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 (UQ)

Abstract. Normal mixture models are often used to cluster continuous data. However, conventional approaches for fitting these models will have problems in producing nonsingular estimates of the component-covariance matrices when the dimension of the observations is large relative to the number of observations. In this case, methods such as principal components analysis (PCA) and the mixture of factor analyzers model can be adopted to avoid these estimation problems. We examine these approaches applied to the Cabernet wine data set of Ashenfelter (1999), considering the clustering of both the wines and the judges, and comparing our results with another analysis. The mixture of factor analyzers model proves particularly effective in clustering the wines, accurately classifying many of the wines by location.

1 Introduction

In recent times much attention has been given in the scientific literature to the use of normal mixture models as a device for the clustering of continuous data; see, for example, McLachlan and Peel (2000b). With this approach, the observed data $\boldsymbol{y}_1, \ldots \boldsymbol{y}_n$, are assumed to have come from the normal mixture distribution,

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

where $\phi(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrix $\boldsymbol{\Sigma}_i$. The normal mixture model (1) can be fitted iteratively to an observed random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ by maximum likelihood (ML) via the expectation-maximization (EM) algorithm of Dempster et al., (1977); see also McLachlan and Krishnan (1997). Frequently, in practice, the clusters in the data are essentially elliptical, so that it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its computational tractability.

Under (1), the posterior probability that an observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture is given by

$$\tau_i(\boldsymbol{y}_j) = \pi_i \phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) / f(\boldsymbol{y}_j; \boldsymbol{\Psi}) \tag{2}$$

M. Gallagher, J. Hogan, and F. Maire (Eds.): IDEAL 2005, LNCS 3578, pp. 302–310, 2005.

for $i = 1, \ldots, g$. The mixture approach gives a probabilistic clustering in terms of these estimated posterior probabilities of component membership. An outright partitioning of the observations into g nonoverlapping clusters $C_1, \ldots, C_g$ is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. Thus the ith cluster C_i contains those observations assigned to the ith component.

The g-component normal mixture model (1) with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. In order for a non-singular estimate of a component-covariance matrix to be obtained, effectively $(p+1)$ observations need to be assigned to that component. Hence problems arise in the fitting of normal mixtures with unrestricted component-covariance matrices, especially if p is large relative to the number of observations n. In microarray experiments, for example, p can be several thousand while n may be no greater than 100 or so. This represents an extreme case; there can be problems for p as small as, say, 20 if n is not relatively large. Hence in practice, there is a need for methods that can handle the analysis of high-dimensional data.

Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. A common approach to reducing the the number of dimensions is to perform a principal component analysis (PCA). But projections of the feature data $\boldsymbol{y}_j$ onto the first few principal axes are not always useful in portraying the group structure; see McLachlan and Peel (2000a, Page 239). This point was also stressed by Chang (1983), who showed in the case of two groups that the principal component of the feature vector that provides the best separation between groups in terms of Mahalanobis distance is not necessarily the first component.

Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices is to adopt the mixture of factor analyzers model, as considered in McLachlan and Peel (2000a, 2000b). In this paper, we present an example to demonstrate further the differences between using principal components and mixtures of factor analyzers to cluster high-dimensional data. The example concerns the Cabernet wine data set of Ashenfelter (1999). In this data set, 32 judges ranked 46 wines from nine different countries on a scale of 1 to 46. All the information about the wines is known, but the judges refused to be identified.

2 Mixtures of Factor Analyzers

Factor analysis is commonly used for explaining data, in particular, correlations between variables in multivariate observations. It can be used also for dimensionality reduction, although the method of PCA is more widely used in this role. However, the effectiveness of these two statistical techniques is limited by their global linearity. A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels (factor analyzers) for the distribution of the full observation vector $\boldsymbol{Y}_j$. That is, with the mixture of factor analyzers model,

we can provide a local dimensionality reduction method by assuming that the distribution of the observation $\boldsymbol{Y}_j$ can be modelled by (1), where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{3}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$. The parameter vector $\boldsymbol{\Psi}$ now consists of the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$, along with the mixing proportions π_i $(i = 1, \ldots, g-1)$, on putting $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$. Unlike the PCA model, the factor analysis model (3) enjoys a powerful invariance property: changes in the scales of the feature variables in $\boldsymbol{y}_j$, appear only as scale changes in the appropriate rows of the matrix $\boldsymbol{B}_i$ of factor loadings.

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 jth observation in the ith component (see Section 8.7.4 in McLachlan and Peel, 2000b).

It can be seen that the mixture of factor analyzers model provides a way of controlling the number of parameters through the reduced model (3) for the component-covariance matrices. It thus provides a model intermediate between the independent and unrestricted models. The adequacy of the fit of a mixture of factor analyzers with q factors can be tested using the likelihood ratio statistic, as regularity conditions hold for tests on the value of q for a fixed number of components g. The model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm, whereby the single M-step of the EM algorithm is replaced by a number of computationally simpler conditional maximization (CM) steps and where the specification of the complete data is allowed to be different on each CM-step.

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.

3 Clustering of Wines

The list of wines is given in Table 1, with the numbering of the wine in the first column and the ranking of the wine in the final column. In the scatter plots, the wine numbers are prefixed by a two letter ISO country code, or by CA or WA when the wine is from California or Washington.

For these data there is interest in two clustering problems: the clustering of the wines on the basis of the (judges') scores and the clustering of the judges on the basis of their scores for the wines. We consider first the former problem, by fitting a mixture of $g = 2$ factor analyzers with $q = 2$ factors to the $n = 46$ wines on the basis of the $p = 32$ scores of the judges.

Table 1. List of wines with ranking in final column

1	Quilced Creek Cab. Sauv. '95	21
2	Chateau Latour '95	43
3	l'Ermita, Palacios (Priorat) '95	36
4	Chateau Ceval-Blanc '95	29
5	Ornellaia '94	39
6	Harlan Napa Valley Cab. Sauv. '94	11
7	Gallo Northern Sonoma Cab. Sauv. '94	14
8	Mitchelton Victoria Print Shiraz '95	24
9	Quintessa of Rutherford '95	33
10	Grans Muralles, Torres '96	20
11	Chateau Kefraya Comte de M Cuvee '96	10
12	Dalla Valle Napa Valley Maya '94	6
13	Grace Family Vineyard Napa Valley Cab. Sauv. '94	18
14	Henry Lagarde Lujan de Cuyo Syrah '95	45
15	Dave Nichol Stag's Leap Hillside Reserve Cab. Sauv.'91	37
16	Chateau Pichon-Longueville, Comtesse de Lalande5	28
17	Ridge Vineyards Monte Bello Red Table Wine '95	17
18	Plaisir de Merle Paarl Cab. Sauv.'95	41
19	Stag's Leap Wine Cellars Napa Valley Cab. Sauv.,Cask 23 '95	4
20	Arietta Napa Valley Red Table Wine '96	9
21	Sassicaia '94	34
22	Caymus Vineyard Napa Valley Cab. Sauv. '95	5
23	Chateau Lefite '95	40
24	Chateau Le Pin '95	NR
25	Longridge Hawkes Bay Merlot '95	46
26	Plumpjack Napa Valley Cab. Sauv. '95	13
27	Clark-Clauden Napa Valley Cab. Sauv. '95	1
28	Staglin Napa Valley Cab. Sauv., reserve '95	12
29	Chateau Margaux '95	23
30	Araugo napa Valley Cab. Sauv., Eisele Vineyard '94	3
31	Brant family Vineyard Napa Valley Cab. Sauv. '95	22
32	Chateau Los Boldos Cab. Sauv. Vieille Vignes '97	32
33	Beringer Napa Valley Cab. Sauv., Bancroft Vineyard '94	2
34	Cogin Napa Valley Cab. Sauv., Herb Lamb Vineyard '94	27
35	Penfold's Cab. Sauv., Bin 707 '90	25
36	Ridge Vineyards Geyersville Red Table Wine '95	16
37	Screaming Eagle Napa Valley Cab. Sauv. '95	8
38	Martinelli Jackass Hill Zinfandel '94	35
39	Chateau Petrus '95	26
40	De Lille Cellars Chaleur Estate Red Table Wine '94	19
41	Turley Napa Valley Zinfandel, Aida Vineyard '95	42
42	Chateau hout-Bion '95	31
43	Lionetti Cab. Sauv. '95	7
44	Forman Winery Napa Valley Cab. Sauv. '94	15
45	Tarapaca Maipo Valey Zavala Red Table Wine '96	30
46	Chateau Mouton-Rothschild '95	38
47	Veramonte Primus Casablanca Valley Merlot '96	44

A test of the value of the number of components g corresponding to the number of clusters in the data can be based on the likelihood ratio statistic λ. However, regularity conditions do not hold for $-2\log\lambda$ to have its usual (asymptotic) null distribution of chi-squared with degrees of freedom d equal to the difference between the number of parameters under the null and alternative hypotheses. Thus we adopted a resampling approach (McLachlan, 1987). On the basis of $B = 19$ replications of $-2\log\lambda$, we rejected the null hypothesis $H_0 : g = 1$ versus of 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

$$-2\log\lambda > d\log(n) \tag{4}$$

The case of unequal unrestricted component-covariance matrices was considered but rejected on the basis of BIC in favour of a common covariance matrix for the components. In the latter case, we can fit a mixture of two $p = 32$ dimensional normal components to the $n = 46$ wines, but we decided to work with a mixture of factor analyzers to keep the number of parameters down to a reasonable level compared to n. With the fitting of a mixture of $g = 2$ factor analyzers and $q = 2$ factors and equal component-covariance matrices, the wines fall into two clear groups. To illustrate the separation between the two clusters we plot the first canonical variate of the 46 wines in Figure 1. As there are only two groups, the canonical space is one-dimensional.

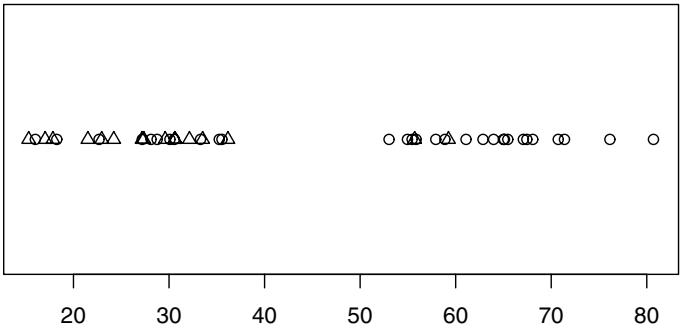

Fig. 1. First canonical variate of the 46 wines. Napa Valley wines are represented by triangles and other wines by circles

The choice of $q = 2$ factors was taken after the test of this value versus $q = 3$ was not significant according to the likelihood ratio statistic with null distribution taken to be chi-squared with 60 degrees of freedom. Although regularity conditions do not hold for the likelihood ratio test on the number of components g, they do for the number factors q for a given g. More specifically, for the test of the null hypothesis that $H_0 : q = q_0$ versus the alternative $H_1 : q = q_0 + 1$, the likelihood ratio statistic $-2\log\lambda$ under H_0 is asymptotically chi-squared with $d = g(p - q_0)$ degrees of freedom.

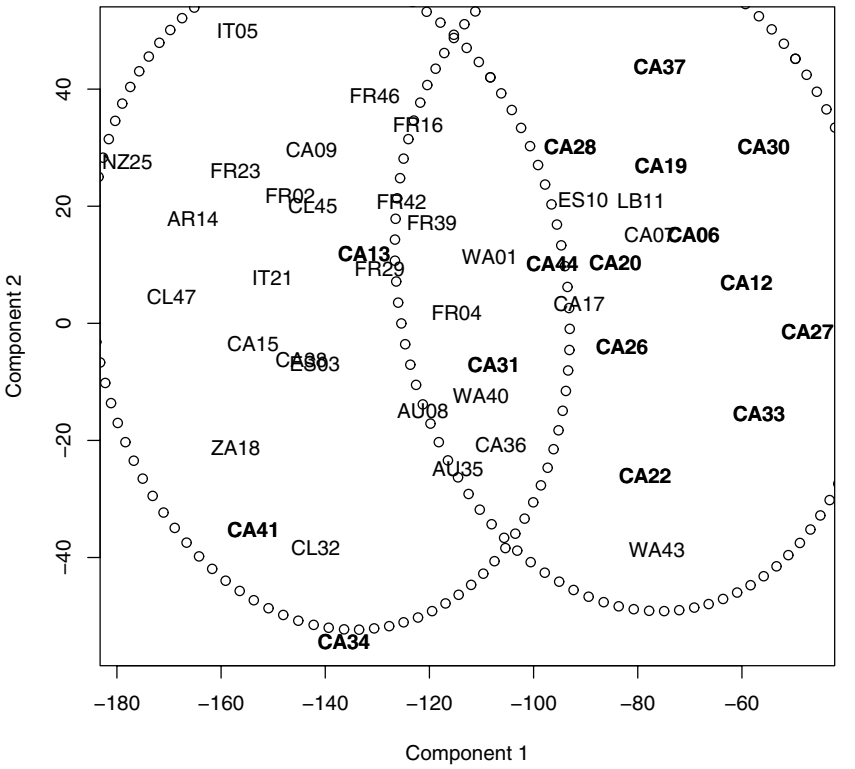

Fig. 2. Plot of the first two principal components of a PCA on the 46 wines. Here CA indicates California, WA indicates Washington, and any other two letter code is an ISO country code. The ellipses indicate the two groups given by fitting a mixture of two normals with equal covariance matrices to the PCs. Napa Valley wines are in bold

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. To this end, we display in Figure 2 a scatter plot of the first two PCs with the wine labels and implied clustering obtained fitting a mixture of two bivariate normals to these PCs. For comparative purposes, we give the clustering obtained by mixtures of factor analyzers in the space of the first two PCs in Figure 3. 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.

4 Cluster Analysis of Judge Scores

Young (2005) indicated that "there are three judges (16, 26, 31) that march to the beat of a different drummer." These atypical judges were 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.

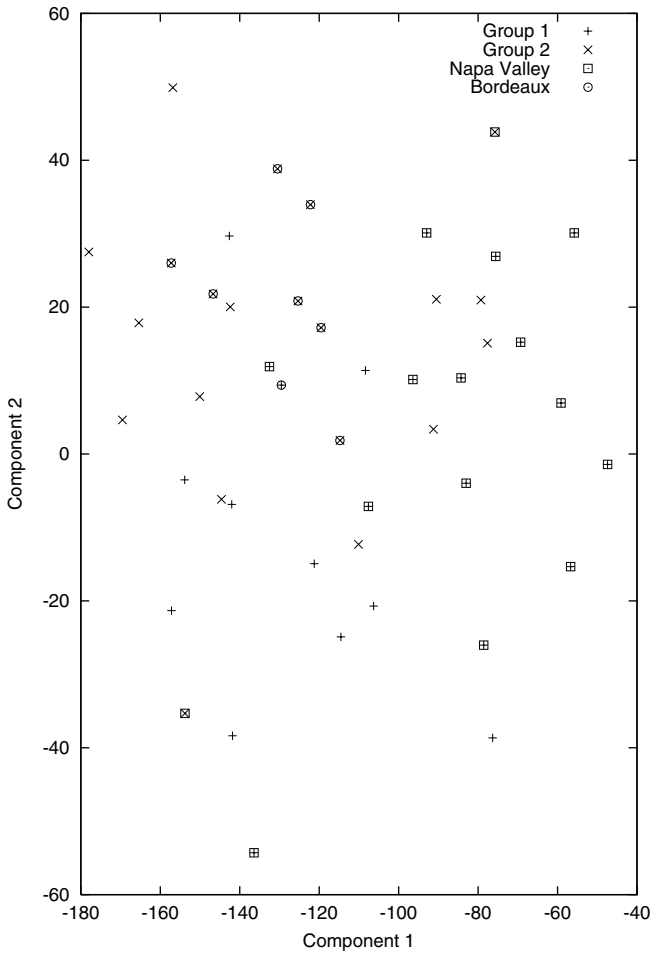

Fig. 3. Clusters obtained by mixtures of factor analyzers together with Napa Valley and Bordeaux wines

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. The plot is given in Figure 4.

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 $p = 46$. 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.

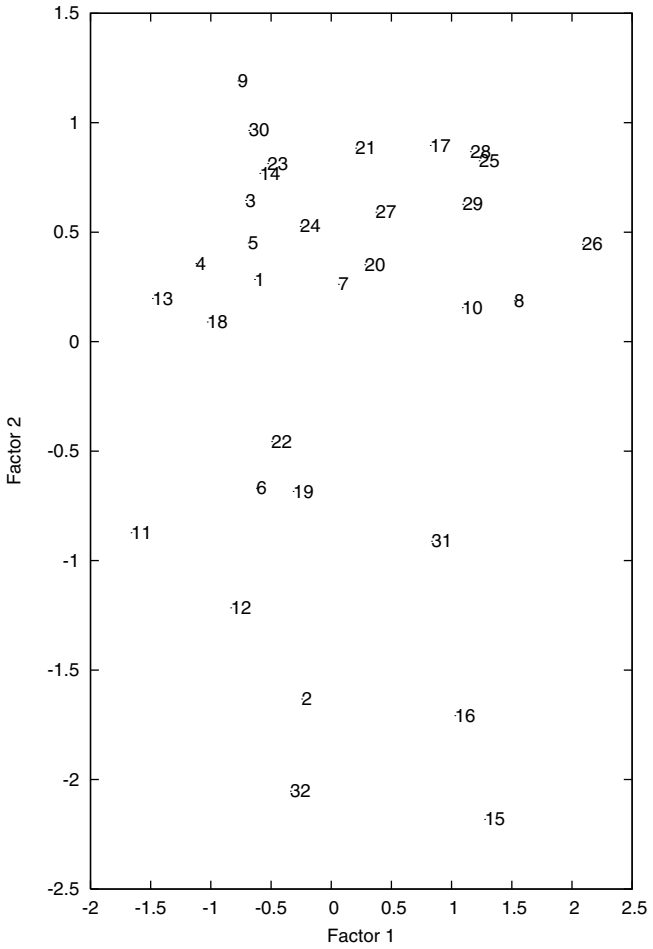

Fig. 4. Plot of the estimated posterior means of the $q = 2$ factors following a single-component factor analysis of the judge scores in the wine data set

References

Ashenfelter, O. (1999). California Versus All Challengers: The 1999 Cabernet Challenge. `http://www.liquidasset.com/report20.html`

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Chang, W.C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* **32**, 267–275.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Healy, M.J.R. (1986). Matrices for Statisticans. Clarendon: Oxford.

Liu, J., Feng, J., and Young, S.S. (2005). PowerMV v0.61. `http://www.niss.org/PowerMV/`

Liu, L., Hawkins, D.M., Ghosh, S., and Young, S.S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences USA* **100**, 13167–13172.

McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

McLachlan, G.J. and Peel, D. (2000a). Mixtures of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, P. Langley (Ed.). San Francisco: Morgan Kaufmann.

McLachlan, G.J. and Peel, D. (2000b). *Finite Mixture Models*. New York: Wiley.

McLachlan, G.J., Peel, D., and Bean, R.W. (2003). Modelling high-dimen-sional data by mixtures of factor analyzers. Comput. Statist. Data Anal. **41**, 379–388.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525.

Young, S. (2005). Private communication.