Model-Based Clustering: An Overview

Paul McNicholas

Department of Mathematics & Statistics, University of Guelph.

Statistics Seminar, McMaster University, October 23, 2007.
Overview

- This talk will focus on model-based clustering via Gaussian mixture models.

- Model-based clustering and Gaussian mixture models are introduced.

- Popular techniques are reviewed.

- New techniques are introduced and demonstrated on real data.
Learning is that process by which knowledge is gained.

Statistical learning can be either supervised or unsupervised.

Models are said to learn in a ‘supervised’ fashion, when the outcome variable is present.

In an ‘unsupervised’ learning situation, the outcome variable may be either absent or non-existent.
Classification Example

- Consider some classification techniques.

- Supervised learning examples.
  - Discriminant analysis.
  - Logistic regression.
  - CART.
  - SVMs.

- Unsupervised learning examples.
  - Association rules.
  - Cluster analysis.
  - Self-organizing maps.
Model-based clustering techniques can be traced at least as far back as Wolfe (1963).

In more recent years model-based clustering has appeared in the statistics literature with increased frequency.

Typically the data are clustered using some assumed mixture modeling structure.

Then the group memberships are ‘learned’ in an unsupervised fashion.
Assume

- The data are collected from a finite collection of populations.
- The data within each population can be modeled using a standard statistical model.

Gaussian mixture models have model density of the form

\[ f(x) = \sum_{g=1}^{G} \pi_g \phi(x | \mu_g, \Sigma_g). \]

- \( \pi_g \) is the probability that an observation belongs to group \( g \).
- \( \phi(x | \mu_g, \Sigma_g) \) is the density of a multivariate Gaussian \( (\mu_g, \Sigma_g) \).
MCLUST & Variable Selection

- MCLUST is probably the most well known model-based clustering technique in the literature.

- Variable selection is a technique that involves repeated application of MCLUST.

- Both are supported by R packages.
  - mclust
  - clustvarsel

The eigenvalue decomposition of the covariance matrix is of the form

$$\Sigma_g = \lambda_g D_g A_g D'_g,$$

where

- $\lambda_g$ is a constant,
- $D_g$ is a matrix consisting of the eigenvectors of $\Sigma_g$, and
- $A_g$ is a diagonal matrix with entries proportional to the eigenvalues of $\Sigma_g$. 

Paul McNicholas

Model-Based Clustering: An Overview
The Models

- This covariance structure allows for a variety of constraints.

<table>
<thead>
<tr>
<th>ID</th>
<th>Volume</th>
<th>Shape</th>
<th>Orient.</th>
<th>Covariance Decomp.</th>
<th>Number of Cov. Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>Equal</td>
<td>Spherical</td>
<td>—</td>
<td>$\lambda I$</td>
<td>1</td>
</tr>
<tr>
<td>VII</td>
<td>Variable</td>
<td>Spherical</td>
<td>—</td>
<td>$\lambda_k I$</td>
<td>$G$</td>
</tr>
<tr>
<td>EEI</td>
<td>Equal</td>
<td>Equal</td>
<td>Ax-Alg</td>
<td>$\lambda A$</td>
<td>$p$</td>
</tr>
<tr>
<td>VEI</td>
<td>Variable</td>
<td>Equal</td>
<td>Ax-Alg</td>
<td>$\lambda_g A_g$</td>
<td>$pG - G + 1$</td>
</tr>
<tr>
<td>EVI</td>
<td>Equal</td>
<td>Variable</td>
<td>Ax-Alg</td>
<td>$\lambda A_g$</td>
<td>$pG$</td>
</tr>
<tr>
<td>VVI</td>
<td>Variable</td>
<td>Variable</td>
<td>Ax-Alg</td>
<td>$\lambda D_k A_k D_k'$</td>
<td>$Gp(p + 1)/2$ - $(G - 1)p$</td>
</tr>
<tr>
<td>EEE</td>
<td>Equal</td>
<td>Equal</td>
<td>Equal</td>
<td>$\lambda D_k A_k D_k'$</td>
<td>$p(p + 1)/2$</td>
</tr>
<tr>
<td>EEV</td>
<td>Equal</td>
<td>Equal</td>
<td>Variable</td>
<td>$\lambda D_k A_k D_k'$</td>
<td>$Gp(p + 1)/2 - (G - 1)(p - 1)$</td>
</tr>
<tr>
<td>VEV</td>
<td>Variable</td>
<td>Equal</td>
<td>Variable</td>
<td>$\lambda_k D_k A_k D_k'$</td>
<td>$Gp(p + 1)/2 - (G - 1)(p - 1)$</td>
</tr>
<tr>
<td>VVV</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>$\lambda_k D_k A_k D_k'$</td>
<td>$Gp(p + 1)/2$</td>
</tr>
</tbody>
</table>

- The non-diagonal constraints have a number of covariance parameters that is quadratic in data-dimensionality $p$. 
Raftery & Dean (2006) propose a variable selection method based on the use of Bayes factors (Kass & Raftery, 1995).

This is essentially a model selection problem.

Two models, $M_1$ and $M_2$ say, for data $X$ are compared using the Bayes factors.
Bayes Factors

- The Bayes factor, $B_{12}$, for model $M_1$ versus model $M_2$, is defined as

$$B_{12} = \frac{p(X \mid M_1)}{p(X \mid M_2)},$$

where

$$p(X \mid M_k) = \int p(X \mid \theta_k, M_k) p(\theta_k \mid M_k) d\theta_k,$$

- $\theta_k$ is the vector of parameters for model $M_k$, and
- $p(\theta_k \mid M_k)$ is the prior distribution of $M_k$ (Kass & Raftery, 2005).

- Variables are then selected based on which model is the ‘best’.
Variable selection is often viewed as an improvement over MCLUST.

Variable selection does not always outperform MCLUST.

In addition to model-based clustering, variable selection is a **data reduction** technique.

Examples are given later...
Factor Analysis

- Introduced by Spearman (1904) following the introduction of Principal Components by Pearson (1901).

- Developed for and by psychologists.

- Laid out as a statistical model by Bartlett (1953).

- Spent much time as “the black sheep of statistical theory” (Lawley & Maxwell, 1962).
Factor Analysis — The Idea

- Consider a $p$-dimensional real-valued data vector $\mathbf{x}$.

- Assume $\mathbf{x}$ can be modeled using a $q$-dimensional vector of real-valued (unobservable) factors $\mathbf{u}$.

- $q \ll p$.

- Data reduction technique.
The Factor Analysis Model

- The model is
  \[ x = \mu + \Lambda u + \epsilon. \]
  - \( \Lambda \) is a \( p \times q \) matrix of factor loadings.
  - \( u \sim N(0, I_q) \) are the factors.
  - \( \epsilon \sim N(0, \Psi) \), where \( \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p) \).

- It follows that the marginal distribution of \( x \) is multivariate Gaussian \( (\mu, \Lambda \Lambda' + \Psi) \).

- \( \Lambda \) is not defined uniquely. If \( \Lambda \) is replaced by \( \Lambda^* = \Lambda D \) where \( D \) is orthonormal, then
  \[ \Lambda \Lambda' + \Psi = (\Lambda^*)(\Lambda^*)' + \Psi. \]
A special case of the factor analysis model, with $\Psi = \psi I_p$.

Therefore, the density of $\mathbf{x}$ is

$$f(\mathbf{x}) = \phi(\mathbf{x} | \mu, \Lambda \Lambda' + \psi I_p).$$

The maximum likelihood estimate (MLE) of $\mu$ is $\bar{x}$.

The MLEs for $\Lambda$ and $\Psi$ are found using the EM algorithm (Dempster et al. 1977).
EM Algorithm for PPCA: E-Step

- The E-step involves calculation of the expected complete-data log-likelihood, denoted $Q$.

- After some mathematics, it follows that $Q$, evaluated with $\mu = \hat{\mu} = \bar{x}$, is given by

$$Q(\Lambda, \Psi) = C + \frac{n}{2} \log |\Psi^{-1}| - \frac{n}{2} \text{tr} \{\Psi^{-1}S\} + n \text{tr} \{\Psi^{-1}\hat{\beta}\hat{S}\}$$

$$- \frac{n}{2} \text{tr} \{\Lambda'\Psi^{-1}\Lambda\Theta\},$$

where $\hat{\beta} = \hat{\Lambda}'(\hat{\Lambda}\hat{\Lambda}'+\hat{\Psi})^{-1}$ and $\Theta = \left(I_q - \hat{\beta}\hat{\Lambda} + \hat{\beta}\hat{S}\hat{\beta}'\right)$. 

Paul McNicholas
Model-Based Clustering: An Overview
EM Algorithm for PPCA: M-Step

- We need to maximize $Q$ with respect to $\Lambda$ and $\Psi$.


\[
\frac{\partial \log |X|}{\partial X} = X^{-1}
\]
\[
\frac{\partial \text{tr}(XA)}{\partial X} = A'
\]
\[
\frac{\partial \text{tr}(AXB)}{\partial X} = B'A'
\]
\[
\frac{\partial \text{tr}(XAXB)}{\partial X} = B'X'A' + A'X'B'
\]
Results of Matrix Differentiation

- Differentiating $Q$ with respect to $\Lambda$ we obtain

$$S_1(\Lambda, \Psi) = \frac{\partial Q}{\partial \Lambda} = n\Psi^{-1}S\hat{\beta}' - n\Psi^{-1}\Lambda\Theta.$$ 

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

$$\hat{\Lambda} = S\hat{\beta}'\Theta^{-1}.$$ 

- Differentiating $Q$ with respect to $\Psi^{-1}$ gives

$$S_2(\Lambda, \Psi) = \frac{\partial Q}{\partial \Psi^{-1}} = \frac{n}{2}\Psi - \frac{n}{2}S' + n\Lambda\hat{\beta}S - \frac{n}{2}\Lambda\Theta'\Lambda'.$$ 

- Solving the equation $S_2(\hat{\Lambda}, \hat{\Psi}) \equiv S_2(\hat{\Lambda}, \hat{\psi}) = 0$ we obtain

$$\hat{\psi} = \frac{1}{p} \text{tr}\{S - \hat{\Lambda}\hat{\beta}S\}.$$
Tipping & Bishop (1999b) develop a mixture of PPCAs model.

MPPCA is actually a special case of the mixture of factor analyzers model (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000).

The MFA model assumes a Gaussian mixture model, with a factor analysis covariance structure;

\[ f(x) = \sum_{g=1}^{G} \pi_g \phi(x \mid \mu_g, \Lambda_g \Lambda'_g + \Psi_g). \]
The parameters $\Lambda_g$ and $\Psi_g$ can be constrained across groups.

There is also the isotropic constraint, $\Psi_g = \psi_g I$.

These constraints lead to eight PGMMs:

<table>
<thead>
<tr>
<th>Model ID</th>
<th>Loading Matrix</th>
<th>Error Variance</th>
<th>Isotropic</th>
<th>Covariance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC</td>
<td>Constrained</td>
<td>Constrained</td>
<td>Const.</td>
<td>${pq - q(q - 1)/2} + 1$</td>
</tr>
<tr>
<td>CCU</td>
<td>Constrained</td>
<td>Constrained</td>
<td>Unconst.</td>
<td>${pq - q(q - 1)/2} + p$</td>
</tr>
<tr>
<td>CUC</td>
<td>Constrained</td>
<td>Unconstrained</td>
<td>Const.</td>
<td>${pq - q(q - 1)/2} + G$</td>
</tr>
<tr>
<td>CUU</td>
<td>Constrained</td>
<td>Unconstrained</td>
<td>Unconst.</td>
<td>${pq - q(q - 1)/2} + Gp$</td>
</tr>
<tr>
<td>UCC</td>
<td>Unconstrained</td>
<td>Constrained</td>
<td>Const.</td>
<td>$G{pq - q(q - 1)/2} + 1$</td>
</tr>
<tr>
<td>UCU</td>
<td>Unconstrained</td>
<td>Constrained</td>
<td>Unconst.</td>
<td>$G{pq - q(q - 1)/2} + p$</td>
</tr>
<tr>
<td>UUC</td>
<td>Unconstrained</td>
<td>Unconstrained</td>
<td>Const.</td>
<td>$G{pq - q(q - 1)/2} + G$</td>
</tr>
<tr>
<td>UUU</td>
<td>Unconstrained</td>
<td>Unconstrained</td>
<td>Unconst.</td>
<td>$G{pq - q(q - 1)/2} + Gp$</td>
</tr>
</tbody>
</table>
The Approach: AECM Algorithm

- ‘Alternating expectation-conditional maximization’ algorithm.

- The PGMMs are fitted using the AECM algorithm (Meng & van Dyk, 1997).

- The AECM algorithm (Meng & van Dyk, 1997) allows a different specification of complete-data for each conditional maximization step.

This missing data are the component membership labels $z_{ng}$.

These are replaced by their expected values

$$
\hat{z}_{ng} \propto \hat{\pi}_g \phi(\mathbf{x}_n | \hat{\mu}_g, \hat{\Lambda}_g \hat{\Lambda}'_g + \hat{\Psi}_g).
$$

This leads to the expected complete-data log-likelihood, $Q_1$.

Maximizing $Q_1$ with respect to $\mu_g$ and $\pi_g$ gives the estimates,

$$
\hat{\mu}_g = \frac{\sum_{n=1}^{N} \hat{z}_{ng} \mathbf{x}_n}{\sum_{n=1}^{N} \hat{z}_{ng}}
$$

and $\hat{\pi}_g = n_g/N$. 

AECM: Stage 2 ($\Lambda_g$ and $\Psi_g$)

- The missing data are the $z_{ng}$ and the latent variables $u_n$.

- Expected complete-data log-likelihood, $Q_2$, is computed.

- Constraints are imposed on $\Lambda_g$ and $\Psi_g$, or not.

- $Q_2$ is then differentiated with respect to $\Lambda_g$ and $\Psi_g^{-1}$; for example, in the UUU case

$$S_1(\Lambda_g, \Psi_g) = \frac{\partial Q(\Lambda_g, \Psi_g)}{\partial \Lambda_g} = \frac{n_g}{2} \left[ \Psi_g^{-1} S_g \hat{\beta}_g' - \Psi_g^{-1} \Lambda_g \Theta_g \right]$$

$$S_2(\Lambda_g, \Psi_g) = \frac{\partial Q(\Lambda_g, \Psi_g)}{\partial \Psi_g^{-1}} = \frac{n_g}{2} \left[ \Psi_g - S'_g + 2 \Lambda_g \hat{\beta}_g S_g - \Lambda_g \Theta_g' \Lambda_g' \right]$$
Further Generalization: 12 Models

Further Generalization of Covariance Structure

More recently, we use

\[ \Sigma_g = \Lambda_g \Lambda'_g + \Psi_g = \Lambda_g \Lambda'_g + \omega_g \Delta_g, \]

where

- \( \omega_g \in \mathbb{R} \),
- \( \Delta_g = \text{diag}\{\phi_1, \phi_2, \ldots, \phi_p\} \), such that \( |\Delta_g| = 1 \).

This leads to 12 models in total, all with a number of covariance parameters that is linear in \( p \).

<table>
<thead>
<tr>
<th>( \Lambda_g = \Lambda )</th>
<th>( \Delta_g = \Delta )</th>
<th>( \omega_g = \omega )</th>
<th>( \Delta = I )</th>
<th>Number of Covariance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
<td>U</td>
<td>U</td>
<td>([pq - q(q - 1)/2] + [G + (p - 1)])</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>U</td>
<td>U</td>
<td>([pq - q(q - 1)/2] + [G + (p - 1)])</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>C</td>
<td>U</td>
<td>([pq - q(q - 1)/2] + [1 + G(p - 1)])</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>C</td>
<td>U</td>
<td>([pq - q(q - 1)/2] + [1 + G(p - 1)])</td>
</tr>
</tbody>
</table>
Italian Wine Data

- Forina et al. (1986) reported twenty-eight chemical properties of Italian wines from the Piedmont region.

- Three specific types: Barolo, Grignolino, Barbera.

- 27 of these 28 properties are available from the UCI Machine Learning Database.
The PGMM family of models were fitted for $G = 1, 2, \ldots, 5$ and $q = 1, 2, \ldots, 5$.

The best model, in terms of both BIC (Schwartz, 1978) and ICL (Biernacki et al., 2000), is a CUU model with $G = 3, q = 4$. 

![Diagram showing the fitting of PGMMs]

Paul McNicholas  
Model-Based Clustering: An Overview
Classification for Best PGMM

- Classification table for the best PGMM.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grignolino</td>
<td>70</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Barbera</td>
<td></td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

- Rand Index = 0.99

- Adjusted Rand Index = 0.98
Results for MCLUST

- Using the `mclust` software, the best MCLUST model was a VVI model with three groups.

- **Classification for MCLUST.**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>58</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Grignolino</td>
<td>4</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>Barbera</td>
<td>48</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Rand Index\(=0.95\)

- Adjusted Rand Index\(=0.90\)
Results for Variable Selection

- Nineteen variables were selected using variable selection via the clustvarsel package (Dean & Raftery, 2006).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>52</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grignolino</td>
<td>17</td>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barbera</td>
<td>1</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Rand Index = 0.91

- Adjusted Rand Index = 0.78
Model Comparison

Comparison of models applied to Italian wine data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Rand Index</th>
<th>Adjusted Rand Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGMM</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>MCLUST</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>Variable Selection</td>
<td>0.91</td>
<td>0.78</td>
</tr>
</tbody>
</table>

The best PGMM model had greater BIC than the best mclust model.

MCLUST does better than Variable Selection.
Crabs Data

- Biological measurements on 200 crabs; 50 male and 50 female, for each of two species; 50 orange and 50 blue.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>Frontal lobe size in millimeters.</td>
</tr>
<tr>
<td>RW</td>
<td>Rear width in millimeters.</td>
</tr>
<tr>
<td>CL</td>
<td>Carapace length in millimeters.</td>
</tr>
<tr>
<td>CW</td>
<td>Carapace width in millimeters.</td>
</tr>
<tr>
<td>BD</td>
<td>Body depth in millimeters.</td>
</tr>
</tbody>
</table>

- The data was sourced from the MASS library in R.

- These data were also analyzed by Raftery & Dean (2006).
Best PGMM

- All twelve PGMMs were fitted for $G = 1, 2, \ldots, 5$ and $q = 1, 2, \ldots, 5$.

- The best model, in terms of both BIC (Schwartz, 1978) and ICL (Biernacki et al., 2000), is a CUUU model ($G = 4, q = 1$).
Comment on Best PGMM

- One latent variable (factor)...

Leptogrus Crabs Data

Paul McNicholas
Model-Based Clustering: An Overview
### Classification for Best PGMM

#### Classification table for the best PGMM.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>40</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td>46</td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Rand Index** = 0.935
- **Adjusted Rand Index** = 0.828
Raftery & Dean (2006) report the results of applying MCLUST and variable selection to the crabs data.

Classification for MCLUST.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blue</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td></td>
</tr>
<tr>
<td><strong>Orange</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>24</td>
<td>21</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rand Index = 0.851

Adjusted Rand Index = 0.533
Results for Variable Selection

Classification for variable selection.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>40</td>
<td>10</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td></td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rand Index = 0.931

Adjusted Rand Index = 0.815
Model Comparison

- Comparison of models applied to crabs data.

<table>
<thead>
<tr>
<th></th>
<th>Rand Index</th>
<th>Adj. Rand Index</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGMM</td>
<td>0.935</td>
<td>0.828</td>
<td>0.07</td>
</tr>
<tr>
<td>MCLUST</td>
<td>0.851</td>
<td>0.533</td>
<td>0.425</td>
</tr>
<tr>
<td>Var. Sel.</td>
<td>0.931</td>
<td>0.815</td>
<td>0.075</td>
</tr>
</tbody>
</table>

- Note that best PGMM model also has higher BIC / ICL than the best MCLUST model.

- Comparison with variable selection via BIC / ICL is not valid.
Consider Longitudinal Data

- How about clustering longitudinal data?
- What type of covariance structure?
- Cholesky decomposition?
- Modified Cholesky decomposition — even better!
Pourahmadi (1999, 2000) exploits the fact that covariance matrix $\Sigma$ of a random variable can be decomposed using the relation

$$T\Sigma T' = D,$$

where

- $T$ is a unique unit lower triangular matrix with diagonal elements $t_{ii} = 1$, and
- $D$ is a unique diagonal matrix with strictly positive entries.

An alternative version of this relationship is written

$$\Sigma^{-1} = T'D^{-1}T.$$
The Decomposition

- **T** and **D** can be interpreted statistically in terms of an autoregressive model.

- This decomposition was also used by Pan & MacKenzie (2003, 2006).

- Pourahmadi *et al.* (2007) extended this decomposition to account for multiple covariance matrices.
Consider the Gaussian mixture model with group covariance structure,

$$\Sigma_g^{-1} = T_g' D_g^{-1} T_g.$$ 

We can impose the following constraints to get a family of 8 models, 6 of which are new.

<table>
<thead>
<tr>
<th>Model</th>
<th>$T_g = T$</th>
<th>$D_g = D$</th>
<th>$D_g = \delta_g I$</th>
<th>Cov. Para's</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNN</td>
<td>Not Constrained</td>
<td>Not Constrained</td>
<td>Not Constrained</td>
<td>$G[p(p-1)/2] + Gp$</td>
</tr>
<tr>
<td>NCN</td>
<td>Not Constrained</td>
<td>Constrained</td>
<td>Not Constrained</td>
<td>$G[p(p-1)/2] + p$</td>
</tr>
<tr>
<td>CNN</td>
<td>Constrained</td>
<td>Not Constrained</td>
<td>Not Constrained</td>
<td>$p(p-1)/2 + Gp$</td>
</tr>
<tr>
<td>CCN</td>
<td>Constrained</td>
<td>Constrained</td>
<td>Not Constrained</td>
<td>$p(p-1)/2 + p$</td>
</tr>
<tr>
<td>NNC</td>
<td>Not Constrained</td>
<td>Not Constrained</td>
<td>Constrained</td>
<td>$G[p(p-1)/2] + G$</td>
</tr>
<tr>
<td>NCC</td>
<td>Not Constrained</td>
<td>Constrained</td>
<td>Constrained</td>
<td>$G[p(p-1)/2] + 1$</td>
</tr>
<tr>
<td>CNC</td>
<td>Constrained</td>
<td>Not Constrained</td>
<td>Constrained</td>
<td>$p(p-1)/2 + G$</td>
</tr>
<tr>
<td>CCC</td>
<td>Constrained</td>
<td>Constrained</td>
<td>Constrained</td>
<td>$p(p-1)/2 + 1$</td>
</tr>
</tbody>
</table>
These models can be fitted using an expectation-conditional maximization (ECM) algorithm (Meng & Rubin, 1993).

The ECM algorithm can be considered a more straightforward version of the AECM algorithm; without the $u$.

A paper based on these 8 models is in preparation.

This family of models has great potential for growth...

The constraints imposed by Pourahmadi et al. (2007) are currently being worked into this family of models.
The Data

- Data on the body weights of rats on one of three different dietary supplements.

- Published by Crowder & Hand (1991).

- 16 rats were put on one of three different diets;
  - 8 rats were on Diet 1,
  - 4 were put on Diet 2, and
  - 4 on Diet 3.
Groups

- The three groups can be seen on the following graph;

- Group 2 has a heavy rat and Group 3 has a light rat.
The clustering for the model with the highest BIC is:

- The Rand index is 0.95 (0.88 adjusted Rand).
Conclusions I

- Data reduction techniques can improve clustering and classification results.

- A family of 12 parsimonious Gaussian mixture models has been introduced, which includes the MFA and MPPCA models as special cases.

- This family of models has been shown to perform favorably when compared to well-established techniques.

- Especially useful for high-dimensional problems; many such problems arise in bioinformatics.
Conclusions II

- Clustering of longitudinal data can also be achieved using Gaussian mixture models.

- A family of 8 mixture models has been introduced, with a modified Cholesky decomposed covariance structure.

- This family of models has been shown to give good results on real data.

- This family has great potential for further expansion.
This work was done in collaboration with Prof. Brendan Murphy, University College Dublin, Ireland.

Part of this work was supported by a Science Foundation of Ireland Basic Research Grant.


