

Missing values in multiple multivariate dataset

E.Morand, J.Josse, F.Husson and J.Pagès

Applied Mathematics Department, Agrocampus Rennes

ASMDA, Chania, Crete, Greece
May, 30th 2007

Overview

- **Presentation, problem, data**
- **Missing values: the state of the art**
- **Methods**
- **Simulation, results**
- **Conclusion, prospects**

The data

- Sensory evaluation
- 12 orange juices, 9 panelists
- Direct collection of sensory distances: napping (*Pagès 2003*)

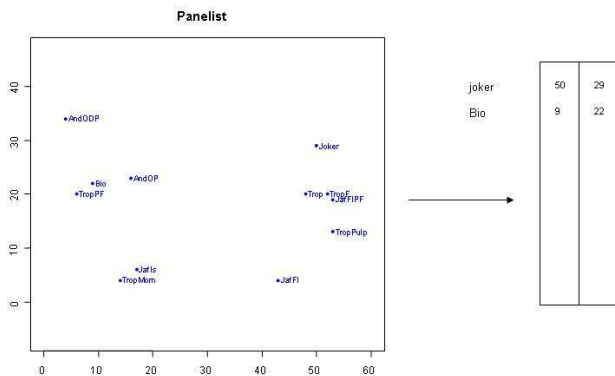


Figure: Napping

Sensory analysis

- Obtain main sensory dimensions
- Is there a consensus configuration?

- Problem: each panelist can't taste more than a certain number of products
- Planned missing products per panelist → a special pattern of missing values

A specific pattern of missing values

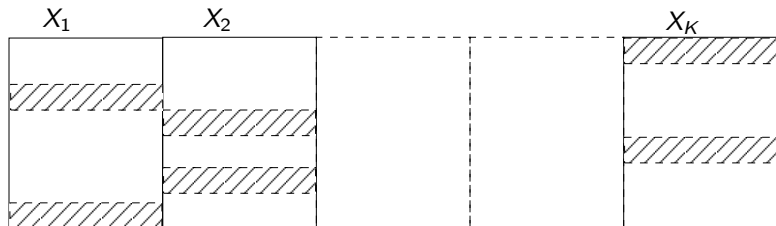


Figure: Multiple multivariate dataset with missing values.

⇒ **Aim** : perform a MFA (Multiple Factor Analysis) on incomplete data

The state of art about missing values

- *Little and Rubin (1987, 2002), Schafer (2002)*
- Complete case, single imputation
- Model based approach → EM (*Dempster, Laird, Rubin, 1977*)
- Multiple imputation (*Rubin, 1987*)
- What we are going to do with our structure of data and our specific missing data pattern?
 - Context of multivariate data analysis? → imputation
 - Hypothesis, distribution? → estimation

Methodology

- Two groups of variables
 - Procrustes analysis
 - RV maximisation
 - Probabilistic MFA
- More than two groups of variables

Procrustes analysis for two configurations

- Procrustes analysis is a method based on rotation, reflection, translation, and dilatation of a set of points in order to fit it to another fixed set of points
- Matches two configurations of points representing the same individuals
- Least squares procrustean adjustment

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | | 2 |
| 4 | 5 | 3 |

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

| | |
|---|---|
| 2 | 3 |
| x | |
| 1 | 4 |

Complete configuration

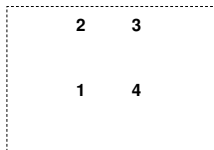
| | | |
|---|---|---|
| 1 | | 2 |
| 4 | 5 | 3 |

| | | |
|---|---|---|
| 1 | | 2 |
| | x | |
| 4 | 5 | 3 |

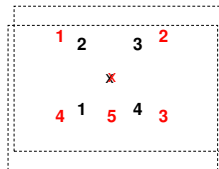
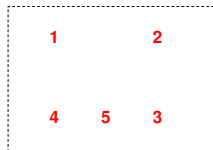
Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration



Complete configuration

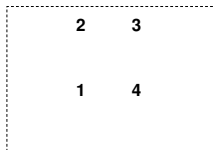


Translation

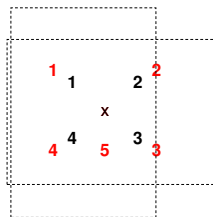
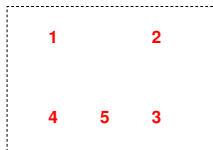
Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration



Complete configuration



Rotation

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | | 2 |
| 4 | 5 | 3 |

| | | |
|---|---|---|
| 1 | | 2 |
| | x | |
| 4 | 5 | 3 |

Dilatation

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | | 2 |
| 4 | 5 | 3 |
| 1 | | 2 |
| | x | |
| 4 | 5 | 3 |

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | 2 | |
| 4 | 5 | 3 |

| | | |
|---|---|---|
| 1 | 2 | |
| | x | |
| 4 | 5 | 3 |

Inverse dilatation

Procustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | 2 | |
| 4 | 5 | 3 |

| | |
|---|---|
| 2 | 3 |
| x | 5 |
| 1 | 4 |

Inverse rotation

Procrustes rotation to deal with missing values

⇒ Imputing a configuration with the help of the other

Incomplete configuration

| | |
|---|---|
| 2 | 3 |
| 1 | 4 |

Complete configuration

| | | |
|---|---|---|
| 1 | 2 | |
| 4 | 5 | 3 |

| | |
|---|---|
| 2 | 3 |
| x | 5 |
| 1 | 4 |

Inverse translation

RV maximisation

A measure of relationship between two sets of variables (*Escoufier 1973*):

$$RV(X_1, X_2) = \frac{\langle W_{X_1}, W_{X_2} \rangle}{\|W_{X_1}\| \|W_{X_2}\|}$$

- $RV(X_1, BX_1 + c) = 1$, B is an orthogonal matrix and c is a constant vector
- Distance between data matrices

$$\begin{aligned} d(X_1, X_2) &= \left\| \frac{X_1^T X_1}{(\text{tr}(X_1^T X_1))^2)^{1/2}} - \frac{X_2^T X_2}{(\text{tr}(X_2^T X_2))^2)^{1/2}} \right\| \\ &= \sqrt{2} \sqrt{1 - RV(X_1, X_2)} \end{aligned}$$

- $RV = 1$ if and only if $d(X_1, X_2) = 0$.

RV maximisation

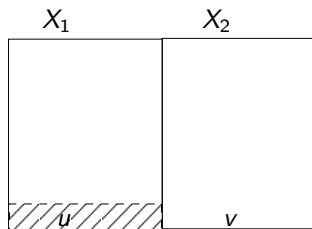


Figure: RV maximization *Crettaz de Roten*.

⇒ Minimization of a distance between two clouds of points = maximize the RV

The probabilistic PCA (Bishop and Tipping, Roweis)

- PCA dates back to long before probabilistic latent variables models (Factor Analysis), and has later been re-interpreted as one.
- PCA a solution of a maximum likelihood estimation.
- Why Probabilistic PCA?
 - provide a solution to perform a PCA with missing data \rightarrow EM algorithm

$$t_{p \times 1} = W_{p \times q} x_{q \times 1} + \mu_{p \times 1} + \varepsilon_{p \times 1}, \quad q \leq p$$

$$x \sim \mathcal{N}(0, I_q)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$$

$$t \sim \mathcal{N}(\mu, WW' + \sigma^2 I_p)$$

Multiple Factor Analysis

- Multiple factor analysis (MFA *Escofier and Pagès 1982*) : analyze a set of observations described by several groups of variables
- The core of MFA is performed in two steps:
 - a principal component analysis (PCA) is performed on each data set $\rightarrow \lambda_1^1, \lambda_1^2$
 - a global PCA is performed on this matrix $(\frac{X_1}{\sqrt{\lambda_1^1}}, \frac{X_2}{\sqrt{\lambda_1^2}})$

The probabilistic MFA

- How to obtain the first eigenvalues of each subgroup?
 - Use the complete case
 - Use the common data
- PPCA on $\left(\frac{X_1}{\sqrt{\lambda_1^1}}, \frac{X_2}{\sqrt{\lambda_1^2}}\right)$

Simulation framework

- Two configurations
- Number of missing values: 10%, 20%, 30%
- Number of individuals : 20, 100
- Number of variables : 2, 5, 9
- Different data structures
 - $\mathcal{N}(0, \Sigma)$, Σ given (real data set and chosen)
 - $X1 = A + \mathcal{N}(0, \sigma^2)$, $X2 = A + \mathcal{N}(0, \sigma^2)$, $\sigma = 0.1, 0.25, 0.5, 1$

Evaluation

- 200 simulations
- Complete data set from which data is randomly removed.
- Imputation \rightarrow missing value estimation results \rightarrow MSE
- Estimation \rightarrow scores obtained with MFA on the complete data set (Z) and on the imputed data set (Z^c) $\rightarrow RV(Z, Z^c)$

Results (1)

| | mean | AP | AP dil | RV | pMFA2 | pMFA3 |
|-------|-------|-------|--------|--------|-------|--------|
| na= 2 | 0.912 | 0.936 | 0.941 | 0.924 | 0.937 | 0.951 |
| na= 4 | 0.841 | 0.855 | 0.867 | 0.8612 | 0.861 | 0.8788 |
| na= 6 | 0.767 | 0.782 | 0.796 | 0.792 | 0.798 | 0.795 |

Table: $n = 20$, $p = 2$, $\mathcal{N}_{p+q}(0, \Sigma)$

$$\Sigma_{11} = I_p,$$

$$\Sigma_{22} = I_q,$$

$$\Sigma_{12} = \Sigma_{21} = C_{0.5}$$

$C_{0.5}$ a $p \times q$ matrix with all elements begin the real numbers 0.5

Results (2)

| | | | | | | |
|-----------|-------|-------|--------|-------|-------|--------|
| sig = 0,1 | mean | AP | AP dil | RV | pMFA2 | pMFA3 |
| na= 2 | 0,951 | 0,982 | 0,983 | 0,977 | 0,983 | 0,9818 |
| na= 4 | 0,901 | 0,966 | 0,967 | 0,959 | 0,962 | 0,9582 |
| na= 6 | 0,842 | 0,945 | 0,946 | 0,933 | 0,916 | 0,9048 |
| sig = 1 | | | | | | |
| na= 2 | 0,924 | 0,924 | 0,938 | 0,842 | 0,936 | 0,9342 |
| na= 4 | 0,836 | 0,847 | 0,865 | 0,752 | 0,855 | 0,8559 |
| na= 6 | 0,748 | 0,780 | 0,798 | 0,713 | 0,772 | 0,7716 |

Table: Subjacent configuration, $n = 20$, $p = 2$

Conclusion (1)

- Simple imputation:
 - Tolerant to amounts of missing values $\leq 30\%$
 - Good performances

- Distort data distribution and relationship \rightarrow inflate correlation between variables
- Uncertainty measure?
- With more than two groups of variables?

Conclusion (2)

K configurations with missing values

- Procrustes rotations, RV maximisation
 - Two by two
 - For each missing value $\rightarrow K - 1$ estimations available
 - Barycenter, median (if outliers), selection?
 - How to select? \rightarrow measure of the relationship between two data matrices?

Conclusion (3)

- PMFA:
 - Good performances
 - Uncertainty measure
 - Estimate the data and the parameters simultaneously via EM algorithm

- Number of components?
- Model, hypothesis.
- PMFA : tolerant to amounts of missing values $\leq 20\%$

Conclusion (4)

K configurations with missing values

- Probabilistic MFA: a global analysis
 - Take simultaneously into account all configurations
 - Calculate the first eigenvalues
 - Use the common data
 - A global PPCA \rightarrow eigenvalues \rightarrow a PMFA
 - Iterative algorithm: a global PPCA \rightarrow eigenvalues \rightarrow a PMFA \rightarrow eigenvalues \rightarrow a PMFA ...

Procrustes analysis for two configurations

Procrustes was the leader of a band of brigands in Greek mythology. He was in the habit of putting his victims in a bed and to stretch or cut their limbs in such a way that they fit in the bed.

By analogy, procrustes analysis is a method based on rotation, reflection, translation, and dilatation of a set of points in order to fit it to another fixed set of points (*Gower, 1971*).

Procuste analysis for two configurations

- Matches two configuration of points representing the same individuals
- Least square procustean rotation

$$\min \text{Tr}(X_1 - X_2 T)(X_1 - X_2 T)' \text{ with } T' T = I$$

$$\min \text{tr}(X_1 X_1') + \text{tr}(X_2 T T' X_2) - 2 \text{tr}(X_1' X_2 T)$$

$$\max \text{tr}(S_{12} T) \text{ with } T' T = I$$

$$S_{12} = V S U' \text{ with } T = V U'$$

- Similar image of the n individuals
- Procuste rotation to deal with missing values

⇒ Imputing a configuration with the help of the others

Index Lingoes and Schöneman (RLS)

$$RLS(X_1, X_2) = \frac{\text{tr}(S_{12}S_{21})^{1/2}}{\sqrt{\text{tr}(S_{11})\text{tr}(S_{22})}}$$

$$\begin{aligned} d(X_1, X_2) &= \left\| \frac{X_1}{\text{tr}(X_1^T X_1)} - \frac{X_2}{\text{tr}(X_2^T X_2)} \right\| \\ &= \sqrt{2} \sqrt{1 - RLS(X_1, X_2)} \end{aligned}$$

- $0 \leq RLS(X_1, X_2) \leq 1$
- $RLS(X_1, X_2) = 0$ if and only if $X_1^T X_2 = 0$
- if $p = q = 1$ $RLS(X_1, X_2) = |\rho(X_1, X_2)|$
- $X_2 = BX_1 + c$ B is a $q \times p$ matrix such that $B^T B = I_p$ and where c is any $q \times 1$ constant vectors, then $RLS(X_1, X_2) = 1$

$$\frac{1}{\sqrt{pq}} RLS^2 \geq RV \geq \sqrt{pq} RLS^2$$

Distance between data matrices

- $X_1 X_1^T$ relative positions of the n points in \mathbb{R}^p .

$$\begin{aligned}d(X_1, X_2) &= \left\| \frac{X_1 X_1^T}{(\text{tr}(X_1 X_1^T))^2)^{1/2}} - \frac{X_2 X_2^T}{(\text{tr}(X_2 X_2^T))^2)^{1/2}} \right\| \\ &= \sqrt{2} \sqrt{1 - RV(X_1, X_2)}\end{aligned}$$

- $RV = 1$ if and only if $d(X_1, X_2) = 0$.

\Rightarrow Minimization of a distance between two clouds of points = maximize the RV

RV maximisation *F.Crettaz de roten*

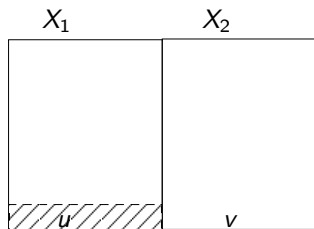


Figure: RV maximization.

$$S_n = \frac{n-2}{n-1} S_{n-1} + \frac{1}{n} (x_n - \bar{x}_{n-1})' (x_n - \bar{x}_{n-1})$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{\sqrt{n}} (x_n - \bar{x}_{n-1})'$$

RV maximisation

Write the RV as a function of the missing row u

$$S_n = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} uu' & uv' \\ vu' & vv' \end{pmatrix}$$

$$RV(u) = \frac{2v'A_{21}u + v'vu'u + tr(A_{12}A_{21})}{\sqrt{(tr(A_{22}^2) + 2v'A_{22}v + (v'v)^2)((u'u)^2 + 2u'A_{11}u + tr(A_{11}^2))}}$$

The probabilistic PCA (2) (Bishop and Tipping, Roweis)

- The model parameters can be solved by maximizing the likelihood:

$$L = \frac{n}{2} p \ln(2\pi) + \ln(|C|) + \text{tr}(C^{-1}S)$$

$$C = WW' + \sigma^2 I_p$$

$$S = \frac{1}{n} \sum_{i=1}^n (t_i - \mu)(t_i - \mu)'$$

- The stationary points can be found analytically as:
 - $\hat{\mu}$ mean of the data.
 - $\hat{W}_{ML} = U_q(\Lambda_q - \sigma^2 I)^{1/2} R$
 - $\hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^p \lambda_j$

The probabilistic PCA (3) (Bishop and Tipping, Roweis)

$$P(x|t) = \frac{P(t|x)P(x)}{P(t)} \sim \mathcal{N}(M^{-1}W'(t - \mu), \sigma^2 M^{-1})$$

$$M = W'W + \sigma^2 I$$

- $\hat{x} = \mathbb{E}[(x|t)] = M^{-1}W'(t - \mu)$
- $\sigma \rightarrow 0 \quad \hat{x} = (W'W)^{-1}W'(t - \mu)$

EM algorithm for PCA (4) (Bishop and Tipping, Roweis)

- In practice, the likelihood can be estimated using an EM algorithm

$$L_C = \sum_{i=1}^n \ln p(t_n, x_n)$$

- E step : estimate the expected values for the latent variables x
 $\mathbb{E}(L_C) \rightarrow \hat{x}$
- M step : find model parameters that maximize the likelihood given the values of $\hat{x} \rightarrow \hat{W}$
- Computational advantages
- Handling with missing data

The RV distribution

- Null hypothesis of no association between the random vectors, significance common structure
- Asymptotic distribution have been obtained when the parent distribution is in the class of elliptical distribution (*Cléroux and Ducharme (1989)*)
- Permutation test (*Kazi Aoual(1993)*)