General overview of methods of analysis of multi-group datasets

Aida Eslami*, El Mostafa Qannari**
Achim Kohler***, Stéphanie Bougeard*

*French Agency for Food, Environmental and Occupational Health Safety,
BP53, F-22440, Ploufragan, France
aida.eslami@anses.fr, stephanie.bougeard@anses.fr,
http://www.anses.fr
** LUNAM University, ONIRIS, Sensometrics and Chemometrics Laboratory,
Nantes, F-44307, France; INRA, Nantes, F-44307, France
elmostafa.qannari@oniris-nantes.fr
http://www.oniris-nantes.fr
*** Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences
and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway
achim.kohler@umb.no

Abstract. Methods of analysis of a dataset where the individuals are partitioned into groups are discussed. These methods encompass known strategies of analysis and a new method called dual generalized Procrustes analysis. The emphasis is put on how the methods used in the context of multi-block data analysis can be adapted to the present context of multi-group setting. The similarities and the differences between the various approaches of analysis are highlighted and illustrated on the basis of three datasets.

1 Introduction

Very often, it occurs that the same J variables are measured on a set of individuals partitioned in M groups. We shall refer to this setting as multi-group datasets. In order to investigate the structure of the data in the groups, principal components analysis (PCA) (Jolliffe, 2002), which is an extensively used tool for the reduction of the dimensionality in multivariate analysis, can be performed on each group separately. Clearly, this strategy of analysis yields a large number of parameters which is likely to lead to an instability problem of the solution because of a lack of sufficient data to accurately estimate all the parameters. Moreover, this strategy of analysis entails a difficulty in interpreting the outcomes and in comparing the results across the groups. It is also possible to perform PCA on the concatenated dataset where the rows refer to the individuals from all the groups. However, in this case the total variance recovered by the principal components mix up both the between and within-group variances.

In order to counteract these problems, several procedures have been proposed using more parsimonious models than separate PCA on the M groups. For instance, common principal