Chap 11: Exploratory Multivariate Analysis

Modern Applied Statistics

Seoul National University

Outline

Introduction

Introduction

Multivariate analysis is concerned with datasets that have more than one response variable for each observational or experimental unit.

Data matrices : $X : n \times p$

- row : observation
- columns : variables
- x : the variables of a case by the row vector

Introduction

In pattern-recognition terminology the distinction is between **supervised** and **unsupervised** methods.

- **Example. iris data**
  - Iris data is collected by Anderson (1935) and given and analysed by Fisher (1936). This has 150 cases, which are stated to be 50 of each of the three species *Iris setosa, virginica and versicolor*. Each case has four measurements on the length and width of its petals and sepals.
  - **supervised** : to classify it to use measurements on a future case , and perhaps to ask how the variables vary among the species.
  - **unsupervised** : to identify species by grouping the cases.

Introduction

Running example 1: *Iris* data

- iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.
- Variables : Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species

Introduction

Running example 2: *Leptograpsus variegatus* crabs data

- Leptograpsus variegatus crabs data set gives body measurements, respectively, for 50 crabs from each of 2 species and 2 sex of crabs
- The species are blue and oragne.
- Variables : sp(B / O), sex(M / F), FL(frontal lobe size), RW(rear width), CL(carapace length), CW(carapace width), BD(body depth)

Visualization Methods

Visualization Methods

There are many **visualization methods** that can be viewed as projection methods for particular definitions of 'interestingness'.

Visualization Methods

- Principal component analysis
- Exploratory projection pursuit
- Distance methods
- Self-Organizing maps
- Biplots
- Independent component analysis
- Glyph Representations
- Parallel coordinate plots

Principal Component Analysis

1. **Principal component analysis(PCA)**
   - **PCA** has a number of different interpretations.
   - The simplest is a projection method finding projections of maximal variability. That is, it seeks linear combinations of the columns of X with maximal (or minimal) variance.

Principal Component Analysis

- Let S denote the covariance matrix of the data X

$$nS = (X - n^{-1}\mathbf{1}\mathbf{1}^T X)^T (X - n^{-1}\mathbf{1}\mathbf{1}^T X) = (X^T X - n\bar{x}\bar{x}^T)$$

$$\text{where } \bar{x} = \mathbf{1}^T X/n$$

- Then the sample variance of a linear combination $\mathbf{xa}$ of a row vector $\mathbf{x}$ is $\mathbf{a^T}\Sigma\mathbf{a}$ and this is to be maximized (or minimized) subject to $\|\mathbf{a}\|^2 = \mathbf{a^T a} = 1$.
- Since Σ is a non-negative definite, it has an eigendecomposition

$$\Sigma = C^T \Lambda C$$

where Λ is a diagonal matrix of (non-negative) eigenvalues in decreasing order.

Principal Component Analysis

- Let $\mathbf{b} = C\ \mathbf{a}$, then $\|\mathbf{b}\|^2 = \|\mathbf{a}\|^2$ ($\because$ C is orthogonal)
- The problem is then equivalent to maximizing $\mathbf{b^T}\Lambda\mathbf{b} = \sum \lambda_i b_i^2$ subject to $\sum b_i^2 = 1$.
- Clearly the variance is maximized by taking $\mathbf{b}$ to be the first unit vector, or equivalently taking $\mathbf{a}$ to be the column eigenvector corresponding to the largest eigenvalue of Σ.
- The i th principal component is then the i th linear combination picked by this procedure.

Principal Component Analysis

- The first k principal components span a subspace containing the 'best' k-dimensional view of the data.
  It has a maximal covariance matrix.
- It also best approximates the original points with minimizing the sum of squared distances from the points to their projections.

Principal Component Analysis

- The first few principal components are often useful to reveal structure in the data.
- The principal components corresponding to the smallest eigenvalues are the most nearly constant combinations of the variables, and can also be of interest.

Principal Component Analysis

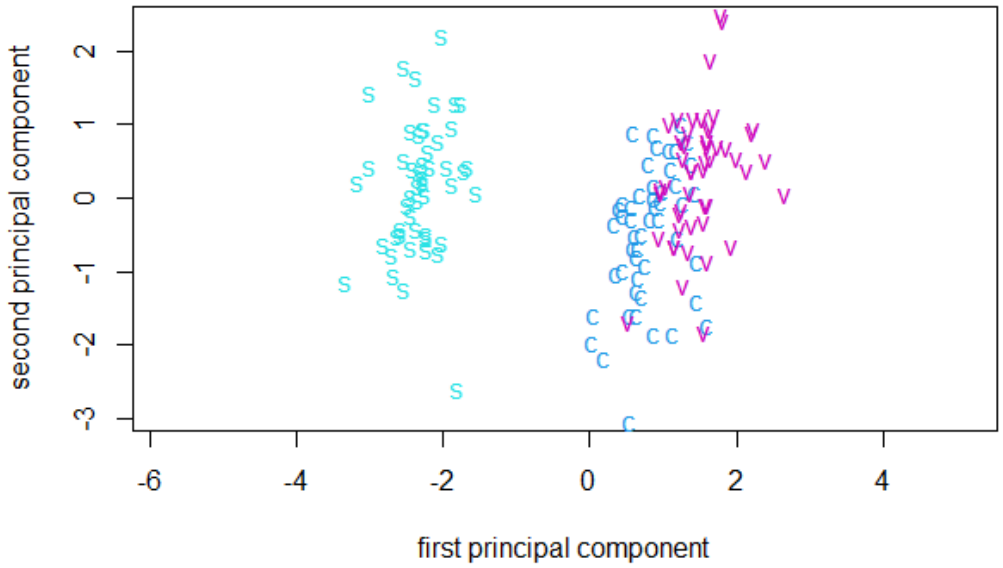

Figure 1: First two principal components for the log-transforrmed iris data.

Figure 1 shows the first two principal components for the **iris** data based on the covariance matrix, revealing the group structure if it had not already been known.

Principal Component Analysis

A *warning*: principal component analysis will reveal the gross features of the data, which may already be known, and is often best applied to residuals after the known structure has been removed.

2. Exploratory projection pursuit

- Using projection pursuit in XGobi or GGobi allows us to examine the data much more thoroughly.

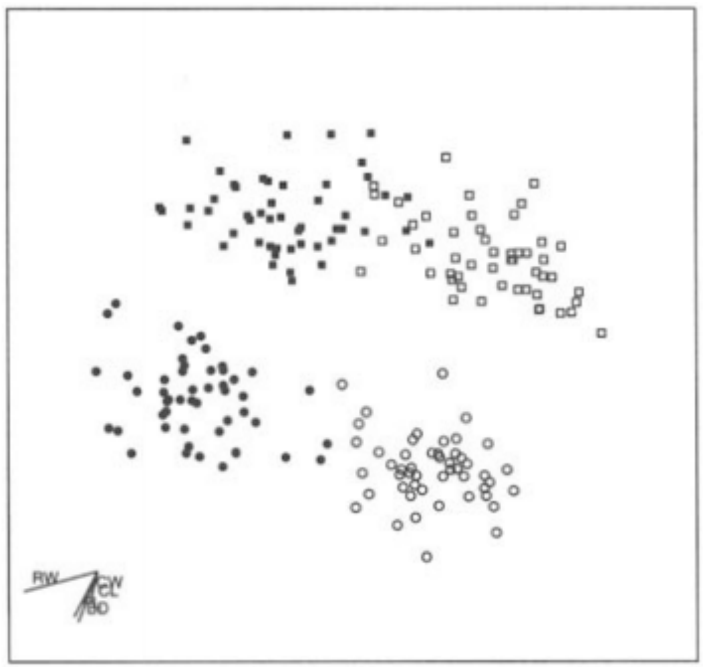

Figure 2: Projection pursuit view of the crabs data. Males are coded as filled symbols, females as open symbols, the blue colour form as squares and the orange form as circles.

3. Distance methods

- This is a class of methods based on representing the cases in a low-dimensional Euclidean space so that their proximity reftects the similarity of their variables.
- We can think of 'squeezing' a high-dimensional point cloud into a small number of dimensions (2, perhaps 3) whilst preserving as well as possible the inter-point distances.
- To do so we have to produce a measure of (dis)similarity.

Distance methods

Distances are often called *dissimilarities*. Jardine and Sibson (1971) discuss several families of similarity and dissimilarity measures.

- For categorical variables most dissimilarities are measures of agreement.
- The simple matching coefficient is the proportion of categorical variables on which the cases differ.
- The Jaccard coefficient applies to categorical variables with a preferred level. It is the proportion of such variables with one of the cases at the preferred level in which the cases differ.
- The binary method of dist is of this family, being the Jaccard coefficient if all non-zero levels are preferred. Applied to logical variables on two cases it gives the proportion of variables in which only one is true among those that are true on at least one case.

Distance methods

The most obvious of the distance methods is multidimensional scaling (MDS), which seeks a configuration in $\mathbb{R}^d$ such that distances between the points best match (in a sense to be defined) those of the distance matrix.

Distance methods

For the iris data:

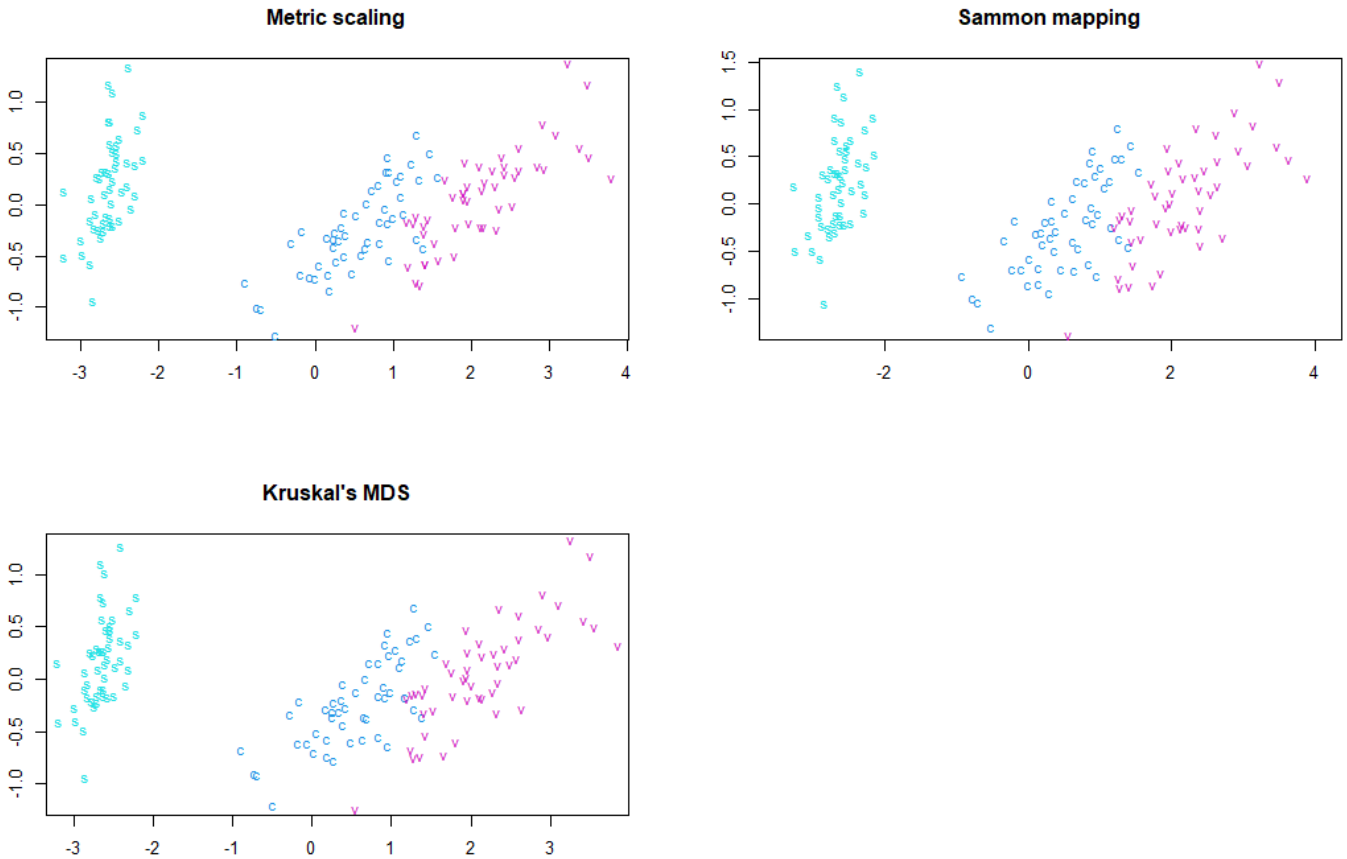

Figure 3: Distance-based representations of the iris data. The top left plot is by multidimensional scaling, the top right by Sammon's non-linear mapping, the bottom left by Kruskal's isotonic multidimensional scaling. Note that each is defined up to shifts, rotations and reflections.

Distance methods

- Note that a configuration can be determined only up to translation, rotation and reflection, since Euclidean distance is invariant under the group of rigid motions and reflections.
- Using classical multidimensional scaling with a **Euclidean distance** as here is equivalent to plotting the first k principal components (without rescaling to correlations).

Distance methods

- Sammon's (1969) non-linear mapping
  - It is given a dissirnilarity d on n points constructs a k-dimensional configuration with distances $\widetilde{d}$ to minimize a weighted 'stress'

$$E_{\text{Sammon}}\left(d, \tilde{d}\right) = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{\left(d_{ij} - \tilde{d}_{ij}\right)^2}{d_{ij}}$$

  - We have to drop duplicate observations to make sense of $E(d, \widetilde{d})$

- classical MDS
  - It is applied to a Euclidean configuration of points (but not in general), which minimizes

$$E_{\text{classical}}(d, \widetilde{d}) = \sum_{i \neq j} \left[d_{ij}^2 - \widetilde{d}_{ij}^2\right] / \sum_{i \neq j} d_{ij}^2$$

Distance methods

- A more thoroughly non-metric version of multidimensional scaling goes back to Kruskal and Shepard in the 1960s.
- The idea is to choose a configuration to minimize

$$\text{STRESS}^2 = \sum_{i \neq j} \left[\theta\left(d_{ij}\right) - \widetilde{d}_{ij} \right]^2 / \sum_{i \neq j} \widetilde{d}_{ij}^2$$

  over both the configuration of points and an increasing function θ.
- Now the location, rotation, reflection and scale of the configuration are all indeterminate.
- The optimization task is difficult and this can be quite slow.

MDS plots of the crabs data

- MDS plots of the `crabs` data tend to show just large and small crabs, so we have to remove the dominant effect of size.
- We used the carapace area as a good measure of size, and divided all measurements by the square root of the area.
- It is also necessary to account for the sex differences, which we can do by analysing each sex separately, or by subtracting the mean for each sex.

Distance methods

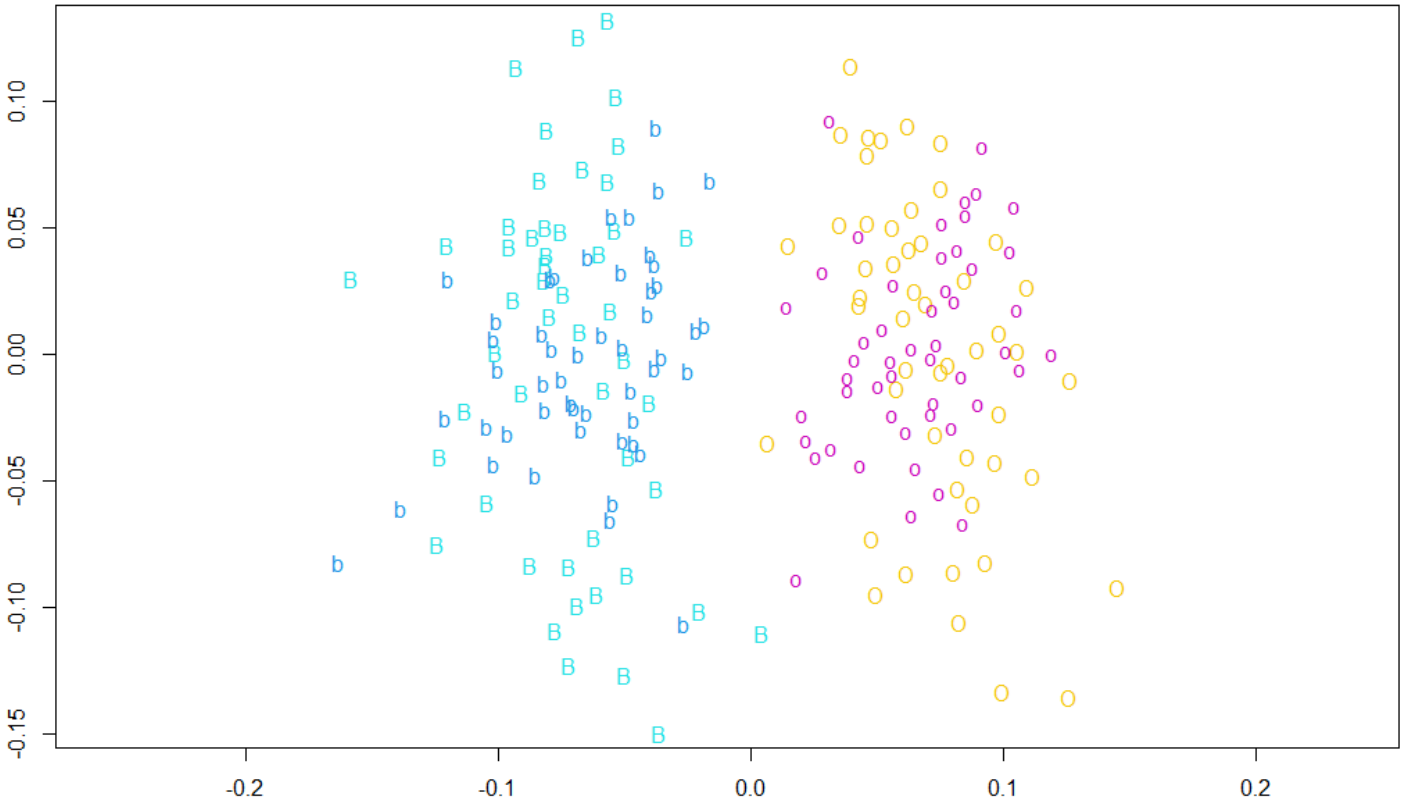

Figure 4: Sammon mapping of crabs data adjusted for size and sex. Males are coded as capitals, females as lower case, colours as the initial letter of blue or orange.

4. Self-Organizing maps

All multi dimensional scaling algorithms are slow, not least because they work with all the distances between pairs of points and so scale at least as $O(n^2)$ and often worse. Engineers have looked for methods to find maps from many more than hundreds of points, of which the best known is 'Self-Organizing maps' (Kohonen, 1995).

Self-Organizing maps

▶ Kohonen's SOM

- Kohonen wanted an algorithm that would effectively map similar patterns (pattern vectors close to each other in the input signal space) onto contiguous locations in the output space.
- However, he interpreted 'contiguous' *via* a rectangular or hexagonal 2-D lattice of representatives $\mathbf{m}_j$, with representatives at nearby points on the grid that are more similar than those that are widely separated.
- Data points are then assigned to the nearest representative (in Euclidean distance).
- Since Euclidean distance is used, pre-scaling of the data is important.

- **Batch SOM**
  - If all the data are available at once, the preferred method is batch SOM (Kohonen, 1995, §3.14).
  - For a single iteration, assign all the data points to representatives, and then update all the representatives by replacing each by the mean of all data points assigned to that representative or one of its neighbours (possibly using a distance-weighted mean).
  - The algorithm proceeds iteratively, shrinking the neighbourhood radius to zero over a small number of iterations.

Self-Organizing maps

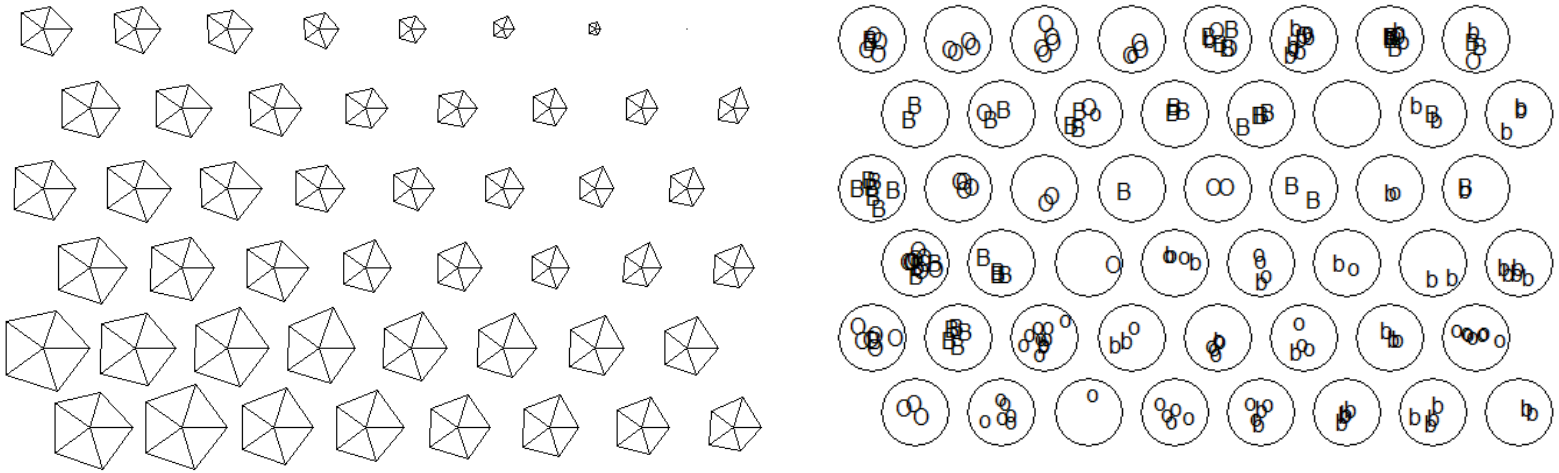

Figure 5: Batch SOM applied to the crabs dataset. The left plot is a stars plot of the representatives, and the right plot shows the assignments of the original points, coded as in 11.4 and placed randomly within the circle. (Plots from R)

The initialization used is to select a random subset of the data points. Different runs give different patterns but do generally show the gradation for small to large animals shown in the left panel of Figure 5.

- **Traditional SOM**

  - Traditional SOM uses an on-line algorithm, in which examples are presented in turn until convergence, usually by sampling from the dataset.
  - Whenever an example x is presented, the closest representative $\mathbf{m}_j$ is found.
  - Then

$$\boldsymbol{m}_i \leftarrow \boldsymbol{m}_i + \alpha\,[\boldsymbol{x} - \boldsymbol{m}_i] \qquad \text{for all neighbours } i$$

    Both the constant α and the definition of 'neighbour' change with time.

5. Biplots

- The biplot (Gabriel, 1971) is a method to represent both the cases and variables.
- Suppose that X has been centered to remove column means.
- The biplot represents X by two sets of vectors of dimensions n and p producing a rank-2 approximation to X.
- The best (in the sense of least squares) such approximation is given by replacing Λ in the singular value decomposition of X by D, a diagonal matrix setting λ_3, ' .. to zero, so

$$X \approx \tilde{X} = [\boldsymbol{u}_1 \boldsymbol{u}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \end{bmatrix} = GH^T$$

where the diagonal scaling factors can be absorbed into G and H in a number of ways.

Biplots

- For example, we could take

$$G = n^{a/2}\,[\boldsymbol{u}_1\boldsymbol{u}_2]\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}^{1-\lambda}, \quad H = n^{-a/2}\,[\boldsymbol{v}_1\boldsymbol{v}_2]\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}^{\lambda}$$

- The biplot then consists of plotting the $n + p$ two-dimensional vectors that form the rows of G and H.
- The interpretation is based on inner products between vectors from the two sets, which give the elements of $\widetilde{X}$. For $\lambda = a = 0$ this is just a plot of the first two principal components and the projections of the variable axes.

Biplots

- The most popular choice is $\lambda = a = 0$ (which Gabriel, 1971, calls the principal component biplot).
- Then G contains the first two principal components scaled to unit variance.
- So the Euclidean distances between the rows of G represent the Mahalanobis distances between the observations and the inner products between the rows of H represent the covariances between the (possibly scaled) variables (Jolliffe, 1986, pp. 77-8)
- Thus the lengths of the vectors represent the standard deviations.

Biplots

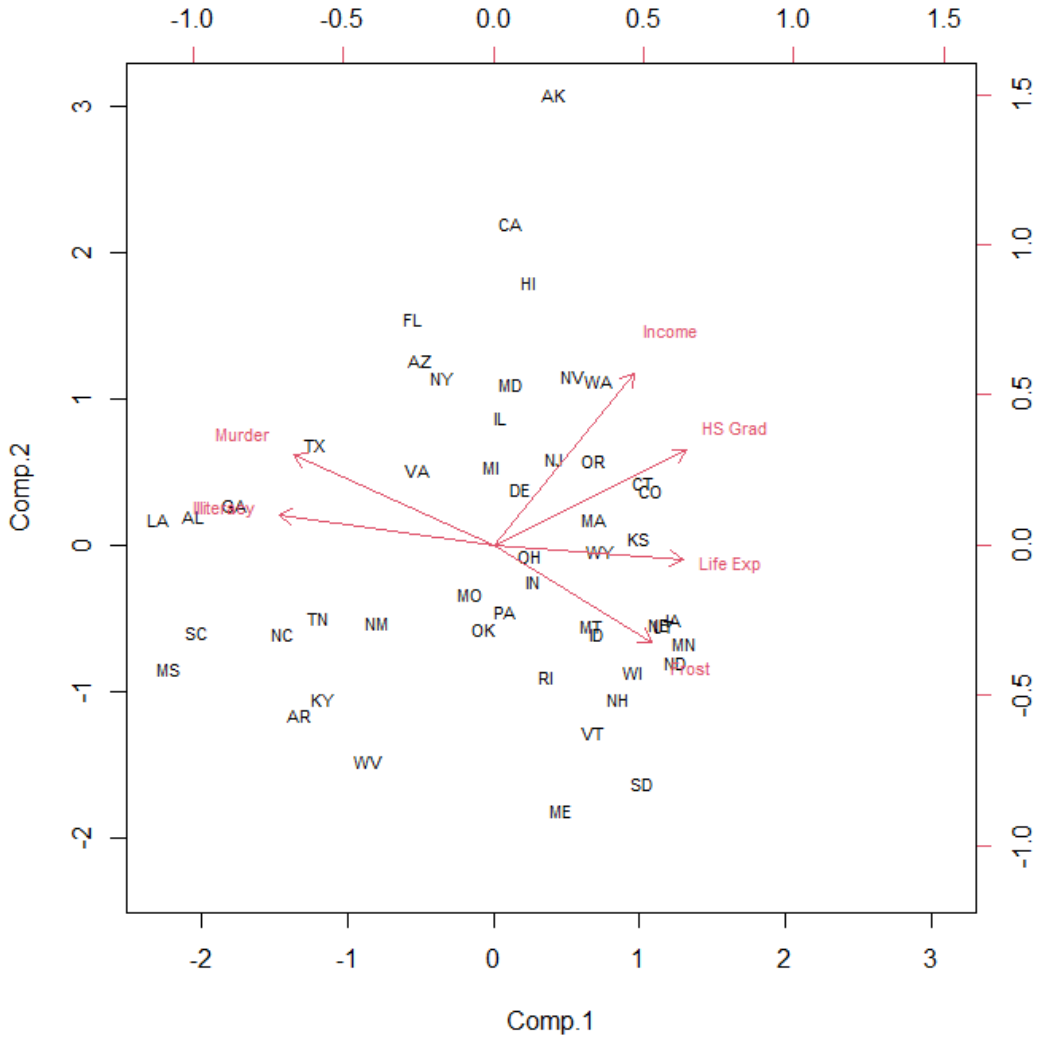

Figure 6: Principal component biplot of the part of the state data. Distances between states represent Mahalanobis distance, and inner products between variables represent correlations. (The arrows extend 80% of the way along the variable's vector.)

6. Independent Component Analtsis

- Independent Component Analtsis(ICA) looks for rotations of sphered data that have approximately independent coordinates.
- This will be true (in theory) for all rotations of samples from multivariate normal distribution, so ICA is of most interest for distributions that are far from normal.

Independent component analysis

The original context for ICA was 'unmixing' of signals.

- Suppose there are $k \leq p$ independent sources in a data matrix S, and we observe the p linear combinations $X = SA$ with mixing matrix A.
- The 'unmixing' problem is to recover S.
- But there are identifiability problems that we cannot recover the amplitudes or the labels of the signals.
- So we may as well suppose that the signals have unit variances.

Unmixing is often illustrated by the problem of listening to just one speaker at a party. Note that this is a 'no noise' model: all the randomness is assumed to come from the signals.

Independent component analysis

- Suppose the data X have been sphered; by assumption S is sphered and so X has variance $A^T A$ and we look for an orthogonal matrix A.
- Thus ICA algorithms can be seen as exploratory projection pursuit in which the measure of interestingness emphasises independence (not just uncorrelatedness), say as the sum of the entropies of the projected coordinates.

Independent component analysis

We can illustrate this for the `crabs` data, where the first and fourth signals shown in Figure 7 seem to pick out the two colour forms and two sexes respectively.

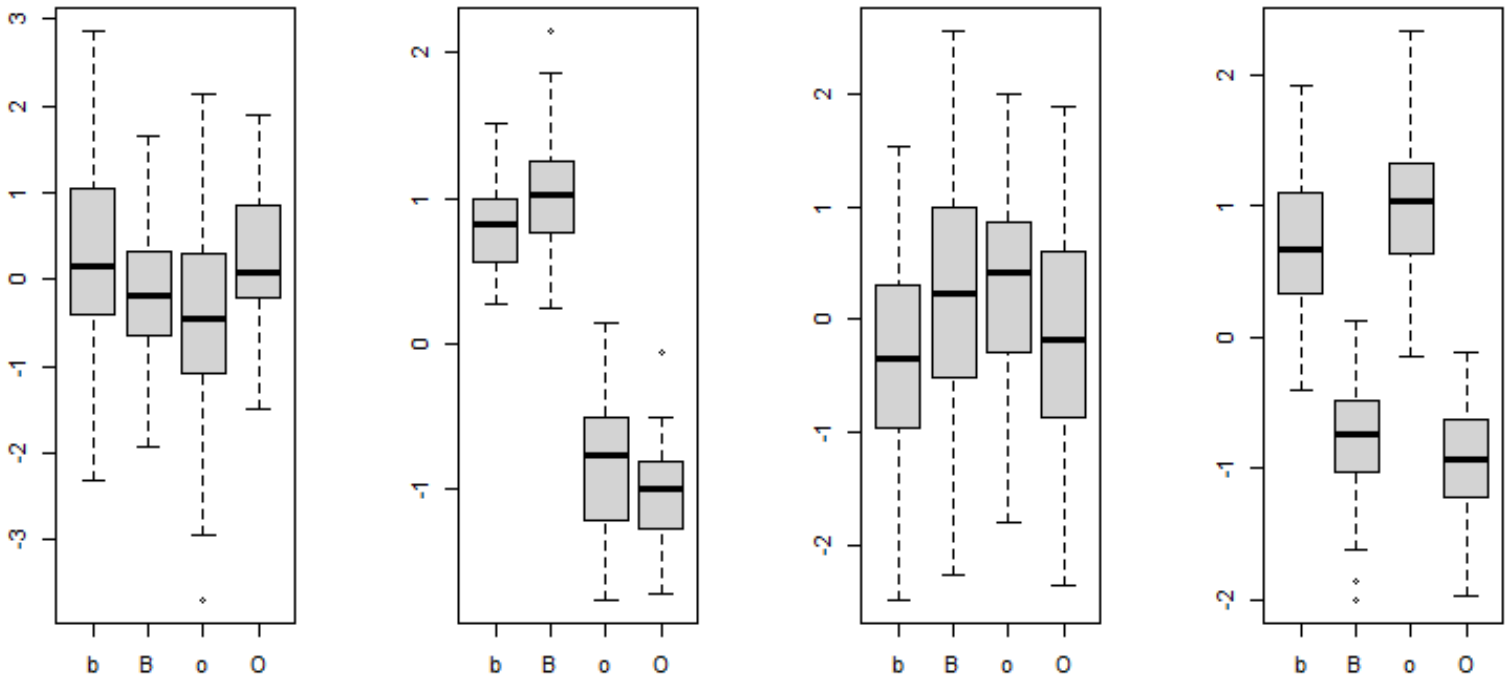

Figure 7: Boxplots of four 'signals' recovered by ICA from the crabs data.

7. Glyph Representations

- There is a wide range of ways to trigger multiple perceptions of a figure, and we can use these to represent each of a moderately large number of rows of a data matrix by an individual figure.
- The best known of these are Chernoff's faces and the star plots.
- they do depend on the ordering of variables and perhaps also their scaling, and they do rely on properties of human visual perception.
- So they have rightly been criticised as subject to manipulation, and one should be aware of the possibility that the effect may differ by viewer.

Glyph Representations

As an example, a stars plot for the **state.x77** dataset with variables in the order showing up in the biplot of Figure 6 can be drawn by

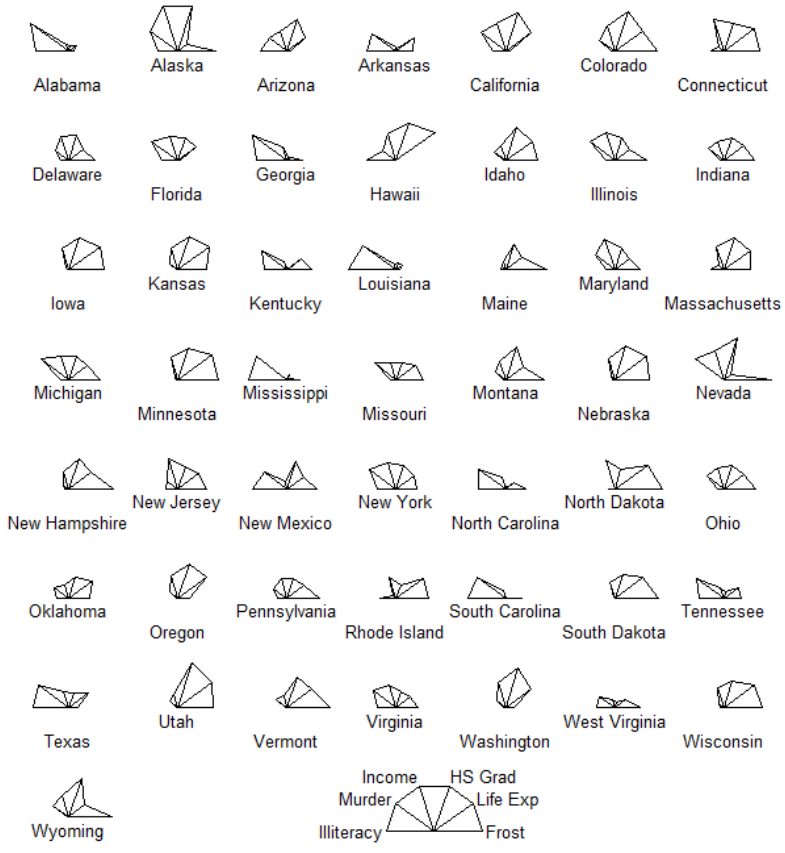

Figure 8: R version of stars plot of the **state.x77** dataset.

8. Parallel coordinate plots

- Parallel coordinates plots (Inselberg, 1984; Wegman, 1990) join the same points across a set of parallel axes.
- We can show the **state.x77** dataset in the order showing up in the biplot of Figure 6 by

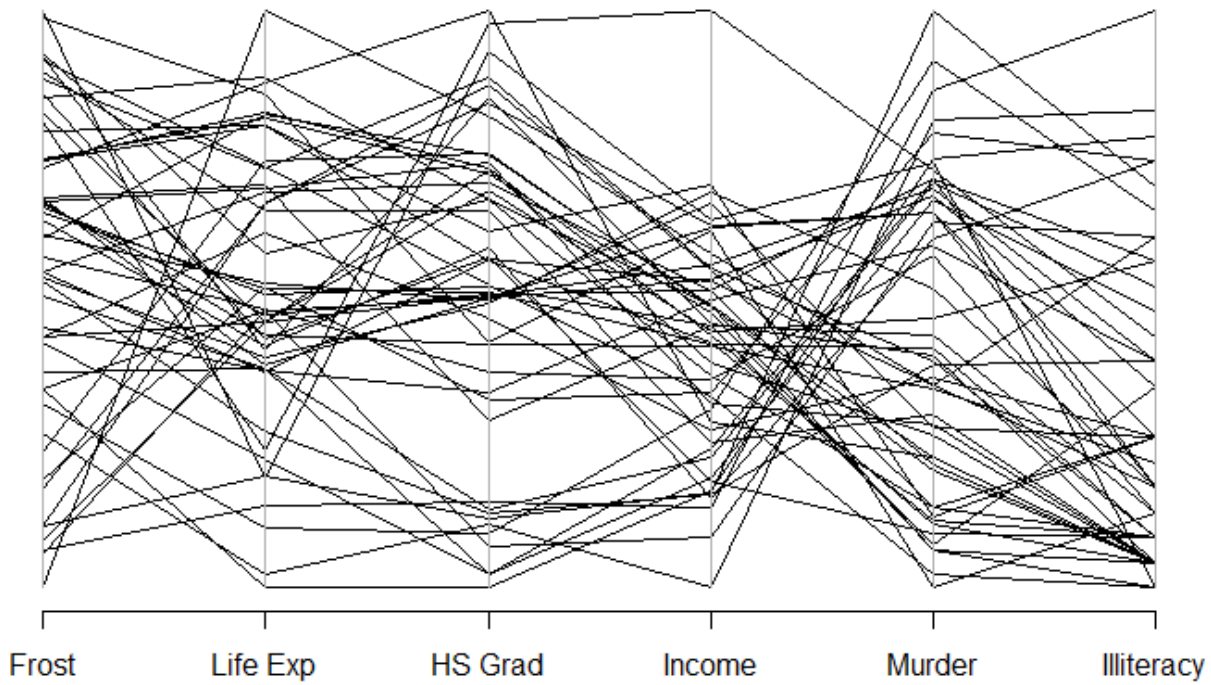

Figure 9: R version of parallel coordinate plot of the **state.x77** dataset

Parallel coordinate plots

- Such plots are often too 'busy' without a means of interaction to identify observations, sign-change and reorder variables, brush groups and so on.
- As an example of a revealing parallel coordinate plot :

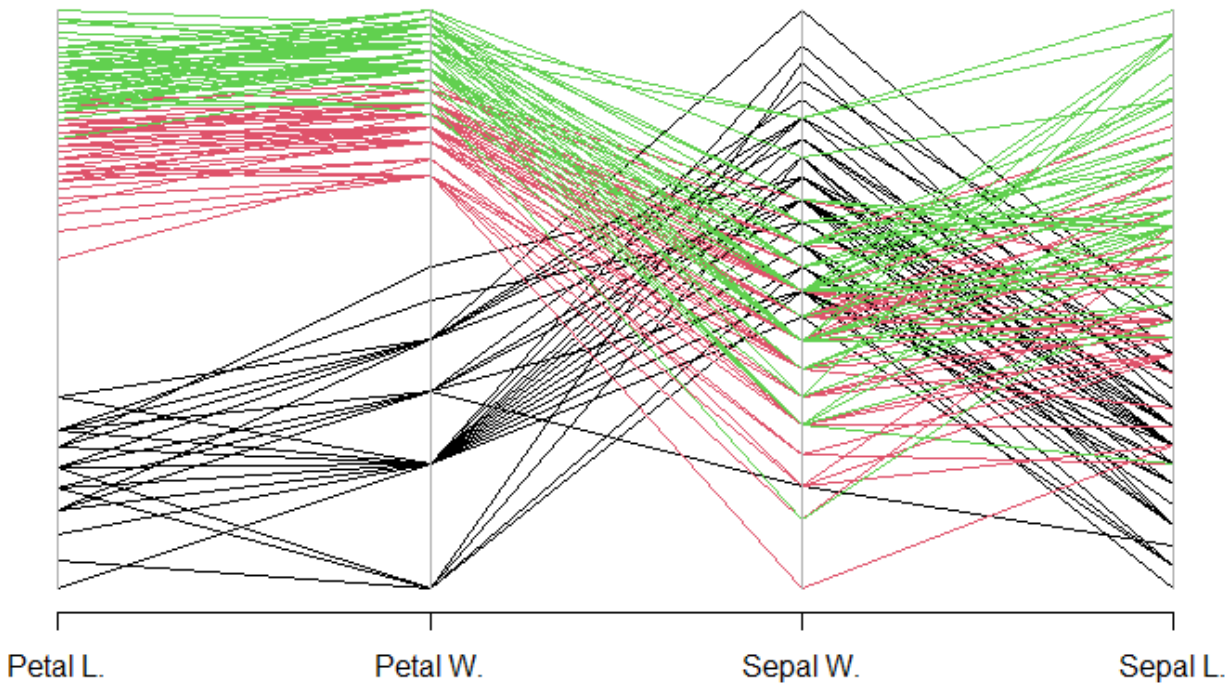

Figure 10: R version of parallel coordinate plot of the iris dataset.

Cluster Analysis

Cluster Analysis

Cluster analysis is concerned with discovering groupings among the cases of our n by p matrix. A comprehensive general reference is Gordon (1999); Kaufman and Rousseeuw (1990) give a good introduction and their methods are available in package cluster for **R** Clustering methods can be clustered in many different ways.

Cluster Analysis

- Agglomerative hierarchical methods.
  - Produces a set of clusterings, usually one with k clusters for each $k = n, \cdots, 2$, successively amalgamating groups.
  - Main differences are in calculating group-group dissimilarities from point-point dissimilarities.
  - Computationally easy.
- Optimal partitioning methods.
  - Produces a clustering for fixed K.
  - Need an initial clustering.
  - Lots of different criteria to optimize, some based on probability models.
  - Can have distinct 'outlier' group(s) .

Cluster Analysis

- Divisive hierarchical methods.
  - Produces a set of clusterings, usually one for each $k = 2, ..., K \ll n$.
  - Computationally nigh-impossible to find optimal divisions (Gordon, 1999, p.90).
  - Most available methods are *monothetic* (split on one variable at each stage).

▶ Do not assume that 'clustering' methods are the best way to discover interesting groupings in the data. There are many different clustering methods, often giving different answers, and so the danger of over-interpretation is high.

Cluster Analysis

Many methods are based on a measure of the similarity or dissimilarity between cases, but some need the data matrix itself.

dissimilarity coefficient d

- d is symmetric (i.e. $d(A, B) = d(B, A)$)
- d is non-negative and $d(A, A) = 0$
- Dissimilarities may be *metric*

$$d(A, C) \leq d(A, B) + d(B, C)$$

or *ultrametric*

$$d(A, B) \leq \max(d(A, C), d(B, C))$$

but need not be either.

Cluster Analysis

- Hierarchical clustering methods can be thought of as approximating a dissimilarity by an ultrametric dissimilarity.
- Jardine and Sibson (1971) argue that one method, single-link clustering, uniquely has all the desirable properties of a clustering method. This measures distances between clusters by the dissimilarity of the closest pair, and agglomerates by adding the shortest possible link (that is, joining the two closest clusters).

- Kaufman and Rousseeuw (1990, §5.2) give a different set of desirable properties leading uniquely to their preferred method, which views the dissimilarity between clusters as the average of the dissimilarities between members of those clusters.
- Another popular method is complete-linkage, which views the dissimilarity between clusters as the maximum of the dissimilarities between members.

Cluster Analysis

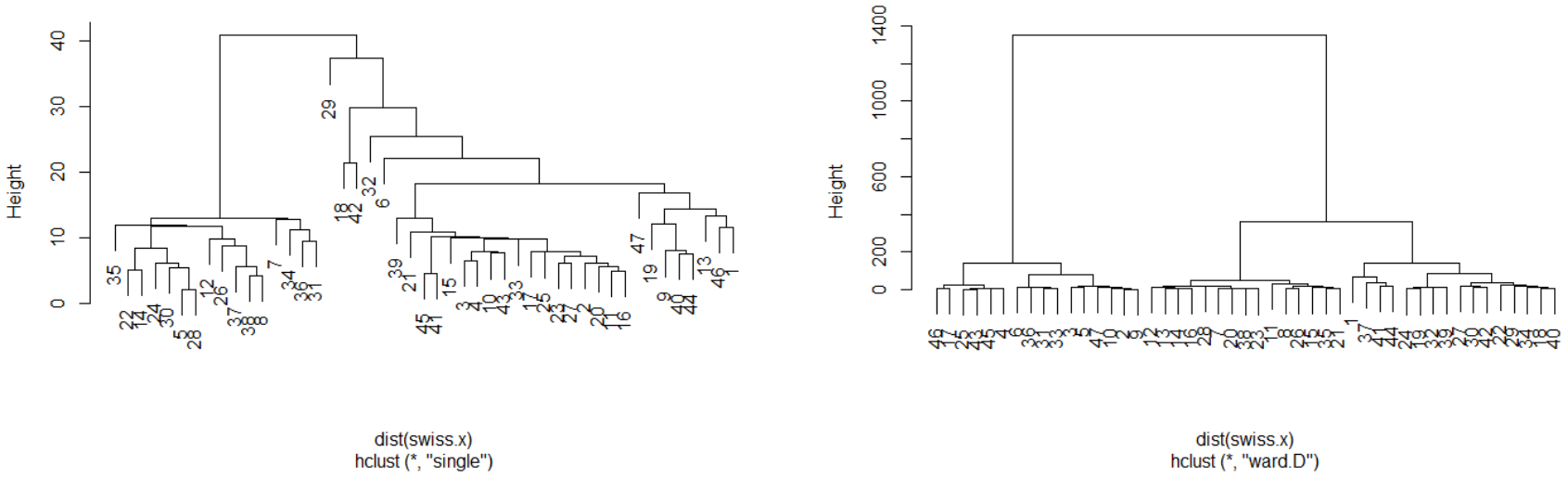

Figure 11: Dendrograms for the socio-economic data on Swiss provinces computed by single-link clustering (top) and divisive clustering (bottom).

- The hierarchy of clusters in a dendrogram is obtained by cutting it at different heights.
- The left plot suggests three main clusters. Note that there appear to be two main groups.
- The right plot was produced by *divisive* clustering.

Partitioning methods

- The K-means clustering algorithm (MacQueen, 1967; Hartigan, 1975; Hartigan and Wong, 1979) chooses a pre-specified number of cluster centres to minimize the within-class sum of squares from those centres.
- As such it is most appropriate to continuous variables, suitably scaled.
- The algorithm needs a starting point, so we choose the means of the clusters identified by group-average clustering.

Partitioning methods

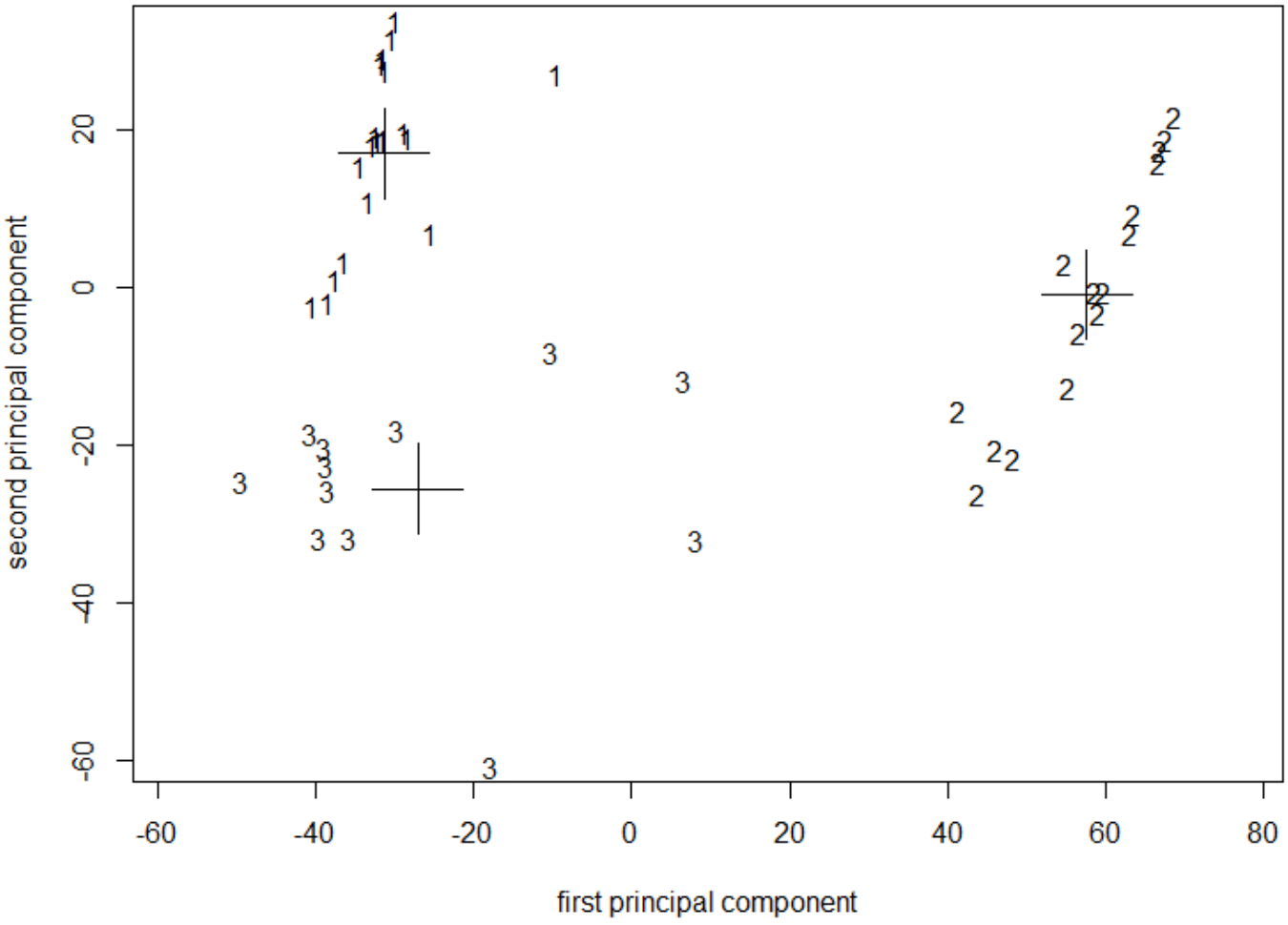

Figure 12: The Swiss provinces data plotted on its first two principal components. The labels are the groups assigned by K-means; the crosses denote the group means. Five points are labelled with smaller symbols.

Partitioning methods

- By definition, K-means clustering needs access to the data matrix and uses Euclidean distance. We can apply a similar method using only dissimilarities if we confine the cluster centres to the set of given examples. This is known as the k-medoids criterion (of Vinod, 1969) implemented in `pam` and `clara`.
- The function `fanny` implements a 'fuzzy' version of the k-medoids criterion. Rather than point i having a membership of just one cluster v, its membership is partitioned among clusters as positive weights u_{iv} summing to one. The criterion then is

$$\min_{(u_{iv})} \sum_v \frac{\sum_{i,j} u_{iv}^2 u_{jv}^2 d_{ij}}{2 \sum_i u_{iv}^2}$$

Partitioning methods

- The 'hard' clustering is formed by assigning each point to the cluster for which its membership coefficient is highest.
- Other partitioning methods are based on the idea that the data are independent samples from a series of group populations, but the group labels have been lost, so the data can be regarded as from a mixture distribution. The idea is then to find the mixture distribution, usually as a mixture of multivariate normals, and to assign points to the component for which their posterior probability of membership is highest.

Partitioning methods

- Library section **mclust** provides hierarchical clustering via functions `hc` and `mhclass`. Then for a given number k of clusters the fitted mixture can be optimized by calling `me` (which here does not change the classification).
- Function `EMclust` automates the whole cluster process, including choosing the number of clusters and between different **modelName**'s. It chooses lots of clusters (see Figure 13).
- Another possibility is to use function `EMclustN` to fit a cluster model including a background 'noise' term, that is a component that is a uniform Poisson process, controlled by argument **noise**.

Partitioning methods

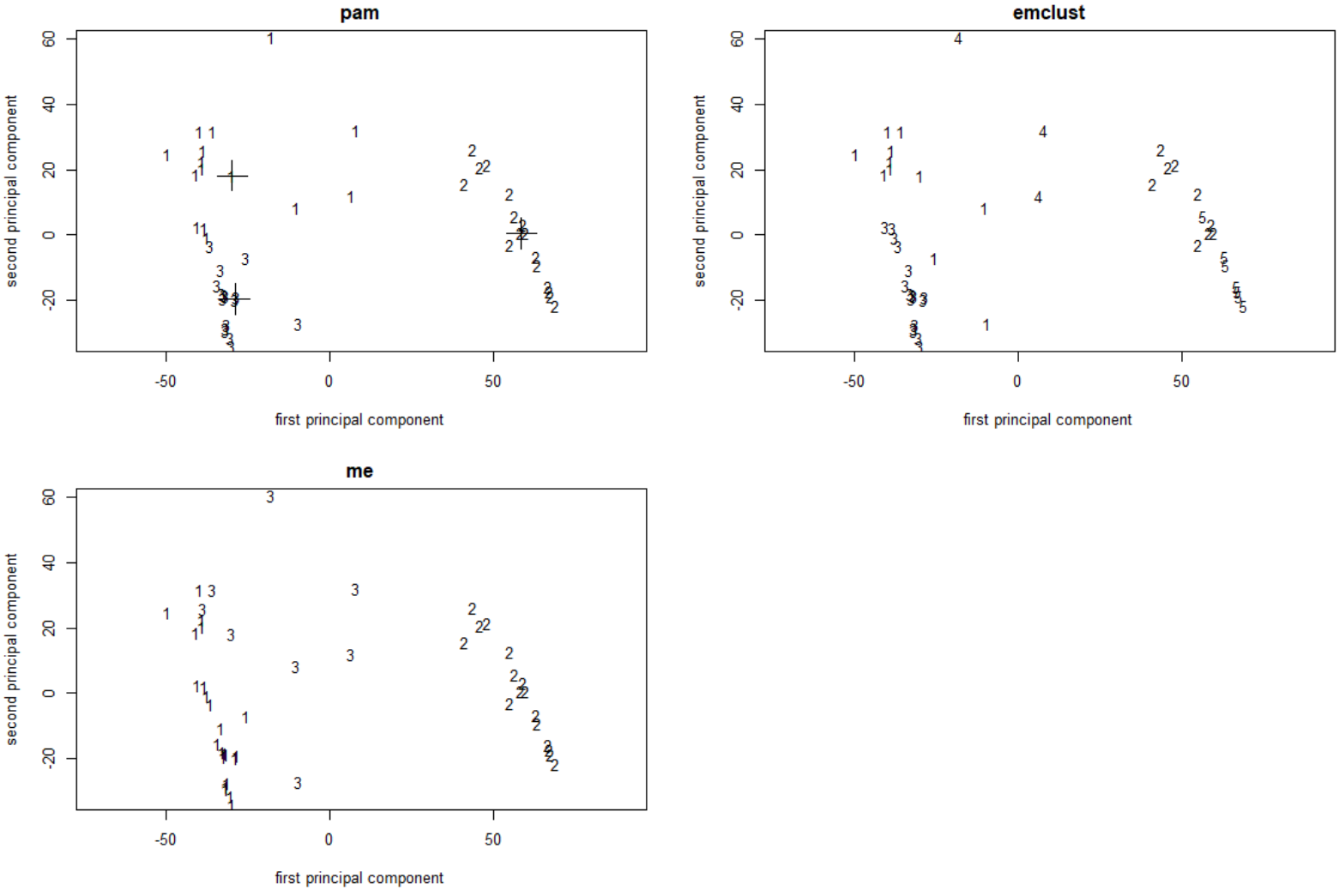

Figure 13: Clustering of the Swiss provinces data by pam with three clusters (with the medoids marked by crosses), me with three clusters and emclust with up to nine clusters (it chose nine).

Factor Analysis

Factor Analysis

- Principal component analysis looks for linear combinations of the data matrix $\boldsymbol{X}$ that are uncorrelated and of high variance.
- Independent component analysis seeks linear combinations that are independent.
- *Factor analysis* seeks linear combinations of the variables, called *factors*, that represent underlying fundamental quantities of which the observed variables are expressions.
- The examples tend to be controversial ones such as 'intelligence' and 'social deprivation', the idea being that a small number of factors might explain a large number of measurements in an observational study. Such factors are to be inferred from the data.

Factor Analysis

We can think of both the factors of factor analysis and the signals of independent component analysis as latent variables, unobserved variables on each experimental unit that determine the patterns in the observations.
The difference is that it is not the factors that are assumed to be independent, but rather the observations conditional on the factors.

The factor analysis model for a single common factor $\boldsymbol{f}$ is

$$x = \mu + \lambda f + u \tag{1}$$

where $\boldsymbol{\lambda}$ is a vector known as the *loadings* and $\boldsymbol{u}$ is a vector of *unique* (or *specific*) factors for that observational unit.

Factor Analysis

To help make the model identifiable, we assume that the factor f has mean zero and variance one, and that u has mean zero and unknown *diagonal* covariance matrix Ψ. For $k < P$ common factors we have a vector $\boldsymbol{f}$ of common factors and a loadings matrix Λ, and

$$x = \mu + \Lambda f + u \tag{2}$$

where the components of $\boldsymbol{f}$ have unit variance and are uncorrelated and $\boldsymbol{f}$ and $\boldsymbol{u}$ are taken to be uncorrelated.

Note that all the correlations amongst the variables in $\boldsymbol{x}$ must be explained by the common factors; if we assume joint normality the observed variables x will be conditionally independent given $\boldsymbol{f}$.

Factor Analysis

Principal component analysis also seeks a linear subspace like $\Lambda \boldsymbol{f}$ to explain the data, but measures the lack of fit by the sum of squares of the $\boldsymbol{u}_i$. Since factor analysis allows an arbitrary diagonal covariance matrix Ψ, its measure of fit of the $\boldsymbol{u}_i$ depends on the problem and should be independent of the units of measurement of the observed variables. (Changing the units of measurement of the observations does not change the common factors if the loadings and unique factors are re-expressed in the new units.)

Factor Analysis

Equation (2) and the conditions on $\boldsymbol{f}$ express the covariance matrix Σ of the data as

$$\Sigma = \Lambda\Lambda^T + \Psi \tag{3}$$

Conversely, if (3) holds, there is a k-factor model of the form (2).

Note that the common factors $G^T\boldsymbol{f}$ and loadings matrix ΛG give rise to the same model for Σ, for any $k \times k$ orthogonal matrix G. Choosing an appropriate G is known as choosing a rotation. All we can achieve statistically is to fit the space spanned by the factors, so choosing a rotation is a way to choose an interpretable basis for that space. Note that if

$$s = \frac{1}{2}p(p+1) - \left[p(k+1) - \frac{1}{2}k(k-1)\right] = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) < 0$$

we would expect an infinity of solutions to (3). This value is known as the *degrees of freedom*, and comes from the number of elements in Σ minus the number of parameters in Ψ and Λ (taking account of the rotational freedom in Λ since only $\Lambda\Lambda^T$ is determined). Thus it is usual to assume $s \geq 0$; for $s = 0$ there may be a unique solution, no solution or an infinity of solutions (Lawley and Maxwell, 1971, pp. 10-11).

The variances of the original variables are decomposed into two parts, the *communality* $h_i^2 = \Sigma_j \lambda_{ij}^2$ and uniqueness ψ_{ii} which is thought of as the 'noise' variance.

Factor rotations

The usual aim of a rotation is to achieve 'simple structure', that is a pattern of loadings that is easy to interpret with a few large and many small coefficients.

The varimax criterion maximizes the sum over factors of the variances of the (normalized) squared loadings.

$$\sum_{i,j} \left(d_{ij} - \bar{d}_{\cdot j} \right)^2 \quad \text{where} \quad d_{ij} = \lambda_{ij}^2 / \sum_j \lambda_{ij}^2 \tag{4}$$

with $\overline{d}_{\cdot j}$ is the mean of the d_{ij}.

The normalizing factors are the communalities that are invariant under orthogonal rotations.

Factor rotations

Following Bartholomew & Knott, we illustrate the **oblimin** criterion which minimizes the sum over all pairs of factors of the covariance between the squared loadings for those factors.

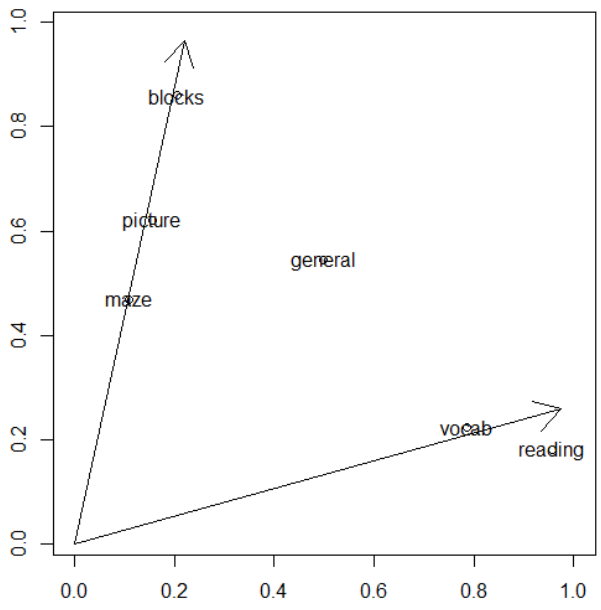

Figure 14: The loadings for the intelligence test data after varimax rotation, with the axes for the oblimin rotation shown as arrows.

Discrete Multivariate Analysis

Discrete Multivariate Analysis

Most work on visualization and most texts on multivariate analysis implicitly assume continuous measurements. However, large-scale categorical datasets are becoming much more prevalent, often collected through surveys or '**CRM**' (customer relationship management: that branch of data mining that collects information on buying habits, for example on shopping baskets) or insurance questionnaires.

Discrete Multivariate Analysis

There are some useful tools available for exploring categorical data.

- Mosaic plots
- Correspondence analysis
- Multiple correspondence analysis

But it is often essential to use models to understand the data, most often log-linear models.

1. Mosaic plots

- There are a few ways to visualize low-dimensional contingency tables.
- Mosaic plots divide the plotting surface recursively according to the proportions of each factor in turn (so the order ofthe factors matters).
- For an example, consider **Fisher**'s (1940) data on colours of eyes and hair of people in Caithness, Scotland:

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Mosaic plots

- Figure 15 shows mosaic plots for these data and for the housing data we used in Section 7.3

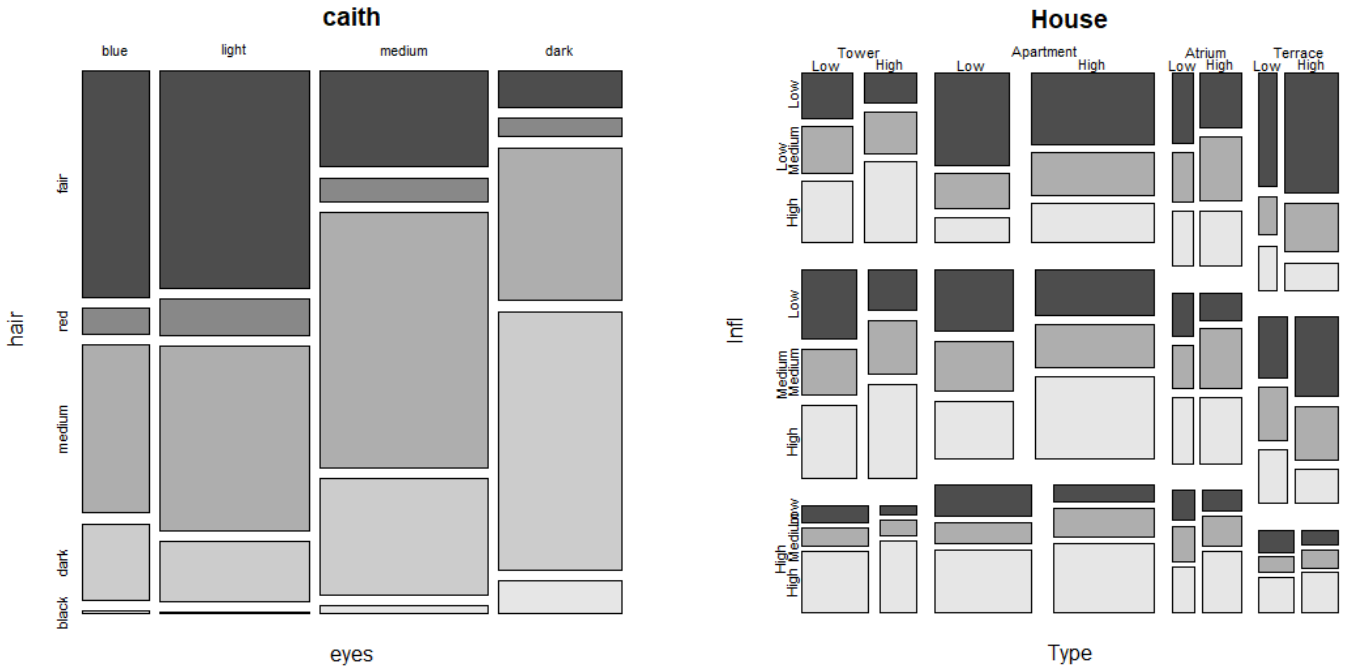

Figure 15: Mosaic plots for (top) Fisher's data on people from Caithness and (bottom) Copenhagen housing satisfaction data.

2. Correspondence analysis

- Correspondence analysis is applied to two-way tables of counts.

- Suppose we have an r x c table N of counts.
- Correspondence analysis seeks 'scores' $\boldsymbol{f}$ and g for the rows and columns which are maximally correlated.

$\Rightarrow$ The maximum correlation is one, attained by constant scores, so we seek the largest non-trivial solution.

Correspondence analysis

- Consider the singular value decomposition of their correlation matrix

$$X_{ij} = \frac{n_{ij}/n - (n_{i\cdot}/n)\,(n_{\cdot j}/n)}{\sqrt{(n_{i\cdot}/n)\,(n_{\cdot j}/n)}} = \frac{n_{ij} - n r_i c_j}{n\sqrt{r_i c_j}}$$

  where $r_i = n_{i\cdot}/n$ and $c_j = n_{\cdot j}/n$
- Let D_r and D_c be the diagonal matrices of r and c.
- Correspondence analysis corresponds to selecting the first singular value and left and right singular vectors of X_{ij} and rescaling by $D_r^{-1/2}$ and $D_c^{-1/2}$, respectively.

Correspondence analysis

Question : Can we make use of the subsequent singular values?

- In what Gower and Hand(1996) call 'classical CA' we consider $A = D_r^{-1/2} U \Lambda$ and $B = D_c^{-1/2} V \Gamma$.
- Then the first columns of A and B are what we have termed the row and column scores *scaled by* ρ, the first canonical correlation.
- More generally, we can see distances between the rows of A as approximating the distances between the row profiles (rows resealed to unit sum) of the table N, and analogously for the rows of B and the column profiles.

Correspondence analysis

- Classical CA plots the first two columns of A and B on the same figure. This is a form of a biplot. This is sometimes known as a 'symmetric' plot.
- The asymmetric plot for the rows is a plot of the first two columns of A with the column labels plotted at the first two columns of $\Gamma = D_c^{-1/2}V$; the corresponding plot for the columns has columns plotted at B and row labels at $\Pi = D_r^{-1/2}U$.
- The most direct interpretation for the row plot is that

$$A = D_r^{-1}N\Gamma$$

so A is a plot of the *row profiles* (the rows normalized to sum to one) as convex combinations of the column vertices given by Γ.

Correspondence analysis

The two-dimensional forms of the plot are shown in Figure 16 for Fisher's data on people from Caithness.

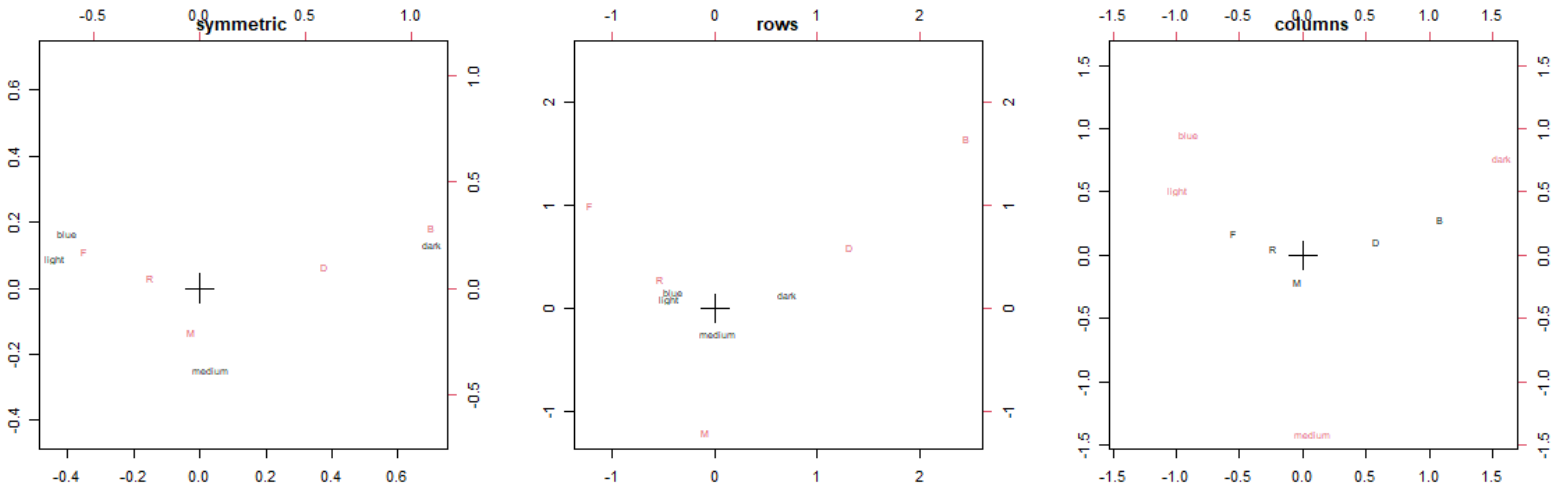

Figure 16: Three variants of correspondence analysis plots from Fisher's data on people in Caithness: (left) 'symmetric', (middle) 'row asymmetric' and (right) 'column asymmetric'.

Note that the 'left' plot has the row points from the 'middle' plot and the column points from the 'right' plot superimposed on the same plot (but with different scales).

3. Multiple correspondence analysis

Multiple correspondence analysis (`MCA`) is a method for visualizing the joint properties of $p \geq 2$ categorical variables that does *not* reduce to correspondence analysis (CA) for $p = 2$, although the methods are closely related (see, for example, Gower and Hand, 1996, §10.2).

Multiple correspondence analysis

- Suppose we have n observations on the p factors with l total levels.
- Consider G, the n x l indicator matrix whose rows give the levels of each factor for each observation. Then all the row sums are p.
- `MCA` is often defined as CA applied to the table G, that is the singular-value decomposition of $D_r^{-1/2}(G/\sum_{ij} g_{ij})D_c^{-1/2}$.
- Note that $D_r = pI$ since all the row sums are p, and $\sum_{ij} g_{ij} = np$, so this amounts to the SVD of $p^{-1/2}GD_c^{-1/2}/pn$.

Multiple correspondence analysis

- An alternative point of view is that `MCA` is a principal components analysis of the data matrix $X = G(pD_c)^{-1/2}$
- it is usual to center the data, but it transpires that the largest singular value is one and the corresponding singular vectors account for the means of the variables.
- A simple plot for `MCA` is to plot the first two principal components of X

Multiple correspondence analysis

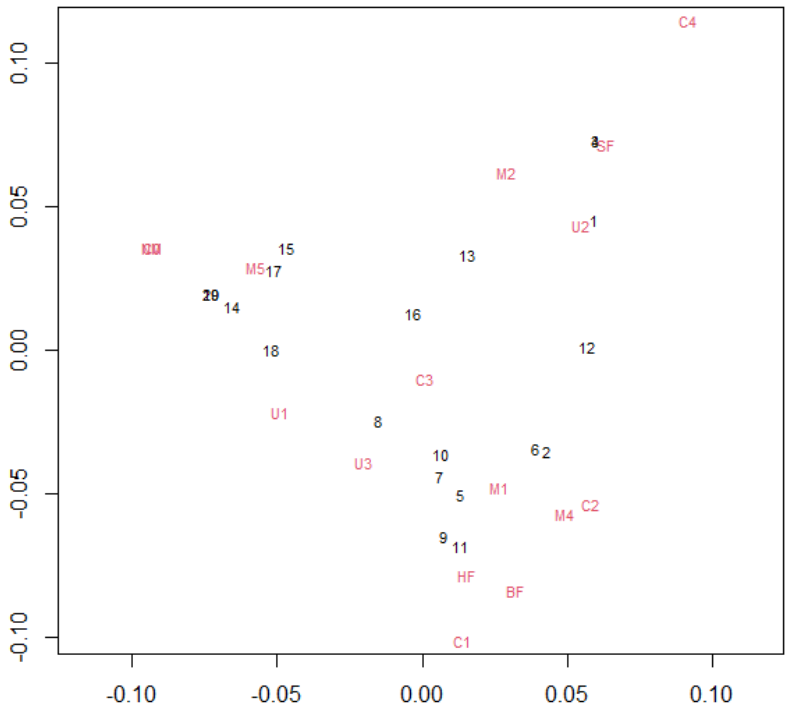

Figure 17: Multiple correspondence analysis plot of dataset farms on 20 farms on the Dutch island of Terschelling. Numbers represent the farms and labels levels of moisture (Mi, $M2$, $M4$ and $M5$), grassland usage ($U1$, $U2$ and $U3$), manure usage (CO to $C4$) and type of grassland management (SF: standard, BF: biological, HF: hobby farming, NM: nature conservation). Levels CO and NM are coincident (on the extreme left), as are the pairs of farms 3 & 4 and 19 & 20.