Bayesian mixture models for cytometry data analysis

Lin Lin, Boris Hejblum

Bayesian Mixture Models for Cytometry Data Analysis

Lin Lin[*], Boris P Hejblum[†‡]

Article Type:

Advanced Review

Abstract

Bayesian mixture models are increasingly used for model-based clustering and the follow-up analysis on the clusters identified. As such, they are of particular interest for analyzing cytometry data where unsupervised clustering and association studies are often part of the scientific questions. Cytometry data are large quantitative data measured in a multi-dimensional space that typically ranges from a few dimensions to several dozens, and which keeps increasing due to innovative high-throughput biotechonologies. We present several recent parametric and nonparametric Bayesian mixture modeling approaches, and describe advantages and limitations of these models under different research context for cytometry data analysis. We also acknowledge current computational challenges associated with the use of Bayesian mixture models for analysing cytometry data, and we draw attention to recent developments in advanced numerical algorithms for estimating large Bayesian mixture models, which we believe have the potential to make Bayesian mixture model more applicable to new types of single-cell data with higher dimensions.

[*]Department of Statistics, Pennsylvania State University

[†]University of Bordeaux, INSERM Bordeaux Population Health Research Center U1219, INRIA SISTM, F-33000 Bordeaux, France.

[‡]Vaccine Research Institute, F-94000 Créteil, France.

1 Introduction

The investigation of single cell biology is crucial for improving our understanding of the immune system (De Rosa, Herzenberg, Herzenberg, & Roederer, 2001; Perfetto, Chattopadhyay, & Roederer, 2004; Stubbington, Rozenblatt-Rosen, Regev, & Teichmann, 2017). Over the past few decades, flow cytometry (FCM) has become one of the most widely used techniques for single cell measurements in many immunological studies and clinical trials. This is due to its ability to quantitatively monitor complex cellular immune responses, such as cell phenotype, activation or maturation status, intracellular cytokine or other effector molecule concentrations. This cellular information is critical for the understanding of the immune system, for the development of effective vaccines, and for the discovery of diagnostic or prognostic biomarkers in clinical trials (Darrah et al., 2007; Corey et al., 2015; Lin et al., 2015; Seshadri et al., 2015). Historically, the blood cells were evaluated manually using microscope. The flow cytometer – invented by Mack Fulwyler (Fulwyler, 1965) thanks to Wallace H. Coulter's earlier breakthroughs (Robinson, 2005, 2013) – made a revolution by combining optical and computer techniques to automatically measure a tremendous amount of cells in a sample within a very short period of time. Briefly, FCM is a high-throughput, laser-based single-cell technique for measuring the individual cell surface and intracellular marker molecules. The cell sample (typically the blood or tissue sample) is first stained with one or more fluorochromes that have been made specific to the cell surface or intracellular proteins of interest; also known as markers. Then FCM measures the cell light scattering and fluorescent intensities. The former provides information about the cell size and complexity, and the latter are related with the amount of fluorochrome found in the cell or attached at its surface. The higher the fluorescent intensities, the more expressed the corresponding molecular marker. An introduction for FCM can be found in Shapiro, 2005 for instance.

One of the fundamental uses of FCM is the identification and quantification of distinct cell subsets with phenotypes characterized by the density of cell surface and intracellular markers (Cossarizza et al., 2017). The recent technological advancements allow FCM to measure up to 28 fluorochromes simultaneously on a single cell (Mair & Prlic, 2018; Nettey, Giles, & Chattopadhyay, 2018), with the promise of soon reaching 50 parameters (BD Biosciences–

US, 2019). Meanwhile, Cytometry by Time-Of-Flight mass spectrometry (CyTOF), a new concurrent technology to FCM, that is also called Mass Cytometry and which is based on ion counts, has been developed and could in theory measure up to 100 different cellular markers at once (Nowicka et al., 2017). Combining many different cell surface and intracellular marker measurements is critical for identifying cellular populations: the cell subsets identified through FCM can then be tested for their functional properties. For example, the earliest uses of FCM helped to identify major cell lineages, such as T and B cells which play a fundamental role in the immune system. As FCM now allows more and more markers to be measured, a higher resolution of immune cells profiling can be achieved. For example, we now realize that T cells can be further distinguished into regulatory T cells, follicular helper T cells, and natural killer T cells, only to name a few. In most studies, the sample sizes of FCM data are large, reaching several million of cells being processed from one blood draw (or other biological tissues), although, in many cases the cell subsets of interest are typically in low frequencies (e.g. $\sim$0.01% of total cells). Hence, there is a need for detecting cell heterogeneity, and especially very low frequency cell subsets, for downstream analysis such as association studies to help understand the link between cellular heterogeneity and disease progression.

Several notable approaches have stood out for cell subsets identification ranging from manual gating to mixture model-based clustering. The manual gating method is a manual process that uses expert knowledge about the lineage, maturation and activation of cells (e.g., Roederer, Brenchley, Betts, & De Rosa, 2004; Perfetto et al., 2004) to manually delineates cells into sequential bounded regions (called gates) on 1-D histogram or 2-D scatter plots pseudo-colored by density. Cells within the region defined by the gates are identified as a specific cell subset. A simplified example to illustrate this sequential process is the task of discriminating CD4+ T cells, which is a type of T cells particularly important in the adaptive immune system. A sequence of subsetting procedures could be performed. Two physical markers, forward and side light-scatter, are first used to construct a 2-D scatter plot for distinguishing lymphocytes from all the live cells. Lymphocytes can then be further partitioned based on 3 fluorescence parameters: CD3, CD4 and CD8 cell-surface markers. CD4+ T cells are the subclass of lymphocytes having high values of CD3 and CD4 but low

value of CD8. In the case of markers for lineage, activation, exhaustion and function, it is common to dichotomize cells being positive (+) and negative (-) for each marker (driven by the underlying absence or presence of the cell functionality associated with this marker), based on an appropriate negative control or in some cases by eyeballing the data. In Boolean gating, cell subsets are then defined as all possible binary combinations of these markers. Hence, if the number of markers is d, then in theory 2^d number of cell subsets could be defined. In practice, many of such cell subsets are empty.

Despite its popular usage in the analysis of (low-dimensional) cytometry data, manual gating has serious limitations including being heavily relying on local expertise, time consuming, hard to reproduce, and cumbersome in analyzing higher dimensions since the number of possible 1-D and/or 2-D projections that need to be examined increases rapidly. This partly underlies the drive for automatic cell subset identification to overcome the limitations of manual gating. In particular, the approach of using statistical mixture modeling is increasingly adopted (e.g. Chan et al., 2008; Lo, Brinkman, & Gottardo, 2008; Finak, Bashashati, Brinkman, & Gottardo, 2009; Pyne et al., 2009), as it is a major statistical framework for performing density estimation and model-based clustering (Bouveyron, Celeux, Murphy, & Raftery, 2019). For FCM, a sequence of compensation, transformation, and standardization of the raw data are usually performed before fitting any statistical model (Finak, Perez, Weng, & Gottardo, 2010; O'Neill, Aghaeepour, Špidlen, & Brinkman, 2013). A general finite mixture model has the density function:

$$g(x|\theta) = \sum_{k=1}^{K} \pi_k f(x|\theta_k), \tag{1}$$

where $x \in \mathcal{R}^d$ is a random vector of length d representing a single cell with d measured markers, π_k is the mixture component probability with the constraint that $\sum_k \pi_k = 1$, and $f(\cdot|\theta_k)$ denotes the multivariate density function parameterized by θ_k for the k^{th} mixture component. The Gaussian distribution is commonly used as the base density, but skewed and heavy-tailed distributions, such as the (skew) t-distribution (Azzalini, Browne, Genton, & McNicholas, 2016), can be applied directly on the un-transformed data which can better accomodate outliers. Figure 1 displays an example of the mixture of two Gaussian distributions.

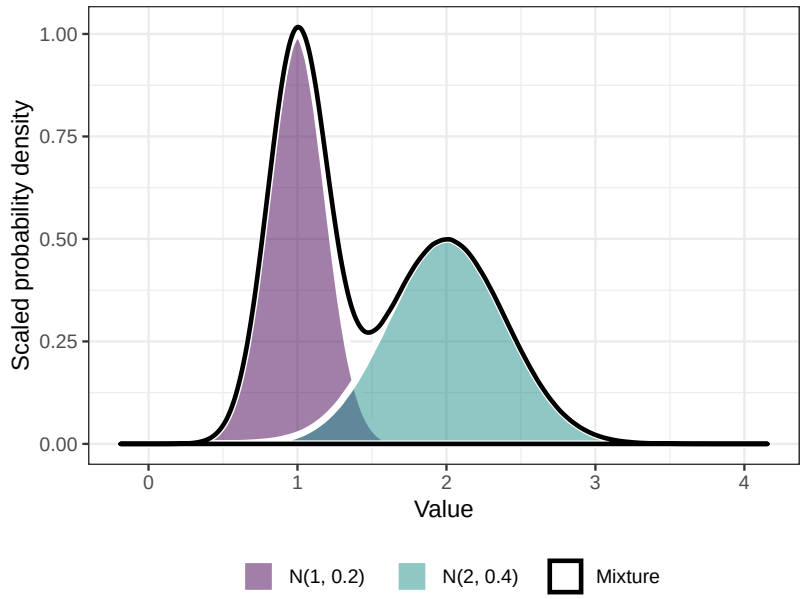

Figure 1: **An example of a mixture of two Gaussian distributions**

The number of mixture components K is typically estimated through model selection, e.g. by minimizing Bayesian information criterion (BIC) or Akaike information criterion (AIC) (Burnham & Anderson, 2004). The mixture model can be interpreted as arising from a clustering procedure by introducing a latent indicator $z \in \{1, 2, ..., K\}$ with $P(z = k) = \pi_k$, for $k = 1, ..., K$. Conditioning on $z = k$, x follows the k^{th} component distribution. Hence, Eq. (1) can be equivalently represented as:

$$g(x|\theta, z = k) = f(x|\theta_k), \; P(z = k) = \pi_k. \tag{2}$$

If we marginalize over z, the original mixture formulation in Eq. (1) is recovered. z can be interpreted as the mixture component identity of x. Thus z can be used for clustering by computing the posterior probability $P(z = k|x)$ and assigning x to the component with the maximum posterior probability. However, this approach by simply equating cluster to mixture component is inadequate when modeling clusters of arbitrary shapes. One major strategy developed in the literature is to perform clustering by mode association. The idea is to merge multiple mixture components for a better and more flexible representation of any individual cluster (e.g., Li, Ray, & Lindsay, 2007; Chan et al., 2008; Finak et al., 2009; Lin, Chan, & West, 2016). Despite the popularity of standard mixture models, there is an

increasing need to design and implement novel mixture models that can flexibly and robustly fit data, in addition to identifying biologically relevant but low probability component structures that deviate from the bulk of the data. Bayesian mixture models have demonstrated not only their abilities in accurately quantifying cell subset populations, but also their flexibility in modeling specific features arising from different contexts (e.g., Lin, 2012; Lin & Chan, 2017).

Thanks to biotechnology advances, FCM data keep getting bigger. In studies where several hundreds of samples are collected (for instance, the DALIA clinical trial for a therapeutic vaccine against HIV collected more than 4,000 tubes by monitoring 19 HIV positives individuals over 18 time points and processing 13 different cellular tubes at each time-points (Lévy et al., 2014)) and several hundreds of thousands of cells are available per tube, manual gating can take several weeks or even months. Especially, as more and more markers can be measured at once, the data dimension is also increasing, allowing for a theoretical 2^{30} (more than a billion) cellular populations to be isolated from 30 markers if considered binary (i.e. if each marker is considered intrinsically either functionally present or absent for each cell). While this amount of data calls for an automated processing, their sheer size and dimension can make this computationally challenging. In particular, estimation of complex Bayesian models, such as mixture models, is often done using simulation-based inference approaches, and in particular using Markov Chain Monte Carlo (MCMC) algorithms (Turkman, Paulino, & Müller, 2019). MCMC are sampling algorithms which target the posterior distribution. In Bayesian mixture models, they can approximate the posterior mixing distribution and provide a clustering estimate. In the context of FCM however, because of the size and dimension of the data, specific algorithmic strategies must be developed to explore the model space and overcome numerical difficulties (Lin et al., 2013; Hejblum, Alkhassim, Gottardo, Caron, & Thiébaut, 2019).

The goal of this review is to guide the reader towards effective understanding and use of core Bayesian mixture methods. We first introduce the standard Bayesian finite mixture model. The goal is to provide the general Bayesian technical context. We then provide the general Bayesian nonparametric mixture model framework. In that context, we examine the different modeling methodologies that can be used for characterizing FCM data. Finally,

we describe the numerical challenges associated with estimation of large Bayesian mixture models for cytometry data.

2 Bayesian finite mixture models

The idea of Bayesian modeling is to treat every parameter θ from any probabilistic model as a random variable with a prior distribution denoted by $P(\theta)$. Such $P(\theta)$ could be either informed by domain knowledge and/or historical data (referred to as "informative prior") or left diffuse (referred to as "weakly-informative prior"). Given data x, the prior distribution for θ can be updated and form the posterior distribution using Bayes' theorem:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}. \tag{3}$$

In most cases, the posterior distribution is typically computed through MCMC algorithms, which generate samples from the posterior distribution.

We start by providing a detailed example to illustrate Bayesian modeling with a finite Gaussian mixture model (GMM). Let X be a matrix of size $n \times d$ containing the FCM measurements from one sample, where n is the total number of cells measured in one sample and d is the number of markers measured. Let $X_i = (X_{i1}, ..., X_{id})$ denote the i^{th} row of X, which is the $d-$dimensional vector containing the i^{th} cell measurements. In this review, we assume the FCM data are already pre-processed. Following the notations in Eq. (1), the finite GMM with K components may be used for modeling each X_i, $i = 1, ..., n$, independently as:

$$X_i \sim \sum_{k=1}^{K} \pi_k f(X_i|\theta_k) = \sum_{k=1}^{K} \pi_k N(X_i|\mu_k, \Sigma_k), \tag{4}$$

where the general density form $f(\cdot|\theta_k)$ is replaced by the multivariate normal density commonly denoted as $N(\cdot|\theta_k)$, $\theta_k = (\mu_k, \Sigma_k)$ with μ_k being the d-dimensional mean vector and Σ_k the $d \times d$ covariance matrix.

GMM can be viewed as a generative model through the following Bayesian formulation:

$$\begin{aligned} \Sigma_k &\sim IW(\nu, \Phi), \\ \mu_k|\Sigma_k &\sim N(m, \lambda\Sigma_k), \\ (\pi_1, ..., \pi_d) &\sim \text{Dir}(\alpha/K, ..., \alpha/K), \\ P(z_i = k) &= \pi_k, \\ X_i|z_i = k &\sim N(X_i|\mu_k, \Sigma_k), \end{aligned} \tag{5}$$

where we assume a missing (i.e. unobserved) indicator $z = (z_1, ..., z_n)$ for $(X_1, ..., X_n)$. Each latent z_i indicates the mixture component from which the corresponding X_i has been generated. The set of parameters associated with GMM $\{\pi_{1:K}, \mu_{1:K}, \Sigma_{i:K}\}$ are generated from their corresponding conjugate priors: $\text{Dir}(\alpha/K, ..., \alpha/K)$ is a symmetric Dirichlet distribution, where α is a positive constant. The component specific mean μ_k and covariance matrix Σ_k are jointly distributed according to a (conjugate) Normal-inverse-Wishart distribution. $\alpha, m, \lambda, \nu, \Phi$ are hyper-parameters that may be either fixed or have their own prior distributions. Such modeling approach can be easily extended to other mixture distributions. For example, Frühwirth-Schnatter & Pyne, 2010 developed a Bayesian framework for mixtures of multivariate skew normal and skew t-distributions for the automated analysis of FCM data. The use of skewed distributions can result in a more parsimonious mixture model as fewer mixture components may be needed to fit the data. On the other hand, GMM may be more appropriate in fitting rare cell subsets.

Finite mixture model often served as a building block to form a Bayesian hierarchical model for more flexible analysis of FCM data: another usage of mixture model is to form a mixture of mixture distribution, so that cluster distributions can be more flexibly modeled (Malsiner-Walli, Frühwirth-Schnatter, & Grn, 2017). For a FCM sample X, let

$$X_i \sim \sum_{k=1}^{K} \eta_k f_k(X_i|\theta_k), \text{ where } f_k(X_i|\theta_k) = \sum_{l=1}^{L} w_{kl} N(X_i|\mu_{kl}, \Sigma_{kl}), \tag{6}$$

so that each vector X_i are drawn independently from a $K-$component mixture distribution with each component distribution being a $L-$component GMM. A special hierarchical prior is needed to prevent the non-identifiability of such mixture of mixture model, as there are

$(K \times L)!$ ways to permute the component labels that give rise to K different cluster densities without ever changing the likelihood. Such priors are constructed as follows:

$$\eta, \theta_1, ..., \theta_K \sim \text{Dir}_K(\eta|e_0) \prod_{k=1}^{K} \{\text{Dir}_L(w|d_0) N(\mu_{kl}|b_{0k}, \sqrt{\Lambda_k} B_0 \sqrt{\Lambda_k}) \text{IW}(\Sigma_{kl}|c_0, C_{0k})\}. \quad (7)$$

First, the block independence prior structure for θ_k ensures that the $K-$component cluster distributions of the upper level mixture (6) are invariant to permutations. Second, the hierarchical "random effects" prior is formulated to capture the dependence among K clusters: cluster-specific random hyper-parameters (C_{0k}, b_{0k}) and $\Lambda_k = \text{diag}(\Lambda_{k1}, ..., \Lambda_{Kd})$, are generated independently for each $k = 1, ..., K$ from a set of three independent base distributions: Wishart, multivariate normal and gamma distributions for C_{0k}, b_{0k} and Λ_k, respectively. The parameters associated with lower level mixture (μ_{kl}, Σ_{kl}) are then generated conditionally on the cluster-specific hyper-parameters. Such hierarchical Bayesian mixture models can also be used to model several FCM samples at once. For example Johnsson, Wallin, & Fontes, 2016 developed BayesFlow, a pipeline which allows the joint modeling of multiple FCM samples so that relations between samples can be more properly modeled. Let Y_{ij} denote vector valued measurements for the i^{th} cell in the FCM sample j, $i = 1, ..., n_j$ and $j = 1, ..., J$, with n_j the number of cells in sample j, and J the total number of samples. Let

$$Y_{ij} \sim \sum_{k=1}^{K} \pi_{jk} N(Y_{ij}|\mu_{jk}, \Sigma_{jk}) + \pi_{j0} N(Y_{ij}|\mu_0, \Sigma_0), \quad (8)$$

where the last component has identical parameters μ_0 and Σ_0 across samples and is used to model outliers, while the vector $\pi_j = (\pi_{j0}, ..., \pi_{jK})$ denotes the mixing proportions specific to the j^{th} sample. An additional latent layer is introduced to connect cell populations between samples by allowing (μ_{jk}, Σ_{jk}) to share a common Normal-inverse-Wishart distribution as in Model (5), and π_j share a common Dirichlet distribution. One issue that remains challenging when analysing multiple FCM samples together is cross-sample calibration, i.e. to align cell populations across biological samples, for comparative analysis. Various approaches have been proposed to i) either align data before (Hahne et al., 2010) or during clustering (Lee et al., 2016; Soriano & Ma, 2019; Freulon et al., 2020), ii) or post-process independent clustering results to align matching clusters (Courtot et al., 2014; Commenges et al., 2018). Hierarchical Bayesian mixture models provide a natural framework to jointly calibrate and

cluster FCM data as individual sample deviation from a common can be allowed for each cluster through an additional hierarchical layer (Cron et al., 2013; Lee et al., 2016; Soriano & Ma, 2019; Gorsky, Chan, & Ma, 2020).

The standard MCMC sampling methods, such as the Gibbs sampler and the Metropolis-Hastings algorithm, are effective for finite Bayesian mixture model and the associated hierarchical model. However, to obtain a meaningful MCMC-based inference, especially the component-specific inference, it is necessary to account for the possible permutations of component labels during sampling – an issue known as label switching – and properly align the component labels across MCMC iterations in a procedure called relabeling (Jasra, Holmes, & Stephens, 2005). However, by considering partitions of the data instead of labelled clusterings (thanks to a label invariant loss function such as Binder's loss function (Binder, 1981)) one can effectively circumvent such a label switching problem (Caron, Neiswanger, Wood, Doucet, & Davy, 2017; Hejblum et al., 2019). A partition of the data can be characterized by a series of co-clustered observation sets, that sums up a clustering in a label-free manner. The posterior distribution of the partitions can then be characterized by a posterior similarity matrix, which is a symmetric $n \times n$ matrix of posterior probabilities of co-clustering for each observation pair.

3 Bayesian nonparametric mixture models

For finite mixture models, one major task is to estimate the number of mixture components K, typically through model selection approach. The nonparametric Bayesian version of the mixture model allows K to be estimated from the data by treating it directly as a random variable within the model. This is done through the use of a Dirichlet process (DP) prior on the mixing distribution, which can be understood as "distributions over distributions" (Ferguson, 1973; Antoniak, 1974). To view this, we can rewrite the finite mixture model in Eq. (1) as:

$$\sum_{k=1}^{K} \pi_k f(X_i|\theta_k) = \int f(X_i|\theta) dG(\theta), \tag{9}$$

where $G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta)$ is the discrete mixing distribution. $\delta_{\theta_k}(\theta)$ is a Dirac delta function equal to 1 if $\theta = \theta_k$ and 0 otherwise. But if we treat $G(\cdot)$ as a random distribution instead, and use DP as a prior for $G(\cdot)$, the right side of Eq. (9) then results in a Dirichlet process mixture model (DPMM) with a countable number of mixtures, which has the following hierarchical form:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0), \\ \theta_i | G &\sim G, \\ X_i | \theta_i &\sim f(X_i | \theta_i), \end{aligned} \tag{10}$$

where each observation X_i is modeled by a distribution f parameterized by the observation-specific parameters θ_i, for $i = 1, ..., n$. Each parameter θ_i is then independently and identically drawn from a discrete distribution G, and DP is used as a prior over G. The DP has two parameters denoted by α and G_0. The concentration parameter α is a positive-valued scalar that directly controls the number of non-empty components of the mixture model: the larger α, the more (non empty) mixture components. G_0 is the base distribution where the random draws G will be centered. Typically, G_0 is chosen to be conjugate to the parametric distribution f, and α can be sampled from a Gamma hyper-prior (Escobar & West, 1995). This is an important modeling feature that ensures that the number of non-empty components is properly estimated in a data-driven manner. The number of expected non-empty clusters a posteriori will grow at a logarithm rate with the amount of data (Teh, 2010). Miller & Harrison, 2014 have therefore argued that DPMM are unable to correctly recover the correct number of clusters in a dataset, because their posterior distribution is inconsistent for the number of non-empty clusters. However, that is only true if α is fixed beforehand. When α is rather treated as a parameter of the Bayesian model, e.g. with a weakly-informative prior, DPMMs are able to adequately recover the number of well separated clusters from a dataset (Hejblum et al., 2019 provide some evidence of that behavior using the augmentation sampling technique from Escobar & West, 1995). A random distribution that follows a DP prior can be expressed as an infinite sum $G(\cdot) = \sum_{k=1}^{\infty} \pi_j \delta_{(\theta_k^*)}(\cdot)$. Due to the discreteness of the resulting G, draws from G will have a positive probability to share the same values (Teh, 2010). Hence, the corresponding X_i's will share the same distribution, which gives rise to

the clustering property of the joint distribution of X_i.

There is an equivalent constructive definition of the DP, which allows to get actual draws from it. This is called the stick-breaking process (Sethuraman, 1994), and has the following hierarchical representation:

$$\begin{aligned} V_k &\sim Be(1, \alpha), \\ \pi_k &= V_k \prod_{l=1}^{k-1}(1 - V_l), \\ \theta_k &\sim G_0, \\ X_i &\sim \sum_{k=1}^{\infty} \pi_k f(X_i|\theta_k), \end{aligned} \tag{11}$$

where Be denotes the beta distribution. Figure 2 illustrate this stick breaking process construction of the DPMM.

Many algorithms have been developed for posterior sampling of the Gaussian DPMM where f is the probability density function of a multivariate normal distribution (e.g., Neal, 2000; Blei & Jordan, 2004; Daume III, 2007; Papaspiliopoulos & Roberts, 2008; Kalli, Griffin, & Walker, 2011). (Hejblum et al., 2019) provide a Gibbs sampler for a DPMM of skew Normal and skew t- distributions combining (Frühwirth-Schnatter & Pyne, 2010) and (Caron et al., 2017) implementations. However, there is a need for faster algorithms when analyzing large-scale FCM data (see section 5 for a discussion). On the other hand, one way to potentially reduce the computational complexity in fitting DPMMs is to use a truncated DP prior (Ishwaran & James, 2001). More specifically, G in model (11) is replaced by $G_J = \sum_{k=1}^{J} \pi_k \delta_{\theta_k}$, where J is some fixed (large) upper bound on the number of effective components to be determined before fitting the model. When J goes to infinity, G_J converges almost surely to $DP(\alpha, G_0)$. The standard blocked Gibbs sampler (Ishwaran & James, 2001) and Bayesian EM algorithm (Lin et al., 2016) can both be used to effectively estimate this truncated model, and can be efficiently implemented for either serial or parallel computation (Suchard et al., 2010).

While the DPMM is a natural extension of the finite mixture model for cell subsets identification, the dependent Dirichlet processes (extensions of the DP) are more powerful for modeling collection of distributions that may vary in time (e.g., for repeated FCM mea-

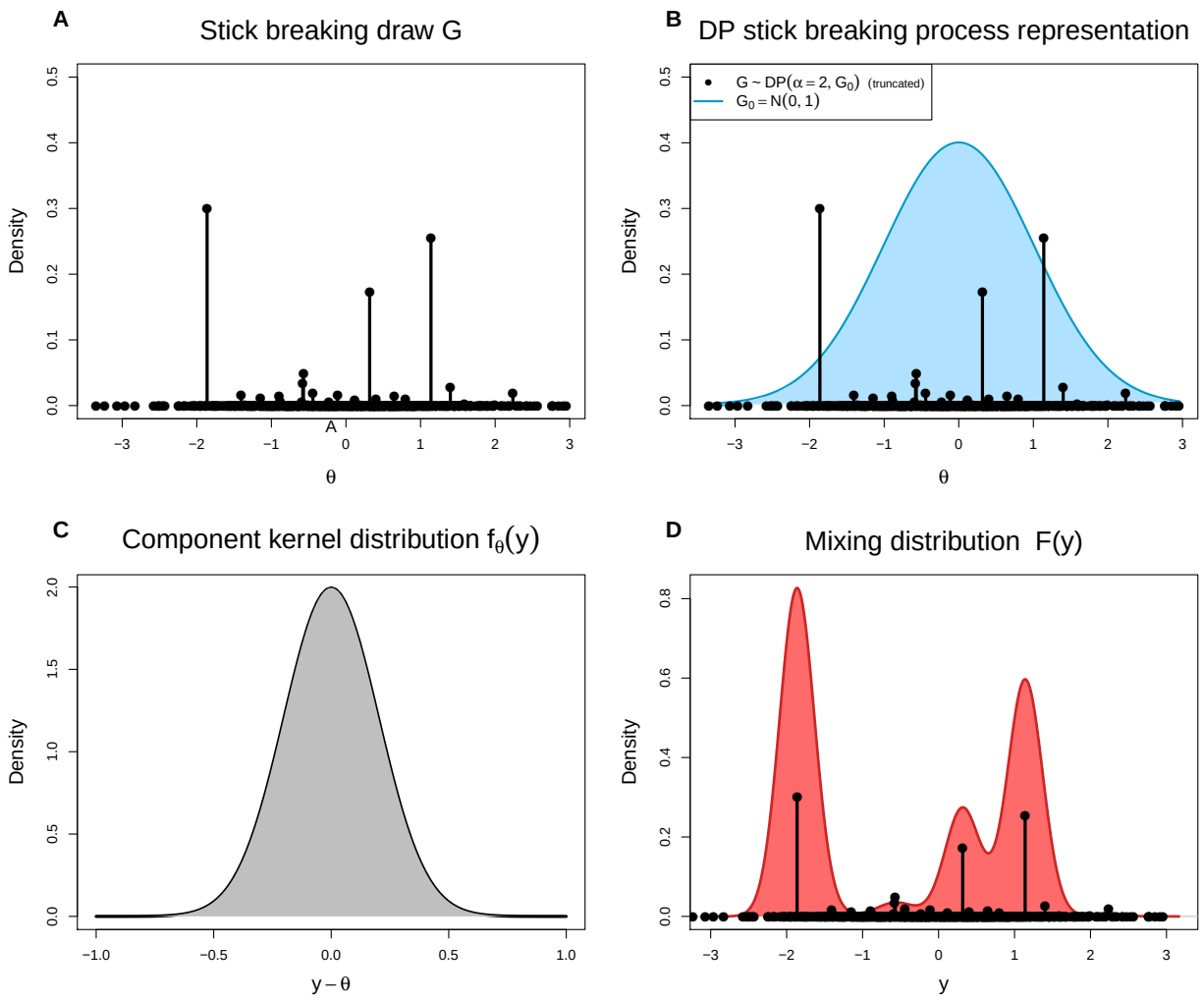

Figure 2: **Illustration of the Dirichlet process mixture model through the stick breaking process representation.** Panel **A** displays a truncated draw (only the first 1,000 most frequent θ values) from a stick breaking process using G_0 as its base distribution and a concentration parameter $\alpha = 2$ with the density probability of each θ value which can be interpreted as a given component frequency. Panel **B** overlays the base distribution G_0. Panel **C** display the kernel density function f_θ that represents the probability density distribution assumed for each component. Panel **D** displays the resulting mixture density that constitutes the DPMM.

surements), or in grouped data settings (e.g., across different FCM samples). For modeling multiple FCM samples, the hierarchical DP (HDP) prior, which is based on two levels of DP

priors, introduces a sharing mechanism across samples (Teh, Jordan, Beal, & Blei, 2005). By constraining mixture components to have the same parameters across all samples, while allowing variation in the component weights across samples, the resulting HDP mixture model is a compromise between the extremes of no sharing (all samples are treated independently) and complete sharing (all samples are assumed to come from the same distribution). Hence, HDP facilitates the identification of extremely rare cell subsets in some samples but less so in others. Such rare cell subsets might otherwise be missed when either fitting separate mixture models to each sample independently, or fitting all the samples together by a common mixture model.

Explicitly, the HDP model is defined for modeling the measurement of i^{th} cell in j^{th} FCM sample Y_{ij} as:

$$\begin{aligned} G_0 &\sim \text{DP}(\gamma, H), \\ G_j &\sim \text{DP}(\alpha_j, G_0), \\ \theta_{ij} &\sim G_j, \\ Y_{ij} &\sim f(Y_{ij}|\theta_{ij}), \end{aligned} \tag{12}$$

where each cell and sample-specific parameter θ_{ij} is independently and identically drawn from a sample-specific random distribution G_j. All the G_js follow a DP prior and share a common global random probability measure G_0 distributed as $DP(\gamma, H)$ while the concentration parameter α_j controls dispersion of the G_j around G_0 and is sample-specific. The baseline distribution H is the marginal prior distribution for θ_{ij}. The distribution G_0 varies around H, with the amount of variability controlled by γ. Cron et al., 2013 pioneered the use of HDP as a prior for GMM to jointly model multiple FCM samples, such hierarchical model enables the automatic alignment of cell subsets across different samples for comparative analysis.

For combinatorially encoded FCM samples which expand the number of antigen-specific T cells that can be measured (Hadrup & Schumacher, 2010), an alternative formulation of the HDP enables us to flexibly model the partitions on the cell's feature space for efficient quantification of multiple extremely rare antigen-specific T-cell subsets. The basic idea of combinatorial encoding is that by using multiple different fluorescent labels for any single epitope, many more types of antigen-specific T-cells can be identified by decoding the color

combinations of their bound multimer reporters. For example, using r colors, we can in principle encode $2^r - 1$ different epitope specificities. With such data sets, X_i now contains both functional FCM phenotypic markers and the light emitted by the fluorescent reporters of multimers binding to specific receptors on the cell surface. The direct application of standard DP mixture model fails to detect low probability antigen-specific T-cell subsets due to the masking by the large background data (Lin et al., 2013). To overcome such masking issue, Lin et al., 2013 proposed to partition X_i into two parts: $X_i = (b_i, t_i)$, where b_i is the subvector of phenotypic marker measurements and t_i is the subvector of fluorescent intensities of each of the multimers being reported via the combinatorial encoding strategy. Each measurement X_i can then be modeled hierarchically:

$$X_i \sim f(b_i\Theta)f(t_i|b_i,\Theta) = \sum_{j=1}^{J} \pi_j N(b_i|\mu_{bj},\mu_{bj},\Sigma_{bj})f(t_i|b_i,\Theta), \tag{13}$$

where Θ contains all relevant and needed parameters and the truncated Dirichlet process Gaussian mixture model is used to model b_i, which can be augmented by introducing the latent indicators z_{bi} for each b_i. This mixture model allows the understanding of substructure in the data reflecting differences in cell phenotypes. Based on their phenotypic markers, $f(t_i|b_i,\Theta)$ facilitates the understanding of subtypes within that, now based on multimer binding that defines finer substructure among T-cell features:

$$f(t_i|b_i,\Theta) = \sum_{k=1}^{K} w_{jk} N(t_i|\mu_{tk},\Sigma_{tk}), \tag{14}$$

where the distribution of t_i depends on the latent indicator z_{bi}. HDP is a natural choice of prior for the parameters in Eq. (14), as HDP can hierarchically model grouped data. Lin & Li, 2017 extend the two-subvector model to accommodate more than two groups of variables, where the sequential dependence of groups of variables are modeled by hidden Markov model. This new mixture model is called Hidden Markov Model on Variable Blocks (HMM-VB).

Under the situation where there are repeated measurements of the same individual, as is often the case in clinical trials, Hejblum et al., 2019 propose to use previous time points or previous samples results as prior information to leverage all the information available to estimate the mixture model. Suppose we have observations from the t^{th} and $t+1^{\text{th}}$ sample,

then the observation from the $t+1^{\text{th}}$ sample can be modeled conditioning on the t^{th} sample:

$$\begin{aligned} \alpha_{t+1}|Y_{1:n_t,t} &\sim Ga(\hat{a}_1, \hat{b}_1), \\ G|\alpha_{t+1}, Y_{1:n_t,t} &\sim DP(\alpha, \hat{G}_1), \\ \theta_{i,t+1} &\sim G, \\ Y_{i,t+1} &\sim f(Y_{i,t+1}|\theta_{i,t+1}), \end{aligned} \tag{15}$$

where the prior for the random measure G is conditionally dependent on the previous sample and Ga denotes the Gamma distribution. $\hat{a}_1$ and $\hat{b}_1$ are the maximum likelihood estimates from the MCMC samples α_t, and $\hat{G}_1$ is a parametric approximation of the posterior mixing distribution.

4 Case studies and practical challenges

In this section, we apply multiple Bayesian models introduced in previous Sections that have R packages available on either CRAN or Bioconductor to analyze a set of three FCM samples. These three samples corresponds to the first replicates of Standford center in the T-cell Lyoplate panel of the SeraCare cell HIPC study for each of the three available patients denoted "P1228R1", "P1349R1" and "P1369R1" (Maecker, McCoy, & Nussenblatt, 2012; Finak et al., 2016). The original raw FCS data files are available on the immunspace plateform at https://www.immunespace.org/project/HIPC/Lyoplate/begin.view?. The following preprocessing steps were applied before they were fitted to the models: i) first, removal of dead cells and doublets to focus only on manually gated cells ; ii) second, standardization of each cellular marker (i.e. features), so that each has a mean of 0 and a standard deviation of 1 – This facilitates the specification of hyper-parameters. After this pre-processing, the datasets P1228R1, P1349R1 and P1369R1 respectively contain $30,427$, $31,228$ and $32,948$ cells, characterized across 7 cellular markers (namely CCR7, CD4, CD45RA, CD3, HLADR, CD38, and CD8). In addition, we have a reference manual gating of those cells into 8 mutually exclusive populations (2 additional gated populations – namely "CD4 Activated" and "CD8 Activated" – overlap with the other cell populations and therefore were not considered). See Figure 3 for a descriptive representation of the 1228R1 sample (before

standardization).

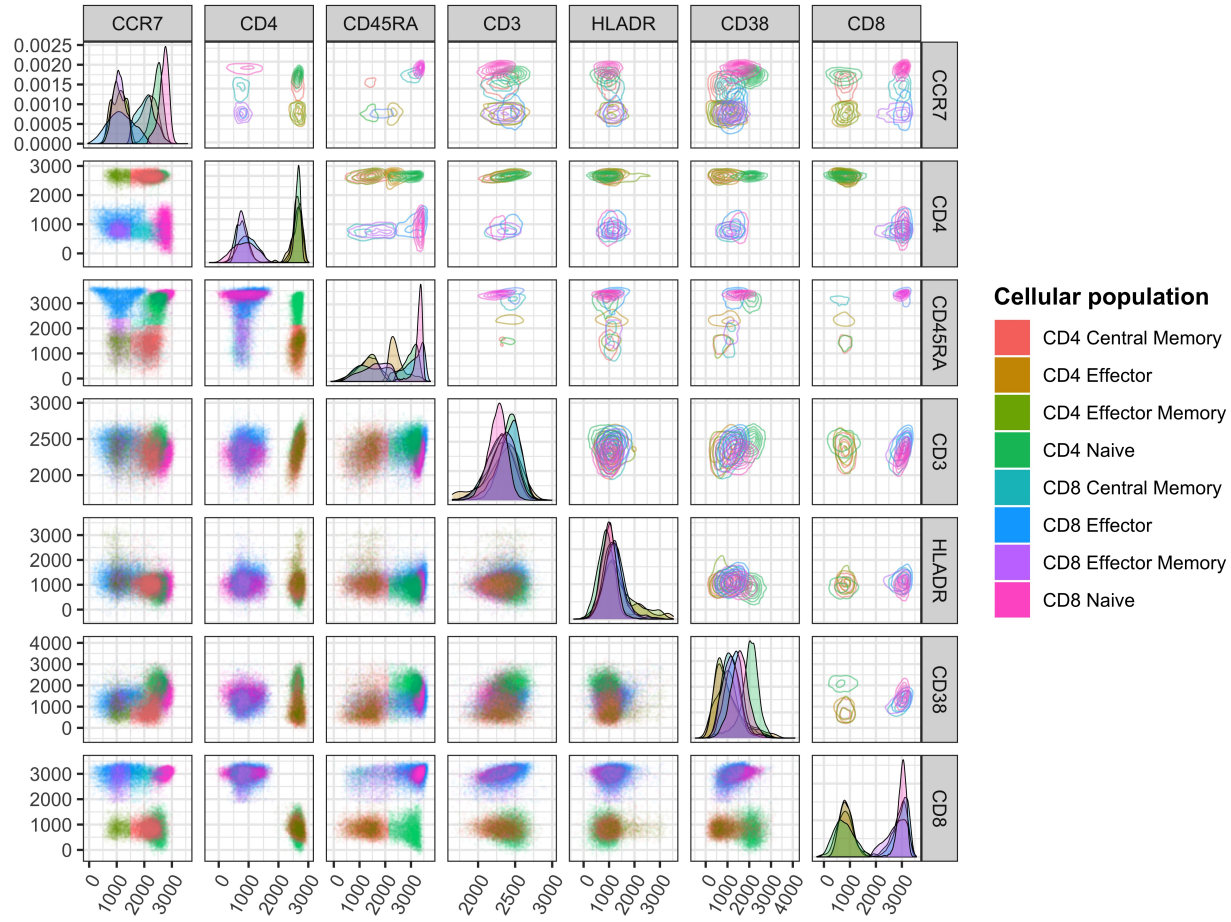

Figure 3: **Visualization of the data from one FCM sample with manual gating**. This FCM sample is the replicate 1 of patient 1228 processed at Standford from the T-cell panel in the HIPC Lyoplate study. 30,427 cells are displayed before standardization of the features. Diagonal plots represent marginal densities per cell population, lower triangle plots are 2D scatter plots of gated cells, and upper triangle plots represent bi-variate densities per cell population.

For clustering analysis, we use the F-measure (i.e. the harmonic mean of precision and recall) to compare any clustering result to manual gating, similarly as in Aghaeepour et al., 2013; Van Gassen et al., 2015; Weber & Robinson, 2016. An F-measure of 1 means the clustering result is a perfect reproduction of the manual gating result, and the worst value of F-measure is 0. The F-measure is computed using the `FMeasure` function from the R

package `FlowSOM` (Van Gassen et al., 2015). In this example section, we analyze all three FCM samples independently.

We first perform the simplest analysis by clustering each individual sample using GMM in Eq. (5). To simplify the analysis, we assume the number of clusters (components) is known according to the manual gating result. Thus, we specify $K = 8$ instead of searching for an optimal value using BIC. We use R package `bayesm` for fitting Bayesian GMM. One advantage of using Bayesian approach is its flexibility in incorporating prior knowledge into the modeling framework. Thus, we fit two Bayesian GMMs, one with non-informative prior and the other using (relatively) informative prior. For both priors, we let α equal to 8, $\nu = d + 2 = 9$, and m being a zero vector because of the normalization step. For the non-informative prior, we let $\lambda = 5$. We also specify $\Phi = \nu I$, which is similar to the default specification as in `bayesm`. Here I stands for the identity matrix. For the relatively informative prior, we specify $\lambda = 10$, as Fig. 3 shows some clusters are more separated from each other. We further specify $\Phi = 10\Sigma$, where Σ is the covariance matrix derived from the data. The MCMC computations were initialized based on the above specified prior distributions and run for a total of $20,000$ iterations. The analysis saved the last $10,000$ MCMC draws for summary inferences. Fig. 4 shows that the label switching issue of MCMC exists, posing difficulty for parameter estimation using posterior mean. In this analysis, instead of performing relabeling analysis, we use maximum a posterior (MAP) estimator to overcome the label switching problem. In addition, to further motivate the Bayesian modeling approach, we also fit a GMM in a frequentist approach by using the R package `mclust` (Scrucca, Fop, Murphy, & Raftery, 2016). The clustering results for the above three models compared with manual gating and their computational times are summarized in Table 1. For all three datasets, the Bayesian GMM using informative prior consistently outperform the other two models in terms of the F-measure.

The above analysis assumes that each cluster is normally distributed, which is a strong assumption and it is not fully supported by Fig. 3 while assuming the manual gating results are gold-standard. The mixture of mixture distribution is one approach to relax such distributional assumption. However, since there is no R package currently implementing the method proposed in Eq. (6), we present here the clustering results obtained by merging mul-

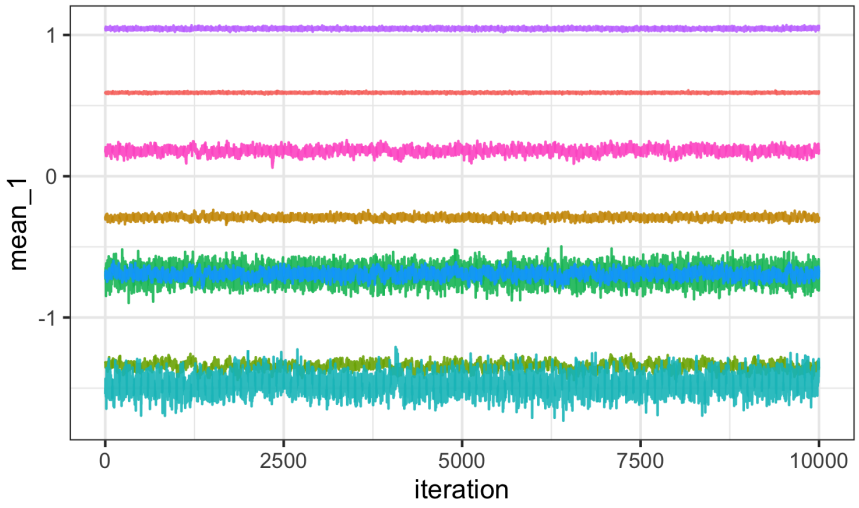

Figure 4: **Trace plot of posterior samples for the first dimension of component means of one FCM sample.**

Dataset	Non-informative GMM	Informative GMM	Frequentist GMM
P1228R1	0.793 (7.87 min)	0.825 (8.63 min)	0.771 (52.87 s)
P1349R1	0.881 (8.93 min)	0.917 (8.80 min)	0.756 (45.52 s)
P1369R1	0.727 (9.05 min)	0.771 (9.25 min)	0.680 (38.75 s)

Table 1: Comparison of the three models (Bayesian GMM using non-informative prior, informative prior and frequentist GMM) in terms of their clustering accuracy and computational time (numbers in parentheses).

tiple mixture components of a GMM into one cluster through mode association (denoted by modal GMM) using the R package `HDclust` (Lin & Li, 2017). In addition, motivated by the gold-standard manual gating analysis strategy where the cellular markers are divided into groups based on prior/biological information and examined sequentially, we further perform the analysis by partitioning the 7 markers into two groups with CD3, CD4 and CD8 in one group, and the rest in the second group. Eq. (14) is developed for such modeling task, however there is also no R package implementing this currently available. Instead, we run HMM-VB, which is a frequentist version of Eq. (14) by using the R package `HDclust`. The clustering results are summarized in Table 2. First, model GMM achieves higher F-measures

for all three data than GMM in Table 1. This suggests the advantage of relaxing normality assumption for each cluster. Second, HMM-VB achieves even better clustering performance than modal GMM, suggesting the advantage of leveraging sequential dependence among groups of variables in more effective clustering. Third, since HMM-VB does not consistently perform better than GMM with informative prior in Table 1, this suggests the need to have relevant R packages developed for Bayesian modeling approach such as Eq. (14).

Dataset	**Modal GMM**	**HMM-VB**
P1228R1	0.818 (10.50 s)	0.896 (17.02 s)
P1349R1	0.834 (12.68 s)	0.857 (18.26 s)
P1369R1	0.803 (11.66 s)	0.830 (24.06 s)

Table 2: Comparison of modal GMM and HMM-VB in terms of their clustering accuracy and computational time (numbers in parentheses).

In all above analyses, the number of clusters is considered known. However, in practice, the number of identifiable cellular populations in an FCM sample is not necessarily known, and DPMM presented in Eq. (10) estimate this number of clusters as an internal parameter. Table 3 shows the performance of such Bayesian nonparametric models on these 3 dataset. As expected given the shape of the data (visualized in Fig. 3), skew t mixtures are performing much better than Gaussian mixtures. Indeed, the latter largely over-estimate the number of clusters to fit the observed data, while skew t DPMMs are able to approximately recover the right number of clusters which then translates in higher concordance though the F-measure.

5 Estimation of large Bayesian mixture models

There can be up to a few million of cells from a simple blood draw. Most often, flow (or mass) cytometry analyses are performed on several hundred of thousand of Peripheral Blood Mononuclear Cells (PBMCs) from a single blood sample. This makes cytometry data relatively large data, especially when considering several samples in a given study. In addition, the dimension of newly generated cytometry data keeps increasing, from 16 to

Dataset	Gaussian DPM	Skew t DPM
P1228R1	0.543 (7.600 h) – 23 clusters	0.713 (4.87 h) – 6 clusters
P1349R1	0.567 (8.847 h) – 26 clusters	0.600 (4.95 h) – 8 clusters
P1369R1	0.540 (7.374 h) – 24 clusters	0.740 (5.89 h) – 8 clusters

Table 3: Comparison of Gaussian and skew t-distribution DPMM in terms of their clustering accuracy and computational time (numbers in parentheses). Markov chains were initialized with 30 non-empty clusters.

50 color channels available on recent flow-cytometers, while up to 100 markers can now be measured with CyTOF. This large size and increasing dimension results in numerical challenges for Bayesian mixture model estimation.

Bayesian mixture models, due to their complexity, are usually estimated through sophisticated MCMC algorithms (Marin, Mengersen, & Robert, 2005). However standard MCMC algorithms, such as the Gibbs sampler, are not very efficient for large data, especially with non-conjugate distributions such as t- or skew t-distributions which have been showed to better fit FCM data (due to the presence of heavy tails and skewness in cellular sub-populations) (Pyne et al., 2009; Hejblum et al., 2019). Gorsky et al., 2020 have proposed to apply recent developments in "coarsened" posterior inference (in short, using the power likelihood in the posterior numerator) to Bayesian mixtures models for FCM analysis as additional way of dealing with model mis-specification (Miller & Dunson, 2018). Current MCMC implementations of Bayesian mixture models requires sweeping over all data-points at each Markov iteration, and therefore computation time can become an issue with such large data. In addition, large dimensional spaces are challenging to explore through MCMC. Because of the curse of dimensionality, there is a rapid expansion of low probability density regions in the sampling space as the dimension increases. Thus, standard MCMC algorithms often feature mixing issues and can struggle to reach convergence. Recent innovations in Monte Carlo algorithm such as Sequential Monte Carlo (SMC) or Hamiltonian Monte Carlo (HMC) have focus on improving convergence and performances in high-dimensional settings (Turkman et al., 2019) but can also bear an important computational cost (Wang & Blei, 2019). van

Havre, White, Rousseau, & Mengersen, 2015 have proposed an approach using parallel tempering (running multiple Markov chains in parallel at different temperatures with moves that can be accepted from a different chain) in order to mitigate identifiability and mixing issues of current MCMC algorithms in the context of univariate Gaussian mixtures. Such approaches require additional computing power (in the form of additional parallel CPU threads) and have yet to be applied for multivariate mixtures in the context of Flow cytometry.

Meanwhile, scaling-up MCMC algorithms for large data remains an active research topic (Hastie, Liverani, & Richardson, 2015; Bardenet, Doucet, & Holmes, 2017; Robert, Elvira, Tawn, & Wu, 2018; Soriano & Ma, 2019). One of the solutions that has been largely considered is to use sub-sampling of the data. However, "there is no free lunch", and Bardenet et al., 2017 show that some information is then lost, which can be especially challenging for cytometry data where rare cell populations are of particular interest (Cron et al., 2013; Srivastava, Li, & Dunson, 2018). Other approaches relying on importance sampling seem promising, e.g. using coresets, but require a first estimate of the clustering structure that can be equally expensive to compute (Campbell & Broderick, 2019).

In addition, the number of clusters is a key parameter for interpreting the results of mixture models. Yet, in cytometry experiments the number of cell sub-populations is generally unknown, which means that the number of clusters in the fitted mixture model must then be estimated, a challenging task in practice (Cai, Campbell, & Broderick, 2017) that will have a non negligible computational cost. In the finite mixture model, the estimation of the number of mixture components is often done using model selection criteria, requireing several models to be estimated and compared, thus adding to the computational burden. In nonparametric mixtures, the number of non empty clusters is estimated directly within the model (in practice, state-of-the-art MCMC algorithms for DPMM rely on an data-augmentation scheme to efficiently sample the concentration parameter α with a weakly-informative hyperprior and tune it to the data (Escobar & West, 1995)). But because it is highly improbable to successfully introduce a stable non-empty new cluster at any MCMC iteration, MCMC algorithms for sampling DPMM must be initialized with much more clusters than anticipated in the data (Hastie et al., 2015), which also adds to the computational burden, especially during the first iterations.

Variational Bayes inference represents a scalable alternative to MCMC algorithms for approaching posterior distributions of Bayesian models (Blei, Kucukelbir, & McAuliffe, 2017). This optimization technique finds the distribution that minimizes the Kullback-Leibler divergence with the exact posterior distribution among a family of parametric densities. Counterbalancing its computational speed and scalability, variational Bayes inference currently suffers from three shortcomings: i) theoretical guarantees for the posterior approximation to be correct are only available in a few restrictive and parametric cases up until now (Chérief-Abdellatif & Alquier, 2018; Wang & Blei, 2019); ii) it has a tendency to underestimate the posterior variance (Blei et al., 2017; Wang & Blei, 2019); iii) it often requires more parametric assumptions and deeper mathematical analysis of the posterior to derive the optimization algorithm compared to a Gibbs sampler, and is not necessarily faster or more scalable (Lim & Wang, 2018).

6 Conclusions

Bayesian mixture models are powerful tools in analyzing cytometry data. In particular, structured hierarchical mixture models are increasingly needed to accommodate different research contexts. Moreover, as sample size increases, generic mixture models can often lack the ability to identify rare events due to the dominance of much of the data. One general technique for addressing such issue is to encourage a sparse mixture model structure. Given multiple data samples, hierarchical, or multi-level mixture models can represent individual observation in a data set as being organized into successively higher units. The critical idea is how to design the model so that it is able to borrow information across samples.

Identifying and quantifying cell populations is often only the first step in the analysis of cytopmetry data. Once cell subsets have been identified, "differential expression" analysis can be conducted to assess the significance of proportion variations in the cell subsets identified across different samples and conditions (e.g., stimulated vs. un-stimulated). Assume there are two conditions where u and s stands for un-stimulated and stimulated samples respectively. Each subject i ($i = 1, \ldots, I$) has the cell counts n_{cik} for each cell subset/category k ($k = 1, \ldots, K$), and c condition, $c = s, u$. It is of interest to identify which cell subsets

are differentially represented in terms of their proportions between the two conditions and for each subject. Such identified cellular heterogeneity could potentially be used as cellular biomarkers which can be associated with subject's outcome variable. Finak et al., 2013 developed a Bayesian hierarchical framework based on a beta-binomial mixture model, which is essentially a Bayesian version of Fisher's 2×2 test. While the model allows the inference to be subject specific, it only performs differential analysis on each cell subset one at a time. Hence, multiple comparisons across cell subsets have to be taken into account to adequately control the error rate. Lin et al., 2015 proposed a formal Bayesian hierarchical framework based on a structured Dirichlet multinomial mixture model. The developed model also uses a Bayesian variable selection approach to automatically identify relevant cell subsets, allows each subject to be responding to stimulation in none, some or all of the subsets, accounts for the different number of cells observed in the different samples, and jointly models all the subjects and cell subsets to allow information sharing to improve the power in detecting weak signals. To the best of our knowledge, the problem of performing differential expression analysis in longitudinal setting as well as across more than 2 conditions have not been studied in existing literature.

Because cytometry data are large data whose dimension keeps increasing, estimation of Bayesian mixture model on those data can be computationally challenging. In addition, the lack of prior knowledge about the number of clusters translates into additional computation time, regardless of whether a finite or a nonparametric mixture model is being estimated. In such case, non-probabilistic solutions, such as SWIFT (Naim, Datta, Sharma, Cavenaugh, & Mosmann, 2010) and cytometree (Commenges et al., 2018) can be worthwhile alternatives, as they run very efficiently and provide good performance for automatic gating. This constitutes yet another motivation for accelerating Bayesian mixture model estimation methods while the latest innovations such as the use of coresets are promising. While this review has focused on flow and mass cytometry data, thanks to new developments in microfluifdics and sequencing technologies a new kind of cytometry data is on the rise: single-cell RNA-sequencing (scRNA-seq) data. Those data share many of the challenging features of flow and mass cytometry data, but only exacerbated – namely a high proportion of zeros, a high-dimension (several thousands of genes measured), a large number of cells, and an unknown

number of clusters. Therefore, while Bayesian mixture models seem like a natural tool to analyze such data as well, their current numerical shortcomings might become critical to tackle those issues.

References

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., ... Consortium, T. D. (2013, Mar 01). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, *10*(3), 228-238.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian non-parametric problems. *The annals of statistics*, 1152–1174.

Azzalini, A., Browne, R. P., Genton, M. G., & McNicholas, P. D. (2016). On nomenclature for, and the relative merits of, two formulations of skew distributions. *Statistics and Probability Letters*, *110*, 201-206. doi: 10.1016/j.spl.2015.12.008

Bardenet, R., Doucet, A., & Holmes, C. (2017). On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research*, *18*(1), 1515–1557.

BD Biosciences–US. (2019). *Bd facsymphony flow cytometer.* Retrieved from http://www.bdbiosciences.com/us/instruments/research/cell-analyzers/bd-facsymphony/m/6022968/overview

Binder, M. A. (1981). Approximations to bayesian clustering rules. *Biometrika*, *68*, 275–285.

Blei, D. M., & Jordan, M. I. (2004). Variational methods for the dirichlet process. In *Proceedings of the twenty-first international conference on machine learning* (p. 12).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877. doi: 10.1080/01621459.2017.1285773

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in r* (Vol. 50). Cambridge University Press.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Cai, D., Campbell, T., & Broderick, T. (2017). Finite mixture models are typically inconsistent for the number of components. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference (AABI).*

Campbell, T., & Broderick, T. (2019). Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, *20*(1), 551–588.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., & Davy, M. (2017). Generalized pólya urn for time-varying pitman-yor processes. *Journal of Machine Learning Research*, *18*(27).

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., & Kepler, T. B. (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, *73*(8), 693–701.

Chérief-Abdellatif, B.-E., & Alquier, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, *12*(2), 2995-3035. doi: 10.1214/18-EJS1475

Commenges, D., Alkhassim, C., Gottardo, R., Hejblum, B. P., & Thiébaut, R. (2018). cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry: Part A*, *93*(11), 1132–1140. doi: 10.1002/cyto.a.23601

Corey, L., Gilbert, P. B., Tomaras, G. D., Haynes, B. F., Pantaleo, G., & Fauci, A. S. (2015). Immune correlates of vaccine protection against hiv-1 acquisition. *Science translational medicine*, *7*(310), 310rv7.

Cossarizza, A., Chang, H.-D., Radbruch, A., Akdis, M., Andrä, I., Annunziato, F., ... Zimmermann, J. (2017). Guidelines for the use of flow cytometry and cell sorting in immunological studies. *European Journal of Immunology*, *47*(10), 1584-1797. doi: 10.1002/eji.201646632

Courtot, M., Meskas, J., Diehl, A. D., Droumeva, R., Gottardo, R., Jalali, A., ... Brinkman, R. R. (2014, 12). flowCL: ontology-based cell population labelling in flow cytometry. *Bioinformatics*, *31*(8), 1337-1339. doi: 10.1093/bioinformatics/btu807

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., ... Chan,

C. (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS computational biology*, *9*(7), e1003130.

Darrah, P. A., Patel, D. T., De Luca, P. M., Lindsay, R. W., Davey, D. F., Flynn, B. J., ... Roederer, M. (2007). Multifunctional th1 cells define a correlate of vaccine-mediated protection against leishmania major. *Nature medicine*, *13*(7), 843.

Daume III, H. (2007). Fast search for dirichlet process mixture models. In *Artificial intelligence and statistics* (pp. 83–90).

De Rosa, S. C., Herzenberg, L. A., Herzenberg, L. A., & Roederer, M. (2001). 11-color, 13-parameter flow cytometry: identification of human naive t cells by phenotype, function, and t-cell receptor diversity. *Nature medicine*, *7*(2), 245.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, *90*(430), 577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Finak, G., Bashashati, A., Brinkman, R., & Gottardo, R. (2009). Merging mixture components for cell population identification in flow cytometry. *Advances in bioinformatics*, *2009*.

Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., ... McCoy, J. P. (2016). Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Scientific reports*, *6*, 20686.

Finak, G., McDavid, A., Chattopadhyay, P., Dominguez, M., De Rosa, S., Roederer, M., & Gottardo, R. (2013). Mixture models for single-cell assays with applications to vaccine studies. *Biostatistics*, *15*(1), 87–101.

Finak, G., Perez, J.-M., Weng, A., & Gottardo, R. (2010). Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics*, *11*(1). doi: 10.1186/1471-2105-11-546

Freulon, P., Bigot, J., & Hejblum, B. P. (2020). Cytopt: Optimal transport with domain adaptation for interpreting flow cytometry data. *arXiv*, 2006.09003.

Frühwirth-Schnatter, S., & Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, *11*(2),

317–336.

Fulwyler, M. J. (1965). Electronic separation of biological cells by volume. *Science*, *150*(3698), 910–911.

Gorsky, S., Chan, C., & Ma, L. (2020). Coarsened mixtures of hierarchical skew normal kernels for flow cytometry analyses. *arXiv preprint arXiv:2001.06451*.

Hadrup, S. R., & Schumacher, T. N. (2010). Mhc-based detection of antigen-specific cd8+ t cell responses. *Cancer Immunology, Immunotherapy*, *59*(9), 1425–1433.

Hahne, F., Khodabakhshi, A. H., Bashashati, A., Wong, C.-J., Gascoyne, R. D., Weng, A. P., ... others (2010). Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, *77*(2), 121–131.

Hastie, D. I., Liverani, S., & Richardson, S. (2015). Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, *25*(5), 1023–1037.

Hejblum, B. P., 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. *The Annals of Applied Statistics*, *13*(1), 638-660. doi: 10.1214/18-AOAS1209

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005, 02). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67.

Johnsson, K., Wallin, J., & Fontes, M. (2016). Bayesflow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics*, *17*(1), 25.

Kalli, M., Griffin, J. E., & Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, *21*(1), 93-105. doi: 10.1007/s11222-009-9150-y

Lee, S. X., McLachlan, G. J., & Pyne, S. (2016). Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure. *Cytometry Part A*, *89*(1), 30–43. doi: 10.1002/cyto.a.22789

Lévy, Y., Thiébaut, R., Montes, M., Lacabaratz, C., Sloan, L., King, B., ... Palucka, K. (2014). Dendritic cell-based therapeutic vaccine elicits polyfunctional hiv-specific t-cell immunity associated with control of viral load. *European journal of immunology*, *44*(9), 2802–2810.

Li, J., Ray, S., & Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, *8*(Aug), 1687–1723.

Lim, K.-L., & Wang, H. (2018). Fast approximation of variational Bayes Dirichlet process mixture using the maximization–maximization algorithm. *International Journal of Approximate Reasoning*, *93*, 153-177. doi: 10.1016/j.ijar.2017.11.001

Lin, L. (2012). *Bayesian variable selection in clustering and hierarchical mixture modeling* (Unpublished doctoral dissertation). Citeseer.

Lin, L., & Chan, C. (2017). Quantitative methods and bayesian models for flow cytometry analysis in hiv/aids research. In *Quantitative methods for hiv/aids research* (pp. 135–156). CRC Press.

Lin, L., Chan, C., Hadrup, S. R., Froesig, T. M., Wang, Q., & West, M. (2013). Hierarchical bayesian mixture modelling for antigen-specific t-cell subtyping in combinatorially encoded flow cytometry studies. *Statistical applications in genetics and molecular biology*, *12*(3), 309–331.

Lin, L., Chan, C., & West, M. (2016). Discriminative variable subsets in bayesian classification with mixture models, with application in flow cytometry studies. *Biostatistics*, *17*(1), 40-53.

Lin, L., Finak, G., Ushey, K., Seshadri, C., Hawn, T. R., Frahm, N., ... Gottardo, R. (2015). Compass identifies t-cell subsets correlated with clinical outcomes. *Nature biotechnology*, *33*(6), 610-616.

Lin, L., & Li, J. (2017). Clustering with hidden markov model on variable blocks. *The Journal of Machine Learning Research*, *18*(1), 3913–3961.

Lo, K., Brinkman, R. R., & Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, *73*(4), 321–332.

Maecker, H. T., McCoy, J. P., & Nussenblatt, R. (2012). Standardizing immunophenotyping

for the human immunology project. *Nature Reviews Immunology*, *12*(3), 191–200.

Mair, F., & Prlic, M. (2018). Omip-044: 28-color immunophenotyping of the human dendritic cell compartment. *Cytometry Part A*, *93*(4), 402–405.

Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grn, B. (2017). Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics*, *26*(2), 285-295.

Marin, J.-M., Mengersen, K., & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In D. Dey & C. Rao (Eds.), *Bayesian thinking* (Vol. 25, p. 459 - 507). Elsevier. doi: https://doi.org/10.1016/S0169-7161(05)25016-2

Miller, J. W., & Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*.

Miller, J. W., & Harrison, M. T. (2014). Inconsistency of pitman-yor process mixtures for the number of components. *The Journal of Machine Learning Research*, *15*(1), 3333–3370.

Naim, I., Datta, S., Sharma, G., Cavenaugh, J. S., & Mosmann, T. R. (2010). Swift: scalable weighted iterative sampling for flow cytometry clustering. In *2010 ieee international conference on acoustics, speech and signal processing* (pp. 509–512).

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249-265.

Nettey, L., Giles, A. J., & Chattopadhyay, P. K. (2018). Omip-050: A 28-color/30-parameter fluorescence flow cytometry panel to enumerate and characterize cells expressing a wide array of immune checkpoint molecules. *Cytometry Part A*, *93*(11), 1094–1096.

Nowicka, M., Krieg, C., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., ... Robinson, M. D. (2017). CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, *6*, 748. doi: 10.12688/f1000research.11622.2

O'Neill, K., Aghaeepour, N., Špidlen, J., & Brinkman, R. (2013). Flow cytometry bioinformatics. *PLoS computational biology*, *9*(12), e1003365.

Papaspiliopoulos, O., & Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, *95*(1), 169–186.

Perfetto, S. P., Chattopadhyay, P. K., & Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, *4*(8), 648.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., ... Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, *106*(21), 8519–8524.

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(5), e1435. doi: 10.1002/wics.1435

Robinson, J. P. (2005). Mack fulwyler in his own words. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, *67*(2), 61–67.

Robinson, J. P. (2013). Wallace h. coulter: Decades of invention and discovery. *Cytometry Part A*, *83A*(5), 424-438. doi: 10.1002/cyto.a.22296

Roederer, M., Brenchley, J. M., Betts, M. R., & De Rosa, S. C. (2004). Flow cytometric analysis of vaccine responses: how many colors are enough? *Clinical immunology*, *110*(3), 199–205.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, *8*(1), 289.

Seshadri, C., Lin, L., Scriba, T. J., Peterson, G., Freidrich, D., Frahm, N., ... Mahomed, H. (2015). T cell responses against mycobacterial lipids and proteins are poorly correlated in south african adolescents. *The Journal of Immunology*, *195*(10), 4595–4603.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, 639–650.

Shapiro, H. M. (2005). *Practical flow cytometry*. John Wiley & Sons.

Soriano, J., & Ma, L. (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis*, *14*(1), 161–180.

Srivastava, S., Li, C., & Dunson, D. B. (2018). Scalable Bayes via Barycenter in Wasserstein Space. *Journal of Machine Learning Research*, *19*(8), 1-35.

Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A., & Teichmann, S. A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. *Science*,

358(6359), 58–63.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., & West, M. (2010). Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of computational and graphical statistics*, *19*(2), 419–438.

Teh, Y. W. (2010). Dirichlet Process. In *Encyclopedia of Machine Learning* (p. 280-287). Boston, MA: Springer US. doi: 10.1007/978-0-387-30164-8_219

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems* (pp. 1385–1392).

Turkman, M. A. A., Paulino, C. D., & Müller, P. (2019). *Computational bayesian statistics: An introduction* (Vol. 11). Cambridge University Press.

van Havre, Z., White, N., Rousseau, J., & Mengersen, K. (2015, 07). Overfitting bayesian mixture models with an unknown number of components. *PLOS ONE*, *10*(7), 1–27. doi: 10.1371/journal.pone.0131739

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, *87*(7), 636-645.

Wang, Y., & Blei, D. M. (2019). Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, *114*(527), 1147-1161. doi: 10.1080/01621459.2018.1473776

Weber, L. M., & Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, *89*(12), 1084-1096.