

Prise en compte des données manquantes en ACP

Julie Josse, François Husson

Laboratoire de mathématiques appliquées, Agrocampus Ouest, Rennes

3 octobre 2011

Laboratoire de mathématiques appliquées



Laboratoire de mathématiques appliquées

- Recherche
 - Analyse factorielle, multi-tableaux
 - modélisation en grande dimension - tests multiples
 - Application : analyse de données sensorielles et génomiques
- Enseignement
 - L3 : modèle linéaire, analyse de données, plan d'expériences
 - Spécialisation et Master statistique appliquée : sensométrie, tableaux multiples, données génomiques
 - Livres : Analyse de données avec R, Statistique avec R, Analyse factorielle simple et multiple, Statistique générale
- Autres activités
 - Packages R : FactoMineR, SensoMineR, FAMT, missMDA
 - Formation continue : statistique avec R, statistiques générales
 - Congrès : useR!2009, CARME 2011, Sensometrics 2012

⇒ Visitez notre site internet!

Plan

- 1 Rappel d'ACP
- 2 Prise en compte des données manquantes en ACP
- 3 Zone de confiance
- 4 Conclusion

Plan

- 1 Rappel d'ACP
- 2 Prise en compte des données manquantes en ACP
⇒ Mise en œuvre avec le package R missMDA
- 3 Zone de confiance
⇒ Mise en œuvre avec le package R missMDA
- 4 Conclusion

Plan

- 1 Rappel d'ACP
- 2 Données manquantes en ACP
- 3 Zone de confiance
- 4 Conclusion

Les données jus d'orange

| | Intensité odeur | Typicité odeur | Caractère pulpeux | Intensité goût | Caractère acide | Caractère amer | Caractère sucré |
|----------------|--------------------|-------------------|----------------------|-------------------|--------------------|-------------------|--------------------|
| Pampryl amb. | 2.82 | 2.53 | 1.66 | 3.46 | 3.15 | 2.97 | 2.6 |
| Tropicana amb. | 2.76 | 2.82 | 1.91 | 3.23 | 2.55 | 2.08 | 3.32 |
| Fruvita fr. | 2.83 | 2.88 | 4 | 3.45 | 2.42 | 1.76 | 3.38 |
| Joker amb. | 2.76 | 2.59 | 1.66 | 3.37 | 3.05 | 2.56 | 2.8 |
| Tropicana fr. | 3.2 | 3.02 | 3.69 | 3.12 | 2.33 | 1.97 | 3.34 |
| Pampryl fr. | 3.07 | 2.73 | 3.34 | 3.54 | 3.31 | 2.63 | 2.9 |

⇒ Données recueillies au laboratoire

ACP

⇒ Données : tableau individus \times variables quantitatives

⇒ Statistique descriptive multidimensionnelle (résumer ; visualiser)

⇒ Objectifs :

- typologie des individus
- bilan des liaisons entre variables
- caractérisation des individus à partir des variables

Nuage des individus



- Les individus vivent dans \mathbb{R}^p
- Etudier la forme du nuage des individus

Ajustement du nuage

Trouver le sous-espace qui fournit la meilleure représentation des données

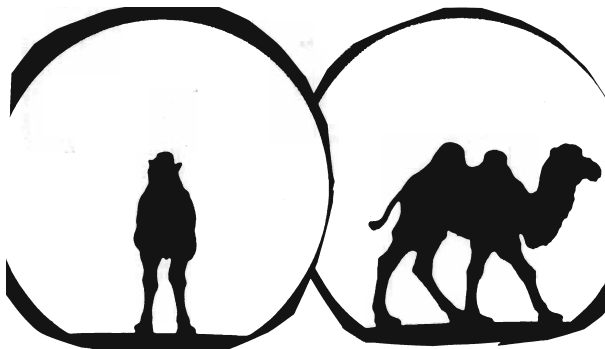
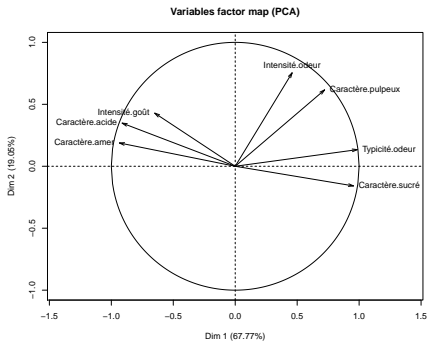
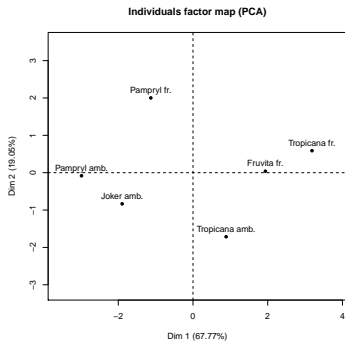


Figure: Chameau ou dromadaire? source J.P. Fenelon

- ⇒ Meilleure approximation par projection
- ⇒ Meilleure représentation de la diversité, de la variabilité

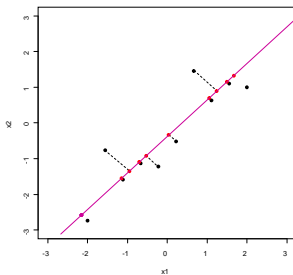
ACP sur les jus d'orange



Ajustement du nuage

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 |



⇒ Minimisation de la distance entre les individus et leur projection

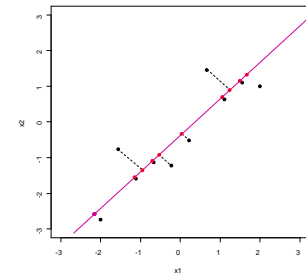
⇒ Minimise $\|X - \hat{X}\|$

Ajustement du nuage

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 |

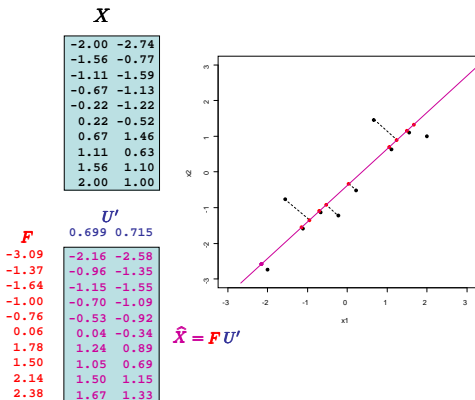


\hat{X}

⇒ Minimisation de la distance entre les individus et leur projection

⇒ Minimise $\|X - \hat{X}\|$

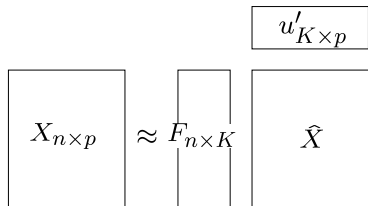
Reconstitution en ACP



$\Rightarrow \hat{X} = F U'$ (produit matriciel utilisant les coordonnées des individus et les coordonnées des variables issues de l'ACP)

Minimiser l'erreur de reconstitution

⇒ Approximation de X par une matrice de rang $K < p$



$$\begin{aligned} \mathcal{C} &= \|\mathbf{X}_{n \times p} - \mathbf{F}_{n \times K} \mathbf{U}'_{K \times p}\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^p \left(x_{ij} - \sum_{k=1}^K f_{ik} u_{jk} \right)^2 \end{aligned}$$

- $\hat{\mathbf{U}}$ axes principaux (normés à 1)
- $\hat{\mathbf{F}}$ composantes principales (normées à la valeur propre)

⇒ Diagonalisation de la matrice de variance-covariance ou de produit-scalaire

Plan

- 1 Rappel d'ACP
- 2 Données manquantes en ACP
- 3 Zone de confiance
- 4 Conclusion

Données manquantes

A diagram illustrating a data matrix with missing values. The matrix is represented as a light blue rectangle with a black border. The horizontal axis is labeled "Variables" and has indices 1, j , and p . The vertical axis is labeled "Individus" and has indices 1, i , and n . Inside the rectangle, several question marks (?) are scattered, representing missing data points. The question marks are located at various intersections of the rows and columns, indicating that data is missing for those specific individuals and variables.

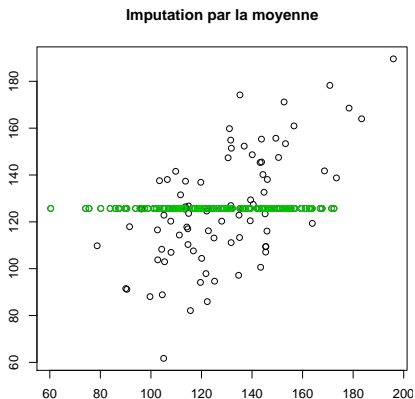
- ⇒ Données manquantes sont (malheureusement!) très fréquentes
- ⇒ La plupart des méthodes statistiques ne peuvent pas être mises en oeuvre directement sur un tableau de données incomplet

Problématique des données manquantes

⇒ Shaefer (1997), Little et Rubin (1987, 2002)

- Méthode très utilisée : suppression
 - ⇒ perte d'information
 - ⇒ étude du dispositif : supprimer une variable?
- Autres méthodes très utilisées : méthodes d'imputation

Imputation par la moyenne



⇒ Déformation des liaisons entre variables

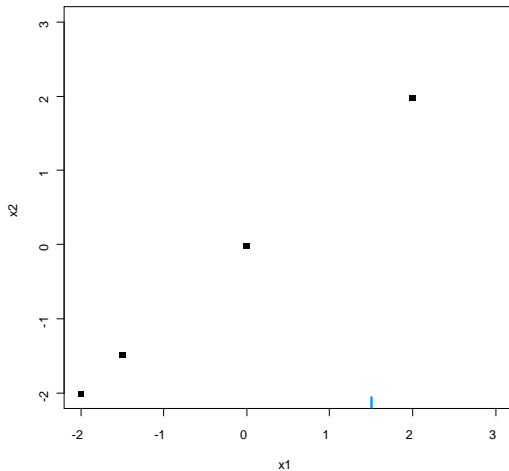
⇒ Problème en ACP!!!

Prise en compte des données manquantes en ACP

Description de l'algorithme d'ACP
itérative - ACP EM

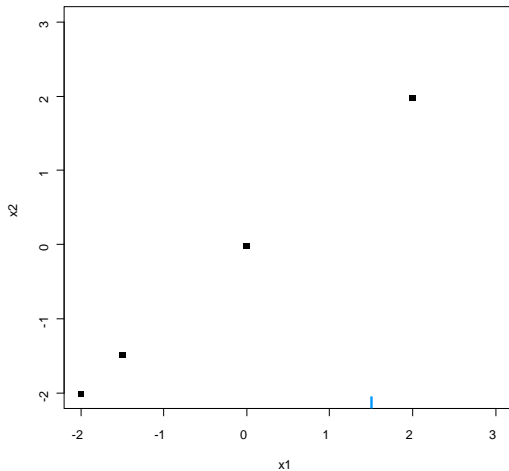
ACP itérative

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



ACP itérative

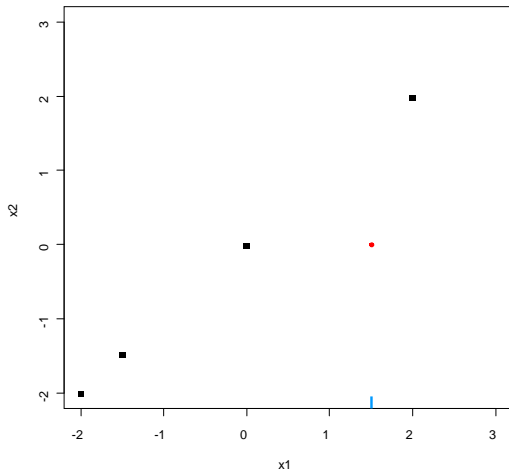
| x1 | x2 |
|------|-------|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |



ACP itérative

| 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 |

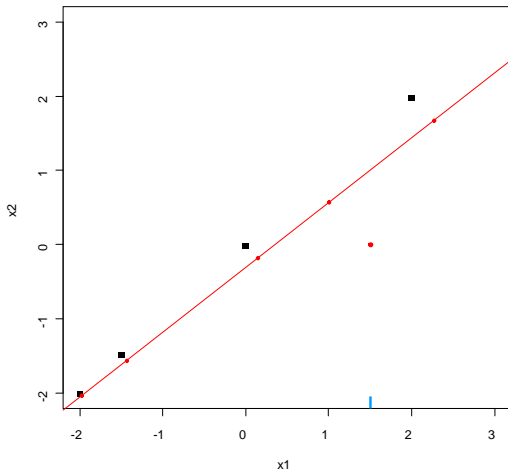


ACP itérative

| 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 |

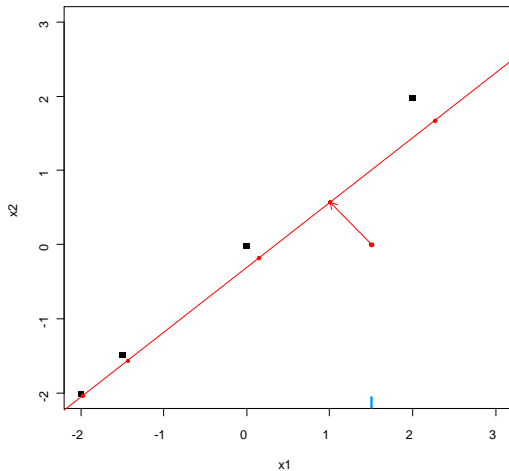


ACP itérative

| 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 |



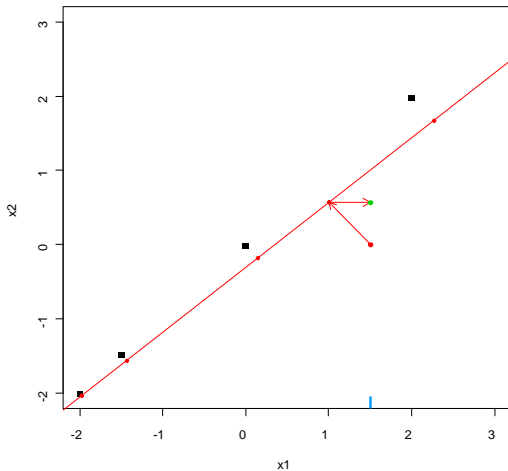
ACP itérative

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

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

| \hat{x}_1 | \hat{x}_2 |
|-------------|-------------|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |

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

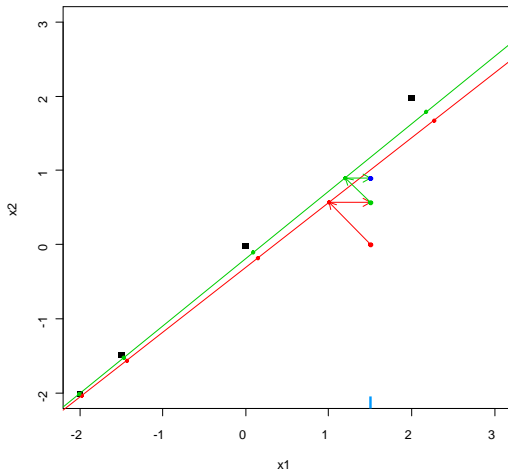


ACP itérative

| 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 |



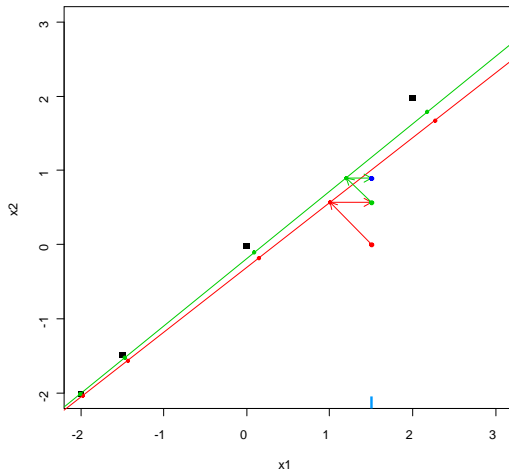
ACP itérative

| 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 |



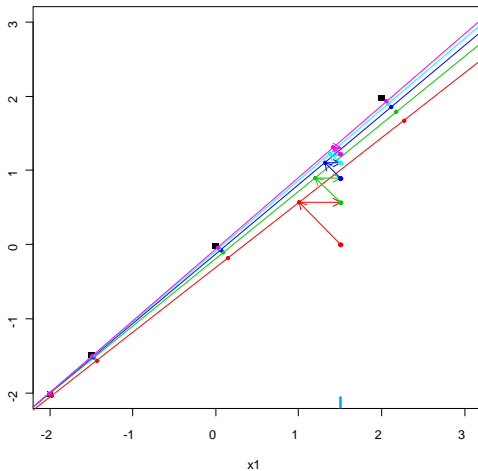
ACP itérative

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

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

| \hat{x}_1 | \hat{x}_2 |
|-------------|-------------|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |

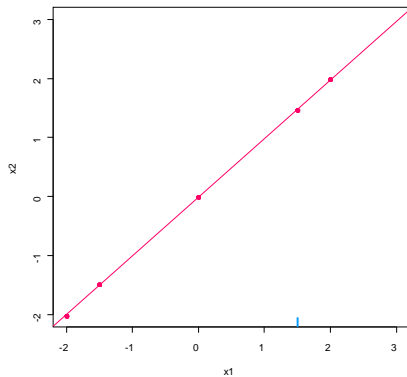
| x1 | x2 |
|------|-------|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |



ACP itérative - convergence

```
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
```



ACP itérative

- 1 initialisation $\ell = 0$: \mathbf{X}^0 (imputation par la moyenne)
- 2 étape ℓ :
 - (a) ACP sur le tableau complété $\Rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$;
 K dimensions conservés
 - (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell r}$;
nouveau tableau : $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$; avec $w_{ij} = 0$ si x_{ij} est manquant et $w_{ij} = 1$ sinon
- 3 étapes répétées jusqu'à convergence

Origine : Nora-Chouteau en AFC (1974); Kiers (1997)

ACP itérative

- 1 initialisation $\ell = 0$: \mathbf{X}^0 (imputation par la moyenne)
- 2 étape ℓ :
 - (a) ACP sur le tableau complété $\Rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$;
 K dimensions conservés
 - (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell T}$;
nouveau tableau : $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$; avec $w_{ij} = 0$ si x_{ij} est manquant et $w_{ij} = 1$ sinon
- 3 étapes répétées jusqu'à convergence

\Rightarrow Le nombre de dimensions K doit être choisi *a priori*

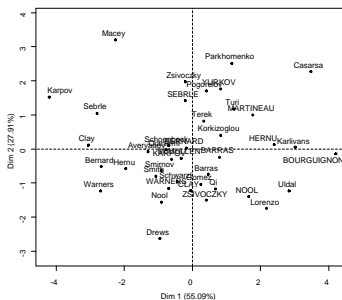
Origine : Nora-Chouteau en AFC (1974); Kiers (1997)

Propriétés

- Résultats de l'ACP obtenus à partir des données observées uniquement : graphe des individus et graphe des variables
⇒ On "saute" les données manquantes, l'ACP itérative minimise $\|\mathbf{W} * (\mathbf{X} - \mathbf{FU}')\|^2$
- Imputation :
 - prend en compte les ressemblances entre individus et les liaisons entre variables
 - le tableau imputé peut être utilisé (avec précaution) pour réaliser d'autres analyses
- Problème de surajustement

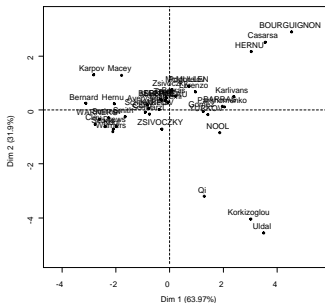
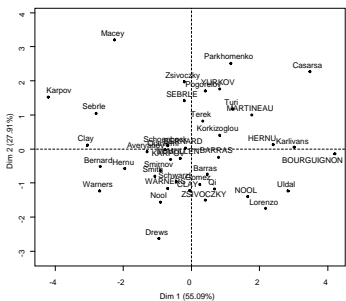
Surajustement

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



Surajustement

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



Surajustement

⇒ Bon ajustement et mauvaise prédiction

- Trop de paramètres sont estimés par rapport au nombre de données observées : le nombre de dimension S et le nombre de données manquantes sont grands
- Faibles liaisons entre variables

① Diminuer le nombre S

② Early stopping

③ Regularisation ⇒ ACP itérative régularisée

ACP itérative régularisée

⇒ Initialisation, étape d'estimation par ACP- étape d'imputation.

L'étape d'imputation $\hat{\mathbf{X}} = \mathbf{FU}'$:

$$\hat{x}_{ij}^{\ell} = \sum_{k=1}^K \hat{f}_{ik}^{\ell} \hat{u}_{jk}^{\ell} = \sum_{k=1}^K \frac{\hat{f}_{ik}^{\ell}}{\|\hat{\mathbf{f}}_k^{\ell}\|} (\sqrt{\lambda_k}) \hat{u}_{jk}^{\ell},$$

est remplacée par une étape d'imputation "shrinkée" :

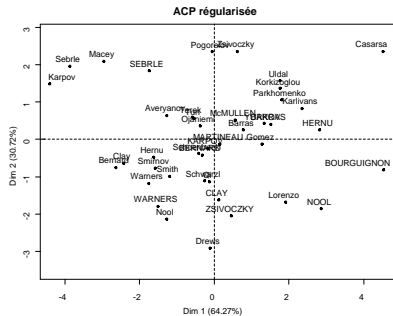
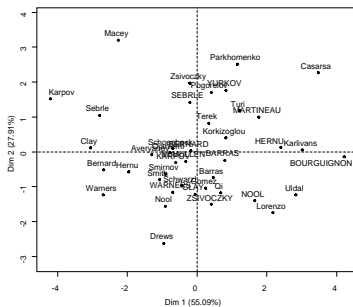
$$\hat{x}_{ij}^{\ell} = \sum_{k=1}^K \frac{\hat{f}_{ik}^{\ell}}{\|\hat{\mathbf{f}}_k^{\ell}\|} \left(\sqrt{\lambda_k} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_k}} \right) \hat{u}_{jk}^{\ell}$$

avec $\hat{\sigma}^2 = \frac{1}{p-K} \sum_{k=K+1}^p \lambda_k$.

⇒ Supprimer le bruit pour éviter l'instabilité dans les prédictions

Surajustement

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



$$\|(1 - \mathbf{W}) * (\mathbf{X} - \hat{\mathbf{X}})\| = 0.67$$

Mise en œuvre logiciel

Utilisation de missMDA et de
FactoMineR

Importer les données

```
library(missMDA)
orange=read.table("orange.csv", header=T, sep=";")
```

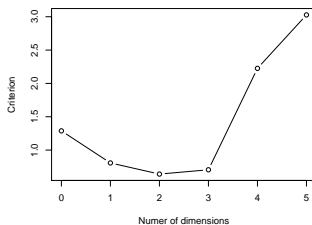
```
      Sweet Acid ... Bitter Pulp Typicity
1      NA  NA ...  2.83  NA      5.21
2  5.46 4.13 ...  3.54 4.62      4.46
3      NA 4.29 ...  3.17 6.25      5.17
..
12 4.88 5.29 ...  4.17 1.50      3.50
```

Etape 1 : Estimer le nombre de dimensions

⇒ estimation de K par validation croisée; approximation de type GCV

```
> 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 |



Etape 2: Imputation des données manquantes

```
> 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 |

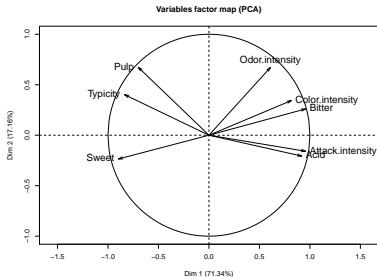
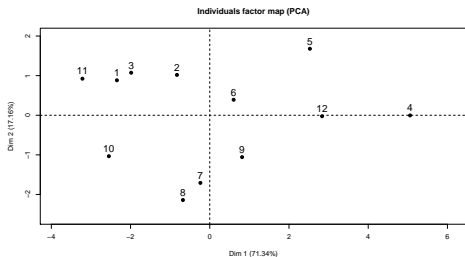
```
> res.comp$completeObs
```

| 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 |

Etape 3: ACP sur le tableau de données complété

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

```
⇒ library FactoMineR
```



```
> res.pca$ind$coord # (composantes principales)
```

```
> res.pca$var$coord
```

Tutoriaux

- Film pour installer R et installer des packages
- Film pour faire une ACP
- Film pour faire une ACP avec données manquantes

⇒ Site internet du laboratoire de mathématiques appliquées et de François Husson

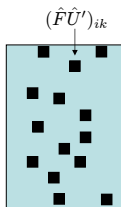
Quel crédit accorder aux résultats
obtenus?

Plan

- 1 Rappel d'ACP
- 2 Données manquantes en ACP
- 3 Zone de confiance**
- 4 Conclusion

Imputation multiple en ACP

⇒ ACP itérative : une méthode d'imputation simple

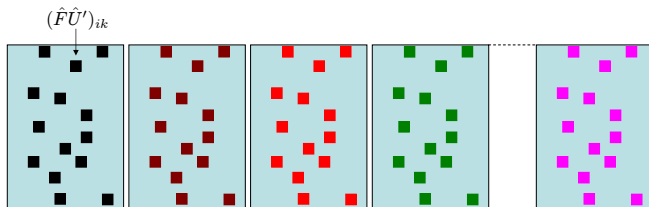


⇒ Une valeur unique ne peut pas refléter la variabilité de prédiction

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

Imputation multiple en ACP

⇒ ACP itérative : une méthode d'imputation simple



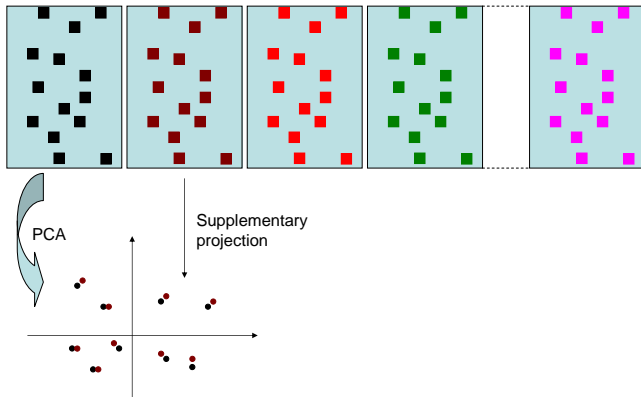
⇒ Une valeur unique ne peut pas refléter la variabilité de prédiction

⇒ Imputation multiple : générer plusieurs valeurs plausibles pour chaque valeur manquantes

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

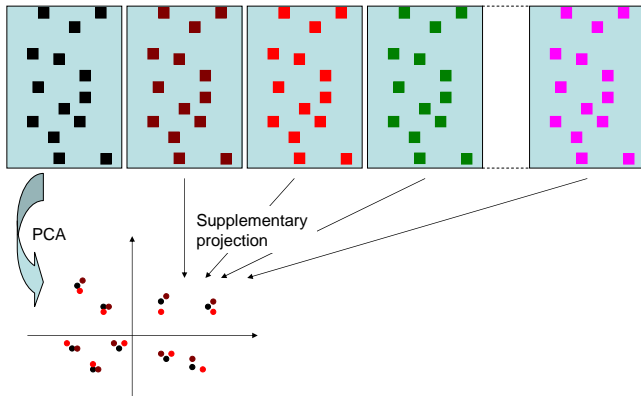
Projection en supplémentaire

⇒ Position des individus (et des variables) avec d'autres prédictions



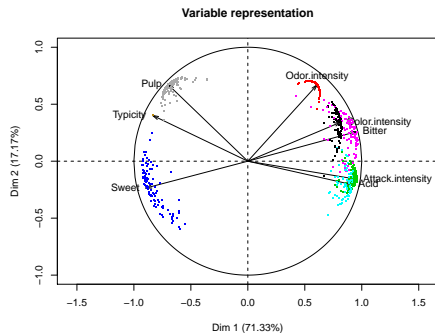
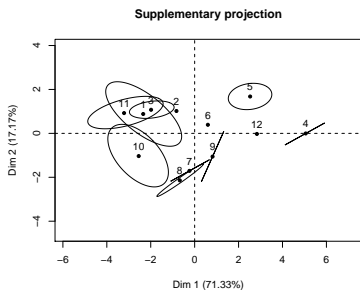
Projection en supplémentaire

⇒ Position des individus (et des variables) avec d'autres prédictions



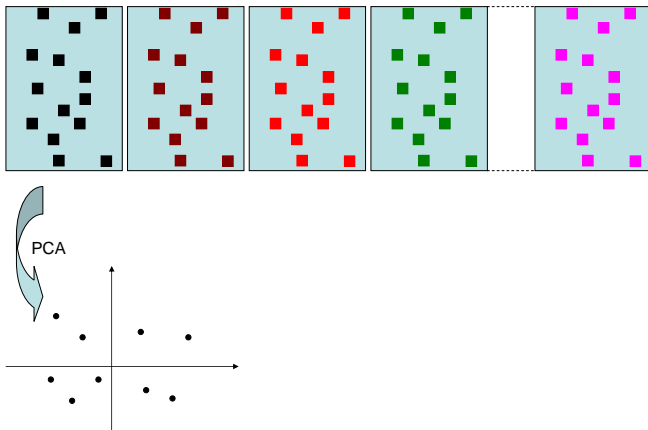
Projection en supplémentaire

```
> plot(mi)
```



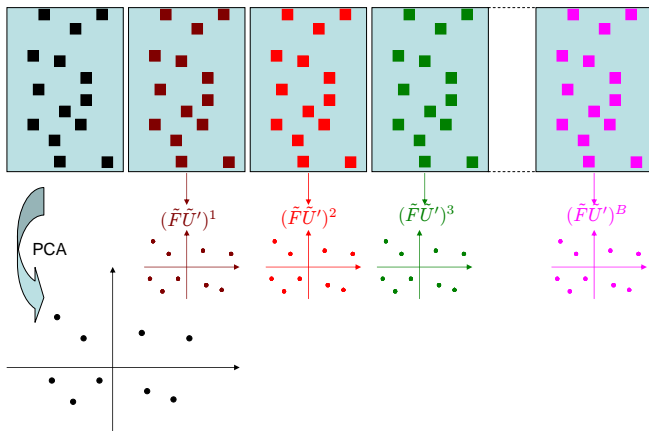
Variabilité inter-imputation

⇒ Influence des différentes prédictions sur les axes et composantes ACP sur chaque tableau)



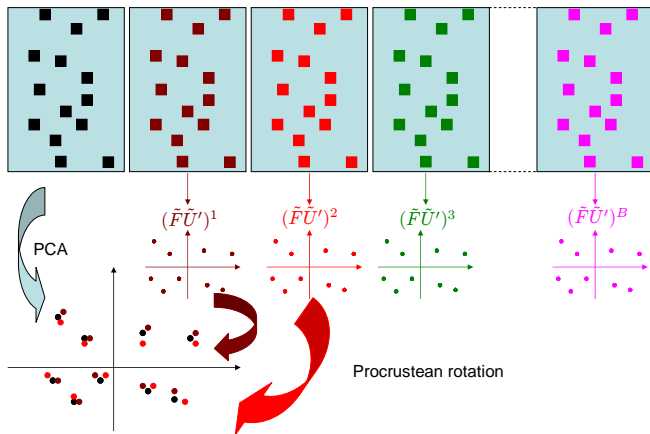
Variabilité inter-imputation

⇒ Influence des différentes prédictions sur les axes et composantes ACP sur chaque tableau)



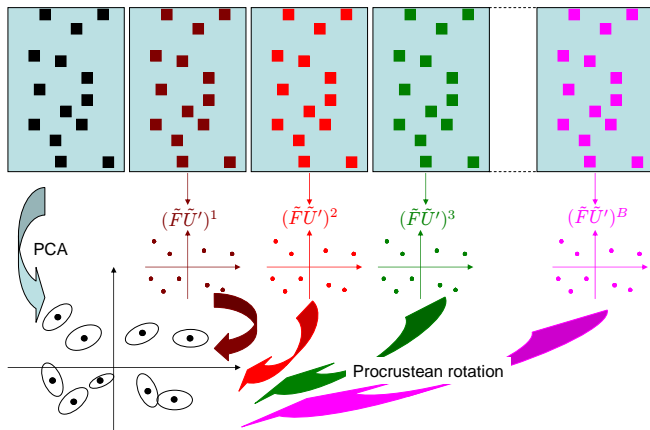
Variabilité inter-imputation

⇒ Influence des différentes prédictions sur les axes et composantes ACP sur chaque tableau)

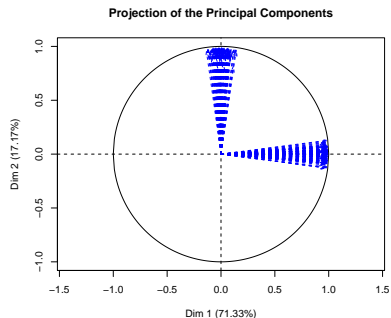
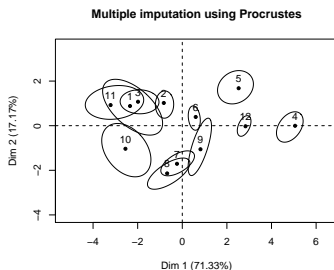


Variabilité inter-imputation

⇒ Influence des différentes prédictions sur les axes et composantes ACP sur chaque tableau)



Variabilité inter-imputation



⇒ Voir le tutorial sur le site internet

Plan

- 1 Rappel d'ACP
- 2 Données manquantes en ACP
- 3 Zone de confiance
- 4 Conclusion**

Conclusion

- ACP avec données manquantes
- Imputation multiple en ACP
 - ⇒ Package R missMDA

⇒ Extension à l'ACM (variables qualitative)

⇒ Extension à l'AFM (groupe de variables quantitatives/qualitatives) et aux données mixtes

Conclusion

- Méthode d'imputation simple pour des variables quantitatives et qualitatives
- Imputation multiple pour des variables quantitatives : une alternative aux packages R mice ou Amelia?

Références

- Josse, J., Husson, F. & Pagès, J. (2011). Multiple imputation in PCA. *Advances in data analysis and classification*. 5 (3) pp. 231-246.
- Josse, J., Husson, F. & Pagès, J. (2009). Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la SFdS*. 150 (2), pp. 28-51.
- Schafer J. L & Graham J. W.(2002) Missing data: our view of the state of the art. *Psychol Methods*. 7(2) pp. 147-77.