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.

Statistical Learning

- 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

- 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.

Finite Mixture Models

- 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

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

  - π_g is the probability that an observation belongs to group g.
  - $\phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian $\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$.

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 Covariance Structure

- Banfield & Rafterey (1993), Celeux & Govaert (1995) and Fraley & Raftery (1998, 2002) exploit an eigenvalue decomposition of the group covariance matrices for the Gaussian mixture model.

- The eigenvalue decomposition of the covariance matrix is of the form

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}'_g,$$

  where

  - λ_g is a constant,
  - $\mathbf{D}_g$ is a matrix consisting of the eigenvectors of $\boldsymbol{\Sigma}_g$, and
  - $\mathbf{A}_g$ is a diagonal matrix with entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$.

The Models

- This covariance structure allows for a variety of constraints.

ID	Volume	Shape	Orient.	Covariance Decomp.	Number of Cov. Parameters
EII	Equal	Spherical	—	$\lambda\mathbf{I}$	1
VII	Variable	Spherical	—	$\lambda_k\mathbf{I}$	G
EEI	Equal	Equal	Ax-Alg	$\lambda\mathbf{A}$	p
VEI	Variable	Equal	Ax-Alg	$\lambda_g\mathbf{A}$	$p+G-1$
EVI	Equal	Variable	Ax-Alg	$\lambda\mathbf{A}_g$	$pG-G+1$
VVI	Variable	Variable	Ax-Alg	$\lambda_g\mathbf{A}_g$	pG
EEE	Equal	Equal	Equal	$\lambda\mathbf{DAD}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{D}_k\mathbf{AD}'_k$	$Gp(p+1)/2-(G-1)p$
VEV	Variable	Equal	Variable	$\lambda_k\mathbf{D}_k\mathbf{AD}'_k$	$Gp(p+1)/2-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k$	$Gp(p+1)/2$

- The non-diagonal constraints have a number of covariance parameters that is quadratic in data-dimensionality p.

The Idea

- 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 using 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

$$p(X \mid M_k) = \int p(X \mid \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k \mid M_k) d\boldsymbol{\theta}_k,$$

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

- Variables are then selected based on which model is the 'best'.

Comments

- 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
$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{u} + \boldsymbol{\epsilon}.$$
    - $\boldsymbol{\Lambda}$ is a p x q matrix of factor loadings.
    - $\mathbf{u} \sim N(0, \mathbf{I}_q)$ are the factors.
    - $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \psi_2, \ldots, \psi_p)$.
- It follows that the marginal distribution of $\mathbf{x}$ is multivariate Gaussian $(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$.
- $\boldsymbol{\Lambda}$ is not defined uniquely. If $\boldsymbol{\Lambda}$ is replaced by $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{D}$ where $\mathbf{D}$ is orthonormal, then
$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = (\boldsymbol{\Lambda}^*)(\boldsymbol{\Lambda}^*)' + \boldsymbol{\Psi}.$$

The PPCA Model (Tipping & Bishop, 1999*a*)

- A special case of the factor analysis model, with $\mathbf{\Psi} = \psi \mathbf{I}_p$.
- Therefore, the density of $\mathbf{x}$ is

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

- The maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ is $\bar{\mathbf{x}}$.
- The MLEs for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ 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 $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, is given by

$$Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = C + \frac{n}{2} \log |\boldsymbol{\Psi}^{-1}| - \frac{n}{2} \operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\mathbf{S}\right\} + n \operatorname{tr}\left\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\hat{\beta}\mathbf{S}\right\} - \frac{n}{2} \operatorname{tr}\left\{\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}\right\},$$

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

EM Algorithm for PPCA: M-Step

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

- Graybill (1983), Lütkepohl (1996) and Magnus & Neudecker (1999) give helpful results.

$$\frac{\partial \log|\mathbf{X}|}{\partial \mathbf{X}} = \mathbf{X}^{-1}$$
$$\frac{\partial \operatorname{tr}(\mathbf{XA})}{\partial \mathbf{X}} = \mathbf{A}'$$
$$\frac{\partial \operatorname{tr}(\mathbf{AXB})}{\partial \mathbf{X}} = \mathbf{B}'\mathbf{A}'$$
$$\frac{\partial \operatorname{tr}(\mathbf{XAXB})}{\partial \mathbf{X}} = \mathbf{B}'\mathbf{X}'\mathbf{A}' + \mathbf{A}'\mathbf{X}'\mathbf{B}'$$

Results of Matrix Differentiation

- Differentiating Q with respect to $\mathbf{\Lambda}$ we obtain

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

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

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

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

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

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

$$\hat{\psi}=\frac{1}{p}\,\mathrm{tr}\{\mathbf{S}-\hat{\mathbf{\Lambda}}\hat{\beta}\mathbf{S}\}.$$

MFA Model

- Tipping & Bishop (1999*b*) 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;

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g).$$

Eight PGMMs

- The parameters $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ can be constrained across groups.
- There is also the isotropic constraint, $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}$.
- These constraints leads to eight PGMMs:

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

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.
- McLachlan & Peel (2000) give extensive details of the fitting algorithm in the UUU case.

AECM: Stage 1 (π_g and $\boldsymbol{\mu}_g$)

- This missing data are the component membership labels z_{ng}.
- These are replaced by their expected values

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

- This leads to the expected complete-data log-likelihood, Q_1.
- Maximizing Q_1 with respect to $\boldsymbol{\mu}_g$ and π_g gives the estimates,

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

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

AECM: Stage 2 ($\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$)

- The missing data are the z_{ng} and the latent variables $\mathbf{u}_n$.
- Expected complete-data log-likelihood, Q_2, is computed.
- Constraints are imposed on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$, or not.
- Q_2 is then differentiated with respect to $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g^{-1}$; for example, in the UUU case

$$S_1(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial Q(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)}{\partial \boldsymbol{\Lambda}_g} = \frac{n_g}{2}\left[\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g\hat{\beta}_g' - \boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g\right]$$

$$S_2(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial Q(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)}{\partial \boldsymbol{\Psi}_g^{-1}} = \frac{n_g}{2}\left[\boldsymbol{\Psi}_g - \mathbf{S}_g' + 2\boldsymbol{\Lambda}_g\hat{\beta}_g\mathbf{S}_g - \boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g'\boldsymbol{\Lambda}_g'\right]$$

Further Generalization of Covariance Structure

- More recently, we use

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \omega_g \mathbf{\Delta}_g,$$

  where

  - $\omega_g \in \mathbb{R}$,
  - $\mathbf{\Delta}_g = \text{diag}\{\phi_1, \phi_2, \ldots, \phi_p\}$, such that $|\mathbf{\Delta}_g| = 1$.

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

$\mathbf{\Lambda}_g = \mathbf{\Lambda}$	$\mathbf{\Delta}_g = \mathbf{\Delta}$	$\omega_g = \omega$	$\mathbf{\Delta} = I$	Number of Covariance Parameters
C	C	U	U	$[pq - q(q-1)/2] + [G + (p-1)]$
U	C	U	U	$G[pq - q(q-1)/2] + [G + (p-1)]$
C	U	C	U	$[pq - q(q-1)/2] + [1 + G(p-1)]$
U	U	C	U	$G[pq - q(q-1)/2] + [1 + G(p-1)]$

Italian Wine Data

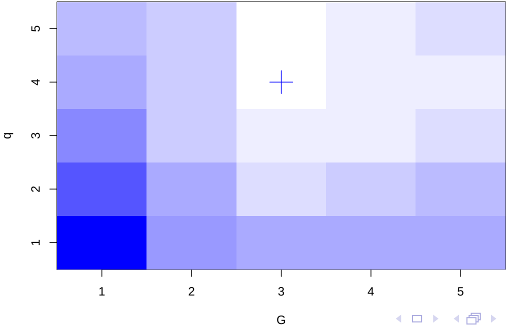
- Forina et al. (1986) reported twenty-eight chemical properties of Italian wines from the Piedmont region.

- Three specific types: Barolo, Grignolino, Barbera.

- 27 of these 28 properties are available from the UCI Machine Learning Database.

Best PGMM

- 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 $G = 3$, $q = 4$.



Classification for Best PGMM

- Classification table for the best PGMM.

	1	2	3
Barolo	59		
Grignolino		70	1
Barbera			48

- 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.

	1	2	3
Barolo	58	1	
Grignolino	4	66	1
Barbera			48

- 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).

	1	2	3	4
Barolo	52	7		
Grignolino		17	54	
Barbera		1		47

- Rand Index=0.91

- Adjusted Rand Index=0.78

Model Comparison

- Comparison of models applied to Italian wine data.

Model	**Rand Index**	**Adjusted Rand Index**
PGMM	0.99	0.98
MCLUST	0.95	0.90
Variable Selection	0.91	0.78

- 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.

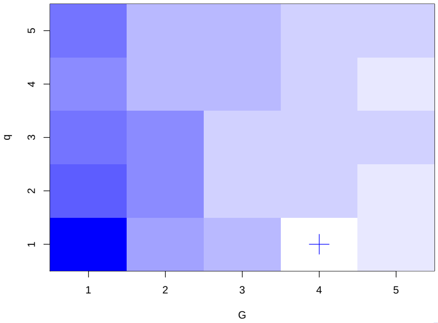
Variable	Measurement
FL	Frontal lobe size in millimeters.
RW	Rear width in millimeters.
CL	Carapace length in millimeters.
CW	Carapace width in millimeters.
BD	body depth in millimeters.

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

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

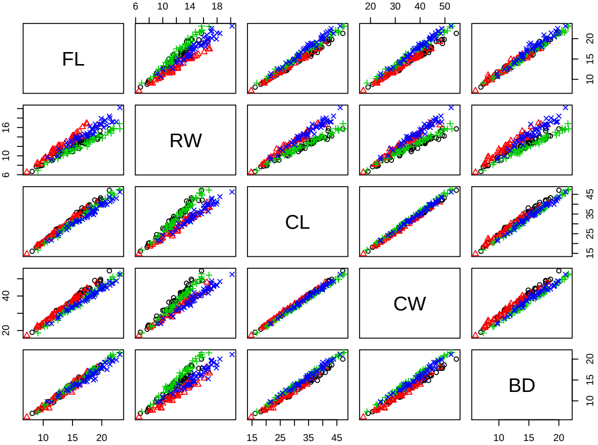
Best PGMM

- 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)...



Classification for Best PGMM

- Classification table for the best PGMM.

		1	2	3	4
Blue	Male	40	10		
	Female		50		
Orange	Male			50	
	Female			4	46

- Rand Index=0.935
- Adjusted Rand Index=0.828

Results for MCLUST

- Raftery & Dean (2006) report the results of applying MCLUST and variable selection to the crabs data.

- Classification for MCLUST.

		1	2	3	4	5	6	7
Blue	Male	32					18	
	Female		31				19	
Orange	Male			28				22
	Female				24	21		5

- Rand Index=0.851

- Adjusted Rand Index=0.533

Results for Variable Selection

- Classification for variable selection.

		1	2	3	4
Blue	Male	40	10		
	Female		50		
Orange	Male			50	
	Female			5	45

- Rand Index=0.931
- Adjusted Rand Index=0.815

Model Comparison

- Comparison of models applied to crabs data.

	Rand Index	Adj. Rand Index	Error Rate
PGMM	0.935	0.828	0.07
MCLUST	0.851	0.533	0.425
Var. Sel.	0.931	0.815	0.075

- 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!.

The Decomposition

- Pourahmadi (1999, 2000) exploits the fact that covariance matrix $\boldsymbol{\Sigma}$ of a random variable can be decomposed using the relation

$$\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' = \mathbf{D},$$

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

- An alternative version of this relationship is written

$$\boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}.$$

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.

Constraints

- Consider the Gaussian mixture model with group covariance structure,

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

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

Model	$\mathbf{T}_g = \mathbf{T}$	$\mathbf{D}_g = \mathbf{D}$	$\mathbf{D}_g = \delta_g\mathbf{I}$	Cov. Para's
NNN	Not Constrained	Not Constrained	Not Constrained	$G[p(p-1)/2] + Gp$
NCN	Not Constrained	Constrained	Not Constrained	$G[p(p-1)/2] + p$
CNN	Constrained	Not Constrained	Not Constrained	$p(p-1)/2 + Gp$
CCN	Constrained	Constrained	Not Constrained	$p(p-1)/2 + p$
NNC	Not Constrained	Not Constrained	Constrained	$G[p(p-1)/2] + G$
NCC	Not Constrained	Constrained	Constrained	$G[p(p-1)/2] + 1$
CNC	Constrained	Not Constrained	Constrained	$p(p-1)/2 + G$
CCC	Constrained	Constrained	Constrained	$p(p-1)/2 + 1$

Model Fitting & Development

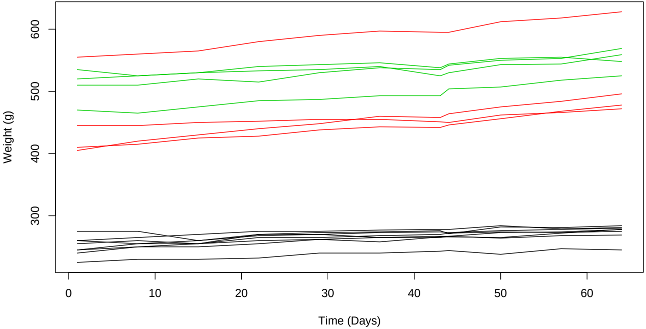
- These models can be fitted using an expectation-conditional maximization (ECM) algorithm (Meng & Rubin, 1993).
- The ECM algorithm can be considered a more straightforward version of the AECM algorithm; without the $\mathbf{u}$.
- A paper based on these 8 models is in preparation.
- This family of 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

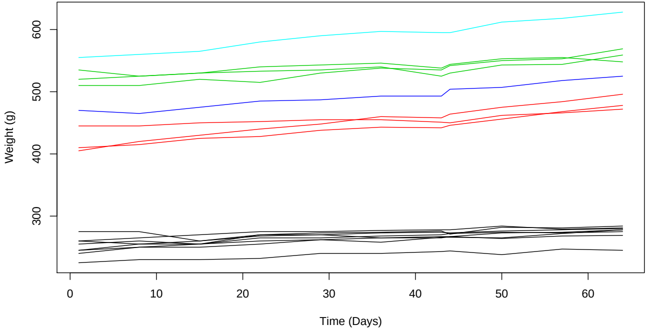
- The three groups can be see on the following graph;



- Group 2 has a heavy rat and Group 3 has a light rat.

Results

- 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.

Acknowledgements

- 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.

Selected Bibliography

- Bartholomew, D. & Knott, M. (1999), *Latent Variable Models and Factor Analysis*, Kendall's Library of Statistics, second edn, Arnold, London.
- Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *JASA* **97**(458), 611–612.
- Ghahramani, Z. & Hinton, G. E. (1997), The EM algorithm for factor analyzers, Technical Report CRG-TR-96-1, University Of Toronto, Toronto.
- Lütkepohl, H. (1996), *Handbook of Matrices*, John Wiley & Sons, Chicester.
- Magnus, J. R. & Neudecker, H. (1999), *Matrix differential calculus with applications in statistics and econometrics*, Revised edn, John Wiley & Sons, Chicester.
- McLachlan, G. J. & Peel, D. (2000*b*), Mixtures of factor analyzers, *in* 'Seventh International Conference on Machine Learning', San Francisco.
- Meng, X. L. & van Dyk (1997), 'The EM algorithm - an old folk song sung to the fast tune (with discussion)', *JRSS Series B* **59**, 511–567.
- Pourahmadi, M., Daniels, M. & Park, T. (2007), 'Simultaneous modelling of the Cholesky decomposition of several covariance matrices', *Journal of Multivariate Analysis* **98**, 568–587.
- Tipping, T., E. & Bishop, C. M. (1999*a*), 'Mixtures of probabilistic principle component analysers', *Neural Computation* **11**(2), 443–482.
- Tipping, T., E. & Bishop, C. M. (1999*b*), 'Probabilistic principle component analysers', *JRSS* **61**, 611–622.