PCA with missing data using the missMDA R package

François Husson

Applied Mathematics Department, Rennes Agrocampus

husson@agrocampus-ouest.fr

Using missMDA to deal with missing data

```
> library(missMDA)
> data(orange)
```

	Color intensity	Odor intensity	Attack intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.79	5.29	NA	NA	NA	2.83	NA	5.21
2	4.58	6.04	4.42	5.46	4.13	3.54	4.62	4.46
3	4.71	5.33	NA	NA	4.29	3.17	6.25	5.17
4	6.58	6.00	7.42	4.17	6.75	NA	1.42	3.42
5	NA	6.17	5.33	4.08	NA	4.38	3.42	4.42
6	6.33	5.00	5.38	5.00	5.50	3.63	4.21	4.88
7	4.29	4.92	5.29	5.54	5.25	NA	1.29	4.33
8	NA	4.54	4.83	NA	4.96	2.92	1.54	3.96
9	4.42	NA	5.17	4.62	5.04	3.67	1.54	3.96
10	4.54	4.29	NA	5.79	4.38	NA	NA	5.00
11	4.08	5.13	3.92	NA	NA	NA	7.33	5.25
12	6.50	5.88	6.13	4.88	5.29	4.17	1.50	3.50

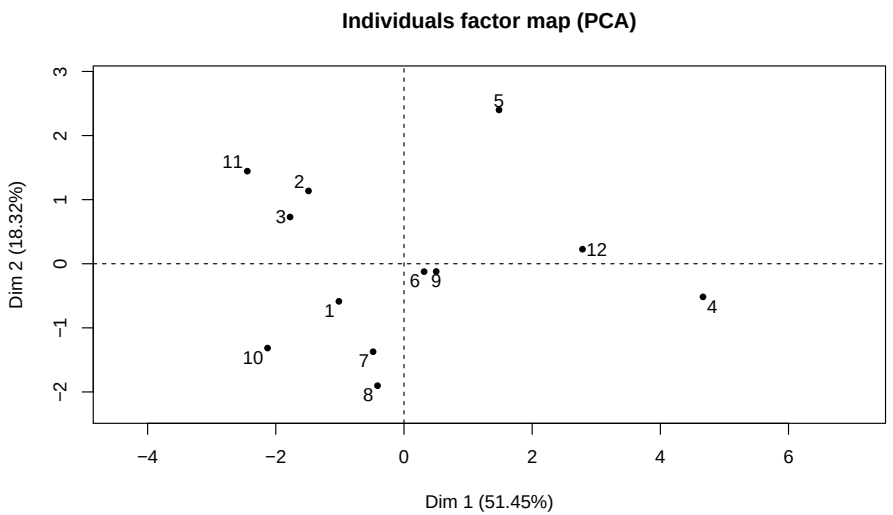
Some (bad) easy methods

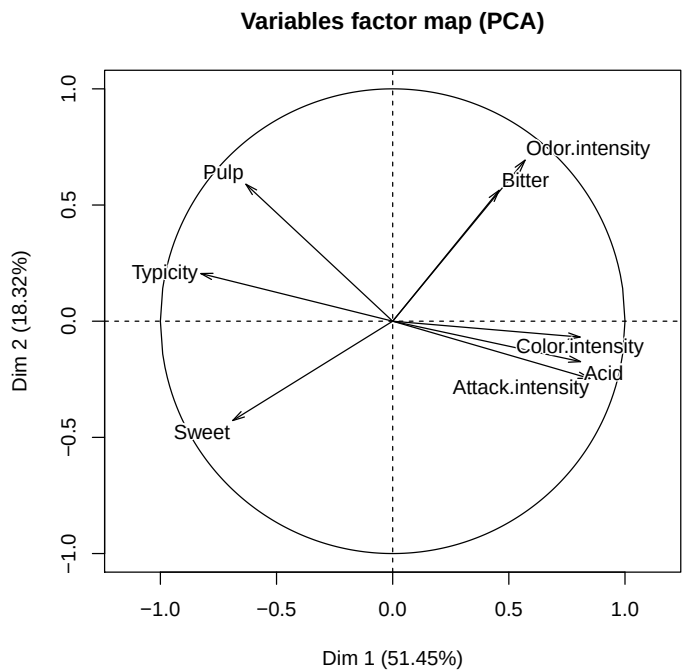
- Delete individuals or variables with missing data : usually not a good idea
- Replace missing data with the mean (default in several packages including FactoMineR)

Some (bad) easy methods

- Delete individuals or variables with missing data : usually not a good idea
- Replace missing data with the mean (default in several packages including FactoMineR)

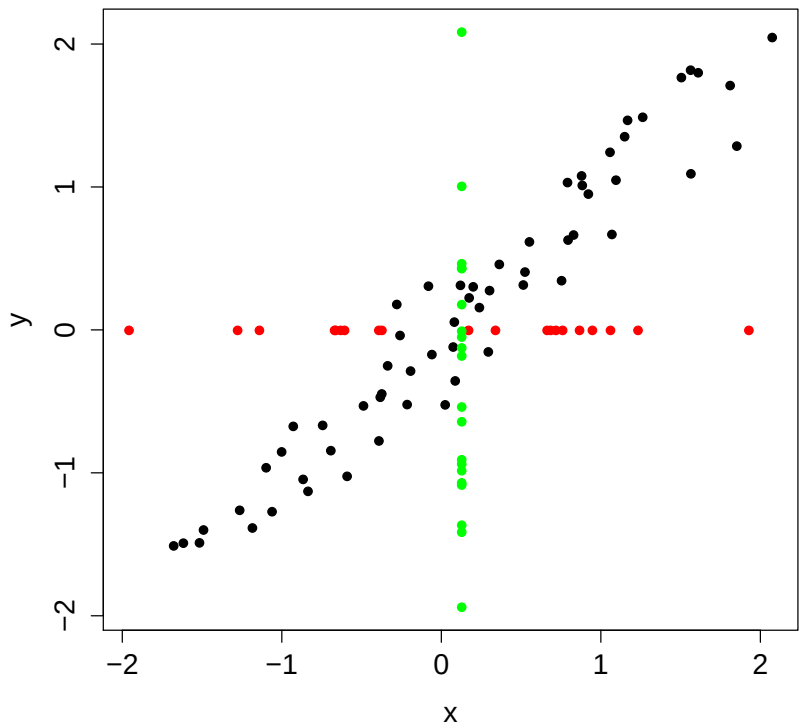
```
> res.pca <- PCA(orange)
```





Some (bad) easy methods

- Delete individuals or variables with missing data : usually not a good idea
- Replace missing data with the mean (default in several packages including FactoMineR)



Big distortion of links between variables

Iterative PCA

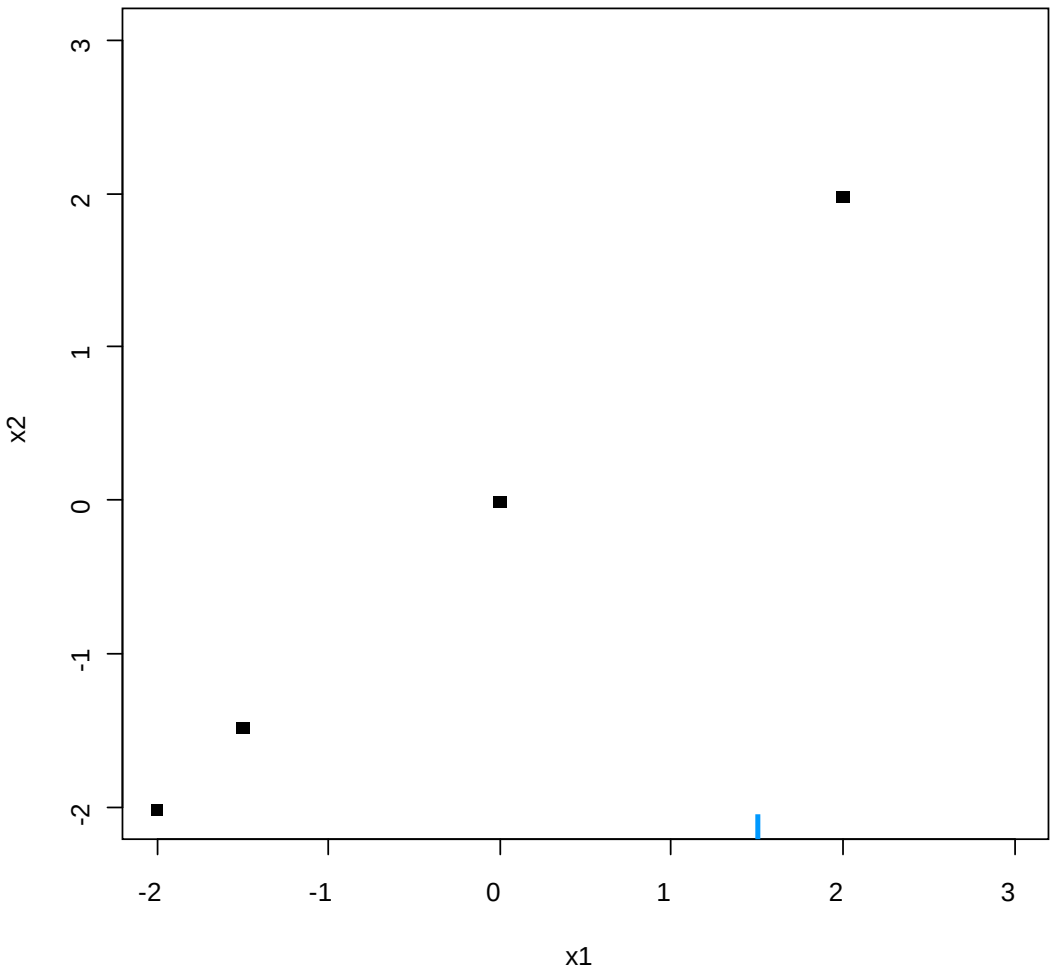
Ideas :

- As x and y strongly correlated : impute missing y value using x value
- if individuals i and j have similar values for all variables, impute missing i value using j value for that variable

$\Rightarrow$ takes into account global similarity between individuals and links between variables

Iterative PCA

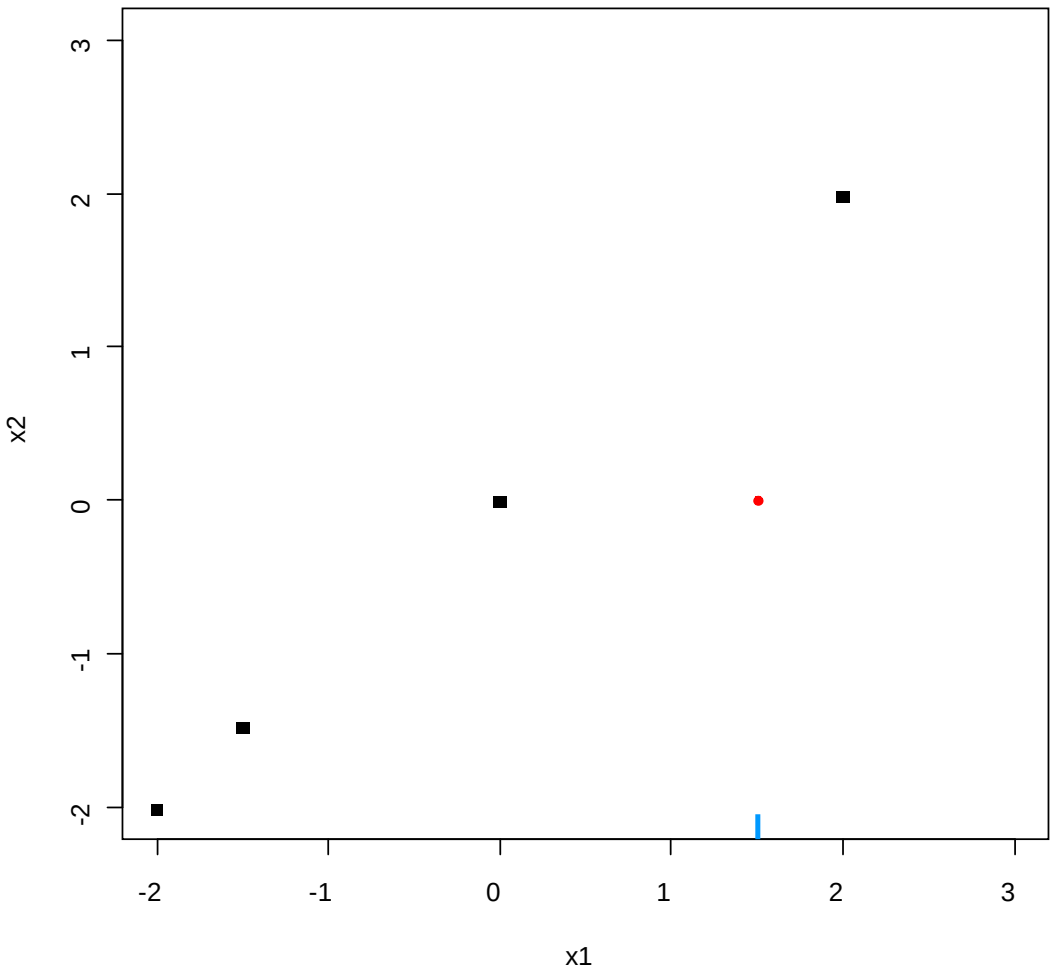
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



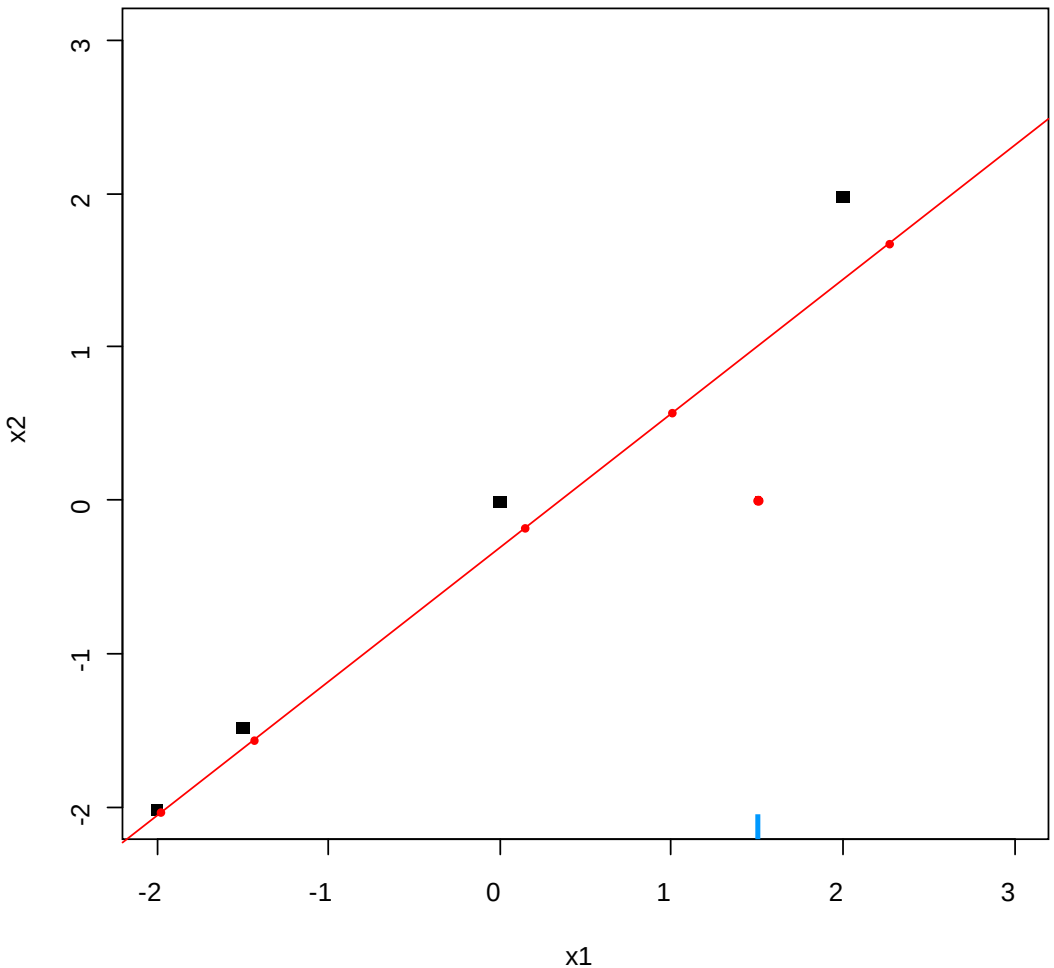
Initialize : impute the mean

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



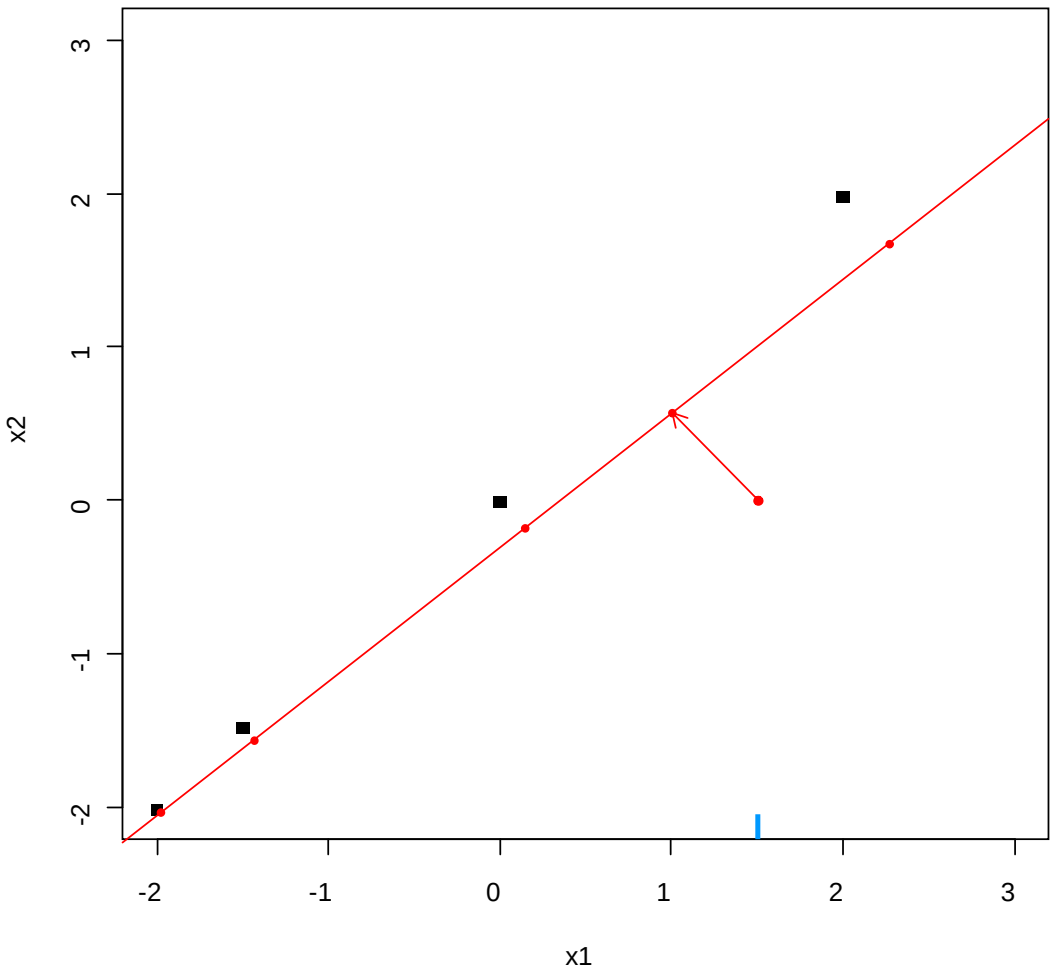
Do PCA on imputed table $\rightarrow$ axes and components ;

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 data imputed using PCA

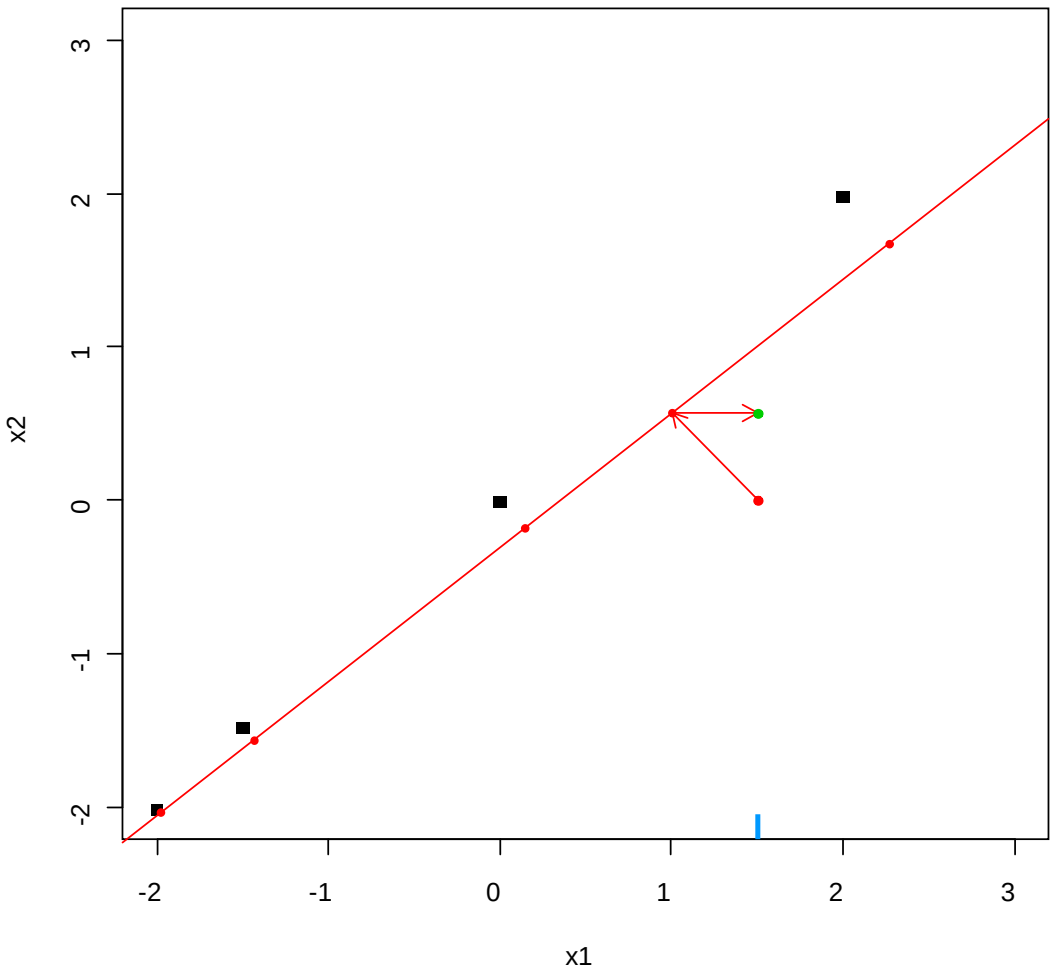
Iterative PCA

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



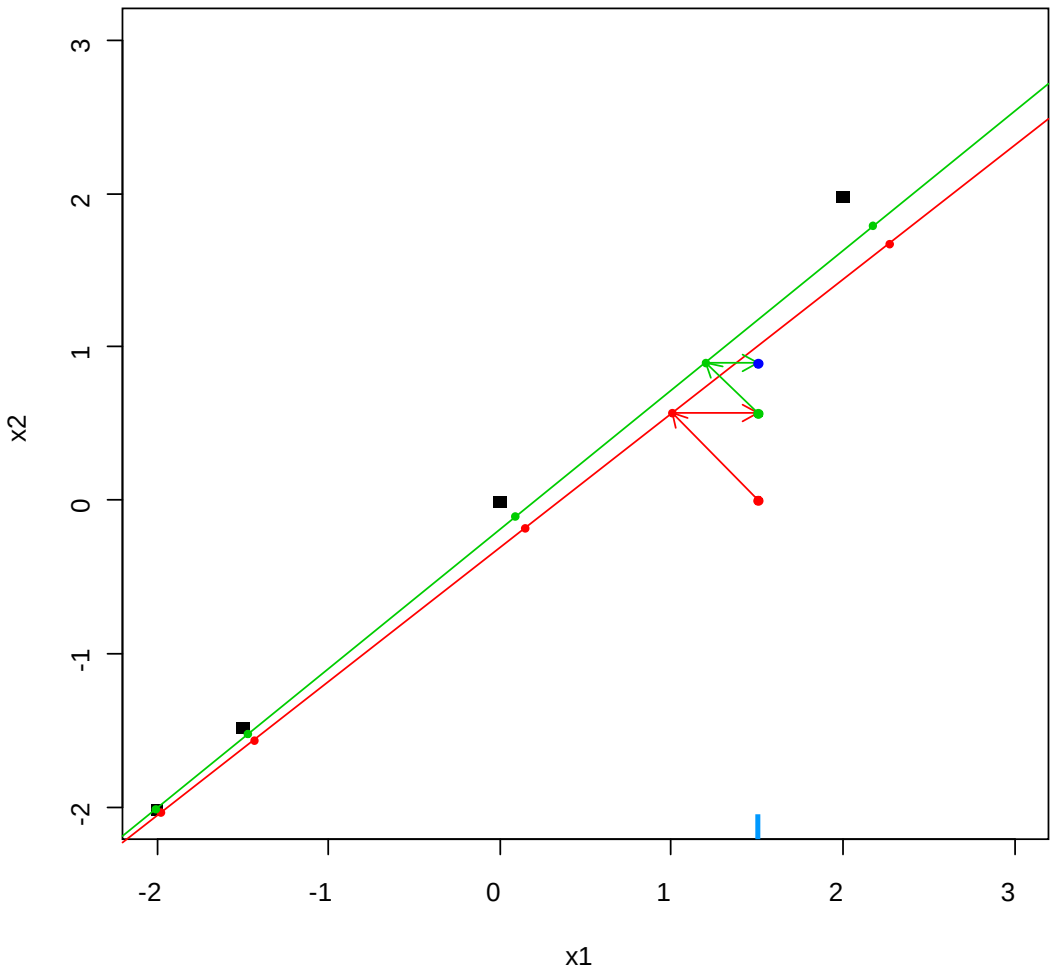
New imputed data table

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



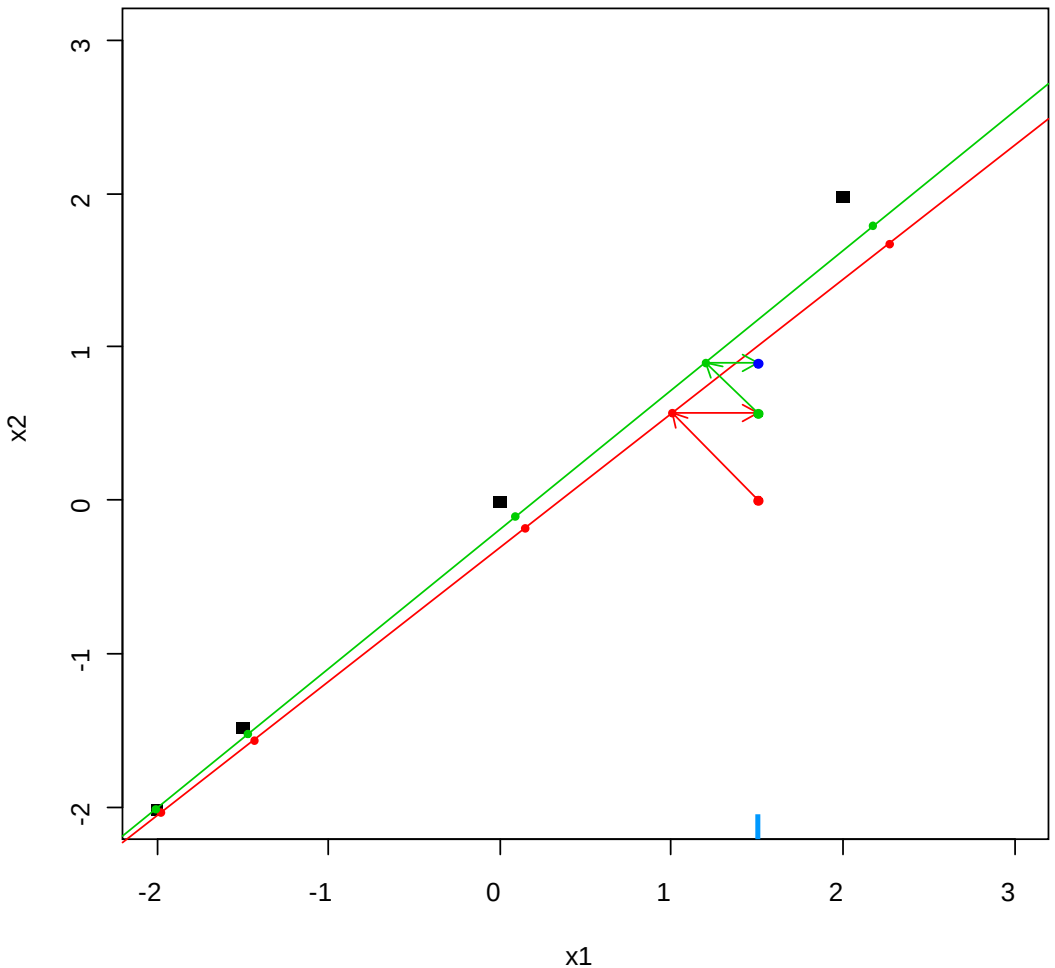
Iterative PCA

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

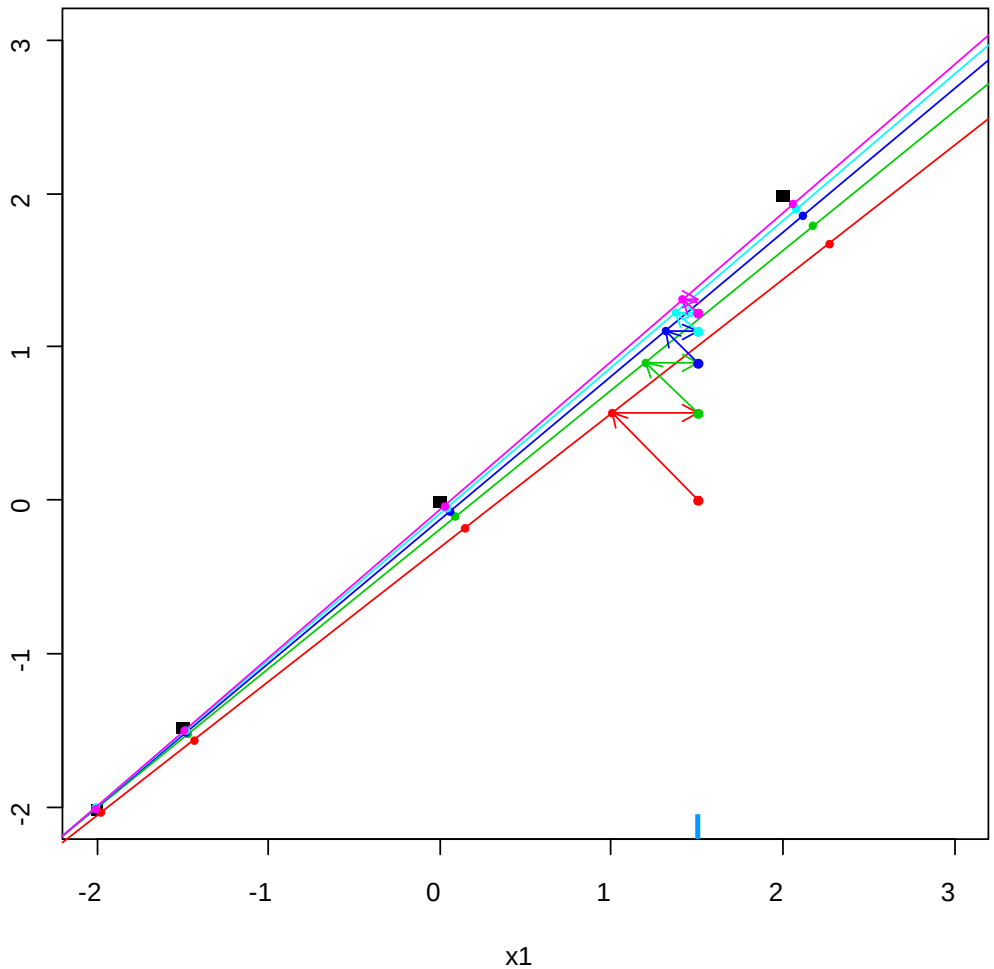
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

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

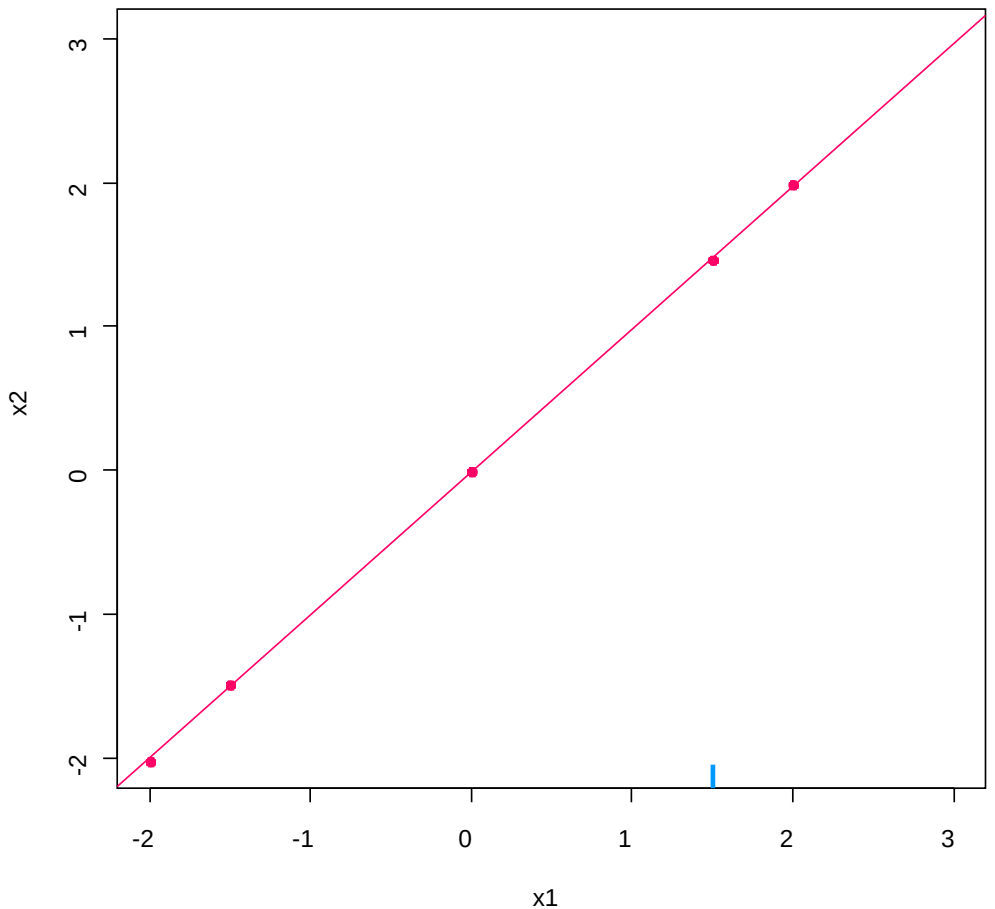
x2



Repeat these steps 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	1.48
2.0	1.98

Do PCA on imputed data table

Iterative PCA

1. initialization : impute using the mean
2. Step ℓ :
   (a) do PCA on imputed data table
   S dimensions retained
   (b) missing data imputed using PCA
   (c) means (and standard deviations) updated
3. iterate the estimation and imputation steps

Iterative PCA

1. initialization : impute using the mean
2. Step ℓ :
   (a) do PCA on imputed data table
   S dimensions retained
   (b) missing data imputed using PCA
   (c) means (and standard deviations) updated
3. iterate the estimation and imputation steps

Overfitting problem due to believing too much in links between variables

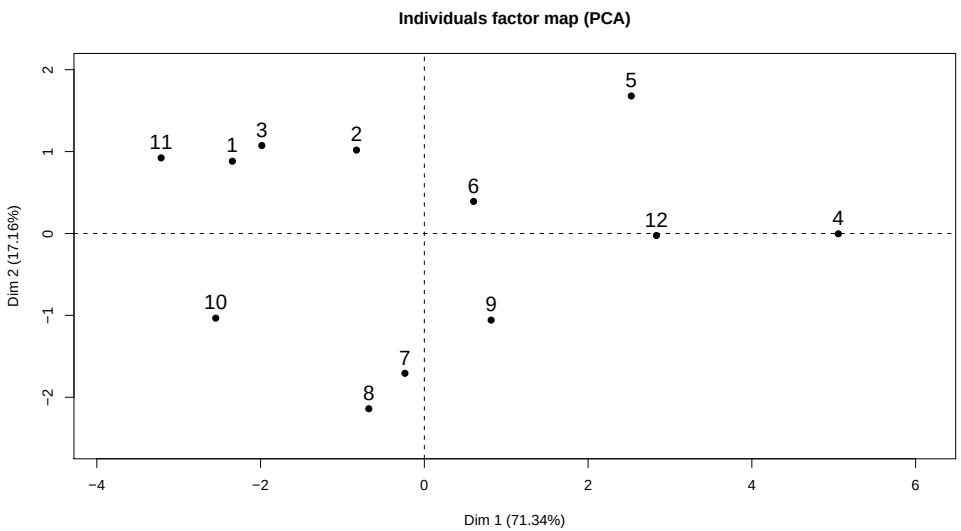
$\Rightarrow$ regularized iterative PCA

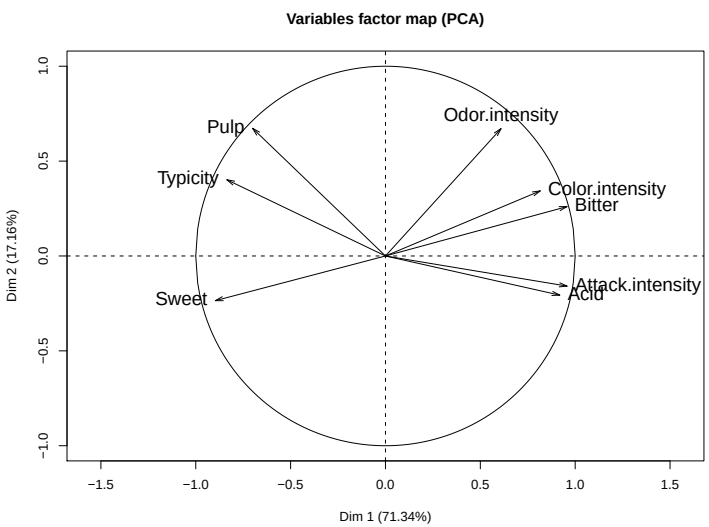
Running missMDA in R

```
> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange, scale=TRUE)          ## Estimate no. of dimensions
> comp <- imputePCA(orange, ncp=2, scale=TRUE)    ## Impute the table
> res.pca <- PCA(comp$completeObs)                ## Do the PCA
```

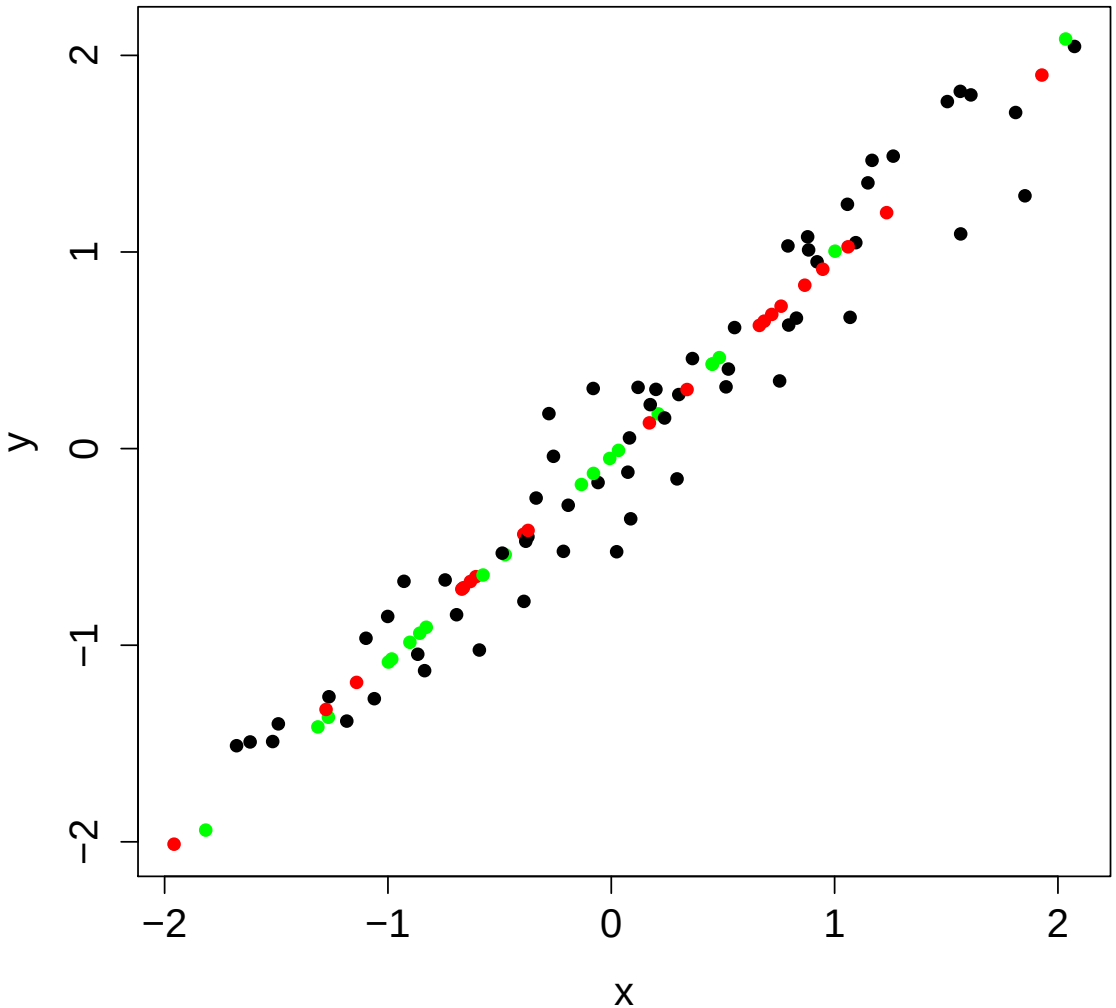
```
> orange
 Sweet Acid Bitter Pulp Typicity
  NA   NA  2.83   NA      5.21
5.46 4.13  3.54 4.62      4.46
  NA 4.29  3.17 6.25      5.17
...
4.88 5.29  4.17 1.50      3.50
```

```
> 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
...
 4.88 5.29   4.17 1.50     3.50
```





Is running the imputation algorithm once sufficient ?

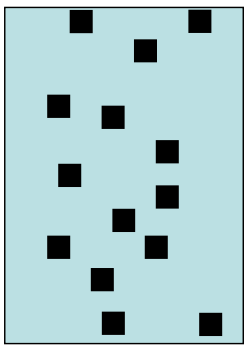


$\Rightarrow$ Reinforces links between variables

Visualizing uncertainty due to missing data

What confidence can we give to the results ? Idea of variance ?

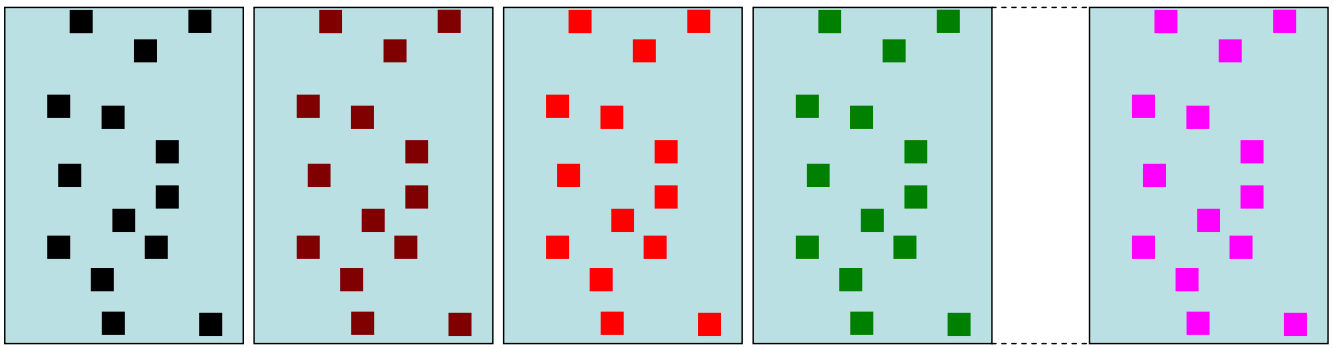
$\Rightarrow$ A single value cannot show variability in the predicted value



Visualizing uncertainty due to missing data

What confidence can we give to the results ? Idea of variance ?

$\Rightarrow$ A single value cannot show variability in the predicted value



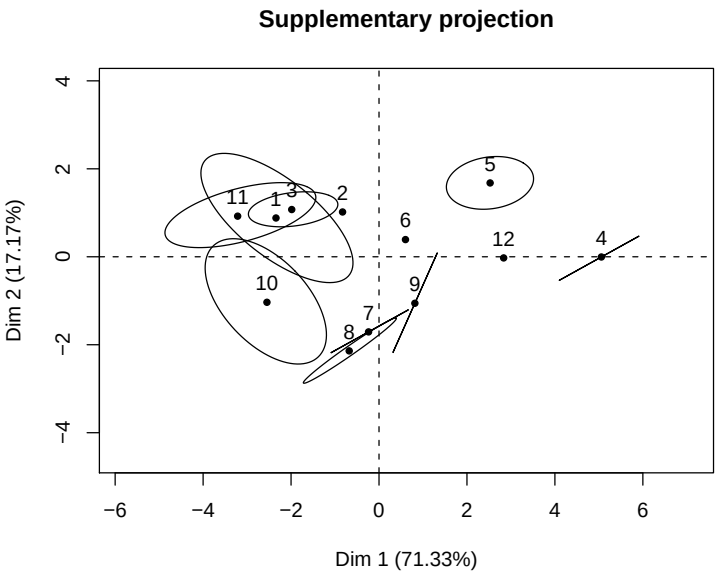
$\Rightarrow$ Multiple imputation : generate several plausible values for each missing data point

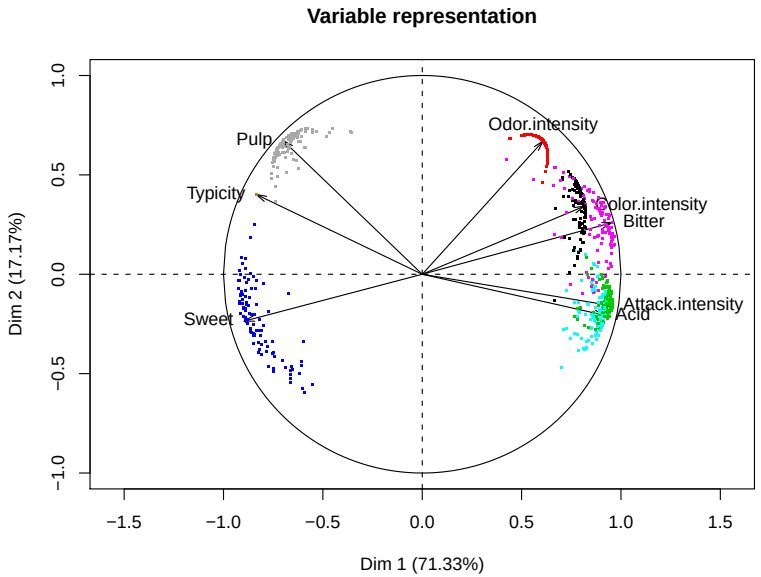
Visualizing uncertainty due to missing data

```
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> plot(mi)
```

Visualizing uncertainty due to missing data

```
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> plot(mi)
```





Visualizing uncertainty due to missing data

```
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> plot(mi)
```