Multivariate analysis of multi-group datasets

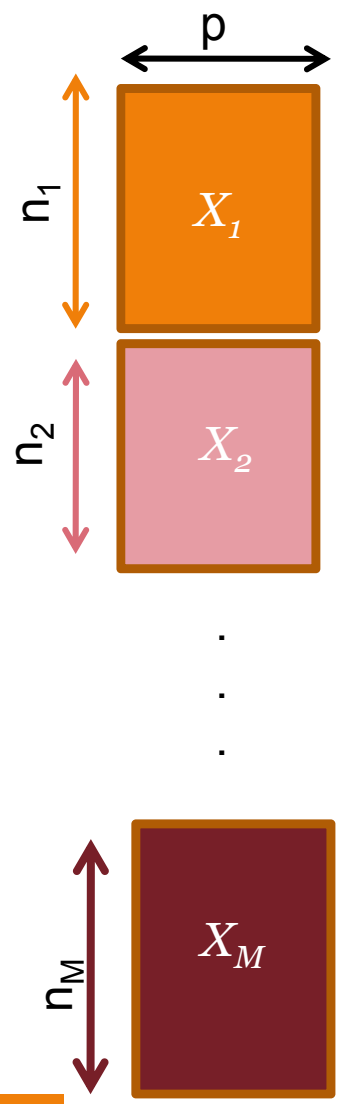
El Mostafa Qannari

Aida Eslami

Achim Kohler

Stéphanie Bougeard

Multi-group datasets



- Setting: the same variables measured on individuals portioned into several groups:
  - Sensory analysis
  - Epidemiology
  - Environmental studies
  - ....

- The same setting as in discriminant analysis but the main aim herein is to investigate the relationships among individuals within the various groups.

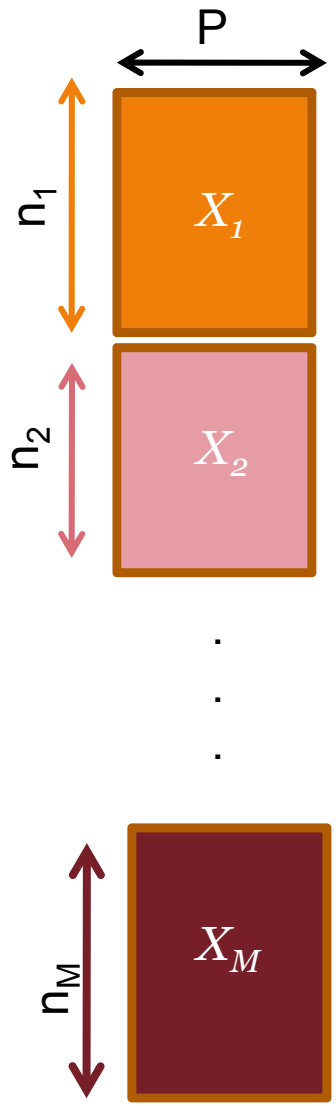
How to investigate the relationships among individuals within the various groups?

- **Perform PCA on each group separately**.
  - Too many parameters (stability and interpretation problems)
- **Perform PCA on the concatenated dataset.**
  - The total variance recovered by the principal components mix up both the between and within groups variances.
- **Common Principal Components Analysis (CPCA, Flury 1984)**
  - The principal components in the various groups are constrained to have the same vectors of loadings.
  - However, Flury's CPCA assumes a multi-normal setting and the algorithm (maximum likelihood estimation) is complex, time consuming and may have convergence problems.

Aim of the talk

- Discuss and compare alternative strategies to Flury's CPCA;
- Discuss their applications in discrimination and classification;
- Outline extensions of the method to investigate relationships between several multi-group datasets

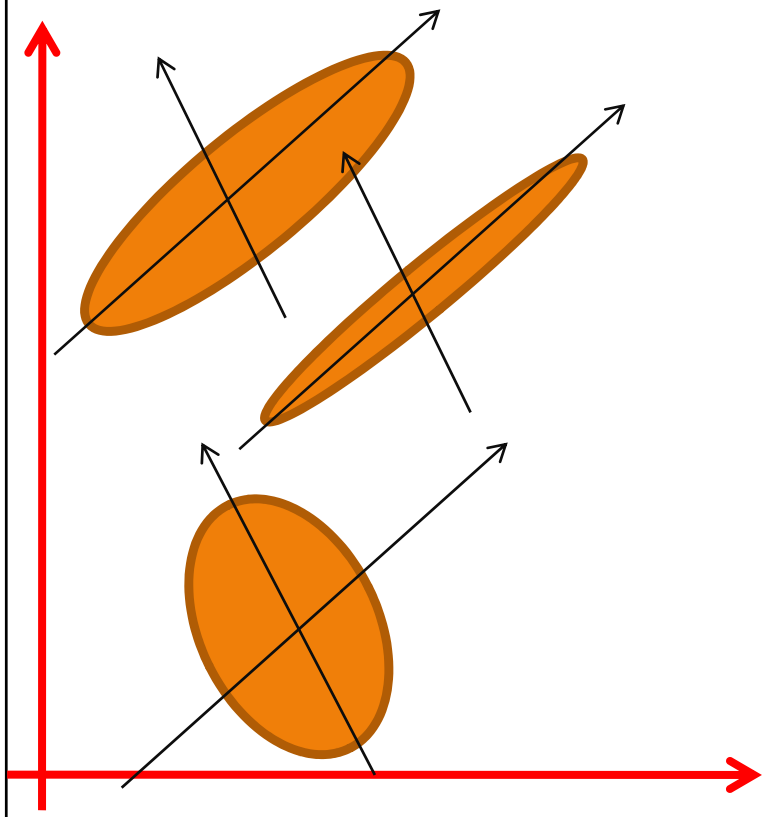
Multi-group datasets



- ✓ *N* individuals partitioned into M groups known a priori.
- ✓X_m are assumed to be centered.
- ✓ V_m : variance covariance matrix in group m.
- ✓ *W*=within group variance covariance matrix :

$$W = \frac{1}{N}\sum n_m V_m$$

CPCA model



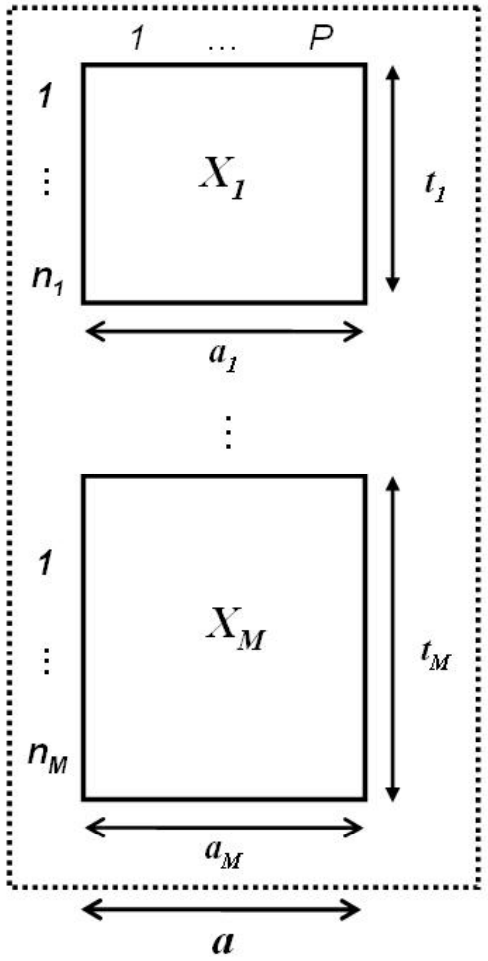
- The dispersion ellipsoids are identically oriented but differently inflated.

$$V_m = A \Lambda_m A^T$$

Flury, B. (1984). Common principal components in k groups. Journal of the American Statistical Association.

Stepwise determination of the common vectors of loadings

Two important remarks



A vector of loadings associated with X_m is given by:

$$a_m = X_m^T t_m$$

Relationship between a (common vector of loadings) and λ_m (specific variance to group m):

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

First optimization problem

- Find a common vector of loadings, a, so as to maximize:

$$\sum_m < a_m , a >^2 \quad with \quad a_m = X_m^T t_m$$

$$\|a_m\| = \|a\| = 1$$

- This is linked to Generalized Canonical Analysis of the spaces spanned by the rows in X_m ($m=1,\ldots,M$).

- Short cut to the method of analysis proposed by Krzanowski (1979): *between-groups comparison of principal components*. *JASA*.

- Subsequent vectors of loadings can be determined following the same strategy + orthogonality constraints.

- Limitations.

Second optimization problem

- Find a common vector of loadings, a, so as to maximize:

$$\sum_m < a_m, a >^2 \quad with \quad a_m = X_m^T t_m$$

$$\|t_m\| = 1 \quad and \quad \|a\| = 1$$

- The solution amounts to performing PCA on the within groups variance covariance matrix, W.
  - Multi-group PCA (Krazanowski, 1984);
  - Dual MFA (Lê et al., 2007)
  - …

Multi-group-PCA

Equivalent criteria to multi-group PCA

- Maximize:

$$\sum_m n_m a^T V_m a \qquad with \quad \|a\| = 1$$

- Maximize:

$$\sum_m n_m \, var(t_m) \qquad with \quad t_m = X_m a \qquad and \qquad \|a\| = 1$$

Third optimization problem

- CPCA model:

$$V_m = A\Lambda_m A^T = \sum_h \lambda_m^{(h)} a^{(h)} a^{(h)T}$$

- Find a rank one approximation (then a rank 2 approximation, etc.):

- Minimize:

$$\sum_{m=1}^{M} \left\| V_m - \lambda_m^{(1)} a^{(1)} a^{(1)T} \right\|^2$$

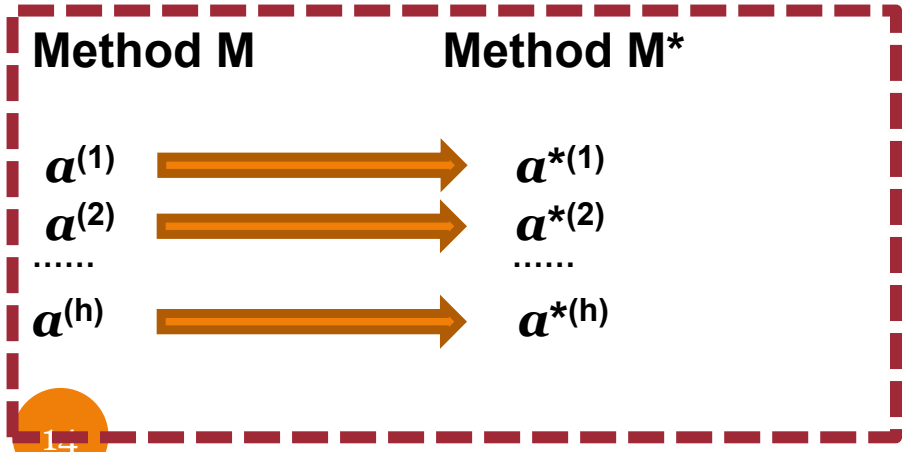
- Pertains to (dual) Common Components and Specific Weights Analysis (Qannari et al., 2000, 2008)

Comparison of methods

Similarity index between two methods

Do the methods lead to parallel common vectors of loadings?

- Given two methods M and M* of determination of the parameters of a *CPCA* model.
- For h=1 to P:

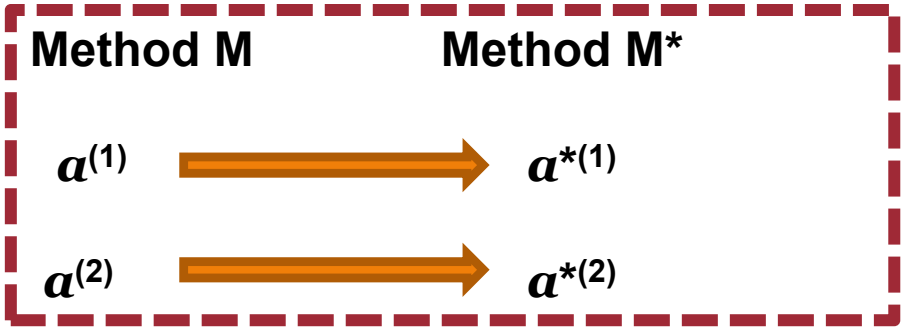


$$S^{(h)} = \frac{1}{h}\sum_{l=1}^{h} \left| \cos\left(a^{(l)}, a^{*(l)}\right) \right|$$

Olive oil data

- Olive oils from 9 regions in Italy.
- Variables : 8 fatty acids.

Similarities between the first two common vectors of loadings



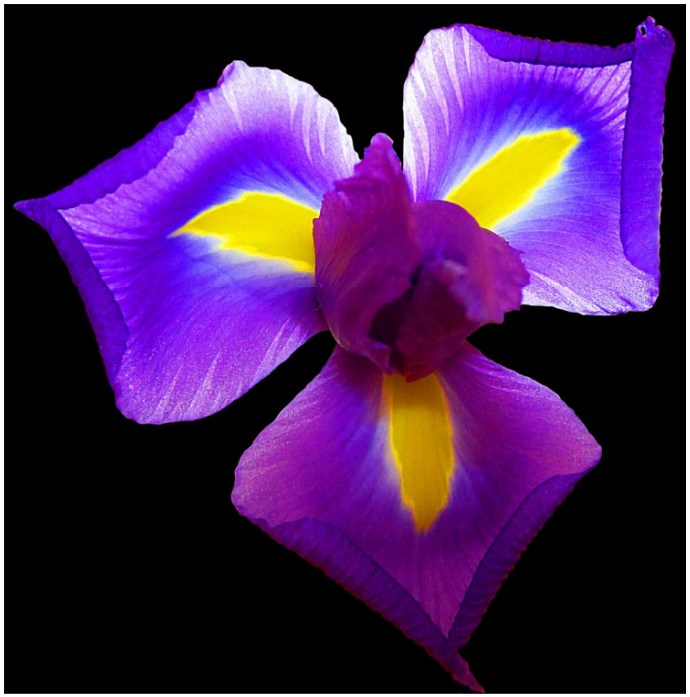
	Dual GPA	Dual CCSW	Dual STATIS	Multi-Grp PCA	Flury's CPCA	Between Grpe Compar.
Dual GPA	1.000					
Dual CCSW	0.997	1.000				
Dual STATIS	0.998	0.999	1.000			
Multi-Grp PCA	0.999	0.999	0.999	1.000		
Flury's CPCA	1.000	0.995	0.996	0.999	1.000	
Between Grpe Compar.	0.967	0.959	0.960	0.965	0.966	1.000

Iris data

50 flowers from each of 3 species.

Four variables:

- length and width of sepals.
- length and width of petals.



Similarities between the first and second common vectors of loadings

	Dual GPA	Dual CCSW	Dual STATIS	Multi-Grp PCA	Flury's CPCA	Between Grpe Compar.
Dual GPA	1.00					
Dual CCSW	0.98	1.00				
Dual STATIS	0.98	0.96	1.00			
Multi-Grp PCA	0.99	1.00	0.96	1.00		
Flury's CPCA	0.94	0.98	0.89	0.97	1.00	
Between Grpe Compar.	0.93	0.92	0.98	0.91	0.83	1.00

Application in discrimination and classification

A strategy of analysis halfway between QDA and SIMCA

QDA:

Use V^{-1}_k

$$USE$$

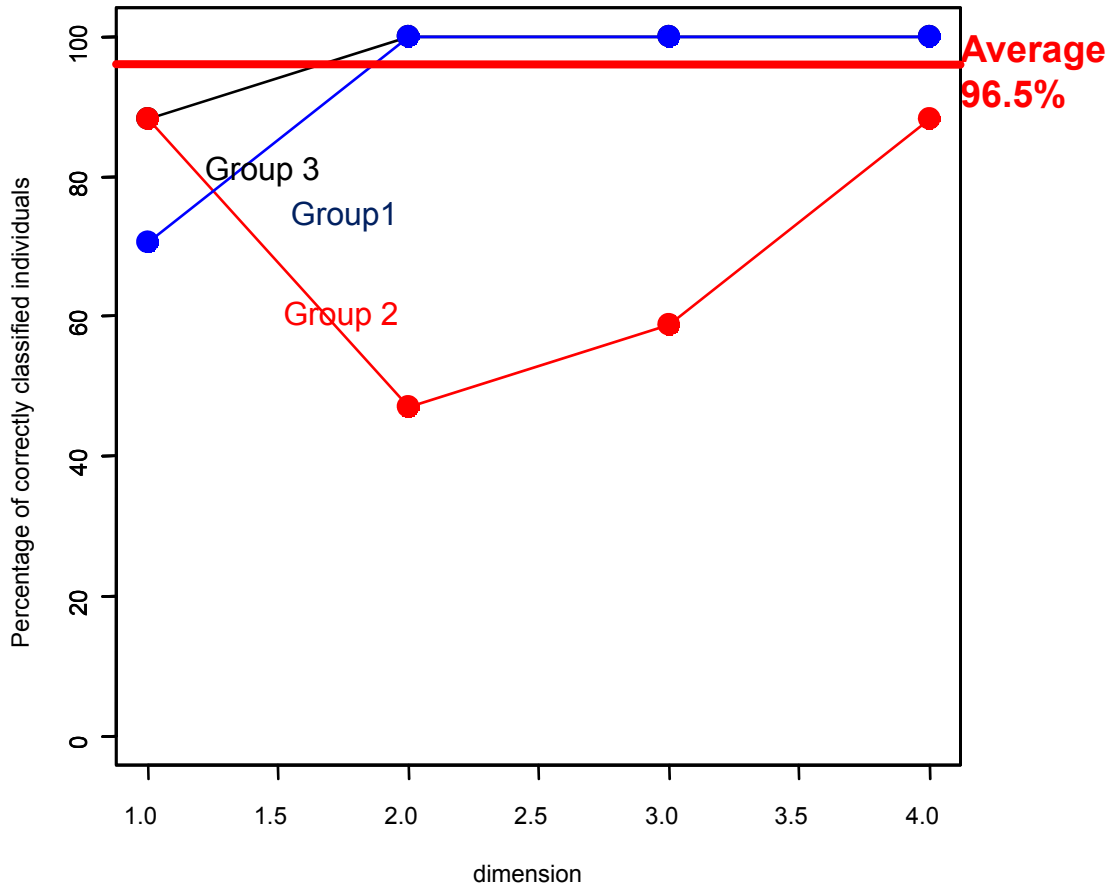
$$A\Lambda_m^{-1}A^T$$

SIMCA:

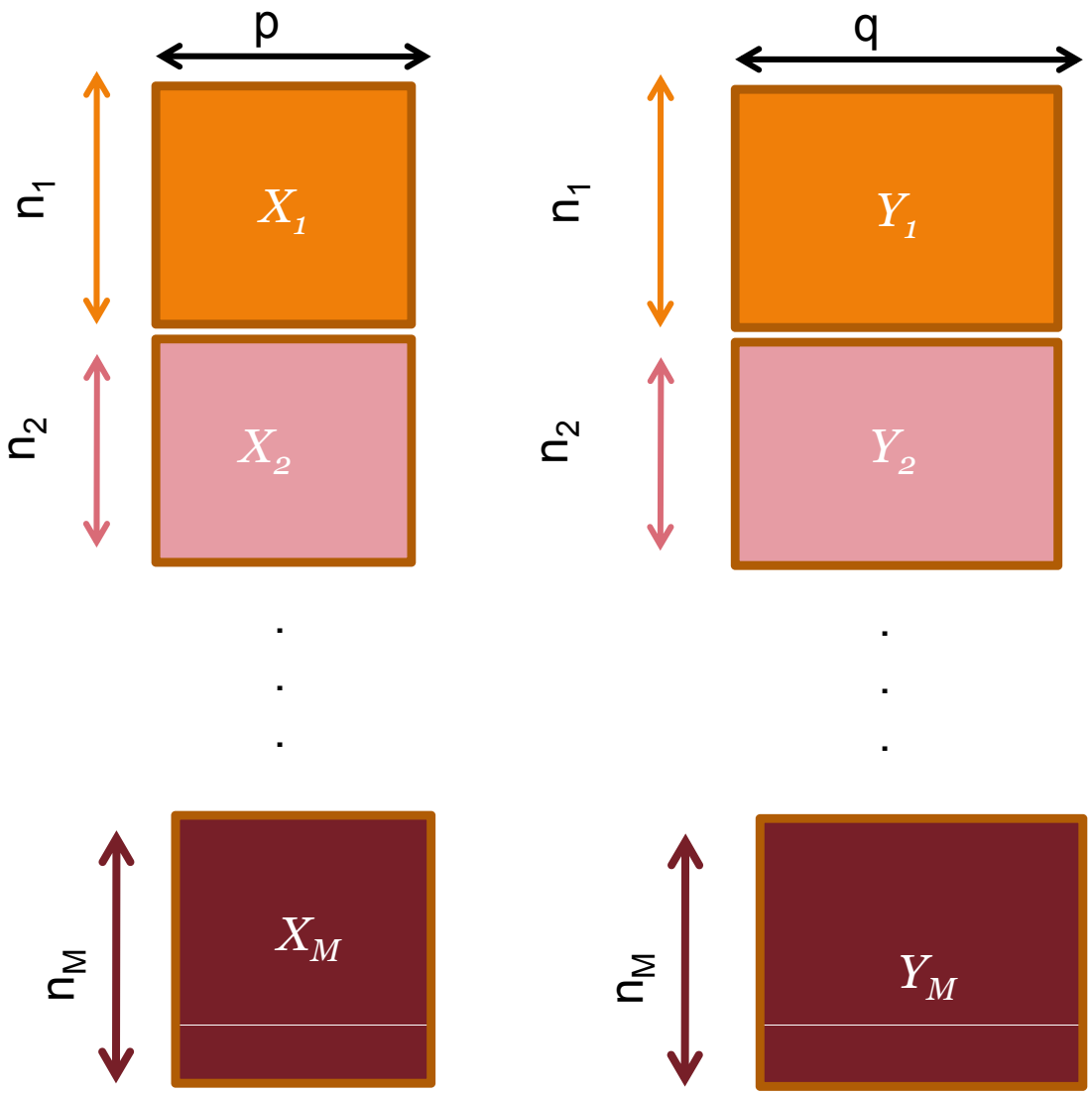
Perform a PCA on each group

Extensions

Application to Iris data



Two block multi-group datasets



First optimization problems

$$\sum_m n_m \, cov^2(X_m a, Y_m b_m) \qquad \|X_m a\| = 1 \quad and \quad \|b_m\| = 1$$

Multi-group redundancy analysis

$$Maximize \quad \sum_m n_m \frac{a^T X_m^T Y_m Y_m^T X_m a}{a^T X_m^T X_m a}$$

Kiers, H. (1995). Maximization of sums of quotients of quadratic-forms and some generalizations. Psychometrika,

Second optimization problems

$$\sum_m n_m \, cov^2(X_m a, Y_m b_m) \qquad \|a\| = 1 \quad and \quad \|b_m\| = 1$$

Multi-group PLS2 analysis

$$Maximize \quad \sum_m n_m a^T X_m^T Y_m Y_m^T X_m a$$

Eigen-analysis solution

Conclusion

- In multivariate data analysis, the complexity may arise from the structure of the data.
- There are still interesting aspects of multi-group data analysis to be investigated:
  - In connection with discrimination
  - Multi-level multi-group datasets
- Move on to multi-block multi-group data analysis.

谢