*Wiley Interdisciplinary Reviews: Computational Statistics*

WILEY

WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW OPEN ACCESS

# Challenges and Adaptations of Model-Based Clustering for Flow and Mass Cytometry

Ultán P. Doherty[1] | Rachel M. McLoughlin[2] | Arthur White[1]

[1]School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin, Dublin, Ireland | [2]Host-Pathogen Interactions Group, School of Biochemistry and Immunology, Trinity Biomedical Sciences Institute, Trinity College Dublin, The University of Dublin, Dublin, Ireland

**Correspondence:** Ultán P. Doherty (dohertyu@tcd.ie)

**Received:** 14 March 2024 | **Revised:** 27 January 2025 | **Accepted:** 30 January 2025

**Commissioning Editor:** James E. Gentle | **Review Editor:** David W. Scott

**Funding:** This study was supported by Taighde Éireann—Research Ireland (GOIPG/2021/1374) and the Science Foundation Ireland Investigator Award (15/IA/3041) under grants awarded to Ultán P. Doherty and Rachel M. McLoughlin, respectively.

**Keywords:** cluster analysis | flow cytometry | mass cytometry | mixture models | model-based clustering

## ABSTRACT

Model-based clustering is a statistical approach to cluster analysis, which has been successfully deployed in a number of domains due to its principled framework, clear assumptions, and adaptability. For these reasons, there has been substantial interest in applying model-based clustering methods to flow cytometry and mass cytometry data. The identification of relevant cell populations is a crucial step in the analysis of cytometry data for immunological research. Technological advances have led to a rapid increase in the dimensionality and complexity of cytometry data, prompting significant interest in the use of clustering algorithms in place of traditional manual data analysis techniques for cell population identification. This article highlights how model-based clustering methods, such as mixture models, have been adapted to meet the many interesting and unusual challenges that present themselves to the researcher when analyzing flow and mass cytometry data. These innovations demonstrate that there is considerable potential for further methodological development and collaboration between the cytometry and model-based clustering research communities.

## 1 | Introduction

Flow cytometry and mass cytometry are high-throughput single-cell analysis techniques used in immunological research to indirectly measure the expression level of a range of protein markers for each cell in a tissue sample. In flow and mass cytometry, population identification is the process of attempting to group together events according to their cell type based on the similarity of their protein marker expression levels. This allows the frequency of each cell type and other population-specific statistics to be computed so that the properties of corresponding cell populations from different samples can be compared. The traditional and most popular approach for population identification is sequential manual gating. This process involves choosing a pair of variables, visualizing the data set in a two-dimensional scatterplot of those variables, and manually drawing a boundary around a subset of cells that contains the population of interest. This subset or gate can be refined by choosing another pair of variables, using them to visualize this subset only, and further subsetting the data using another manually drawn boundary. This process is repeated for each population of interest in the data set. The protein markers that are analyzed are often chosen to facilitate manual gating strategies that will isolate known populations of interest. The development of automated methods for population identification has been motivated by a number of issues with sequential manual gating (Saeys et al. 2016). Manually drawing boundaries can lead to problems with bias and irreproducibility as different researchers are likely to draw

**TABLE 1** | This table lists model-based clustering methods that have been designed specifically for flow cytometry and/or mass cytometry.

| Method (citation) | Component distribution | Multiple samples | Additional details |
|---|---|---|---|
| ASPIRE [doi] (Dundar et al. 2014) | Normal | Joint | Dirichlet process directly estimates the number of populations. |
| BayesFlow [doi] (Johnsson et al. 2016) | Normal + merging | Joint | Hierarchical Bayesian multi-sample model. |
| CYBERTRACK [doi, doi] (Minoura et al. 2019, 2021) | Normal | Pooled | Designed for longitudinal data. |
| CytoFA [doi] (Lee 2019) | Skew–normal | None | Factor analysis and trimming improve computational efficiency. |
| FLAME [doi] (Pyne et al. 2009) | Skew-*t* | Meta | PAM meta-clustering of separately clustered samples. |
| flowClust [doi] (Lo et al. 2008) | Box–Cox transformed *t* | None | Transformation attempts to normalize data. |
| flowMerge [doi] (Finak et al. 2009) | Box–Cox *t* + merging | None | flowClust with post hoc cluster merging. |
| HDPGMM [doi] (Cron et al. 2013) | Normal + merging | Joint | Dirichlet process directly estimates the number of populations. |
| HMM-VB [doi] (Lin and Li 2017) | Normal + merging | None | Hidden Markov model on variable blocks motivated by rare clusters. |
| immunoClust [doi] (Sörensen et al. 2015) | *t* + merging | Meta | Divisive model fitting suits rare populations. |
| JCM [doi] (Pyne et al. 2014) | Skew-*t* | Joint | Random effects joint clustering model. |
| LAMBDA [doi] (Abe et al. 2020) | Normal | None | Model can incorporate clinical information as a covariate. |
| NPflow [doi] (Hejblum et al. 2019) | Skew-*t* | Sequential | Dirichlet process. Sequential posterior-as-prior approach. |
| SWIFT [doi] (Naim et al. 2014) | Normal + split + merge | None | Iterative sampling approach suits rare populations. |
| Tailor [doi] (Ionita et al. 2021) | Normal + merging | Pooled | Data binning reduces computation time. |

*Note:* For each method, the probability distribution of its mixture model components is listed, its approach for multiple samples is mentioned, and additional details are provided. A reference and link are also given for each method.

different boundaries (Finak et al. 2016). Two-dimensional plots fail to display the complex interactions between large sets of variables. This process' sequential subsetting may prevent populations that are excluded early in the sequence from being analyzed with respect to later variable pairs. Finally, for data sets with high numbers of variables, the number of variable pairs is too large for experts to examine them all. To avoid these problems, algorithms have been designed to carry out population identification using data clustering methods from statistics and machine learning.

Population identification for cytometry is a complex clustering problem that has a substantial real-world impact through its role in immunological research. Stubbington et al. (2017) identify immunophenotyping through flow cytometry as the most prominent technique for the categorization of immune cells, while Saeys et al. (2016) describe flow cytometry as the "workhorse for high-throughput quantitative analysis of cells".

Among the challenges posed by cytometry data are that some cell populations do not follow a multivariate Normal distribution, that clustering solutions must be consistent across multiple samples within an experiment, that the number of populations is not known a priori, that there are rare populations, and that the large size of the data sets can cause computational issues. Model-based clustering is one of the most prominent statistical approaches to clustering. Its mathematical foundations and uncertainty quantification make it an appealing approach to this problem. Conversely, cytometry data is increasingly used as motivation for novel model-based methodology, particularly in relation to the development of non-normal mixture models. Cytometry data motivated the skew-*t* mixture model introduced by Pyne et al. (2009) and was used for the primary examples in the "Beyond the Multivariate Gaussian Distribution" section of Gormley et al. (2023) and the "Non-Gaussian Model-based Clustering" chapter of Bouveyron et al. (2019). Our review examines the interdisciplinary overlap between these two

research topics. In particular, it provides an overview and discussion of the challenges involved in employing a model-based clustering approach for population identification in flow and mass cytometry. We begin by providing preliminary information on model-based clustering in Section 2, which introduces it in the context of population identification for flow and mass cytometry data. We discuss some of the benefits and drawbacks of model-based clustering in Section 2.1. Then, in Sections 3–7, we put forward a series of challenges associated with clustering cytometry data and discuss the adaptations that model-based clustering algorithms designed for population identification have developed or adopted to address these issues. Table 1 provides a list of model-based clustering methods which have been developed specifically for flow and mass cytometry. Some of the problems mentioned can be viewed as examples of general challenges in clustering, such as those discussed in Melnykov (2013). Challenges in clustering cytometry data that arise from using manual gating as a benchmark are discussed in Section 8.

Model-based clustering is only one of the clustering approaches that has been employed for population identification. The challenges of analyzing flow cytometry and the potential for statistical methods to replace manual gating have been discussed for a number of years (Lock 2003). Leading machine learning algorithms include FlowSOM (van Gassen et al. 2015), which clusters using a self-organizing map (Kohonen 1990), and PhenoGraph (Levine et al. 2015), which performs nonparametric density-based clustering using a form of nearest-neighbors community detection. Several methods based on $k$-means clustering have also been developed, such as flowMeans (Aghaeepour et al. 2011), as well as tree-based methods, such as cytometree (Commenges et al. 2018). A 2013 review by the FlowCAP Consortium applied a wide range of model-based and non-model-based clustering algorithms to a selection of flow cytometry data sets (The FlowCAP Consortium, The DREAM Consortium, et al. 2013). A further FlowCAP review was carried out by Aghaeepour et al. (2016). Weber and Robinson (2016) applied another set of methods, which included some of those featured in FlowCAP-I, to a set of mass cytometry data sets with multiple populations of interest and also two flow cytometry data sets with a single population of interest only. A similar review was carried out by Liu et al. (2019). These three reviews involved quantitatively comparing the performance of clustering algorithms on benchmark cytometry data sets, rather than discussing the methodology of these algorithms. More recently, Lin and Hejblum (2021) reviewed how different Bayesian parametric and nonparametric mixture modeling approaches can be applied to cytometry data. In this review, we highlight the methodological and computational adaptations that have been made across the model-based clustering literature which address the challenges of analyzing flow cytometry data.

### 1.1 | Mass Cytometry Data Example

The following plots display a 32-dimensional mass cytometry data set which we will refer to as Levine32 (Levine et al. 2015). This data set is publicly available via the HDCytoData R package (Weber and Soneson 2019). It consists of 118,888 events. 72,463 of these events have been manually assigned to 14 gates and the remaining 46,425 events have been left unassigned. These gates vary dramatically in size, ranging from 182 events to 18,131 events. Figure 1 shows a pairs-style plot of three variables corresponding to the surface marker expression of CD19, CD8, and CD3. The seven largest gates are individually colored while the seven smallest gates are plotted in gray. Differences in the expression levels of these three markers between the large gates can be seen. For example, Gate 8 (Green) appears to have higher levels of CD3 than Gate 9 (Yellow) and Gate 10 (Light Blue), and higher levels of CD8 than Gate 7 (Orange). However, it does seem that it would be difficult to identify seven distinct gates within the gray points. Gates 7, 8, 9, and 10 all contain more than 10,000 events (18,131, 13,205, 12,320, and 16,111 respectively), whereas the smallest seven gates collectively contain less than 5000 events (4676). It can also be observed that some of the boundaries between the gates are straight lines, which may seem somewhat artificial. These are a result of the manual gating approach used to obtain these gates.

Figure 2 is a scatter plot of a two-dimensional UMAP representation of the Levine32 data set. The same coloring scheme as that of Figure 1 has been used to highlight the seven largest manual gates. Nonlinear dimension reduction methods such as UMAP (McInnes et al. 2020) and t-SNE (van der Maaten and Hinton 2008) can be useful tools for exploring these high-dimensional data sets (Becht et al. 2019). Machine learning algorithms have also been developed that cluster the data based on a lower-dimensional representation of the data from a nonlinear dimension reduction method (Shekhar et al. 2014). However, it can be challenging to tune the parameters for nonlinear dimension reduction algorithms (Kobak and Berens 2019; Wattenberg et al. 2016). The UMAP algorithm is able to distinguish between the large gates, with some overlap between similar pairs of gates. However, as mentioned previously with respect to Figure 1, it is not clear how the gray points would be divided into seven clusters.

## 2 | Model-Based Clustering

Clustering is a form of unsupervised learning that involves grouping together observations within a data set (Hartigan 1975; Hennig et al. 2015). However, there is no universally agreed definition of a cluster (Hennig 2015; McNicholas 2016) and as a result, there are a variety of different clustering approaches. One popular approach is to assume each cluster can be represented using a probability distribution and then attempt to estimate the parameters of these distributions based on the data. This is called model-based clustering and the combination of probability distributions is known as a mixture model (Celeux and Govaert 1995; Fraley and Raftery 2002; McLachlan and Peel 2000). Model-based clustering is a well-established field of research dating back at least as far as Pearson (1894), with key developments in Fisher (1922) and Dempster et al. (1977). McNicholas (2016), Bouveyron et al. (2019), McLachlan et al. (2019), and Gormley et al. (2023) provide detailed reviews of model-based clustering and mixture models.

In a cytometry context, model-based clustering methods view each cell population as a sample drawn from a probability distribution. That is, they assume that the spread of cells in a population can be approximately modeled by a probability density function. The form of this function will be predetermined by the

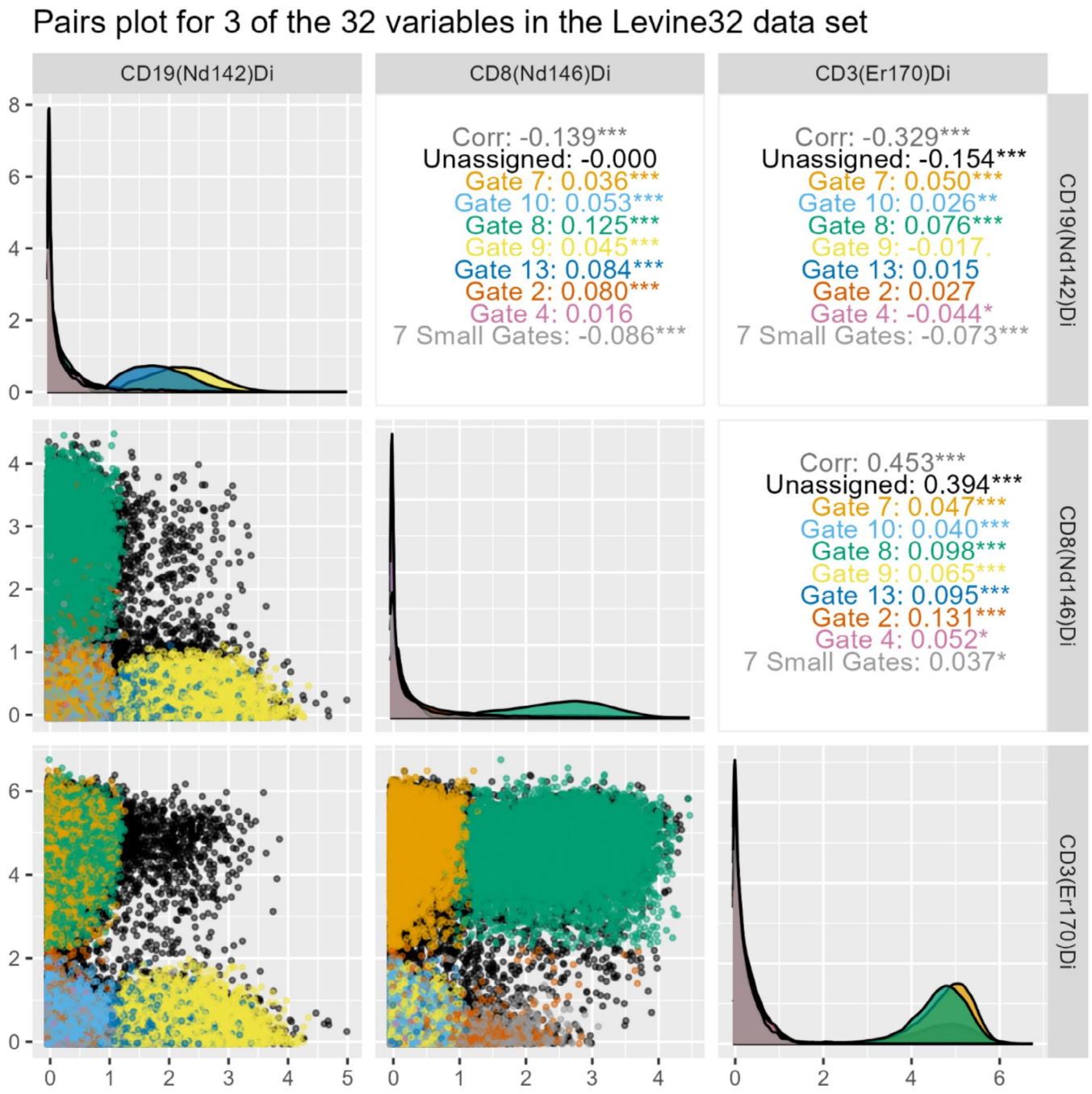


**FIGURE 1** | This pairs-style plot displays three of the 32 variables from the Levine32 data set (Weber and Soneson 2019). These variables correspond to the surface expression of the CD19, CD8, and CD3 markers, which are important for distinguishing between the cell populations in this data set. Only events from the seven largest manually identified cell populations or gates are colored. Unassigned events are shown in black and events from the seven smaller gates are shown in gray.

choice of probability distribution, but its parameter values will be computed based on the data. The probability density function for a finite mixture model with K components is given by Equation (1). Ideally, each component would correspond to a separate cell population. However, in some cases, a clustering solution in which a cell population is modeled by several mixture components may provide a better model fit. This can occur if the shape of that cell population violates the assumptions of the probability distribution used in that mixture model. The mixing proportion, $\pi_k$, is the prior probability of any unspecified event belonging to component $k$. In other words, $\pi_k$ is the proportion of events in component $k$. The parameters of component $k$ are denoted by $\boldsymbol{\theta}_k$ and its probability density function is $f_k(\bullet \mid \boldsymbol{\theta}_k)$. $f_k(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k)$ is the density of component $k$ evaluated at event $i$, $\boldsymbol{y}_i$.

$$f(\boldsymbol{y}_i|\boldsymbol{\theta},\pi) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k) \tag{1}$$

The most common and traditional model-based clustering method is a Normal/Gaussian mixture model. This is a finite mixture model with a probability density function of the form

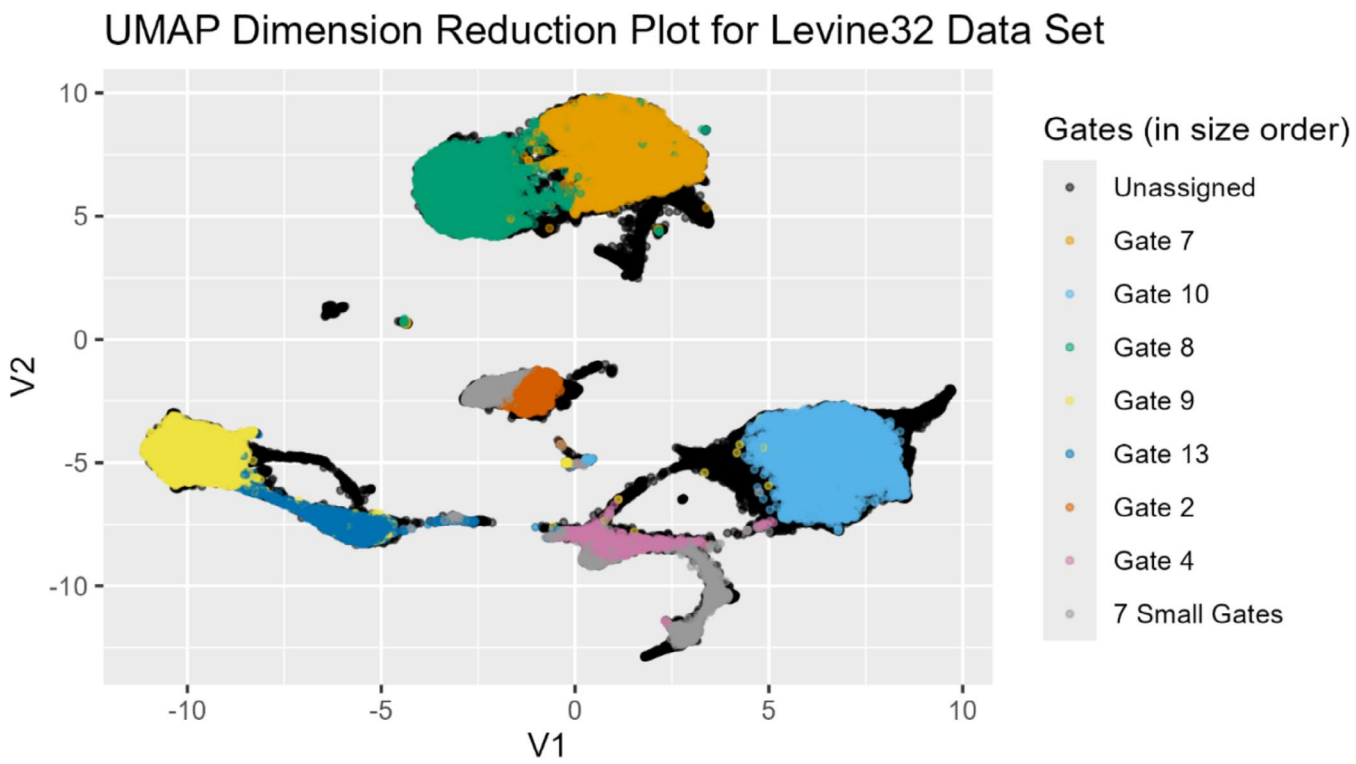


**FIGURE 2** | This scatter plot displays a reduced dimension representation of the Levine32 data set (Weber and Soneson 2019). There are 32 variables in this data set and 14 manually identified cell populations or gates. Events assigned to the seven largest gates are colored. The unassigned events are shown in black and events from the seven smallest gates are shown in gray.

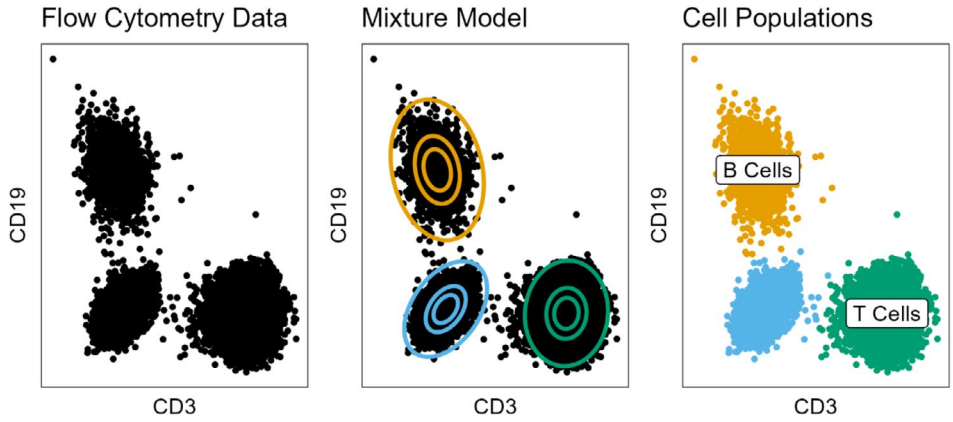


**FIGURE 3** | Three CD3 versus CD19 scatterplots of a flow cytometry sample from the healthyFlowData R package (Azad 2013). The plots show how the data is clustered by a three-component, bivariate Normal mixture model, implemented by the mclust R package (Fraley et al. 2022). They display the two stages of fitting a mixture model: Estimating each cluster's distribution and assigning each event to the cluster with the most suitable distribution. In the left scatterplot, the data is uncolored but three distinct clusters are visible. In the middle scatterplot, three colored, concentric ellipses are imposed over each cluster, showing the 33rd, 66th, and 99th percentile level sets of each cluster's density according to a Normal distribution. In the right scatterplot, the three clusters are colored and the CD3+ CD19− and CD3− CD19+ clusters are labeled as "T Cells" and "B Cells", respectively.

given by Equation (1), where each component is modeled by a multivariate Normal distribution. In $d$ dimensions, the multivariate Normal distribution is parametrized by a $d \times 1$ mean vector, $\boldsymbol{\mu}$, and a positive-definite $d \times d$ variance–covariance matrix, $\boldsymbol{\Sigma}$. The level sets of its probability density function, that is, the curves along which the density is equal, are given by symmetrical ellipsoids, centered at $\boldsymbol{\mu}$ with shape determined by $\boldsymbol{\Sigma}$. Equation (2) gives the probability density function for component $k$ of a Normal mixture model evaluated at event $i$, $\boldsymbol{y}_i$.

$$f_k\left(\boldsymbol{y}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right) = (2\pi)^{-\frac{d}{2}}\left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{y}_i-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k\left(\boldsymbol{y}_i-\boldsymbol{\mu}_k\right)\right) \tag{2}$$

$$\hat{z}_{ik} = \mathbb{P}\left(\text{event } i \in \text{component } k|\boldsymbol{y},\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\theta}}\right) = \frac{\hat{\pi}_k f_k\left(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_k\right)}{\sum_{l=1}^{K}\hat{\pi}_l f_l\left(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_l\right)} \tag{3}$$

Mixture models produce a probabilistic or soft clustering. This means that they compute the probability of each event being assigned to any of the components. Equation (3) displays the formula for the estimated posterior probability of event $i$ being assigned to component $k$ given the data, where $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\theta}}$ represent estimates for the model parameters. To obtain a hard clustering, we assign each event to the component to which it has the highest posterior probability of being assigned. These component assignment probabilities also give us some insight into the level of uncertainty for each individual event and therefore, how well the model fits the data. Figure 3 shows a scatterplot of the CD3 and CD19 protein markers for the healthyFlowData flow cytometry data set (Azad 2013) before clustering, with the symmetrical ellipsoids of a three-component, bivariate Normal mixture model, and colored according to a hard clustering from that mixture model.

### 2.1 | Advantages of Model-Based Clustering

One of the main benefits of employing model-based clustering is its strong statistical framework. As well as the richer, event-specific clustering solution given by probabilistic clustering compared to hard clustering algorithms, using a model-based clustering approach also allows us to be clear about our algorithm's assumptions and provides mathematical justification for our decisions. When choosing a clustering algorithm and interpreting any clustering solution it produces, it is important to understand the type of clusters that it tends to construct. Model-based clustering explicitly assumes the probability distribution of each component, which provides transparency around the form of its clusters. Its mathematical foundations can also provide reassurance about the validity of model selection decisions with respect to the number of components or component distributions. The lack of a true class structure in cytometry makes model-based clustering's clear assumptions and statistical framework particularly appealing.

Another benefit of model-based clustering is its adaptability, as discussed in Sections 3–7. Changing the probability distribution of a mixture model influences the type of clusters that the model will construct. Extending a sample-specific model to form a hierarchical multi-sample model allows information to propagate between samples. Factor analysis includes dimension reduction as a part of the clustering process, allowing each cluster to have its own projection to a lower-dimensional space. The model fitting process can also be modified to search for rare populations. These adaptations, which will be discussed in more detail in the following sections, allow mixture models to be adjusted to suit the complexities of flow and mass cytometry data, without losing the benefits of their mathematical origins.

One drawback of model-based clustering is its computational cost. Complex models require the estimation of a high number of parameters, which can be relatively slow, depending on the available computing power. As a result, fitting a mixture model and obtaining a soft, probabilistic clustering will often be more computationally demanding than obtaining a hard clustering from a more straightforward clustering algorithm, such as $k$-means. However, mixtures of factor analyzers and other parsimonious mixture models can be used to reduce the number of parameters to be estimated.

Another disadvantage is that the form of the clusters identified by a mixture model can be limited by the distributions chosen for its components. That is, if the shapes of the cell populations violate the assumptions underpinning the mixture model, then it may not be flexible enough to cluster the data satisfactorily. This can be partially mitigated by choosing a more flexible and robust distribution for the components, as discussed in the next section.

## 3 | Non-Normal Populations: Alternative Distributions

Model-based clustering operates within a strong statistical and probabilistic framework compared to other clustering approaches. This mathematical foundation is obtained by making assumptions about the distribution of the data within each cluster. For example, the Normal mixture model defined by Equations (1) and (2) assumes that each cluster occupies a symmetric ellipsoidal region and that events are dispersed through this region according to a multivariate Normal distribution. Mixture models can still be effective when their assumptions are not fully satisfied, however, there can be undesirable consequences if the assumptions are severely violated. An example of this is the tendency of Normal mixture models to model a single non-normal cluster using multiple components (Baudry et al. 2010). One of the challenges associated with using model-based clustering for population identification is that the cell populations may not be approximately Normally distributed either in terms of their asymmetrical shape or the concentration of events away from the center of the population. In model-based clustering, these issues are referred to as skewness and heavy-tailedness, respectively. However, by choosing a more sophisticated distribution for our mixture components than the Normal distribution, it is possible to address these concerns.

The level sets of a $t$ distribution's probability density are symmetric ellipsoids, similar to those of a Normal distribution (Andrews and McNicholas 2012; McLachlan and Peel 1998). However, the probability density of a $t$ distribution decreases less rapidly away from the mean than that of a Normal distribution. Hence, sampling from a $t$ distribution is more likely to result in points far from the mean than sampling from a Normal distribution. As a result, such points can be accommodated more comfortably by $t$ distributions. Skewed distributions are generalizations of symmetric distributions which allow for asymmetry. Whereas Normal and $t$ mixture models assume that the shape of each population can be approximated by a symmetric ellipsoid, their skewed generalizations, skew–normal and skew-$t$ mixture models, include a skewness parameter to explicitly account for asymmetry in the shape of a cluster. Lee and McLachlan (2022) provide a detailed review of mixtures of skewed distributions. Many non-normal mixture models were not necessarily intended specifically for cytometry data but instead were developed as a generic clustering method (Franczak et al. 2014; Lee and McLachlan 2016; O'Hagan et al. 2016; Zhu and Melnykov 2018). Among others, Bouveyron et al. (2019) credit McLachlan and Peel (1998) for the $t$ mixture, Lin (2009) for the skew–normal mixture, and Pyne

et al. (2009) for the skew-*t* mixture. However, the development of the skew–normal distribution itself is attributed to Azzalini and Dalla Valle (1996). Interestingly, Pyne et al. (2009) do explicitly intend for the skew-*t* mixture to be used for flow cytometry data. The software implementation corresponding to this publication is named FLow analysis with Automated Multivariate Estimation (FLAME). Furthermore, the "Non-Gaussian Model-based Clustering" chapter of Bouveyron et al. (2019) uses flow cytometry data as an example.

Several methods for flow and mass cytometry have used these more flexible distributions to account for the non-normal populations commonly found in this data. flowClust (Lo et al. 2008) addresses asymmetry in the data by applying a Box–Cox transformation (Box and Cox 1964) intended to make the populations more symmetric (Lo and Gottardo 2012), then uses a standard *t* mixture model as a more outlier-friendly alternative to a Normal mixture model. FLAME (Pyne et al. 2009), JCM (Pyne et al. 2014), and NPflow (Hejblum et al. 2019) take a more direct approach by using the skew-*t* distribution. However, the symmetric Normal remains the most commonly used distribution. It is the basis of HDPGMM (Cron et al. 2013), ASPIRE (Dundar et al. 2014), SWIFT (Naim et al. 2014), BayesFlow (Johnsson et al. 2016), CYBERTRACK (Minoura et al. 2019, 2021), LAMBDA (Abe et al. 2020), and Tailor (Ionita et al. 2021). Meanwhile, immunoClust (Sörensen et al. 2015) and CytoFA (Lee 2019) utilize a *t* distribution and a skew–Normal distribution, respectively, both choosing an intermediate option between the symmetric Normal and the skew-*t*.

## 4 | Multiple Samples: Hierarchical Models

One of the most interesting challenges associated with clustering data from flow or mass cytometry is the need to cluster multiple samples in a coherent and consistent manner. Flow and mass cytometry are often employed in immunological studies involving multiple sets of samples from different control or treatment groups. Population identification methods must identify corresponding clusters of events in each sample, despite the biological differences between the patients and any technical variation between the samples. In particular, these biological differences are often the focus of the experiment and must not be eliminated. Figure 4 displays scatterplots of the CD4 and CD8 protein markers for four flow cytometry samples from the healthyFlowData package (Azad 2013), where each sample was obtained from a different subject. The four samples are clustered according to four-component Normal mixture models implemented by the mclust package (Fraley et al. 2022). This figure demonstrates how the shape, size, and location of corresponding cell populations can differ between different cytometry samples.

The most straightforward approach to a multi-sample clustering problem would be to combine the samples into one large, pooled sample and apply the clustering algorithm once. This is the approach taken by Tailor (Ionita et al. 2021) when dealing with multiple samples. It relies on the assumption that the shape and location of populations do not vary substantially between samples. If this assumption is violated, then distinct populations from different samples can overlap in the pooled sample despite not overlapping in any individual sample. However, pooling does allow for some sharing of information between samples. A rare population could contain too few events to be identified in some samples but become identifiable in the pooled sample when its events from all samples are considered.

Another possibility is to cluster each sample individually and apply a meta-clustering algorithm to match groups of similar clusters from different samples. FLAME (Pyne et al. 2009)

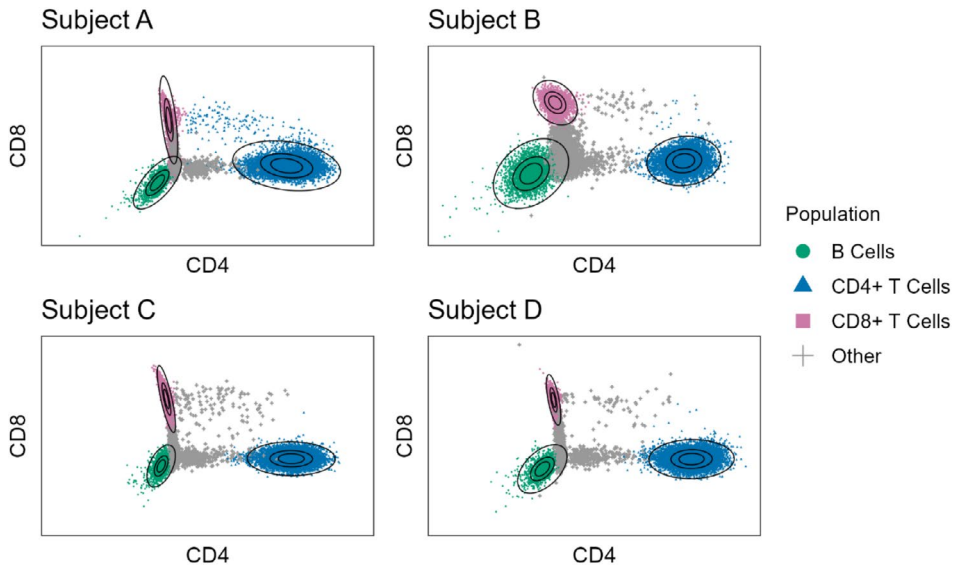


**FIGURE 4** | Four CD4 versus CD8 scatterplots of healthyFlowData (Azad 2013) samples from four individuals. The first sample from each subject is shown. These plots illustrate that corresponding cell populations from different subjects in the same study can have different shapes, sizes, and locations despite being analyzed in the same way. Each plot is colored according to a four-component Normal mixture model implemented by the mclust R package (Fraley et al. 2022). Three concentric ellipses are imposed over three of the four clusters. These are the 33rd, 66th, and 99th percentile level sets of the Normal density for the clusters identified as "B Cells," "CD4+ T Cells," and "CD8+ T Cells".

and immunoClust (Sörensen et al. 2015) both include meta-clustering extensions to their original clustering algorithms. Several standalone meta-clustering methods have also been released which are intended to be used after any clustering algorithm has been applied to each sample individually, for example, flowMatch (Azad et al. 2012), FlowMap-FR (Hsiao et al. 2016), QFMatch (Orlova et al. 2018), and MetaCyto (Hu et al. 2018). One popular approach to meta-clustering, used by FLAME and immunoClust, is to treat the cluster centers from all samples as a new set of data points and apply a clustering algorithm to them. Another approach is to construct a bipartite graph from two sets of clusters and match the clusters based on pairwise dissimilarities, similar to flowMatch. Figure 5 demonstrates how flowMatch can identify sets of corresponding clusters from different samples and construct template Normal components for each of these meta-clusters. It can be difficult to correctly match corresponding populations in different samples. Meta-clustering does not benefit from any sharing of information between samples when constructing the clusters.

Model-based clustering also facilitates a joint clustering approach. By defining a hierarchical model, it is possible to allow information to propagate between samples without pooling them. These models jointly cluster all samples simultaneously. They construct a model for each sample, similar to model-based clustering with meta-clustering, but consider these sample-specific models to be variations of a cross-sample template model. As a result, information from one sample, for example, evidence of the existence of a rare population, can affect the clustering solution constructed for the other samples by influencing the form of the template model. JCM (Pyne et al. 2014) and ASPIRE (Dundar et al. 2014) implement joint clustering models based on random effects, whereas BayesFlow (Johnsson et al. 2016) jointly clusters using a hierarchical Bayesian model.

Another model-based approach to the multi-sample problem is sequential clustering. If a Bayesian model is applied to a single sample, then a posterior distribution for the model parameters will be obtained which can be used as a prior distribution for the next sample's model. NPflow (Hejblum et al. 2019) is an example of a model that uses this sequential Bayesian approach.

## 5 | Unknown Population Number: Model Selection

Due to the complexity of the population identification problem, it may be difficult to choose an appropriate number of clusters before applying a clustering algorithm. Therefore, having a data-driven procedure to inform this choice is desirable. However, many clustering algorithms do not have natural methods to select the number of clusters. In contrast, model-based clustering methods are able to take advantage of mathematically derived likelihood-based model selection criteria such as the Bayesian Information Criterion (BIC; Schwarz 1978) or the Integration Classification Likelihood (ICL; Biernacki et al. 2000) to select the number of components to be used in their mixture model (Fraley and Raftery 2002). The likelihood for the model parameters increases as the number of components increases, so these criteria include model complexity penalty terms. When comparing two models with different numbers of components, the criterion will only select the model with a higher number of components if the increase in the likelihood is larger than the increase in the complexity penalty term. Baudry et al. (2010) define the BIC and ICL as shown in Equations (4) and (5), respectively, where $\hat{\pi}_k$ and $\hat{\boldsymbol{\theta}}_k$ are the estimated mixing proportion and parameters for component $k$, $\hat{z}_{ik}$ is the estimated probability of event $i$ being assigned to component $k$, and $\nu_K$ is the number of estimated parameters required by the model with $K$ components. They also state that the ICL is approximately equal to the BIC with an additional mean entropy penalty term. This means that the ICL more severely penalizes clustering solutions with higher component assignment uncertainties. In practice, this means that the ICL is less likely to choose a model with overlapping or adjacent components than the BIC, since events near the boundary between these components could reasonably be assigned to either. As a result, the ICL tends to select a lower number of components than the BIC. Another model selection approach

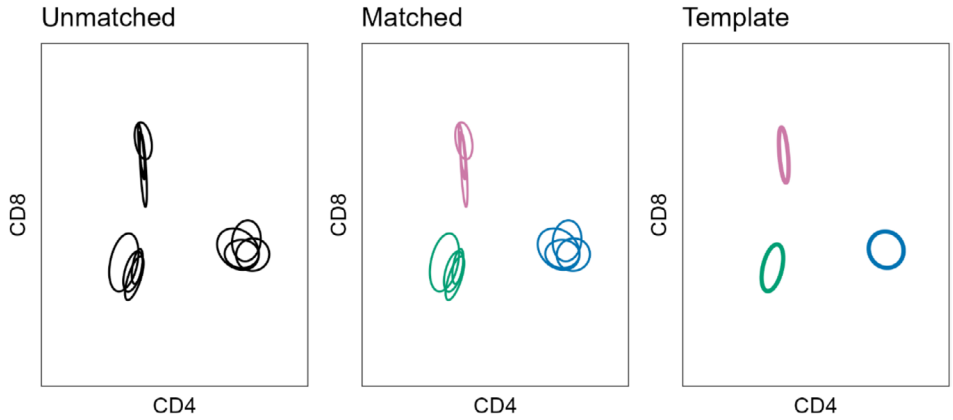


**FIGURE 5** | Three plots illustrating how flowMatch (Azad et al. 2012) matches clusters from different samples and constructs a template Normal mixture model with one component per meta-cluster, or set of clusters. A four-component Normal mixture model was fitted to the first sample of each of the four subjects from the healthyFlowData R package (Azad 2013) using the mclust R package (Fraley et al. 2022). Three components from each sample's mixture model were provided to flowMatch. *Left*: 66th percentile ellipses for 12 components from four samples are overlayed on the same plot. *Middle*: The same 12 ellipses colored as three sets of four clusters according to a meta-cluster assignment by flowMatch. *Right*: 66th percentile ellipses for the three template components constructed by flowMatch for each meta-cluster.

is to use the Bayes factor (Kass and Raftery 1995), although we are not aware of this being used by any cytometry-specific methods. flowClust (Lo et al. 2008), JCM (Pyne et al. 2014), and LAMBDA (Abe et al. 2020) utilize the BIC, whereas immunoClust (Sörensen et al. 2015) employs a modified version of the ICL which is intended to serve as a compromise between the two criteria. The main drawback of using these model selection criteria is that models with different numbers of components can only be compared after being fitted to the data, and the process of fitting a range of models and only using one can be considered time-consuming and wasteful.

$$\mathrm{BIC}(K) = \sum_{i=1}^{N} \log\left( \sum_{k=1}^{K} \pi_k f_k\left(\boldsymbol{y}_i | \widehat{\boldsymbol{\theta}}_k\right) \right) - \frac{\nu_K}{2} \log(n) \tag{4}$$

$$\mathrm{ICL}(K) = \sum_{i=1}^{N} \sum_{k=1}^{K} \widehat{z}_{ik} \log\left( \pi_k f_k\left(\boldsymbol{y}_i | \widehat{\boldsymbol{\theta}}_k\right) \right) - \frac{\nu_K}{2} \log(n) \tag{5}$$

Bayesian nonparametric mixture models offer an alternative approach (Ghahramani 2013). They incorporate selecting the number of components into the model fitting process so that it is carried out as the parameters of the model are being estimated. This allows them to fit a single model to the data without prespecifying the number of components to be used. However, these methods often utilize Markov Chain Monte Carlo (MCMC) sampling schemes, which can be computationally demanding. Summarizing the output of Bayesian nonparametric models can also require additional consideration, as they sample from the posterior distribution over the space of partitions instead of converging to a single clustering solution. Label-switching issues can be resolved using loss-based relabeling approaches (Jasra et al. 2005; Stephens 2000), while more recently Wade and Ghahramani (2018) outlined a general approach for identifying an optimal partition of the data using MCMC output. Lin and Hejblum (2021) provide a discussion of Bayesian nonparametric mixture modeling. Several methods for flow and mass cytometry have taken this approach using Dirichlet process priors (Ferguson 1973; Neal 2000), for example, HDPGMM (Cron et al. 2013), ASPIRE (Dundar et al. 2014), and NPflow (Hejblum et al. 2019).

Another approach for determining the number of cell populations is to fit a relatively high number of mixture components to the data and apply a merging procedure to combine multiple components for each cell population. It is well-established in model-based clustering that if the shape of a cluster is not compatible with the mixture components' probability distribution, then the model-fitting process may compensate by modeling a single cluster with more than one component (Fraley and Raftery 2002; Pyne et al. 2009). A variety of approaches have been proposed for merging mixture components (Hennig 2010). flowMerge (Finak et al. 2009) applies the entropy-based merging approach of Baudry et al. (2010) to mixture components fitted by flowClust (Lo et al. 2008). Another approach is to associate each component with a local density maximum or mode and then merge any components that share the same mode. The Hidden Markov Model on Variable Blocks (HMM-VB) method (Lin and Li 2017) employs a modal clustering approach as part of the model-fitting process. BayesFlow (Johnsson et al. 2016) and SWIFT (Naim et al. 2014) both merge unimodal components, while SWIFT also splits multimodal components. immunoClust (Sörensen et al. 2015) and HDPGMM (Cron et al. 2013) carry out merging in the context of dealing with multiple samples. Tailor (Ionita et al. 2021) annotates each component based on their expression levels for every marker and merges components with the same annotation. Over-clustering with a mixture model and then merging some of the resulting mixture components is appealingly flexible since a union of components can form a wider variety of cluster shapes than a single component. However, proceeding in a fully principled manner when merging components is challenging, because each merging approach must be motivated by subjective choices about what constitutes a cluster (Hennig 2010). Another drawback of this approach is that the flexibility obtained by representing a cluster as a union of components comes at the expense of interpretability as the cluster will not be described by the model parameters of a single component and the process through which it was constructed is more convoluted.

Despite these model-based approaches, selecting the correct number of clusters remains a difficult challenge in analyzing flow or mass cytometry data. There is no objectively correct number of populations in a cytometry sample since hierarchical relationships exist between many cell populations. As a result, distinct solutions with different levels of cluster resolution may be suitable for particular problems. In other words, experts may require a large number of small clusters or a small number of large clusters, depending on the context of their research. Although the intermediate solutions through which an iterative merging procedure transitions could be interpreted as representing a hierarchy of clustering solutions, it is also common for these methods to only present the final outcome. Neither likelihood-based criteria nor nonparametric models address the dual issues of cluster resolution and cell hierarchy as they both result in a single number of components being chosen. While mathematical justification exists for this choice, it may not suit the specific needs of the researcher.

## 6 | Rare Populations: Novel Model Fitting Procedures

Cytometry data sets can contain both rare and abundant populations. Some model-based clustering methods for cytometry data have developed alternative model-fitting approaches designed to better identify these rare populations.

SWIFT (Naim et al. 2014) uses a novel procedure called Weighted Iterative Sampling which is designed to prevent large populations from dominating the model fitting process. SWIFT starts by fitting a mixture model to a uniform subsample from the data. The $p$ most populous components are included in all subsequent models with their original parameter values fixed. Another subsample is computed from the data with events sampled according to their probability of not being assigned to any of the previously fixed components. Then, a new model is fitted to this subsample, which includes the previously fixed components, and the $p$ most populous components from this model are also fixed. This process of constructing a weighted subsample which avoids events from previous components and fitting a new model to obtain the

next $p$ components is repeated until all components' parameter values have been fixed.

immunoClust (Sörensen et al. 2015) adopts a hierarchical divisive clustering approach. It fits models with a range of component numbers to a weighted subsample from the full data set, including a single component model, and uses a modified version of the ICL criterion defined by Equation (5) to select one of these models. This model is then used to partition the data set and a new range of models is fitted to each of the resulting subsets. This process is repeated until the modified ICL criterion selects the single component model for each remaining subset.

HMM-VB (Lin and Li 2017) is a model-based clustering approach that is designed to identify rare clusters in large data sets, and is partly motivated by cytometry data. This model involves arranging the variables into a sequence of disjoint sets and fitting a Gaussian mixture model to each variable block with a sequential dependence between these blocks controlled by a Hidden Markov Model. The model-fitting procedure is carried out using a modal version of the Baum–Welch algorithm, which is related to the EM algorithm. This modification involves merging mixture components with the same mode.

Initialization is an important aspect of model-fitting, particularly for mixture models that are fitted to data via an EM algorithm (Baudry and Celeux 2015). Model-based hierarchical clustering can be used to obtain an initial partition of the data (Fraley 1998; Fraley and Raftery 2002; Scrucca and Raftery 2015). This is the initialization technique used by immunoClust, flowClust, and flowMerge. Another approach is to test multiple initializations and retain the one that achieved the highest likelihood after a small number of EM iterations (O'Hagan et al. 2012). FLAME, HMM-VB, and SWIFT all utilize multiple initial partitions. FLAME partitions the data by applying the k-means clustering algorithm, while SWIFT generates random partitions, and HMM-VB uses a combination of both types of partition. Tailor's binning approach, which is discussed in the next section, is in part an initialization technique as the representatives of the largest bins are used as initial component means. An intriguing and potentially promising idea for cytometry data sets would be to obtain an initial partition by clustering a reduced dimension representation of the data obtained from a nonlinear dimension reduction technique, such as t-SNE and UMAP. However, we have not come across any methods which utilize this approach. Researchers may have been put off by the possibility of irregularly shaped initial clusters as a result of the nonlinear nature of the dimension reduction, or it could simply be an underexplored area of research.

## 7 | Large Data Sets: Binning and Factor Analysis

The size of flow and mass cytometry data sets can cause computational challenges both in relation to dimensionality and sample size, that is, the number of variables and the number of events. CytoFA (Lee 2019) applies a factor analysis approach as a dimensionality reduction technique while Tailor (Ionita et al. 2021) employs binning to construct a smaller representative data set used for model fitting.

CytoFA adopts a factor analysis approach for a skew–normal mixture model. Factor analysis is a popular model-based method for clustering high-dimensional data which incorporates dimension reduction into the model fitting process (Ghahramani and Hinton 1996; McNicholas and Murphy 2008; Murphy et al. 2020). It reduces the number of parameter values that need to be estimated to reduce the computational complexity of the model. This approach assumes that the variation present in the given data can be modeled by a simpler covariance structure than that of a traditional, high-dimensional model. Instead of modeling each cluster as a set of data points sampled from a high-dimensional random variable, factor analysis models each cluster as a high-dimensional projection of a set of data points sampled from a low-dimensional random variable.

Whereas CytoFA uses a dimensionality reduction approach to address the high number of variables, Tailor instead tries to reduce the sample size by binning the events and replacing each bin with a representative data point. Each variable is split univariately to create an orthogonal partitioning of the data space. Highly populous regions are further divided into bins using $k$-means, sparsely populated regions are discarded, and moderately populated regions are used as bins directly. A single weighted representative data set is obtained from each of the bins. A weighted EM algorithm is used to fit a Normal mixture model to this representative data set. Other methods, including SWIFT (Naim et al. 2014) and immunoClust (Sörensen et al. 2015), attempt to reduce the number of events by subsampling, where a subset of the original events are selected, rather than constructing new data points as representatives.

## 8 | Other Challenges Related to Manual Gating

To assess how well a clustering algorithm performs when applied to a flow cytometry or mass cytometry data set, its clustering solution will often be compared to a manual gating of the same data set. The $F_1$ measure (Rijsbergen 1979) is commonly used to quantify the similarity between the set of clusters produced by the algorithm and a set of gates produced by an expert. Evaluating the performance of population identification algorithms by comparing their clustering solutions to manual gating using external validation indices, such as the $F_1$ measure, assumes that the manual gating represents the true class structure of the data. However, if a researcher is focused on a particular cell type, they may not be interested in distinguishing between every subpopulation of another cell type. Hence, the manual gating may only represent one of the valid class structures for a particular data set. If an algorithm discovers a structure in the data that was not identified in the manual gating, then this will reduce its $F_1$ score because its clusters will be less similar to the gates.

It may be beneficial to allow the expert to provide information to the clustering algorithm about which cell populations they are trying to identify. This could increase the practicality of an algorithm in a research context and improve the performance of the algorithm with respect to external validation indices in demonstrations and applied comparisons. The ACDC algorithm developed by Lee et al. (2017) utilizes expert information


**FIGURE 6** | A graphical summary of the challenges and adaptations of using model-based clustering for flow cytometry. Five challenges (green) are centered around "Model-Based Clustering for Cytometry" (pink). Two adaptations (purple) arise from each of these challenges.

in the form of a table with one row per cell type and one column per protein marker, whose entries indicate whether the cell type is positive, negative, or neutral for that marker. Ji et al. (2018) inform their Bayesian tree algorithm using the same table. However, neither of these methods were model-based. To the best of our knowledge, other than some discussion on informative priors by BayesFlow (Johnsson et al. 2016), the use of expert information has not yet been addressed by model-based clustering algorithms designed for flow and mass cytometry data.

Another challenge relating to replicating manual gating is the practice of leaving events unassigned. As a result of the sequential subsetting nature of manual gating, many events are not assigned to any of the final gates. This allows the manual gating process to naturally perform outlier detection, as well as excluding borderline events that could belong to more than one population. For some data sets, half of events can be left unassigned (Levine et al. 2015). Traditional mixture models are designed to assign every event to one of their mixture components. However, methods have been developed for outlier identification in model-based clustering. Mixture models can be modified to include a single uniform component across the region occupied by the data to which outliers should be assigned (Banfield and Raftery 1993). Another approach, called trimming, involves identifying and excluding outliers during the model fitting procedure (Gallegos and Ritter 2005; García-Escudero et al. 2008). In this case, events are usually classified as outliers if they do not fit the model well. In other words, if the value of the mixture model's density function at that event is low. Chan et al. (2008) discuss labeling as unassigned any events that lie outside a prespecified confidence region for their assigned component or whose posterior probability falls below a prespecified threshold. flowClust (Lo et al. 2008) classifies outliers based on the Mahalanobis distance from an event to the mean of each component. CytoFA (Lee 2019) trims the events with the lowest contribution to the mixture model. We are not aware of any other model-based clustering algorithms that attempt to address the prevalence of unassigned events in cytometry data.

## 9 | Conclusion

Model-based clustering is a statistical approach to data clustering that can be used for population identification in flow and mass cytometry. This approach's appeal lies in its clear assumptions and mathematical foundations. Population identification is a particularly challenging problem. Among the difficulties associated with this task are the non-normality of the populations, the multi-sample nature of cytometry studies, the absence of a known number of populations, the imbalance between the sizes of abundant and rare populations, and the high dimensionality and large sample size of the data sets. We have discussed a number of ways in which mixture models have been modified to address these issues, which are displayed in Figure 6. These include the adoption of more robust and flexible distributions, the design of hierarchical, multi-sample models, the application of model selection criteria and nonparametric mixtures, the development of novel model fitting procedures, and the use of data reduction techniques, such as binning and factor analysis. The range and potential of these adaptations demonstrate the capability of model-based clustering to evolve to overcome the challenges posed by population identification. One source of issues with cytometry data that was not discussed in this article is the requirement for preprocessing and transformation of cytometry data (Liechti et al. 2021).

In Section 8, we highlighted that the practice of treating manual gating as the ground truth for external validation could penalize algorithms for identifying a different but potentially valid class structure in the data. We believe that the development of population identification algorithms that allow the expert to inform the model is an area that has been under-explored by the model-based clustering community. As well as improving the quantitative performance of the algorithm in comparisons, this also has the potential for real-world benefits by allowing immunological researchers to tailor the algorithm to their specific needs.

Our review has discussed the challenges of population identification for flow and mass cytometry data and highlighted a number of adaptations that have been developed to aid model-based clustering in addressing these challenges. We believe that the statistical foundations and adaptability of model-based clustering make it a suitable choice for the problem of population identification.

### Author Contributions

**Ultán P. Doherty:** conceptualization (lead), funding acquisition (lead), investigation (lead), visualization (lead), writing – original draft (lead), writing – review and editing (equal). **Rachel M. McLoughlin:** conceptualization (supporting), funding acquisition (equal), supervision (supporting), writing – review and editing (equal). **Arthur White:** conceptualization (supporting), funding acquisition (equal), investigation (supporting), supervision (lead), writing – review and editing (equal).

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

### Related WIREs Articles

Challenges in model-based clustering

Bayesian mixture models for cytometry data analysis

### References


Abe, K., K. Minoura, Y. Maeda, H. Nishikawa, and T. Shimamura. 2020. "Model-Based Clustering for Flow and Mass Cytometry Data With Clinical Information." *BMC Bioinformatics* 21, no. Suppl 13: 393. https://doi.org/10.1186/s12859-020-03671-7.

Aghaeepour, N., P. Chattopadhyay, M. Chikina, et al. 2016. "A Benchmark for Evaluation of Algorithms for Identification of Cellular Correlates of Clinical Outcomes." *Cytometry, Part A* 89, no. 1: 16–21. https://doi.org/10.1002/cyto.a.22732.

Aghaeepour, N., R. Nikolic, H. H. Hoos, and R. R. Brinkman. 2011. "Rapid Cell Population Identification in Flow Cytometry Data." *Cytometry, Part A* 79A, no. 1: 6–13. https://doi.org/10.1002/cyto.a.21007.

Andrews, J. L., and P. D. McNicholas. 2012. "Model-Based Clustering, Classification, and Discriminant Analysis via Mixtures of Multivariate t-Distributions." *Statistics and Computing* 22, no. 5: 1021–1029. https://doi.org/10.1007/s11222-011-9272-x.

Azad, A. 2013. "healthyFlowData: Healthy dataset used by the flowMatch package. Doi:10.18129/B9.bioc.healthyFlowData, R package version 1.40.0 [Computer software]." https://bioconductor.org/packages/healthyFlowData.

Azad, A., S. Pyne, and A. Pothen. 2012. "Matching Phosphorylation Response Patterns of Antigen-Receptor-Stimulated T Cells via Flow Cytometry." *BMC Bioinformatics* 13, no. S2: S10. https://doi.org/10.1186/1471-2105-13-S2-S10.

Azzalini, A., and A. Dalla Valle. 1996. "The Multivariate Skew-Normal Distribution." *Biometrika* 83, no. 4: 715–726.

Banfield, J. D., and A. E. Raftery. 1993. "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics* 49, no. 3: 803–821. https://doi.org/10.2307/2532201.

Baudry, J.-P., and G. Celeux. 2015. "EM for Mixtures: Initialization Requires Special Care." *Statistics and Computing* 25, no. 4: 713–726. https://doi.org/10.1007/s11222-015-9561-x.

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, no. 2: 332–353. https://doi.org/10.1198/jcgs.2010.08111.

Becht, E., L. McInnes, J. Healy, et al. 2019. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology* 37, no. 1: 38–44. https://doi.org/10.1038/nbt.4314.

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, no. 7: 719–725. https://doi.org/10.1109/34.865189.

Bouveyron, C., G. Celeux, T. B. Murphy, and A. E. Raftery. 2019. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press. https://doi.org/10.1017/9781108644181.

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B: Methodological* 26, no. 2: 211–252.

Celeux, G., and G. Govaert. 1995. "Gaussian Parsimonious Clustering Models." *Pattern Recognition* 28, no. 5: 781–793. https://doi.org/10.1016/0031-3203(94)00125-6.

Chan, C., F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler. 2008. "Statistical Mixture Modeling for Cell Subtype Identification in Flow Cytometry." *Cytometry, Part A* 73, no. 8: 693–701. https://doi.org/10.1002/cyto.a.20583.

Commenges, D., C. Alkhassim, R. Gottardo, B. Hejblum, and R. Thiébaut. 2018. "Cytometree: A Binary Tree Algorithm for Automatic Gating in Cytometry Analysis." *Cytometry, Part A* 93, no. 11: 1132–1140. https://doi.org/10.1002/cyto.a.23601.

Cron, A., C. Gouttefangeas, J. Frelinger, et al. 2013. "Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment Across Flow Cytometry Samples." *PLoS Computational Biology* 9, no. 7: e1003130. https://doi.org/10.1371/journal.pcbi.1003130.

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: Methodological* 39, no. 1: 1–38.

Dundar, M., F. Akova, H. Z. Yerebakan, and B. Rajwa. 2014. "A Non-parametric Bayesian Model for Joint Cell Clustering and Cluster Matching: Identification of Anomalous Sample Phenotypes With Random Effects." *BMC Bioinformatics* 15, no. 1: 314. https://doi.org/10.1186/1471-2105-15-314.

Ferguson, T. S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics* 1, no. 2: 209–230. https://doi.org/10.1214/aos/1176342360.

Finak, G., A. Bashashati, R. Brinkman, and R. Gottardo. 2009. "Merging Mixture Components for Cell Population Identification in Flow Cytometry." *Advances in Bioinformatics* 2009: 247646. https://doi.org/10.1155/2009/247646.

Finak, G., M. Langweiler, M. Jaimes, et al. 2016. "Standardizing Flow Cytometry Immunophenotyping Analysis From the Human ImmunoPhenotyping Consortium." *Scientific Reports* 6, no. 1: 20686. https://doi.org/10.1038/srep20686.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character* 222: 309–368.

Fraley, C. 1998. "Algorithms for Model-Based Gaussian Hierarchical Clustering." *SIAM Journal on Scientific Computing* 20, no. 1: 270–281. https://doi.org/10.1137/S1064827596311451.

Fraley, C., and A. E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97, no. 458: 611–631. https://doi.org/10.1198/016214502760047131.

Fraley, C., A. E. Raftery, L. Scrucca, T. B. Murphy, and M. Fop. 2022. "mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation (Version 6.0.0) [Computer software]." https://cran.r-project.org/web/packages/mclust/index.html.

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, no. 6: 1149–1157. https://doi.org/10.1109/TPAMI.2013.216.

Gallegos, M. T., and G. Ritter. 2005. "A Robust Method for Cluster Analysis." *Annals of Statistics* 33, no. 1: 347–380.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar. 2008. "A General Trimming Approach to Robust Cluster Analysis." *Annals of Statistics* 36, no. 3: 1324–1345. https://doi.org/10.1214/07-AOS515.

van Gassen, S., B. Callebaut, M. J. van Helden, et al. 2015. "FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data: FlowSOM." *Cytometry, Part A* 87, no. 7: 636–645. https://doi.org/10.1002/cyto.a.22625.

Ghahramani, Z. 2013. "Bayesian Non-parametrics and the Probabilistic Approach to Modelling." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371, no. 1984: 20110553. https://doi.org/10.1098/rsta.2011.0553.

Ghahramani, Z., and G. E. Hinton. 1996. The EM Algorithm for Mixtures of Factor Analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Gormley, I. C., T. B. Murphy, and A. E. Raftery. 2023. "Model-Based Clustering." *Annual Review of Statistics and Its Application* 10, no. 1: 573–595. https://doi.org/10.1146/annurev-statistics-033121-115326.

Hartigan, J. A. 1975. *Clustering Algorithms*. 99th ed. John Wiley & Sons, Inc.

Hejblum, B. P., C. Alkhassim, R. Gottardo, F. Caron, and R. Thiébaut. 2019. "Sequential Dirichlet Process Mixtures of Multivariate Skew *t*-Distributions for Model-Based Clustering of Flow Cytometry Data." *Annals of Applied Statistics* 13, no. 1: 638–660. https://doi.org/10.1214/18-AOAS1209.

Hennig, C. 2010. "Methods for Merging Gaussian Mixture Components." *Advances in Data Analysis and Classification* 4, no. 1: 3–34. https://doi.org/10.1007/s11634-010-0058-3.

Hennig, C. 2015. "What are the true clusters? arXiv:1502.02555 [Stat]." http://arxiv.org/abs/1502.02555.

Hennig, C., M. Meila, F. Murtagh, and R. Rocci. 2015. *Handbook of Cluster Analysis*. Chapman & Hall/CRC. https://www.routledge.com/Handbook-of-Cluster-Analysis/Hennig-Meila-Murtagh-Rocci/p/book/9780367570408.

Hsiao, C., M. Liu, R. Stanton, M. McGee, Y. Qian, and R. H. Scheuermann. 2016. "Mapping Cell Populations in Flow Cytometry Data for Cross-Sample Comparison Using the Friedman-Rafsky Test Statistic as a Distance Measure: FCM Cross-Sample Comparison." *Cytometry, Part A* 89, no. 1: 71–88. https://doi.org/10.1002/cyto.a.22735.

Hu, Z., C. Jujjavarapu, J. J. Hughey, et al. 2018. "MetaCyto: A Tool for Automated Meta-Analysis of Mass and Flow Cytometry Data." *Cell Reports* 24, no. 5: 1377–1388. https://doi.org/10.1016/j.celrep.2018.07.003.

Ionita, M., R. Schretzenmair, D. Jones, J. Moore, L.-S. Wang, and W. Rogers. 2021. "Tailor: Targeting Heavy Tails in Flow Cytometry Data With Fast, Interpretable Mixture Modeling." *Cytometry, Part A* 99, no. 2: 133–144. https://doi.org/10.1002/cyto.a.24307.

Jasra, A., C. C. Holmes, and D. A. Stephens. 2005. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science* 20, no. 1: 50–67. https://doi.org/10.1214/088342305000000016.

Ji, D., E. Nalisnick, Y. Qian, R. H. Scheuermann, and P. Smyth. 2018. "Bayesian Trees for Automated Cytometry Data Analysis." In Proceedings of the 3rd Machine Learning for Healthcare Conference, pp. 465–483. https://proceedings.mlr.press/v85/ji18a.html.

Johnsson, K., J. Wallin, and M. Fontes. 2016. "BayesFlow: Latent Modeling of Flow Cytometry Cell Populations." *BMC Bioinformatics* 17, no. 1: 25. https://doi.org/10.1186/s12859-015-0862-z.

Kass, R. E., and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90, no. 430: 773–795. https://doi.org/10.2307/2291091.

Kobak, D., and P. Berens. 2019. "The Art of Using t-SNE for Single-Cell Transcriptomics." *Nature Communications* 10, no. 1: 5416. https://doi.org/10.1038/s41467-019-13056-x.

Kohonen, T. 1990. "The Self-Organizing Map." *Proceedings of the IEEE* 78, no. 9: 1464–1480. https://doi.org/10.1109/5.58325.

Lee, H.-C., R. Kosoy, C. E. Becker, J. T. Dudley, and B. A. Kidd. 2017. "Automated Cell Type Discovery and Classification Through Knowledge Transfer." *Bioinformatics* 33, no. 11: 1689–1695. https://doi.org/10.1093/bioinformatics/btx054.

Lee, S. X. 2019. "CytoFA: Automated Gating of Mass Cytometry Data via Robust Skew Factor Analzyers." In *Advances in Knowledge Discovery and Data Mining*, edited by Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, and S.-J. Huang, 514–525. Springer International Publishing. https://doi.org/10.1007/978-3-030-16148-4_40.

Lee, S. X., and G. J. McLachlan. 2016. "Finite Mixtures of Canonical Fundamental Skew t-Distributions." *Statistics and Computing* 26, no. 3: 573–589. https://doi.org/10.1007/s11222-015-9545-x.

Lee, S. X., and G. J. McLachlan. 2022. "An Overview of Skew Distributions in Model-Based Clustering." *Journal of Multivariate Analysis* 188: 104853. https://doi.org/10.1016/j.jmva.2021.104853.

Levine, J. H., E. F. Simonds, S. C. Bendall, et al. 2015. "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-Like Cells That Correlate With Prognosis." *Cell* 162, no. 1: 184–197. https://doi.org/10.1016/j.cell.2015.05.047.

Liechti, T., L. M. Weber, T. M. Ashhurst, et al. 2021. "An Updated Guide for the Perplexed: Cytometry in the High-Dimensional Era." *Nature Immunology* 22, no. 10: 1190–1197. https://doi.org/10.1038/s41590-021-01006-z.

Lin, L., and B. P. Hejblum. 2021. "Bayesian Mixture Models for Cytometry Data Analysis." *WIREs Computational Statistics* 13, no. 4: e1535. https://doi.org/10.1002/wics.1535.

Lin, L., and J. Li. 2017. "Clustering With Hidden Markov Model on Variable Blocks." *Journal of Machine Learning Research* 18, no. 110: 1–49.

Lin, T. I. 2009. "Maximum Likelihood Estimation for Multivariate Skew Normal Mixture Models." *Journal of Multivariate Analysis* 100, no. 2: 257–265. https://doi.org/10.1016/j.jmva.2008.04.010.

Liu, X., W. Song, B. Y. Wong, et al. 2019. "A Comparison Framework and Guideline of Clustering Methods for Mass Cytometry Data." *Genome Biology* 20, no. 1: 297. https://doi.org/10.1186/s13059-019-1917-7.

Lo, K., R. R. Brinkman, and R. Gottardo. 2008. "Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering." *Cytometry, Part A* 73A, no. 4: 321–332. https://doi.org/10.1002/cyto.a.20531.

Lo, K., and R. Gottardo. 2012. "Flexible Mixture Modeling via the Multivariate t Distribution With the Box-Cox Transformation: An Alternative to the Skew-t Distribution." *Statistics and Computing* 22, no. 1: 33–52. https://doi.org/10.1007/s11222-010-9204-1.

Lock, M. 2003, March 8. "A Density-based Clustering Method for Flow Cytometry Data. JSM 2003 Online Program." https://ww2.amstat.org/meetings/jsm/2003/onlineprogram/index.cfm?fuseaction=abstract_details&abstractid=302400.

van der Maaten, L., and G. Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 9, no. 86: 2579–2605.

McInnes, L., J. Healy, and J. Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv:1802.03426 [Cs, Stat]. http://arxiv.org/abs/1802.03426.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake. 2019. "Finite Mixture Models." *Annual Review of Statistics and Its Application* 6, no. 1: 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325.

McLachlan, G. J., and D. Peel. 1998. "Robust Cluster Analysis via Mixtures of Multivariate t-Distributions." In *Advances in Pattern Recognition*, edited by A. Amin, D. Dori, P. Pudil, and H. Freeman, 658–666. Springer. https://doi.org/10.1007/BFb0033290.

McLachlan, G. J., and D. Peel. 2000. *Finite Mixture Models*. Wiley. https://doi.org/10.1002/0471721182.

McNicholas, P. D. 2016. "Model-Based Clustering." *Journal of Classification* 33, no. 3: 331–373. https://doi.org/10.1007/s00357-016-9211-9.

McNicholas, P. D., and T. B. Murphy. 2008. "Parsimonious Gaussian Mixture Models." *Statistics and Computing* 18, no. 3: 285–296. https://doi.org/10.1007/s11222-008-9056-0.

Melnykov, V. 2013. "Challenges in Model-Based Clustering." *WIREs Computational Statistics* 5, no. 2: 135–148. https://doi.org/10.1002/wics.1248.

Minoura, K., K. Abe, Y. Maeda, H. Nishikawa, and T. Shimamura. 2019. "Model-Based Cell Clustering and Population Tracking for Time-Series Flow Cytometry Data." *BMC Bioinformatics* 20, no. Suppl 23: 633. https://doi.org/10.1186/s12859-019-3294-3.

Minoura, K., K. Abe, Y. Maeda, H. Nishikawa, and T. Shimamura. 2021. "CYBERTRACK2.0: Zero-Inflated Model-Based Cell Clustering and Population Tracking Method for Longitudinal Mass Cytometry Data." *Bioinformatics* 37, no. 11: 1632–1634. https://doi.org/10.1093/bioinformatics/btaa873.

Murphy, K., C. Viroli, and I. C. Gormley. 2020. "Infinite Mixtures of Infinite Factor Analysers." *Bayesian Analysis* 15, no. 3: 937–963. https://doi.org/10.1214/19-BA1179.

Naim, I., S. Datta, J. Rebhahn, J. S. Cavenaugh, T. R. Mosmann, and G. Sharma. 2014. "SWIFT—Scalable Clustering for Automated Identification of Rare Cell Populations in Large, High-Dimensional Flow Cytometry Datasets, Part 1: Algorithm Design." *Cytometry, Part A* 85, no. 5: 408–421. https://doi.org/10.1002/cyto.a.22446.

Neal, R. M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics* 9, no. 2: 249–265. https://doi.org/10.1080/10618600.2000.10474879.

O'Hagan, A., T. B. Murphy, and I. C. Gormley. 2012. "Computational Aspects of Fitting Mixture Models Via the Expectation–Maximization Algorithm." *Computational Statistics & Data Analysis* 56, no. 12: 3843–3864. https://doi.org/10.1016/j.csda.2012.05.011.

O'Hagan, A., T. B. Murphy, I. C. Gormley, P. D. McNicholas, and D. Karlis. 2016. "Clustering With the Multivariate Normal Inverse Gaussian Distribution." *Computational Statistics & Data Analysis* 93: 18–30. https://doi.org/10.1016/j.csda.2014.09.006.

Orlova, D. Y., S. Meehan, D. Parks, et al. 2018. "QFMatch: Multidimensional Flow and Mass Cytometry Samples Alignment." *Scientific Reports* 8, no. 1: 1. https://doi.org/10.1038/s41598-018-21444-4.

Pearson, K. 1894. "Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences* 185: 71–110.

Pyne, S., X. Hu, K. Wang, et al. 2009. "Automated High-Dimensional Flow Cytometric Data Analysis." *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 21: 8519–8524. https://doi.org/10.1073/pnas.0903028106.

Pyne, S., S. X. Lee, K. Wang, et al. 2014. "Joint Modeling and Registration of Cell Populations in Cohorts of High-Dimensional Flow Cytometric Data." *PLoS One* 9, no. 7: e100334. https://doi.org/10.1371/journal.pone.0100334.

Rijsbergen, C. J. V. 1979. *Information Retrieval*. Butterworths.

Saeys, Y., S. van Gassen, and B. N. Lambrecht. 2016. "Computational Flow Cytometry: Helping to Make Sense of High-Dimensional Immunology Data." *Nature Reviews Immunology* 16, no. 7: 449–462. https://doi.org/10.1038/nri.2016.56.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6, no. 2: 461–464. https://doi.org/10.1214/aos/1176344136.

Scrucca, L., and A. E. Raftery. 2015. "Improved Initialisation of Model-Based Clustering Using Gaussian Hierarchical Partitions." *Advances in Data Analysis and Classification* 9, no. 4: 447–460. https://doi.org/10.1007/s11634-015-0220-z.

Shekhar, K., P. Brodin, M. M. Davis, and A. K. Chakraborty. 2014. "Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 1: 202–207. https://doi.org/10.1073/pnas.1321405111.

Sörensen, T., S. Baumgart, P. Durek, A. Grützkau, and T. Häupl. 2015. "immunoClust—An Automated Analysis Pipeline for the Identification of Immunophenotypic Signatures in High-Dimensional Cytometric

Datasets." *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* 87, no. 7: 603–615. https://doi.org/10.1002/cyto.a.22626.

Stephens, M. 2000. "Dealing With Label Switching in Mixture Models." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 62, no. 4: 795–809. https://doi.org/10.1111/1467-9868.00265.

Stubbington, M. J. T., O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann. 2017. "Single-Cell Transcriptomics to Explore the Immune System in Health and Disease." *Science* 358, no. 6359: 58–63. https://doi.org/10.1126/science.aan6828.

The FlowCAP Consortium, The DREAM Consortium, N. Aghaeepour, et al. 2013. "Critical Assessment of Automated Flow Cytometry Data Analysis Techniques." *Nature Methods* 10, no. 3: 228–238. https://doi.org/10.1038/nmeth.2365.

Wade, S., and Z. Ghahramani. 2018. "Bayesian Cluster Analysis: Point Estimation and Credible Balls (With Discussion)." *Bayesian Analysis* 13, no. 2: 559–626. https://doi.org/10.1214/17-BA1073.

Wattenberg, M., F. Viégas, and I. Johnson. 2016. "How to Use t-SNE Effectively." *Distill* 1, no. 10: e2. https://doi.org/10.23915/distill.00002.

Weber, L. M., and M. D. Robinson. 2016. "Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data." *Cytometry, Part A* 89, no. 12: 1084–1096. https://doi.org/10.1002/cyto.a.23030.

Weber, L. M., and C. Soneson. 2019. "HDCytoData: Collection of High-Dimensional Cytometry Benchmark Datasets in Bioconductor Object Formats." *F1000Research* 8: 1459. https://doi.org/10.12688/f1000research.20210.2.

Zhu, X., and V. Melnykov. 2018. "Manly Transformation in Finite Mixture Modeling." *Computational Statistics & Data Analysis* 121: 190–208. https://doi.org/10.1016/j.csda.2016.01.015.