Chapter 5

On the analysis of grouped data

Contents

Objectives

- ✓ To learn how to form groups when the data set is compositional.
- ✓ To introduce how to calculate the linear discriminant function as a log-contrast.
- ✓ To properly analyse the difference between the centres of several groups using the MANOVA test.

5.1. Cluster analysis

The family of statistical *cluster analysis* methods focuses on the question: according to the values taken by the collected samples, can we make data groups? The answer to this question has application in many fields, including ecology, medicine, marketing, and social sciences. Broadly speaking, clustering methods are a wide range of multivariate methods to make *heterogeneous* groups or clusters composed of *homogeneous* samples. That is, we want samples which belong to the same cluster to be *similar*, and they should be *dissimilar* from samples in other groups. But, what does *similar* and *dissimilar* mean? Fig. 5.1 shows one possible strategy. The idea is to evaluate the homogeneity in each group when *comparing* samples with the corresponding mean of the group; and evaluate the heterogeneity among groups when *comparing* the group means with the overall mean. With this approach, one "simply" needs to decide how to make these *comparisons*.

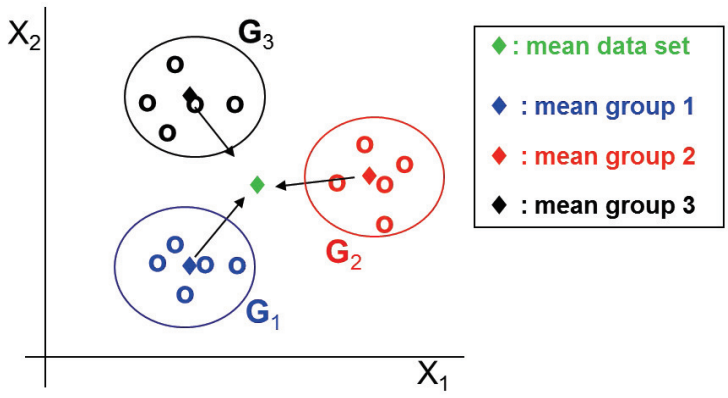

Figure 5.1. One strategy for evaluating heterogeneity among groups.

However, this strategy is not unique. A preliminary classification of clustering methods could be:

- *Parametric or model-based* methods: a probability distribution model is assumed for the data (e.g. normality of the *olr*-coordinates). The techniques to make groups are usually based on maximum likelihood methods [CMM16].