Sixth International Conference on
Intelligent Data Engineering and Automated Learning
IDEAL'05

Cluster Analysis Session (Stream 2)
Thursday 7th July 2005, 4pm
Hawken Engineering Building, Room 2
University of Queensland

Cluster Analysis of High-Dimensional Data: A Case Study

Richard Bean[1] and Geoff McLachlan[1,2]

[1]ARC Centre in Bioinformatics, Institute for Molecular Bioscience, UQ
[2]Department of Mathematics, University of Queensland

Outline of Talk

- Normal Mixture Models
- Principal Components Analysis (PCA)
- Mixtures of Factor Analyzers
- Ashenfelter Wine Data
  - ❖ Clustering of Wines
  - ❖ Clustering of Judges

Basic Definition

We let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability density function $f(\mathbf{y}_j)$

$$f(\mathbf{y}_j) = \pi_1 f_1(\mathbf{y}_j) + \cdots + \pi_g f_g(\mathbf{y}_j)$$

where the $f_i(\mathbf{y}_j)$ are densities and the π_i are nonnegative quantities that sum to one.

Mixture distributions are applied to data with two main purposes in mind:

- To provide an appealing semiparametric framework in which to model unknown distributional shapes, as an alternative to, say, the kernel density method.
- To use the mixture model to provide a model-based clustering. (In both situations, there is the question of how many components to include in the mixture.)

Normal Mixtures

- Computationally convenient for multivariate data
- Provide an arbitrarily accurate estimate of the underlying density with *g* sufficiently large
- Provide a probabilistic clustering of the data into *g* clusters - outright clustering by assigning a data point to the component to which it has the greatest posterior probability of belonging

Normal Mixtures

- The observed data $y_1,\ldots,y_n$ are assumed to have come from the normal distribution.

Mixture of g normal components

$$f(\mathbf{y};\Psi) = \pi_1\phi(\mathbf{y};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \ldots\ldots + \pi_g\phi(\mathbf{y};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$$

where

$$-2\log\phi(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) + \text{ constant}$$

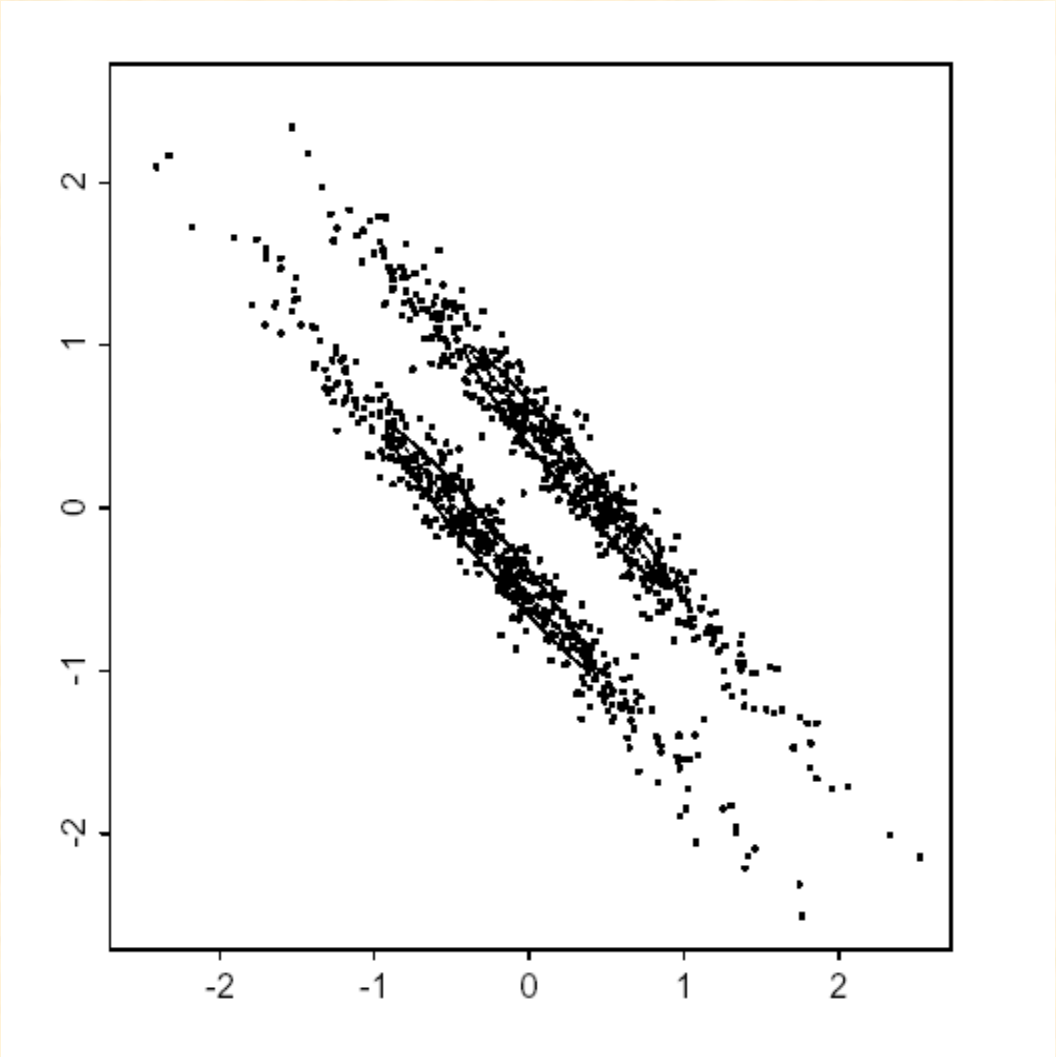
MAHALANOBIS DISTANCE

$$(\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{y}-\boldsymbol{\mu})$$

EUCLIDEAN DISTANCE

In exploring high-dimensional data sets for group structure, it is typical to rely on Principal Component Analysis (PCA).

Two Groups in Two Dimensions. All cluster information would be lost by collapsing to the first principal component. The principal ellipses of the two groups are shown as solid curves.



Mixtures of Factor Analyzers

A normal mixture model without restrictions on the component-covariance matrices may be viewed as too general for many situations in practice, in particular, with high dimensional data.

One approach for reducing the number of parameters is to work in a lower dimensional space by adopting mixtures of factor analyzers (Ghahramani & Hinton, 1997)

$$f(\boldsymbol{y}_j;\Psi) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i),$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1,\ldots,g),$$

$\boldsymbol{B}_i$ is a p x q matrix and

$\boldsymbol{D}_i$ is a diagonal matrix.

Single-Factor Analysis Model

$$Y_j = \mu + B\,U_j + e_j \qquad (j = 1,\ldots,n),$$

where U_j is a q - dimensional $(q < p)$ vector of latent or unobservable variables called factors and B is a p x q matrix of factor loadings.

The U_j are iid $N(O, I_q)$ independently of the errors e_j, which are iid as $N(O, D)$, where D is a diagonal matrix

$$D = \text{diag}\,(\sigma_1^2, \ldots, \sigma_p^2)$$

Mixtures of Factor Analyzers

A single-factor analysis model provides only a global linear model.

A global nonlinear approach by postulating a mixture of linear submodels…

Conditional on ith component membership of the mixture,

$$Y_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \ (i = 1, \ldots, g).$$

where $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are independent, identically distributed (iid) $N(\boldsymbol{O}, \boldsymbol{I}_q)$, independently of the $\boldsymbol{e}_{ij}$, which are iid $N(\boldsymbol{O}, \boldsymbol{D}_i)$, where $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$.

If the number of factors q is chosen sufficiently small relative to the number of observations n, then there will be no singularity problems in fitting a mixture of factor analyzers for equal component-covariance matrices. For unrestricted component-covariance matrices, there may still be some problems if the clusters are small in size; in which case, they can be avoided by specifying the diagonal matrices $\boldsymbol{D}_i$ to be the same.

Ashenfelter Wine Data

- St Francis Hotel, San Francisco 1999
- 32 judges; 47 Cabernet wines from 11 countries (France, Lebanon, Chile, Australia, New Zealand, Argentina, Spain, France, Italy, South Africa, USA (California and Washington))

- Anonymous judges: “all members of the West Coast wine establishment”

http://www.liquidasset.com/report20.html

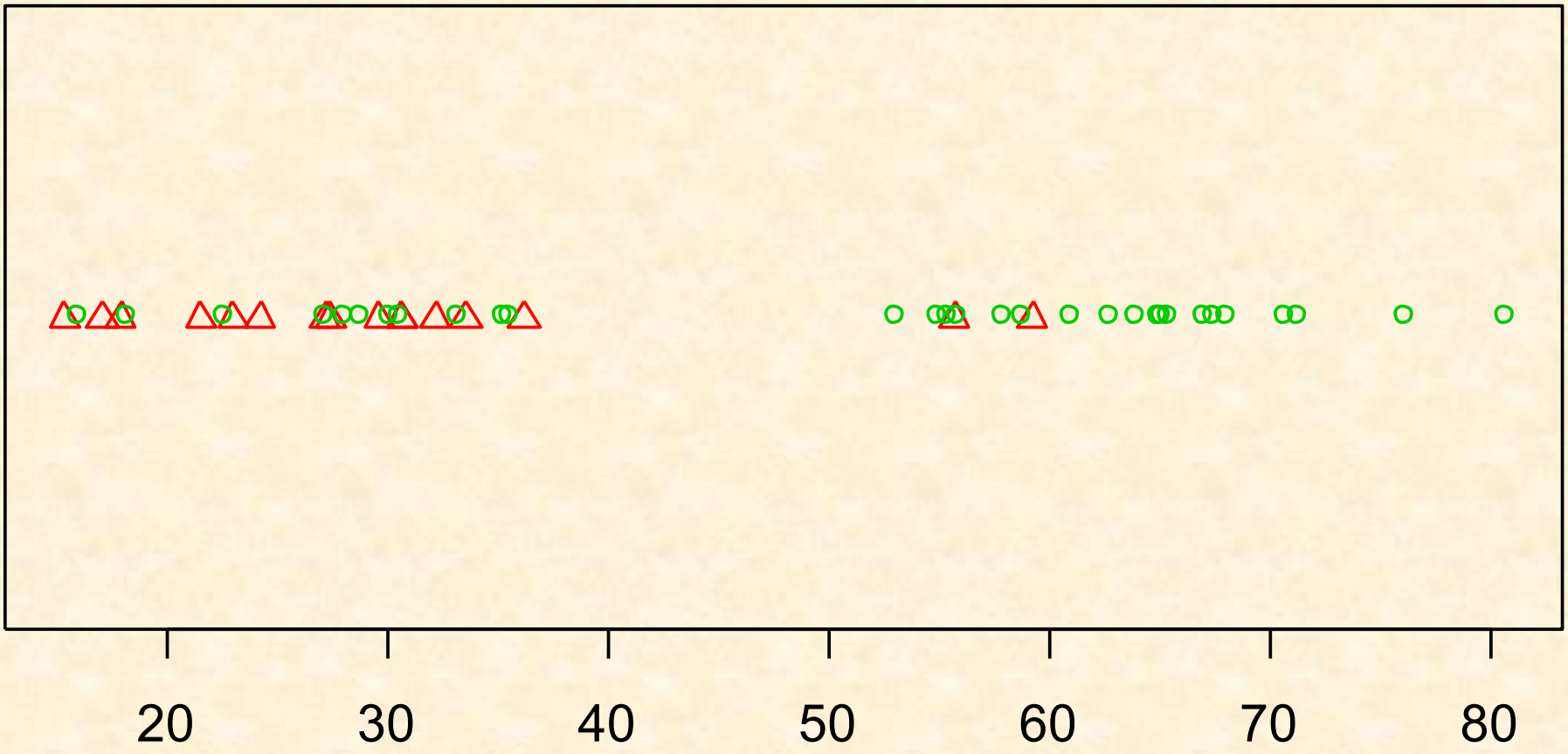
Wine Data

- Distinctive Clusters
  - ❖ 16 Napa Valley wines
  - ❖ 8 Bordeaux wines
- 1 missing data point imputed via *k*-nearest neighbours algorithm with *k*=5
- No data for one wine (i.e. data for 46 wines)

Clustering of Wines

Using a resampling approach, on the basis of B=19 replications of the LRTS $-2 \log \lambda$, we rejected the null hypothesis $H_0 : g = 1$ versus the alternative $H_1 : g = 2$ at the 5% level. The null hypothesis of a single normal component was rejected also using the Bayesian information criterion (BIC) since it was found that $-2 \log \lambda > d \log(n)$.

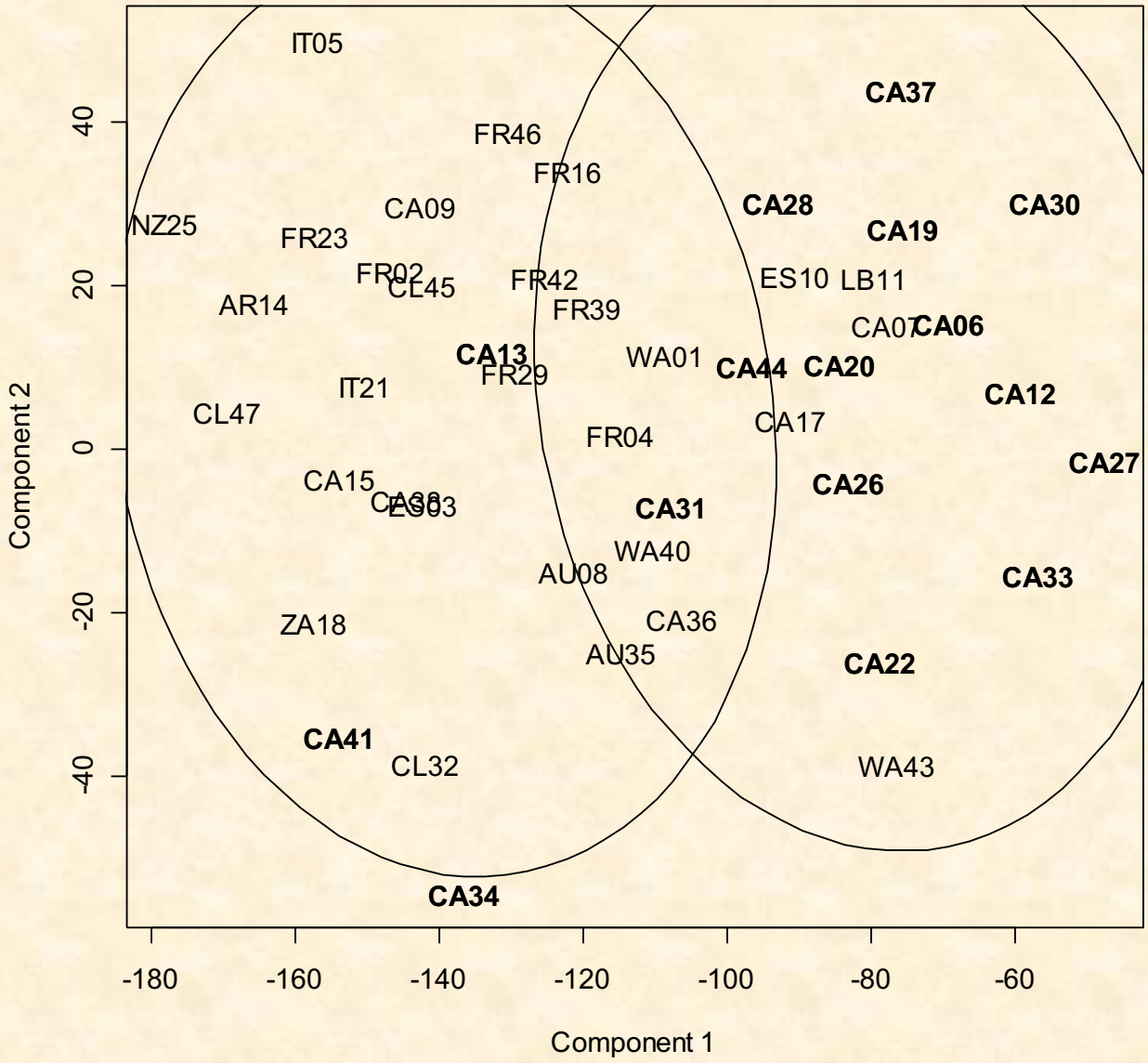
q=2 factors chosen over q=3 factors after similar test.



First canonical variate of the 46 wines. Napa Valley wines are represented by triangles and the other wines by circles.

Clustering of Wines

It is of interest to compare the clustering obtained using mixtures of $g = 2$ factor analyzers ($q = 2$ factors) with that obtained using a mixture of two normals fitted to the first two PCs.

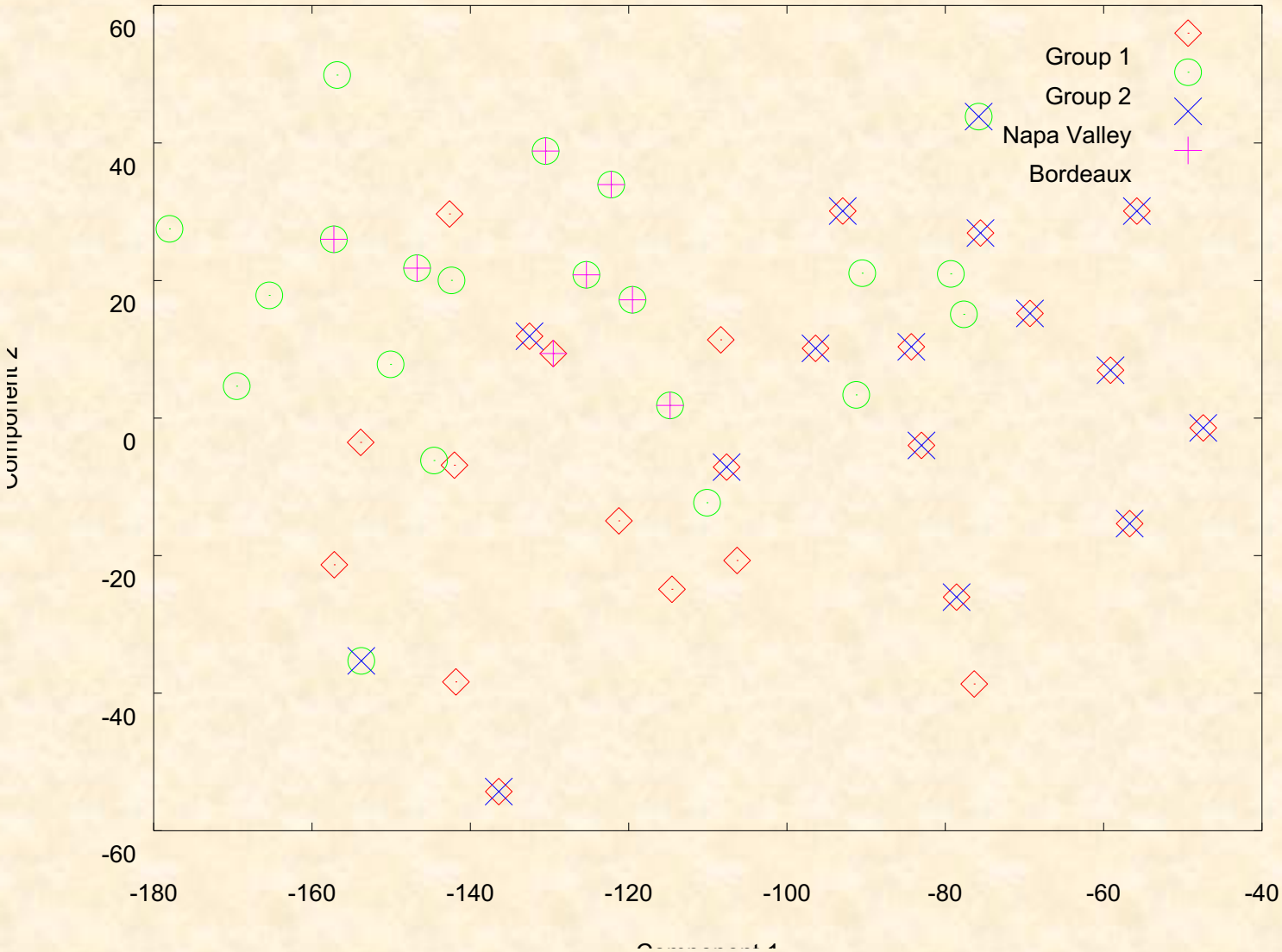


Plot of the first two principal components of a PCA on the 46 wines. Napa Valley wines in **bold;** ellipses two groups given by fitting a mixture of two normals with equal covariance matrices.

Clustering of Wines

For comparative purposes, we give the clustering obtained by mixtures of factor analyzers in the space of the first two PCs.

Clustering obtained by mixtures of factor analyzers with Groups 1 and 2 superimposed with Napa Valley and Bordeaux Wines.



Clustering obtained by mixtures of factor analyzers with Groups 1 and 2 superimposed with Napa Valley and Bordeaux Wines.

Clustering of Wines

The larger cluster obtained using mixtures of factor analyzers contains 14 of the 16 Napa Valley wines from California, while the smaller cluster obtained using mixtures of normals fitted to the first two PCs contains 12 of the 16 Napa Valley wines.

Clustering of Judges

- Young (2005) – three atypical judges (16, 26, 31) detected with the PowerMV program of Liu et al. (2005) using an R/G plot of the outer product of the right and left eigenvectors.
- It is of interest to consider the clustering of the 32 judges on the basis of their scores for 46 wines. For this clustering problem, we now have $n = 32$ and $p = 46$.

Clustering of Judges

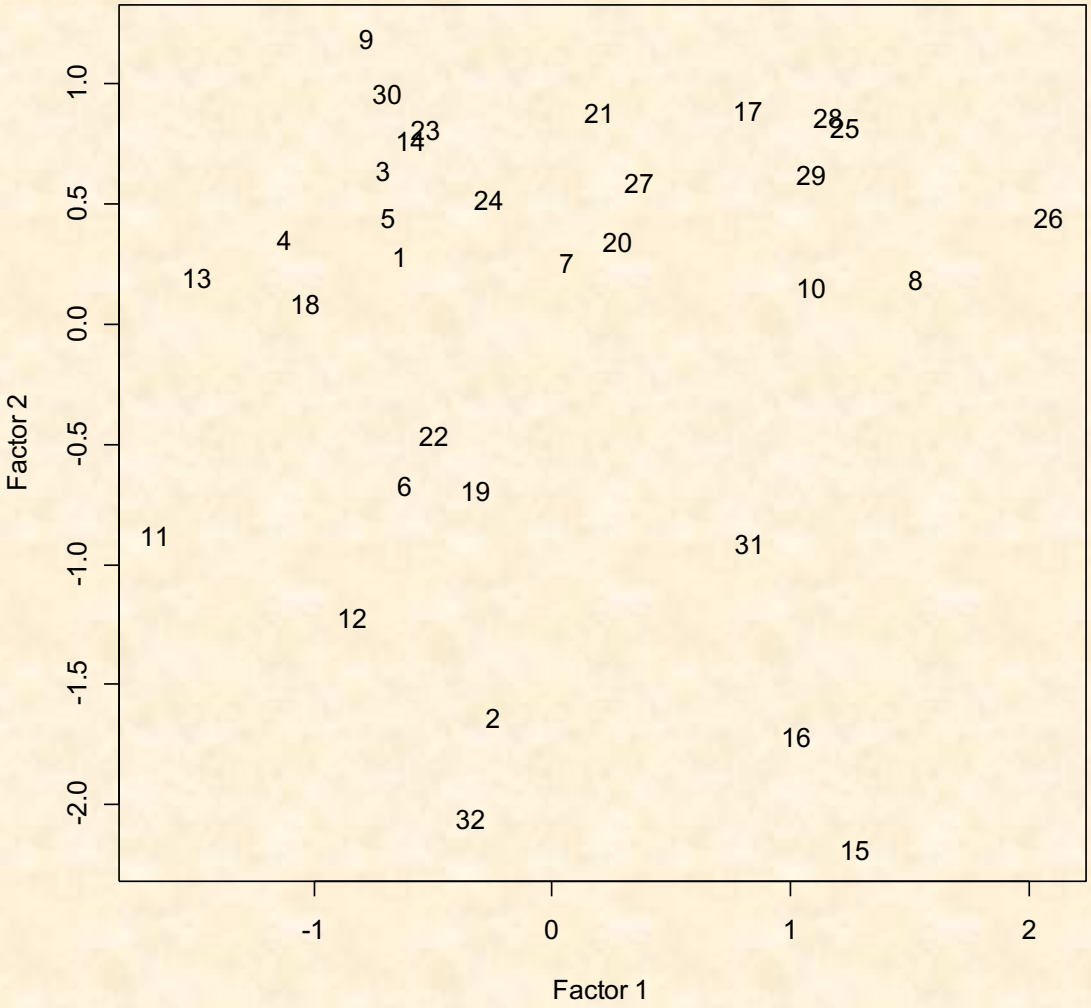
- Using equal covariance matrices and fitting $g = 2$ factor analyzers with $q = 2$ resulted in two clusters each of size 16, placing judges 16 and 26 in one cluster and judge 31 in the other.
- A resampling approach with $B = 19$ replications, as above, showed that the test of $g = 1$ versus $g = 2$ groups was significant at the 5% level.

Posterior Means

We can represent an original data point $\boldsymbol{y}_j$ in q-dimensional space by plotting the estimated conditional expectation of each factor given $\boldsymbol{y}_j$ and its component membership, that is, the (estimated) posterior mean of the factor $\boldsymbol{U}_{ij}$ ($i = 1, \ldots, g$; $j = 1, \ldots, n$), where $\boldsymbol{U}_{ij}$ is the latent factor corresponding to the j^{th} observation in the i^{th} component (see Section 8.7.4 in McLachlan and Peel, "Finite Mixture Models").

Judges' Plot

A plot of the estimated posterior means of the (unobservable) factors from fitting a single factor analysis model with $q = 2$ factors also suggests that these judges (plus judge 15) are quite distinct from the others in their scores.



Plot of the estimated posterior means of the $q = 2$ factors following a single-component factor analysis of the judge scores in the wine data set.