Parsimonious Gaussian Mixture Models *

Paul David McNicholas
Trinity College Dublin, Ireland

Thomas Brendan Murphy
Trinity College Dublin, Ireland

Abstract

Parsimonious Gaussian mixture models are developed using latent Gaussian models which are closely related to the factor analysis model. These models provide a unified modeling framework which includes the mixture of probabilistic principal component analyzers and mixture of factor of analyzers models as special cases.

Initially, the basic underlying factor analysis and probabilistic principal components analysis models are described and the use of the EM algorithm to find maximum likelihood estimates for these models is reviewed.

Then, a class of eight parsimonious Gaussian mixture models which are based on the mixtures of factor analyzers model are introduced and the maximum likelihood estimates for the parameters in these models are found using an AECM algorithm.

The use of these models is demonstrated on the analysis of chemical and physical properties of Italian wines; the models are shown to give excellent results and they reveal an interesting feature of these data.

Keywords: Mixture models, factor analysis, probabilistic principal components analysis, cluster analysis.

---

*

Paul David McNicholas is a graduate student, Department of Statistics, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland. E-mail: mcnichop@tcd.ie; Web: http://www.tcd.ie/Statistics/postgraduate/paulmcnicholas.shtml.
Thomas Brendan Murphy is a lecturer, Department of Statistics, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland. E-mail: murphybt@tcd.ie; Web: http://www.tcd.ie/Statistics/staff/brendanmurphy.shtml.
This research was funded by the SFI Basic Research Grant (04/BR/M0057). Part of this work was completed during a visit to the Center for Statistics in the Social Sciences which was supported by NIH grant (8 R01 EB002137- 02).
The authors would like to thank Prof. Adrian Raftery and the members of the Working Group on Model-Based Clustering at the University of Washington for useful suggestions and inputs into this work.

1 Introduction

Mixture models assume that data are collected from a finite collection of populations and that the data within each population can be modeled using a standard statistical model (e.g. Gaussian, Poisson or binomial). As a result, mixture models are particularly useful when modeling data collected from multiple sources. The Gaussian mixture model has received particular attention in the statistical literature; this model assumes a Gaussian structure for each population. In particular, the model density is of the form

$$f(x) = \sum_{g=1}^{G} \pi_g \phi(x|\mu_g, \mathbf{\Sigma}_g), \tag{1}$$

where $\phi(x|\mu_g, \mathbf{\Sigma}_g)$ is the density of a multivariate Gaussian with mean μ_g and covariance $\mathbf{\Sigma}_g$.

Extensive reviews of mixtures are contained in Titterington et al. (1985), McLachlan and Basford (1988) and McLachlan and Peel (2000); these books examine many different types of mixture models, but emphasis is given to Gaussian mixtures. Additionally, Fraley and Raftery (2002) provide an excellent review of Gaussian mixture models with applications to cluster analysis, discriminant analysis and density estimation.

The general Gaussian mixture model (1) has a total of $(G-1)+Gp+Gp(p+1)/2$ parameters of which $Gp(p+1)/2$ are used in the group covariance matrices $\mathbf{\Sigma}_g$. A simpler form of the mixture assumes that the covariances are constrained to be equal across groups, which reduces to a total of $(G-1)+Gp+p(p+1)/2$ parameters of which $p(p+1)/2$ are used for the common group covariance matrix $\mathbf{\Sigma}_g = \mathbf{\Sigma}$.

Banfield and Raftery (1993), Celeux and Govaert (1995) and Fraley and Raftery (1998, 2002) exploit an eigenvalue decomposition of the group covariance matrices to give a wide range of covariance structures that use between one and $Gp(p+1)/2$ parameters. The eigenvalue decomposition of the covariance matrix is of the form $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$ where λ_g is a constant, $\mathbf{D}_g$ is a matrix of eigenvectors of $\mathbf{\Sigma}_g$ and $\mathbf{A}_g$ is a diagonal matrix with entries proportional to the eigenvalues of $\mathbf{\Sigma}_g$. The model-based clustering methods that they have developed allow for constraining the components of the eigenvalue decomposition of $\mathbf{\Sigma}_g$ across components of the mixture model; such constraints are easily interpreted in terms of the contours of the densities of the mixture components. Fraley and Raftery (2002) demonstrate that the parsimonious mixture models derived in this model-based clustering framework give excellent results in clustering, discriminant analysis and density estimation applications.

We develop a new class of Gaussian mixture models with parsimonious covariance structure. These models are based on assuming a latent Gaussian model structure for each population; the latent Gaussian model is closely related to the mixture of factor analyzers model (Ghahramani and Hinton 1997). The mixture of factor analyzers model assumes a covariance structure of the form $\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g$, where the loading matrix $\mathbf{\Lambda}_g$ is a $(p \times q)$ matrix of parameters typically with $q \ll p$ and the noise matrix $\mathbf{\Psi}_g$ is a diagonal matrix; a detailed development of this covariance structure is given in Section 2. The loading and noise terms of the covariance matrix can be constrained to be equal or unequal across groups and the noise term can be further restricted to give a collection of eight parsimonious covariance structures. These covariance structures can have as few as $pq - q(q-1)/2 + 1$ parameters or as many as $G(pq - q(q-1)/2 + p)$ parameters with $q = 0, 1, 2, \ldots$. The full development

of the collection of parsimonious Gaussian mixture models and methods for fitting these models are given in Section 3.

The issue of model selection for the class of parsimonious Gaussian mixture models is addressed in Section 4.

In Section 5 we apply the models to the analysis of data recording chemical and physical properties of Italian wines collected from three areas of the Piedmont region. We find that the mixture models capture the three area structure of the data as well as a feature of the data collection process. In addition, the results of this analysis are compared with other methods of analysis.

We conclude, in Section 6 with a discussion of the methods and results of this work.

2 Latent Gaussian Models

In this section, we review the statistical methodology and describe it in terms of a latent Gaussian model (Section 2.1). The connection between factor analysis and probabilistic principal components analysis (Tipping and Bishop 1999b) is also shown. Methods for fitting factor analysis and probabilistic principal components analysis models using the EM algorithm are reviewed in Section 2.2.

2.1 Factor Analysis Model

Factor analysis (Spearman 1904) is a data reduction technique that aims to find unobserved factors that explain the variability in the data. The model (see Bartholomew and Knott 1999, Chapter 3) assumes that a p-dimensional random vector X is modeled using a q-dimensional vector of latent (unobserved) factors, U, where $q < p$.

The model is $X = \mu + \mathbf{\Lambda}U + \epsilon$ where $\mathbf{\Lambda}$ is a $p \times q$ matrix of factor loadings, the factors $U \sim N(0, \mathbf{I}_q)$ and $\epsilon \sim N(0, \mathbf{\Psi})$, where $\mathbf{\Psi} = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$. It follows from this model that the marginal distribution of X is $N(\mu, \mathbf{\Lambda\Lambda'} + \mathbf{\Psi})$.

The probabilistic principal components analysis model (Tipping and Bishop 1999b) is a special case of the factor analyzers model but it assumes that the distribution of the errors (ϵ) are isotropic so that $\mathbf{\Psi} = \psi\mathbf{I}_p = \text{diag}(\psi, \psi, \ldots, \psi)$.

Therefore, the density of observation x_i is

$$f(x_i) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Lambda\Lambda'} + \mathbf{\Psi}|^{1/2}} \exp\left\{-\frac{1}{2}(x_i - \mu)'(\mathbf{\Lambda\Lambda'} + \mathbf{\Psi})^{-1}(x_i - \mu)\right\}. \quad (2)$$

Hence,

$$\log f(x_i) = -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Lambda\Lambda'} + \mathbf{\Psi}| - \frac{1}{2}(x_i - \mu)'(\mathbf{\Lambda\Lambda'} + \mathbf{\Psi})^{-1}(x_i - \mu).$$

Therefore, under the *iid* assumption, the log-likelihood for data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$\begin{aligned}\sum_{i=1}^{n}\log f(x_i) &= -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\mathbf{\Lambda\Lambda'} + \mathbf{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)'(\mathbf{\Lambda\Lambda'} + \mathbf{\Psi})^{-1}(x_i - \mu) \\ &= -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\mathbf{\Lambda\Lambda'} + \mathbf{\Psi}| - \frac{n}{2}\,\text{tr}\left\{\mathbf{S}(\mathbf{\Lambda\Lambda'} + \mathbf{\Psi})^{-1}\right\},\end{aligned}$$

where $\mathbf{S} = (1/n)\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)'$ is the sample covariance matrix; in fact the data only appears in the model through $\mathbf{S}$.

Clearly, the maximum likelihood estimate of μ is $\hat{\mu} = \overline{x}$. Finding the maximum likelihood estimates for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ is more problematic. However, estimates of

these parameters can be found using the EM algorithm (Dempster et al. 1977), this approach is described in Section 2.2.

2.2 EM Algorithm for Factor Analysis Models

The EM algorithm (Dempster et al. 1977) for the factor analysis model involves two steps, the E-step involves computing the expected value of the complete-data log-likelihood and the M-step involves maximizing the expected complete-data log-likelihood. Following Dempster et al. (1977), we take $\mathbf{x}$ as the observed data and let $\mathbf{u}$ be missing data; the values of $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ being the unobserved latent values. The complete-data consist of $(\mathbf{x}, \mathbf{u})$, the observed and missing data.

The complete-data log-likelihood $l_c(\mu, \mathbf{\Lambda}, \mathbf{\Psi}) = \log f(\mathbf{x}, \mathbf{u})$ can be derived as follows. The density of observation x_i given the value of the underlying latent variable value u_i is

$$f(x_i|u_i) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Psi}|^{1/2}} \exp\left\{-\frac{1}{2}(x_i - \mu - \mathbf{\Lambda}u_i)'\mathbf{\Psi}^{-1}(x_i - \mu - \mathbf{\Lambda}u_i)\right\}.$$

Hence,

$$\begin{aligned}
\log f(x_i|u_i) &= -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Psi}| - \frac{1}{2}(x_i - \mu - \mathbf{\Lambda}u_i)'\mathbf{\Psi}^{-1}(x_i - \mu - \mathbf{\Lambda}u_i) \\
&= -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Psi}| - \frac{1}{2}(x_i - \mu)'\mathbf{\Psi}^{-1}(x_i - \mu) + (x_i - \mu)'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i \\
&\quad -\frac{1}{2}u_i'\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i \\
&= -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Psi}| - \frac{1}{2}(x_i - \mu)'\mathbf{\Psi}^{-1}(x_i - \mu) + (x_i - \mu)'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i \\
&\quad -\frac{1}{2}\operatorname{tr}\left\{\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i u_i'\right\} \\
&= -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Psi}| - \frac{1}{2}\operatorname{tr}\left\{\mathbf{\Psi}^{-1}(x_i - \mu)(x_i - \mu)'\right\} + (x_i - \mu)'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i \\
&\quad -\frac{1}{2}\operatorname{tr}\left\{\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i u_i'\right\}.
\end{aligned}$$

Recalling that $u_i \sim N(0, \mathbf{I}_q)$, we can now write down $l_c(\mu, \mathbf{\Lambda}, \mathbf{\Psi})$ as

$$\begin{aligned}
l_c(\mu, \mathbf{\Lambda}, \mathbf{\Psi}) = \log f(\mathbf{x}, \mathbf{u}) &= \sum_{i=1}^{n} \log[f(x_i \mid u_i) f(u_i)] \\
&= C - \frac{n}{2}\log|\mathbf{\Psi}| - \frac{1}{2}\operatorname{tr}\left\{\mathbf{\Psi}^{-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)'\right\} + \sum_{i=1}^{n}(x_i - \mu)'\mathbf{\Psi}^{-1}\mathbf{\Lambda}u_i \\
&\quad -\frac{1}{2}\operatorname{tr}\left\{\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}\sum_{i=1}^{n}u_i u_i'\right\},
\end{aligned}$$

where C is a constant function of $\mu, \mathbf{\Lambda}$, and $\mathbf{\Psi}$.

The expected value of u_i conditional on x_i and the current model parameters is

$$\mathbb{E}(u_i|x_i, \mu, \mathbf{\Lambda}, \mathbf{\Psi}) = \mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}(x_i - \mu) = \boldsymbol{\beta}(x_i - \mu),$$

where $\boldsymbol{\beta} = \mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}$.

The expected value of $u_i u_i'$ conditional on x_i and the current model parameters is

$$\mathbb{E}(u_i u_i'|x_i, \mu, \mathbf{\Lambda}, \mathbf{\Psi}) = \mathbf{I}_q - \boldsymbol{\beta}\mathbf{\Lambda} + \boldsymbol{\beta}(x_i - \mu)(x_i - \mu)'\boldsymbol{\beta}'.$$

The expected complete-data log-likelihood Q, evaluated with $\mu = \hat{\mu}$, is

$$\begin{aligned}
Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) &= C - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})'\right\} \\
&\quad + \sum_{i=1}^{n}(x_i-\hat{\mu})'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\mathbb{E}(u_i|x_i,\hat{\mu},\hat{\boldsymbol{\Lambda}},\hat{\boldsymbol{\Psi}}) \\
&\quad - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\sum_{i=1}^{n}\mathbb{E}(u_iu_i'|x_i,\hat{\mu},\hat{\boldsymbol{\Lambda}},\hat{\boldsymbol{\Psi}})\right\} \\
&= C - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})'\right\} \\
&\quad + \operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})'\right\} - \frac{n}{2}\operatorname{tr}\left\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Lambda}})\right\} \\
&\quad - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\left[\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})'\hat{\boldsymbol{\beta}}'\right]\right\} \\
&= C + \frac{n}{2}\log|\boldsymbol{\Psi}^{-1}| - \frac{n}{2}\operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\mathbf{S}\right\} + n\operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\mathbf{S}\right\} - \frac{n}{2}\operatorname{tr}\left\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}\right\},
\end{aligned}$$

where $\boldsymbol{\Theta} = \left(\mathbf{I}_q - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}\mathbf{S}\hat{\boldsymbol{\beta}}'\right)$ is a symmetric $q \times q$ matrix. We need to maximize Q with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ in the M-step of the EM algorithm.

Differentiating Q with respect to $\boldsymbol{\Lambda}$, utilizing results from Lütkepohl (1996, Chapter 10, Section 3), Magnus and Neudecker (1988, Chapter 9, Section 11) and Graybill (1983, Chapter 9, Section 1), gives

$$\begin{aligned}
S_1(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) &= \frac{\partial Q}{\partial \boldsymbol{\Lambda}} \\
&= n\,\frac{\partial}{\partial \boldsymbol{\Lambda}}\operatorname{tr}\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\mathbf{S}\} - \frac{n}{2}\,\frac{\partial}{\partial \boldsymbol{\Lambda}}\operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}\} \\
&= n(\boldsymbol{\Psi}^{-1})'(\hat{\boldsymbol{\beta}}\mathbf{S})' - \frac{n}{2}\,\frac{\partial}{\partial \boldsymbol{\Lambda}}\operatorname{tr}\{\boldsymbol{\Lambda}\boldsymbol{\Theta}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\} \\
&= n\boldsymbol{\Psi}^{-1}\mathbf{S}'\hat{\boldsymbol{\beta}}' - \frac{n}{2}\left[(\boldsymbol{\Psi}^{-1})'\boldsymbol{\Lambda}\boldsymbol{\Theta}' + \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}\right] \\
&= n\boldsymbol{\Psi}^{-1}\mathbf{S}\hat{\boldsymbol{\beta}}' - n\,\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}.
\end{aligned}$$

Solving the equation $S_1(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Psi}) = 0$ we obtain

$$\hat{\boldsymbol{\Lambda}} = \mathbf{S}\hat{\boldsymbol{\beta}}'\boldsymbol{\Theta}^{-1}. \tag{3}$$

Now, differentiating Q with respect to $\boldsymbol{\Psi}^{-1}$, using the same matrix differential results gives

$$\begin{aligned}
S_2(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) &= \frac{\partial Q}{\partial \boldsymbol{\Psi}^{-1}} \\
&= \frac{n}{2}\,\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\log|\boldsymbol{\Psi}^{-1}| - \frac{n}{2}\,\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\operatorname{tr}\{\boldsymbol{\Psi}^{-1}\mathbf{S}\} + n\,\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\operatorname{tr}\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\mathbf{S}\} \\
&\quad - \frac{n}{2}\,\frac{\partial}{\partial \boldsymbol{\Psi}^{-1}}\operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}\} \\
&= \frac{n}{2}\,\boldsymbol{\Psi} - \frac{n}{2}\,\mathbf{S}' + n\,(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\mathbf{S})' - \frac{n}{2}\,(\boldsymbol{\Lambda}')'(\boldsymbol{\Lambda}\boldsymbol{\Theta})' \\
&= \frac{n}{2}\,\boldsymbol{\Psi} - \frac{n}{2}\,\mathbf{S}' + n\,\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\mathbf{S} - \frac{n}{2}\,\boldsymbol{\Lambda}\boldsymbol{\Theta}'\boldsymbol{\Lambda}'.
\end{aligned}$$

Solving the equation $S_2(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}) = 0$ and recalling that $\hat{\mathbf{\Psi}}$ is a diagonal matrix, we obtain

$$\hat{\mathbf{\Psi}} = \operatorname{diag}\{\mathbf{S} - \hat{\mathbf{\Lambda}}\hat{\boldsymbol{\beta}}\mathbf{S}\}. \tag{4}$$

Hence, the maximum likelihood estimates for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ can be obtained by iteratively applying (3) and (4) and updating the values of $\hat{\boldsymbol{\beta}}$ and $\mathbf{\Theta}$ as required.

In the PPCA model, we find

$$\hat{\mathbf{\Psi}} = \frac{1}{p}\operatorname{tr}\{\mathbf{S} - \hat{\mathbf{\Lambda}}\hat{\boldsymbol{\beta}}\mathbf{S}\}.$$

3 Parsimonious Gaussian Mixture Models

Ghahramani and Hinton (1997) extended the factor analysis model (Section 2.1) by developing the mixture of factor analyzers model which assumes a mixture of Gaussian distributions model with a factor analysis covariance structure for each Gaussian component distribution; this work was further developed by McLachlan and Peel (2000). Additionally, Tipping and Bishop (1999a) developed a mixture of probabilistic principal components model.

Under the general mixture of factor analyzers model, the density of an observation in group g is of the form given in (2) with mean parameter μ_g, loading matrix $\mathbf{\Lambda}_g$ and noise matrix $\mathbf{\Psi}_g$. If we let the probability of membership of group g be π_g, then this leads to the mixture of factor analyzers model with density

$$f(x_i) = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}|\mathbf{\Psi}_g|^{1/2}} \exp\left\{-\frac{1}{2}(x_i - \mu_g - \mathbf{\Lambda}_g u_i)'\mathbf{\Psi}_g^{-1}(x_i - \mu_g - \mathbf{\Lambda}_g u_i)\right\}. \tag{5}$$

It is worth noting that the mixtures of factor analyzers model can differ in whether the $\mathbf{\Psi}_g$ term is constrained to be equal across groups or not. Ghahramani and Hinton (1997) assume equal noise and McLachlan and Peel (2000) and McLachlam et al. (2003) assume unequal noise, however they comment that assuming equal noise can give more stable results. In the context of the mixture of probabilistic principal components analyzers model, Tipping and Bishop (1999a) assume unequal, but isotropic, noise $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$.

We propose extending and unifying these Gaussian mixture models by allowing constraints across groups on the $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ matrices and on whether or not $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$. The full range of possible constraints provides a class of eight different parsimonious Gaussian mixture models (PGMM) (Table 1).

The Alternating Expectation Conditional Maximization (AECM) (Meng and VanDyk 1997) algorithm is used for fitting these models; McLachlan and Krishnan (1997) also give a detailed review of the AECM algorithm. This algorithm is an extension of the EM algorithm that uses different definitions of missing data at different stages. For the PGMM, when estimating π_g and μ_g the missing data are the unobserved group labels $\mathbf{z}$ and when estimating $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ the missing data are the group labels $\mathbf{z}$ and the unobserved latent factors $\mathbf{u}$.

At the first stage of the algorithm, when estimating π_g and μ_g, we let $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ be the group labels of the observations, where $z_{ig} = 1$ if observation i belongs to group g and $z_{ig} = 0$ otherwise. Hence, the complete-data likelihood for the mixture model is

$$L_1(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{g=1}^{G} \left[\pi_g f(x_i|\mu_g, \mathbf{\Lambda}_g, \mathbf{\Psi}_g)\right]^{z_{ig}}.$$

Table 1: Parsimonious covariance structures derived from the mixture of factor analyzers model.

ModelID	Loading Matrix	Error Variance	Isotropic	Covariance Parameters
CCC	Constrained	Constrained	Constrained	$\{pq - q(q-1)/2\} + 1$
CCU	Constrained	Constrained	Unconstrained	$\{pq - q(q-1)/2\} + p$
CUC	Constrained	Unconstrained	Constrained	$\{pq - q(q-1)/2\} + G$
CUU	Constrained	Unconstrained	Unconstrained	$\{pq - q(q-1)/2\} + Gp$
UCC	Unconstrained	Constrained	Constrained	$G\{pq - q(q-1)/2\} + 1$
UCU	Unconstrained	Constrained	Unconstrained	$G\{pq - q(q-1)/2\} + p$
UUC	Unconstrained	Unconstrained	Constrained	$G\{pq - q(q-1)/2\} + G$
UUU	Unconstrained	Unconstrained	Unconstrained	$G\{pq - q(q-1)/2\} + Gp$

Hence, the complete-data log-likelihood for the mixture model is

$$\begin{aligned} l_1(\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \left[\log \pi_g + \log f(x_i|\mu_g, \mathbf{\Lambda}_g, \mathbf{\Psi}_g)\right] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \left[\log \pi_g - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g| \right. \\ &\quad \left. -\frac{1}{2}\operatorname{tr}\left\{(x_i - \mu_g)(x_i - \mu_g)'(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}\right\}\right]. \end{aligned}$$

Hence, we find the expected complete-data log-likelihood is of the form

$$\begin{aligned} Q_1(\mu_g, \pi_g) &= \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2}\log 2\pi - \sum_{g=1}^{G}\frac{n_g}{2}\log|\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g| \\ &\quad -\sum_{g=1}^{G} n_g \operatorname{tr}\left\{\frac{1}{n_g}\sum_{i=1}^{n} z_{ig}(x_i - \mu_g)(x_i - \mu_g)'(\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g)^{-1}\right\} \\ &= \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2}\log 2\pi - \sum_{g=1}^{G}\frac{n_g}{2}\log|\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g| \\ &\quad -\sum_{g=1}^{G} n_g \operatorname{tr}\left\{\mathbf{S}_g(\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g)^{-1}\right\}, \end{aligned}$$

where

$$\hat{z}_{ig} = \frac{\hat{\pi}_g \phi(x_i|\hat{\mu}_g, \hat{\mathbf{\Lambda}}_g, \hat{\mathbf{\Psi}}_g)}{\sum_{g'=1}^{G}\hat{\pi}_{g'}\phi(x_i|\hat{\mu}_{g'}, \hat{\mathbf{\Lambda}}_{g'}, \hat{\mathbf{\Psi}}_{g'})}, \tag{6}$$

$n_g = \sum_{i=1}^{n}\hat{z}_{ig}$ and $\mathbf{S}_g = (1/n_g)\sum_{i=1}^{n}\hat{z}_{ig}(x_i - \mu_g)(x_i - \mu_g)'$.

Again, the values of $\mathbf{x}$ only appear in this function through $\mathbf{S}_g$. Maximizing the expected complete-data log-likelihood with respect to μ_g and π_g yields

$$\hat{\mu}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}x_i}{\sum_{i=1}^{n}\hat{z}_{ig}} \text{ and } \hat{\pi}_g = \frac{n_g}{n}.$$

At the second stage of the AECM algorithm, when estimating $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$, we take the group labels $\mathbf{z}$ and the latent factors $\mathbf{u}$ to be the missing data. Therefore,

the complete data log likelihood is

$$
\begin{aligned}
l_2(\mathbf{x}, \mathbf{z}, \mathbf{u}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\big[\log \pi_g + \log f(x_i|u_i, \mu_g, \mathbf{\Lambda}_g, \mathbf{\Psi}_g) + \log f(u_i)\big] \\
&= C + \sum_{g=1}^{G}\big[n_g \log \pi_g - \frac{n_g}{2}\log|\mathbf{\Psi}_g| - \frac{n_g}{2}\operatorname{tr}\big\{\mathbf{\Psi}_g^{-1}\mathbf{S}_g\big\} \\
&\quad + \sum_{i=1}^{n} z_{ig}(x_i - \mu_g)'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g u_i - \frac{1}{2}\operatorname{tr}\big\{\mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g \sum_{i=1}^{n} z_{ig}u_i u_i'\big\}\big],
\end{aligned}
$$

where C is a constant function of μ_g, $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$.

It follows that the expected complete-data log-likelihood evaluated with $\mu_g = \hat{\mu}_g$ and $\pi_g = \hat{\pi}_g$ is of the form

$$
\begin{aligned}
Q(\mathbf{\Lambda}_g, \mathbf{\Psi}_g) &= C + \sum_{g=1}^{G}\Bigg[-\frac{n_g}{2}\log|\mathbf{\Psi}_g| - \frac{n_g}{2}\operatorname{tr}\big\{\mathbf{\Psi}_g^{-1}\mathbf{S}_g\big\} \\
&\quad + \sum_{i=1}^{n}\hat{z}_{ig}(x_i - \hat{\mu}_g)'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\mathbb{E}\big[u_i|x_i, \hat{\mu}_g, \hat{\mathbf{\Lambda}}_g, \hat{\mathbf{\Psi}}_g\big] \\
&\quad - \frac{1}{2}\operatorname{tr}\Bigg\{\mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\sum_{i=1}^{n}\hat{z}_{ig}\mathbb{E}(u_i u_i'|x_i, \hat{\mu}_g, \hat{\mathbf{\Lambda}}_g, \hat{\mathbf{\Psi}}_g)\Bigg\}\Bigg] \\
&= C + \sum_{g=1}^{G}\Bigg[-\frac{n_g}{2}\log|\mathbf{\Psi}_g| - \frac{n_g}{2}\operatorname{tr}\big\{\mathbf{\Psi}_g^{-1}\mathbf{S}_g\big\} + \sum_{i=1}^{n}\hat{z}_{ig}(x_i - \hat{\mu}_g)'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\hat{\boldsymbol{\beta}}_g(x_i - \hat{\mu}_g) \\
&\quad - \frac{1}{2}\operatorname{tr}\Bigg\{\mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\sum_{i=1}^{n}\hat{z}_{ig}\mathbb{E}(u_i u_i'|x_i, \hat{\mu}_g, \hat{\mathbf{\Lambda}}_g, \hat{\mathbf{\Psi}}_g)\Bigg\}\Bigg] \\
&= C + \sum_{g=1}^{G} n_g\Bigg[\frac{1}{2}\log|\mathbf{\Psi}_g^{-1}| - \frac{1}{2}\operatorname{tr}\big\{\mathbf{\Psi}_g^{-1}\mathbf{S}_g\big\} + \operatorname{tr}\big\{\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\big\} \\
&\quad - \frac{1}{2}\operatorname{tr}\big\{\mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g\mathbf{\Theta}_g\big\}\Bigg],
\end{aligned}
$$

where $\mathbf{\Theta}_g = (\mathbf{I}_q - \hat{\boldsymbol{\beta}}_g\hat{\mathbf{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{S}_g\hat{\boldsymbol{\beta}}_g')$ is a symmetric $q \times q$ matrix and the $\hat{z}_{ig}$ are computed as in (6) with the estimates of $\hat{\mu}_g$ and $\hat{\pi}_g$ as calculated in the first stage of the AECM algorithm. The resulting estimates when we impose constraints (Table 1) on the $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ matrices can be easily derived from the expression for $Q(\mathbf{\Lambda}_g, \mathbf{\Psi}_g)$ given above.

The first stage of the AECM where μ_g and π_g are estimated and the second stage where $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ are estimated are iterated until convergence. The resulting values give maximum likelihood estimates of the parameters in the model. The resulting $\hat{z}_{ig}$ values are estimates of the *a posteriori* probability of group membership for each observation.

We illustrate how the maxima of $Q(\mathbf{\Lambda}_g, \mathbf{\Psi}_g)$ are computed for the CCC (Section 3.1) and CUU (Section 3.2) cases. The remaining cases follow the same principles and are excluded to save space. The CUU case is included because it is more difficult than any other case. In any case, the resulting calculations required to compute the estimates for every one of the models are given in Appendix A.

We adopt the following notation:

$$
\tilde{\mathbf{S}} = \sum_{g=1}^{G}\hat{\pi}_g\mathbf{S}_g,
$$

$$\tilde{\boldsymbol{\Theta}} = (\mathbf{I}_q - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}').$$

3.1 Model CCC

For Model CCC we have $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi} = \psi\mathbf{I}_p$. Therefore, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\psi}\mathbf{I}_p)^{-1}$ and it follows that $\tilde{\mathbf{S}}$ is the sample covariance matrix (instead of $\mathbf{S}_g$) and so $\boldsymbol{\Theta}_g$ is replaced by $\tilde{\boldsymbol{\Theta}}$. Therefore, Q can be written as

$$\begin{aligned} Q(\boldsymbol{\Lambda}, \psi) &= C + n\left[\frac{1}{2}\log|\psi^{-1}\mathbf{I}_p| - \frac{1}{2}\operatorname{tr}\{\psi^{-1}\mathbf{I}_p\tilde{\mathbf{S}}\} + \operatorname{tr}\{\psi^{-1}\mathbf{I}_p\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\} - \frac{1}{2}\operatorname{tr}\{\boldsymbol{\Lambda}'\psi^{-1}\mathbf{I}_p\boldsymbol{\Lambda}\tilde{\boldsymbol{\Theta}}\}\right] \\ &= C + \frac{n}{2}\left[p\log\psi^{-1} - \psi^{-1}\operatorname{tr}\{\tilde{\mathbf{S}}\} + 2\psi^{-1}\operatorname{tr}\{\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\} - \psi^{-1}\operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\tilde{\boldsymbol{\Theta}}\}\right]. \end{aligned}$$

Now, we can maximize Q with respect to $\boldsymbol{\Lambda}$ and ψ to get maximum likelihood estimates of $\boldsymbol{\Lambda}$ and ψ. Differentiating Q with respect to $\boldsymbol{\Lambda}$ and ψ^{-1} respectively gives the score functions below.

$$\begin{aligned} S_1(\boldsymbol{\Lambda}, \psi) &= \frac{\partial Q(\boldsymbol{\Lambda}, \psi)}{\partial \boldsymbol{\Lambda}} = \psi^{-1}n\left[\tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}' - \boldsymbol{\Lambda}\boldsymbol{\Theta}\right] \\ S_2(\boldsymbol{\Lambda}, \psi) &= \frac{\partial Q(\boldsymbol{\Lambda}, \psi)}{\partial \psi^{-1}} = \frac{n}{2}\left[p\psi - \operatorname{tr}\{\tilde{\mathbf{S}}\} + 2\operatorname{tr}\{\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\} - \operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\tilde{\boldsymbol{\Theta}}\}\right] \end{aligned}$$

Solving $S_1(\hat{\boldsymbol{\Lambda}}^{\text{new}}, \hat{\psi}) = 0$ for $\hat{\boldsymbol{\Lambda}}^{\text{new}}$gives

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Theta}}^{-1}.$$

Solving $S_2(\hat{\boldsymbol{\Lambda}}^{\text{new}}, \hat{\psi}^{\text{new}}) = 0$ for $\hat{\psi}^{\text{new}}$ we obtain

$$\begin{aligned} \hat{\psi}^{\text{new}} &= \frac{1}{p}\operatorname{tr}\{\tilde{\mathbf{S}}' - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}} + \hat{\boldsymbol{\Lambda}}^{\text{new}}\tilde{\boldsymbol{\Theta}}'(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\} \\ &= \frac{1}{p}\operatorname{tr}\{\tilde{\mathbf{S}}' - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}} + \hat{\boldsymbol{\Lambda}}^{\text{new}}\tilde{\boldsymbol{\Theta}}'(\tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Theta}}^{-1})'\} \\ &= \frac{1}{p}\operatorname{tr}\{\tilde{\mathbf{S}}' - \hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\}. \end{aligned}$$

3.2 Model CUU

For Model CUU we have $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$. Therefore, $\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}_g)^{-1}$ and $\boldsymbol{\Theta}_g = (\mathbf{I}_q - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}_g\mathbf{S}_g\hat{\boldsymbol{\beta}}_g')$, so that Q can be written as

$$Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g) = C + \sum_{g=1}^{G} n_g\left[\frac{1}{2}\log|\boldsymbol{\Psi}_g^{-1}| - \frac{1}{2}\operatorname{tr}\{\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g\} + \operatorname{tr}\{\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\} - \frac{1}{2}\operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g\}\right].$$

Now, we maximize Q to get maximum likelihood estimates of the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g$. Differentiating Q with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g^{-1}$ respectively gives the score functions

$$\begin{aligned} S_1(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g) &= \frac{\partial Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Lambda}} = \sum_{g=1}^{G} n_g\left[\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g' - \boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g\right], \\ S_2(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g) &= \frac{\partial Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g)}{\partial \boldsymbol{\Psi}_g^{-1}} = n_g\left[\frac{1}{2}\boldsymbol{\Psi}_g - \frac{1}{2}\mathbf{S}_g' + \boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g - \frac{1}{2}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g'\boldsymbol{\Lambda}'\right]. \end{aligned}$$

Unfortunately, we must solve $S_1\big(\hat{\boldsymbol{\Lambda}}^{\text{new}}, \hat{\boldsymbol{\Psi}}_g\big) = 0$ for $\hat{\boldsymbol{\Lambda}}^{\text{new}}$ in a row-by-row manner. We obtain

$$(\hat{\boldsymbol{\Lambda}}^{\text{new}})_j = \left[\sum_{g=1}^{G} \frac{n_g}{\hat{\psi}_{g(j)}} (\mathbf{S}_g \hat{\boldsymbol{\beta}}_g')_j\right] \left[\sum_{g=1}^{G} \frac{n_g}{\hat{\psi}_{g(j)}} \boldsymbol{\Theta}_g\right]^{-1},$$

where $\hat{\psi}_{g(j)}$ is the j^{th} element along the diagonal of $\hat{\boldsymbol{\Psi}}_g$ and $j = 1, 2, \ldots, p$.

We solve $S_2\big(\hat{\boldsymbol{\Lambda}}^{\text{new}}, \hat{\boldsymbol{\Psi}}_g^{\text{new}}\big) = 0$ for $\hat{\boldsymbol{\Psi}}_g^{\text{new}}$ to get

$$\begin{aligned} & n_g \left[\frac{1}{2}\hat{\boldsymbol{\Psi}}_g^{\text{new}} - \frac{1}{2}\mathbf{S}_g' + \hat{\boldsymbol{\Lambda}}^{\text{new}} \hat{\boldsymbol{\beta}}_g \mathbf{S}_g - \frac{1}{2}\hat{\boldsymbol{\Lambda}}^{\text{new}} \boldsymbol{\Theta}_g' (\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right] = 0 \\ \Rightarrow \quad & \hat{\boldsymbol{\Psi}}_g^{\text{new}} = \mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}} \hat{\boldsymbol{\beta}}_g \mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}} \boldsymbol{\Theta}_g (\hat{\boldsymbol{\Lambda}}^{\text{new}})'. \end{aligned}$$

But $\boldsymbol{\Psi}_g^{\text{new}}$ is a diagonal matrix, therefore

$$\hat{\boldsymbol{\Psi}}_g^{\text{new}} = \text{diag}\left\{\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}} \hat{\boldsymbol{\beta}}_g \mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}} \boldsymbol{\Theta}_g (\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right\}.$$

4 Model Selection & Performance

The Bayesian Information Criterion (Schwartz 1978) is proposed for selecting an appropriate parsimonious Gaussian mixture model.

For a model with parameters θ, the Bayesian Information Criterion (BIC) is given by

$$\text{BIC} = 2l(x, \hat{\theta}) - m \log n,$$

where $l(x, \hat{\theta})$ is the maximized log-likelihood, $\hat{\theta}$ is the maximum likelihood estimate of θ and m is the number of free parameters in the model. The use of the BIC can be motivated through an asymptotic approximation of the log posterior probability of the models (Kass and Raftery 1995). The usual regularity conditions for the asymptotic approximation used in the development of BIC are not generally satisfied by mixture models. However, Keribin (1998, 2000) shows that BIC gives consistent estimates of the number of components of a mixture model. In addition, Fraley and Raftery (1998, 2002) provide practical evidence that BIC performs well as a model selection criterion for mixture models.

We propose using BIC to choose the model type (CCC, CCU, ..., UUU), the number of latent factors (q) and the number of components (G) for the parsimonious Gaussian mixture models. In addition, BIC can be used to choose between the PGMM and model-based clustering as implemented by `mclust`.

The performance of the PGMM in revealing group structures in data is measured using the Rand index (Rand 1971) and Adjusted Rand index (Hubert and Arabie 1985). These indices are computed on a cross tabulation of the maximum *a posteriori* (MAP) classification of the observations with the true group membership. Large values of these indices indicate strong agreement between the true groupings and the classifications proposed by the mixture model. The adjusted Rand index corrects the Rand index for agreements between the MAP classification and the true groupings which occur by chance.

5 Application: Italian Wines

Forina et al. (1986) used data recording twenty eight chemical and physical properties of wines from the Piedmont region of Italy to classify the wines to their geographic origin (Barolo, Grignolino, Barbera). A subset of thirteen variables (Table

2) are commonly available in the UCI Machine Learning data repository (Newman et al. 1998) and as part of the `gclus` library (Hurley 2004) for `R` (R Development Core Team 2004).

Table 2: Thirteen of the chemical and physical properties of the Italian wine used in this study.

Chemical Properties		
Alcohol	Malic Acid	Ash
Alcalinity of Ash	Magnesium	Total Phenols
Flavanoids	Nonflavanoid Phenols	Proanthocyanins
Color Intensity	Hue	OD280/OD315 Of Diluted Wines
Proline		

Forina et al. (1986), who originally analyzed these data, describe the data collection process in some detail in their paper. It is worth observing that the wines in this study were collected over the time period of 1970–1979 (Table 3). Furthermore, the wines were collected over a ten year period and the Barbera wines are predominantly from a period which is later than the Barolo or Grignolino wines.

Table 3: A table showing the year of production of the wine samples in the data.

	Year of Production									
	70	71	72	73	74	75	76	77	78	79
Barolo		19		20	20					
Grignolino	9	9	7	9	16	9	12			
Barbera					9		5		29	5

The parsimonious Gaussian mixture models are demonstrated on the analysis of these Italian wines (Section 5.1). Alternative methods of analysis are shown in Section 5.2 and the methods are compared in Section 5.3.

5.1 Parsimonious Gaussian Mixture Models

All eight parsimonious Gaussian mixture models (CCC, CCU, ..., UUU) were fitted to the data for $G = 1, 2, \ldots, 8$ and $q = 1, 2, \ldots, 5$. Hence, a total of 320 different parsimonious Gaussian mixture models were fitted to the data. The BIC value for each model was computed and the model with the highest BIC was selected; this was the CUU model with $G = 4$ and $q = 2$. Figure 1 shows the maximum BIC for the eight parsimonious models for each pair (G, q).

While $q = 0$ is possible, the PGMM with $q = 0$ correspond to the `mclust` model-based clustering models (EII, EEI, VII and VVI), so the case where $q = 0$ is not considered at this point but it is considered in Section 5.2.1.

A cross tabulation of the MAP classifications from the best PGMM versus the true wine type is given in Table 4. This model classifies the Barolo and Barbera observations exactly into Cluster 1 and Cluster 4 respectively but Grignolino is spread largely across Cluster 2 and Cluster 3 with a couple of observations in Cluster 4.

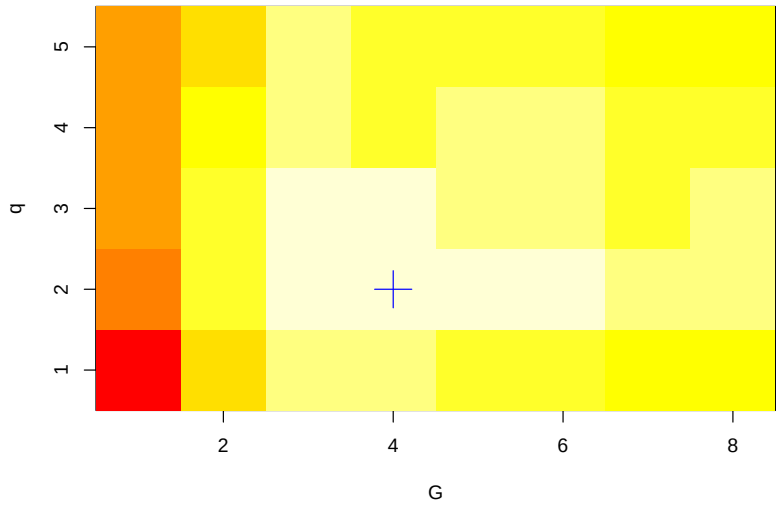

Figure 1: A heat map representation of the maximum BIC value for each value of (G, q) where the maximum is taken over the eight models (CCC,CCU,..., UUU).

Table 4: A classification table for the PGMM with highest BIC value. The table shows a cross tabulation of the MAP classification with the true wine type.

	Cluster			
	1	2	3	4
Barolo	59			
Grignolino		38	31	2
Barbera				48

The fact that the model is picking up an extra group may not actually be an error; this is explored in more detail below.

A cross tabulation of the MAP classification results for the PGMM versus wine type and year is given in Table 5. The table shows that Cluster 2 consists primarily of Grignolino wines from 1971–1974 and Cluster 3 consists primarily of Grignolino wines from 1974–1976. Therefore, the mixture model appears to be dividing the Grignolino wines into two groups according to year of production.

Incidentally, due to the data collection process, it is not possible to completely determine if the correspondence of the Barbera wines and Cluster 4 is due to the wine type or the year of production. The Barbera wines were primarily from 1978 and 1979 whereas the other wine types were from 1970–1976.

5.2 Alternative Methods

5.2.1 Model-Based Clustering (`mclust`)

Model-based clustering was also completed on the wine data using the `mclust` software (Fraley and Raftery 1999, 2003) for R (R Development Core Team 2004). The model with the highest BIC value (BIC = −5470.0) is an eight component mixture with the VEI covariance structure; detailed explanations of the covariance structures are given in Fraley and Raftery (2002). The VEI covariance is a diagonal covariance structure which implies elliptical group structure with the equally shaped, but

Table 5: A table of the MAP classifications from the PGMM versus the wine type and year.

	Barolo			**Grignolino**							**Barbera**			
	71	73	74	70	71	72	73	74	75	76	74	76	78	79
Cluster 1	19	20	20											
Cluster 2				7	8	4	8	8	2	1				
Cluster 3				2	1	2	1	8	7	10				
Cluster 4						1				1	9	5	29	5

different sized ellipses that are aligned to the axes.

The MAP classifications were calculated for the eight component VEI model; a cross tabulation of the classifications and the wine types is shown in Table 6.

Table 6: The classification table for the best model found using `mclust`.

	Cluster							
	1	2	3	4	5	6	7	8
Barolo	40	18	1					
Grignolino			21	22	27	1		
Barbera						17	4	27

While model-based clustering found eight groups, it is clear that multiple groups are being used within each wine type. Only Cluster 3 and Cluster 7 contain wines of more than one type. In these cases, there is only a single wine of a different type.

Furthermore, applying the information in Table 3 to the results of the `mclust` model gave some insight into why there were eight groups in the data. However, although Table 7 does show a block diagonal structure, these results do not yield an entirely effective clustering.

A cross tabulation of cluster membership versus wine type and year of production (Table 7) shows that the subgroups found by model-based clustering are not based on year of production.

Table 7: `mclust` classifications of wines produced by year.

	Barolo			**Grignolino**							**Barbera**			
	71	73	74	70	71	72	73	74	75	76	74	76	78	79
Cluster 1	9	17	14											
Cluster 2	10	2	6											
Cluster 3		1		3	4	2		3	3	6				
Cluster 4				1		2	3	9	4	3				
Cluster 5				5	5	2	6	4	2	3				
Cluster 6						1					5	5	7	
Cluster 7											4			
Cluster 8													22	5

Table 8: Classification table for the model-based clustering with variable selection method.

	Cluster		
	1	2	3
Barolo	51		8
Grignolino	4	64	3
Barbera		1	47

5.2.2 Model-based Clustering with Variable Selection

Raftery and Dean (2004) propose a version of model-based clustering with variable selection. This method was applied to the wine data and the results of this analysis are given in Table 8. Interestingly, only four of the variables were selected: Malic Acid, Proline, Flavanoids and Color Intensity.

While the model-based clustering with variable selection methods does select three clusters, there is some overlap between the clusters and the wine types; no wine type is uniquely assigned to one cluster nor is any cluster only of one wine type.

5.3 Model Comparison

An informal comparison of the clustering results on the wine data indicates that all of the methods are very good at discovering the group structure in the wine data. The model-based clustering with variable selection selects the correct number of groups. The PGMM and `mclust` methods separate the wines into subgroups with the subgroups being predominantly wines of one type.

The BIC values of the PGMM and `mclust` results can be compared and the BIC is higher for the PGMM. BIC cannot be used to choose between the model-based clustering with variable selection model and the other models, this issue is discussed in Raftery and Dean (2004).

The clustering performance of all of the models can be quantified using the Rand Index and Adjusted Rand Index, as shown in Table 9, and the PGMM has the highest scores for both of these criteria.

Table 9: Rand Index, Adjusted Rand Index and BIC values for the PGMM, `mclust` and model-based clustering with variable selection models.

Model	Rand Index	Adjusted Rand	BIC
PGMM	0.91	0.79	-5305.5
`mclust`	0.80	0.48	-5470.0
Variable Selection	0.88	0.74	

Hence, even though the PGMM found four groups, the Rand and Adjusted Rand indices show that the MAP classification for this model is in closer agreement to the true wine types than the other two methods. In addition, the BIC value suggests that the PGMM should be used instead of the `mclust` model.

6 Discussion

A new family of parsimonious Gaussian mixture models (PGMM) has been proposed. These models are closely related to the mixture of factor analyzers and mixture of principal components analyzers models and includes them as special cases.

The number of covariance parameters in the PGMM grows linearly with data dimension which is especially important in high-dimensional situations. In model-based clustering as implemented in `mclust`, the number of parameters in any of the diagonal covariance structures is linear (or constant) in the data dimension, however the number of parameters in the non-diagonal covariance structures is quadratic in data dimension. This means that PGMM may offer a more flexible modeling structure for high-dimensional data than `mclust`.

Chang (1983) shows that the principal components corresponding to the highest eigenvalues do not necessarily contain the most group information. Directly modeling data using the PGMM avoids the problems of implementing data reduction using principal components analysis before clustering.

The application of the PGMM to the wine data indicates that the models gives excellent clustering performance. The clusters found using the model showed greater ability to capture the group structure of the data than other methods. Interestingly, choosing between PGMM and model-based clustering with `mclust` using the Bayesian Information Criterion selected PGMM which gave superior clustering performance on these data; this is in agreement with the results (Fraley and Raftery 2002) that show that choosing models within the `mclust` framework using BIC give models with good classification and clustering performance.

A Estimation Procedure for Eight Covariance Structures

The resulting equations for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Lambda}}$ and $\hat{\Psi}$ for each of the eight parsimonious models are given in the following sections.

A.1 Model CCC

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\psi}\mathbf{I}_p)^{-1}$$

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Theta}}^{-1}$$

$$\hat{\psi}^{\text{new}} = \frac{1}{p}\operatorname{tr}\left\{\tilde{\mathbf{S}}' - \hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\right\}$$

A.2 Model CCU

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}})^{-1}$$

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \tilde{\mathbf{S}}\hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Theta}}^{-1}$$

$$\hat{\boldsymbol{\Psi}}^{\text{new}} = \operatorname{diag}\left\{\tilde{\mathbf{S}} - \hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}\tilde{\mathbf{S}}\right\}$$

A.3 Model CUC

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\psi}_g\mathbf{I}_p)^{-1}$$

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_g}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\right]\left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_g}\boldsymbol{\Theta}_g\right]^{-1}$$

$$\hat{\psi}_g^{\text{new}} = \frac{1}{p}\operatorname{tr}\left\{\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right\}$$

A.4 Model CUU

We solve for the loading matrix in a row-by-row manner.

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}_g)^{-1}$$

$$(\hat{\boldsymbol{\Lambda}}^{\text{new}})_j = \left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_{g(j)}}(\mathbf{S}_g\hat{\boldsymbol{\beta}}_g')_j\right]\left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_{g(j)}}\boldsymbol{\Theta}_g\right]^{-1}$$

$$\hat{\boldsymbol{\Psi}}^{\text{new}} = \operatorname{diag}\left\{\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right\}$$

Where $(\hat{\boldsymbol{\Lambda}}^{\text{new}})_j$ is the jth row of the loading matrix and $\hat{\psi}_{g(j)}$ is the j^{th} element along the diagonal of $\hat{\boldsymbol{\Psi}}_g$.

A.5 Model UCC

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g'(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g' + \hat{\psi}\mathbf{I}_p)^{-1}$$

$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\boldsymbol{\Theta}_g^{-1}$$

$$\hat{\psi}^{\text{new}} = \frac{1}{p}\sum_{g=1}^{G}\hat{\pi}_g\operatorname{tr}\left\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right\}$$

A.6 Model UCU

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g'(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g' + \hat{\boldsymbol{\Psi}})^{-1}$$

$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\boldsymbol{\Theta}_g^{-1}$$

$$\hat{\boldsymbol{\Psi}}^{\text{new}} = \sum_{g=1}^{G}\hat{\pi}_g\operatorname{diag}\left\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right\}$$

A.7 Model UUC

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g'(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g' + \hat{\psi}_g\mathbf{I}_p)^{-1}$$

$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\boldsymbol{\Theta}_g^{-1}$$

$$\hat{\psi}_g^{\text{new}} = \frac{1}{p}\operatorname{tr}\left\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right\}$$

A.8 Model UUU

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g'(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g' + \hat{\boldsymbol{\Psi}}_g)^{-1}$$
$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\boldsymbol{\Theta}_g^{-1}$$
$$\hat{\boldsymbol{\Psi}}_g^{\text{new}} = \text{diag}\left\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right\}$$

References

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

Bartholomew, D. J. and Knott, M. (1999), *Latent variable models and factor analysis*, London: Edward Arnold, 2nd ed.

Celeux, G. and Govaert, G. (1995), "Gaussian parsimonious clustering models," *Pattern Recognition*, 28, 781–793.

Chang, W.-C. (1983), "On using principal components before separating a mixture of two multivariate normal distributions," *Journal of the Royal Statistical Society. Series C. 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," *Journal of the Royal Statistical Society. Series B*, 39, 1–38, with discussion.

Forina, M., Armanino, C., Castino, M., and Ubigli, M. (1986), "Multivariate data analysis as a discriminating method of the origin of wines," *Vitis*, 25, 189–201.

Fraley, C. and Raftery, A. E. (1998), "How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis," *Computer Journal*, 41, 578–588.

— (1999), "MCLUST: Software for model-based clustering," *Journal of Classification*, 16, 297–306.

— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–612.

— (2003), "Enhanced model-based clustering, density estimation and discriminant analysis software: MCLUST," *Journal of Classification*, 20, 263–296.

Ghahramani, Z. and Hinton, G. E. (1997), "The EM algorithm for factor analyzers," Tech. Rep. CRG-TR-96-1, University Of Toronto, Toronto.

Graybill, F. A. (1983), *Matrices with applications in statistics*, Belmont, California: Wadsworth, 2nd ed.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Hurley, C. (2004), "Clustering visualizations of multivariate data," *Journal of Computational and Graphical Statistics*, 13, 788–806.

Kass, R. E. and Raftery, A. E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

Keribin, C. (1998), "Estimation consistante de l'ordre de modèles de mélange," *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique*, 326, 243–248.

— (2000), "Consistent estimation of the order of mixture models," *Sankhyā. The Indian Journal of Statistics. Series A*, 62, 49–66.

Lütkepohl, H. (1996), *Handbook of Matrices*, Chicester: John Wiley & Sons.

Magnus, J. R. and Neudecker, H. (1988), *Matrix differential calculus with applications in statistics and econometrics*, Chicester: John Wiley & Sons.

McLachlam, G. J., Peel, D., and Bean, R. W. (2003), "Modelling high-dimensional data by mixtures of factor analyzers," *Computational Statistics and Data Analysis*, 41, 379–388.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, New York: Marcel Dekker Inc.

McLachlan, G. J. and Krishnan, T. (1997), *The EM algorithm and extensions*, New York: John Wiley & Sons Inc.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture models*, New York: John Wiley & Sons.

Meng, X.-L. and VanDyk, D. (1997), "The EM algorithm – an old folk song sung to a new fast tune (with discussion)," *Journal of the Royal Statistical Society, Series B*, 59, 511–567.

Newman, D. J., Hettich, S., Blake, C. L., and Merz, C. J. (1998), "UCI Repository of machine learning databases," .

R Development Core Team (2004), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raftery, A. E. and Dean, N. (2004), "Variable Selection for Model-Based Clustering," Tech. Rep. 452, Department of Statistics, University of Washington, Seattle, Washington.

Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 31–38.

Spearman, C. (1904), "The proof and measurement of association between two things," *American Journal of Psychology*, 15, 72–101.

Tipping, M. E. and Bishop, C. M. (1999a), "Mixtures of probabilistic principal component analysers," *Neural Computation*, 11, 443–482.

— (1999b), "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society Series B*, 61, 611–622.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Chichester: John Wiley & Sons.