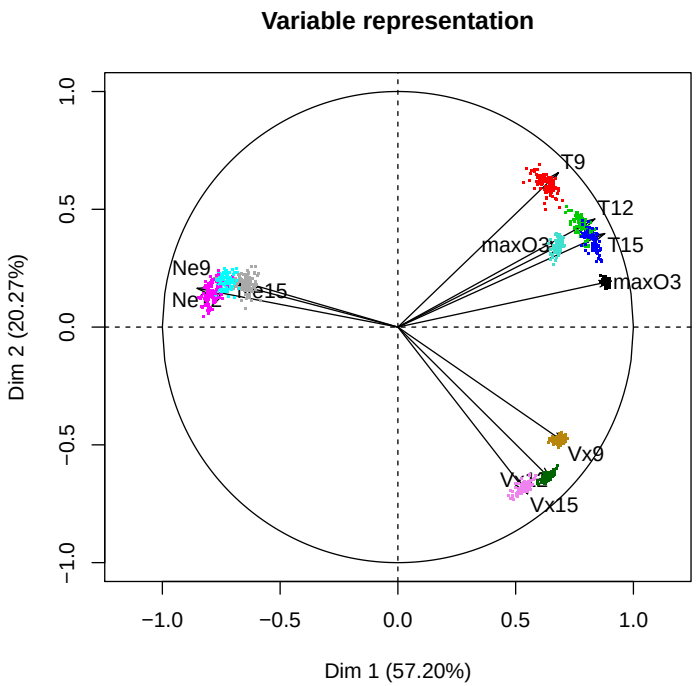

```
> imp <- cbind.data.frame(res.comp$completeObs, ozo[, 12])
> res.pca <- PCA(imp,quanti.sup = 1, quali.sup = 12)
> plot(res.pca, hab =12, lab = "quali"); plot(res.pca, choix = "var")
> res.pca$ind$coord #scores (principal components)
```

Multiple imputation in practice

⇒ Step 2: visualization

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

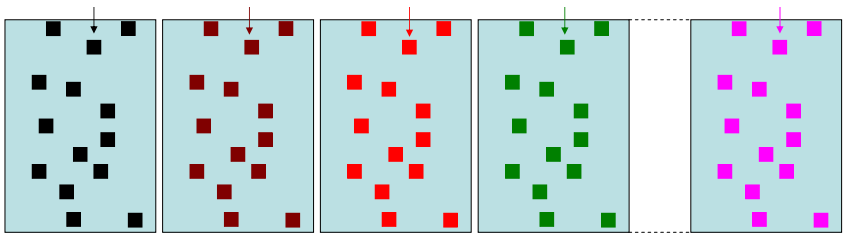




Multiple imputation

Single imputation: a single value can't reflect the uncertainty of prediction $\Rightarrow$ underestimate the standard errors

1. Generating M imputed data sets



2. Performing the analysis on each imputed data set
3. Combining: variance = within + between imputation variance

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

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

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
```

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(missMDA)
#res.MIPCA <- MIPCA(don, ncp = 2, nboot  = 100)
#require(mice)
imp<-prelim(res.mi = res.MIPCA, X = don)#creating a mids object
fit <- with(data=imp,exp=lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v
res.pool <-pool(fit)
summary(res.pool)


names(res.amelia$imputations)
res.amelia$imputations$imp1# the first imputed data set
resamelia <- lapply(res.amelia$imputations, as.data.frame)
# A regression on each imputed dataset
fitamelia<-lapply(resamelia,lm, formula="maxO3~ T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx
# Pool is a function from mice to aggregate the results according
poolamelia <- pool(as.mira(fitamelia))
summary(poolamelia)
```

*"**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)