Mixture Model Averaging for Clustering

Yuhong Wei[*]and Paul D. McNicholas[†]

Abstract

In mixture model-based clustering applications, it is common to fit several models from a family and report clustering results from only the 'best' one. In such circumstances, selection of this best model is achieved using a model selection criterion, most often the Bayesian information criterion. Rather than throw away all but the best model, we average multiple models that are in some sense close to the best one, thereby producing a weighted average of clustering results. Two (weighted) averaging approaches are considered: averaging the component membership probabilities and averaging models. In both cases, Occam's window is used to determine closeness to the best model and weights are computed within a Bayesian model averaging paradigm. In some cases, we need to merge components before averaging; we introduce a method for merging mixture components based on the adjusted Rand index. The effectiveness of our model-based clustering averaging approaches is illustrated using a family of Gaussian mixture models on real and simulated data.

1   Introduction

Model averaging approaches provide a coherent mechanism for accounting for model uncertainty through combining parameter estimates across a set of competing models. Bayesian model averaging (BMA; cf. Madigan and Raftery, 1994; Raftery et al., 1995) is a popular approach to model averaging that has been applied successfully to many statistical models, including linear regression (Raftery et al., 1998), generalized linear models (Raftery, 1996), Cox regression models (Hoeting, Raftery, and Madigan, 1999), and survival analysis (Volinsky et al., 1997). A comprehensive review of Bayesian model averaging is given by Hoeting, Madigan, Raftery, and Volinsky (1999). Despite its prevalence, BMA has not previously been applied in clustering applications, where there may be different clusterings for the same data set emanating from different models.

While BMA has not been used, other approaches for combining several clustering results to give a single clustering have been tried. One such method, based on least-squares, was proposed by Dahl (2006) and has been applied to Bayesian mixture models (e.g., Molitor et al., 2010). Another class of methods, called consensus clustering, has also been proposed. Fred and Jain (2005) summarize various clusterings in a co-association matrix, which indicates the strength of association between objects by analyzing how often each pair of objects appears in the same cluster. The final clustering is then determined using a voting-type algorithm. Strehl et al. (2002) propose three graph-based consensus clustering algorithms: the cluster-based similarity partitioning algorithm (CSPA), the hyper-graph partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA). A problem arises when the various clusterings have different numbers of clusters, and the same problem arises in our BMA approach; we resolve it by merging mixture components (Section 3.1).

Model-based clustering is an idiom that is often used to describe the application of a mixture model for clustering. Dating at least as far back as Wolfe (1963), interest in model-based clustering is increasing

---

[*]Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.

[†]Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada, L8S 4L8. Tel.: 905-525-9140, ext. 23419. Email `mcnicholas@math.mcmaster.ca`

steadily in application areas such as food authenticity, social networks, and microarray gene expression analyses (e.g., Yeung et al., 2001; Wehrens et al., 2004; Krivitsky et al., 2009; McNicholas and Murphy, 2010). In model-based clustering applications, it is common to fit many mixture models within a family (cf. Section 2) and then report clustering results based only on the best one. Such criterion-based methods of model selection have the general feature that the smaller (or larger, as the case may be) the value of the criterion, the stronger the 'evidence' for the model (i.e., the covariance structure and number of components, as well as number of latent variables if relevant).

In this paper, we argue that criterion-based approaches that throw away all but the best model are not necessarily optimal. One may argue, inferentially, that it should matter that many models have been fitted when reporting clustering results. Furthermore, criterion-based approaches may be unreliable when the difference between the best value of a criterion is close to one or more other values. The goal of this paper is to depart from the paradigm of selecting a single best model within the mixture model family setting. Instead, we report clustering results based on an averaging of models. While model averaging in itself is not a novel idea, its use within the model-based clustering literature marks a significant departure from the accepted norm. We consider two different approaches to averaging herein: averaging models to produce one interpretable model and directly averaging *a posteriori* probabilities.

The remainder of this paper is outlined as follows. In Section 2, background on model-based clustering is provided. In Section 3, we describe our model averaging schemata, including a novel approach for merging components, before illustrating our approaches on real data (Section 4). The paper concludes with discussion and suggestions for future work (Section 5).

2 Model-Based Clustering

2.1 Gaussian Parsimonious Clustering Models

A p-dimensional random vector $\mathbf{X}$ arises from a parametric finite mixture distribution if its density can be written $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \, f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$, where G is the number of components, π_g are mixing proportions $(\pi_g > 0, \sum_{g=1}^{G} \pi_g = 1)$, and $\boldsymbol{\vartheta}$ denotes the model parameters, i.e., $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$. Until very recently, Gaussian mixtures have dominated the model-based clustering literature. The likelihood for p-dimensional $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a Gaussian mixture model is

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

where $\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\vartheta}$ once again denotes the model parameters.

The density in (1) has a total of $(G-1) + Gp + Gp(p+1)/2$ free parameters and so it is usually necessary to introduce parsimony. A total of $Gp(p+1)/2$ of these parameters are in the component covariances, so imposing the isotropic constraint on the component covariances, i.e., $\boldsymbol{\Sigma}_g = \sigma_g \mathbf{I}_p$, is a very simple way to reduce the number of parameters from quadratic to linear in p. Of course, this constraint — implying spherical components with different volumes — will not be practical for most applications and so less restrictive constraints are needed. Several such approaches have been tried, usually based on imposing constraints upon a decomposed component covariance structure. The most famous such approach is based on an eigen-decomposition of the component covariance matrices (Banfield and Raftery, 1993), i.e., $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, where $\mathbf{D}_g$ is the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, $\mathbf{A}_g$ is a diagonal matrix, with $|\mathbf{A}_g| = 1$, containing elements proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$, and λ_g is the associated constant. Celeux and Govaert (1995) impose constraints on λ_g, $\mathbf{D}_g$, and $\mathbf{A}_g$ to obtain a family of 14 Gaussian parsimonious clustering models (GPCMs; Table 1). The `mixture` package (Browne and McNicholas, 2013) for the R software (R Core Team, 2013) provides an implementation of all 14 members of the GPCM family. Parameter

estimation for 12 of the members of the GPCM family is carried out using the expectation-maximization (EM) algorithm (Dempster et al., 1977) and details are given by Celeux and Govaert (1995). An MM algorithm (cf. Hunter and Lange, 2004) is used for the other two models (EVE and VEE), and extensive details are provided by Browne and McNicholas (2014).

Table 1: Nomenclature, covariance structure, and number of covariance parameters for each member of the GPCM family.

	Volume	Shape	Orientation	$\mathbf{\Sigma}_g$	Free cov. paras.
EII	Equal	Equal	NA	$\lambda\mathbf{I}$	1
VII	Variable	Equal	NA	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Cord. Axes	$\lambda\mathbf{A}$	p
VEI	Variable	Equal	Cord. Axes	$\lambda_g\mathbf{A}$	$p + G - 1$
EVI	Equal	Variable	Cord. Axes	$\lambda\mathbf{A}_g$	$Gp - G + 1$
VVI	Variable	Variable	Cord. Axes	$\lambda_g\mathbf{A}_g$	Gp
EEE	Equal	Equal	Equal	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{D}_g\mathbf{A}\mathbf{D}'_g$	$Gp(p+1)/2 - (G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g\mathbf{D}_g\mathbf{A}\mathbf{D}'_g$	$Gp(p+1)/2 - (G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}'_g$	$Gp(p+1)/2$
EVE	Equal	Variable	Equal	$\lambda\mathbf{D}\mathbf{A}_g\mathbf{D}'$	$p(p+1)/2 + (G-1)(p-1)$
VVE	Variable	Variable	Equal	$\lambda_g\mathbf{D}\mathbf{A}_g\mathbf{D}'$	$p(p+1)/2 + (G-1)p$
VEE	Variable	Equal	Equal	$\lambda_g\mathbf{D}\mathbf{A}\mathbf{D}'$	$p(p+1)/2 + (G-1)$
EVV	Equal	Variable	Variable	$\lambda\mathbf{D}_g\mathbf{A}_g\mathbf{D}'_g$	$Gp(p+1) - (G-1)$

A typical application of the GPCM family of models consists of running each of the models (Table 1) for a range of values of G. Then, the best of these models is selected using some criterion and the associated classifications are reported. The most popular criterion for this purpose is the Bayesian information criterion (BIC; Schwarz, 1978), i.e.,

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

where $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, $l(\mathbf{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, and ρ is the number of free parameters. Leroux (1992) and Keribin (2000) give theoretical results that, under certain regularity conditions, support the use of the BIC for choosing the number of components in a mixture model. In addition, model selection has been based on the BIC in a wide range of model-based clustering applications (recent examples include Bouveyron et al., 2007, McNicholas and Murphy, 2008, Andrews and McNicholas, 2011, and Vrbik and McNicholas, 2014). Nonetheless, the model with the smallest BIC does not necessarily give the best predicted classifications or the 'correct' number of components (see Biernacki et al., 2000, for discussion).

2.2 *A Posteriori* Merging of Components

In Gaussian model-based clustering, clusters are often taken as synonymous with mixture components. However, a cluster might itself be well modelled by a mixture of Gaussian distributions. In such cases, mixture components can be *a posteriori* merged to obtain clusters, and different approaches have been considered to do this, e.g., Baudry et al. (2010) and Hennig (2010). Baudry et al. (2010) proposed finding the number of clusters with a hierarchical merging procedure that minimizes the entropy of the clustering solution obtained by fitting the MCLUST family to the data — the MCLUST family is a subset of 10 of the GPCM models (cf. Fraley and Raftery, 2002). Hennig (2010) followed a similar hierarchical merging procedure, where some aggregation criteria are proposed based on either modality (i.e., merging components that produce a unimodal distribution) or on misclassification probabilities (i.e., merging components to minimize misclassification probabilities). In our merging procedure, we maximize the adjusted Rand index (ARI; Hubert and Arabie, 1985) with a reference model.

Recent work using non-Gaussian mixture models for clustering has brought into sharp relief the fact that, although it can work very well in some cases, merging Gaussian components is not a "get out of jail free card" (McNicholas and Browne, 2013). The method of Baudry et al. (2010), and accompanying software, is discussed by Vrbik and McNicholas (2014), who compare the *a posteriori* merging approach of Baudry et al. (2010) with results from fitting mixtures of skew-normal distributions and mixtures of skew-t distributions. In these experiments, the merging approach gave inferior results to fitting mixtures with more flexible component densities. Similar results around the inferiority of merging components over fitting non-Gaussian mixtures have been reported by others, including Murray et al. (2014) and Franczak et al. (2014). While not considered herein, our averaging approaches could be used to average non-Gaussian mixtures or *a posteriori* probabilities arising therefrom. Of course, great care would need to be taken when averaging mixtures of skewed distributions, especially in cases where each component has its own skewness parameter.

2.3 Bayesian Model Averaging

Statistical inference with families of mixture models typically involves selection of a model using some criteria and then proceeding as if the selected model has generated the data, without considering the additional uncertainty introduced by ignoring all other models. Difficulties arise with this approach because it is practically impossible to decide what to do when the difference between the values of the criteria for two different models is 'small'. Model averaging, including frequentist model averaging (e.g., Hjort and Claeskens, 2003) and BMA, takes model uncertainty into consideration by combining parameter estimates across different models. Bayesian model averaging is a popular technique for model averaging, cf. Hoeting, Madigan, Raftery, and Volinsky (1999). Borrowing their notation, suppose we have models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$, and Δ is the quantity of interest. The posterior distribution of Δ given data D is

$$\text{pr}(\Delta \mid D) = \sum_{i=1}^{K} \text{pr}(\Delta \mid \mathcal{M}_i, D)\text{pr}(\mathcal{M}_i \mid D), \tag{2}$$

where $\text{pr}(\Delta \mid \mathcal{M}_i, D)$ is the posterior distribution of Δ under model $\mathcal{M}_i$, $\text{pr}(\mathcal{M}_i \mid D)$ is the posterior probability for model $\mathcal{M}_i$, i.e.,

$$\text{pr}(\mathcal{M}_i \mid D) = \frac{\text{pr}(D \mid \mathcal{M}_i)\text{pr}(\mathcal{M}_i)}{\sum_{i=1}^{K} \text{pr}(D \mid \mathcal{M}_i)\text{pr}(\mathcal{M}_i)}, \tag{3}$$

and

$$\text{pr}(D \mid \mathcal{M}_i) = \int_{\boldsymbol{\Theta}_i} \text{pr}(D \mid \boldsymbol{\theta}_i, \mathcal{M}_i)\text{pr}(\boldsymbol{\theta}_i \mid \mathcal{M}_i)d\boldsymbol{\theta}_i, \tag{4}$$

where $\boldsymbol{\theta}_i$ is the vector of parameters of model $\mathcal{M}_i$, $\text{pr}(\boldsymbol{\theta}_i \mid \mathcal{M}_i)$ is the prior for $\boldsymbol{\theta}_i$ under model $\mathcal{M}_i$, and $\text{pr}(\mathcal{M}_i)$ is the prior probability of model $\mathcal{M}_i$.

Madigan and Raftery (1994) show that BMA can give better predictive performance than any single model; however, BMA has two major implementation difficulties. One is that the number of models in the summation in (2) can be prohibitively large. Another is that the posterior model probabilities are hard to compute because they involve very high-dimensional integrals. To overcome the former problem, Madigan and Raftery (1994) proposed using Occam's window to choose a set of models. The idea is that if a given model is a far worse fit to the data than the best model, then it is discarded and no longer considered. More formally, models not in Occam's window

$$\left\{ \mathcal{M}_i : \frac{\max_l \left\{\text{pr}(\mathcal{M}_l \mid D)\right\}}{\text{pr}(\mathcal{M}_i \mid D)} \leq c \right\} \tag{5}$$

are discarded, where c is some positive number. Madigan and Raftery (1994) use $c = 20$ in their analyses, drawing an analogy to a p-value of 0.05, and we use the same value in our analyses to determine which models to average (Section 4).

4

In the case of mixture models, the BIC can be used to approximate the integral in (4) (Kass and Raftery, 1995; Dasgupta and Raftery, 1998). Specifically, the BIC approximation can be used to compute $\text{pr}(D \mid \mathcal{M}_i)$ via

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

where BIC_i is the BIC value for model $\mathcal{M}_i$. Therefore, under equal prior probabilities, $\text{pr}(\mathcal{M}_i \mid D)$ can be computed based on (3), i.e.,

$$\text{pr}(\mathcal{M}_i \mid D) = \frac{\exp\left\{-\frac{1}{2}\text{BIC}_i\right\}}{\sum_{i=1}^{K}\exp\left\{-\frac{1}{2}\text{BIC}_k\right\}}, \tag{6}$$

and Occam's window (5) is equivalent to

$$\left\{\mathcal{M}_i : \text{BIC}_i - \min_l\{\text{BIC}_l\} \leq 2\log c\right\}. \tag{7}$$

Herein, we use (6) to compute the weights for our averaging approaches and we use Occam's window (7) to decide which models to average. Before describing our averaging approaches, we need to consider how to merge mixture components; this will be necessary when models in Occam's window have different numbers of components and we want to use all of the models.

3 Methodology

3.1 Merging Mixture Components

In this section, we introduce a procedure for merging mixture components. Suppose we have a G-component mixture model and we want to merge components to produce an H-component mixture, where $H < G$. The density of the resulting mixture is just another representation of the original model, which can be denoted

$$f(\mathbf{x}) = \sum_{j=1}^{H}\pi_j^* f_j^*(\mathbf{x}) = \sum_{g=1}^{G}\pi_g\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{8}$$

where each π_j^* is equivalent to one of or the sum of some of the mixing proportions $\pi_1, \ldots, \pi_G$, and each $f_j^*(\mathbf{x})$ is given by one of or a mixture of some of the component densities $\phi(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, \phi(\mathbf{x} \mid \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)$. For example, suppose that $G = 3$ and $H = 2$; then, $f_1^*(\mathbf{x})$ could be a convex combination of the first two Gaussian components, in which case $f_2^*(\mathbf{x})$ would equal the third one. This gives us a straightforward method of thinking about merged components but does not tell us which components should be merged with one another, e.g., when we considered the example with $G = 3$ and $H = 2$, we looked at only one of three possible merging outcomes.

We develop a mixture model component merging criterion based on the ARI, which is the Rand index (Rand, 1971) corrected for chance agreement. The Rand index compares two data partitions based on pairwise agreement and is given by the number of pairwise agreements divided by the total number of pairs. The correction leading to the ARI is performed to account for the fact that some cases will be correctly classified by chance if classification is performed randomly. An ARI value of 1 corresponds to perfect class agreement and the expected value of the ARI under random classification is 0. Some properties of the ARI are discussed by Steinley (2004).

Herein, we use the ARI to merge components based on a 'reference model'. The purpose of merging components is that models in Occam's window have the same number of components as the reference model so that we can use them all if desired. Two cases are considered.

- In Case I, the model with the smallest BIC is the reference model. Merging is performed for models in Occam's window with more components than the reference model; however, models in Occam's window with fewer components than the reference model are discarded.

- In Case II, the model with the fewest components is the reference model and merging is performed on other models in Occam's window, as needed, to give the same number of components.

Of course, Case I and Case II are equivalent when the model with the smallest BIC also has the fewest components. Note that in Case I we are effectively assuming that the BIC will only overestimate the number of components; however, we take a more flexible position in this regard in Case II.

Our merging process is best illustrated with an example. Suppose that one model inside Occam's window has seven components and that the reference model is a four-component model. The partition corresponding to the reference model is used as the underlying 'true' classification in the merging process and is called the 'reference partition'. In what follows here, to avoid confusion, we use $\{1, 2, \ldots, 7\}$ to denote components of the seven-component model and $\{a, b, c, d\}$ to denote components in the reference model. The following steps illustrate how our merging strategy works.

1. A combination matrix $\mathbf{A}$ of size 35×4 is generated. Note that $\binom{7}{4} = 35$, where each row represents a partial clustering, e.g., if the 22nd row of $\mathbf{A}$ is $\mathbf{a}_{22} = (2, 3, 4, 6)$, then component 2 goes into new component a, component 3 goes into new component b, component 4 goes into new component c, and component 6 goes into new component d.

2. For each row in $\mathbf{A}$, we must deal with the remaining components, e.g., $\{1, 5, 7\}$ in the case of row 22 with respect to the components a, b, c, d introduced above. A permutation matrix $\mathbf{B}$ (64×3) is created to include all possibilities. One sample row is $\mathbf{b}_j = (a, a, d)$, which puts component 1 into component a, component 5 into component a, and component 7 into component d, so the model after merging has the new components $\{a, b, c, d\} = \{1 \cup 2 \cup 5, 3, 4, 6 \cup 7\}$.

3. The ARIs between the reference partition and all partitions arising from the model after merging are calculated and recorded in the 35×64 matrix $\mathbf{C}$. For example, in step 1, $\mathbf{a}_{22} = (2, 3, 4, 6)$ is the 22nd row in matrix $\mathbf{A}$, in step 2, $\mathbf{b}_j = (a, a, d)$ is the 4th row in matrix $\mathbf{B}$, and the ARI value between reference partition and this model after merging is stored at the 22th row and the 4th column of the matrix $\mathbf{C}$.

Once we have every element of the matrix $\mathbf{C}$, the best merging combination is chosen to correspond to the largest ARI value, i.e., the largest element in $\mathbf{C}$.

3.2 Averaging *A Posteriori* Probabilities

In model-based clustering applications, we use z_{ig} to denote component membership, where $z_{ig} = 1$ if observation i is in component g and $z_{ig} = 0$ otherwise. After parameters have been estimated, e.g., via the EM algorithm, the *a posteriori* probabilities (or, equivalently, expected values)

$$\hat{z}_{ig} := \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{h=1}^G \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)} \tag{9}$$

are used to produce the predicted classifications. Typically, the $\hat{z}_{ig}$ are hardened to either 0 or 1 based on maximum *a posteriori* (MAP) probabilities, i.e., $\text{MAP}(\hat{z}_{ig}) = 1$ if $\max_h\{\hat{z}_{ih}\}$ occurs in component $h = g$, and $\text{MAP}(\hat{z}_{ig}) = 0$ otherwise. See Basford and McLachlan (1985) for a discussion of allocation of observations into clusters in the model-based clustering context. Within our averaging approaches, if components of a model in Occam's window need to be merged, then, using the notation in (8), the post-merging $\hat{z}_{ij}^*$ will correspond to one of or the sum of pre-merging $\hat{z}_{ig}$, for $j = 1, \ldots, H$ and $g = 1, \ldots, G$. For example if $\hat{f}_1^*(\mathbf{x}_i)$ is a convex combination of what were the first two components, then we simply have $\hat{z}_{i1}^* = \hat{z}_{i1} + \hat{z}_{i2}$.

After any necessary merging has been carried out for models within Occam's window, and models are dropped if necessary (i.e., if in Case I there are fewer components than the reference model), we compute a weighted average of the *a posteriori* probabilities. For each i, we average the *a posteriori* probabilities

from the models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$ using the weights $\Pr(\mathcal{M}_i \mid D)$, cf. (6). Averages of the *a posteriori* probabilities can then be hardened to give predicted classifications.

3.3 Model Averaging

In addition to averaging *a posteriori* model probabilities, we also consider a direct averaging of models. This time, we only consider models within Occam's window that have the same number of components as the model with the smallest BIC. We compute a weighted average of the parameter estimates for each parameter in these models $\mathcal{M}_i$ using the weights $\Pr(\mathcal{M}_i \mid D)$, cf. (6). Expected values $\hat{z}_{ig}$ are then computed based on the averaged parameter estimates and the predicted classifications are based on the associated maximum *a posteriori* probabilities.

This model averaging approach is quite different from averaging *a posteriori* probabilities (Section 3.2). Rather than focusing on a weighted average of *a posteriori* probabilities, possibly after merging, here we are concerned with producing classifications from one interpretable model. That is, the classifications from our model averaging approach will come from one parameterized, interpretable mixture model, e.g., if we average G-component models from the GPCM family, the result of model averaging is a G-component Gaussian mixture model.

Note that for all of our averaging approaches, we need to be careful to correctly match components across models, e.g., component 1 in the best model might be equivalent to component 2 in the second best model. This issue is related to the problem of label switching in mixture models (cf. Stephens, 2000). In both averaging strategies, we match components based on minimum distance between estimated component means.

3.4 Comments

Although more generally applicable, we develop and demonstrate our averaging approaches for Gaussian mixture models. Naturally, questions arise around what sort of situations might benefit from these approaches and what other methods are appropriate for comparison. Recall that the objective of our averaging approaches is to improve on the clustering performance of a single, i.e., the best, (Gaussian) mixture model. In the case of model averaging (Section 3.3), the result is itself a Gaussian mixture model. Accordingly, model averaging is expected to work well for clustering in the same kind of situations where Gaussian mixture models would be expected to work well. When averaging *a posteriori* probabilities in Case I (Section 3.2), the best model is the reference model, models with fewer components are discarded, and merging is used for models with more components than the reference model before averaging is carried out. Because the best model is the reference model, averaging *a posteriori* probabilities in Case I tends to produce similar clustering results to model averaging, and this is borne out in our real and simulated data analyses (Section 4).

When averaging *a posteriori* probabilities in Case II (Section 3.2), the model with the fewest components is the reference model and all models in Occam's window are used for averaging (after merging is carried out where needed). Unless the best model is also the model with the fewest components, averaging *a posteriori* probabilities in Case II produces clusters that may be quite different from those produced by model averaging or averaging *a posteriori* probabilities in Case I. There are two reasons for this: the number of components is different and more models are used in the averaging. That is not to say that the clustering results arising from averaging *a posteriori* probabilities in Case II will be very different to those produced by a single Gaussian mixture model — after all, we know there is a well-fitting (as determined by being inside Occam's window) single Gaussian mixture with the same number of components — but rather that clustering results arising from averaging *a posteriori* probabilities in Case II might be quite different to those produced by the best model or one of our other averaging approaches. This is also borne out in our real and simulated data analyses (Section 4).

It is possible that averaging *a posteriori* probabilities in Case II might successfully capture clusters that are not well-described by a Gaussian distribution; however, it can only do this to the extent that a model

within Occam's window has selected the appropriate number of components (e.g., Section 4.3, Scenario III). Because of this limitation, we do not proffer averaging *a posteriori* probabilities in Case II, or either of our other averaging approaches, as techniques for finding clusters that cannot be captured using a single Gaussian distribution. In such cases, a non-Gaussian mixture model could be used, *a posteriori* merging of components (cf. Section 2.2) could be applied to the best model or to the model that results from model averaging (e.g., Section 4.3, Scenario III), or the approaches used herein could be applied to average non-Gaussian mixture models.

4 Data Analyses

4.1 Performance Assessment

All of our examples are treated as genuine clustering examples, i.e., no prior knowledge of the labels or the number of components is used. However, the true labels are known in each case and we can use them to evaluate classification performance. There are many indices that can be used to compare agreement between true and predicted classes. In a comprehensive comparison, Milligan and Cooper (1986) recommended the ARI, which has become the index of choice in model-based clustering applications. As mentioned in Section 3, Steinley (2004) discusses some properties of the ARI. We use the ARI to evaluate classification performance herein.

4.2 Real Data Analyses

4.2.1 Italian wine

Forina et al. (1986) present data on chemical and physical properties of 178 samples of three varieties (Barolo, Barbera, and Grignolino) of wine from the Piedmont region of Italy. A data set containing 27 physical and chemical properties is available in the `pgmm` package (McNicholas et al., 2014) for R. We fit the GPCM models to these data using `mixture`, under the default settings, for $G = 1, 2, \ldots, 9$. The VVI model with $G = 3$ components is the best model in terms of the BIC (12103.74), but the EVI model has an almost identical BIC value (12103.81). This is an example where choosing the 'best' model in terms of the BIC and ignoring all other models is clearly questionable because the BIC values are so close. Although the BIC values are extremely close, the classifications associated with the respective models are different, i.e., the respective MAP classifications disagree on four of the wine samples.

We apply our averaging procedures to these data. Only the aforementioned two models lie in Occam's window and, because both have $G = 3$ components, merging components is not required. The results (Table 2) indicate that averaging *a posteriori* probabilities leads to a very slight improvement in the ARI over the best model (going from 0.8951 to 0.8962), while model averaging gives a slightly inferior ARI (0.8456) when compared to the best model.

Table 2: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model, from averaging *a posteriori* probabilities (AAP), and from model averaging (MA), respectively, for the Italian wine data.

Occam's Window			$\Pr(\mathcal{M}_i \mid D)$	ARI Values		
Model	BIC	G		Best	AAP	MA
VVI	12103.74	3	0.5088	0.8951	0.8962	0.8456
EVI	12103.81	3	0.4912			

4.2.2 Female voles data

Flury (1997) discusses seven measurements on female voles from two species, *Microtus californicus* and *Microtus ochrogaster*. The data are available within the R package `Flury` (Flury, 2012). The GPCM models are fitted to these data using `mixture`, for $G = 1, 2, \ldots, 9$. The BIC selected a VEE model with $G = 3$ components and ARI = 0.6577. We apply our averaging procedures to these data. There are three models within Occam's window, i.e., the best model and two models with $G = 2$ components. When averaging *a posteriori* probabilities, we merge components of the best model to give a two-component model (i.e., Case II). Averaging *a posteriori* probabilities leads to much better classification performance (ARI = 0.9081) when compared with the best model (ARI = 0.6577), cf. Table 3. Because we only consider model averaging when there are models within Occam's window with the same number of components as the best model, it is not applicable here.

Table 3: A summary of the models in Occam's window along with ARI values for the true classifications versus classifications from the best model and from averaging *a posteriori* probabilities (AAP), respectively, for the female voles data.

Occam's Window			$\Pr(\mathcal{M}_i \mid D)$	ARI Values	
Model	BIC	G		Best	AAP
VEE	1309.37	3	0.9540		
EEE	1316.70	2	0.0244	0.6577	0.9081
VEE	1316.95	2	0.0216		

4.2.3 Swiss bank notes

The Swiss bank note data are available in the `mclust` package (Fraley et al., 2013), and comprise six measurements on 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes. The GPCM models are fitted to these data using `mixture`, for $G = 1, 2, \ldots, 9$. The BIC selects an EEE model with $G = 4$ components and ARI = 0.6790. This model uses two components for each of the true labels (i.e., genuine and counterfeit). We apply our averaging procedures to these data. This time, three models lie in Occam's window: the first (i.e., the best) and third models have $G = 4$ components and the second model has $G = 3$ components.

When averaging *a posteriori* probabilities, we must either ignore the three-component model (Case I) or merge components within the four-component models to give three-component models (Case II). The results (Table 4) indicate that averaging *a posteriori* probabilities under Case I leads to the same classification performance as given by the best model (i.e., ARI = 0.6790); however, averaging *a posteriori* probabilities under Case II leads to a great improvement in classification performance (ARI = 0.9068). When carrying out model averaging, we only consider the two models with $G = 4$ components, i.e., the first and third models, and the results (Table 4) show that model averaging has led to an improved classification performance (ARI = 0.7602) when compared with the best model (ARI = 0.6790). However, the classification improvement using model averaging is notably less than that from averaging *a posteriori* probabilities in Case II.

4.2.4 Iris data

Fisher (1936) discussed four measurements, in centimetres, of the length and width of sepals and petals from three species of irides (*Iris setosa*, *Iris virginica*, and *Iris versicolor*), originally collected by Anderson (1935). The data consist of 50 samples from each of three species and is available in the `datasets` package for R. The GPCM models are fitted to these data using `mixture`, for $G = 1, 2, \ldots, 9$. The BIC selected an VEV model with $G = 3$ components and ARI = 0.9222. Three models lie in Occam's window and they do not have the same number of components. Similar to our analysis of the Swiss bank notes data

Table 4: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model, from averaging *a posteriori* probabilities (AAP), and from model averaging (MA), respectively, for the bank note data.

| Occam's Window | | | Pr($\mathcal{M}_i \mid D$) | | ARI Values | | | |
Model	BIC	G	Case I / MA	Case II	Best	AAP Case I	Case II	MA
EEE	2651.92	4	0.9819	0.6736				
VEE	2653.44	3		0.3141	0.6790	0.6790	0.9068	0.7602
VEE	2659.91	4	0.0181	0.0124				

(Section 4.2.3), we must either ignore one model (Case I) or merge components within the three-component models to give two-component models (Case II). Averaging *a posteriori* probabilities under Case I leads to the same classifications as given by the best model (i.e., ARI = 0.9222), while Case II gives an inferior ARI (0.5681) when compared to the best model (cf. Table 5). Because there are three species of iris in these data, it is not surprising that Case II does not yield a favourable result here. Model averaging is also performed and gives a slightly inferior ARI (0.9039) when compared to the best model (Table 5).

Table 5: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model, from averaging *a posteriori* probabilities (AAP), and from model averaging (MA), respectively, for the iris data.

| Occam's Window | | | Pr($\mathcal{M}_i \mid D$) | | ARI Values | | | |
Model	BIC	G	Case I / MA	Case II	Best	AAP Case I	Case II	MA
VEV	789.37	3	0.9833	0.6531				
VVV	790.70	2		0.3358	0.9222	0.9222	0.5681	0.9039
VVV	797.52	3	0.0167	0.0111				

4.2.5 Coffee data

The coffee data comprise 43 observations on twelve chemical properties of coffee samples collected from two species (Arabica and Robusta) from 29 countries around the world. These data are available in the `pgmm` package for R. The `mixture` package is used to fit the GPCM family to the twelve chemical properties for $G = 1, \ldots, 9$. The VEI model with $G = 2$ was chosen by the BIC, and there is one other model within Occam's window, i.e., a $G = 2$ component EEI model. Performing model averaging via *a posteriori* probabilities or model averaging leads to perfect classification (ARI = 1; Table 6), as does the best model.

Table 6: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model, from averaging *a posteriori* probabilities (AAP), and from model averaging (MA), respectively, for the coffee data.

| Occam's Window | | | Pr($\mathcal{M}_i \mid D$) | ARI Values | | |
Model	BIC	G		Best	AAP	MA
VEI	1334.22	2	0.6843			
EEI	1335.72	2	0.3157	1	1	1

4.2.6 Examples with only one model in Occam's window: breast cancer data and olive oil data

In a Wisconsin breast cancer study, several features are recorded for 681 cases of potentially cancerous tumours, of which 238 are actually malignant. Mangasarian et al. (1995) use these data to establish whether fine needle aspiration can determine tumour status. These data are available in the `faraway` package (Faraway, 2011) for R.

The olive oil data, available in the `pgmm` package for R, comprise 572 observations on the percent composition of olive oil with respect to eight fatty acids found by lipid fractionation from nine areas across three regions (Southern Italy, Sardinia, and Northern Italy) in Italy.

Fitting the GPCM models to each of these data sets using `mixture`, for $G = 1, 2, \ldots, 9$, results in only one model lying within Occam's window. For the olive oil data, the BIC chooses an EVE model with $G = 9$ components. For the breast cancer data, the BIC chooses an EVE model with $G = 6$ components. We note that $G = 9$ is the upper end of the range of values that was used for G (i.e., $G = 1, \ldots, 9$); accordingly, we repeated the analysis of these data using the range $G = 1, \ldots, 20$. The same result was obtained, i.e., the EVE model with $G = 9$ components was the only model inside Occam's window. Of course, when there is only one model in Occam's window, averaging is equivalent to reporting classifications from the model with the best BIC.

4.3 Simulated Data

Three simulation scenarios are considered. In each case, data are generated via the `genRandomClust()` function from the R package `clusterGeneration` (Qiu and Joe, 2012). The `genRandomClust()` function generates random clusters based on the method proposed by Qiu and Joe (2006).

Figure 1: A pairs plot for one of the 30 data sets from Scenario I.

In Scenario I, we generate 30 data sets with $p = 3$ variables, $n = 480$ observations, and substantially overlapping clusters by setting `numClust=4`, `sepVal=0.03`, `numNonNoisy=3`, and `clustszind=2` (e.g., Figure 1). This is a very difficult clustering problem and perfect classification results are not expected. We use `mixture` to fit each of the 30 data sets for $G = 1, 2, \ldots, 9$. Over the 30 runs, the ARI values achieved by the

best model, averaging *a posteriori* probabilities (Cases I and II), and model averaging were all similar (Table 7). However, the performance of averaging *a posteriori* probabilities in Case II might be considered slightly preferable to the others because it has the largest mean ARI as well as the smallest standard deviation.

Table 7: The summary for the best model, averaging *a posteriori* probabilities (AAP), and model averaging (MA) among 30 runs from Scenario I.

ARI	Best	AAP		MA
		Case I	Case II	
Average	0.7345	0.7373	0.7528	0.7442
Std. Deviation	0.0624	0.0631	0.0367	0.0604

In Scenario II, we generate $p = 3$ variables for $n = 644$ observations and we add one noise variable (by setting `numNoisy=1`). We used `mixture` to fit the GPCM models to these data for $G = 1, 2, \ldots, 9$. The BIC selects a VEE model with $G = 4$ and ARI = 0.9061, and there are four models within Occam's window (Table 8). One of the models (EEI, $G = 6$) has more components than the best model and so merging is required before averaging *a posteriori* probabilities. Both model averaging (ARI = 0.9196) and averaging *a posteriori* probabilities (ARI = 0.9092) lead to a slight improvement in classification performance, with the former being slightly better.

Table 8: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model, from averaging *a posteriori* probabilities (AAP), and from model averaging (MA), respectively, for the simulated data from Scenario II.

Occam's Window			$\Pr(\mathcal{M}_i \mid D)$		ARI Values		
Model	BIC	G	AAP	MA	Best	AAP	MA
VEE	11981.63	4	0.4410	0.4504			
EEE	11982.15	4	0.3401	0.3473	0.9061	0.9092	0.9196
EVE	11983.23	4	0.1990	0.2024			
EEI	11987.83	6	0.0199				

In Scenario III, we generated four components. Two are simulated from bivariate Gaussian distributions, and the other two are uniformly generated from triangles via the rejection method (Figure 2). There are 500 points, 100 in each of the triangles and 150 in each of the Gaussian components. These data were generated to illustrate our merging approaches when the model generating the data is very clearly not within the family of models being used for analysis.

We fit the GPCM models to these data using `mixture`, for $G = 1, \ldots, 9$. A VVV model with $G = 6$ components and ARI = 0.9061 was selected by the BIC (Figure 3), and there are four other models in Occam's window (with $G = 4, 5, 6$, and 7 components, respectively). Averaging *a posteriori* probabilities in Case I leads to some improvement in ARI (0.9344), and in Case II leads to perfect classification (Table 9, Figure 3). Model averaging gives the same classification performance as the best model (Table 10); notably, if we merge the components of the resulting model, perfect classification results are obtained (Figure 3). This is a nice illustration of how model averaging can be followed by component merging to model non-Gaussian clusters.

4.4 Comparison with Strehl's Consensus Clustering Algorithms

In this section, our methodology is compared with the consensus clustering algorithms proposed by Strehl et al. (2002). Consensus clustering is defended as forming a single consolidated clustering through combining multiple clusterings of a set of objects without considering the features or algorithms that determined these clusterings. Strehl et al. (2002) introduced three efficient heuristics to solve the cluster ensemble problem:

Figure 2: Scatterplot of the simulated data from Scenario III.

Table 9: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model and from averaging *a posteriori* probabilities (AAP), respectively, for the simulated data from Scenario III.

Occam's Window			$\Pr(\mathcal{M}_i \mid D)$		ARI Values		
Model	BIC	G	Case I	Case II	Best	AAP	
						Case I	Case II
VVV	4168.84	6	0.6820	0.5471			
VVI	4170.39	6	0.3145	0.2523			
VVV	4171.37	5		0.1547	0.9061	0.9344	1
VVV	4173.92	4		0.0431			
VVI	4179.37	7	0.0035	0.0028			

Table 10: A summary of the models in Occam's window along with ARI values for the true labels versus predicted classifications from the best model and from model averaging (MA), respectively, for the simulated data from Scenario III.

Occam's Window & $G = 6$			$\Pr(\mathcal{M}_i \mid D)$	ARI Values	
Model	BIC	G		Best	MA
VVV	4168.84	6	0.6844	0.9061	0.9061
VVI	4170.39	6	0.3156		

CSPA, HGPA, and MCLA (cf. Section 1). These three algorithms transform the given multiple clusterings into a suitable graph representation. Note that MCLA uses hyperedge collapsing operations to determine soft cluster membership values for each object and is a more efficient algorithm than CSPA or HGPA. MATLAB (MATLAB, 2011) code for these three algorithms is available at `http://strehl.com`. To facilitate comparison of Strehl's algorithms to our methodology, the partitions corresponding to the models in Occam's

Figure 3: The best model (VVV, $G = 6$, left-hand side) and one other model from within Occam's window (VVV, $G = 4$, right-hand side) for Scenario III.

window are taken to be the multiple clusterings that need to be consolidated.

We compare our methods to those using Strehl's consensus clustering algorithms for each of the five real data sets already considered that had multiple models inside Occam's window (Table 11). For the coffee data, there is a three-way tie between averaging *a posteriori* probabilities, model averaging, and MCLA. In the other four cases, one of our averaging *a posteriori* probabilities approaches outperforms all other approaches. In the four cases were model averaging was applied, it gave the same classification performance as MCLA in one case and outperformed all consensus clustering approaches in the other three cases. Looking at the three consensus clustering algorithms, we note that MCLA always outperforms both CSPA and HGPA. The very poor performance of both CSPA and HGPA on the coffee data is also noteworthy, especially when one considers that all other methods give perfect classification results.

Table 11: A summary of ARI values for our methodology and Strehl's algorithms on five real data sets. The highest ARI for each data set is highlighted.

	AAP		MA	CSPA	HGPA	MCLA
	Case I	Case II				
Wine	**0.8962**		0.8456	0.7590	0.1522	0.8625
Female voles		**0.9081**		0.4488	0.3802	0.7060
Bank note	0.6790	**0.9068**	0.7602	0.4868	0.0315	0.6790
Iris	**0.9222**	0.5681	0.9039	0.8508	0.2889	0.9039
Coffee	**1**		**1**	0.1100	0.0792	**1**

5   Discussion

This paper proposes a departure from the 'single best model' paradigm that has heretofore dominated the model-based clustering literature. Two averaging approaches are used and both focus on averaging models within Occam's window: one averages *a posteriori* probabilities and the other averages the models' parameters to give a single interpretable model. Averaging *a posteriori* probabilities may require some

14

components to be merged, and we introduced an ARI-based method to do this. This method requires a reference model, which is either taken to be the best model within Occam's window or the model within Occam's window with the fewest components.

While we present our approaches as either averaging *a posteriori* probabilities or averaging models, one could also think of them as either using all models in Occam's window or using just the models with the same number of components as the best model. When using model averaging, we use the models with the same number of components as the best model. When averaging *a posteriori* probabilities in Case I, we use the models with at least as many components as the best model. When averaging *a posteriori* probabilities in Case II, we use all models to carry out the averaging. The performance of averaging *a posteriori* probabilities will depend on the efficacy of the component merging approach used — the method proposed herein (Section 3.1) clearly worked very well on the real and simulated data considered in Section 4. The performance also depends on how good the choice of the number of components is; in fact, this is a central issue with all three of our approaches. Instead of choosing the number of components corresponding to the best model or choosing the fewest components, other approaches could be used to select the number of components.

We used three simulation scenarios and seven real data sets to illustrate our approaches. On the seven real data sets, the performance of our averaging techniques was similar to, or slightly better than, that of the best model. In all but one case, averaging *a posteriori* probabilities in Case II gave comparable or better performance to the best model and the other averaging approaches. The exception arose in the analysis of the iris data, where the number of components in the reference model was smaller than the true number of classes. Our approaches performed favourably when compared to the three consensus clustering algorithms introduced by Strehl et al. (2002). Averaging *a posteriori* probabilities in Case II has the advantage of using all models within Occam's window and perhaps has the greatest potential to notably improve classification performance; however, there is the risk that a model with too few components will result. Using model averaging or averaging *a posteriori* probabilities in Case I avoids the risk of underestimating the number of components relative to the best model; however, the potential for improvement in classification performance may not be as great.

One may argue that averaging is not needed, and that one should just report the set of competing models within Occam's window. Of course, there may be particular applications where this is preferable; however, end users generally want a single clustering result and our averaging approaches deliver this — in fact, similar to model-based clustering in general, they deliver this in both hard and soft varieties. In terms of speed, it is noteworthy that the merging and averaging steps we use, i.e., the steps that we carry out above and beyond the usual model-based clustering procedure, are very fast and can be carried out for very large values of n, G, and p (in fact, the value of latter is irrelevant in this respect). This means that our averaging approaches will not be a limiting factor in performing model-based clustering for data with larger values of n, G, and/or p, i.e., our averaging approaches can be applied to any model-based clustering technique that itself works well for the data at hand. In such cases, it might be worth considering the LASSO-penalized BIC (Bhattacharya and McNicholas, 2014) as an alternative to the BIC, and this will be explored in future work.

The approaches introduced herein are perhaps the beginning of a corpus of work on averaging for families of mixture models in model-based clustering. We used the GPCM family of models to illustrate our approaches, but they could equally well be applied to non-Gaussian mixtures and to other families. We opted for $c = 20$ in the definition of Occam's window (5), and the use of other values will be explored in future work. Alternatives to the BMA framework we consider could also be explored, leading to different weights and/or a different window. We focused on clustering applications herein, but our merging methods could also be applied, in an analogous fashion, for model-based classification (e.g., Dean et al., 2006; McNicholas, 2010; Andrews et al., 2011) and model-based discriminant analysis (Hastie and Tibshirani, 1996); this will also be a focus of future work.

Acknowledgements

References

Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society 59*, 2–5.

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

Andrews, J. L., P. D. McNicholas, and S. Subedi (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis 55*(1), 520–529.

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

Basford, K. E. and G. J. McLachlan (1985). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association 80*(390), 286–293.

Baudry, J.-P., A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics 19*(2), 332–353.

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

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725.

Bouveyron, C., S. Girard, and C. Schmid (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis 52*(1), 502–519.

Browne, R. P. and P. D. McNicholas (2013). *mixture: Mixture Models for Clustering and Classification.* R package version 1.0.

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

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

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In K.-A. Do, P. Müller, and M. Vannucci (Eds.), *Bayesian Inference for Gene Expression and Proteomics.* Cambridge University Press.

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

Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society: Series C 55*(1), 1–14.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 39*(1), 1–38.

Faraway, J. (2011). *faraway: Functions and datasets for books by Julian Faraway.* R package version 1.0.5.

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

Flury, B. (1997). *A First Course in Multivariate Statistics.* New York: Springer.

Flury, B. (2012). *Flury: Data Sets from Flury, 1997.* R package version 0.1-3.

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

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C., A. E. Raftery, and L. Scrucca (2013). *mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.* R package version 4.2.

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

Fred, A. L. N. and A. K. Jain (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transaction on Pattern Analysis and Machine Intelligence 27*, 835–850.

Hastie, T. and R. Tibshirani (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B 58*, 155–176.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification 4*, 3–34.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 879–899.

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

Hoeting, J. A., A. E. Raftery, and D. Madigan (1999). Bayesian simultaneous variable and transformation selection in linear regression. Technical Report 9905, Department of Statistics, Colorado State University.

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

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

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

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

Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks 31*(3), 204–213.

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

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

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research 43*, 570–577.

MATLAB (2011). *version 7.12.0.635 (R2011a)*. Natick, Massachusetts: The MathWorks Inc.

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 R. P. Browne (2013). Discussion of 'How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification'. *Journal of the Royal Statistical Society: Series C 62*(3), 352–353.

McNicholas, P. D., K. R. Jampani, A. F. McDaid, T. B. Murphy, and L. Banks (2014). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.1.

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*(21), 2705–2712.

Milligan, G. W. and M. C. Cooper (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*(4), 441–458.

Molitor, J., M. Papathomas, M. Jerrett, and S. Richardson (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics 11*(3), 484–498.

Murray, P. M., R. B. Browne, and P. D. McNicholas (2014). Mixtures of skew-t factor analyzers. *Computational Statistics and Data Analysis 77*, 326–335.

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

Qiu, W. and H. Joe (2012). *clusterGeneration: random cluster generation (with specified degree of separation)*. R package version 1.2.9.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika 83*(2), 251–266.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1998). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Raftery, A. E., D. Madigan, and C. T. Volinsky (1995). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 323–349. Oxford University Press.

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

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

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods 9*, 386–396.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B 62*, 795–809.

Strehl, A., J. Ghosh, and C. Cardie (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research 3*, 583–617.

Volinsky, C. T., D. Madigan, A. E. Raftery, and R. A. Kronmal (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C 46*(4), 433–448.

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

Wehrens, R., L. M. Buydens, C. Fraley, and A. E. Raftery (2004). Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification 21*, 231–253.

Wolfe, J. H. (1963). Object cluster analysis of social areas. Master's thesis, University of California, Berkeley.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics 17*(10), 977–987.