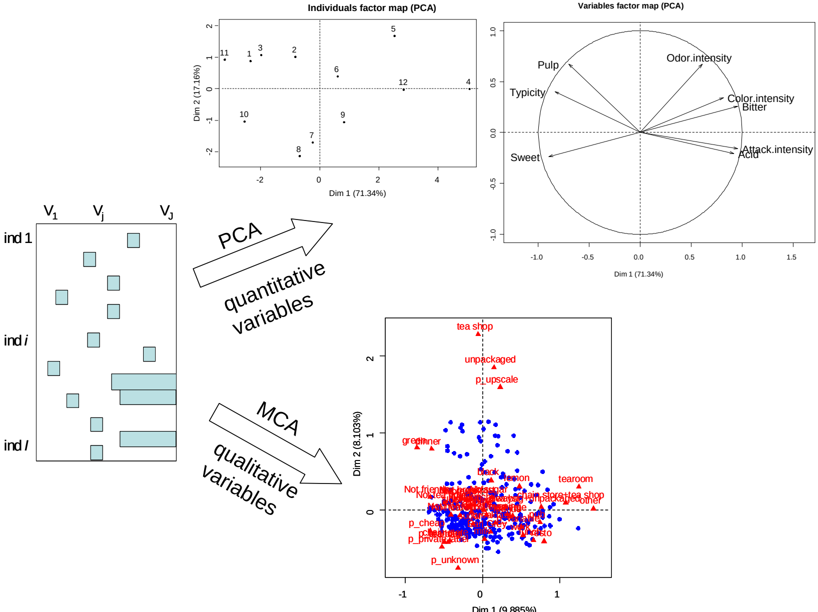
missMDA: a package to handle missing values in Multivariate exploratory Data Analysis methods

Julie Josse & François Husson

Applied Mathematics Department
Agrocampus Rennes - France

useR! 2011, Warwick, 17 August 2011

Aim


Individuals factor map (PCA)
Dim 1 (71.34%)
Dim 2 (17.16%)
Variables factor map (PCA)
Pulp
Odor.intensity
Typicity
Color.intensity
Bitter
Sweet
Attack.intensity
Acid
Dim 1 (71.34%)
V1
Vj
VJ
ind 1
ind i
ind I
PCA
quantitative variables
MCA
qualitative variables
tea shop
unpackaged
p_upscale
tearoom
p_cheap
p_unknown
Dim 2 (8.103%)
Dim 1 (9.885%)

Handling missing values in PCA

⇒ Minimization of:

$$\mathcal{C} = \|\mathbf{X}_{I\times J} - \mathbf{F}_{I\times S}\mathbf{U}^t_{S\times J}\|^2$$

⇒ With missing values:

$$\mathcal{C} = \|\mathbf{W} * (\mathbf{X} - \mathbf{F}\mathbf{U}^t)\|^2,$$

with $w_{ij} = 0$ if x_{ij} is missing, $w_{ij} = 1$ otherwise.

⇒ Criss-cross multiple regression (Gabriel & Zamir, 1979), iterative PCA (Kiers, 1997)

Iterative PCA

1. initialization $\ell = 0$: $\mathbf{X}^0$ (mean imputation)
2. step ℓ:
   (a) PCA is performed on the completed data set $\rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$; S dimensions are kept
   (b) missing values are imputed with the model matrix $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell\prime}$; the new imputed dataset is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$
   (c) means (and standard deviations) are updated
3. steps are repeated until convergence

$\Rightarrow$ The number of dimensions S has to be chosen *a priori*

$\Rightarrow$ Imputation method

$\Rightarrow$ EM algorithm of $x_{ij} = \sum_{s=1}^{S} f_{is} u_{js} + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

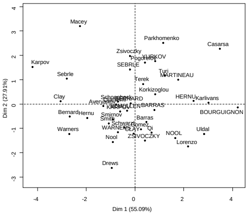
Iterative PCA

1. initialization $\ell = 0$: $\mathbf{X}^0$ (mean imputation)
2. step ℓ:
   (a) PCA is performed on the completed data set $\rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$;
   S dimensions are kept
   (b) missing values are imputed with the model matrix $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell\prime}$;
   the new imputed dataset is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$
   (c) means (and standard deviations) are updated
3. steps are repeated until convergence

$\Rightarrow$ The number of dimensions S has to be chosen *a priori*
$\Rightarrow$ Imputation method
$\Rightarrow$ EM algorithm of $x_{ij} = \sum_{s=1}^{S} f_{is} u_{js} + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

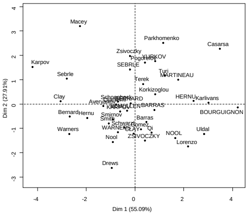
Overfitting

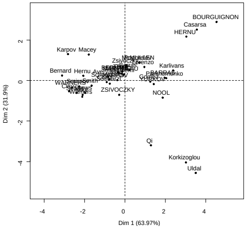
$$X_{41\times 6} = \mathbf{F}_{41\times 2}\mathbf{U}'_{2\times 6} + \mathcal{N}(0, 0.5);$$



Overfitting

$$X_{41\times 6} = \mathbf{F}_{41\times 2}\mathbf{U}'_{2\times 6} + \mathcal{N}(0, 0.5);\ 50\%\text{ of NA}$$

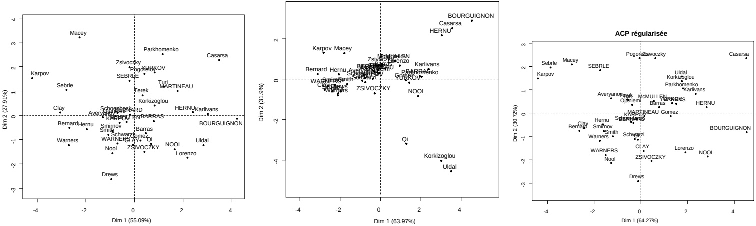




$$||\mathbf{W} * (\mathbf{X} - \hat{\mathbf{X}})|| = 0.48;\ ||(1 - \mathbf{W}) * (\mathbf{X} - \hat{\mathbf{X}})|| = 5.58$$

Overfitting

$X_{41\times 6} = \mathbf{F}_{41\times 2}\mathbf{U}'_{2\times 6} + \mathcal{N}(0, 0.5)$; 50% of NA



$||\mathbf{W} * (\mathbf{X} - \hat{\mathbf{X}})|| = 0.48$; $||(1 - \mathbf{W}) * (\mathbf{X} - \hat{\mathbf{X}})|| = 5.58$

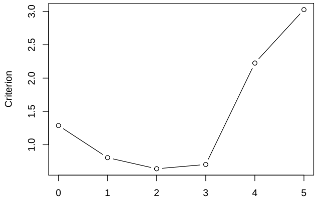
$\Rightarrow$ Regularized iterative PCA: $||(1 - \mathbf{W}) * (\mathbf{X} - \hat{\mathbf{X}})|| = 0.67$

Step 1: Estimation of the number of dimensions

```
  Sweet Acid ... Bitter Pulp Typicity
1    NA   NA ...   2.83   NA     5.21
2  5.46 4.13 ...   3.54 4.62     4.46
3    NA 4.29 ...   3.17 6.25     5.17
..           ...
12 4.88 5.29 ...   4.17 1.50     3.50
```

⇒ EM cross-validation (Bro, 2008); GCV (Josse & Husson, 2011)

```
> nb <- estim_ncpPCA(orange)
> nb$ncp        #2
> nb$criterion
        0         1         2         3         4         5
1.2884873 0.8069719 0.6400517 0.7045074 2.2257738 3.0274337
```



Step 2: Imputation of the missing values

```
> res.comp <- imputePCA(orange,ncp=2,
    scale=TRUE,method="regularized")
```

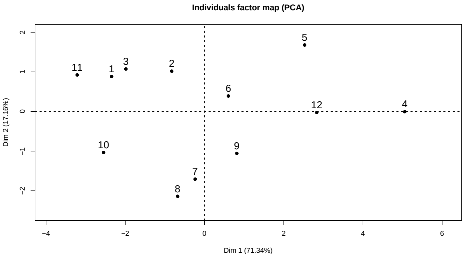
```
> orange
 Sweet Acid Bitter Pulp Typicity
  NA    NA   2.83    NA      5.21
5.46 4.13   3.54 4.62      4.46
  NA 4.29   3.17 6.25      5.17
4.17 6.75     NA 1.42      3.42
...
  NA    NA     NA 7.33      5.25
4.88 5.29   4.17 1.50      3.50
```

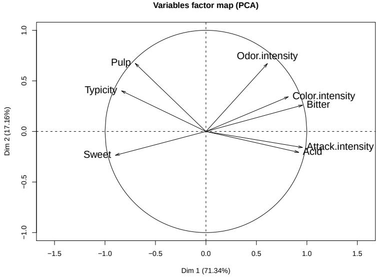
```
> res.comp$completeObs
Sweet Acid Bitter Pulp Typicity
 5.54 4.13   2.83 5.89     5.21
 5.46 4.13   3.54 4.62     4.46
 5.45 4.29   3.17 6.25     5.17
 4.17 6.75   4.73 1.42     3.42
...
 5.71 3.87   2.80 7.33     5.25
 4.88 5.29   4.17 1.50     3.50
```

Step 3: PCA on the completed data set

```
> res.pca <- PCA(res.comp$completeObs)
```

⇒ library FactoMineR

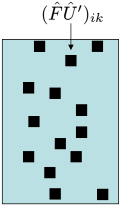





```
> res.pca$ind$coord #scores (principal components)
> res.pca$var$coord
```

MI-PCA

⇒ Iterative PCA: single imputation method

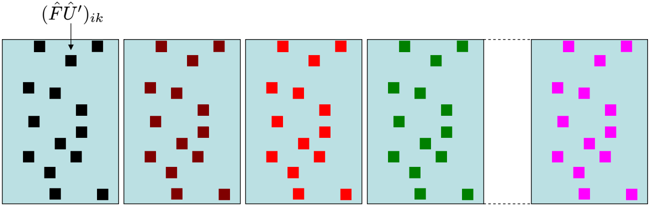


⇒ A unique value cannot reflect the variability of prediction

```
> mi <- MIPCA(orange, scale = TRUE, method = "Regularized", ncp=2)
> mi$res.MI
```

MI-PCA

⇒ Iterative PCA: single imputation method



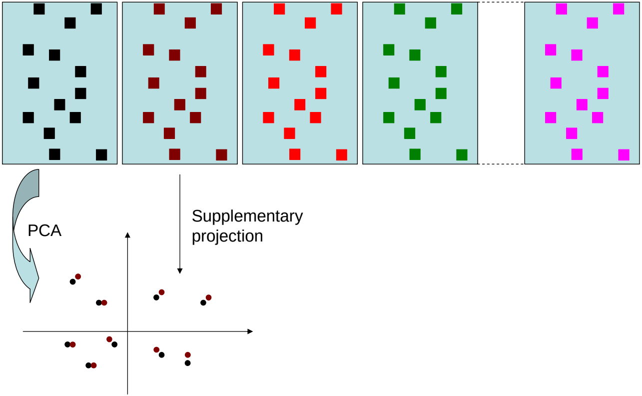
⇒ A unique value cannot reflect the variability of prediction

⇒ Multiple imputation: generating plausible values for each missing value

```
> mi <- MIPCA(orange, scale = TRUE, method = "Regularized", ncp=2)
> mi$res.MI
```

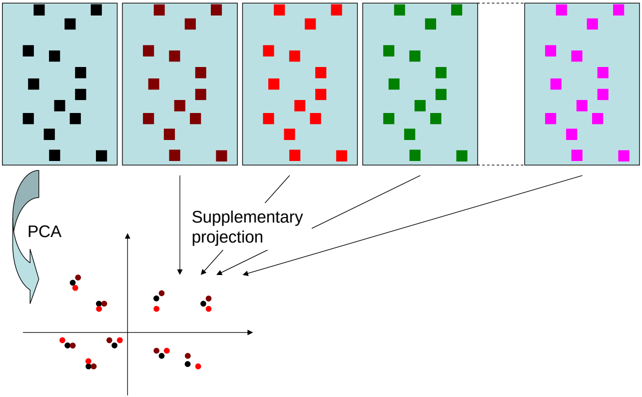
Supplementary projection

⇒ Individuals position (and variables) with other predictions



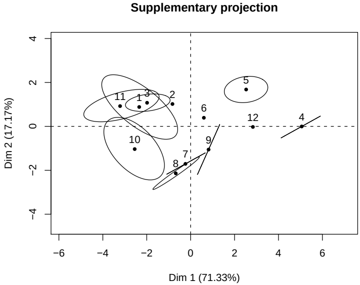
Supplementary projection

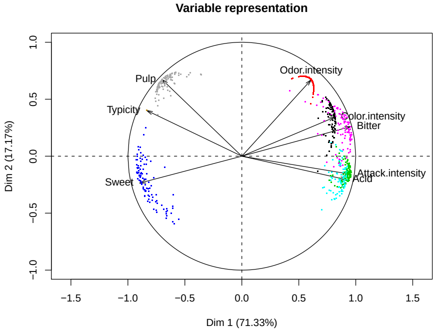
⇒ Individuals position (and variables) with other predictions



Supplementary projection

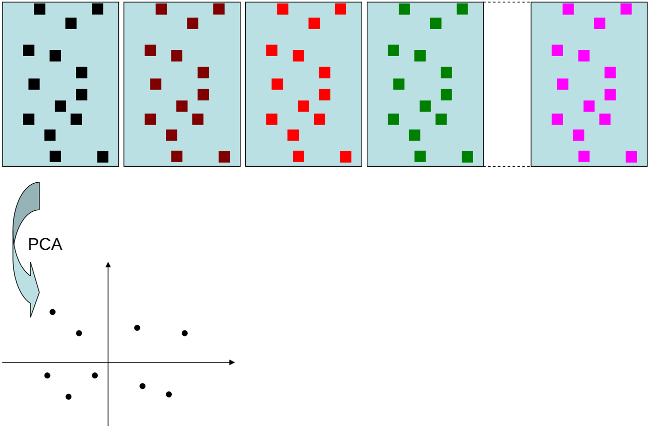
```
> plot(mi)
```





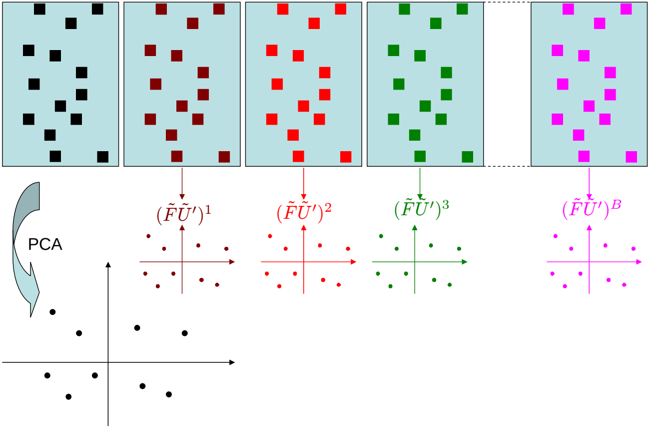
Between imputation variability

⇒ Influence of the different predictions on the parameters (PCA on each table)



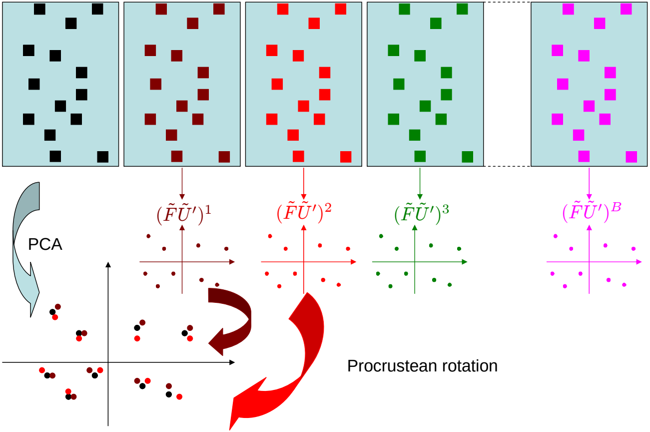
Between imputation variability

⇒ Influence of the different predictions on the parameters (PCA on each table)



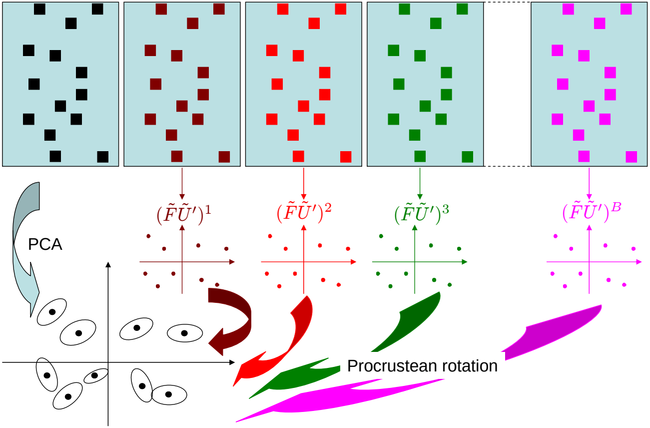
Between imputation variability

⇒ Influence of the different predictions on the parameters (PCA on each table)

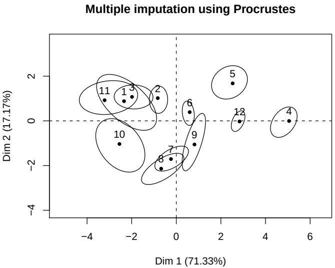


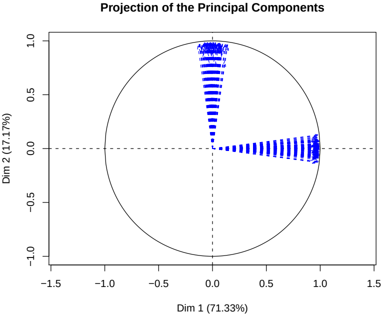
Between imputation variability

⇒ Influence of the different predictions on the parameters (PCA on each table)



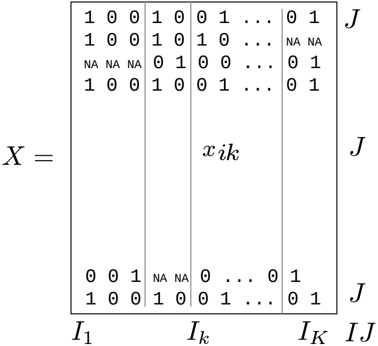
Between imputation variability





Handling missing values in MCA

MCA is a PCA on the indicator matrix **X** with specific rows and columns weights



$\Rightarrow$ Regularized iterative MCA

1. Initialization: imputation of the indicator matrix (proportion)
2. Iterate until convergence
   (a) Estimation of $\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell$: MCA on the completed indicator matrix
   (b) Imputation of the missing values with the model matrix
   (c) Column margins are updated

Imputation of the indicator matrix

```
> data(vnf)
> ncp <- estim_ncpMCA(vnf)
> tab.disj <- imputeMCA(vnf,ncp=4)
```

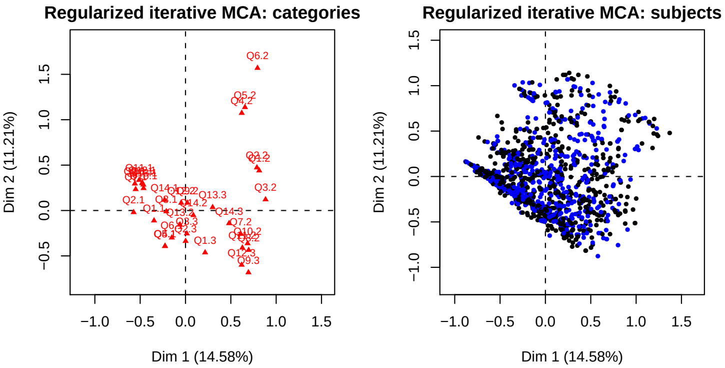
	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

MCA using the completed indicator matrix

```
> res.mca <- MCA(vnf,tab.disj=tab.disj)
```

⇒ library FactoMineR




```
> res.mca$ind$coord #scores
> res.mca$var$coord
```

Conclusion

⇒ missMDA handles missing values in PCA and MCA and also in multi-way methods (imputeMFA)

- Single imputation for continuous and categorical variables
- Multiple imputation: an alternative to mice or Amelia packages?