



Handling missing values in PCA - Single and multiple imputation

Julie Josse

Applied mathematics department, Agrocampus Ouest, Rennes, France

Leiden, 20 June 2011



Applied mathematics department





Applied mathematics department

- Research
 - principal component methods - multi tables
 - inferential issues for high-throughput data - multiple testing
 - application fields: sensory and genomic data
- Teaching
 - Bachelor's degree: linear model, principal component methods, experimental design
 - Master's degree: sensometric, multi-tables, genomic data
 - Books: Exploratory Multivariate Analysis by Example using R, R for Statistics (Chapman & Hall)
- Other activities
 - R Packages: FactoMineR, SensoMineR, FAMT, missMDA
 - Conference: useR!2009, CARME 2011, Sensometrics 2012



Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas
 - Complete case
 - Incomplete case: multiple imputation
- 4 Choosing the number of dimensions
- 5 Conclusion



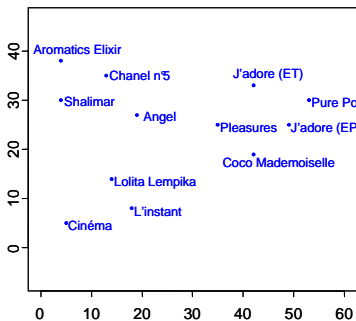
Motivations

- Sensory evaluation:



99 judges, 12 perfumes

- Direct collection of sensory distances: Napping (*Pagès 2003*)



	X	Y
Angel	19	27
Chanel	13	35
Coco	42	19
...
Shalimar	4	30

Sensory problem

	1	K_1			1	K_J
1						
i						
I						

Figure: Missing values pattern.

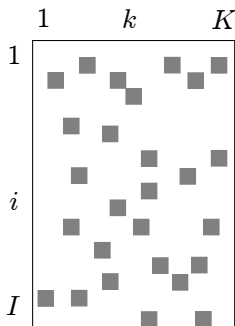
⇒ Perform a Multiple Factor Analysis (MFA, Escofier & Pagès, 1982)

- Problem: each panelist can't evaluate more than a certain number of products → saturation
- Planned missing products per panelist (BIB) → a special pattern of missing values

What we'd like to do... what we have done

Missing values in multi-tables (MFA)

... missing values in one table (PCA)



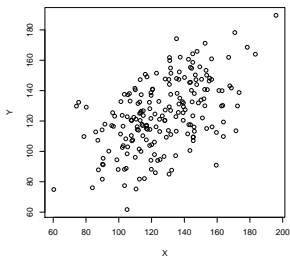


Missing values

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

⇒ Single imputation methods

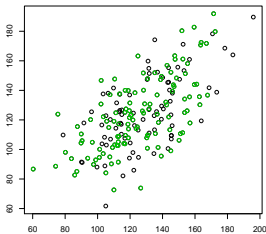
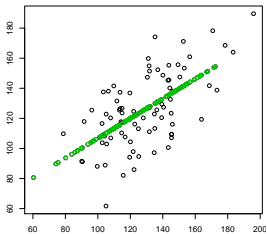
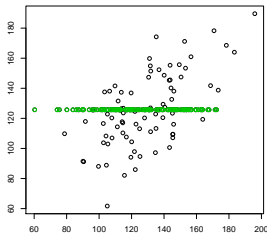
⇒ Shaefer (2002); a sample (x_i, y_i) , $i = (1, \dots, 50)$ drawn from a normal distribution ($\mu_x = \mu_y = 125$, $\sigma_x = \sigma_y = 25$, $\rho = 0.6$)



⇒ MCAR (Rubin, 1976): 73% NA in Y completely at random

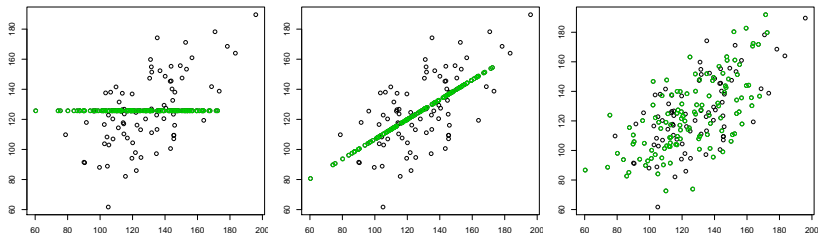


Single imputation methods





Single imputation methods



⇒ A unique value cannot reflect the uncertainty of the prediction of the missing value

⇒ Standard errors of the parameters calculated from the imputed data set are underestimated ($\hat{\mu}_y = 125.02$, $\hat{\sigma}_y = 24.58$, $\hat{\rho} = 0.59$; coverage for μ_y is 70.8%)



Recommended methods

⇒ Multiple imputation (Rubin, 1987):

- generating plausible values for each missing values
- performing the analysis on each imputed data set
- combining the results

⇒ Maximum likelihood: EM algorithm (Dempster *et al.*, 1977) to obtain point estimates + SEM to obtain estimation of their variability

⇒ Common aim: provide the best estimation of the parameters and of their variability (taken into account the variability due to missing values)



Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas
- 4 Number of dimensions
- 5 Conclusions and perspectives



Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas
 - Complete case
 - Incomplete case: multiple imputation
- 4 Choosing the number of dimensions
- 5 Conclusion



Weighted least squares

⇒ Minimization of:

$$\mathcal{C} = \|\mathbf{X}_{I \times K} - \mathbf{M}_{I \times K} - \mathbf{F}_{I \times S} \mathbf{U}_{S \times K}^t\|^2$$

⇒ With missing values:

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

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

⇒ Many algorithms: criss-cross multiple regression (Gabriel & Zamir, 1979); iterative PCA (Kiers, 1997)



iterative PCA

- 1 initialization $\ell = 0$: \mathbf{X}^0 ; $\hat{\mathbf{M}}^0$ is computed;
- 2 step ℓ :
 - (a) $(\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell) = \operatorname{argmin}_{(\mathbf{F}, \mathbf{U})} \|\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1} - \mathbf{F}\mathbf{U}'\|^2$; S dim are kept
 - (b) $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell'} + \hat{\mathbf{M}}^{\ell-1}$. The new imputed data set is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
 - (c) $\hat{\mathbf{M}}^\ell$ is computed on \mathbf{X}^ℓ .
- 3 steps (a), (b) and (c) are repeated until convergence.

⇒ Remark: the mean (the standard deviation!) is updated



iterative PCA

- 1 initialization $\ell = 0$: \mathbf{X}^0 ; $\hat{\mathbf{M}}^0$ is computed;
- 2 step ℓ :
 - (a) $(\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell) = \operatorname{argmin}_{(\mathbf{F}, \mathbf{U})} \|\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1} - \mathbf{F}\mathbf{U}'\|^2$; S dim are kept
 - (b) $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell'} + \hat{\mathbf{M}}^{\ell-1}$. The new imputed data set is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
 - (c) $\hat{\mathbf{M}}^\ell$ is computed on \mathbf{X}^ℓ .
- 3 steps (a), (b) and (c) are repeated until convergence.

⇒ Remark: the mean (the standard deviation!) is updated



iterative PCA

- 1 initialization $\ell = 0$: \mathbf{X}^0 ; $\hat{\mathbf{M}}^0$ is computed;
- 2 step ℓ :
 - (a) $(\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell) = \operatorname{argmin}_{(\mathbf{F}, \mathbf{U})} \|\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1} - \mathbf{F}\mathbf{U}'\|^2$; S dim are kept
 - (a') $(\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$ reduce $\|\mathbf{X}^{\ell-1} - \mathbf{F}\mathbf{U}' - \hat{\mathbf{M}}^{\ell-1}\|^2$

$$\begin{aligned}\hat{\mathbf{U}}^\ell &= (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1})' \hat{\mathbf{F}}^{\ell-1} (\hat{\mathbf{F}}^{\ell-1}' \hat{\mathbf{F}}^{\ell-1})^{-1}, \\ \hat{\mathbf{F}}^\ell &= (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1}) \hat{\mathbf{U}}^\ell (\hat{\mathbf{U}}^{\ell'} \hat{\mathbf{U}}^\ell)^{-1}.\end{aligned}$$

- (b) $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell'} + \hat{\mathbf{M}}^{\ell-1}$. The new imputed data set is
 $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
 - (c) $\hat{\mathbf{M}}^\ell$ is computed on \mathbf{X}^ℓ .

- 3 steps (a), (b) and (c) are repeated until convergence.

⇒ Remark: the mean (the standard deviation!) is updated



iterative PCA = EM-PCA

$$\text{Model: } x_{ik} = m_k + \sum_{s=1}^S f_{is} u_{js} + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

$$\text{Likelihood: } L_c(\mathbf{F}, \mathbf{U}, \sigma^2) = -\frac{IK}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{X} - \mathbf{M} - \mathbf{F}\mathbf{U}^t\|^2$$

- Step E: expectation \Rightarrow **imputation** with $\hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell'} + \hat{\mathbf{M}}^{\ell-1}$
- Step M: maximization \Rightarrow **estimation** of the parameters: PCA on the completed data set
- Step M': increase \Rightarrow one step of ALS
 \Rightarrow GEM-PCA (generalized expectation maximization)



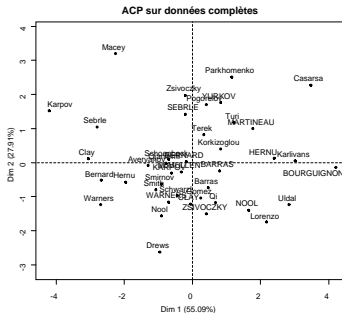
Properties

- Imputation (matrix completion problem, Netflix prize)
- Reduction of the variability (imputation by $\hat{\mathbf{F}}\hat{\mathbf{U}}'$)
- Solutions are not nested: number of dimensions?
- Overfitting



Overfitting

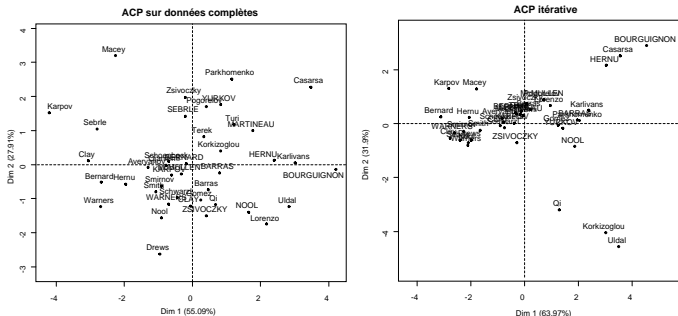
$$X_{41 \times 6} = F_{41 \times 2} U'_{2 \times 6} + \mathcal{N}(0, 0.5);$$





Overfitting

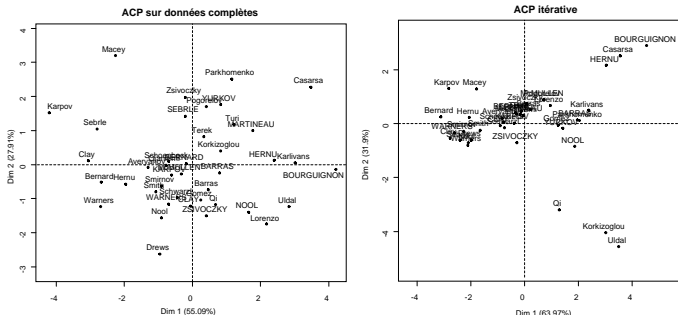
$$X_{41 \times 6} = \mathbf{F}_{41 \times 2} \mathbf{U}'_{2 \times 6} + \mathcal{N}(0, 0.5); \text{ 50\% of NA}$$





Overfitting

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



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

⇒ fitting error low; prediction error high

⇒ convergence problems of EM → overfitting problems



Overfitting

⇒ Fitting is good, prediction is bad

- Many parameters are estimated with respect to the number of observed values: the number of dimensions S and the number of missing values are important
- The correlations between variables are not strong

- 1 Reduce S
- 2 Early stopping
- 3 Shrinkage methods: ridge regressions

$$\begin{aligned}\hat{\mathbf{U}}^\ell &= (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{l-1})' \hat{\mathbf{F}}^{\ell-1} (\hat{\mathbf{F}}^{\ell-1}' \hat{\mathbf{F}}^{\ell-1})^{-1}, \\ \hat{\mathbf{F}}^\ell &= (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{l-1}) \hat{\mathbf{U}}^\ell (\hat{\mathbf{U}}^{\ell'} \hat{\mathbf{U}}^\ell)^{-1}.\end{aligned}$$

⇒ Ridge parameters: Probabilistic PCA



Probabilistic PCA (*Lawley, 1953, Tipping & Bishop, 1999; Roweis, 1998*)

⇒ A specific case of an exploratory factor analysis model

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

- Distribution for the observations:

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

- Explicit solution:

- $\hat{\sigma}^2 = \frac{1}{K-S} \sum_{s=S+1}^K \lambda_j$
- $\hat{\Gamma} = U_S (\Lambda_S - \sigma^2 I_S)^{1/2}$



Probabilistic PCA via the EM algorithm

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

- Step E: Conditional expectation

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

- Step M: Maximization of $\mathbb{E}[L_c]$ with respect to Γ and σ^2

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

⇒ Ridge regressions

⇒ A ridge GEM-PCA algorithm:

- Estimation of Z and Γ
- Imputation with $\hat{Z} \hat{\Gamma}'$



Regularized iterative PCA (Josse *et al.*, 2009)

1 initialization $\ell = 0$: $\hat{\mathbf{X}}^0$

2 iteration ℓ :

(a) $(\mathbf{F}^\ell, \mathbf{U}^\ell)$ minimize $\|\mathbf{X}^{\ell-1} - \mathbf{F}\mathbf{U}'\|^2$; S dimensions are kept

$$(b) \hat{x}_{ik}^\ell = \sum_{s=1}^S \frac{\hat{F}_{is}^\ell}{\|\hat{\mathbf{F}}_s^\ell\|} \left(\sqrt{\lambda_k^\ell} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s^\ell}} \right) \hat{u}_{ks}^\ell$$

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

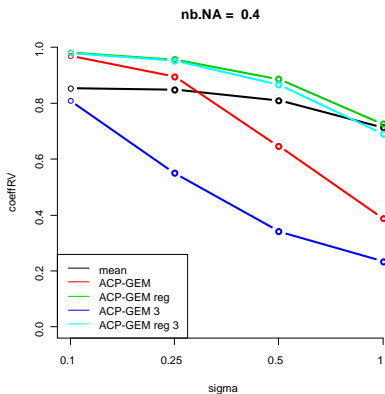
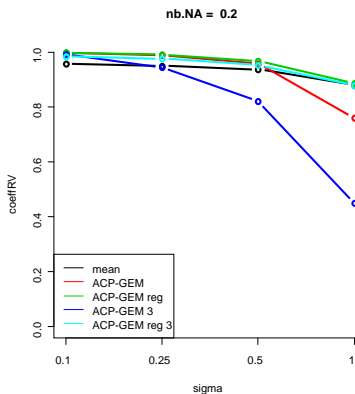
3 steps (a) and (b) are repeated until convergence

\Rightarrow Regularized svd: Ilin & Raiko (2009); Mazumder, Hastie & Tibshirani (2009)



Simulations

- $\mathbf{X}_{21 \times 10} = \mathbf{F}_{21 \times 2} \mathbf{U}'_{2 \times 10} + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma)$
- *RV* coefficient between configurations (complete / incomplete)





Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas**
 - Complete case
 - Incomplete case: multiple imputation
- 4 Choosing the number of dimensions
- 5 Conclusion



Stability in PCA

⇒ Exploratory framework

- PCA on a random sample from a population
 - ⇒ Individuals resampling (Timmerman *et al*, 2007): non parametric bootstrap
 - ⇒ Sampling variability
 - ⇒ All the dimensions are bootstrapped
 - ⇒ Confidence areas around the position of the variables
- PCA on a full population data?



Model in PCA

$$x_{ik} = \text{signal} + \text{noise}$$

- Random effect model: factor analysis, Probabilistic PCA
 - ⇒ individuals are exchangeable
 - ⇒ relationship between variables
- Fixed effect model: Lawley (1941), Caussinus (1986)
 - ⇒ individuals have different expectations
 - ⇒ study the individuals and the variables

$$x_{ik} = m_k + \sum_{s=1}^S f_{is} u_{ks} + \varepsilon_{ik}, \text{ where } \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$



Residuals bootstrap

- Semi-parametric bootstrap
 - ⇒ fluctuations due to the noise
 - ⇒ only the last dimensions are bootstrapped: "the noise goes everywhere"
 - ⇒ confidence areas around the position of the individuals and the variables



Residuals bootstrap

- ① PCA on $\mathbf{X} \Rightarrow \hat{\mathbf{M}}_{I \times K}$, $\hat{\mathbf{F}}_{I \times S}$ and $\hat{\mathbf{U}}_{K \times S}$ (S dimensions are kept);
- ② Model matrix $\hat{\mathbf{X}} = \hat{\mathbf{M}} + \hat{\mathbf{F}}\hat{\mathbf{U}}'$ and residuals $\hat{\varepsilon} = \mathbf{X} - \hat{\mathbf{X}}$;
- ③ Bootstrap procedure: repeat B times the step
 - a residuals are bootstrapped: ε^b
 \rightarrow drawn from $\mathcal{N}(0, \hat{\sigma}^2)$
 - b $\mathbf{X}^b = \hat{\mathbf{X}} + \varepsilon^b$
 - c PCA on \mathbf{X}^b to obtain $\hat{\mathbf{M}}^b$, $\hat{\mathbf{F}}^b$ and $\hat{\mathbf{U}}^b$ $\Rightarrow B$ couples $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1), \dots, (\hat{\mathbf{F}}^B, \hat{\mathbf{U}}^B)$



Residuals bootstrap

- ① PCA on $\mathbf{X} \Rightarrow \hat{\mathbf{M}}_{I \times K}$, $\hat{\mathbf{F}}_{I \times S}$ and $\hat{\mathbf{U}}_{K \times S}$ (S dimensions are kept);
- ② Model matrix $\hat{\mathbf{X}} = \hat{\mathbf{M}} + \hat{\mathbf{F}}\hat{\mathbf{U}}'$ and residuals $\hat{\varepsilon} = \mathbf{X} - \hat{\mathbf{X}}$;
 \Rightarrow Number of dimensions?
- ③ Bootstrap procedure: repeat B times the step
 - a residuals are bootstrapped: ε^b
 \rightarrow drawn from $\mathcal{N}(0, \hat{\sigma}^2)$
 \Rightarrow Under-estimation of the residuals?
 - b $\mathbf{X}^b = \hat{\mathbf{X}} + \varepsilon^b$
 - c PCA on \mathbf{X}^b to obtain $\hat{\mathbf{M}}^b$, $\hat{\mathbf{F}}^b$ and $\hat{\mathbf{U}}^b$ $\Rightarrow B$ couples $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1), \dots, (\hat{\mathbf{F}}^B, \hat{\mathbf{U}}^B)$



Residuals bootstrap

- ① PCA on $\mathbf{X} \Rightarrow \hat{\mathbf{M}}_{I \times K}$, $\hat{\mathbf{F}}_{I \times S}$ and $\hat{\mathbf{U}}_{K \times S}$ (S dimensions are kept);
- ② Model matrix $\hat{\mathbf{X}} = \hat{\mathbf{M}} + \hat{\mathbf{F}}\hat{\mathbf{U}}'$ and residuals $\hat{\varepsilon} = \mathbf{X} - \hat{\mathbf{X}}$;
 \Rightarrow Number of dimensions?
- ③ Bootstrap procedure: repeat B times the step
 - a residuals are bootstrapped: ε^b
 \rightarrow drawn from $\mathcal{N}(0, \hat{\sigma}^2)$
 \Rightarrow Under-estimation of the residuals?
 - b $\mathbf{X}^b = \hat{\mathbf{X}} + \varepsilon^b$
 - c PCA on \mathbf{X}^b to obtain $\hat{\mathbf{M}}^b$, $\hat{\mathbf{F}}^b$ and $\hat{\mathbf{U}}^b$ \Rightarrow B couples $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1), \dots, (\hat{\mathbf{F}}^B, \hat{\mathbf{U}}^B)$
 \Rightarrow Visualization?



Uncertainty due to missing values

⇒ A new source of variability to take into account

Iterative PCA: single imputation ⇒ a residual bootstrap procedure applied on the completed data set will lead to an underestimation of the variability

⇒ Multiple imputation

- 1 Generating B imputed data sets
- 2 Performing the analysis on each imputed data set
- 3 Combining the results: Total variability = Within imputation variability + Between imputation variability



Uncertainty due to missing values

⇒ A new source of variability to take into account

Iterative PCA: single imputation ⇒ a residual bootstrap procedure applied on the completed data set will lead to an underestimation of the variability

⇒ Multiple imputation

- 1 Generating B imputed data sets
- 2 Performing the analysis on each imputed data set
- 3 Combining the results: Total variability = Within imputation variability + Between imputation variability



Idea to generate B imputed data sets

$$x_{ik} = m_k + \sum_{s=1}^S f_{is} u_{ks} + \varepsilon_{ik}, \text{ with } \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

Iterative PCA on the incomplete data set $\Rightarrow (\hat{\mathbf{F}}, \hat{\mathbf{U}})$

\Rightarrow A first idea to generate different imputations:

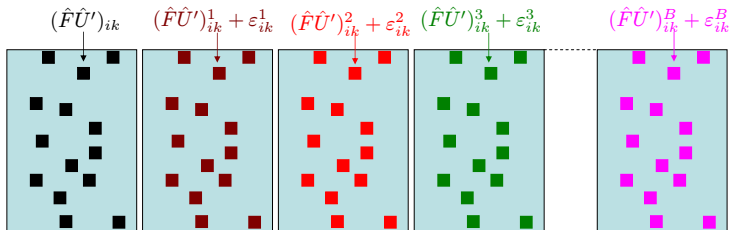
For $b = 1, \dots, B$, impute missing values x_{ik}^b by drawing from the predictive distribution $\mathcal{N}\left((\hat{\mathbf{F}}\hat{\mathbf{U}}')_{ik}, \hat{\sigma}^2\right)$

\Rightarrow “improper” imputation (Rubin, 1987)



“proper” multiple imputation

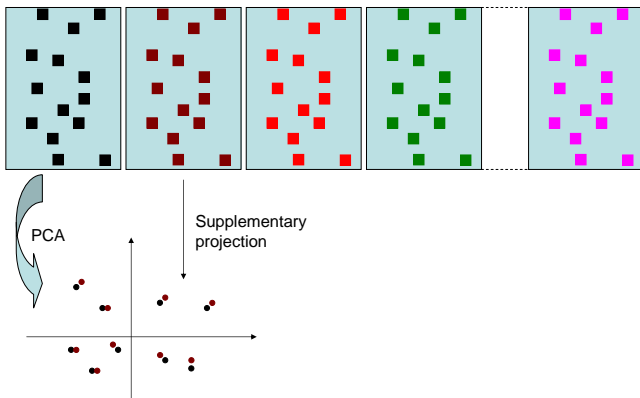
- 1 Variability of the parameters: obtaining B plausible sets of parameters, $(\hat{\mathbf{F}}, \hat{\mathbf{U}}')^1, \dots, (\hat{\mathbf{F}}, \hat{\mathbf{U}}')^B \Rightarrow$ residuals bootstrap
- 2 Noise : for $b = 1, \dots, B$, missing values x_{ik}^b are imputing by drawing from the predictive distribution $\mathcal{N}\left((\hat{\mathbf{F}}\hat{\mathbf{U}}')_{ik}^b, \hat{\sigma}^2\right)$





Supplementary projection

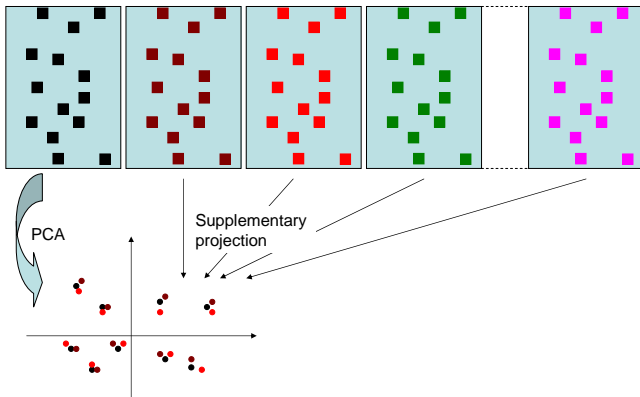
⇒ Individuals position (and variables) with other predictions





Supplementary projection

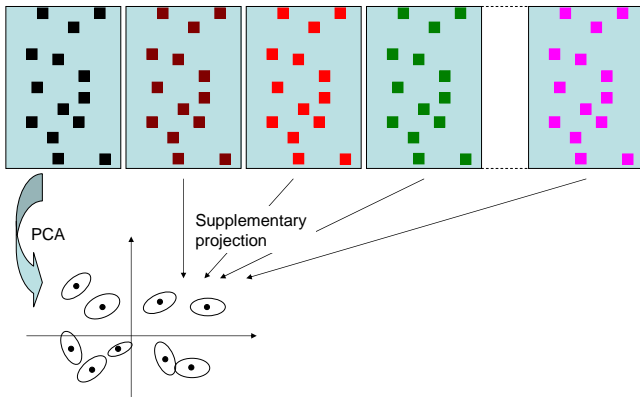
⇒ Individuals position (and variables) with other predictions





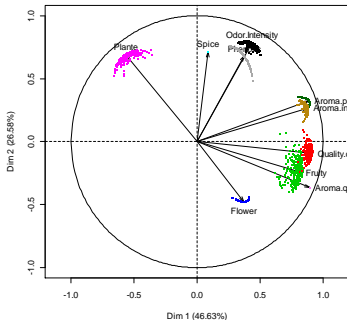
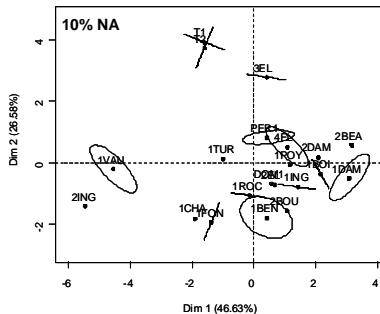
Supplementary projection

⇒ Individuals position (and variables) with other predictions





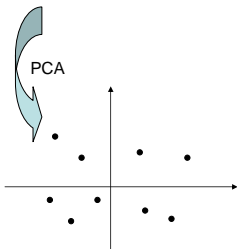
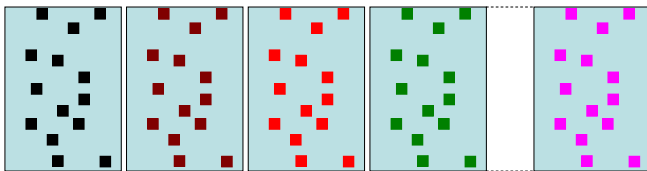
Supplementary projection





Between imputation variability

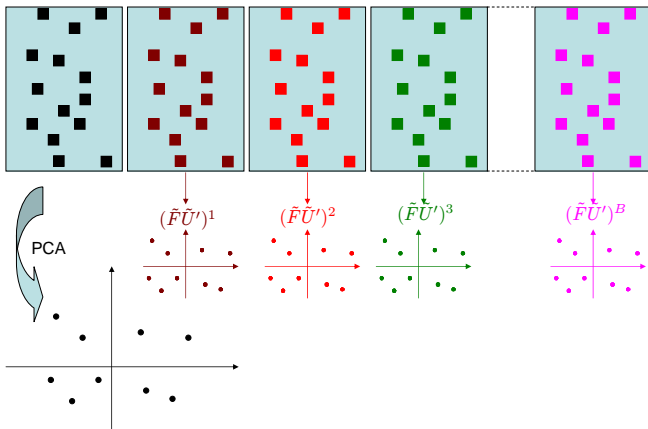
⇒ Influence of the different predictions on the parameters (PCA on each table)





Between imputation variability

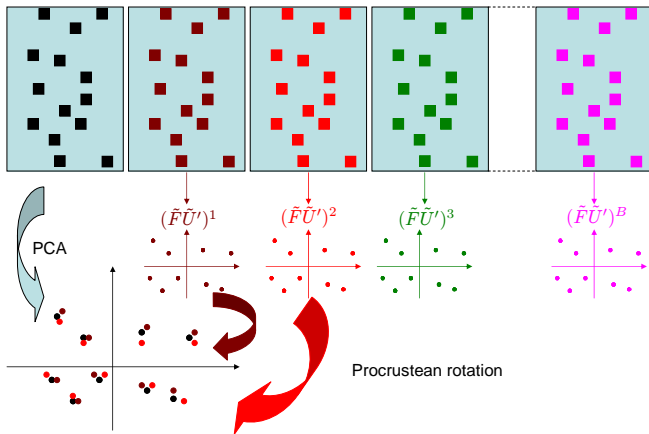
⇒ Influence of the different predictions on the parameters (PCA on each table)





Between imputation variability

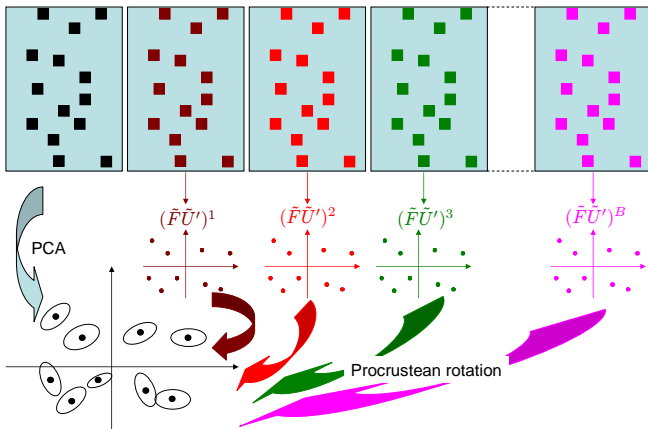
⇒ Influence of the different predictions on the parameters (PCA on each table)





Between imputation variability

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

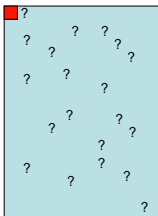


Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas
 - Complete case
 - Incomplete case: multiple imputation
- 4 Choosing the number of dimensions
- 5 Conclusion



Cross-validation



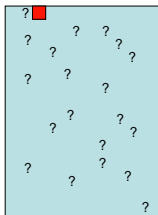
⇒ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_S = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

⇒ Computational costly



Cross-validation



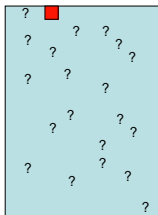
⇒ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_S = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

⇒ Computational costly



Cross-validation



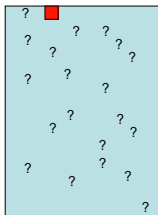
⇒ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_S = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

⇒ Computational costly



Cross-validation



⇒ EM-CV (Bro *et al.* 2008)

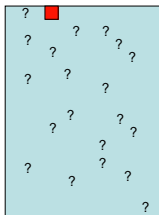
$$\text{MSEP}_S = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

⇒ Computational costly

⇒ In regression $\hat{y} = \mathbf{P}y$ (Craven & Whaba, 1979):

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - \mathbf{P}_{i,i}}$$

Cross-validation



⇒ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_S = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

⇒ Computational costly

⇒ In regression $\hat{y} = \mathbf{P}y$ (Craven & Whaba, 1979):

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - \mathbf{P}_{i,i}}$$

⇒ Aim: write PCA as $\hat{\mathbf{X}} = \mathbf{P}\mathbf{X}$ (complete case approximation)

$$\hat{x}_{ik}^{-ik} - x_{ik} \simeq \frac{\hat{x}_{ik} - x_{ik}}{1 - \mathbf{P}_{ik,ik}}$$



Projection in PCA

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

⇒ 2 projection matrices

$$\begin{cases} \hat{\mathbf{U}}' = (\hat{\mathbf{F}}' \hat{\mathbf{F}})^{-1} \hat{\mathbf{F}}' \mathbf{X} & \Rightarrow \mathbf{P}_{\mathbf{F}} = \hat{\mathbf{F}} (\hat{\mathbf{F}}' \hat{\mathbf{F}})^{-1} \hat{\mathbf{F}}' \\ \hat{\mathbf{F}} = \mathbf{X} \hat{\mathbf{U}} (\hat{\mathbf{U}}' \hat{\mathbf{U}})^{-1} & \Rightarrow \mathbf{P}_{\mathbf{U}} = \hat{\mathbf{U}} (\hat{\mathbf{U}}' \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}' \end{cases}$$

⇒ Bilinear form

$$\text{Model matrix} \quad \hat{\mathbf{X}}^{(S)} = \hat{\mathbf{F}} \hat{\mathbf{U}}' \quad \Rightarrow \hat{\mathbf{X}}^{(S)} = \mathbf{P}_{\mathbf{F}} \mathbf{X} = \mathbf{X} \mathbf{P}_{\mathbf{U}}$$

$$\text{Residual matrix} \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{X} - \hat{\mathbf{X}}^{(S)} \quad \Rightarrow \hat{\boldsymbol{\varepsilon}} = (\mathbb{I}_I - \mathbf{P}_{\mathbf{F}}) \mathbf{X} (\mathbb{I}_K - \mathbf{P}_{\mathbf{U}})$$

⇒ This equation can be developed; the model matrix is extracted:

$$\hat{\mathbf{X}} - \hat{\mathbf{X}}^{(S)} = \mathbf{X} - (\mathbb{I}_I \mathbf{X} \mathbf{P}_{\mathbf{U}} + \mathbf{P}_{\mathbf{F}} \mathbf{X} \mathbb{I}_K - \mathbf{P}_{\mathbf{F}} \mathbf{X} \mathbf{P}_{\mathbf{U}}).$$



Big projection matrix

$$\text{vec}(\hat{\mathbf{X}}^{(S)}) = \text{vec}(\mathbb{I}_I \mathbf{X} \mathbf{P}_U) + \text{vec}(\mathbf{P}_F \mathbf{X} \mathbb{I}_K) - \text{vec}(\mathbf{P}_F \mathbf{X} \mathbf{P}_U)$$

$$\text{vec}(\hat{\mathbf{X}}^{(S)}) = (\mathbf{P}'_U \otimes \mathbb{I}_I) \text{vec}(\mathbf{X}) + (\mathbb{I}'_K \otimes \mathbf{P}_F) \text{vec}(\mathbf{X}) - (\mathbf{P}'_U \otimes \mathbf{P}_F) \text{vec}(\mathbf{X})$$

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

$$\mathbf{P}_{IK \times IK}^{(S)} = (\mathbf{P}'_U \otimes \mathbb{I}_I) + (\mathbb{I}'_K \otimes \mathbf{P}_F) - (\mathbf{P}'_U \otimes \mathbf{P}_F)$$

- Number of independent parameters:

$$\text{tr}(\mathbf{P}^{(S)}) = S \times I + K \times S - S^2$$

- Ddl residuals: $\text{tr}(\mathbb{I}_{IK} - \mathbf{P}^{(S)}) = IK - (SI + KS - S^2)$

$$\hat{\sigma}_{cor}^2 = \frac{\|\mathbf{X} - \hat{\mathbf{F}}\hat{\mathbf{U}}'\|^2}{IK - (IS + KS - S^2)}$$



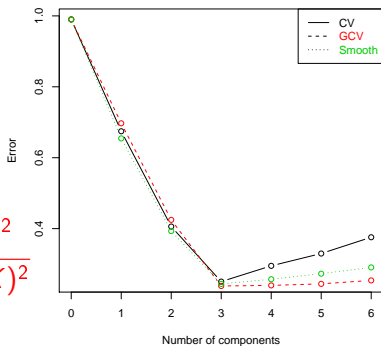
Approximations (Josse & Husson, 2010 submitted)

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

$$CV(S) = \frac{1}{IK} \sum_{i,k} (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

$$ACV(S) = \frac{1}{IK} \sum_{i,k} \left(\frac{\hat{x}_{ik} - x_{ik}}{1 - P_{ik,ik}} \right)^2$$

$$GCV(S) = \frac{1}{IK} \times \frac{\sum_{i,k} (\hat{x}_{ik} - x_{ik})^2}{(1 - \text{tr}(\mathbf{P}^{(S)})/IK)^2}$$



CV (600 s); ACV (0.019 s); GCV (0.006 s)



Approximation incomplete case

$$\text{GCV complete}(S) = \frac{IK \|\mathbf{X} - \mathbf{FU}\|^2}{(IK - (IS + KS - S^2))^2}$$

⇒ Number of dimensions in the incomplete case

$$\text{GCV incomplete}(S) = \frac{IK \|\mathbf{W} * (\mathbf{X} - \mathbf{FU})\|^2}{(IK - \text{nb missing} - (IS + KS - S^2))^2}$$



Outline

- 1 Introduction
- 2 Point estimates
- 3 Confidence areas
 - Complete case
 - Incomplete case: multiple imputation
- 4 Choosing the number of dimensions
- 5 Conclusion



Conclusion

- Point estimates in PCA \Rightarrow regularized iterative PCA
- Multiple imputation in PCA
- Number of dimensions in the incomplete case

\Rightarrow R package missMDA (function in the package FactoMineR)

- Missing values in multiple correspondence analysis
 \Rightarrow Regularized iterative MCA (Josse *et al*, 2011, submitted)



Perspectives

- Multiple imputation:
 - study the global variability: van Ginkel, Kiers (2010)
 - bootstrap with missing values
 - evaluation of the proposed method as a MI method
 - variability in MCA
- Regularization:
 - missing values in multi-table, three-way methods
 - regularization in the complete framework (Takane, 2006-2011)