Model-based clustering using mixtures of t-factor analyzers: A food authenticity example

Jeffrey L. Andrews

Ph.D. Candidate
Department of Mathematics & Statistics
University of Guelph
Guelph, Ontario, Canada

July 26, 2010

Welcome

- This presentation will focus on model-based clustering using a 6-member family of mixtures of multivariate t-distribution models as introduced by Andrews and McNicholas (2010).
- Parameter estimation, model selection, and model performance will be discussed.
- The 6-member MMtFA family will be illustrated via an application to two food authenticity data sets.

Italian Wines

The `wine` dataset from the `gclus` library in R:

- 13 chemical properties;
- 178 samples of wine;
- 3 varieties of wine: Barolo, Barbera, and Grignolino.

Can we objectively cluster types of wine according to their chemical properties?

Table: Thirteen of the chemical and physical properties of the Italian wines.

Alcohol	Proline	OD_{280}/OD_{315} of diluted wines
Malic acid	Ash	Alcalinity of ash
Hue	Total phenols	Magnesium
Flavonoids	Nonflavonoid phenols	Proanthocyanins
Color intensity		

Mixtures of Multivariate t-Distributions

The model density is of the form

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g),$$

where

$$f_t(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma\left(\frac{\nu}{2}\right)\left\{1 + \frac{\delta(\mathbf{x},\boldsymbol{\mu}|\boldsymbol{\Sigma})}{\nu}\right\}^{\frac{1}{2}(\nu+p)}}$$

is the multivariate t-distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν. π_g are the mixing proportions.

Mixtures of Multivariate t-Factor Analyzers

The model density is of the form

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

MMtFAs adjust the covariance structure of the density such that

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g$$

This is the factor analysis covariance structure.

Extensions

- McLachlan et al. (2007) develop the unconstrained case:

$$\mathbf{\Sigma_g} = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g.$$

- Zhao and Jiang (2006) develop a version of the PPCA constraint:

$$\mathbf{\Sigma_g} = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \psi_g \mathbf{I}_p.$$

- We will consider:
  - constraining the degrees of freedom parameter, or $\nu_g = \nu$;
  - the PPCA constraint, or $\mathbf{\Psi}_g = \psi_g \mathbf{I}$;
  - the loading matrix constraint, or $\mathbf{\Lambda}_g = \mathbf{\Lambda}$.

EM Algorithms

- The expectation-maximization (EM) algorithm is an iterative procedure used to find maximum likelihood estimates in the presence of missing or incomplete data.
- The expectation-conditional maximization (ECM) algorithm replaces the maximization (M) step with a series of computationally simpler conditional maximization (CM) steps.
- The alternating expectation-conditional maximization (AECM) algorithm permits the complete data vector to vary, or alternate, on each CM-step.
- Parameters are estimated using the AECM algorithm in the t-factors case because there are three types of missing data.

BIC and ICL

- Model selection is performed using the Bayesian information criterion (BIC) and the integrated completed likelihood (ICL):

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

$$\text{ICL} = \text{BIC} + \sum_{i=1}^{n} \sum_{g=1}^{G} \text{MAP}(\hat{z}_{ig}) \ln(\hat{z}_{ig}).$$

- Note that

$$\text{MAP}(\hat{z}_{ig}) = \begin{cases} 1 & \text{if } \max_g \{z_{ig}\} \text{ occurs at group } g, \\ 0 & \text{otherwise.} \end{cases}$$

Adjusted Rand Index

- Clustering performance will be evaluated using the adjusted Rand index.
- The Rand index is calculated by

$$\frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}},$$

where 'number of agreements/disagreements' are based on pairwise comparisons.

- The adjusted Rand index corrects for chance, recognizing that clustering performed randomly would correctly classify some pairs.

MM*t*FA Family Development

- Three constraints will now be introduced that lead to a family of six mixture models.

Constraining $\nu_g = \nu$

- Constraining the degrees of freedom to be equal across groups ($\nu_g = \nu$) effectively assumes that each group can be modelled using the same distributional shape.
- The savings in parameter estimation are quite small ($G - 1$), however in practice constraining the degrees of freedom can lead to better clustering performance (Andrews and McNicholas, 2010).
- This is likely due to a more stable estimation of the degrees of freedom parameter under n samples rather than n_g.

Constraining $\mathbf{\Psi}_g = \psi_g \mathbf{I}$

- Utilizing the isotropic constraint ($\mathbf{\Psi}_g = \psi_g \mathbf{I}$) assumes that each group contains a unique, scalar error in the variance estimation under the factor analysis structure.
- As $\mathbf{\Psi}_g$ is a diagonal matrix, Gp parameters are normally needed for estimation.
- Under this constraint, only G parameters are estimated: a significant reduction, especially under high-dimensional data sets.

Constraining $\mathbf{\Lambda}_g = \mathbf{\Lambda}$

- Constraining the loading matrices to be equal across groups ($\mathbf{\Lambda}_g = \mathbf{\Lambda}$) assumes that each group's covariance estimates are identical.
- As $\mathbf{\Lambda}_g$ is a $p \times q$ matrix, $G[pq - q(q-1)/2]$ parameters are normally needed for estimation.
- Under this constraint, only $pq - q(q-1)/2$ parameters are estimated: a large reduction in free parameters.

The Six Models

- Covariance structures derived from the mixtures of t-factor analyzers model (C=Constrained, U=Unconstrained):

Model	$\mathbf{\Lambda}$	$\psi_g\mathbf{I}$	ν	Covariance and DF Parameters
CCC	C	C	C	$[pq - q(q-1)/2] + G + 1$
CCU	C	C	U	$[pq - q(q-1)/2] + G + G$
UCC	U	C	C	$G[pq - q(q-1)/2] + G + 1$
UCU	U	C	U	$G[pq - q(q-1)/2] + G + G$
UUC	U	U	C	$G[pq - q(q-1)/2] + Gp + 1$
UUU	U	U	U	$G[pq - q(q-1)/2] + Gp + G$

Overview

- The MM*t*FA family will be compared to established model-based clustering techniques:
  - Parsimonious Gaussian mixture models (PGMMs, McNicholas and Murphy, 2008);
  - MCLUST (Fraley and Raftery, 1999);
  - and Variable selection (Dean and Raftery, 2006).
- A brief summary of these methods follows...

PGMMs

- McNicholas and Murphy (2008) introduce PGMMs, a family based on mixtures of factor analyzers
- The model density is

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

  where $\phi(\cdot)$ is the multivariate Gaussian density.
- Constraining...
  - $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$,
  - $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$,
  - and/or $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$ leads to a family of 8 mixture models

MCLUST

- Fraley and Raftery (1999) introduce MCLUST, a family based on the eigendecomposition of the multivariate Gaussian covariance structure
- The model density is

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g),$$

- Constraining...
  - $\lambda_g = \lambda$,
  - $\lambda = 1$,
  - $\mathbf{D}_g = \mathbf{D}$,
  - $\mathbf{A}_g = \mathbf{A}$,
  - or replacing $\mathbf{A}$ and/or $\mathbf{D}$ with the identity matrix leads to a family of 10 mixture models.

Variable Selection

- Dean and Raftery (2006) introduce `clustvarsel`, a variable selection package for the R computing environment.
- `clustvarsel` runs multiple MCLUST models on different subsets of variables.
- The best subset of variables are determined using Bayes factors.

The Data

Recall...

The `wine` dataset from the `gclus` library in R:

- 13 chemical properties
- 178 samples of wine
- 3 varieties of wine: Barolo, Barbera, and Grignolino

Can we objectively cluster types of wine according to their chemical properties?

Table: Thirteen of the chemical and physical properties of the Italian wines.

Alcohol	Proline	OD_{280}/OD_{315} of diluted wines
Malic acid	Ash	Alcalinity of ash
Hue	Total phenols	Magnesium
Flavonoids	Nonflavonoid phenols	Proanthocyanins
Color intensity		

The Method

- We run t-factors for
  - G components from 1–5.
  - q factors from 1–6.
- Choose model according to the largest BIC/ICL.
- Compare clustering results with PGMMs, `mclust`, and `clustvarsel`.

Clustering Results

Classification table for the fully unconstrained model on the `wine` dataset:

	1	2	3
Barolo	58	1	0
Grignolino	1	70	0
Barbera	0	0	48

Comparison

Adjusted Rand indices for different model-based clustering techniques:

Model	Adjusted Rand Index
UUC	0.98
UUU	0.96
CCU	0.95
UCU	0.93
UCC	0.90
CCC	0.84
PGMMs	0.79
`clustvarsel`	0.78
MCLUST	0.48

Coffee Data

- Streuli (1973) reported thirteen chemical properties of coffee from across 28 countries and of two types; Robusta and Arabica.
- The following chemical properties were recorded.

Chemical Properties		
Water	Bean Weight	Extract Yield
pH Value	Free Acid	Mineral Content
Fat	Caffeine	Trigonellin
Chlorogenic Acid	Neochlorogenic Acid	Isochlorogenic Acid
Total Chlorogenic Acid		

- The data was sourced from `www.parvus.unige.it`.

The Method

- We run t-factors for
  - G components from 1–4.
  - q factors from 1–4.
- Choose model according to the largest BIC/ICL.
- Compare clustering results with PGMMs, `mclust`, and `clustvarsel`.

Comparison

Adjusted Rand indices for different model-based clustering techniques:

Model	Adjusted Rand Index
UUC	1.00
UUU	1.00
UCU	1.00
UCC	1.00
PGMMs	1.00
MCLUST	1.00
CCU	0.38
CCC	0.38
`clustvarsel`	0.23

Overview

- Mixtures of t-factors give better clustering results than all other considered methods on the `wine` dataset.
- In fact, the entire MMtFA family choose the right number of groups, and each has an adjusted Rand of 0.84 or higher.
- The MMtFA model chosen, as well as the majority of the family, perform as well as PGMMs and MCLUST on the coffee data.

- A full family of 16 mixtures of multivariate t-factor analyzers is forthcoming.

- Mixture models can also be used under a classification framework (McNicholas, 2010, Andrews et al., 2010); incorporation of this framework for these models is also forthcoming.

Thank you

This collaboration with Paul D. McNicholas is supported by:

- Compusense
- The Natural Sciences and Engineering Research Council of Canada (NSERC)
  - Discovery Grant
  - Postgraduate Scholarship (PGS-D)
- The Canada Foundation for Innovation (CFI)
  - Leaders Opportunity Fund
- The Ontario Research Fund
  - Research Infrastructure Program

Selected Bibliography

Andrews, J. L. and McNicholas, P. D. (2010), 'Extending mixtures of multivariate t-factor analyzers', *Statistics and Computing* . To appear, DOI: 10.1007/s11222-010-9175-2.

Andrews, J. L., McNicholas, P. D. and Subedi, S. (2010), 'Model-based classification via mixtures of multivariate t-distributions', *Computational Statistics and Data Analysis* . To appear, DOI: 10.1016/j.csda.2010.05.019.

Dean, N. and Raftery, A. E. (2006), *The clustvarsel Package*. R package version 0.2-4.

Fraley, C. and Raftery, A. E. (1999), 'MCLUST: Software for model-based cluster analysis', *Journal of Classification* **16**, 297–306.

McLachlan, G. J., Bean, R. W. and Jones, L. B.-T. (2007), 'Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution', *Computational Statistics and Data Analysis* **51**(11), 5327–5338.

McNicholas, P. D. (2010), 'Model-based classification using latent Gaussian mixture models', *Journal of Statistical Planning and Inference* **140**(5), 1175–1181.

McNicholas, P. D. and Murphy, T. B. (2008), 'Parsimonious Gaussian mixture models', *Statistics and Computing* **18**, 285–296.