Overview of methods of analysis of multi-group datasets

Application to the chemical composition of olive oils

A. Eslami[1], E. M. Qannari[2], A. Kohler[3] and S. Bougeard[1]

(1) *French agency for food, environmental and occupational health safety (Anses), Department of Epidemiology, Ploufragan, France*

(2) *Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering (Oniris), Department of Chemometrics and Sensometrics, Nantes, France*

(3) *Norwegian University of Life Sciences, Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Ås, Norway.*

The 12th European Symposium on Statistical Methods for the Food Industry
28-29 February and 1-2 March 2012

Table of contents

Characteristics of multi-group datasets

The same set of variables is measured on individuals which are *a priori* divided into several groups:

- Sensory analysis (panels within countries)
- Epidemiology (animals within farms)
- Environmental studies (plants within sites)

Aim: investigate the relationships among variables while taking account of the group structure.

Risks to ignore the group structure

- PCA is applied to each group separately:
  - Lots of parameters to estimate
  - Difficulty in summing up results
  - Results may be unstable in case of small groups

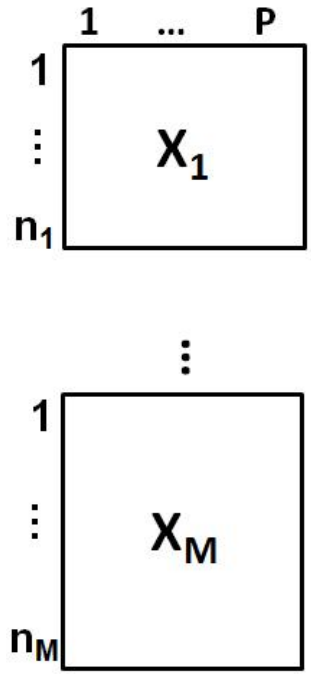
- PCA on the concatenated dataset:
  - Between group variance may dominate the within group variance

Table of contents

Data and notations

- P variables
- N individuals partitioned into M groups known *a priori*, $N = \sum_{m=1}^{M} n_m$
- X_m are assumed to be centered
- V_m is the variance-covariance matrix of group m

$$V_m = \frac{1}{n_m} X_m^T X_m$$



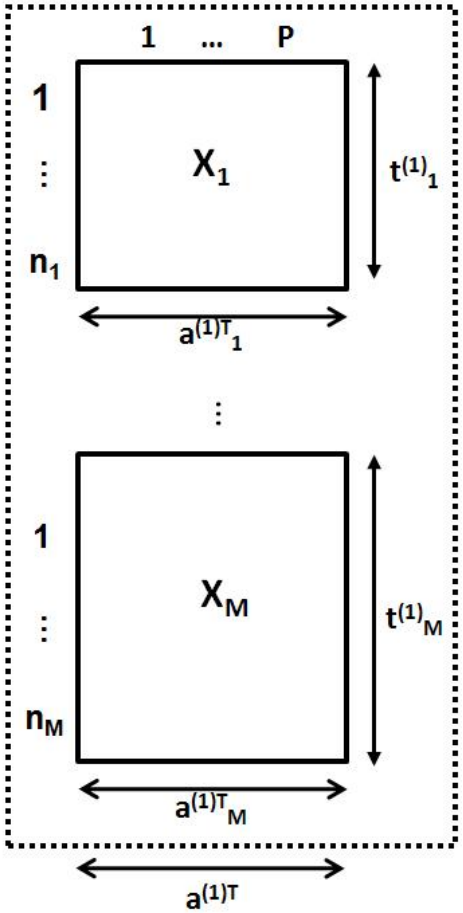
Aim: describe the groups by common characteristics such as common loadings

Data and notations

- P variables
- N individuals partitioned into M groups known *a priori*, $N = \sum_{m=1}^{M} n_m$
- X_m are assumed to be centered
- V_m is the variance-covariance matrix of group m

$$V_m = \frac{1}{n_m} X_m^T X_m$$

- $a^{(h)}$: common vector of loadings associated with dimension $h = (1, \ldots, H)$- $A = [a^{(1)}, \ldots, a^{(H)}]$
- $a_m^{(h)}$: group vector of loadings associated with dimension h and group m
- $t_m^{(h)} = X_m a^{(h)}$: group component



Aim: describe the groups by common characteristics such as common loadings

Flury's common principal components analysis- CPCA

Flury, 1984

CPCA is as a generalization of PCA to several groups under the assumptions:

- The PC in the various groups are constrained to have the same vectors of loadings
- The PCs of each group have specific variances

$$V_m = A\Lambda_m A^T \text{ with } A^T A = I \text{ (identity matrix)}$$

- solution: maximum likelihood estimation by F-G algorithm (Flury and Gautschi 1986)
- advantages: hypothesis testing framework
- restriction: complexity of calculation, multivariate normal assumption, time consuming and may have convergence problems

Remark 1

a vector of loadings associated with group m is a linear combination of the rows of matrix X_m^T:

$$a_m^{(1)} = X_m^T t_m^{(1)}$$

Remark 2

all the methods used within the framework of multi-block datasets can be adopted to multi-group datasets

Stepwise determination of the common vectors of loadings

First optimization problem-Between-groups comparison (BGC)

Find a common vector of loadings a to maximize:

$$\sum_{m=1}^{M} n_m < a_m^{(1)}, a^{(1)} >^2$$

$$\text{with } a_m^{(h)} = (X_m)^T t_m^{(h)} \quad \| a_m \| = 1, \quad \| a \| = 1$$

- refers to generalized canonical correlation analysis (Carroll, 1968)
- solution: $a^{(1)}$ is the eigenvector of $\sum_{m=1}^{M} n_m P_m$: $P_m = X_m^T (X_m X_m^T)^{-1} X_m$
- More straightforward than Between-groups comparison of principal components (Krzanowski, 1979)
- Subsequent vectors of loadings can be determined following the same strategy by adding orthogonality constraints.

Stepwise determination of the common vectors of loadings

Second optimization problem-Multi-group PCA (MGPCA)

Find a common vector of loadings a to maximize:

$$\sum_{m=1}^{M} n_m < a_m^{(1)}, a^{(1)} >^2$$

$$\text{with } a_m^{(h)} = (X_m)^T t_m^{(h)} \quad \| t_m \| = 1, \quad \| a \| = 1$$

- refers: multiple co-inertia analysis (Hanafi et al., 2011)
- solution: PCA on the within groups variance-covariance matrix, $W = \sum_{m=1}^{M} \frac{n_m}{N} V_m$
  - Principal component analysis in the presence of group structure-Krzanowski, 1984
  - Dual multiple factor analysis (Lê et al., 2010)
  - ...

MGPCA criterion is equivalent to

$$\sum_{m=1}^{M} n_m (a^{(1)})^T V_m a^{(1)}$$

This highlights the difference between MGPCA and BGC.

Dual STATIS

DSTATIS, Lavit et al., 1994

Dual STATIS is based on compromise variance-covariance matrix V_c:

$$\min_{V_c} \sum_{m=1}^{M} \| \alpha_m V_m - V_c \|^2$$

$$\text{with } V_c = \sum_{m=1}^{M} \alpha_m V_m, \quad \sum_{m=1}^{M} \alpha_m^2 = 1$$

- solution: α_m: eigenanalysis of $R = (r_{jk})$, where $r_{jk} = trace(V_j V_k)$ for $(j, k = 1, \ldots, M)$ and then eigenanalysis of V_c
- advantages: take account of the similarities between the variance-covariance matrices of the groups

Dual generalized Procrustes analysis
DGPA

Dual generalized Procrustes analysis is based on X_m^T instead of X_m and seek to minimize:

$$\sum_{m=1}^{M} \| \frac{1}{\sqrt{n_m}} X_m^T H_m - C \|^2$$

where $\frac{1}{\sqrt{n_m}} X_m^T$ is rotated to the common matrix C by H_m, the rotation matrix associated with group m.

- solution: Once C is determined, the common vectors of loadings are calculated as the left singular vectors of C.

Comparison criteria

- Criterion 1: assess whether these methods lead to similar vectors of loadings:

Let $A = [a^{(1)}, \ldots, a^{(H)}]$ and $A^* = [a^{*(1)}, \ldots, a^{*(H)}]$ are the common loadings associated with two methods:

$$S^{(h)} = \frac{1}{h} \sum_{r=1}^{h} |(a^{(r)})^T a^{*(r)}|$$

$0 \leq S^{(h)} \leq 1$

$S^{(h)} \to 1$ similar vectors of loadings
$S^{(h)} \to 0$ orthogonal vectors of loadings

Comparison criteria

- Criterion 1: assess whether these methods lead to similar vectors of loadings:

Let $A = [a^{(1)}, \ldots, a^{(H)}]$ and $A^* = [a^{*(1)}, \ldots, a^{*(H)}]$ are the common loadings associated with two methods:

$$S^{(h)} = \frac{1}{h} \sum_{r=1}^{h} |(a^{(r)})^T a^{*(r)}|$$

- Criterion 2: total variance recovered by the principal components

Let $\lambda_m = var(X_m a) = a^T V_m a$

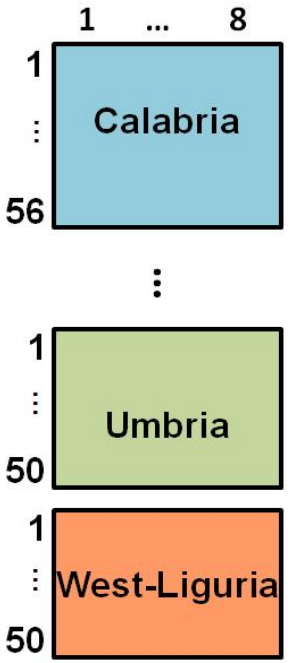
$$I_m^{(h)} = \frac{\lambda_m^{(h)}}{trace(V_m)}$$

Table of contents

Chemical composition of olive oil

Forina et al. (1983)

- ($N = 527$) samples of olive oils from ($M = 9$) regions of Italy
- Variables: ($P = 9$) fatty acids



Objective: describe the regions by common characteristics (common loadings)

Similarity matrix $S^{(h)}$ for dimension$(h = 1, 2)$

$$S^{(h)} = \frac{1}{h} \sum_{r=1}^{h} |(a^{(r)})^T a^{*(r)}|$$

The first dimension (h=1)					
	DGPA	DSTATIS	MGPCA	CPCA	BGC
DGPA	1.000				
DSTATIS	0.998	1.000			
MGPCA	1.000	0.999	1.000		
CPCA	1.000	0.997	1.000	1.000	
BGC	0.978	0.967	0.976	0.980	1.000
Average	**0.994**	**0.990**	**0.993**	**0.994**	**0.975**

The first two dimensions (h=1, 2)					
DGPA	1.000				
DSTATIS	0.998	1.000			
MGPCA	0.999	0.999	1.000		
CPCA	1.000	0.996	0.999	1.000	
BGC	0.967	0.960	0.965	0.966	1.000
Average	**0.991**	**0.988**	**0.990**	**0.990**	**0.965**

For the first two dimensions ($h = 1$ *and* 2), the methods DGPA, MGPCA and CPCA show, on average, the highest similarity with the other methods, whereas BGC shows the least similarity with the other methods. Where $0 \leq S \leq 1$, more similar the methods, the higher value of S.

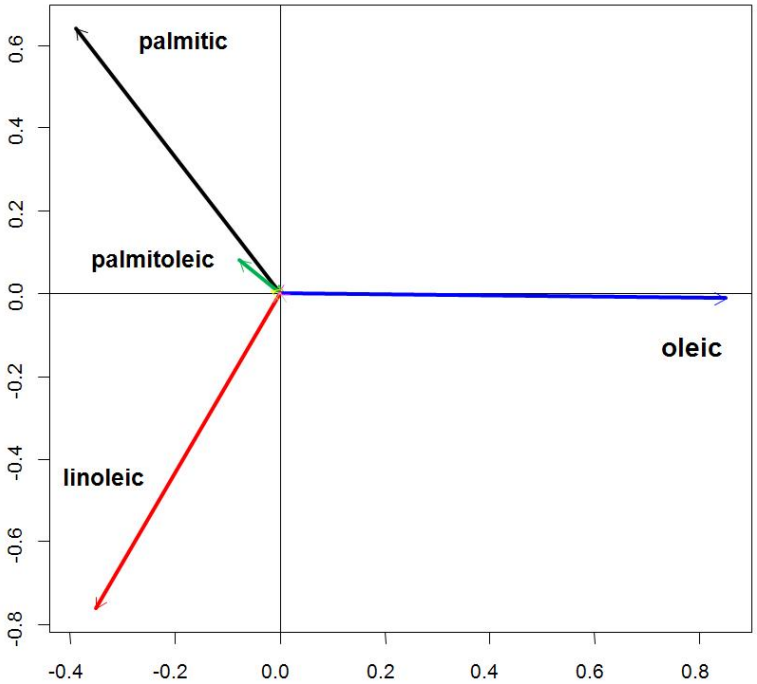
The group percentages of total variance recovered by the principal components

In the first dimension, the only notable difference is indicated by BGC

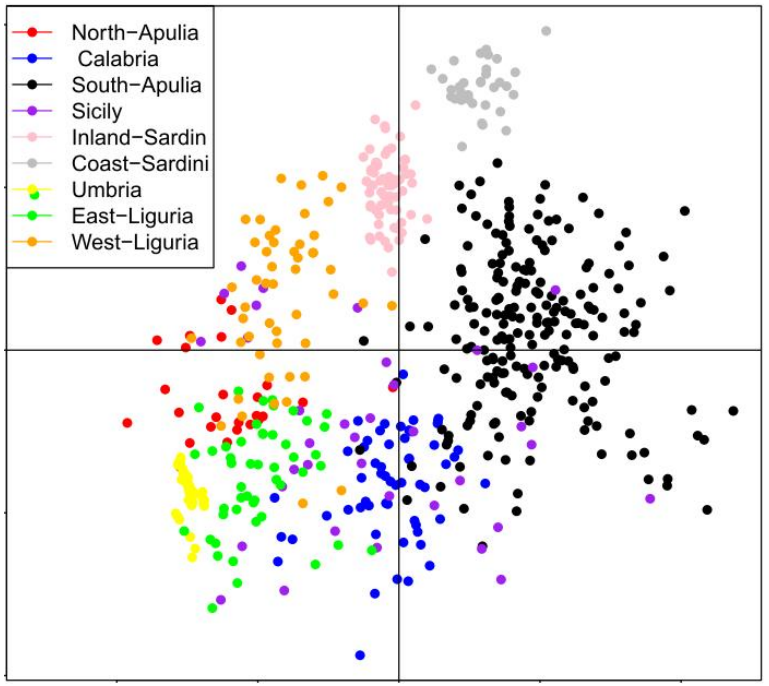
$$I_m^{(h)} = \frac{\lambda_m^{(h)}}{trace(V_m)}$$

The first dimension (h=1)					
Groups Methods	DGPA	DSTATIS	MGPCA	CPCA	BGC
North-Apulia	78.0	76.5	77.7	78.3	79.6
Calabria	82.4	81.1	82.1	82.5	82.3
South-Apulia	79.1	79.0	79.0	78.8	72.9
Sicily	48.2	47.8	48.1	48.2	47.6
Inland-Sardin	87.9	87.7	87.8	88.0	84.2
Coast-Sardini	90.1	90.6	90.2	89.9	84.6
Umbria	82.1	81.9	82.2	82.1	79.0
East-Liguria	28.9	28.8	29.0	28.5	24.5
West-Liguria	59.0	58.1	58.9	59.1	61.0
Average	**70.6**	**70.2**	**70.6**	**70.6**	**68.4**
The second dimension (h=2)					
North-Apulia	14.7	15.6	15.0	14.5	12.8
Calabria	10.7	11.4	10.9	10.5	11.1
South-Apulia	15.6	14.6	15.1	15.9	19.5
Sicily	33.3	32.5	32.9	33.6	31.5
Inland-Sardin	7.0	7.0	6.9	7.1	9.1
Coast-Sardini	7.5	7.5	7.6	7.6	11.5
Umbria	14.7	14.9	14.8	14.4	17.0
East-Liguria	28.8	28.0	28.1	28.4	39.9
West-Liguria	30.1	30.8	30.4	30.0	27.6
Average	**18.1**	**18.0**	**18.0**	**18.0**	**20.0**
Cumulated total variance ($h = 1, 2$)	**88.7**	**88.8**	**88.6**	**88.6**	**88.4**

Loadings and score plots associated with MGPCA method



(a) Common loadings plot



(b) Scores plot

Figure: Graphical display of the common loadings $(a^{(1)}, a^{(2)})$ and individual scores associated with the first two common latent variables $(t^{(1)}, t^{(2)})$ for MGPCA method.

Veterinary epidemiological data

- $N = 884$, $n_m \simeq 120$ animals
- $M = 7$ farms
- $P = 19$ risk factors
- The variables are centered and standardised within groups

The first dimension (h=1)					
	DGPA	DSTATIS	MGPCA	CPCA	BGC
DGPA	1.000				
DSTATIS	0.967	1.000			
MGPCA	0.916	0.984	1.000		
CPCA	0.893	0.817	0.711	1.000	
BGC	0.823	0.678	0.582	0.771	1.000
Average	**0.900**	**0.862**	**0.798**	**0.798**	**0.713**
The first two dimensions (h=1, 2)					
DGPA	1.000				
DSTATIS	0.963	1.000			
MGPCA	0.909	0.983	1.000		
CPCA	0.646	0.521	0.452	1.000	
BGC	0.428	0.346	0.359	0.630	1.000
Average	**0.737**	**0.703**	**0.676**	**0.562**	**0.441**

DGPA, DSTATIS and MGPCA seem to be in relatively high agreement
CPCA and BGC analysis show only a fair agreement between them and with the other methods. Moreover, they explain less variation in the groups.

Table of contents

Conclusions & perspectives

Conclusions

Multi-group datasets analysis:

- hight agreement between: DGPA, DTATIS, MGPCA and CPCA
  - CPCA: multivariate normal assumption
- Different result: BGC
  - does not aim at recovering the total variance in the groups

Perspectives

- Application to discriminant analysis
- Extension to multi-block multi-group data analysis

Conclusions & perspectives

Conclusions

Multi-group datasets analysis:

- hight agreement between: DGPA, DTATIS, MGPCA and CPCA
  - CPCA: multivariate normal assumption
- Different result: BGC
  - does not aim at recovering the total variance in the groups

Perspectives

- Application to discriminant analysis
- Extension to multi-block multi-group data analysis

Overview of methods of analysis of multi-group datasets

Application to the chemical composition of olive oils

A. Eslami[1], E. M. Qannari[2], A. Kohler[3] and S. Bougeard[1]

(1) *French agency for food, environmental and occupational health safety (Anses), Department of Epidemiology, Ploufragan, France*

(2) *Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering (Oniris), Department of Chemometrics and Sensometrics, Nantes, France*

(3) *Norwegian University of Life Sciences, Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Ås, Norway.*

The 12th European Symposium on Statistical Methods for the Food Industry
28-29 February and 1-2 March 2012