

Prise en compte des données manquantes en analyse exploratoire des données

Julie Josse

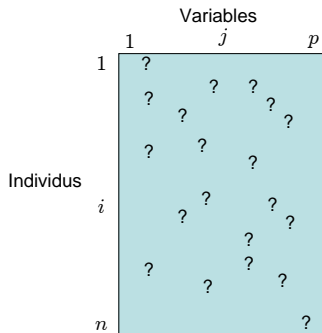
Laboratoire de mathématiques appliquées, Agrocampus Ouest

Rennes, 22 octobre 2010

Plan

- 1 Introduction
- 2 Estimation ponctuelle
 - cas complet
 - cas incomplet
- 3 Zone de confiance
 - cas complet
 - cas incomplet : imputation multiple
- 4 Choix de la dimension
- 5 Conclusion

Contexte



⇒ Etude et mise en œuvre des méthodes factorielles en présence de données manquantes : **ACP** (variables quantitatives), **ACM** (variables qualitatives)

⇒ Objectif exploratoire

Problématique des données manquantes

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

⇒ Méthode très utilisée : suppression

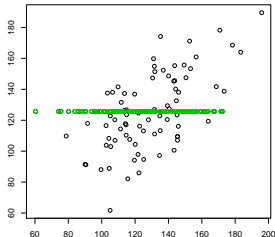
Traitement des données manquantes dépend du :

- dispositif des données manquantes : non structuré
- mécanisme conduisant à l'apparition de données manquantes (Rubin, 1976) : hypothèse Missing (Completely) At Random

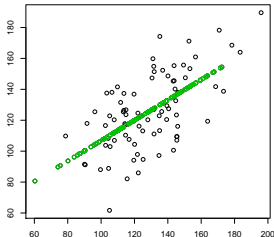
Méthodes d'imputation simple

⇒ Autres méthodes très utilisées : méthodes d'imputation

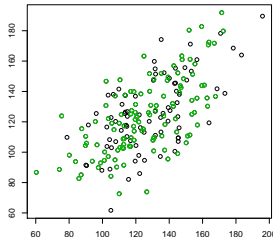
Imputation par la moyenne



Imputation par régression



Imputation par régression aléatoire



Méthodes d'imputation simple

⇒ Autres méthodes très utilisées : méthodes d'imputation



Une valeur unique ne reflète pas l'incertitude sur la prévision

⇒ Sous-estimation de la variance des estimateurs calculée à partir du tableau imputé

Méthodes préconisées

⇒ Imputation multiple (Rubin, 1987) :

- générer plusieurs imputations plausibles
- réaliser l'analyse sur chaque tableau de données complété
- combiner les résultats

⇒ Maximum de vraisemblance : algorithme EM (Dempster *et al.*, 1977) pour l'estimation ponctuelle

⇒ Objectif : estimer les paramètres ponctuellement et par intervalle en présence de données manquantes avec des variances qui prennent en compte la variabilité supplémentaire due aux données manquantes

Plan

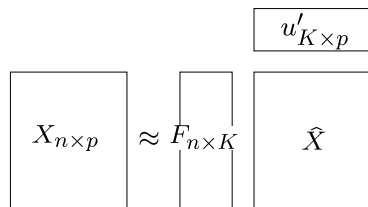
- 1 Introduction
- 2 Estimation ponctuelle des axes et composantes
- 3 Zone de confiance
- 4 Choix de la dimension
- 5 Conclusion et perspectives

Plan

- 1 Introduction
- 2 Estimation ponctuelle
 - cas complet
 - cas incomplet
- 3 Zone de confiance
 - cas complet
 - cas incomplet : imputation multiple
- 4 Choix de la dimension
- 5 Conclusion

Minimiser l'erreur de reconstitution

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



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

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

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

⇒ Algorithmes itératifs

ACP via NIPALS (Non linear Iterative PARTial Least Squares)

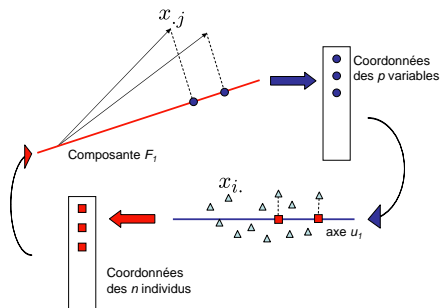
- Meilleure approximation de rang 1 (Wold, 1966, 1969)

$$C_1 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - F_{i1} u_{j1})^2$$

⇒ 2 régressions simples

$$\hat{u}_{j1} = \frac{\sum_i (x_{ij} \times F_{i1})}{\sum_i F_{i1}^2}$$

$$\hat{F}_{i1} = \frac{\sum_j (x_{ij} \times u_{j1})}{\sum_j u_{j1}^2}$$



- Déflation : une fois (\hat{F}_1, \hat{u}_1) trouvé, on cherche (\hat{F}_2, \hat{u}_2) premier axe et première composante de $\hat{\epsilon}_1 = X - \hat{F}_1 \hat{u}'_1$

ACP via la recherche directe du sous-espace

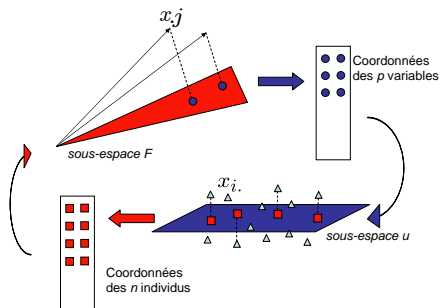
- Recherche directe du sous-espace de dimension K avec $K > 1$

$$C = \|X_{n \times p} - F_{n \times K} u'_{K \times p}\|^2$$

⇒ 2 régressions multiples

$$\hat{u} = X'F(F'F)^{-1}$$

$$\hat{F} = Xu(u'u)^{-1}$$



Moindres carrés pondérés

$$C = \|W * (X - Fu')\|^2 = \sum_{i=1}^n \sum_{j=1}^p (w_{ij}x_{ij} - \sum_{k=1}^K F_{ik}w_{ij}u_{jk})^2$$

avec W matrice de poids, $w_{ij} = 0$ si x_{ij} manquant, $w_{ij} = 1$ sinon

⇒ Mêmes algorithmes mais on “saute” les données manquantes

- NIPALS : 2 régressions simples pondérées (Christofferson, 1969) $\hat{u}_{j1} = \frac{\sum_i (w_{ij}x_{ij}F_{i1})}{\sum_i w_{ij}F_{i1}^2}$; $\hat{F}_{i1} = \frac{\sum_j (w_{ij}x_{ij}u_{j1})}{\sum_j w_{ij}u_{j1}^2}$
- Recherche directe du sous-espace : 2 régressions multiples pondérées (Gabriel & Zamir, 1979)

Moindres carrés pondérés

$$C = \|W * (X - Fu')\|^2 = \sum_{i=1}^n \sum_{j=1}^p (w_{ij}x_{ij} - \sum_{k=1}^K F_{ik}w_{ij}u_{jk})^2$$

avec W matrice de poids, $w_{ij} = 0$ si x_{ij} manquant, $w_{ij} = 1$ sinon

⇒ Mêmes algorithmes mais on “saute” les données manquantes

- NIPALS : 2 régressions simples pondérées (Christofferson, 1969) $\hat{u}_{j1} = \frac{\sum_i (w_{ij}x_{ij}F_{i1})}{\sum_i w_{ij}F_{i1}^2}$; $\hat{F}_{i1} = \frac{\sum_j (w_{ij}x_{ij}u_{j1})}{\sum_j w_{ij}u_{j1}^2}$ ⇒ pas optimal pour C
- Recherche directe du sous-espace : 2 régressions multiples pondérées (Gabriel & Zamir, 1979)

ACP itérative

⇒ Nora-Chouteau en AFC (1974) : estimation/imputation

① initialisation $\ell = 0 : X^0$

② itération ℓ :

(a) (F^ℓ, u^ℓ) minimisent $\|X^{\ell-1} - Fu'\|^2$; K dimensions retenues

(b) $\hat{X}^\ell = \hat{F}^\ell \hat{u}^{\ell'}$ ⇒ $X^\ell = W * X + (1 - W) * \hat{X}^\ell$

③ les étapes (a) et (b) sont répétées jusqu'à convergence

⇒ Kiers (1997) : ACP itérative minimise $\|W * (X - Fu')\|^2$

ACP itérative

⇒ Nora-Chouteau en AFC (1974) : estimation/imputation

① initialisation $\ell = 0$: X^0

② itération ℓ :

(a) (F^ℓ, u^ℓ) minimisent $\|X^{\ell-1} - Fu'\|^2$; K dimensions retenues

(a') (F^ℓ, u^ℓ) diminuent $\|X^{\ell-1} - Fu'\|^2$

$$\hat{u}^\ell = X^{\ell-1'} \hat{F}^{\ell-1} (\hat{F}^{\ell-1'} \hat{F}^{\ell-1})^{-1}$$

$$\hat{F}^\ell = X^{\ell-1} \hat{u}^\ell (\hat{u}^{\ell'} \hat{u}^\ell)^{-1}$$

(b) $\hat{X}^\ell = \hat{F}^\ell \hat{u}^{\ell'}$ ⇒ $X^\ell = W * X + (1 - W) * \hat{X}^\ell$

③ les étapes (a) et (b) sont répétées jusqu'à convergence

⇒ Kiers (1997) : ACP itérative minimise $\|W * (X - Fu')\|^2$

ACP itérative = ACP-EM

Modèle (Causinus, 1986) : $x_{ij} = \sum_{k=1}^K F_{ik} u_{jk} + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Vraisemblance : $L_c(F, u, \sigma^2) = -\frac{np}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|X - Fu'\|^2$

- Etape E : $\mathbb{E}(x_{ij} | X_{obs}, \hat{F}^\ell, \hat{u}^\ell, \hat{\sigma}^\ell) \Rightarrow$ **imputation** par $\hat{F}\hat{u}'$
 - Etape M : maximise l'espérance de $L_c \Rightarrow$ **estimation** des paramètres par l'ACP du tableau de données complété
 - Etape M' : augmente l'espérance de $L_c \Rightarrow$ moindres carrés alternés
- \Rightarrow ACP GEM (Generalized Expectation Maximization)

Propriétés

⇒ Deux algorithmes = deux points de vues

- 1 ACP-itérative impute
- 2 Recherche directe "saute" les données manquantes

Propriétés

⇒ Deux algorithmes = deux points de vues

- 1 ACP-itérative impute ⇒ “saute” les données manquantes (données imputées n’ont pas d’influence)
- 2 Recherche directe “saute” les données manquantes ⇒ impute implicitement

Propriétés

⇒ Deux algorithmes = deux points de vues

- ① ACP-itérative **impute** ⇒ “saute” les données manquantes (données imputées n’ont pas d’influence)
 - ② Recherche directe “saute” les données manquantes ⇒ **impute** implicitement
- Décentrage : recentrage
 - Réduction de la variabilité (imputation par $\hat{F}\hat{u}'$)

Propriétés

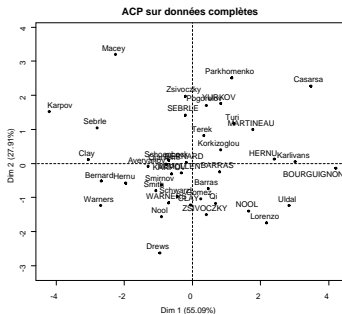
⇒ Deux algorithmes = deux points de vues

- ① ACP-itérative **impute** ⇒ “saute” les données manquantes (données imputées n’ont pas d’influence)
 - ② Recherche directe “saute” les données manquantes ⇒ **impute** implicitement
- Décentrage : recentrage
 - Réduction de la variabilité (imputation par $\hat{F}\hat{u}'$)
 - Solutions non emboîtées : choix du nombre d’axes (considéré pour l’instant connu)
 - Surajustement



Surajustement

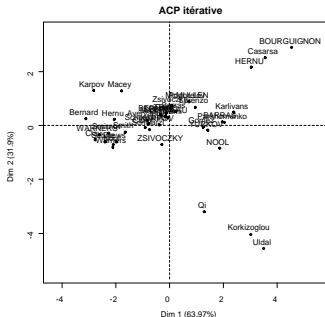
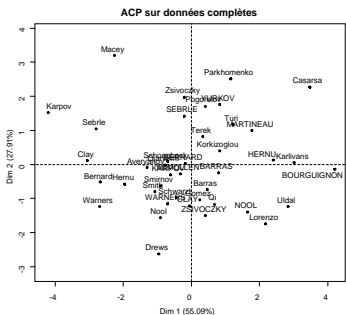
$$X_{41 \times 6} = F_{41 \times 2} u'_{2 \times 6} + \mathcal{N}(0, 0.5);$$





Surajustement

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

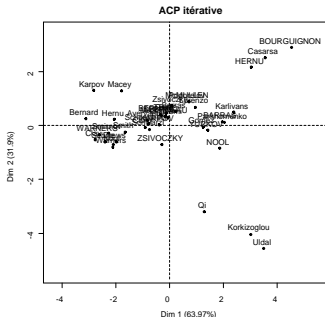
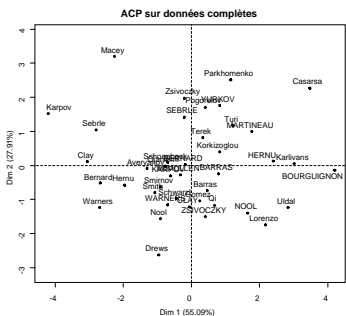


$$\text{ACP EM} : \|W * (X - \hat{X})\| = 0.48; \|(1 - W) * (X - \hat{X})\| = 5.58$$



Surajustement

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



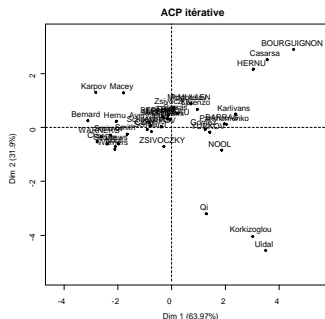
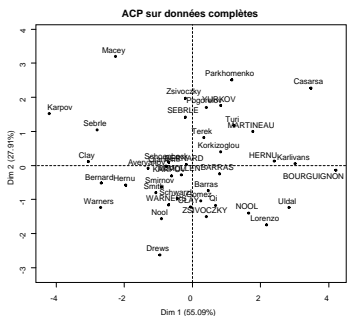
$$\text{ACP EM} : \|W * (X - \hat{X})\| = 0.48; \|(1 - W) * (X - \hat{X})\| = 5.58$$

- Diminuer K



Surajustement

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



$$\text{ACP EM} : \|W * (X - \hat{X})\| = 0.48; \|(1 - W) * (X - \hat{X})\| = 5.58$$

- Diminuer K
- Régulariser les deux régressions \Rightarrow ACP Probabiliste

ACP Probabiliste (*Tipping & Bishop, 1999; Roweis, 1998*)

⇒ Modèle d'analyse en facteurs communs et spécifiques particulier

$$x_{j.} = \Gamma_{p \times K} z_{j.} + \varepsilon_{j.}, \quad z_{j.} \sim \mathcal{N}(0, I_K), \quad \varepsilon_{j.} \sim \mathcal{N}(0, \sigma^2 I_p)$$

- Distribution des observations :

$$x_{j.} \sim \mathcal{N}(0, \Sigma) \text{ avec } \Sigma_{p \times p} = \Gamma_{p \times K} \Gamma'_{K \times p} + \sigma^2 I_p$$

- Solution explicite :

- $\hat{\sigma}^2 = \frac{1}{p-K} \sum_{j=K+1}^p \lambda_j$
- $\hat{\Gamma} = u_K (\Lambda_K - \sigma^2 I_K)^{1/2}$

ACP Probabiliste via l'algorithme EM

$$z_i | x_i \sim \mathcal{N}((\Gamma' \Gamma + \sigma^2 I)^{-1} \Gamma' x_i, V)$$

- Etape E : Espérance conditionnelle

$$\hat{Z}' = (\hat{\Gamma}' \hat{\Gamma} + \hat{\sigma}^2 I)^{-1} \hat{\Gamma}' X'$$

- Etape M : Maximise $\mathbb{E}[L_c]$ par rapport à Γ et σ^2

$$\hat{\Gamma}' = (\hat{Z}' \hat{Z} + n \hat{V})^{-1} \hat{Z}' X$$

⇒ Régressions ridges

⇒ Vers un algorithme d'ACP-GEM régularisé :

- estimer Z et Γ
- imputer par $\hat{Z} \hat{\Gamma}'$

ACP itérative régularisée

1 initialisation $\ell = 0 : X^0$

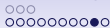
2 itération ℓ :

(a) (F^ℓ, u^ℓ) minimisent $\|X^{\ell-1} - Fu'\|^2$; K dimensions retenues

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

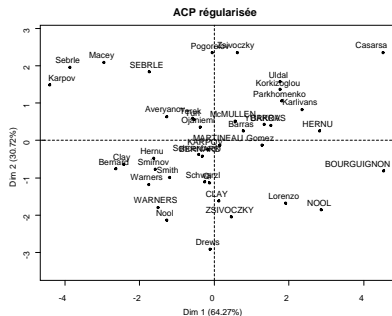
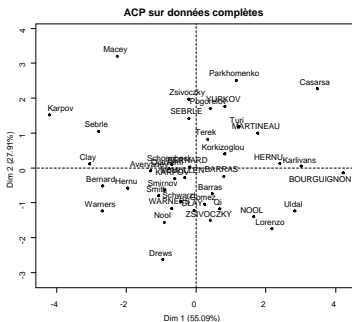
nouvelle imputation : $X^\ell = W * X + (1 - W) * \hat{X}^\ell$;

3 les étapes (a) et (b) sont répétées jusqu'à convergence



Surajustement

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

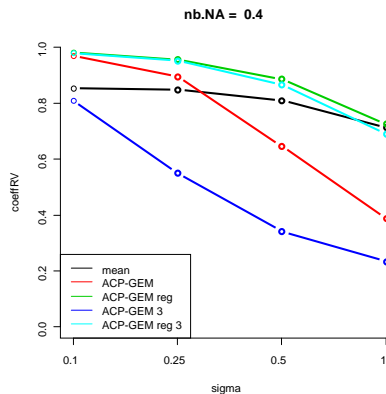
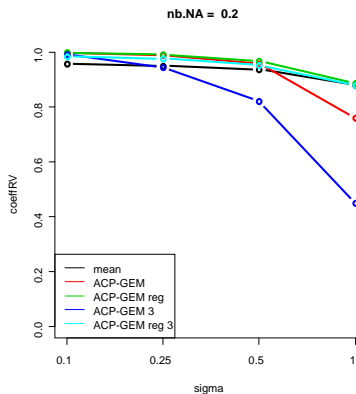


$$\|(1 - W) * (X - \hat{X})\| = 0.67$$



Simulations

- $X_{21 \times 10} = F_{21 \times 2} u'_{2 \times 10} + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma)$
- Coefficient *RV* entre configurations (complète / incomplète)



Plan

- 1 Introduction
- 2 Estimation ponctuelle
 - cas complet
 - cas incomplet
- 3 Zone de confiance
 - cas complet
 - cas incomplet : imputation multiple
- 4 Choix de la dimension
- 5 Conclusion

Stabilité en ACP

- Epreuves de validité en analyse exploratoire
- Rééchantillonnage des individus (Lebart, 1996) : bootstrap non-paramétrique
 - ⇒ fluctuations dues à l'échantillonnage
 - ⇒ bootstrap toutes les dimensions
 - ⇒ zones de confiance autour de la position des variables
- Zone de confiance quand l'ACP est réalisée sur une population d'individus ?

Modèle en ACP

$$x_{ij} = \text{structure} + \text{bruit}$$

- Modèle à effets aléatoires (structurel) : analyse en facteurs, ACP Probabiliste (ACPP)
 - ⇒ les individus sont interchangeables
 - ⇒ étude des liaisons entre variables
- Modèle à effets fixes (fonctionnel) : Caussinus (1986)
 - ⇒ les individus ont des espérances différentes
 - ⇒ étude des individus et des variables

$$x_{ij} = \sum_{k=1}^K F_{ik} u_{jk} + \varepsilon_{ij}, \text{ avec } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Bootstrap des résidus

- Rééchantillonnage des résidus : bootstrap semi-paramétrique
 - ⇒ fluctuations dues au bruit
 - ⇒ bootstrap sur les dernières dimensions : “le bruit va partout”
 - ⇒ zones de confiance autour de la position des individus et des variables

Bootstrap des résidus

- ① ACP sur $X \Rightarrow \hat{F}_{n \times K}$ et $\hat{u}_{p \times K}$ (K dimensions retenues);
 - ② Données reconstituées $\hat{X} = \hat{F}\hat{u}'$ et résidus $\hat{\varepsilon} = X - \hat{X}$;
 - ③ Procédure bootstrap : répéter B fois les étapes
 - a Bootstrapper les résidus : ε^b
→ tirer dans une $\mathcal{N}(0, \hat{\sigma}^2)$
 - b $X^b = \hat{F}\hat{u}' + \varepsilon^b$
 - c ACP sur X^b pour obtenir \hat{F}^b et \hat{u}^b
- ⇒ B couples $(\hat{F}^1, \hat{u}^1), \dots, (\hat{F}^B, \hat{u}^B)$



Bootstrap des résidus

- ① ACP sur $X \Rightarrow \hat{F}_{n \times K}$ et $\hat{u}_{p \times K}$ (K dimensions retenues);
- ② Données reconstituées $\hat{X} = \hat{F}\hat{u}'$ et résidus $\hat{\varepsilon} = X - \hat{X}$;
 \Rightarrow Choix de la dimension ?
- ③ Procédure bootstrap : répéter B fois les étapes
 - a Bootstrapper les résidus : ε^b
 \rightarrow tirer dans une $\mathcal{N}(0, \hat{\sigma}^2)$
 \Rightarrow Sous-estimation des résidus ?
 - b $X^b = \hat{F}\hat{u}' + \varepsilon^b$
 - c ACP sur X^b pour obtenir \hat{F}^b et \hat{u}^b \Rightarrow B couples $(\hat{F}^1, \hat{u}^1), \dots, (\hat{F}^B, \hat{u}^B)$

Bootstrap des résidus

- ① ACP sur $X \Rightarrow \hat{F}_{n \times K}$ et $\hat{u}_{p \times K}$ (K dimensions retenues);
- ② Données reconstituées $\hat{X} = \hat{F}\hat{u}'$ et résidus $\hat{\varepsilon} = X - \hat{X}$;
 \Rightarrow Choix de la dimension ?
- ③ Procédure bootstrap : répéter B fois les étapes
 - a Bootstrapper les résidus : ε^b
 \rightarrow tirer dans une $\mathcal{N}(0, \hat{\sigma}^2)$
 \Rightarrow Sous-estimation des résidus ?
 - b $X^b = \hat{F}\hat{u}' + \varepsilon^b$
 - c ACP sur X^b pour obtenir \hat{F}^b et \hat{u}^b \Rightarrow B couples $(\hat{F}^1, \hat{u}^1), \dots, (\hat{F}^B, \hat{u}^B)$
 \Rightarrow Visualisation ?



Incertitude supplémentaire due aux données manquantes

⇒ Source de variabilité supplémentaire

ACP itérative : imputation simple ⇒ bootstrap des résidus sur le tableau imputé sous-estimerait la variabilité des paramètres

⇒ Imputation multiple

- 1 Générer B tableaux de données imputés
- 2 Réaliser l'analyse sur chaque tableau imputé
- 3 Combiner les résultats : Variance totale \approx Variance intra imputation + Variance inter imputation

Incertitude supplémentaire due aux données manquantes

⇒ Source de variabilité supplémentaire

ACP itérative : imputation simple ⇒ bootstrap des résidus sur le tableau imputé sous-estimerait la variabilité des paramètres

⇒ Imputation multiple

- 1 Générer B tableaux de données imputés
- 2 Réaliser l'analyse sur chaque tableau imputé
- 3 Combiner les résultats : Variance totale \approx Variance intra imputation + Variance inter imputation

Idée pour générer B tableaux imputés

$$x_{ij} = \sum_{k=1}^K F_{ik} u_{jk} + \varepsilon_{ij}, \text{ avec } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

ACP itérative sur le tableau de données incomplet $\Rightarrow (\hat{F}, \hat{u})$

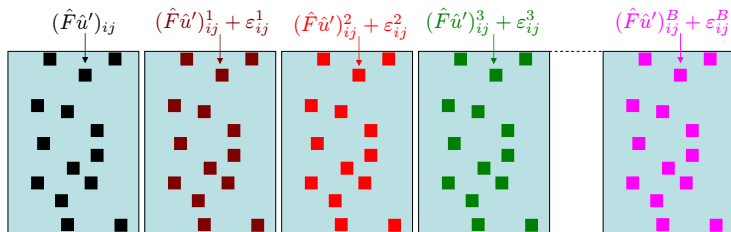
\Rightarrow Première idée pour générer différentes imputations:

Pour $b = 1, \dots, B$, imputer les valeurs manquantes x_{ij}^b en tirant dans la distribution prédictive $\mathcal{N}\left((\hat{F}\hat{u}')_{ij}, \hat{\sigma}^2\right)$

\Rightarrow Imputation "improper" (Rubin, 1987)

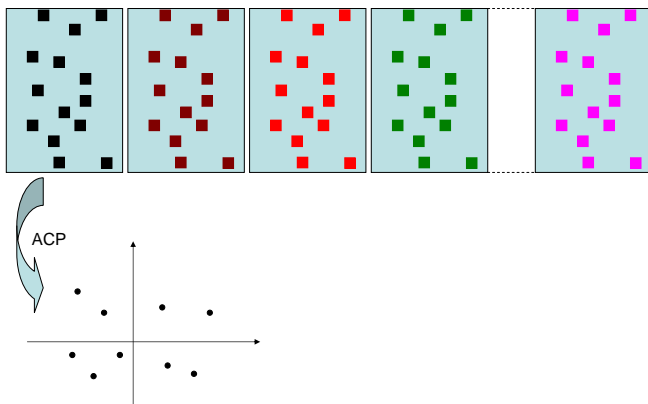
Imputation multiple “proper”

- 1 Variance d'estimation : obtenir B couples $(\hat{F}\hat{u}')^1, \dots, (\hat{F}\hat{u}')^B$
 \Rightarrow bootstrap des résidus
- 2 Bruit : pour $b = 1, \dots, B$, imputer les valeurs manquantes x_{ij}^b
 en tirant dans la distribution prédictive $\mathcal{N}\left((\hat{F}\hat{u}')_{ij}^b, \hat{\sigma}^2\right)$



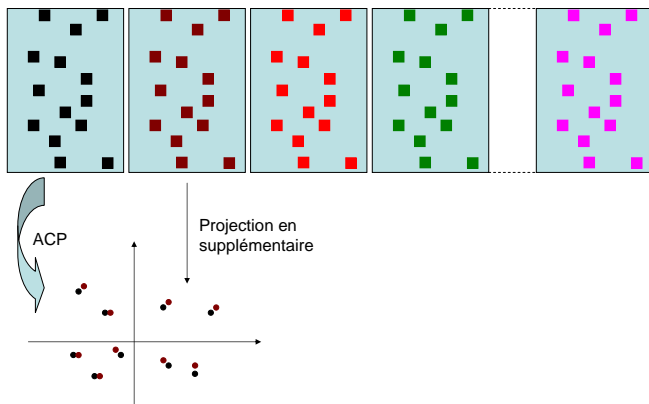
Projection en supplémentaire

⇒ Instabilité des individus (et des variables) due aux données manquantes (projection en supplémentaire)



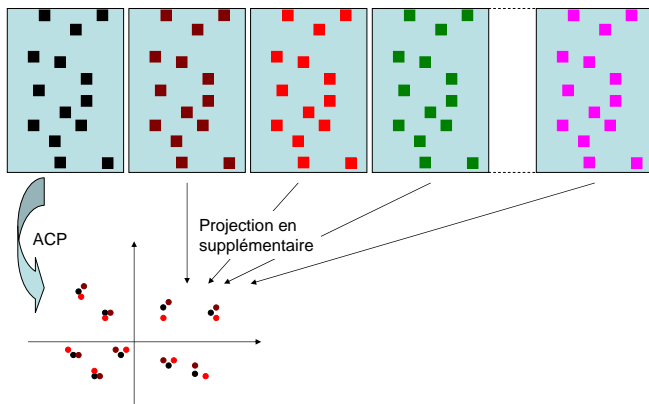
Projection en supplémentaire

⇒ Instabilité des individus (et des variables) due aux données manquantes (projection en supplémentaire)



Projection en supplémentaire

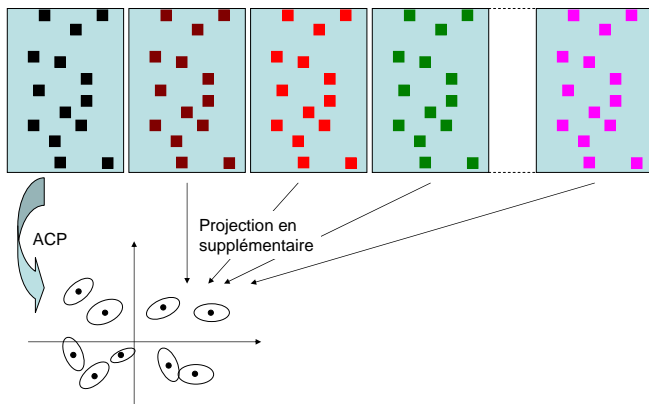
⇒ Instabilité des individus (et des variables) due aux données manquantes (projection en supplémentaire)

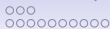




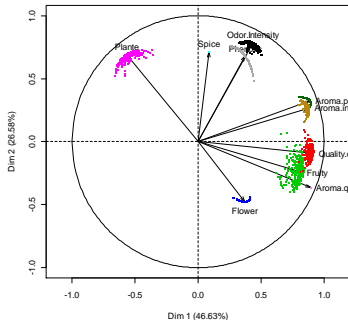
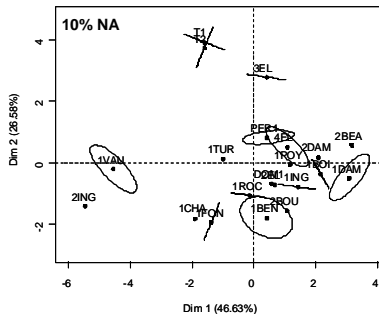
Projection en supplémentaire

⇒ Instabilité des individus (et des variables) due aux données manquantes (projection en supplémentaire)



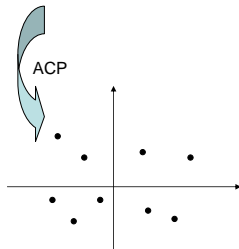
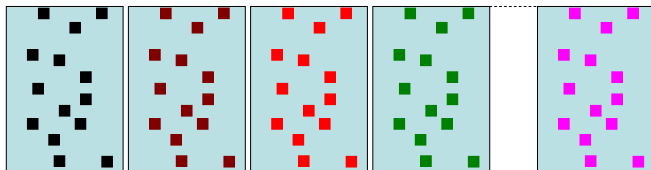


Projection en supplémentaire



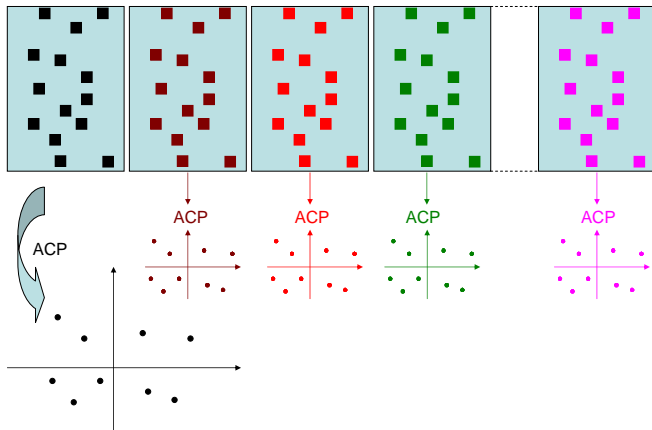
Variance inter-imputation

⇒ Impact des données manquantes sur la construction des axes et des composantes (ACP sur chaque tableau)



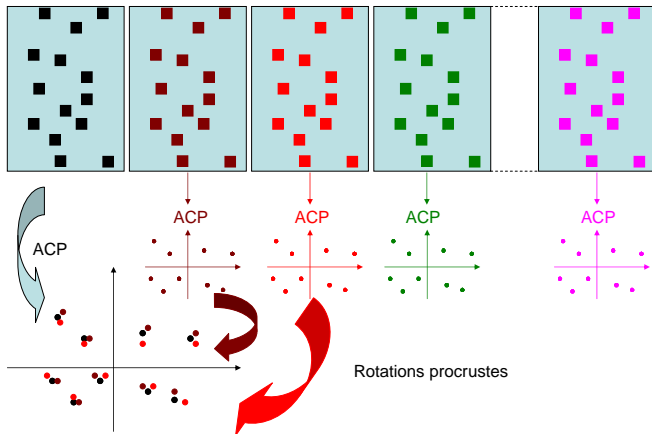
Variance inter-imputation

⇒ Impact des données manquantes sur la construction des axes et des composantes (ACP sur chaque tableau)



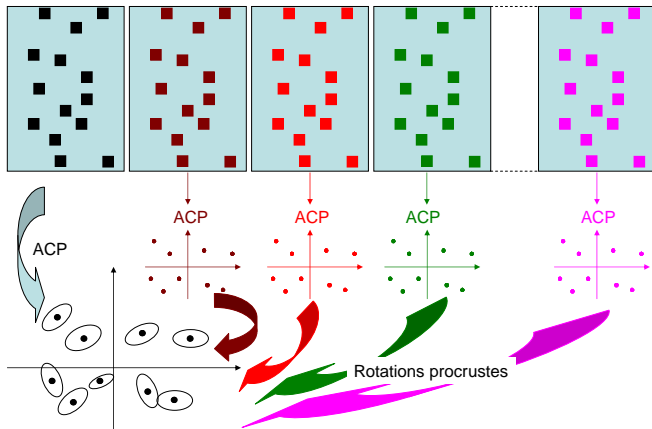
Variance inter-imputation

⇒ Impact des données manquantes sur la construction des axes et des composantes (ACP sur chaque tableau)



Variance inter-imputation

⇒ Impact des données manquantes sur la construction des axes et des composantes (ACP sur chaque tableau)



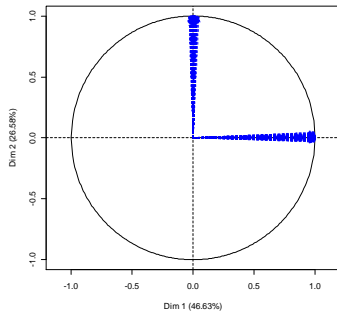
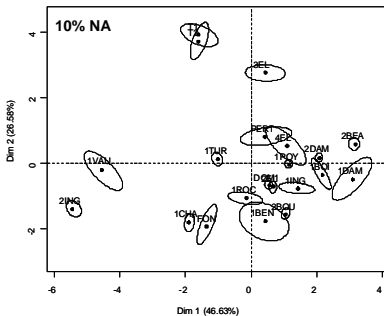
○○○

○○○○

○○○○○○○○○○

○○○○○○●

Variance inter-imputation



Plan

- 1 Introduction
- 2 Estimation ponctuelle
 - cas complet
 - cas incomplet
- 3 Zone de confiance
 - cas complet
 - cas incomplet : imputation multiple
- 4 Choix de la dimension
- 5 Conclusion

Méthode de validation croisée



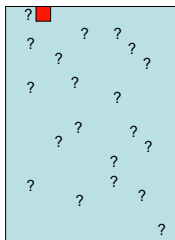
$$\text{MSEP}_K = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij}^{-ij})^2$$

- ⇒ Coûteux en temps de calcul
- ⇒ Approximations cas complet

⇒ Objectif : écrire l'ACP comme $\hat{X} = PX$

$$\hat{x}_{ij}^{-ij} - x_{ij} \simeq \frac{\hat{x}_{ij} - x_{ij}}{1 - P_{ij,ij}}$$

Méthode de validation croisée



$$\text{MSEP}_K = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij}^{-ij})^2$$

- ⇒ Coûteux en temps de calcul
- ⇒ Approximations cas complet

⇒ Objectif : écrire l'ACP comme $\hat{X} = PX$

$$\hat{x}_{ij}^{-ij} - x_{ij} \simeq \frac{\hat{x}_{ij} - x_{ij}}{1 - P_{ij,ij}}$$

Méthode de validation croisée



$$\text{MSEP}_K = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij}^{-ij})^2$$

- ⇒ Coûteux en temps de calcul
- ⇒ Approximations cas complet

⇒ Objectif : écrire l'ACP comme $\hat{X} = PX$

$$\hat{x}_{ij}^{-ij} - x_{ij} \simeq \frac{\hat{x}_{ij} - x_{ij}}{1 - P_{ij,ij}}$$

Projections en ACP

$$\mathcal{C} = \|X_{n \times p} - F_{n \times K} U'_{K \times p}\|^2$$

⇒ 2 matrices de projections

$$\begin{cases} \hat{u}' = (\hat{F}'\hat{F})^{-1}\hat{F}'X & \Rightarrow P_F = \hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}' \\ \hat{F} = X\hat{u}(\hat{u}'\hat{u})^{-1} & \Rightarrow P_U = \hat{u}(\hat{u}'\hat{u})^{-1}\hat{u}' \end{cases}$$

⇒ Forme bilinéaire

$$\text{Matrice ajustée} \quad \hat{X} = \hat{F}\hat{u}' \quad \Rightarrow \hat{X} = P_F X = X P_U$$

$$\text{Matrice des résidus} \quad \hat{\varepsilon} = X - \hat{X} \quad \Rightarrow \hat{\varepsilon} = (\mathbb{I}_n - P_F)X(\mathbb{I}_p - P_U)$$

⇒ On développe les résidus et on extrait \hat{X}

$$X - \hat{X} = X - (\mathbb{I}_n X P_U + P_F X \mathbb{I}_p - P_F X P_U)$$

Matrice de projection

$$\text{vec}(\hat{X}) = \text{vec}(\mathbb{I}_n X P_u) + \text{vec}(P_F X \mathbb{I}_p) - \text{vec}(P_F X P_u)$$

$$\text{vec}(\hat{X}) = (P'_u \otimes \mathbb{I}_n) \text{vec}(X) + (\mathbb{I}'_p \otimes P_F) \text{vec}(X) - (P'_u \otimes P_F) \text{vec}(X)$$

$$\text{vec}(\hat{X}) = P \text{vec}(X)$$

$$P_{np \times np} = (P'_u \otimes \mathbb{I}_n) + (\mathbb{I}'_p \otimes P_F) - (P'_u \otimes P_F)$$

- Nombre de paramètres équivalent :

$$\text{tr}(P) = K \times n + p \times K - K^2$$

- Degré de liberté des résidus : $\text{tr}(\mathbb{I}_{np} - P) = (n - K)(p - K)$

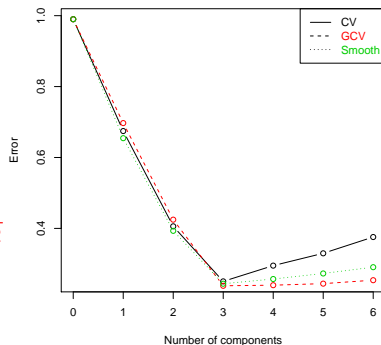
$$\hat{\sigma}_{cor}^2 = \frac{\|X - \hat{F}\hat{u}'\|^2}{np - (nK + pK - K^2)}$$

Approximations

$$\hat{x}_{ij}^{-ij} - x_{ij} \simeq \frac{\hat{x}_{ij} - x_{ij}}{1 - P_{ij,ij}}$$

$$S(K) = \frac{1}{np} \sum_{i,j} \left(\frac{\hat{x}_{ij} - x_{ij}}{1 - P_{ij,ij}} \right)^2$$

$$GCV(K) = \frac{1}{np} \times \frac{SCR(K)}{(1 - \text{tr}(P)/np)^2}$$



Simulations

Pourcentage de réponses correctes		
9 variables	$n = 30$	$n = 50$
RV	49.06	68.31
CV	59.08	70.58
Smooth	59.97	65.21
GCV	56.06	62.86
Parallel	49.72	58.05
Kaiser	46.58	50.80

Pourcentage de réponses correctes		
18 variables	$n = 60$	$n = 100$
RV	92.61	95.36
CV	91.06	96.39
Smooth	92.28	95.81
GCV	74.93	81.40
Parallel	76.11	82.72
Kaiser	37.17	48.58

Plan

- 1 Introduction
- 2 Estimation ponctuelle
 - cas complet
 - cas incomplet
- 3 Zone de confiance
 - cas complet
 - cas incomplet : imputation multiple
- 4 Choix de la dimension
- 5 Conclusion

Conclusion

- Estimation ponctuelle en ACP
- Imputation multiple en ACP

Conclusion

- Estimation ponctuelle en ACP → extension à l'ACM
- Imputation multiple en ACP

Conclusion

- Estimation ponctuelle en ACP → extension à l'ACM
- Imputation multiple en ACP
- Besoin : choix de la dimension et test sur le coefficient RV

⇒ Enrichissement du point de vue exploratoire par le point de vue probabiliste

- Création d'un package R missMDA et de fonctions dans le package FactoMineR

Perspectives

- Zone de confiance en ACM
- Nombre de dimensions dans le cas incomplet
- Prise en compte des données manquantes en AFM