

Multiple imputation with principal component methods

Vincent Audigier

Inserm ECSTRA team, Saint-Louis Hospital, Paris

ISPED, March 14, 2016

- ① Introduction
- ② Single imputation based on principal component methods
- ③ Multiple imputation for continuous data with PCA
- ④ Multiple imputation for categorical data with MCA
- ⑤ Conclusion

Missing values

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| NA | | | | | NA | NA | |
| | | | NA | | | | |
| | NA | | | | | | NA |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | NA | NA | NA | | | |

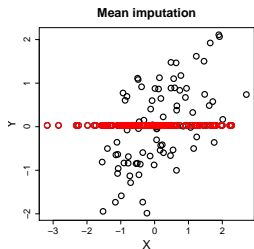
- Aim: inference on a quantity θ from incomplete data
→ point estimate $\hat{\theta}$ and associated variability T

Missing values

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| NA | | | | | NA | NA | |
| | | | NA | | | | |
| | NA | | | | | | NA |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | NA | NA | NA | | | |

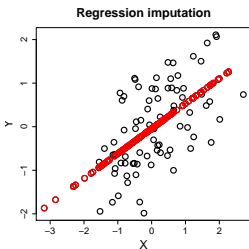
- Aim: inference on a quantity θ from incomplete data
→ point estimate $\hat{\theta}$ and associated variability T
- R : response indicator (known)
 $X = (X^{obs}, X^{miss})$: data (partially known)
MAR assumption: $P(R|X) = P(R|X^{obs})$
- Weighting procedures
- Likelihood approaches → EM, SEM
- Multiple Imputation → $P(X^{miss}|X^{obs})$

Why is single imputation inadequate?

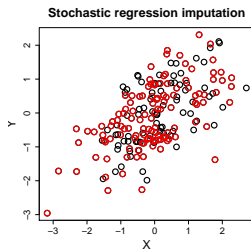


$$\begin{aligned} \mu_y &= 0 \\ \sigma_y &= 1 \\ \rho &= 0.6 \\ CI_{\mu_y} 95\% \end{aligned}$$

| |
|------|
| 0.01 |
| 0.5 |
| 0.30 |
| 39.4 |



| |
|------|
| 0.01 |
| 0.72 |
| 0.78 |
| 61.6 |



| |
|------|
| 0.01 |
| 0.99 |
| 0.59 |
| 70.8 |

⇒ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated from the imputed data set are underestimated

Generating imputed data sets

To simulate $P(X^{miss}|X^{obs}, \psi)$: Joint modelling or Fully conditional specification:

- JM: define $P(X, \psi)$, draw from $P(X^{miss}|X^{obs}, \hat{\psi}_1)$,
 $P(X^{miss}|X^{obs}, \hat{\psi}_2), \dots, P(X^{miss}|X^{obs}, \hat{\psi}_M)$
- FCS: define $P(X_k|X_{-k}, \psi_{-k})$, draw from $P(X_k^{miss}|X_{-k}^{obs}, \hat{\psi}_{-k})$ for all k . Repeat with $(\hat{\psi}_{-k}^2)_{1 \leq k \leq K}, \dots, (\hat{\psi}_{-k}^M)_{1 \leq k \leq K}$.

| | <i>Theory</i> | <i>Fit</i> | <i>Time</i> |
|------------|---------------|------------|-------------|
| <i>JM</i> | + | - | + |
| <i>FCS</i> | - | + | - |

However... $I < K$? high dependence? high dimensionality?

Generating imputed data sets

To simulate $P(X^{miss}|X^{obs}, \psi)$: Joint modelling or Fully conditional specification:

- JM: define $P(X, \psi)$, draw from $P(X^{miss}|X^{obs}, \hat{\psi}_1)$,
 $P(X^{miss}|X^{obs}, \hat{\psi}_2), \dots, P(X^{miss}|X^{obs}, \hat{\psi}_M)$
- FCS: define $P(X_k|X_{-k}, \psi_{-k})$, draw from $P(X_k^{miss}|X_{-k}^{obs}, \hat{\psi}_{-k})$ for all k . Repeat with $(\hat{\psi}_{-k}^2)_{1 \leq k \leq K}, \dots, (\hat{\psi}_{-k}^M)_{1 \leq k \leq K}$.

| | <i>Theory</i> | <i>Fit</i> | <i>Time</i> |
|------------|---------------|------------|-------------|
| <i>JM</i> | + | - | + |
| <i>FCS</i> | - | + | - |

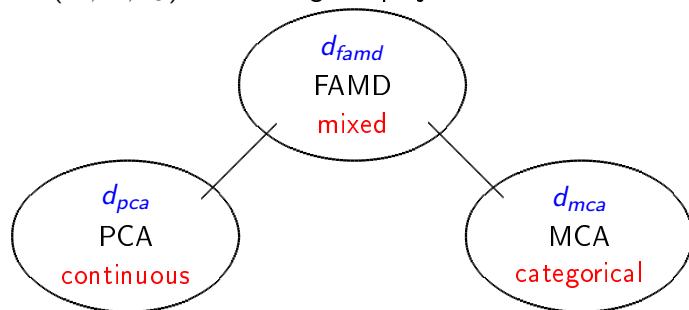
However... $I < K$? high dependence? high dimensionality?

Could principal component methods provide another way to deal with missing values?

Principal component methods

Dimensionality reduction:

- individuals are seen as elements of \mathbb{R}^K
- a distance d on \mathbb{R}^K
- $\text{Vect}(v_1, \dots, v_S)$ maximising the projected inertia



$$d_{famd}^2 = d_{pca}^2 + d_{mca}^2$$

- 1 Introduction
- 2 Single imputation based on principal component methods**
- 3 Multiple imputation for continuous data with PCA
- 4 Multiple imputation for categorical data with MCA
- 5 Conclusion

How to perform FAMD?

FAMD can be seen as the SVD of \mathbf{X} with weights for

- the continuous variables and categories: $(\mathbf{D}_\Sigma)^{-1}$
- the individuals: $\frac{1}{I} \mathbb{1}_I$

$$\rightarrow SVD \left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I \right)$$

$X =$

| | | | | | | | | |
|-------|-----|------|---|---|-----|---|---|---|
| 11.04 | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 2.04 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| 11.06 | | 2.01 | 0 | 1 | | 0 | 0 | 1 |
| 10.95 | | 1.67 | 0 | 1 | | 0 | 1 | 0 |

$D_\Sigma =$

| | | | |
|----------------|----------|----------------|-----------|
| σ_{x_1} | | | |
| | \ddots | | |
| | | σ_{x_k} | |
| | | | 0 |
| 0 | | | I_{k+1} |
| | | | \ddots |
| | | | I_K |

How to perform FAMD?

$$SVD \left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I \right) \longrightarrow \mathbf{X}_{I \times K} = \mathbf{U}_{I \times K} \mathbf{\Lambda}_{K \times K}^{1/2} \mathbf{V}_{K \times K}^\top$$

with $\mathbf{U}^\top \left(\frac{1}{I} \mathbb{1}_I \right) \mathbf{U} = \mathbb{1}_K$
 $\mathbf{V}^\top \mathbf{D}_\Sigma^{-1} \mathbf{V} = \mathbb{1}_K$

- principal components: $\hat{\mathbf{F}}_{I \times S} = \hat{\mathbf{U}}_{I \times S} \hat{\mathbf{\Lambda}}_{S \times S}^{1/2}$
- loadings: $\hat{\mathbf{V}}_{K \times S}^\top$
- fitted matrix: $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} \hat{\mathbf{\Lambda}}_{S \times S}^{1/2} \hat{\mathbf{V}}_{K \times S}^\top$

$$\| \hat{\mathbf{X}} - \mathbf{X} \|_{\mathbf{D}_\Sigma^{-1} \otimes \frac{1}{I} \mathbb{1}}^2 = \text{tr} \left((\hat{\mathbf{X}} - \mathbf{X}) \mathbf{D}_\Sigma^{-1} (\hat{\mathbf{X}} - \mathbf{X})^\top \frac{1}{I} \mathbb{1}_I \right)$$

minimized under the constraint of rank S

Properties of the method

- The distance between individuals is:

$$d^2(i, i') = \sum_{j=1}^k \frac{(x_{ij} - x_{i'j})^2}{\sigma_{x_j}^2} + \sum_{j=k+1}^K \frac{1}{I_j} (x_{ij} - x_{i'j})^2$$

- The principal component \mathbf{F}_s maximises:

$$\sum_{var \in \text{continuous}} r^2(\mathbf{F}_s, var) + \sum_{var \in \text{categorical}} \eta^2(\mathbf{F}_s, var)$$

FAMD with missing values

⇒ FAMD: least squares

$$\|\mathbf{X}_{I \times K} - \mathbf{U}_{I \times S} \mathbf{\Lambda}_{S \times S}^{\frac{1}{2}} \mathbf{V}_{K \times S}^T\|^2$$

⇒ FAMD with missing values: weighted least squares

$$\|\mathbf{W}_{I \times K} * (\mathbf{X}_{I \times K} - \mathbf{U}_{I \times S} \mathbf{\Lambda}_{S \times S}^{\frac{1}{2}} \mathbf{V}_{K \times S}^T)\|^2$$

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

Many algorithms developed for PCA such as NIPALS (Christoffersson, 1970) or iterative PCA (Kiers, 1997)

FAMD with missing values

Iterative FAMD algorithm:

- 1 initialization: imputation by mean/proportion
- 2 iterate until convergence
 - (a) estimation of the parameters of FAMD
 \rightarrow SVD of $(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I)$
 - (b) imputation of the missing values with
 $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} \hat{\Lambda}_{S \times S}^{1/2} \hat{\mathbf{V}}_{K \times S}^\top$
 - (c) \mathbf{D}_Σ is updated

| | | | | | |
|-------|-----|------|-----|-----|-----|
| NA | ... | 2.07 | A | ... | A |
| 10.76 | ... | 1.86 | A | ... | A |
| 11.02 | ... | NA | A | ... | NA |
| 11.02 | ... | 1.92 | B | ... | B |
| ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | NA | ... | C |
| NA | ... | 1.67 | B | ... | B |

\rightarrow

| | | | | | | | | |
|-------|-----|------|-----|-----|-----|-----|-----|-----|
| NA | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | NA | 1 | 0 | ... | NA | NA | NA |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | NA | NA | ... | 0 | 0 | 1 |
| NA | ... | 1.67 | 0 | 1 | ... | 0 | 1 | 0 |

FAMD with missing values

Iterative FAMD algorithm:

- 1 initialization: imputation by mean/proportion
- 2 iterate until convergence
 - (a) estimation of the parameters of FAMD
 \rightarrow SVD of $(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I)$
 - (b) imputation of the missing values with
 $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} \hat{\Lambda}_{S \times S}^{1/2} \hat{\mathbf{V}}_{K \times S}^\top$
 - (c) \mathbf{D}_Σ is updated

| | | | | | |
|-------|-----|------|-----|-----|-----|
| NA | ... | 2.07 | A | ... | A |
| 10.76 | ... | 1.86 | A | ... | A |
| 11.02 | ... | NA | A | ... | NA |
| 11.02 | ... | 1.92 | B | ... | B |
| ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | NA | ... | C |
| NA | ... | 1.67 | B | ... | B |

\rightarrow

| | | | | | | | | |
|-------|-----|------|------|------|-----|------|------|------|
| 11.01 | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 1.89 | 1 | 0 | ... | 0.61 | 0.19 | 0.20 |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | 0.32 | 0.68 | ... | 0 | 0 | 1 |
| 11.01 | ... | 1.67 | 0 | 1 | ... | 0 | 1 | 0 |

FAMD with missing values

Iterative FAMD algorithm:

- 1 initialization: imputation by mean/proportion
- 2 iterate until convergence
 - (a) estimation of the parameters of FAMD
 \rightarrow SVD of $(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I)$
 - (b) imputation of the missing values with
 $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} \hat{\Lambda}_{S \times S}^{1/2} \hat{\mathbf{V}}_{K \times S}^\top$
 - (c) \mathbf{D}_Σ is updated

| | | | | | |
|-------|-----|------|-----|-----|-----|
| NA | ... | 2.07 | A | ... | A |
| 10.76 | ... | 1.86 | A | ... | A |
| 11.02 | ... | NA | A | ... | NA |
| 11.02 | ... | 1.92 | B | ... | B |
| ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | NA | ... | C |
| NA | ... | 1.67 | B | ... | B |

\rightarrow

| | | | | | | | | |
|-------|-----|------|------|------|-----|------|------|------|
| 11.04 | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 2.04 | 1 | 0 | ... | 0.81 | 0.05 | 0.14 |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | 0.25 | 0.75 | ... | 0 | 0 | 1 |
| 10.95 | ... | 1.67 | 0 | 1 | ... | 0 | 1 | 0 |

Single imputation with FAMD (Audigier et al., 2014)

Iterative FAMD algorithm:

- 1 initialization: imputation by mean/proportion
- 2 iterate until convergence
 - (a) estimation of the parameters of FAMD
 \rightarrow SVD of $(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I)$
 - (b) imputation of the missing values with
 $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} \hat{\Lambda}_{S \times S}^{1/2} \hat{\mathbf{V}}_{K \times S}^\top$
 - (c) \mathbf{D}_Σ is updated

| | | | | | |
|-------|-----|------|-----|-----|-----|
| 11.04 | ... | 2.07 | A | ... | A |
| 10.76 | ... | 1.86 | A | ... | A |
| 11.02 | ... | 2.04 | A | ... | A |
| 11.02 | ... | 1.92 | B | ... | B |
| ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | B | ... | C |
| 10.95 | ... | 1.67 | B | ... | B |



| | | | | | | | | |
|-------|-----|------|------|------|-----|------|------|------|
| 11.04 | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 2.04 | 1 | 0 | ... | 0.81 | 0.05 | 0.14 |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | 0.25 | 0.75 | ... | 0 | 0 | 1 |
| 10.95 | ... | 1.67 | 0 | 1 | ... | 0 | 1 | 0 |

\Rightarrow the imputed values can be seen as degree of membership

Single imputation with FAMD (Audigier et al., 2014)

Iterative FAMD algorithm:

- 1 initialization: imputation by mean/proportion
- 2 iterate until convergence
 - (a) estimation of the parameters of FAMD
 \rightarrow SVD of $(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I} \mathbb{1}_I)$
 - (b) imputation of the missing values with
 $\hat{\mathbf{X}}_{I \times K} = \hat{\mathbf{U}}_{I \times S} f(\hat{\lambda}_{S \times S}^{1/2}) \hat{\mathbf{V}}_{K \times S}^\top$
 - (c) \mathbf{D}_Σ is updated

$$f(\hat{\lambda}_s^{1/2}) = \hat{\lambda}_s^{1/2} - \frac{\hat{\sigma}^2}{\hat{\lambda}_s^{1/2}}$$

| | | | | | |
|-------|-----|------|-----|-----|-----|
| 11.04 | ... | 2.07 | A | ... | A |
| 10.76 | ... | 1.86 | A | ... | A |
| 11.02 | ... | 2.04 | A | ... | A |
| 11.02 | ... | 1.92 | B | ... | B |
| ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | B | ... | C |
| 10.95 | ... | 1.67 | B | ... | B |

←

| | | | | | | | | |
|-------|-----|------|------|------|-----|------|------|------|
| 11.04 | ... | 2.07 | 1 | 0 | ... | 1 | 0 | 0 |
| 10.76 | ... | 1.86 | 1 | 0 | ... | 1 | 0 | 0 |
| 11.02 | ... | 2.04 | 1 | 0 | ... | 0.81 | 0.05 | 0.14 |
| 11.02 | ... | 1.92 | 0 | 1 | ... | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11.06 | ... | 2.01 | 0.25 | 0.75 | ... | 0 | 0 | 1 |
| 10.95 | ... | 1.67 | 0 | 1 | ... | 0 | 1 | 0 |

\Rightarrow the imputed values can be seen as degree of membership

Simulation results

Single imputation with FAMD shows a high quality of prediction compared to random forests (Stekhoven and Bühlmann, 2012)

- on real data
- when the relationships between continuous variables are linear
- for rare categories
- with MAR/MCAR mechanism

Can impute mixed, continuous or categorical data

Simulation results

Single imputation with FAMD shows a high quality of prediction compared to random forests (Stekhoven and Bühlmann, 2012)

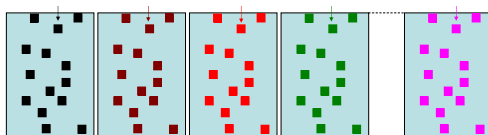
- on real data
- when the relationships between continuous variables are linear
- for rare categories
- with MAR/MCAR mechanism

Can impute mixed, continuous or categorical data

But a single imputation method only

From single imputation to multiple imputation

$$P(X^{miss}|X^{obs}, \psi_1) \quad \dots \quad P(X^{miss}|X^{obs}, \psi_M)$$



- 1 Reflect the variability on the parameters of the imputation model

$$\rightarrow \left(\left(\hat{\mathbf{U}}_{I \times S}, \hat{\Lambda}_{S \times S}^{1/2}, \hat{\mathbf{V}}_{K \times S}^T \right)_1, \dots, \left(\hat{\mathbf{U}}_{I \times S}, \hat{\Lambda}_{S \times S}^{1/2}, \hat{\mathbf{V}}_{K \times S}^T \right)_M \right)$$

Bayesian or Bootstrap

- 2 Add a disturbance on the prediction by $\hat{X}_m = \hat{\mathbf{U}}_m \hat{\Lambda}_m^{1/2} \hat{\mathbf{V}}_m^T$
 → need to distinguish continuous and categorical data

- 1 Introduction
- 2 Single imputation based on principal component methods
- 3 Multiple imputation for continuous data with PCA**
- 4 Multiple imputation for categorical data with MCA
- 5 Conclusion

PCA model (Caussinus, 1986)

Model

$$\begin{aligned}\mathbf{X}_{I \times K} &= \tilde{\mathbf{X}}_{I \times K} + \varepsilon_{I \times K} \\ &= \mathbf{U}_{I \times S} \mathbf{\Lambda}_{S \times S}^{\frac{1}{2}} \mathbf{V}_{K \times S}^T + \varepsilon_{I \times K} \text{ with } \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_K)\end{aligned}$$

Maximum Likelihood:

$$\hat{\mathbf{X}}^S = \hat{\mathbf{U}}_{I \times S} \hat{\mathbf{\Lambda}}_{S \times S}^{\frac{1}{2}} \hat{\mathbf{V}}_{K \times S}^T \rightarrow \hat{\sigma}^2 = \|\mathbf{X} - \hat{\mathbf{X}}^S\|^2 / \text{degrees of f.}$$

Bayesian formulation:

- Hoff (2007): Uniform prior for \mathbf{U} and \mathbf{V} , Gaussian on $(\lambda_s)_{s=1 \dots S}$
- Todeschini et al. (2013): Exponential prior on $(\lambda_s)_{s=1 \dots S}$
- Verbanck et al. (2013): Prior on $\tilde{\mathbf{X}}$

Bayesian PCA (Verbanck et al., 2013)

$$\begin{aligned}
 \text{Model: } \mathbf{X}_{I \times K} &= \tilde{\mathbf{X}}_{I \times K} + \varepsilon_{I \times K} \\
 x_{ik} &= \tilde{x}_{ik} + \varepsilon_{ik}, \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \\
 &= \sum_{s=1}^S \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ik} \\
 &= \sum_{s=1}^S \tilde{x}_{ik}^{(s)} + \varepsilon_{ik}
 \end{aligned}$$

$$\text{Prior: } \tilde{x}_{ik}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$$

$$\text{Posterior: } \left(\tilde{x}_{ik}^{(s)} | x_{ik}^{(s)} \right) \sim \mathcal{N}(\Phi_s x_{ik}^{(s)}, \Phi_s \sigma^2) \text{ with } \Phi_s = \frac{\tau_s^2}{\tau_s^2 + \sigma^2}$$

$$\text{Empirical Bayes for } \tau_s^2: \hat{\tau}_s^2 = \left(\hat{\lambda}_s - \hat{\sigma}^2 \right)$$

$$\hat{\Phi}_s = \frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s} = \frac{\text{signal variance}}{\text{total variance}} \quad (\text{Efron and Morris, 1972})$$

Multiple imputation with Bayesian PCA (Audigier et al., 2015)

- 1 Variability of the parameters, M plausible $(\tilde{x}_{ij})^1, \dots, (\tilde{x}_{ij})^M$
 - Posterior distribution: Bayesian PCA

$$\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$$

- 2 Imputation according to the PCA model using the set of M parameters $x_{ij}^{miss} \leftarrow \mathcal{N}(\hat{x}_{ij}, \hat{\sigma}^2)$

Multiple imputation with Bayesian PCA (Audigier et al., 2015)

- 1 Variability of the parameters, M plausible $(\tilde{x}_{ij})^1, \dots, (\tilde{x}_{ij})^M$
 - Posterior distribution: Bayesian PCA

$$\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$$

- Data Augmentation (Tanner and Wong, 1987)
- 2 Imputation according to the PCA model using the set of M parameters $x_{ij}^{miss} \leftarrow \mathcal{N}(\hat{x}_{ij}, \hat{\sigma}^2)$

Multiple imputation with Bayesian PCA (Audigier et al., 2015)

Data augmentation

- a Gibbs sampler
- simulate $(\psi, X^{miss} | X^{obs})$ from
 - (I) $(X^{miss} | X^{obs}, \psi)$: imputation
 - (P) $(\psi | X^{obs}, X^{miss})$: draw from the posterior
- convergence checked by graphical investigations

For Bayesian PCA:

- initialisation: ML estimate for $\tilde{\mathbf{X}}$
- for ℓ in $1 \dots L$
 - (I) Given $\tilde{\mathbf{X}}$, $x_{ij}^{miss} \leftarrow \mathcal{N}(\tilde{x}_{ij}, \hat{\sigma}^2)$
 - (P) $\tilde{x}_{ij} \leftarrow \mathcal{N}\left(\sum_s \hat{\phi}_s x_{ij}^{(s)}, \frac{\hat{\sigma}^2 \sum_s \hat{\phi}_s}{l-1}\right)$

MI methods for continuous data

Generally based on normal distribution:

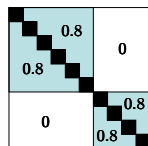
- JM: $X_{I \times K} : x_i. \sim \mathcal{N}(\mu, \Sigma)$ (Honaker et al., 2011)
 - 1 Bootstrap rows: X^1, \dots, X^M
 - EM algorithm: $(\mu^1, \Sigma^1), \dots, (\mu^M, \Sigma^M)$
 - 2 Imputation: x_i^m drawn from $\mathcal{N}(\mu^m, \Sigma^m)$

- FCS: $\mathcal{N}(\mu_{X_k|X_{(-k)}}, \Sigma_{X_k|X_{(-k)}})$ (Van Buuren, 2012)
 - 1 Bayesian approach: (β^m, σ^m)
 - 2 Imputation: stochastic regression x_{ij}^m drawn from $\mathcal{N}(X_{(-k)}\beta^m, \sigma^m)$

Simulations

- Quantities of interest: $\theta_1 = \mathbb{E}[Y]$, $\theta_2 = \beta_1$, $\theta_3 = \rho$
- 1000 simulations

- data set drawn from $\mathcal{N}_p(\mu, \Sigma)$ with a two-block structure, varying I (30 or 200), K (6 or 60) and ρ (0.3 or 0.9)



- 10% or 30% of missing values using a MCAR mechanism
- multiple imputation using $M = 20$ imputed arrays
- Criteria
 - bias
 - CI width, coverage

Results for the expectation

| | parameters | | | | confidence interval width | | | coverage | | |
|----|------------|-----|--------|-----|---------------------------|-------|------------|----------|-------|------------|
| | I | K | ρ | % | JM | FCS | BiassMIPCA | JM | FCS | BiassMIPCA |
| 1 | 30 | 6 | 0.3 | 0.1 | 0.803 | 0.805 | 0.781 | 0.955 | 0.953 | 0.950 |
| 2 | 30 | 6 | 0.3 | 0.3 | | 1.010 | 0.898 | | 0.971 | 0.949 |
| 3 | 30 | 6 | 0.9 | 0.1 | 0.763 | 0.759 | 0.756 | 0.952 | 0.95 | 0.949 |
| 4 | 30 | 6 | 0.9 | 0.3 | | 0.818 | 0.783 | | 0.965 | 0.953 |
| 5 | 30 | 60 | 0.3 | 0.1 | | | 0.775 | | | 0.955 |
| 6 | 30 | 60 | 0.3 | 0.3 | | | 0.864 | | | 0.952 |
| 7 | 30 | 60 | 0.9 | 0.1 | | | 0.742 | | | 0.953 |
| 8 | 30 | 60 | 0.9 | 0.3 | | | 0.759 | | | 0.954 |
| 9 | 200 | 6 | 0.3 | 0.1 | 0.291 | 0.294 | 0.292 | 0.947 | 0.947 | 0.946 |
| 10 | 200 | 6 | 0.3 | 0.3 | 0.328 | 0.334 | 0.325 | 0.954 | 0.959 | 0.952 |
| 11 | 200 | 6 | 0.9 | 0.1 | 0.281 | 0.281 | 0.281 | 0.953 | 0.95 | 0.952 |
| 12 | 200 | 6 | 0.9 | 0.3 | 0.288 | 0.289 | 0.288 | 0.948 | 0.951 | 0.951 |
| 13 | 200 | 60 | 0.3 | 0.1 | | 0.304 | 0.289 | | 0.957 | 0.945 |
| 14 | 200 | 60 | 0.3 | 0.3 | | 0.384 | 0.313 | | 0.981 | 0.958 |
| 15 | 200 | 60 | 0.9 | 0.1 | | 0.282 | 0.279 | | 0.951 | 0.948 |
| 16 | 200 | 60 | 0.9 | 0.3 | | 0.296 | 0.283 | | 0.958 | 0.952 |

Properties for BayesMIPCA

A MI method based on a Bayesian treatment of the PCA model

advantages

- captures the structure of the data: good inferences for regression coefficient, correlation, mean
- a dimensionality reduction method: ($I < K$ or $I > K$, low or high percentage of missing values)
- no inversion issue: strong or weak relationships
- a regularization strategy improving stability

remains competitive if:

- the low rank assumption is not verified
- the Gaussian assumption is not true

- ① Introduction
- ② Single imputation based on principal component methods
- ③ Multiple imputation for continuous data with PCA
- ④ Multiple imputation for categorical data with MCA
- ⑤ Conclusion

Multiple imputation for categorical data using MCA

MI for categorical data is **very challenging** for a moderate number of variables

- estimation issues
- storage issues

Multiple imputation for categorical data using MCA

MI for categorical data is **very challenging** for a moderate number of variables

- estimation issues
- storage issues

MI with MCA

- 1 Variability on the parameters of the imputation model

$$\left(\left(\hat{\mathbf{U}}_{I \times S}, \hat{\Lambda}_{S \times S}^{1/2}, \hat{\mathbf{V}}_{K \times S}^T \right)_1, \dots, \left(\hat{\mathbf{U}}_{I \times S}, \hat{\Lambda}_{S \times S}^{1/2}, \hat{\mathbf{V}}_{K \times S}^T \right)_M \right)$$

→ A **non-parametric bootstrap** approach

- 2 Add a disturbance on the MCA prediction $\hat{X}_m = \hat{\mathbf{U}}_m \hat{\Lambda}_m^{1/2} \hat{\mathbf{V}}_m^T$

Multiple imputation with MCA (Audigier et al., 2015)

- Variability of the parameters of MCA ($\hat{\mathbf{U}}_{I \times S}$, $\hat{\Lambda}_{S \times S}^{1/2}$, $\hat{\mathbf{V}}_{K \times S}^T$) using a non-parametric bootstrap:
 - define M weightings $(R_m)_{1 \leq m \leq M}$ for the individuals
 - estimate MCA parameters using SVD of $(\mathbf{X}, \frac{1}{K}(\mathbf{D}_\Sigma)^{-1}, R_m)$
- Imputation:

| $\hat{\mathbf{X}}_1$ | | | $\hat{\mathbf{X}}_2$ | | | $\hat{\mathbf{X}}_M$ | | |
|----------------------|------|-----|----------------------|------|-----|----------------------|------|-----|
| 1 | 0 | ... | 1 | 0 | ... | 1 | 0 | ... |
| 1 | 0 | ... | 1 | 0 | ... | 1 | 0 | ... |
| 1 | 0 | ... | 0.81 | 0.19 | ... | 0.60 | 0.40 | ... |
| 0.25 | 0.75 | ... | 0.26 | 0.74 | ... | 0.20 | 0.80 | ... |
| 0 | 1 | ... | 0 | 1 | ... | 0 | 1 | ... |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |

Draw categories from the values of $(\hat{\mathbf{X}}_m)_{1 \leq m \leq M}$

| | | | | | | | | |
|---|-----|---|---|-----|---|---|-----|---|
| A | ... | A | A | ... | A | A | ... | A |
| A | ... | A | A | ... | A | A | ... | A |
| A | ... | B | A | ... | A | A | ... | B |
| B | ... | C | B | ... | C | B | ... | C |
| B | ... | B | B | ... | B | B | ... | B |

Properties

MCA address the categorical data challenge by

- requiring a small number of parameters
- preserving the essential data structure
- using a regularisation strategy

MIMCA can be applied on various data sets

- small or large number of variables/categories
- small or large number of individuals

MI methods for categorical data

- Log-linear model (Schafer, 1997)

- Hypothesis on $X = (x_{ijk})_{i,j,k}$: $X|\psi \sim \mathcal{M}(n, \psi)$

$$\log(\psi_{ijk}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

- 1 Variability of the parameter ψ : Bayesian formulation
- 2 Imputation using the set of M parameters

- Latent class model (Si and Reiter, 2013)

- Hypothesis: $\mathbb{P}(X = (x_1, \dots, x_K); \psi) = \sum_{\ell=1}^L \left(\psi_{\ell} \prod_{k=1}^K \psi_{x_k}^{(\ell)} \right)$

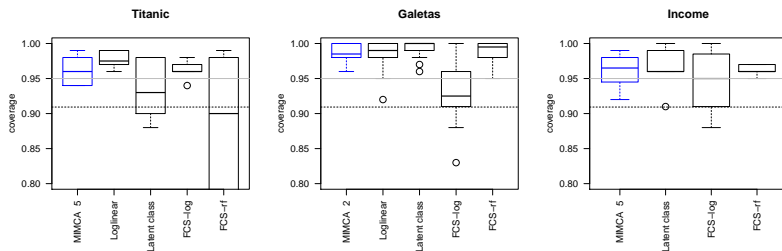
- 1 Variability of the parameters ψ_L and ψ_X : Bayesian formulation
- 2 Imputation using the set of M parameters

- FCS: GLM (Van Buuren, 2012) or Random Forests (Doove et al., 2014; Shah et al., 2014)

Simulations from real data sets

- Quantities of interest: θ = parameters of a logistic model
- Simulation design (repeated 200 times)
 - the real data set is considered as a population
 - drawn one sample from the data set
 - generate 20% of missing values
 - multiple imputation using $M = 5$ imputed arrays
- Criteria
 - bias
 - CI width, coverage
- Comparison with :
 - JM: log-linear model, latent class model
 - FCS: logistic regression, random forests

Results - Inference



| | Titanic | Galetas | Income |
|----------------------|----------|-----------|----------|
| Number of variables | 4 | 4 | 14 |
| Number of categories | ≤ 4 | ≤ 11 | ≤ 9 |

Results - Time

| | Titanic | Galetas | Income |
|--------------------|---------|---------|---------------|
| MIMCA | 2.750 | 8.972 | 58.729 |
| Loglinear | 0.740 | 4.597 | NA |
| Latent class model | 10.854 | 17.414 | 143.652 |
| FCS logistic | 4.781 | 38.016 | 881.188 |
| FCS forests | 265.771 | 112.987 | 6329.514 |

Table: Time consumed in second

| | Titanic | Galetas | Income |
|-----------------------|---------|---------|--------|
| Number of individuals | 2201 | 1192 | 6876 |
| Number of variables | 4 | 4 | 14 |

Conclusion

MI methods using dimensionality reduction method

- captures the relationships between variables
- captures the similarities between individuals
- requires a small number of parameters

Address some imputation issues:

- can be applied on various data sets
- provide correct inferences for analysis model based on relationships between pairs of variables

Available in the R package missMDA, with a user guide on <http://vincentaudigier.weebly.com/links.html>

Perspectives

To go further:

- a MI method based on FAMD
- a MI method for cluster structure
- uncertainty on the number of dimensions S
- with a very large number of individuals?

References I

- V. Audigier, F. Husson, and J. Josse. MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 2016.
- V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 2015.
- V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, pages 1–22, 2014.
- D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.
- J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.