ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Model-Based Clustering

Isobel Claire Gormley,[1] Thomas Brendan Murphy,[1,2] and Adrian E. Raftery[3]

[1]School of Mathematics and Statistics, University College Dublin, Dublin, Ireland; email: claire.gormley@ucd.ie, brendan.murphy@ucd.ie

[2]Collegium de Lyon Institut d'Études Avancées and ERIC, Université de Lyon, Lyon, France

[3]Department of Statistics and Department of Sociology, University of Washington, Seattle, Washington; email: raftery@uw.edu

Annu. Rev. Stat. Appl. 2023. 10:573–95

First published as a Review in Advance on October 21, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-033121-115326

Keywords

clustering, mixture models, expectation–maximization algorithm, Bayesian inference, software

Abstract

Clustering is the task of automatically gathering observations into homogeneous groups, where the number of groups is unknown. Through its basis in a statistical modeling framework, model-based clustering provides a principled and reproducible approach to clustering. In contrast to heuristic approaches, model-based clustering allows for robust approaches to parameter estimation and objective inference on the number of clusters, while providing a clustering solution that accounts for uncertainty in cluster membership. The aim of this article is to provide a review of the theory underpinning model-based clustering, to outline associated inferential approaches, and to highlight recent methodological developments that facilitate the use of model-based clustering for a broad array of data types. Since its emergence six decades ago, the literature on model-based clustering has grown rapidly, and as such, this review provides only a selection of the bibliography in this dynamic and impactful field.

1. INTRODUCTION

Clustering methods address the challenge of collecting observations into a number of homogeneous groups based on quantitative measurements on the observations. While humans seemingly perform such a task effortlessly when the number of observations, measurements, and groups is small, clustering is a difficult problem: The number of clusters is unknown, as is the clustering solution, i.e., the cluster membership of each observation. Many approaches to inferring a clustering solution exist and, in general, can themselves be grouped into heuristic and model-based approaches. Heuristic, algorithmic approaches often rely on dissimilarity measures between observations and clusters to provide a clustering solution. Such approaches are typically intuitive and computationally efficient but require many subjective decisions (e.g., which dissimilarity measure to use, how dissimilar two clusters are, and how many clusters are present), meaning reproducibility and robustness are often lacking. Model-based approaches infer the clustering solution by employing a statistical modeling framework and by applying standard statistical inferential methods, allowing for objective and robust inference.

1.1. A Brief History

Grouping objects and individuals with similar characteristics is inherent to language. Perhaps the first to formalize it was Plato, with his theory of forms, defining a form as an abstract idea, of which there may be many instances in practice. Aristotle, in his *History of Animals*, classified animals into groups based on their characteristics, for the first time drawing heavily on empirical observations. More systematically, Linnaeus (1753) classified biological species based on measured empirical characteristics.

Cluster analysis is something more: the search for groups in quantitative data using systematic numerical methods, first proposed by Czekanowski (1909). Cluster analysis developed rapidly in the 1950s, driven by the needs of biological taxonomy and marketing, yielding heuristic algorithmic methods such as single link hierarchical agglomerative clustering (Sneath 1957), and the average link and complete link methods (Sokal & Michener 1958). These methods were heuristic, and developed largely in isolation from the statistical methods based on probability models that were emerging rapidly at the same time. The heuristic methods worked well overall but still did not answer questions that arose frequently, such as which method should be used in a particular setting, how many clusters there are, and how outliers should be handled.

Answers to questions like these were to come from expressing cluster analysis in terms of a probability model, leading to model-based clustering. This was first done for multivariate discrete data, in the form of the latent class model, in which several discrete characteristics are measured on each object, the objects are assumed to be grouped, and within each group the characteristics are statistically independent (Lazarsfeld 1950a,b).

The dominant model for clustering continuous-valued data is the mixture of multivariate normal distributions, proposed for this purpose by Wolfe (1965, 1967, 1970) and independently by Day (1969). Binder (1978) developed a Bayesian estimation method for the model, while Murtagh & Raftery (1984) developed a model-based clustering method based on the eigenvalue decomposition of the component covariance matrices. This was built on by Banfield & Raftery (1989, 1993), who also coined the term "model-based clustering." Celeux & Govaert (1995) introduced maximum likelihood estimation for this family of models using the expectation–maximization (EM) algorithm (Dempster et al. 1977).

Articles reviewing the literature on model-based clustering include those of Fraley & Raftery (1998, 2002), Ahlquist & Breunig (2012), and McNicholas (2016b). More extensive reviews in

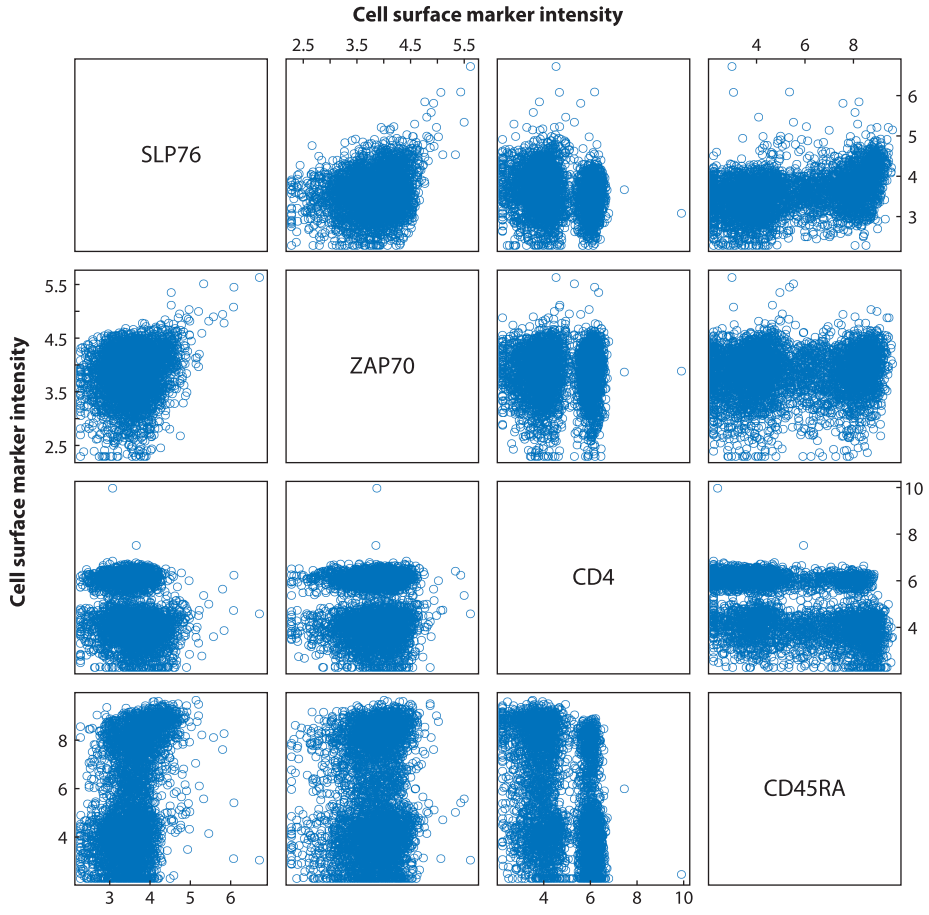

Figure 1

Pairwise scatterplots of the flow cytometry data used to illustrate model-based clustering. The fluorescence intensity at each of four cell surface markers, SLP76, ZAP70, CD4, and CD45RA, is illustrated. Figure adapted with permission from O'Hagan et al. (2016).

book form include those of McLachlan & Basford (1988), McLachlan & Peel (2000), McNicholas (2016a), and Bouveyron et al. (2019).

1.2. A Motivating Example

To motivate and illustrate model-based clustering, we draw on a problem from the field of flow cytometry that allows for rapid single cell analysis by measuring fluorescence intensity across a number of cell surface markers (Maier et al. 2007). **Figure 1** illustrates 4,669 observations with four cell surface markers (SLP76, ZAP70, CD4, and CD45RA) from the T cell phosphorylation data outlined by Pyne et al. (2009). The challenge here is to identify discrete cell populations using the four cell surface markers only, where the number of cell populations (i.e., clusters) is unknown, as is the clustering solution. **Figure 1** illuminates many of the difficulties that clustering poses, even in this relatively low-dimensional example: The number of clusters present is unclear, and different cell markers are suggestive of different numbers of clusters. There are different degrees

of separation between some clusters, and the distribution of data within clusters varies. A model-based approach to clustering these data elegantly addresses each of these challenges in a unified, principled framework that allows for reproducibility and uncertainty quantification.

The clustering solution resulting from the application of model-based clustering to these flow cytometry data is given in Section 2, which reviews the statistical framework underpinning model-based clustering. Section 3 discusses maximum likelihood and Bayesian inferential approaches, and model-selection tools used to identify the number of clusters present and the form of the selected model. Section 4 reviews some recent methodological advances in model-based clustering across a range of data modes, highlighting the preeminent software available to facilitate practical implementation. The article concludes in Section 5 with a discussion, focusing on pending issues in the field of model-based clustering.

2. THE MODEL-BASED CLUSTERING FRAMEWORK

2.1. A Finite Mixture Model

Model-based clustering assumes a probability distribution for the data, typically a finite mixture of G multivariate distributions. Suppose that for each of n observations we have data on d variables, denoted $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})$ for observation i. Thus, the probability model is a weighted average of G probability density functions, i.e.,

$$p(\mathbf{y}_i) = \sum_{g=1}^{G} \tau_g f_g(\mathbf{y}_i|\boldsymbol{\theta}_g), \qquad 1.$$

where the gth mixing proportion τ_g denotes the probability that observation i's data were generated by the gth density, where $\tau_g \geq 0$ for $g = 1, 2, \ldots, G$ and $\sum_{g=1}^{G} \tau_g = 1$. The density of the gth mixture component is $f_g(\cdot|\boldsymbol{\theta}_g)$, where its parameters are collected in $\boldsymbol{\theta}_g$. The "model-based clustering" terminology is often synonymous with the assumption that $f_g(\cdot|\boldsymbol{\theta}_g)$ is a multivariate Gaussian distribution, meaning that Equation 1 is then a Gaussian mixture model. In such a setting, $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$, where $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ denote the mean and covariance of the gth component, respectively (Banfield & Raftery 1993, Celeux & Govaert 1995). Typically each mixture component is taken to correspond to a different cluster, but in cases where a cluster may be better represented by a mixture of Gaussian distributions rather than by a single Gaussian distribution, components can be combined to provide more substantively refined clustering solutions (Baudry et al. 2010, Hennig 2010).

For data with large d, employing a full covariance matrix for each mixture component requires estimation of a large number of parameters, which can lead to a lack of precision, generalizability, and interpretability. Thus it is common to employ more parsimonious component covariance matrices by considering their geometric interpretation. Under the Gaussian density assumption, clusters are ellipsoidal, with their volume, shape, and orientation determined by the component covariance matrices. The widely used software for the implementation of model-based clustering, `mclust` (Scrucca et al. 2016, 2022), considers an eigen-decomposition of the covariance matrix $\boldsymbol{\Sigma}_g$ of the form

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^{\top},$$

where $\boldsymbol{D}_g$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, $\boldsymbol{A}_g$ is a diagonal matrix with elements that are proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$ in descending order, and λ_g is the associated proportionality constant. Geometrically, λ_g controls the volume of the ellipsoid, $\boldsymbol{A}_g$ specifies the shape of the density contours, and $\boldsymbol{D}_g$ determines the orientation of the corresponding ellipsoid (Banfield & Raftery 1993, Celeux & Govaert 1995). The volume, shape, and orientation of the cluster densities can be constrained to be equal (E) or variable (V) across clusters.

Table 1 Parameterizations of the covariance matrix Σ_g available in `mclust` and the associated number of covariance parameters ν when $G = 9$ for $d = 4$ and $d = 40$

Model name	Description	ν	
		$d = 4$	$d = 40$
EII	Spherical, equal volume	1	1
VII	Spherical, varying volume	9	9
EEI	Diagonal, equal volume and shape	4	40
VEI	Diagonal, equal shape	12	48
EVI	Diagonal, equal volume, varying shape	28	352
VVI	Diagonal, varying volume and shape	36	360
EEE	Ellipsoidal, equal volume, shape and orientation	10	820
VEE	Ellipsoidal, equal shape and orientation	18	828
EVE	Ellipsoidal, equal volume and orientation	34	1132
VVE	Ellipsoidal, equal orientation	42	1140
EEV	Ellipsoidal, equal volume and shape	58	7060
VEV	Ellipsoidal, equal shape	66	7068
EVV	Ellipsoidal, equal volume	82	7372
VVV	Ellipsoidal, varying volume, shape, and orientation	90	7380

The three letters in the model name denote, in order, the characteristics of the volume, shape, and orientation across clusters. Abbreviations: E, equal; I, spherical; V, varying.

Table 1 details the constraints on the volume, shape, and orientation of clusters, and associated model names, for all models currently available in `mclust`. Each model is denoted by a three-letter code, the volume-shape-orientation representation: The first letter denotes whether the volume is constrained to be equal (E) or varies (V) across clusters; the second letter denotes whether the shape is constrained to be equal (E) or varies (V) across clusters or if the clusters are spherical (I); and the final letter, which refers to the clusters' orientation, is subject to a similar interpretation. To demonstrate the parsimony gained by such a representation of $\boldsymbol{\Sigma}_g$, the number of covariance parameters ν in each model is detailed for $d = 4$, as in the flow cytometry data in **Figure 1**, under each `mclust` model with $G = 9$. For additional context, the number of covariance parameters in each model for a higher-dimensional setting where $d = 40$ and $G = 9$ is also detailed.

Figure 2 illustrates the clustering solution resulting from the application of model-based clustering (via `mclust`) to the flow cytometry data. A finite mixture of $G = 9$ ellipsoidal multivariate Gaussian distributions with equal shape (i.e., the VEV `mclust` model) was deemed optimal by the Bayesian information criterion (BIC) (see Section 3.3). However, combining these components hierarchically according to an entropy criterion (Baudry et al. 2010) strongly indicates the six-cluster solution illustrated in **Figure 2**.

2.2. Beyond the Multivariate Gaussian Distribution

Model-based clustering's use of a probability model for the data easily allows for clustering of different data modes by employing an apposite density (or mass function) $f_g(\cdot|\boldsymbol{\theta}_g)$ for the data under study. For example, for multivariate continuous data with heavier tails than allowed for by the Gaussian, a mixture model in which $f_g(\cdot|\boldsymbol{\theta}_g)$ is a multivariate t-distribution may be employed for clustering. For the nonelliptically shaped flow cytometry data in **Figure 1**, the FLAME (flow analysis with automated multivariate estimation) method of Pyne et al. (2009) suggested a

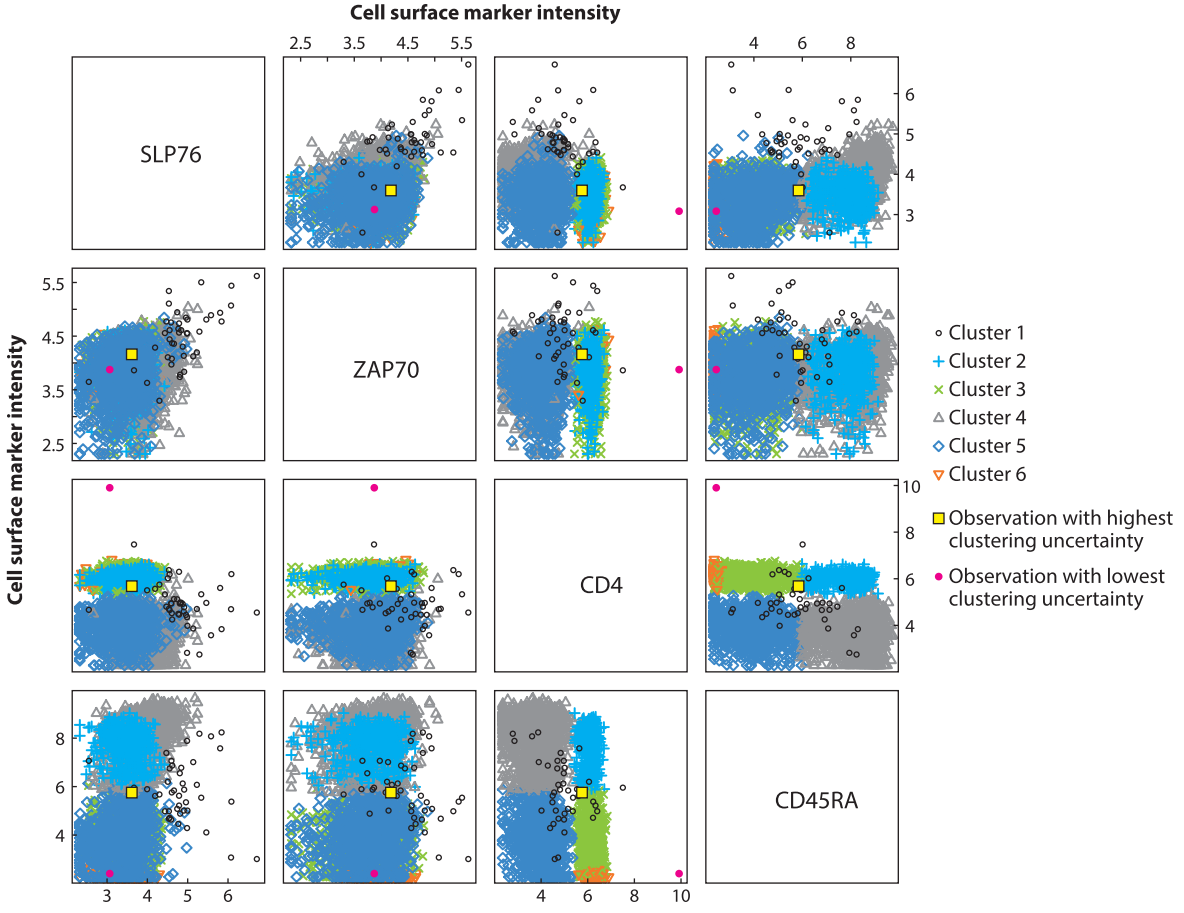

Figure 2

The six-cluster solution resulting from the application of model-based clustering to the flow cytometry data, with subsequent cluster merging. Observations within a cluster are represented by points with the same color and shape. The observation with the lowest clustering uncertainty (see Section 3.1) is illustrated by the magenta circle; it is clustered with observations in cluster 1, illustrated by black circles. The observation with the highest clustering uncertainty is the yellow square; it is clustered with observations in cluster 2, illustrated by light blue crosses. SLP76, ZAP70, CD4, and CD45RA are four cell surface markers at which the fluorescence intensity is measured.

mixture of five skew-*t* distributions as optimal, while O'Hagan et al. (2016) proffer a mixture of six multivariate normal inverse Gaussian distributions. McLachlan et al. (2019) provide a review of finite mixture models with nonnormal but continuous components, including robust mixtures of Gaussian distributions such as Banfield & Raftery's (1993) mixture of Gaussian distributions with a noise component and the trimming approaches of García-Escudero et al. (2008, 2018).

The finite mixture structure can additionally accommodate different forms of data: high-dimensional data (i.e., $n \ll p$) can be clustered using a mixture of latent factor models (McNicholas & Murphy 2008, Montanari & Viroli 2010, Viroli & McLachlan 2019), and mixtures of latent space models permit clustering of nodes in a network (Handcock et al. 2007, Gormley & Murphy 2010). Model-based clustering methods for clustering functional data (Bouveyron & Jacques 2011, Jacques & Preda 2014) or text data (Viroli & Anderlucci 2021) provide further examples. This adaptability of the model-based clustering framework to appositely model a variety

of data types and forms ensures its widespread applicability; Section 4 provides a deeper overview of model-based clustering beyond the Gaussian setting.

2.3. Beyond the Finite Mixture Model

While the finite mixture model is the prevalent basis for model-based approaches to clustering, many modifications or extensions have been developed. Within the Bayesian paradigm, inference for the finite mixture model described in Section 2.1 is straightforward (see Section 3.2) but extensions including sparse finite and infinite mixture models are naturally accommodated. Careful specification of the hyperparameters of a Dirichlet prior on a finite mixture model's mixing proportions gives rise to the sparse finite mixture model (Rousseau & Mengersen 2011, Van Havre et al. 2015, Malsiner-Walli et al. 2016) while the use of a Dirichlet process prior, for example, provides an infinite mixture model (Papastamoulis 2018, Hejblum et al. 2019, Murphy et al. 2020). Frühwirth-Schnatter & Malsiner-Walli (2019) discuss links between sparse finite and Dirichlet process mixtures.

This array of mixture models typically assume the observations are independent, but in certain settings such an assumption may be invalid. Modifying the basic finite mixture model to account for longitudinal measures or panel-type data (Frühwirth-Schnatter 2006) is feasible given its probabilistic basis. In cases where concomitant information is available, mixture-of-experts models can be employed; in these models the mixing proportions and/or the component density parameters are modeled as a function of the concomitant variables (Jordan & Jacobs 1994, Gormley & Frühwirth-Schnatter 2019). Again, Section 4 provides a deeper overview of model-based clustering beyond the basic finite mixture model setting.

3. STATISTICAL INFERENCE FOR MODEL-BASED CLUSTERING

While early approaches to inferring the parameters of the mixture model used the method of moments (Pearson 1894), modern approaches, particularly in the context of model-based clustering, predominantly rely on maximum likelihood estimation or Bayesian inference. The clustering problem can be viewed as an incomplete data problem or a latent variable problem, where the cluster membership of each observation is unobserved or latent. This view lends the clustering problem well to inference via the well-known EM algorithm (McLachlan & Krishnan 2008) in a maximum likelihood setting, or to Bayesian inference (Bensmail et al. 1997, Frühwirth-Schnatter 2006) given its elegant handling of latent variables.

3.1. Maximum Likelihood Inference

Since the formalization of the EM algorithm by Dempster et al. (1977), the majority of model-based clustering applications use the EM algorithm for inference (McLachlan et al. 2019). In this setting the data are viewed as $(\mathbf{y}_i, \mathbf{z}_i)$ for $i = 1, \ldots, n$, where $\mathbf{y}_i$ denotes the observed data on d variables as before, and $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,G})$ is the unobserved portion of the data. Specifically, we define

$$z_{i,g} = \begin{cases} 1 & \text{if } i \text{ belongs to cluster } g \\ 0 & \text{otherwise.} \end{cases}$$

Thus, each of $\mathbf{z}_1, \ldots \mathbf{z}_n$, assumed to be independent and identically distributed, has a multinomial distribution with G categories, with probabilities $\tau_1, \ldots, \tau_G$, respectively. Rather than directly maximizing the observed data likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g f_g(\mathbf{y}_i|\boldsymbol{\theta}_g),$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_G)$, the EM algorithm works with the complete data likelihood function

$$\mathcal{L}_C(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{z}|\mathbf{y}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\tau_g f_g(\mathbf{y}_i|\boldsymbol{\theta}_g)\right]^{z_{i,g}}, \qquad 2.$$

where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$.

The EM algorithm is an iterative algorithm where each iteration consists of an expectation step (E-step) and a maximization step (M-step). In the E-step, the conditional expectation of the complete data log-likelihood function, given the observed data and the current parameter estimates, is computed. In the M-step, the expected complete data log-likelihood function from the E-step is maximized with respect to the model parameters. Iterating these E- and M-steps until convergence achieves at least a local maximum of the observed data likelihood function, under mild regularity conditions (Dempster et al. 1977). As with any iterative algorithm, convergence of the EM algorithm can be assessed in several ways, predominantly by tracking the change in log-likelihood and/or parameter estimates between successive iterations, or using Aitken's acceleration-based stopping criterion (McLachlan & Peel 2000, McLachlan & Krishnan 2008). Starting values for the EM algorithm also require careful consideration; O'Hagan et al. (2012) propose some useful initialization strategies.

In practice, for the complete data log-likelihood function (Equation 2), where $f_g(\cdot|\boldsymbol{\theta}_g)$ is the multivariate Gaussian distribution, the E-step involves computing the conditional expectation of $z_{i,g}$ for $i = 1, \ldots, n$ and $g = 1, \ldots, G$ given the current parameter estimates and the data $\mathbf{y}$. Given the expected values $\hat{\mathbf{z}}$, the M-step involves maximizing the expected complete data log-likelihood function with respect to the mixing proportions and mean parameters for which solutions are available in closed form; closed form solutions for the covariance matrices are available for some covariance parameterizations. On convergence, the value of $\hat{z}_{i,g}$, the conditional expectation of $z_{i,g}$, is the estimated conditional probability that observation i belongs to cluster g. Thus, a hard classification of cluster membership for each observation is available through allocating each observation to the cluster g' for which $\{g'|\hat{z}_{i,g'} = \max_g \hat{z}_{i,g}\}$, and the uncertainty in that cluster membership is quantified by $(1 - \max_g \hat{z}_{i,g})$ for observation i (Bensmail et al. 1997).

Figure 3 illustrates the clustering uncertainty for each of the observations in the flow cytometry data set under the BIC-optimal mixture of $G = 9$ Gaussian distributions. The maximum possible uncertainty, i.e., $1 - 1/G$, is highlighted; in general, the clustering uncertainty is low, and particularly so for some observations. In **Figure 2**, the magenta circle illustrates the observation with the lowest clustering uncertainty ($<10^{-14}$) under the optimal `mclust` $G = 9$ VEV solution; this observation is clustered with observations in cluster 1, illustrated by black circles. The yellow square is the observation with the highest clustering uncertainty (of 0.68) under the optimal `mclust` $G = 9$ VEV solution; this observation is clustered with observations in cluster 2, illustrated by light blue crosses. The yellow square lies on the boundary of several clusters, and thus the cluster membership of this observation is highly uncertain. The model-based approach to clustering facilitates the quantification of this uncertainty within a principled probabilistic framework.

3.2. Bayesian Inference

In the Bayesian framework, it is intuitive to view the finite mixture model as a hierarchical latent variable model where the distribution of each observation's data depends on a discrete latent variable that indicates component membership. These latent variables, $\mathbf{s} = (s_1, \ldots, s_n)$, where $s_i \in \{1, \ldots, G\}$, are typically assumed to be independent and have a multinomial distribution with one

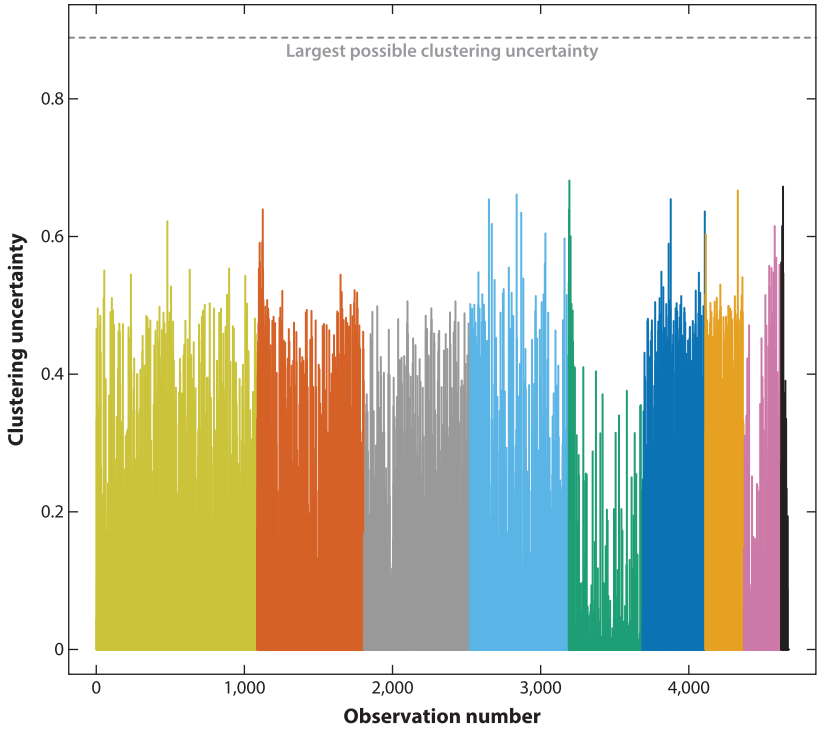

Figure 3

The clustering uncertainty for each of the observations in the flow cytometry data. Each observation is colored according to its cluster membership under the optimal $G = 9$ VEV `mclust` model. VEV indicates that the volume varies across clusters, the shape is constrained to be equal, and the orientation varies.

trial and category probabilities $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_G)$. This gives the complete data likelihood

$$p(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \{p(\mathbf{y}_i|\boldsymbol{\theta}_g)\tau_g\}^{\mathbb{1}_{\{s_i=g\}}}, \qquad 3.$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. In this setting, Bayesian inference on $(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\tau})$ typically proceeds by sampling from the complete data posterior distribution $p(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{y})$. This posterior distribution is available by combining the complete data likelihood (Equation 3) with prior distributions on $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ through Bayes' theorem, i.e.,

$$p(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\theta})p(\boldsymbol{s}|\boldsymbol{\tau})p(\boldsymbol{\theta})p(\boldsymbol{\tau}),$$

where independence between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ is assumed a priori. Through this data augmentation approach (Tanner & Wong 1987), it is straightforward to sample from the posterior distribution using Markov chain Monte Carlo (MCMC) methods, in particular through the use of Gibbs sampling. The importance of Gibbs sampling for Bayesian estimation of mixture models is well established; early work includes that of Diebolt & Robert (1994), Mengersen & Robert (1996), and Raftery (1996), with Frühwirth-Schnatter (2006) and Frühwirth-Schnatter et al. (2019) providing comprehensive resources.

Specification of the prior distributions on the component parameters depends on the data mode. For a finite mixture of Gaussian distributions, Frühwirth-Schnatter & Malsiner-Walli

(2019) recommend putting a normal-gamma prior on component means and a conjugate hierarchical Wishart prior on the component precision matrices. For the mixing proportions, a Dirichlet prior $\mathcal{D}(\boldsymbol{\alpha})$ is typically used, and much work has been devoted to discussion of the choice of $\boldsymbol{\alpha}$. Based on results of Rousseau & Mengersen (2011) and Frühwirth-Schnatter (2011), Malsiner-Walli et al. (2016) encourage small values of $\boldsymbol{\alpha}$ in a sparse finite mixture (i.e., a mixture in which G is most likely larger than the number of clusters) to allow emptying of superfluous components.

The number of components G need not be finite, and the Bayesian framework naturally handles infinite mixture models, such as Dirichlet process mixture models. Using a Dirichlet process prior (Ferguson 1973), Dirichlet process mixtures have been well utilized as model-based clustering tools (e.g., Quintana & Iglesias 2003). Bayesian inference for the Dirichlet process mixture model proceeds via full conditional MCMC sampling (Ishwaran & James 2001). More flexible alternatives, such as the Pitman-Yor process prior (of which the Dirichlet process is a special case), have also been employed in a model-based clustering setting, e.g., by Murphy et al. (2020), where slice sampling (Kalli et al. 2011) is invoked for inference.

Sparse finite mixture models (Malsiner-Walli et al. 2016, Frühwirth-Schnatter & Malsiner-Walli 2019), which are also termed overfitted mixture models in the literature (Rousseau & Mengersen 2011), offer a bridge between finite mixture and Dirichlet process mixture models. Frühwirth-Schnatter & Malsiner-Walli (2019) develop sparse finite mixture models to cluster a broad range of non-Gaussian data; Bayesian inference here is a straightforward extension of inference for a standard finite mixture (Frühwirth-Schnatter 2006), requiring only one additional step. Furthermore, Frühwirth-Schnatter & Malsiner-Walli (2019) unify and compare sparse finite mixtures and Dirichlet process mixtures, highlighting the importance of placing hyperpriors on the Dirichlet parameters and thereby linking the two approaches.

The identifiability of finite mixture distributions (Teicher 1963) requires special attention when performing inference in the Bayesian paradigm. Frühwirth-Schnatter (2006) outlines three types of nonidentifiability: invariance to relabeling the components of the mixture model, nonidentifiability due to potential overfitting, and a generic nonidentifiability that occurs only for certain classes of mixture distributions. To draw valid posterior inference, the invariance to relabeling characteristic requires particular attention. An array of approaches have been developed to address this so-called label-switching problem: Stephens (2000b) proposes a loss function based approach, Celeux et al. (2000) outline a clustering approach, and Murphy et al. (2020) employ a cost-minimizing permutation given by the square assignment algorithm (Carpaneto & Toth 1980). Jasra et al. (2005) provide an encompassing review.

3.3. Model Selection

A central challenge in many clustering problems is inferring the number of clusters G that are present. While the adaptability of model-based clustering to different data modes and forms is advantageous, it can lead to the additional need to choose between competing component densities. The probabilistic model underpinning model-based clustering allows these choices to be jointly addressed using objective and statistically principled model-selection tools.

A typical application of model-based clustering involves fitting a number of models, $\mathcal{M}_1, \ldots, \mathcal{M}_K$, to the data $\boldsymbol{y}$, where each model has a different value of G and a different form of component densities. In `mclust`, for example, models with different values of G and different constraints on the covariance matrix (see **Table 1**) may be fitted. A natural approach to selecting the optimal model is then to consider the model that is most likely a posteriori. Assuming all fitted models are equally probable a priori, this amounts to choosing the model with the highest marginal likelihood. Evaluating the integral that defines the marginal likelihood of model $\mathcal{M}_k$ can

be difficult, but for regular models it can be approximated by the BIC:

$$\text{BIC} \ = 2 \log p(\boldsymbol{y}|\mathcal{M}_k) \approx 2 \log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}) - \nu_{\mathcal{M}_k} \log(n),$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ denote the maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, respectively, and $\nu_{\mathcal{M}_k}$ is the number of independent parameters to be estimated in model $\mathcal{M}_k$. While the necessary regularity conditions do not hold for mixture models in general (Aitkin & Rubin 1985), Keribin (2000) showed that the BIC is consistent for the number of groups, subject to the assumption of a bounded likelihood. While a bounded likelihood is not the case in general for Gaussian mixtures, mild restrictions, such as ensuring a small lower bound on the smallest eigenvalue of the covariance matrices, can ensure this. Moreover, the BIC has been widely used in clustering practice with good results (e.g., Fraley & Raftery 2002, McNicholas & Murphy 2008, McParland et al. 2017, Murphy et al. 2021).

Figure 4 illustrates the BIC values resulting from fitting 210 different `mclust` models to the flow cytometry data. Each model had $G \in \{1, \ldots, 15\}$ and used one of the 14 covariance constrained Gaussian densities detailed in **Table 1**. The optimal model was the $G = 9$ VEV model, which requires estimation of 110 parameters, but there were many closely competing models. The next best model, according to the BIC, was the $G = 12$ VVI model requiring 107 parameters, while the third best was the $G = 11$ VVI model with 103 parameters. This illustrates the model

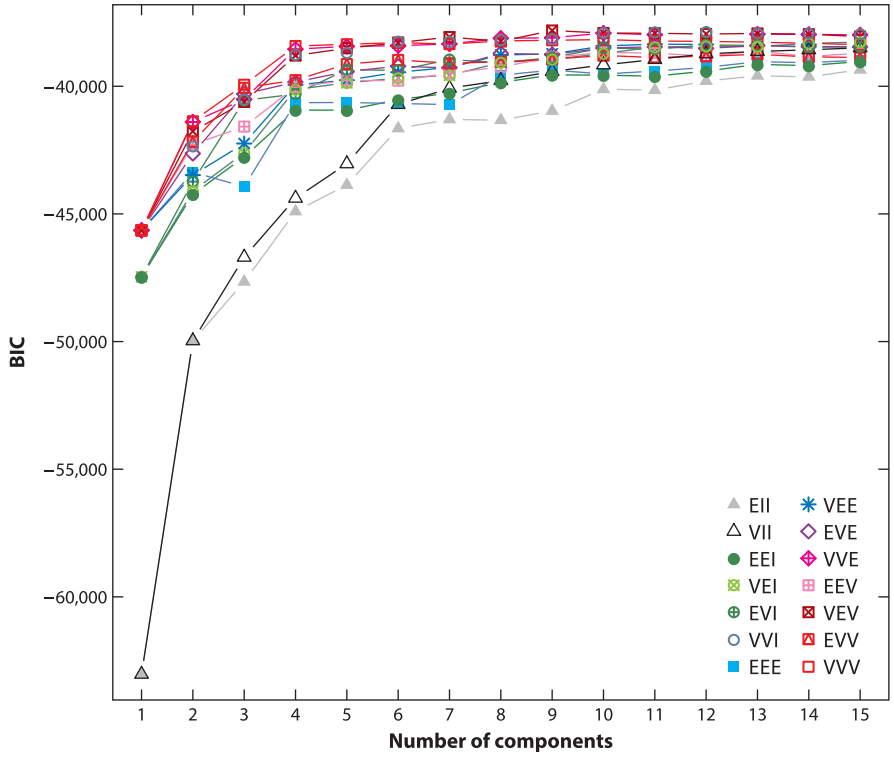

Figure 4

The BIC values for all 14 `mclust` models fitted to the flow cytometry data, with the number of components $G = 1, \ldots, 15$. The optimal model, as deemed by the BIC, is the $G = 9$ VEV model. The three letters in the model name denote, in order, the volume, shape, and orientation across clusters. Abbreviations: BIC, Bayesian information criterion; E, equal; I, spherical; V, varying.

complexity and number of clusters trade-off in model selection in the context of model-based clustering: Where the more parsimonious VVI model is fitted, more clusters are required to provide a good representation of the data. Where the more complex VEV model is fitted, fewer clusters are sufficient.

Many other approaches to model selection have been proposed; Steele & Raftery (2010) provide a review. While the BIC seeks out the number of components in a mixture model, if interest is primarily in the clustering solution, the integrated completed likelihood (Biernacki et al. 2000) provides an alternative to the BIC. The well-known Akaike information criterion (Schwarz 1978) tends to overestimate the number of components in a mixture model; the deviance information criterion (Spiegelhalter et al. 2002), which is an Akaike information criterion–like likelihood penalization criterion, has shown varied performance.

In the Bayesian framework, inferring the number of clusters is naturally available through the posterior mode, or maximum a posteriori estimate of G. For finite mixtures this can be computed by reversible jump MCMC (Richardson & Green 1997) or by birth-death processes (Stephens 2000a), among other methods. For sparse finite and infinite mixture models, inferring the number of clusters is somewhat automatic, for example, by examining the posterior modal number of nonempty clusters. Frühwirth-Schnatter (2006) covers many approaches to inferring the number of components in Bayesian mixture models.

Model-selection criteria have also been used in the Bayesian clustering context. For example, for finite mixtures, Murphy et al. (2020) use the BIC-MCMC of Frühwirth-Schnatter (2011). Raftery et al. (2007) propose the BIC–Monte Carlo (BICM), where $\text{BICM} = 2\log\tilde{\mathcal{L}} - 2s_l^2\log(n)$. Here $\tilde{\mathcal{L}}$ and s_l^2 denote, respectively, the largest value and the sample variance of the log-likelihoods in the posterior sample. Murphy et al. (2020) employ the BICM for a finite mixture of infinite latent factor models; the BICM is particularly useful in the context of such nonparametric models where the number of free parameters can be difficult to quantify.

Including all available variables in a cluster analysis can result in a poor clustering solution if some of the variables cloud the clustering structure. In the model-based framework, selecting the informative clustering variables is naturally framed as a model selection problem; Section 4.9 outlines some approaches to address this issue.

4. MODEL-BASED CLUSTERING IN PRACTICE

The model-based clustering approach has been developed to account for a wide range of data structures beyond the finite Gaussian mixture model described in Section 2.1. Thus, the structure of the data being clustered can be appropriately accounted for using the model-based approach. In this section we outline a selection of such extensions.

A wide range of software packages are available for implementing model-based clustering for different data types. The availability of open-source software has been a particular strong point of the model-based clustering research effort, dating back at least as far as Wolfe (1967). In addition to the software highlighted in this section, the sidebar titled Software describes aspects of five model-based clustering software packages that are particularly widely used.

4.1. Continuous Data

The Gaussian mixture model (Section 2.1) is the most common approach for model-based clustering of multivariate continuous data. The model has been used since the origins of model-based clustering (see also Section 1.1).

The `mclust` software package (Banfield & Raftery 1993, Fraley & Raftery 2002, Scrucca et al. 2016) is one of the most commonly used software packages for model-based clustering of continuous data. The model underlying `mclust` assumes that the data arise from a finite Gaussian

SOFTWARE

While a range of software packages is available to implement model-based clustering, here we describe five widely used packages.

- The preeminent software package for model-based clustering is `mclust` (Scrucca et al. 2016, 2022), which is used to cluster continuous data. The software uses Gaussian mixture models with parsimonious covariance structures and facilitates model-based clustering, classification, density estimation, dimension reduction, and visualization.
- The `poLCA` software package (Linzer & Lewis 2011) is a widely used package for clustering categorical data. Based on the latent class analysis (LCA) model, `poLCA` uses the EM algorithm and Newton-Raphson for model fitting. The package includes latent class regression, which is an extension of LCA to include covariates.
- The `clustMD` software package (McParland & Gormley 2017) clusters mixed data. The software can model data consisting of continuous, binary, ordinal and/or nominal variables using a parsimonious mixture of latent Gaussian variable models.
- The `flexmix` software package (Leisch 2004; Grün & Leisch 2007, 2008) implements model-based clustering for a range of component distributions. The package can handle a wide range of data types (continuous, categorical, count, etc.) and allows for clustering with covariates.
- The `mixtools` software package (Benaglia et al. 2009) contains approaches for clustering different data types and for model-based clustering with covariates.

mixture model where the component covariance matrices are constrained to have a parsimonious structure. Celeux & Govaert (1995) developed a similar model-based clustering approach called Gaussian parsimonious clustering models (GPCM); the `Rmixmod` software (Lebret et al. 2015) implements this approach.

Alternative Gaussian model–based clustering methods have been developed by assuming different covariance structures. Examples include `pgmm` (McNicholas & Murphy 2008, McNicholas et al. 2021), which is based on the mixture of factor analyzers model (Ghahramani & Hinton 1996, McLachlan et al. 2003); `longclust` (McNicholas & Murphy 2010, McNicholas et al. 2019), which uses a constrained Cholesky decomposition for the component covariances; and `mixggm` (Fop et al. 2019), which assumes a sparse covariance structure based on Gaussian graphical models.

Many extensions of model-based clustering for continuous data have been developed by relaxing the Gaussian assumption for the component densities. The multivariate t-distribution has been proposed by McLachlan & Peel (1998) and Stephens (2000a) to accommodate heavy tailed data clusters. The `EMMIX` software (McLachlan et al. 1999) was developed for fitting mixtures of multivariate t-distributions for clustering heavy tailed continuous data.

A wide range of more flexible component densities that accommodate heavy tails and skewness have also been proposed. Examples include model-based clustering methods based on the skew-normal and t-distributions (e.g., Lee & McLachlan 2018) and mixtures of generalized hyperbolic distributions (e.g., Morris & McNicholas 2016, Tang et al. 2018).

Lee & McLachlan (2013), McNicholas (2016a), McLachlan et al. (2019), Bouveyron et al. (2019, chapter 9), and Lee & McLachlan (2022) give an overview of many non-Gaussian mixture models proposed for clustering continuous data.

4.2. Categorical Data

Multivariate categorical data arise in a wide range of application domains. The latent class analysis (LCA) model (Lazarsfeld 1950a,b) (see also Section 1.1) is the primary model that is used

for clustering multivariate categorical data. LCA is based on assuming that the data arise from a finite mixture model, where the component distribution assumes that the categorical variables are conditionally independent given the cluster memberships; this is known as the local independence assumption.

Extensions of the latent class model have been proposed to relax the local independence assumption by allowing for dependence within clusters using latent variables. Examples of this approach include mixtures of Rasch models (Rost 1990), multi-level mixture item response models (Vermunt 2007), and mixtures of latent trait analyzers (Cagnone & Viroli 2012, Gollini & Murphy 2014).

The `poLCA` (Linzer & Lewis 2011) software is widely used for LCA, the `BayesLCA` software (White & Murphy 2014) implements Bayesian LCA, and the `lvm4net` software (Gollini 2019) includes the mixture of latent trait analyzers model.

4.3. Mixed Data

Mixed data arise when observations consist of different data types, including continuous, nominal categorical, and ordinal categorical data, for example. These data arise widely in a range of applications, in particular in survey data.

Everitt (1984, chapter 4) proposed a mixture model for mixed data that assumes local independence between the variables, as used in LCA. The location model (Hunt & Jorgensen 1999, 2003) was developed for model-based clustering of mixed data, but where the local independence assumption is relaxed. The model can accommodate arbitrary dependence structures for the categorical variables, and the continuous variables can be dependent on the categorical variables. The model is structured to accommodate data with a small number of categorical variables and a potentially large number of continuous variables.

McParland et al. (2014, 2017) and McParland & Gormley (2016) developed the `clustMD` framework for clustering mixed data. The `clustMD` model assumes that the observed variables (nominal, ordinal, and continuous) are manifestations of underlying continuous latent variables. This approach facilitates clustering a large number of categorical variables compared with the location model. The `clustMD` (McParland & Gormley 2017) software package implements the methodology. A review of many of the approaches for model-based clustering for mixed data is given by McParland & Murphy (2019).

4.4. Count Data

A number of model-based clustering approaches have been developed for clustering multivariate count data that arise in application domains including ecology, epidemiology, bioinformatics, and text analytics. One of the simplest model-based clustering approaches for multivariate count data uses mixtures with a locally independent Poisson assumption, similar to the local independence assumption used in LCA (Section 4.2). A number of approaches have been proposed to cluster multivariate count data while allowing for dependence and overdispersion in the counts. Karlis & Meligkotsidou (2007) and Karlis (2019) developed model-based clustering methods based on the multivariate Poisson distribution. Subedi & Browne (2020) developed a model-based clustering approach for count data using multivariate Poisson-log normal component distributions. Roick et al. (2021) developed a model-based clustering approach based on integer-valued autoregressive (INAR) models. Ng & Murphy (2021) developed a model-based clustering approach based on the Gaussian Cox process to cluster count process data. Karlis (2019) gave an overview of model-based clustering of discrete data, including count-valued data.

4.5. Sequence Data

Sequence data arise when observations consist of strings of categorical values selected from a finite set. Sequence data arise in many social science applications, in particular, where life-course trajectories are collected, as well as in many other application domains, including computer science and biology.

Erosheva et al. (2014) provided a review of a number of finite mixture models for life-course sequence data, including the group-based trajectory and growth mixture models. More recently, Murphy et al. (2021) developed a distance-based mixture model for clustering life-course trajectory data; this approach was implemented in the `MEDseq` software (Murphy et al. 2022).

A number of approaches for model-based clustering of sequences have been developed which are based on Markov chain models (Melnykov 2016b, Zhang et al. 2021). In this approach, the clustering of observations is focused on the transition between states in the sequences rather than the specific trajectories. The `ClickClust` software (Melnykov 2016a) provides an implementation of Markov chain model-based clustering for sequence data.

4.6. Rank Data

Rank data arise when a set of judges are asked to list a set of objects in order of preference. The resulting data are ordered lists of objects. Rank data arise in a number of contexts, including voting, education, and marketing. A number of model-based clustering methods have been developed for ranking data.

Gormley & Murphy (2006, 2008) developed finite mixtures of Plackett–Luce (Plackett 1975) and Benter (Benter 1994) models for clustering rank data; model inference was implemented in a maximum likelihood framework. Mollica & Tardella (2014, 2017, 2021) further developed the finite mixture of Plackett–Luce modeling framework by considering an extension of the Plackett–Luce model and Bayesian methods for inference. Furthermore, the `PLMIX` software (Mollica & Tardella 2020) facilitates fitting the mixture of Plackett–Luce models (and extensions) to rank data.

An alternative approach to clustering rank data was proposed by Murphy & Martin (2003) using a finite mixture of distance-based models, known as Mallows models (Mallows 1957). This modeling approach was extended by Busse et al. (2007), who developed computationally efficient algorithms for inference. More recently, Bayesian mixtures of Mallows models have been proposed for clustering rank data by Vitelli et al. (2017) and Liu et al. (2019). The `BayesMallows` software (Sørensen et al. 2020) facilitates fitting mixtures of Mallows models, with a number of rank distances. Related to the Mallows modeling approach, Biernacki & Jacques (2013) proposed a model-based clustering approach for rank data based on insertion and sort algorithms.

4.7. Network Data

Network data arise when the connections (or relations) between entities or nodes are recorded. Often the connections are binary in nature, but it is also possible to have weighted connections (showing the strength of the connection); these data arise widely in biology and the social sciences. The emphasis when clustering social network data is on clustering the entities into meaningful groups.

A number of models have been proposed for model-based clustering of network data. For example, Snijders & Nowicki (1997) and Nowicki & Snijders (2001) proposed the stochastic blockmodel for clustering network data. The stochastic blockmodel assumes a latent clustering variable

and that connections are formed independently, conditional on the latent clustering variable. The stochastic blockmodel has been extended in a number of directions, including the mixed membership stochastic blockmodel (Airoldi et al. 2008) and the overlapping stochastic blockmodel (Latouche et al. 2011).

The latent position cluster model (LPCM), developed by Handcock et al. (2007) for clustering network data, is based on a latent Gaussian mixture model. The LPCM was extended by Krivitsky et al. (2009) to account for complex structures that arise in network data. The `latentnet` software (Krivitsky & Handcock 2008, 2020) has been developed to cluster network data using the LPCM. Gormley & Murphy (2010) include node covariates within the LPCM, with implementation available through the `MEclustnet` software package (Gormley & Murphy 2019). Wasserman et al. (2007), Salter-Townshend et al. (2012), and Bouveyron et al. (2019, chapter 10) provide reviews of a number of model-based clustering approaches for network data.

4.8. Clustering with Covariates

Jacobs et al. (1991) developed the mixture-of-experts modeling framework, which provides a systematic way to include covariates in the finite mixture model–based clustering approach. The mixture-of-experts framework has been used by a number of authors to extend model-based clustering methods to the situation where covariate information is also available. Examples include Murphy & Murphy (2020), who developed the `MoEClust` family of models that extend the `mclust` modeling framework to include covariates, and Gormley & Murphy (2008), who extended the Plackett–Luce mixture model (Section 4.6) to include covariates. Yuksel et al. (2012), Gormley & Frühwirth-Schnatter (2019), and Bouveyron et al. (2019, chapter 11) provide reviews of the application of mixture-of-experts models.

4.9. Variable Selection

Many applications of model-based clustering involve high-dimensional data, which are difficult to model. To extend the scope of model-based clustering to high-dimensional settings, a number of approaches have been developed for simultaneously clustering data and performing variable selection.

Raftery & Dean (2006) proposed a model-based clustering approach that includes variable selection; the approach is based on the same Gaussian mixture model as `mclust` (Section 2.1). This approach was further developed and extended by Maugis et al. (2009a,b) and Celeux et al. (2011). The `clustvarsel` software package (Scrucca & Raftery 2018) implements this variable selection approach.

Dean & Raftery (2010) also developed a model-based clustering approach with variable selection for categorical data, using the LCA model (Section 4.2). This approach was extended by Fop et al. (2017) to partition data variables into clustering, redundant, and irrelevant variables; both approaches are implemented in the `LCAvarsel` software package (Fop & Murphy 2017). Detailed reviews of variable selection methods and clustering high-dimensional data are given by Celeux et al. (2014), Fop & Murphy (2018) and McParland & Murphy (2019).

5. DISCUSSION

In this article we have reviewed the well-established yet rapidly developing area of model-based clustering. From its beginnings in the 1960s, model-based clustering has contributed much to the field of statistics, not only through its own developments but also through the impact those developments have had on other areas of statistics. As such, this article does not provide a review of

the total existing literature on model-based clustering, nor does it cover all the emerging themes, of which there are many.

Perhaps the most urgent aspect of model-based clustering that requires development is its scalability to be able to handle the volume, velocity, and variety of data currently being generated. While Section 4 highlights literature that has begun to address these aspects (in particular variety), there is significant scope for advancement. For example, DNA methylation data are collected at a scale of circa 1 million variables being recorded per observation, of which there are typically few. While mixtures of latent variable models have potential given their dimension reduction characteristics (e.g., McLachlan et al. 2002), they break down at such scale. Similarly, for data sets with large n, the reliance on the likelihood function often means the model-based clustering approach is intractable. While Antonazzo et al. (2021) proposed a binned technique for scalable model-based clustering on huge data sets, there is much scope for further novel developments here.

Section 3.3 alluded to the trade-off between the use of complex component densities and the number of clusters. The idea of cluster merging (Baudry et al. 2010), as used to arrive at **Figure 2**, suggests one way of alleviating this tension. Recent work has used mixtures of complex component densities to cluster non-Gaussian continuous data (e.g, Murray et al. 2020, Lee & McLachlan 2022), and this area is ripe for further advancements to be made. Similarly, model-based clustering approaches for object-oriented data analysis through the use of complex component densities require attention.

From an applications viewpoint, the utility of a clustering solution may be low if the resulting clusters are not associated with (e.g., clinical) outcomes of interest. Developing a guided model-based clustering approach that can infer clustering structure with specific outcomes in mind would provide an approach that delivers deeper impact in a range of application domains.

From a software viewpoint, Section 4 and the sidebar titled Software highlight the broad array of software available to facilitate the widespread use of model-based clustering approaches. However, predominantly these software packages employ the maximum likelihood approach to model-based clustering, and there is little on offer from the Bayesian paradigm given the associated technical and computational challenges. However, as approaches such as sparse finite mixture models gain further traction, there is scope for the Bayesian approach to become more accessible to the nonstatistical community through the development of associated open-source software.

As unsupervised learning methods gain traction and popularity in a broad range of domains and application areas, the model-based clustering approach remains strongly competitive given its inherent statistical rigor and flexibility. There is significant scope for the development of bespoke, apposite model-based clustering approaches for data types and problems that we have not yet seen.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank the members of the Working Group in Model-Based Clustering for discussions that strongly informed this work. This work was supported by Science Foundation Ireland grants (12/RC/2289_P2, 16/RC/3835), a stay at Collegium de Lyon, and National Institutes of Health (NIH) grant R01 HD-70936 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).

LITERATURE CITED

Ahlquist JS, Breunig C. 2012. Model-based clustering and typologies in the social sciences. *Political Anal.* 20:92–112

Airoldi EM, Blei DM, Fienberg SE, Xing EP. 2008. Mixed-membership stochastic blockmodels. *J. Mach. Learn. Res.* 9(65):1981–2014

Aitkin M, Rubin DB. 1985. Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. Ser. B* 47(1):67–75

Antonazzo F, Biernacki C, Keribin C. 2021. A binned technique for scalable model-based clustering on huge datasets. In *Book of Short Papers of the 5th International Workshop on Models and Learning for Clustering and Classification (MBC2 2020), Catania, Italy*, ed. S Ingrassia, A Punzo, R Rocci, pp. 11–16. Milan: Ledizioni

Banfield JD, Raftery AE. 1989. *Model-based Gaussian and non-Gaussian clustering.* Tech. Rep. 186, Dep. Stat., Univ. Washington, Seattle, WA

Banfield JD, Raftery AE. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3):803–21

Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. 2010. Combining mixture components for clustering. *J. Comput. Graph. Stat.* 19:332–53

Benaglia T, Chauveau D, Hunter DR, Young D. 2009. `mixtools`: An R package for analyzing finite mixture models. *J. Stat. Softw.* 32(6):1–29

Bensmail H, Celeux G, Raftery AE, Robert CP. 1997. Inference in model-based cluster analysis. *Stat. Comput.* 7(1):1–10

Benter W. 1994. Computer based horse race handicapping and wagering systems: a report. In *Efficiency of Racetrack Betting Markets*, ed. WT Ziemba, VS Lo, DB Haush, pp. 183–98. Singapore: World Sci.

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal.* 22(7):719–25

Biernacki C, Jacques J. 2013. A generative model for rank data based on insertion sort algorithm. *Comput. Stat. Data Anal.* 58:162–76

Binder DA. 1978. Bayesian cluster analysis. *Biometrika* 65(1):31–38

Bouveyron C, Celeux G, Murphy TB, Raftery AE. 2019. *Model-Based Clustering and Classification for Data Science: With Applications in R.* Cambridge, UK: Cambridge Univ. Press

Bouveyron C, Jacques J. 2011. Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.* 5(4):281–300

Busse LM, Orbanz P, Buhmann JM. 2007. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, pp. 113–20. New York: ACM

Cagnone S, Viroli C. 2012. A factor mixture analysis model for multivariate binary data. *Stat. Model.* 12(3):257–77

Carpaneto G, Toth P. 1980. Algorithm 548: Solution of the assignment problem [H]. *ACM Trans. Math. Softw.* 6(1):104–11

Celeux G, Govaert G. 1995. Gaussian parsimonious clustering models. *Pattern Recognit.* 28(5):781–93

Celeux G, Hurn M, Robert CP. 2000. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.* 95(451):957–70

Celeux G, Martin-Magniette ML, Maugis C, Raftery AE. 2011. Letter to the editor: "A framework for feature selection in clustering." *J. Am. Stat. Assoc.* 106(493):383

Celeux G, Martin-Magniette ML, Maugis-Rabusseau C, Raftery AE. 2014. Comparing model selection and regularization approaches to variable selection in model-based clustering. *J. Soc. Fr. Stat.* 155(2):57–71

Czekanowski J. 1909. Zur differential-diagnose der Neadertalgruppe. *Korresp. Bl. Dtsch. Ges. Anthropol. Ethnol. Urgesch.* 40:44–47

Day NE. 1969. Estimating the components of a mixture of two normal distributions. *Biometrika* 56(3):463–74

Dean N, Raftery AE. 2010. Latent class analysis variable selection. *Ann. Inst. Stat. Math.* 62(1):11–35

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. With discussion. *J. R. Stat. Soc. Ser. B* 39(1):1–38

Diebolt J, Robert CP. 1994. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. Ser. B* 56(2):363–75

Erosheva EA, Matsueda RL, Telesca D. 2014. Breaking bad: two decades of life-course data analysis in criminology, developmental psychology, and beyond. *Annu. Rev. Stat. Appl.* 1:301–32
Everitt B. 1984. *An Introduction to Latent Variable Models*. London: Chapman and Hall
Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* 1:209–30
Fop M, Murphy TB. 2017. `LCAvarsel`: variable selection for latent class analysis. *R Package*, version 1.1
Fop M, Murphy TB. 2018. Variable selection methods for model-based clustering. *Stat. Surv.* 12:18–65
Fop M, Murphy TB, Scrucca L. 2019. Model-based clustering with sparse covariance matrices. *Stat. Comput.* 29(4):791–819
Fop M, Smart K, Murphy TB. 2017. Variable selection for latent class analysis with application to low back pain diagnosis. *Ann. Appl. Stat.* 11:2085–115
Fraley C, Raftery AE. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* 41(8):578–88
Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.* 97(458):611–31
Frühwirth-Schnatter S. 2006. *Finite Mixture and Markov Switching Models.* New York: Springer
Frühwirth-Schnatter S. 2011. Dealing with label switching under model uncertainty. In *Mixtures: Estimation and Application*, ed. K Mengersen, CP Robert, DM Titterington, pp. 213–39. New York: Wiley
Frühwirth-Schnatter S, Celeux G, Robert CP. 2019. *Handbook of Mixture Analysis*. Boca Raton, FL: Chapman and Hall/CRC
Frühwirth-Schnatter S, Malsiner-Walli G. 2019. From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv. Data Anal. Classif.* 13(1):33–64
García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A. 2018. Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Adv. Data Anal. Classif.* 12(2):203–33
García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A. 2008. A general trimming approach to robust cluster analysis. *Ann. Stat.* 36(3):1324–45
Ghahramani Z, Hinton GE. 1996. *The EM algorithm for mixtures of factor analyzers.* Tech. Rep. CRG-TR-96-1, Dep. Comput. Sci., Univ. Toronto, Toronto, Can.
Gollini I. 2019. `lvm4net`: Latent variable models for networks. *R Package*, version 0.3
Gollini I, Murphy TB. 2014. Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* 24(4):569–88
Gormley IC, Frühwirth-Schnatter S. 2019. Mixture of experts models. In *Handbook of Mixture Analysis*, ed. S Frühwirth-Schnatter, G Celeux, CP Robert, pp. 271–307. Boca Raton, FL: Chapman and Hall/CRC
Gormley IC, Murphy TB. 2006. Analysis of Irish third-level college applications data. *J. R. Stat. Soc. Ser. A* 169(2):361–79
Gormley IC, Murphy TB. 2008. Exploring voting blocs within the Irish electorate: a mixture modeling approach. *J. Am. Stat. Assoc.* 103(483):1014–27
Gormley IC, Murphy TB. 2010. A mixture of experts latent position cluster model for social network data. *Stat. Methodol.* 7(3):385–405
Gormley IC, Murphy TB. 2019. `MEclustnet`: the mixture of experts latent position cluster model for network data. *R Package*, version 1.2.2
Grün B, Leisch F. 2007. Fitting finite mixtures of generalized linear regressions in R. *Comput. Stat. Data Anal.* 51(11):5247–52
Grün B, Leisch F. 2008. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* 28(4):1–35
Handcock MS, Raftery AE, Tantrum JM. 2007. Model-based clustering for social networks. *J. R. Stat. Soc. Ser. A* 170(2):1–22
Hejblum BP, Alkhassim C, Gottardo R, Caron F, Thiébaut R. 2019. Sequential Dirichlet process mixtures of multivariate skew *t*-distributions for model-based clustering of flow cytometry data. *Ann. Appl. Stat.* 13(1):638–60
Hennig C. 2010. Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.* 4(1):3–34
Hunt L, Jorgensen M. 1999. Theory & methods: mixture model clustering using the MULTIMIX program. *Aust. N. Z. J. Stat.* 41(2):154–71

Hunt L, Jorgensen M. 2003. Mixture model clustering for mixed data with missing information. *Comput. Stat. Data Anal.* 41(3–4):429–40

Ishwaran H, James LF. 2001. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* 96(453):161–73

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. 1991. Adaptive mixtures of local experts. *Neural Comput.* 3(1):79–87

Jacques J, Preda C. 2014. Model-based clustering for multivariate functional data. *Comput. Stat. Data Anal.* 71:92–106

Jasra A, Holmes CC, Stephens DA. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat. Sci.* 20(1):50–67

Jordan MI, Jacobs RA. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* 6(2):181–214

Kalli M, Griffin JE, Walker SG. 2011. Slice sampling mixture models. *Stat. Comput.* 21(1):93–105

Karlis D. 2019. Mixture modelling of discrete data. In *Handbook of Mixture Analysis*, ed. S Frühwirth-Schnatter, G Celeux, CP Robert, pp. 193–218. Boca Raton, FL: Chapman and Hall/CRC

Karlis D, Meligkotsidou L. 2007. Finite mixtures of multivariate Poisson distributions with application. *J. Stat. Plan. Inference* 137(6):1942–60

Keribin C. 2000. Consistent estimation of the order of mixture models. *Sankhya A* 62(1):49–66

Krivitsky PN, Handcock MS. 2008. Fitting latent cluster models for networks with latentnet. *J. Stat. Softw.* 24(5):1–23

Krivitsky PN, Handcock MS. 2020. `latentnet`: Latent position and cluster models for statistical networks. *R Package*, version 2.10.5

Krivitsky PN, Handcock MS, Raftery AE, Hoff PD. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* 31(3):204–13

Latouche P, Birmelé E, Ambroise C. 2011. Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.* 5(1):309–36

Lazarsfeld PF. 1950a. The logical and mathematical foundations of latent structure analysis. In *Studies in Social Psychology in World War II*. Vol. IV: *Measurement and Prediction*, ed. SA Stouffer, L Guttman, EA Suchman, PF Lazarsfeld, pp. 362–412. Princeton, NJ: Princeton Univ. Press

Lazarsfeld PF. 1950b. Some latent structures. In *Studies in Social Psychology in World War II*. Vol. IV: *Measurement and Prediction*, ed. SA Stouffer, L Guttman, EA Suchman, PF Lazarsfeld, pp. 413–73. Princeton, NJ: Princeton Univ. Press

Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G. 2015. `Rmixmod`: the R package of the model-based unsupervised, supervised, and semi-supervised classification mixmod library. *J. Stat. Softw.* 67(6):1–29

Lee SX, McLachlan GJ. 2013. Model-based clustering and classification with non-normal mixture distributions. *Stat. Methods Appl.* 22(4):427–54

Lee SX, McLachlan GJ. 2018. `EMMIXcskew`: an R package for the fitting of a mixture of canonical fundamental skew t-distributions. *J. Stat. Softw.* 83(3):1–32

Lee SX, McLachlan GJ. 2022. An overview of skew distributions in model-based clustering. *J. Multivar. Anal.* 188:104853

Leisch F. 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* 11(8):1–18

Linnaeus C. 1753. *Species Plantarum*. Stockholm: Laurentii Salvii. 1st ed.

Linzer DA, Lewis JB. 2011. `poLCA`: An R package for polytomous variable latent class analysis. *J. Stat. Softw.* 42(10):1–29

Liu Q, Crispino M, Scheel I, Vitelli V, Frigessi A. 2019. Model-based learning from preference data. *Annu. Rev. Stat. Appl.* 6:329–54

Maier LM, Anderson DE, De Jager PL, Wicker LS, Hafler DA. 2007. Allelic variant in CTLA4 alters T cell phosphorylation patterns. *PNAS* 104(47):18607–12

Mallows CL. 1957. Non-null ranking models. *Biometrika* 44(1/2):114–30

Malsiner-Walli G, Frühwirth-Schnatter S, Grün B. 2016. Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* 26:303–24

Maugis C, Celeux G, Martin-Magniette ML. 2009a. Variable selection for clustering with Gaussian mixture models. *Biometrics* 65(3):701–9

Maugis C, Celeux G, Martin-Magniette ML. 2009b. Variable selection in model-based clustering: a general variable role modeling. *Comput. Stat. Data Anal.* 53(11):3872–82

McLachlan G, Peel D. 1998. Robust cluster analysis via mixtures of multivariate *t*-distributions. In *Advances in Pattern Recognition*, ed. A Amin, D Dori, P Pudil, H Freeman, pp. 658–66. Berlin: Springer-Verlag

McLachlan G, Peel D. 2000. *Finite Mixture Models.* New York: Wiley-Interscience

McLachlan GJ, Basford KE. 1988. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker

McLachlan GJ, Bean RW, Peel D. 2002. A mixture model–based approach to the clustering of microarray expression data. *Bioinformatics* 18(3):413–22

McLachlan GJ, Krishnan T. 2008. *The EM Algorithm and Extensions*. New York: Wiley-Interscience. 2nd ed.

McLachlan GJ, Lee SX, Rathnayake SI. 2019. Finite mixture models. *Annu. Rev. Stat. Appl.* 6:355–78

McLachlan GJ, Peel D, Basford KE, Adams P. 1999. The EMMIX software for the fitting of mixtures of normal and *t*-components. *J. Stat. Softw.* 4(2):1–14

McLachlan GJ, Peel D, Bean RW. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data Anal.* 41(3–4):379–88

McNicholas PD. 2016a. *Mixture Model–Based Classification*. Boca Raton, FL: Chapman and Hall/CRC

McNicholas PD. 2016b. Model-based clustering. *J. Classif.* 33:331–73

McNicholas PD, ElSherbiny A, McDaid AF, Murphy TB. 2021. `pgmm`: Parsimonious Gaussian mixture models. *R Package*, version 1.2.5

McNicholas PD, Jampani KR, Subedi S. 2019. `longclust`: Model-based clustering and classification for longitudinal data. *R Package*, version 1.2.3

McNicholas PD, Murphy TB. 2008. Parsimonious Gaussian mixture models. *Stat. Comput.* 18(3):285–96

McNicholas PD, Murphy TB. 2010. Model-based clustering of longitudinal data. *Can. J. Stat.* 38(1):153–68

McParland D, Gormley IC. 2016. Model based clustering for mixed data: clustMD. *Adv. Data Anal. Classif.* 10(2):155–69

McParland D, Gormley IC. 2017. `clustMD`: Model based clustering for mixed data. *R Package*, version 1.2.1

McParland D, Gormley IC, McCormick TH, Clark SJ, Kabudula CW, Collinson MA. 2014. Clustering South African households based on their asset status using latent variable models. *Ann. Appl. Stat.* 8(2):747–76

McParland D, Murphy TB. 2019. Mixture modelling of high-dimensional data. In *Handbook of Mixture Analysis*, ed. S Frühwirth-Schnatter, G Celeux, CP Robert, pp. 39–70. Boca Raton, FL: Chapman and Hall/CRC

McParland D, Phillips CM, Brennan L, Roche HM, Gormley IC. 2017. Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. *Stat. Med.* 36(28):4548–69

Melnykov V. 2016a. `ClickClust`: an R package for model-based clustering of categorical sequences. *J. Stat. Softw.* 74(9):1–34

Melnykov V. 2016b. Model-based biclustering of clickstream data. *Comput. Stat. Data Anal.* 93:31–45

Mengersen KL, Robert CP. 1996. Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, ed. JM Bernardo, JO Berger, AP Dawid, AFM Smith, pp. 255–76. Oxford, UK: Oxford Univ. Press

Mollica C, Tardella L. 2014. Epitope profiling via mixture modeling of ranked data. *Stat. Med.* 33(21):3738–58

Mollica C, Tardella L. 2017. Bayesian Plackett-Luce mixture models for partially ranked data. *Psychometrika* 82(2):442–58

Mollica C, Tardella L. 2020. `PLMIX`: an R package for modelling and clustering partially ranked data. *J. Stat. Comput. Simul.* 90(5):925–59

Mollica C, Tardella L. 2021. Bayesian analysis of ranking data with the extended Plackett-Luce model. *Stat. Methods Appl.* 30(1):175–94

Montanari A, Viroli C. 2010. Heteroscedastic factor mixture analysis. *Stat. Model.* 10(4):441–60

Morris K, McNicholas PD. 2016. Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. *Comput. Stat. Data Anal.* 97:133–50

Murphy K, Murphy TB. 2020. Gaussian parsimonious clustering models with covariates and a noise component. *Adv. Data Anal. Classif.* 14(2):293–325

Murphy K, Murphy TB, Piccarreta R, Gormley IC. 2021. Clustering longitudinal life-course sequences using mixtures of exponential-distance models. *J. R. Stat. Soc. Ser. A* 184(4):1414–51

Murphy K, Murphy TB, Piccarreta R, Gormley IC. 2022. `MEDseq`: Mixtures of exponential-distance models with covariates. *R Package*, version 1.3.3

Murphy K, Viroli C, Gormley IC. 2020. Infinite mixtures of infinite factor analysers. *Bayesian Anal.* 15(3):937–63

Murphy TB, Martin D. 2003. Mixtures of distance-based models for ranking data. *Comput. Stat. Data Anal.* 41(3–4):645–55

Murray PM, Browne RP, McNicholas PD. 2020. Mixtures of hidden truncation hyperbolic factor analyzers. *J. Classif.* 37(2):366–79

Murtagh F, Raftery AE. 1984. Fitting straight lines to point patterns. *Pattern Recognit.* 17(5):479–83

Ng TLJ, Murphy TB. 2021. Model-based clustering of count processes. *J. Classif.* 38(2):188–211

Nowicki K, Snijders TAB. 2001. Estimation and prediction of stochastic blockstructures. *J. Am. Stat. Assoc.* 96(455):1077–87

O'Hagan A, Murphy TB, Gormley IC. 2012. Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Comput. Stat. Data Anal.* 56(12):3843–64

O'Hagan A, Murphy TB, Gormley IC, McNicholas PD, Karlis D. 2016. Clustering with the multivariate normal inverse Gaussian distribution. *Comput. Stat. Data Anal.* 93:18–30

Papastamoulis P. 2018. Overfitting Bayesian mixtures of factor analyzers with an unknown number of components. *Comput. Stat. Data Anal.* 124:220–34

Pearson K. 1894. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. A* 185:71–110

Plackett RL. 1975. The analysis of permutations. *J. R. Stat. Soc. Ser. C* 24(2):193–202

Pyne S, Hu X, Wang K, Rossin E, Lin TI, et al. 2009. Automated high-dimensional flow cytometric data analysis. *PNAS* 106(21):8519–24

Quintana FA, Iglesias PL. 2003. Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B* 65(2):557–74

Raftery AE. 1996. Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, ed. WR Gilks, S Richardson, DJ Spiegelhalter, pp. 163–88. London: Chapman and Hall

Raftery AE, Dean N. 2006. Variable selection for model-based clustering. *J. Am. Stat. Assoc.* 101(473):168–78

Raftery AE, Newton M, Satagopan J, Krivitsky P. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics 8*, ed. JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, et al., pp. 1–45. Oxford, UK: Oxford Univ. Press

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. Ser. B* 59(4):731–92

Roick T, Karlis D, McNicholas PD. 2021. Clustering discrete-valued time series. *Adv. Data Anal. Classif.* 15(1):209–29

Rost J. 1990. Rasch models in latent classes: an integration of two approaches to item analysis. *Appl. Psychol. Meas.* 14(3):271–82

Rousseau J, Mengersen K. 2011. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B* 73(5):689–710

Salter-Townshend M, White A, Gollini I, Murphy TB. 2012. Review of statistical network analysis: models, algorithms, and software. *Stat. Anal. Data Min.* 5(4):243–64

Schwarz G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6(2):461–64

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. `mclust 5`: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8(1):289–317

Scrucca L, Fraley C, Murphy TB, Raftery AE. 2022. *Model-Based Clustering, Classification and Density Estimation Using mclust in R*. Boca Raton, FL: Chapman and Hall/CRC

Scrucca L, Raftery AE. 2018. `clustvarsel`: A package implementing variable selection for Gaussian model-based clustering in R. *J. Stat. Softw.* 84(1):1–28

Sneath PHA. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* 17:201–6

Snijders TAB, Nowicki K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classif.* 14(1):75–100

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38:1409–38

Sørensen Ø, Crispino M, Liu Q, Vitelli V. 2020. `BayesMallows`: an R package for the Bayesian Mallows model. *R J.* 12(1):324–42

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64(4):583–639

Steele RJ, Raftery AE. 2010. Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, ed. M-H Chen, P Müller, D Sun, K Ye, DK Dey, pp. 113–30. New York: Springer

Stephens M. 2000a. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Stat.* 28(1):40–74

Stephens M. 2000b. Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B* 62(4):795–809

Subedi S, Browne RP. 2020. A family of parsimonious mixtures of multivariate Poisson-lognormal distributions for clustering multivariate count data. *Stat* 9(1):e310

Tang Y, Browne RP, McNicholas PD. 2018. Flexible clustering of high-dimensional data via mixtures of joint generalized hyperbolic distributions. *Stat* 7(1):e177

Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82(398):528–40

Teicher H. 1963. Identifiability of finite mixtures. *Ann. Math. Stat.* 34(4):1265–69

Van Havre Z, White N, Rousseau J, Mengersen K. 2015. Overfitting Bayesian mixture models with an unknown number of components. *PLOS ONE* 10(7):e0131739

Vermunt J. 2007. Multilevel mixture item response theory models: an application in education testing. In *Proceedings of the 56th Session of the International Statistical Institute*, *Lisbon*, *Portugal*. Voorburg, Neth.: Int. Stat. Inst.

Viroli C, Anderlucci L. 2021. Deep mixtures of unigrams for uncovering topics in textual data. *Stat. Comput.* 31(3):1–10

Viroli C, McLachlan GJ. 2019. Deep Gaussian mixture models. *Stat. Comput.* 29(1):43–51

Vitelli V, Sørensen Ø, Crispino M, Frigessi A, Arjas E. 2017. Probabilistic preference learning with the Mallows rank model. *J. Mach. Learn. Res.* 18:1–49

Wasserman S, Robins G, Steinley D. 2007. Statistical models for networks: a brief review of some recent research. In *Statistical Network Analysis: Models, Issues, and New Directions*, ed. EM Airoldi, DM Blei, SE Fienberg, A Goldenberg, EP Xing, AX Zheng, pp. 45–56. New York: Springer

White A, Murphy TB. 2014. `BayesLCA`: an R package for Bayesian latent class analysis. *J. Stat. Softw.* 61(13):1–28

Wolfe JH. 1965. *A computer program for the maximum-likelihood analysis of types*. USNPRA Tech. Bull. 65-15, US Naval Pers. Res. Act., San Diego, CA

Wolfe JH. 1967. *NORMIX: computational methods for estimating the parameters of multivariate normal mixture distributions of types*. USNPRA Tech. Bull. 68-2, US Naval Pers. Res. Act., San Diego, CA

Wolfe JH. 1970. Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* 5(3):329–50

Yuksel SE, Wilson JN, Gader PD. 2012. Twenty years of mixture of experts. *IEEE Trans. Neural Netw. Learn. Syst.* 23(8):1177–93

Zhang Y, Melnykov V, Zhu X. 2021. Model-based clustering of time-dependent categorical sequences with application to the analysis of major life event patterns. *Stat. Anal. Data Min.* 14(3):230–40