Non-Gaussian Mixture Model Averaging for Clustering

NON-GAUSSIAN MIXTURE MODEL AVERAGING FOR CLUSTERING

BY

XU XUAN (JOY) ZHANG, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF McMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2016)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Non-Gaussian Mixture Model Averaging for Clustering
AUTHOR:	Xu Xuan (Joy) Zhang B.Sc., (Mathematics)
SUPERVISOR:	Dr. Paul McNicholas
NUMBER OF PAGES:	viii, 36

Abstract

The Gaussian mixture model has been used for model-based clustering analysis for decades. Most model-based clustering analyses are based on the Gaussian mixture model. Model averaging approaches for Gaussian mixture models are proposed by Wei and McNicholas (2015), based on a family of 14 Gaussian parsimonious clustering models (Celeux and Govaert, 1995). In this thesis, we use non-Gaussian mixture models, namely the tEigen family, for our averaging approaches. This paper studies fitting an averaged model from a set of multivariate t-mixture models instead of fitting a best model.

Acknowledgements

I would first like to thank my supervisor Paul D. McNicholas for his on going guidance with my research. It is a wonderful experience under his supervision. I would also like to thank Yuhong Wei for providing me the information on model averaging. I want to thank my family for their consistent support and encouragement. I also want to thank my friends Bonnie Han, Jing Cai, Angela Wang, and Yiliang Zhou for their kind assistance.

Contents

List of Tables

Chapter 1

Introduction

In this thesis, we discuss model-based clustering analysis using the tEigen family (Andrews and McNicholas, 2012). The main purpose is to analyze the classification performance after averaging models, which are selected within the Occam's window, and to compare with the Gaussian analogue. In general, we fit a data set $\mathbf{x}$ using the mixture model density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\pi_g > 0$ is the gth mixing proportion, such that $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the multivariate density for the gth component, with parameters $\boldsymbol{\vartheta} = (\pi_1, ..., \pi_G, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G)$.

In Chapter 2, we introduce model-based clustering and the tEigen family of 28 models. The tEigen family arises from the eigen-decomposition of the multivariate t-distribution scale structure $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, which can be used to generate the 28 mixture models. The parameters of models within tEigen family are estimated through the ECM algorithm. The tEigen family has been implemented in the `R` package `teigen` (Andrews *et al.*, 2016), and details are given in Chapter 2.

There are some methods to select the mixture models to averaging, such as the Bayesian information criterion (BIC; Schwarz 1978) and Akaike information criterion (AIC; Akaike, 1974). Despite its limitations (cf. Bhattacharya and McNicholas, 2014), the BIC is most commonly used criterion and we use it to select models from the tEigen family to put in Occam's Window. When necessary, we merge components of models based on their weights using a merging criterion (Chapter 3). Two averaging approaches are explored: averaging *a posteriori* probabilities and model averaging.

Illustrations for our averaging results for the tEigen models are given in Chapter 4, and a discussion about the averaging approaches is included in Chapter 5.

Chapter 2

Background

2.1 Model-based clustering

Under Gaussian model-based clustering, a p-dimensional random variable $\mathbf{X}$ has G-component mixture density, i.e.,

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.1}$$

where $\pi_g > 0$ is the mixing proportion of the gth component in a mixture model, such that $\sum \pi_g = 1$, with parameters $\boldsymbol{\vartheta} = (\pi_1, ..., \pi_G, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G)$, and $\boldsymbol{\mu}_g$ is the mean and $\boldsymbol{\Sigma}_g$ is the covariance matrix for component g. The component densities are usually the same for all components in a clustering analysis, i.e., $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g) = f(\mathbf{x} \mid \boldsymbol{\theta})$. From (2.1), the likelihood is :

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g). \tag{2.2}$$

Banfield and Raftery (1993) have discussed a parametrization in terms of eigenvalue decomposition for component variance matrix, i.e.:

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g' \tag{2.3}$$

where $\lambda_g = |\boldsymbol{\Sigma}_g|^{1/p}$, $\mathbf{D}_g$ is the orthogonal matrix containing the eigenvectors of $\boldsymbol{\Sigma}_g$, and the determinant of a diagonal matrix $\mathbf{A}_g$ is 1.

Based on Banfield and Raftery (1993), Celeux and Govaert (1995) have built 14 Gaussian parsimonious clustering models (GPCMs; Table 2) in three categories of spherical (EII, VII), diagonal (EEI, VEI, EVI, VVI), and general (EEE, VEE, EVE, EEV,VVE, VEV, EVV, VVV). The constraints on $\boldsymbol{\Sigma}_g$ that are considered are: $\boldsymbol{\Sigma}_g = \mathbf{I}_p, \boldsymbol{\Sigma}_g = \sigma_g \mathbf{I}_p, \boldsymbol{\Sigma}_g = \sigma \mathbf{I}_p$, and $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Details are discussed in Gordon (1981), Banfield and Raftery (1993), and McNicholas (2016, Chapter 2).

The parameters for 12 of the 14 models in the GPCM family were estimated by using the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977), and details are discussed in Celeux and Govaert (1995) and McLachlan and Krishnan (2008). The parameters for the other two models, i.e., EVE and VVE, are estimated by Browne and McNicholas (2014a) using the MM algorithm (Hunter and Lange, 2004). The details of the developed alternative algorithms are given by Browne and McNicholas (2014a), and implemented in the R package `mixture` (Browne and McNicholas, 2014b), which includes all 14 mixture models of the GPCM family for model-based clustering analysis. The idea of averaging mixture models based on GPCMs family is introduced by Wei and McNicholas (2015) as an alternative to selecting the single best model.

Table 2.1: Nomenclature, covariance structure, and number of covariance parameters for each member of the GPCM family, where G denotes the number of components, and p denotes the dimension of the data.

Model	Volume	Shape	Orientation	$\mathbf{\Sigma}_g$	Free covariance parameters
EII	Equal	Equal	NA	$\lambda\mathbf{I}$	1
VII	Variable	Equal	NA	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Coord. Axes	$\lambda\mathbf{A}$	p
VEI	Variable	Equal	Coord. Axes	$\lambda_g\mathbf{I}$	p+G-1
EVI	Equal	Variable	Coord. Axes	$\lambda\mathbf{A}_g$	Gp-G+1
VVI	Variable	Variable	Coord. Axes	$\lambda_g\mathbf{A}_g$	pG
EEE	Equal	Equal	Equal	$\lambda\mathbf{DAD'}$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{D_gAD'_g}$	$Gp(p+1)/2$-(G-1)p
VEV	Variable	Equal	Variable	$\lambda_g\mathbf{D_gAD'_g}$	$Gp(p+1)/2$-(G-1)(p+1)
VVV	Variable	Variable	Variable	$\lambda_g\mathbf{D_gA_gD'_g}$	$Gp(p+1)/2$
EVE	Equal	Variable	Equal	$\lambda\mathbf{DA_gD'}$	$p(p+1)/2$+(G-1)(p-1)
VVE	Variable	Variable	Equal	$\lambda_g\mathbf{DA_gD'}$	$p(p+1)/2$+(G-1)p
VEE	Variable	Equal	Equal	$\lambda_g\mathbf{DAD'}$	$p(p+1)/2$+(G-1)
EVV	Equal	Variable	Variable	$\lambda\mathbf{D_gA_gD'_g}$	$Gp(p+1)$-(G-1)

2.2 The tEIGEN family of models

The departure from the Gaussian mixture model is the mixture of multivariate t-distributions. McLachlan and Peel (1998) and Peel and McLachlan (2000) motivated a heavy-tailed multivariate t-distribution. The gth component density for a p-dimensional mixture of multivariate t-distributions is given by

$$f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \mathbf{\Sigma}_g, \nu_g) = \frac{\Gamma([\nu_g + p]/2)|\mathbf{\Sigma}_g|^{-\frac{1}{2}}}{(\pi\nu_g)^{\frac{\nu_g}{2}}[1+\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \mathbf{\Sigma}_g)/\nu_g]^{\frac{(\nu_g+p)}{2}}}, \tag{2.4}$$

with mean $\boldsymbol{\mu}_g$, scale matrix $\mathbf{\Sigma}_g$, and degrees of freedom ν_g, where

$$\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \mathbf{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)'\mathbf{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)$$

is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$.

The G-component mixture of multivariate t-distributions has density:

$$g(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g), \tag{2.5}$$

with the same notation as before. The eigenvalue decomposition for scale matrices is also applied to the mixture of multivariate t-distributions. Utilizing analogues of the 14 models in GPCM family, in addition to constraining the degrees of freedom ($\nu_g = \nu$; cf. Andrews and McNicholas 2011a), leads to the tEigen family. Note that these models are implemented in the the R package `teigen` (Andrews *et al.*, 2016).

2.3 Performance assessment

The Rand index (Rand, 1971) is sometimes used to measure the agreement of two partitions of one object in a clustering analysis. It can be simply expressed as:

$$\text{Rand index} = \frac{\text{number of pairwise agreements}}{\text{total number of pairs}}. \tag{2.6}$$

Suppose we have a set S of size n. We have two ways of partitionings the set, denoted as set A and set B. Let a be the number of pairs in the same group in A and also in the same group in B; and let b be number of pairs in different groups in A and B. Then, $a + b$ is the number of pairwise agreement, and the total number of pairs is $\binom{n}{2}$. A Rand index of 1 indicates a perfect agreement.

Hubert and Arabie (1985) introduced the adjusted Rand index (ARI), which corrects the Rand index for chance agreement and is given by:

$$\text{Adjusted random index (ARI)} = \frac{\text{Rand Index} - \text{Expected Rand Index}}{\text{Max Rand Index} - \text{Expected Rand Index}}. \quad (2.7)$$

The expected value of the adjusted Rand index under random classification is 0, and for perfect classification, ARI= 1. The ARI is used in the merging criterion and in assessing classification performance of models after averaging in Sections 3.4.

2.4 Model selection

The Bayesian information criterion (Schwarz, 1978) is the most common criterion to use for selecting the best model from a family of mixture models. The BIC is

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\vartheta}}) - m \log n, \quad (2.8)$$

where $l(\mathbf{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, and $\hat{\boldsymbol{\vartheta}}$ is maximum likelihood estimate of $\boldsymbol{\vartheta}$, m is the number of free parameters, and n is the sample size. The uses and applications of BIC for model-based clustering are discussed by Leroux (1992), Kass and Raftery (1995), Kass and Wasserman (1995), and Keribin (2000).

Bayesian model averaging (BMA; Hoeting *et al.*, 1999) is one of the most popular techniques for model averaging. The BMA takes a combination of parameters across the models into consideration. Borrowing the notation of Hoeting *et al.* (1999), suppose we have data D and models $\mathcal{M}_1, ..., \mathcal{M}_K$, with Δ a quantity of interest. The posterior distribution given data D is:

$$\mathrm{pr}(\Delta \mid D) = \sum_{k=1}^{K} \mathrm{pr}(\Delta \mid \mathcal{M}_k, D)\mathrm{pr}(\mathcal{M}_k \mid D), \tag{2.9}$$

where $\mathrm{pr}(\mathcal{M}_k \mid D)$ is the posterior probability of model $\mathcal{M}_k$, i.e.:

$$\mathrm{pr}(\mathcal{M}_k \mid D) = \frac{\mathrm{pr}(D \mid \mathcal{M}_k)\mathrm{pr}(\mathcal{M}_i)}{\sum_{k=1}^{K} \mathrm{pr}(D \mid \mathcal{M}_k)\mathrm{pr}(\mathcal{M}_k)}, \tag{2.10}$$

where

$$\mathrm{pr}(D \mid \mathcal{M}_k) = \int \mathrm{pr}(D \mid \boldsymbol{\theta}_k, \mathcal{M}_k)\mathrm{pr}(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k. \tag{2.11}$$

There are two flaws for BMA. One of them is the that the number of terms in the sum (2.9) can be very large. Another one is the posterior probability $\mathrm{pr}(\mathcal{M}_k \mid D)$, which is hard to compute because of the high-dimensional integrals involved in Equation (2.10). Therefore, Occam's window is proposed by Madigan and Raftery (1994) — models that predict the data less well will not be included in the Occam's window. Note that Occam's window is given by

$$\left\{ \mathcal{M}_i : \frac{\max_l \, \mathrm{pr}(\mathcal{M}_l \mid D)}{\mathrm{pr}(\mathcal{M}_i \mid D)} \leq c \right\}, \tag{2.12}$$

and Madigan and Raftery (1994) choose $c = 20$ by analogy with a p-value of 0.05.

The BIC can be used to compute posterior distribution of D given model,

$$\mathrm{pr}(D \mid \mathcal{M}_i) = \exp\left\{ -\frac{1}{2}\mathrm{BIC}_i \right\}, \tag{2.13}$$

where BIC_i is the BIC for $\mathcal{M}_i$. Therefore, the posterior probability for $\mathcal{M}_i$ is given by

$$\mathrm{pr}(\mathcal{M}_i \mid D) = \frac{\exp\{-\frac{1}{2}\mathrm{BIC}_i\}}{\sum_{k=1}^{K} \exp\{-\frac{1}{2}\mathrm{BIC}_k\}}. \tag{2.14}$$

And we use (2.14) to compute the weights, i.e.,

$$\text{weight of } \mathcal{M}_i = \frac{\exp\{-\frac{1}{2}\Delta_i\}}{\sum_{k=1}^{K} \exp\{-\frac{1}{2}\Delta_k\}}, \tag{2.15}$$

for the averaging approaches, where

$$\Delta_i = \max_l\{\mathrm{BIC}_l\} - \mathrm{BIC}_i$$

for models $\mathcal{M}_1, ..., \mathcal{M}_K$. The models in the Occam's window based on BIC is

$$\{\mathcal{M}_i : \mathrm{BIC}_i \geq \max_l\{\mathrm{BIC}_l\} - 2 \log c\}. \tag{2.16}$$

Equation (2.9) is used for the weight of each model $\mathcal{M}_i$ that are selected within the Occam's window for our analysis in Chapter 4, i.e., (2.10).

Table 2.2: Nomenclature and number of covariance parameters for each member of the tEIGEN.

Model	$\lambda_g = \lambda$	$\mathbf{D}_g = \mathbf{D}$	$\mathbf{A}_g = \mathbf{A}$	$\nu_g = \nu$	Free covariance parameters
CIIC	C	I	I	C	1+1
CIIU	C	I	I	U	1+G
UIIC	U	I	I	C	(G-1)+1
UIIU	U	I	I	U	(G-1)+G
CICC	C	I	C	C	p+1
CICU	C	I	C	U	p+G
UICC	U	I	C	C	p+(G-1)+1
UICU	U	I	C	U	p+(G-1)+G
CIUC	C	I	U	C	Gp-(G-1)+1
CIUU	C	I	U	U	Gp-(G-1)+G
UIUC	U	I	U	C	Gp+1
UIUU	U	I	U	U	Gp+G
CCCC	C	C	C	C	[$p(p+1)/2$]+1
CCCU	C	C	C	U	[$p(p+1)/2$]+G
UCCC	U	C	C	C	[$p(p+1)/2$]+(G-1)+1
UCCU	U	C	C	U	[$p(p+1)/2$]+(G-1)+G
CUCC	C	C	C	C	G[$p(p+1)/2$]-(G-1)(p)+1
CUCU	C	C	C	U	G[$p(p+1)/2$]-(G-1)(p)+G
UUCC	U	C	C	C	G[$p(p+1)/2$]-(G-1)(p-1)+1
UUCU	U	C	C	U	G[$p(p+1)/2$]-(G-1)(p-1)+G
CUUC	C	C	U	C	G[$p(p+1)/2$]-(G-1)+1
CUUU	C	C	U	U	G[$p(p+1)/2$]-(G-1)+G
UUUC	U	C	U	C	G[$p(p+1)/2$]+1
UUUU	U	C	U	U	G[$p(p+1)/2$]+G

Chapter 3

Methodology

3.1 Merging mixture components

We consider the merging criterion for some data sets, such as the vasoconstriction data in Section 4.1.4 and the Flea Beetles in Section 4.1.5. Suppose we need to merge a G-component mixture model to give a H-component mixture model, where $H < G$. We denote the density of mixture model after merging procedure as

$$g(\mathbf{x} \mid \boldsymbol{\vartheta}^*) = \sum_{j=1}^{H} \pi_j^* f_t(\mathbf{x} \mid \boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*, \nu_j^*), \tag{3.1}$$

with the averaging parameter $\boldsymbol{\vartheta}^* = (\pi_1^*, \ldots, \pi_H^*, \boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_H^*, \boldsymbol{\Sigma}_1^*, \ldots, \boldsymbol{\Sigma}_H^*, \nu_1^*, \ldots, \nu_H^*)$, the mixture proportion $\pi_j^* > 0$, and $\sum_{j=1}^{H} \pi_j^* = 1$. Each merged mixture proportion, π_j^* is sum of the mixture proportions π_g that need to be merged, or $\pi_j^* = \pi_g$ sometimes. Each $f_t(\mathbf{x} \mid \boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*, \nu_j^*)$ with the mean $\boldsymbol{\mu}_j^*$, covariance matrix $\boldsymbol{\Sigma}_j^*$, and degree of freedom ν_j^*, is multivariate t-distribution after merging the components. For example, in Section 4.1.4, the models of Vasoconstriction data are required to merge the mixture components from $G = 3$ to $H = 2$.

There are two cases (Case I and Case II) for merging, as proposed by Wei and McNicholas (2015). Both cases are based on a 'reference model' and models in Occam's window. For Case I, the reference model is the one with the maximum BIC in Occam's window For Case II, the reference model is the model with least number of components within the Occam's window. The models in Occam's window with more components than the reference model undergo merging. The models in Occam's window with fewer components than the reference model (Case II only) are discarded.

An approach based on ARI was proposed by Wei and McNicholas (2015). Suppose we want to merge from G to H components (i.e., $G > H$). Then, the steps of the merging procedure are the following.

1. Calculating a combination matrix A of size $\binom{G}{H} \times H$, i.e.:

$$A = \begin{array}{c} A_1 \\ \vdots \\ A_{\binom{G}{H}} \end{array} \begin{pmatrix} a_{1,1} & \cdots & a_{1,H} \\ \vdots & \ddots & \vdots \\ a_{\binom{G}{H},1} & \cdots & a_{\binom{G}{H},H} \end{pmatrix}.$$

   Each row A_i represents a subset of the original G components. For example, we have components $\{1, 2, 3, 4, 5\}$, i.e., $G = 5$, to be merged to {a,b,c}, i.e., $H = 3$. Then the combination matrix A is of size $\binom{5}{3} \times 3$. Suppose $A_i = (1, 3, 4)$, this is assigned to the new components $(a, b, c) := (1, 3, 4)$. The remaining components, $\{2, 5\}$, are going to be assigned in the next step.

2. Calcuate a permutation matrix B of size $H^{(G-H)} \times (G - H)$, i.e.,

$$B = \begin{array}{c} B_1 \\ \vdots \\ B_{H^{(G-H)}} \end{array} \left(\begin{array}{ccc} b_{1,1} & \cdots & b_{1,(G-H)} \\ \vdots & \ddots & \vdots \\ b_{H^{(G-H)},1} & \cdots & b_{H^{(G-H)},(G-H)} \end{array} \right).$$

Each row B_j represents a subset of the H components. Suppose we have $B_j = (a, a)$, then the rest components $(2, 5) := (a, a)$. Now, we have $\{\{1, 2, 5\}, \{3\}, \{4\}\} := \{a, b, c\}$, and components $\{$1,2,5$\}$ need to be merged to one component.

3. Computing a matrix C of ARI with size $\binom{G}{H} \times H^{(G-H)}$ that contains all possibilities, ie.,

$$C = \begin{pmatrix} c_{1,1} & \cdots & c_{1,H^{(G-H)}} \\ \vdots & \ddots & \vdots \\ c_{\binom{G}{H},1} & \cdots & c_{\binom{G}{H},H^{(G-H)}} \end{pmatrix}$$

   There are $\binom{G}{H} \times H^{(G-H)}$ possibilities for merging components. For each element in C, $c_{i,j}$ represents the ARI value between true label with the relabelled data set using A_i and B_j.

We select the maximum ARI value in C as the best combination for merging process because the greater the ARI value, the better the classification performance.

For many data sets, Cases I and II are the same, such as the Iris data (Section 4.1.2). This happens because the model with the best BIC also has the fewest components.

3.2 Matching components

Before applying the averaging approaches, we need to match their components by measuring the Euclidean distance,

$$d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \sqrt{\sum_{p=1}^{P} \sum_{q=1}^{Q} (\mu_{ip} - \mu_{jq})^2}. \tag{3.2}$$

where $\boldsymbol{\mu}_i$ is from a component before merging and $\boldsymbol{\mu}_j$ is from a component after merging. $P = Q$, and represent the number of variables for a given data set. We measure the pairwise distance between two components within a mixture model, the minimum distance between a pair of components will be matched. This matching procedure is required before both averaging approaches.

3.3 Averaging *a posteriori* probabilities $\hat{z}_{ig}$

In the model-based clustering analysis, *a posteriori* probability z_{ig} is used to denote the cluster membership, where

$$z_{ig} = \begin{cases} 1 & \text{if observation } i \text{ belongs to component } g, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

We use the estimated *a posteriori* probability for our model based clustering averaging analysis. For a model j in tEigen family, the generated *a posteriori* probability estimates of observation i by the ECM algorithm is given by:

$$\hat{z}_{ijg} = \frac{\pi_{jg} f_t(\mathbf{x}_i \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}, \nu_{jg})}{\sum_{h=1}^{G} \pi_{jh} f_t(\mathbf{x}_i \mid \boldsymbol{\mu}_{jh}, \boldsymbol{\Sigma}_{jh}, \nu_{jh})} \tag{3.4}$$

Then we directly average the $\hat{z}_{ijg}$ for model j in Occam's window based on its weights $\text{pr}(\mathcal{M}_j \mid D)$ in (2.9). The merging criterion is applicable in this averaging approach. Before we average the models, some data sets require merging of components, e.g., Vasoconstriction in Section 4.1.4. Suppose we need to merge from G to H components for a model in Occam's window, then the averaged *a posteriori* probability (AAP) is given by:

$$\hat{z}^*_{ih} = \sum_{\mathcal{M}_j \text{ to be merged}} \text{pr}(\mathcal{M}_j \mid D)\hat{z}_{ijg}, \tag{3.5}$$

where $\sum_{h=1}^{G} \hat{z}^*_{ih} = 1$.

3.4 Model averaging

We average parameters of models based on their weights $\text{pr}(\mathcal{M}_i \mid D)$. In the model averaing approaches, we only consider the models within Occam's window that have same number of components as the model with largest BIC value. Therefore, the merging criterion is not necessary for the model averaging approach.

The parameters of the models in the tEigen family are estimated via the ECM algorithm (Celeux and Govaert, 1995). In addition, more parameters for characteristic weights u_{ig} and the degrees of freedom ν_{ig} were generated through expectation step (E-step) and conditional step (CM-step) by Andrews and McNicholas (2011). The estimated parameter $\mathbf{\Theta}_{\mathcal{M}_k}$ for models in tEigen family is given by:

$$\mathbf{\Theta}_{\mathcal{M}_k} = \{\hat{\pi}_{kg}, \hat{\boldsymbol{\mu}}_{kg}, \hat{\mathbf{\Sigma}}_{kg}, \hat{\nu}_{kg}\}_{g=1}^{G}. \tag{3.6}$$

For each parameter using weights $\text{pr}(D \mid \mathcal{M}_i)$, i.e. (2.9), we compute a weighted average of parameter estimates, i.e.,

$$\overline{\mathbf{\Theta}}_g = \{\overline{\pi}_g, \overline{\boldsymbol{\mu}}_g, \overline{\Sigma}_g, \overline{\nu}_g\}, \tag{3.7}$$

where

$$\begin{aligned}
\overline{\pi}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\pi}_{kg},\ \hat{\pi}_{kg} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{\sum_{g=1}^{G}\sum_{i=1}^{n} \hat{z}_{ig}}, \\
\overline{\boldsymbol{\mu}}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\boldsymbol{\mu}}_{kg}, \\
\overline{\mathbf{\Sigma}}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\mathbf{\Sigma}}_{kg} \\
\overline{\nu}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\nu}_{kg}
\end{aligned} \tag{3.8}$$

The AAP for each observation i is denoted as

$$\overline{z}_{ig} = \frac{\overline{\pi}_g f_t(\mathbf{x}_i \mid \overline{\mathbf{\Theta}}_g)}{\sum_{h=1}^{G} \overline{\pi}_g f_t(\mathbf{x}_i \mid \overline{\mathbf{\Theta}}_h)}, \tag{3.9}$$

where $f_t(\mathbf{x}_i \mid \overline{\mathbf{\Theta}}_g)$ is the density function of multivariate t-distribution fitted with the averaged parameter $\overline{\mathbf{\Theta}}_g$ for $\mathbf{x}_i$. The AAP $\overline{z}_{ig}$ ranges from 0 to 1.

We estimate the cluster memberships via maximum *a posteriori* probability (MAP) classification:

$$\text{MAP}\{\overline{z}_{ig}\} = \begin{cases} 1 & \text{if } g = \text{arg max}_h\{\overline{z}_{ih}\} \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

For example, we have a set of maximum *a posteriori* probabilities for observation i, where $(\overline{z}_{i1}, \overline{z}_{i2}, \overline{z}_{i3}) = (0.1, 0.7, 0.2)$. After we apply $\text{MAP}\{\overline{z}_{ig}\}$ with $g = 3$ in this example. We assign the membership for observation i to component 2.

Chapter 4

Illustration

4.1 Real data analyses

4.1.1 Bankruptcy data

The `bankruptcy` data set can be found in `R` package `MixGHD` (Tortora *et al.*, 2015). The data contains the ratio of retained earnings (RE)/ total assets, and the ratio of earnings before interests and taxes (EBIT)/ total assets of 66 American firms, and was initially recorded by Altman (1968). The selected firms were assigned into two groups with bankruptcy 0 and financially sound 1.

The *t*EIGEN models in the `teigen` package are used to fit the data. The default gives $G = 1, 2, \ldots, 9$. The UIIU model with $G = 2$ is the best model was selected by BIC (-266.68). There are six models (CUCC, UICU, UCCU, UIIC, UIUU) selected in the Occam's window with all $G = 2$ components. Because all models in Occam's window have the same number of of components, the merging criterion (Section 2.1) is not necessary in this example.

For the bankruptcy data, AAP returns the same ARI value (0.82) as the best model. However, MA provides a better classification performance with the ARI increasing from 0.82 to 0.88, cf. Table 4.1.

Table 4.1: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data bankruptcy data

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
UIIU	-266.68	2	0.52	0.82	0.82	0.88
CUCC	-268.80	2	0.18			
UICU	-269.41	2	0.13			
UCCU	-270.59	2	0.07			
UIIC	-271.06	2	0.06			
UIUU	-272.45	2	0.03			

4.1.2 Iris data

The famous iris data set is discussed by Fisher (1936), and is available in datasets package for R. There are four measurements in centimetres of the length and width of both sepal and petals from all three species of iris (*Iris setosa*, *versicolor*, and *virginica*).

We fit the iris data using the *t*EIGEN models. The BIC selected the UUUC model with two components, and gives the best classification for the data set with the largest BIC value (-795.56). There are three models (UUUC, UUCC, UUUU) in Occam's window. Each of the models has two components except model UUCC, which has three components. Therefore, merging components is required for the iris data. There are two cases to be considered while we are using AAP. The merging procedure for both cases are the same because the model with fewest components is same as the best model. Thus, the UUCC model needs to be merged into the model with two components. However, we discard the model with different components as the best model. Therefore, the weights of the models are different for the two averaging approaches.

As a result, in Table 4.4, the ARI values (0.57) of both model averaging approaches (AAP and MA) remain the same comparing with the classification performance of the best model UUUC.

Table 4.2: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data iris data.

Occam's window			$\Pr(D \mid \mathcal{M}_i)$		ARI values		
Model	BIC	G	Case I/MA		Best	AAP	MA
UUUC	-795.56	2	0.83	0.70	0.57	0.57	0.57
UUCC	-798.60	3		0.15			
UUUU	-798.79	2	0.17	0.14			

4.1.3 Female voles data

Flury (1997) discussed the Female voles data where is available in the R package `Flury` (Flury, 2015). The data consists 86 observations was measured on seven variables of two species of female voles: *Microtus Californicus* and *M. ochrogaster*. The data is fitted in tEigen with $G = 1, 2, ...9$. Based on the BIC, the best model is CCCC with $G = 2$. There are total four models in Occam's window, all with same number of components (i.e., $G = 2$). Hence, the merging procedure is not required in this example. Notably, the models also have close BIC values (i.e., $-132.61, -1324.27, -1326.80, -1328.03$). Both two averaging procedures provide the same classification performance based on the ARI (0.91) comparing with the best model.

Table 4.3: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data female voles data

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
CCCC	-1323.61	2	0.49	0.91	0.91	0.91
UCCC	-1324.27	2	0.35			
UCCU	-1326.80	2	0.10			
CCCU	-1328.03	2	0.05			

4.1.4 Vasoconstriction

The vasoconstriction data is available in the R package `Flury` (Flury, 2015). The data include 39 observations, and was based on the measurements of three variables: the volume of air inspired, the rate of air inspired, and a binary indicator (1 indicates there is vasoconstriction, 0 otherwise). We fit the data using tEigen family with $G = 1, 2, \ldots, 9$.

There were 18 models in the Occam's window; some models have the same scale structure but with a different number of components. For example, the model CIIC with $G = 3$ is selected to be the best model for this data, and model CIIC is also selected in Occam's window with $G = 2$. In this data, both cases are considered for merging, with only models with $G \geq 3$ selected for merging. The weights for Case I in AAP and MA are same becasue the models for averaging have the number of components (i.e., $G = 3$). The ARI value for the best model is 0.0143, this is close to 0 because the true label is binary ($G = 2$). Following the averaging procedure, the classification performance for Case I in AAP is improved from 0.01 to 0.02, and ARI value for Case II in AAP is 0.02. The classification performance is improved for AAP in both cases. The ARI value (i.e., 0.01) for MA procedure remains the same.

Table 4.4: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data Vasoconstriction

Occam's window			$\Pr(D \mid \mathcal{M}_i)$		ARI values			
Model	BIC	G	Case I/MA	Case II	Best	AAP		MA
						Case I	Case II	
CIIC	-224.45	3	0.33	0.18	0.01	0.02	0.02	0.01
CICC	-224.87	3	0.27	0.14				
UICC	-225.30	2		0.11				
UUUU	-225.61	2		0.10				
CICC	-225.66	3	0.18	0.10				
UICC	-226.12	2		0.08				
UUUU	-227.28	2		0.04				
CICC	-227.64	2		0.04				
UICC	-227.65	3	0.07	0.04				
UUUU	-227.72	2		0.03				
CICC	-227.89	3	0.06	0.03				
UICC	-228.27	3	0.05	0.03				
UUUU	-228.96	2		0.02				
CICC	-228.96	2		0.02				
UICC	-229.11	3	0.03	0.02				
UUUU	-229.27	2		0.02				
UICC	-229.78	2		0.01				
CICC	-229.78	2		0.01				

4.1.5 Flea Beetles data

The Flea Beetles data can be also found in R package `Flury`. The data contains 39 observations based on the measurements of 4 variables of two species of flea beetle: *Haltica Oleracea* and *H. Carduourum*. We use the tEigen models with $G = 1, 2, \ldots, 9$. The model CIIC with $G = 3$ is selected by BIC (-429.94). In the Occam's window, there are four models selected by BIC with $G = 3$, except model UUUU because with $G = 1$. In this example, two cases for merging are all considered, and the models in the Occam's window with same number of components or greater than the best model are considered for merging. In this example, models CIIC, CICC and UICC with same number of component (i.e., $G = 1$) are selected by Case I. The merging procedure is not considered here. For the other case, the models CIIC, CICC and UICC are all considered to be merged into one component. The weights are the same for both AAP and MA while we are applying both cases in the following averaging process. The classification performance of AAP for Case I is close to the best model (Table 4.5). And considering Case II of AAP, the ARI value is 0 under the expectation. Therefore, the model after merging is not good for classification in this data set because the true number of classes is $G = 2$. The ARI value is lower than the best model for MA, and indicates that classification performance of the model after averaging parameters is not good for this data set either.

Table 4.5: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data flea beetles data

Occam's window			$\Pr(D \mid \mathcal{M}_i)$		ARI values			
Model	BIC	G	Case I/MA	Case II	Best	AAP		MA
						Case I	Case II	
CIIC	-429.94	3	0.52	0.49	0.48	0.43	0	0.30
CICC	-430.36	3	0.42	0.40				
UICC	-434.36	3	0.06	0.05				
UUUU	-434.39	1		0.05				

4.1.6 Microtus classification (more vole data)

This Microtus classification data is included in the R package `Flury`. The data is fitted in the tEigen models with default $G = 1, 2, \ldots, 9$. There are two models, CCCC and UCCC, selected in Occam's window and they have the same number of components (i.e., $G = 3$). Therefore, the merging criterion is not needed. The best model is CCCC, as selected by BIC. The ARI value has improved from 0.031 to 0.033 for the model after AAP, and the ARI value also improved from 0.031 to 0.032. The classification performance has improved for both averaging procedures.

Table 4.6: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data Microtus classification (more vole data)

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
CCCC	-4347.80	3	0.77	0.031	0.033	0.032
UCCC	-4350.27	3	0.23			

4.2 Special cases

4.2.1 One model in Occam's window

There are many data sets with only one model selected by BIC in the Occam's window under an analogy of p-value of 0.05. One example, the Swiss banknotes data is discussed in the R package `mclust` (Fraley *et al.*, 2016). The data set consists six variables on measurement of 100 genuine and 100 counterfeit from an Swiss bank notes.

Another example is the Simulated minefield data (Dasgupta and Raftery, 1998;

Fraley and Raftery, 1998). This is also included in the R package `mclust`, and was simulated on the 1104 bivariate minefield observations.

We fit the data sets using the tEigen family. In the Swiss bank notes data, the model CCCC was selected by BIC within the Occam's window with $G = 4$ under the default $G = 1, 2, \dots, 9$. In the minefield data, under the similar procedure, the BIC choose only one model CIUU in the Occam's Window with $G = 4$.

4.2.2 Male Twins

The Male Twins data is also introduced in the R package `Flury`. There are six variables based on 89 observations of male twins with a factor indicating whether the twins are monozygotic or dizygotic. We fit the male twins data using the tEigen family with $G = 1, 2, \dots, 9$. There are two models, UCCC and UCCU, selected with the same number of components (i.e., $G = 2$). The best model UCCC was selected by BIC with the (ARI $= -0.01$). The merging criterion is not considered for this data. After AAP and MA, the ARI values (i.e. -0.01) remain the same for both averaging models. This is a special case for ARI value which is negative. The reason is the index of the data is smaller than the expected index.

Table 4.7: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for real data Male Twins

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
UCCC	-524.70	2	0.90	-0.01	-0.01	-0.01
UCCU	-529.09	2	0.10			

4.3 Simulated data

The simulated data is generated from the `R` package `clusterGeneration` (Qiu, 2015). The function `genRandomClust()` is used to generate random cluster data sets based on the method is proposed in Qiu and Joe (2006).

We generate $p = 3$ variables and a noise variable by setting `numNoisy=1` for $n = 440$ observations and fit the tEigen models with $G = 1, 2, \ldots, 9$. We assign the simulated data into 4 components for each model. The models CICC and CIUC with $G = 4$ are selected by the BIC. The ARI value for AAP has significant improvement from 0.89 to 0.91, compared to the best model. And the ARI value for MA remains the same.

Table 4.8: BIC, number of components and weights for models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for a simulated data with noise variable

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
CICC	-4607.67	4	0.95	0.89	0.91	0.89
CIUC	-4613.50	4	0.05			

4.4 Comparison between GPCM and tEIGEN

The six real data sets are used to fit the GPCM and tEigen families. The main purpose of this comparison is to compare the classification performance of two averaging approaches within the two families. In Table 4.5 of the bank data, the merging procedure is required. We only consider the ARI value in Case I for fitting in GPCM family in this example. The Iris data shows that the result of AAP procedure for both families remains the same compared to the best model. However, the ARI value for MA in GPCM family decreases compared to the best model. In the tEigen family, the ARI value for MA remains the same because the true cluster of Iris data is in three groups. However, the best model is selected with $G = 2$ by fitting in tEigen models. In the third female vole data, the ARI value for both AAP and MA have increased in GPCM family, and ARI value for AAP and MA are the same comparing to best model in tEigen. For Vasoconstriction data, the best ARI of best models from both families are same. However, the classification performances of selected models in tEigen are increased for both averaging approaches. Moreover, in Flea Beetles data, the GPCM models has a better classification performance for the averaging approaches. There is a similar result in Microtus classification data set. However, the ARI for models in averaging approaches increases compared to the best model in tEigen family.

Table 4.9: The GPCM vs. tEIGEN models.

	bank			Iris			Female vole		
	Best	AAP	MA	Best	AAP	MA	Best	AAP	MA
tEigen	0.824	0.824	0.881	0.568	0.568	0.568	0.908	0.908	0.908
GPCM	0.679	0.679	0.760	0.922	0.922	0.904	0.658	0.908	0.908

	Vasoconstriction				Flea Beetles			
	Best	AAP		MA	Best	AAP		MA
		Case I	Case II			Case I	Case II	
tEigen	0.01	0.02	0.02	0.01	0.48	0.43	0	0.30
GPCM	0.01	0.01	0	0.03	0.43	0.44	0	0.38

	Microtus classification		
	Best	AAP	MA
tEigen	0.031	0.033	0.032
GPCM	0.037	0.037	0.037

Chapter 5

Summary

This thesis proposed averaging model-based clustering for a family of non-Gaussian mixture models; the *t*EIGEN family. The two averaging approaches are used for the mixture of multivariate *t*-distributions. Both approaches are based on Occam's window. One of the approaches is AAP, and the other is to average the parameters of models selected by BIC which gives a new mixture model (i.e., MA). We saw that AAP may require merging mixture components. The ARI merging methodology was introduced in Wei and McNicholas (2015).

While we were following the merging and both averaging approaches, the results were given from both real and simulated data sets. For some data sets, the BIC values for models in Occam's window are very close, and the new models after averaging are similar to the best model. Therefore, the ARI values are same for these data sets. In the female voles data, the classification performance for the best and averaging criterion are the same (ARI = 0.91). In the Iris data (Table 4.2), the weight for the best model (UUUC) is 0.83, which dominates the model after averaging. The ARI value for the male twins data gives a negative ARI value (-0.01) for the best model and both models after averaging, which is close to 0. Later, we compared the classification performance between the GPCM family and the *t*Eigen family. In most data sets, *t*Eigen family shows a better classification performance of the best model and the models after the two averaging approaches, such as the bank data.

There are many works on clustering and classification using skewed distributions, including shifted asymmetric Laplace mixtures (Franczak *et al.*, 2014), skew-normal mixtures (Vrbik and McNicholas, 2014), skewed-*t* mixtures (Lin, 2010; Lee and McLachlan, 2011; Vrbik and McNicholas, 2012, 2014; Murray *et al.*, 2014) and variance-gamma mixtures (McNicholas *et al.*, 2017). Further research on averaging approaches can be applied to these mixtures. Another avenue for future direction will concern looking at alternatives to Euclidean distance for matching components (Section 3.2); a natural starting point would be the Mahalanobis distance.

Appendix A

Tables

Table A.1: BIC, number of components and weights for GPCM models within Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for the flea beetles data.

Occam's window			$\Pr(D \mid \mathcal{M}_i)$		ARI values			
Model	BIC	G	Case I/MA	Case II	Best	AAP		MA
						Case I	Case II	
EII	425.4649	3	0.5335	0.5115	0.4308	0.4430	0	0.3760
EEI	426.3587	3	0.3412	0.3271				
VEI	430.1288	3	0.0518	0.0497				
EEE	430.2747	1		0.0462				
VII	431.5576	3	0.0482	0.0243				
VII	432.1880	2		0.0177				
EEE	434.1524	2		0.0066				
EVI	435.0225	3	0.0254	0.0043				
EEI	435.0487	4		0.0042				
VEI	435.0534	2		0.0042				
VEE	435.1748	2		0.0040				

Table A.2: BIC, number of components and weights for GPCM models in Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for the voles data.

Occam's window			$\Pr(D \mid \mathcal{M}_i)$	ARI values		
Model	BIC	G		Best	AAP	MA
VEE	4359.95	3	0.9898	0.0370	0.0370	0.0370
EEE	4369.10	3	0.0102			

Table A.3: BIC, number of components and weights for GPCM models within Occam's window; ARI for true labels versus best model, AAP and MA in Occam's window for the vasoconstriction data.

Occam's window			$\Pr(D \mid \mathcal{M}_i)$		ARI values			
Model	BIC	G	Case I/MA	Case II	Best	AAP		MA
						Case I	Case II	
VII	220.4054	3	0.1760	0.1521	0.0144	0.0144	0	0.0344
EVI	220.9546	2		0.1155				
EII	220.9625	3	0.1338	0.1151				
VEI	221.5244	3	0.1332	0.0869				
VEE	221.7105	2		0.0792				
VVI	221.8105	2		0.0753				
EVE	222.9975	2		0.0416				
VII	223.2234	4		0.0372				
EEI	223.2349	2		0.0369				
EEV	223.4699	2		0.0329				
VVI	223.5010	3	0.1006	0.0323				
EEE	223.6589	2		0.0299				
EEI	223.7200	3	0.0916	0.0290				
EVI	224.1791	3	0.0872	0.0230				
VEE	224.9952	3	0.0482	0.0153				
VVE	225.2038	2		0.0138				
VEV	225.5502	3	0.0430	0.0116				
VVE	225.7304	3	0.0428	0.0106				
VEV	225.7730	2		0.0104				
EVV	226.4052	2		0.0076				
VVV	226.4444	2		0.0074				
VEI	226.7958	4		0.0062				
VEI	226.8971	2		0.0059				
EVE	227.1510	3	0.0380	0.0052				
EEE	227.2644	3	0.0374	0.0049				
EII	228.4250	4		0.0028				
EII	228.8137	2		0.0023				
VVI	228.8841	4		0.0022				
EVV	229.9408	3	0.0346	0.0013				
EEI	229.9608	4		0.0013				
EEE	230.0607	1		0.0012				
VVV	230.2185	3	0.0336	0.0011				
EII	230.3190	1		0.0011				
EVE	230.6512	4		0.0009				

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, **23**(4), 589–609.

Andrews, J. L. and McNicholas, P. D. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**(3), 361–373.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Statistics and Computing*, **22**(5), 1021–1029.

Andrews, J. L., Wickins, J. R., Boers, N. M., and McNicholas, P. D. (2016). *teigen: Model-Based Clustering and Classification with the Multivariate t Distribution.* R package version 2.1.1.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

Bhattacharya, S. and McNicholas, P. D. (2014). A LASSO-penalized BIC for mixture model selection. *Advances in Data Analysis and Classification*, **8**(1), 45–61.

Browne, R. P. and McNicholas, P. D. (2014a). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, **8**(2), 217–226.

Browne, R. P. and McNicholas, P. D. (2014b). *mixture: Mixture Models for Clustering and Classification*. R package version 1.1.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, 294–302.

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), 1–38.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

Flury, B. (1997). A first course in multivariate statistics. *Springer, New York*.

Flury, B. (2015). Flury:data sets from Flury, 1997. *R Package Version 0.1-3*.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588.

Fraley, C., Raftery, A. E., and Scrucca, L. (2016). *mclust: Normal Mixture Modeling* *for Model-Based Clustering, Classification, and Density Estimation*. R package version 5.2.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(6), 1149–1157.

Gordon, A. D. (1981). *Classification*. Chapman and Hall, London.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**(4), 382–401.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Hunter, D. L. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, **58**(1), 30–37.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**(431), 928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A*, **62**(1), 49–66.

Lee, S. and McLachlan, G. J. (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. arXiv:1109.4706.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, **20**(3), 343–356.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**(428), 1535–1546.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley, New York, 2 edition.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, volume 1451, pages 658–666. Springer-Verlag, Berlin.

McNicholas, P. D. (2016). *Mixture Model-Based Classification*. Chapman & Hall/CRC Press, Boca Raton.

McNicholas, S. M., McNicholas, P. D., and Browne, R. P. (2017). A mixture of variance-gamma factor analyzers. In A. S. E, editor, *Big and Complex Data Analysis: Methodology and Applications*. Springer, New York.

Murray, P. M., McNicholas, P. D., and Browne, R. B. (2014). A mixture of common skew-t factor analyzers. *Stat*, **3**(1), 68–82.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Qiu, W. (2015). clustergeneration: Random cluster generation (with specified degree of separation). *R Package Version 1.3.4*.

Qiu, W.-L. and Joe, H. (2006). Generation of random clusters with specified degree of separation. *Journal of Classification*, **23**(2), 315–334.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Tortora, C., ElSherbiny, A., Browne, R. P., Franczak, B. C., and McNicholas, P. D. (2015). *MixGHD: Model Based Clustering, Classification and Discriminant Analysis Using the Mixture of Generalized Hyperbolic Distributions*. R package version 2.0.

Vrbik, I. and McNicholas, P. D. (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics and Probability Letters*, **82**(6), 1169–1174.

Vrbik, I. and McNicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*, **71**, 196–210.

Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering. *Advances in Data Analysis and Classification*, **9**(2), 197–217.