Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **72**, 718–738
https://doi.org/10.1093/jrsssc/qlad029
Advance access publication 25 April 2023
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

A Bayesian feature allocation model for identifying cell subpopulations using CyTOF data

Arthur Lui[1], Juhee Lee[2], Peter F. Thall[3], May Daher[4], Katy Rezvani[4] and Rafet Basar[4]

[1]Department of Statistics, Baskin School of Engineering, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, USA
[2]Department of Statistics, University of California at Santa Cruz, Santa Cruz, CA, USA
[3]Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, USA
[4]Department of Stem Cell Transplantation and Cellular Therapy, M.D. Anderson Cancer Center, Houston, TX, USA

Address for correspondence: Arthur Lui, Department of Statistics, Baskin School of Engineering, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA. Email: alui2@ucsc.edu

Abstract

A Bayesian feature allocation model (FAM) is presented for identifying cell subpopulations based on multiple samples of cell surface or intracellular marker expression level data obtained by cytometry by time of flight (CyTOF). Cell subpopulations are characterized by differences in marker expression patterns, and cells are clustered into subpopulations based on their observed expression levels. A model-based method is used to construct cell clusters within each sample by modeling subpopulations as latent features, using a finite Indian buffet process. Non-ignorable missing data due to technical artifacts in mass cytometry instruments are accounted for by defining a static missingship mechanism. In contrast with conventional cell clustering methods, which cluster observed marker expression levels separately for each sample, the FAM-based method can be applied simultaneously to multiple samples, and also identify important cell subpopulations likely to be otherwise missed. The proposed FAM-based method is applied to jointly analyse three CyTOF datasets to study natural killer (NK) cells. Because the subpopulations identified by the FAM may define novel NK cell subsets, this statistical analysis may provide useful information about the biology of NK cells and their potential role in cancer immunotherapy which may lead, in turn, to development of improved NK cell therapies.

Keywords: clustering, latent features, natural killer cells, non-ignorable missing data, subpopulations

1 Introduction

Mass cytometry data have been used for high-throughput characterization of cell subpopulations based on unique combinations of surface or intracellular markers that may be expressed by each cell. Cytometry by time-of-flight (CyTOF), first introduced in 2009, is a technology that can rapidly quantify a large number of biological, phenotypic, or functional markers on single cells through use of metal-tagged antibodies. For example, CyTOF can identify up to 40 cell surface or intracellular markers in less time and at a higher resolution than previously available methods, such as fluorescence cytometry (Cheung & Utz, 2011). Because CyTOF can reveal cellular diversity and heterogeneity that could not be seen previously, it has the potential to rapidly advance the study of cellular phenotype and function in immunology.

Despite the potential of CyTOF, analysis of the data that it generates is computationally expensive and challenging, and statistical tools for making inferences about cell subpopulations identified by CyTOF are quite limited. Manual 'gating' is a traditional method in which homogeneous

Received: March 29, 2023. **Accepted:** April 2, 2023

cell clusters are sequentially identified and refined based on a given set of surface markers. Manual gating has several severe shortcomings, however, including its inherent subjectivity due to the fact that it requires manual analysis, and being unscalable for high-dimensional data with large numbers of markers. While manual gating is commonly used in practice, a variety of computational methods that automatically identify cell clusters have been proposed to analyse high-dimensional cytometry data. Many existing automated methods use dimension reduction techniques and/or clustering methods, such as density-based or model-based clustering. For example, FlowSOM, given by Van Gassen et al. (2015), uses an unsupervised self-organizing map (SOM) for clustering and dimension reduction. A low-dimensional representation of the marker vectors is obtained by using unsupervised neural networks for easy visualization in a graph called a map. FlowSOM is fast and can be used either as a starting point for manual gating, or as a visualization tool after gating has been performed. Other common approaches are density-based clustering methods, including DBSCAN (Ester et al., 1996) and ClusterX (H. Chen et al., 2016), and model-based clustering methods, including flowClust (Lo et al., 2009) and BayesFlow (Johnsson et al., 2016). More sophisticated clustering methods based on Bayesian non-parametric models also have been proposed, see, for example, by Soriano and Ma (2019). Weber and Robinson (2016) performed a study to compare several clustering methods for high-dimensional cytometry data. They analysed six publicly available cytometry datasets and compared identified cell subpopulations to cell population identities known from expert manual gating. They found that, in many scenarios, FlowSOM had significantly shorter runtimes. Moreover, in many studies where manual gating was performed, FlowSOM produced the best clusterings, in terms of various clustering metrics, when compared to cell clustering by manual gating.

While conventional clustering methods identify subgroups of cells with similar marker expression values, they often fail to provide direct inferences that identify and characterize cell subpopulations. Clustering methods put cells in the same cluster if their expression levels are similar, and they assume implicitly that underlying cell subpopulations can be identified and constructed from clusters estimated directly from the marker expression levels. The usefulness of such conventional clustering approaches is limited by the fact that observed numerical marker expression values may differ substantially due to between-sample variability, often due to technical variation in the cytometry measurement process, as well as variability in the expression measurement process. Figure 1 illustrates a toy example. Suppose that the respective log expression levels of markers 1 and 2 are -2 and -4 on a given cell, and that the corresponding values on a second cell are -6 and -4. A negative (or positive) log expression level implies that it is unlikely (or likely) that a surface marker is expressed. Although their expression patterns are qualitatively similar and are from the same subpopulation, a conventional clustering method is unlikely to include these two cells in the same cluster because their marker 1 expression levels are very different. A deeper problem is that cell clusters based on expression values may not serve as a useful surrogate for identifying cell subpopulations. As a result, most existing clustering methods are used to analyse different samples separately.

In this paper, we propose a Bayesian feature allocation model (FAM) to identify and place probabilities on cell subpopulations based on multiple cytometry samples of a vector of cell surface marker expression values. Our proposed FAM characterizes cell subpopulations as latent features defined in terms of their expression patterns, and clusters individual cells to one of the identified subpopulations. We will refer to each latent feature as a 'subpopulation'. With this FAM, a given marker may be expressed in more than one cell subpopulation, and each subpopulation is characterized by a unique marker expression pattern. To characterize subpopulation configurations, we introduce a random matrix $\boldsymbol{Z}$ with rows corresponding to markers and columns to subpopulations, with entry 1 denoting expression and 0 denoting non-expression of a marker in a subpopulation. Unlike traditional clustering methods, the FAM constructs latent subpopulations based on marker expression patterns, as illustrated by the $\boldsymbol{Z}$ matrix in the top figure in Figure 1. It assigns cells 1 and 2 to subpopulation 1, where neither marker is expressed, and it assigns cell 3 to subpopulation 2, where marker 1 is expressed and marker 2 is not expressed. We assume a finite Indian buffet process (IBP) as the prior distribution for $\boldsymbol{Z}$. The IBP is a popular model for latent binary features, and it may be obtained as the infinite limit of a Beta-Bernoulli process (Ghahramani & Griffiths, 2006). Applications of the IBP prior in FAMs for a range of biological applications are given by Hai-son and Bar-Joseph (2011), M. Chen et al. (2013), Xu et al. (2013,

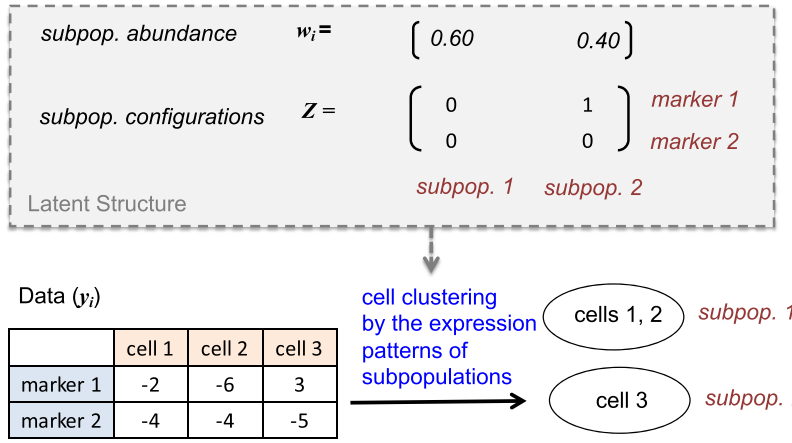


Figure 1. A stylized overview of the proposed feature allocation model. $\boldsymbol{Z}$ is a binary matrix whose columns define latent subpopulations, and $\boldsymbol{w}$ is a vector of cell subpopulation abundances. Two subpopulations are constructed in $\boldsymbol{Z}$ based on their marker expression patterns. Cells are clustered into the subpopulations based on their observed expression level patterns.

2015), Sengupta et al. (2014), Lee et al. (2015, 2016), and Ni et al. (2018). Griffiths and Ghahramani (2011) reviews some earlier applications of the IBP. Furthermore, we introduce a vector of subpopulation abundances $\boldsymbol{w}_i$ for each sample (i). This approach provides a framework for the joint analysis of multiple samples, and includes structures to account for large sample-to-sample variation and abnormalities, such as missing values due to technical artifacts in the cytometry data, while quantifying uncertainty in posterior inferences.

The model and analyses in this paper are motivated by a dataset comprised of three CyTOF samples of surface marker expression levels in umbilical cord blood (UCB)—derived natural killer (NK) cells. NK cells play a critical role in cancer immune surveillance, and are the body's first line of defense against viruses and transformed tumour cells. NK cells have the intrinsic ability to infiltrate cancer tissues. Recently, NK cells have been used therapeutically to treat a variety of diseases (Lanier, 2008; Wu & Lanier, 2003). In particular, NK cells have emerged as a potentially powerful treatment modality for advanced cancers refractory to conventional therapies (Liu et al., 2020; Lupo & Matosevic, 2019; J. S. Miller et al., 2005; Rezvani & Rouce, 2015; Shah et al., 2017; Suck et al., 2016). Because cell-surface protein expression levels are used as markers to describe the behaviour of NK cells, accurate identification of diverse NK-cell subpopulations along with their composition is crucial to the process of obtaining more complete characterizations of their biological processes and functions. The goal of our statistical analysis is to identify and characterize NK cell subpopulations and functions across heterogeneous collections of these cells. This may provide critical information for guiding selective *ex vivo* expansion of UCB-derived NK cells for treating specific cancers.

The remainder of this paper is organized as follows. We present the proposed statistical model in Section 2, simulation studies in Section 3, and an analysis of the NK cell mass cytometry data in Section 4. We close with concluding remarks in Section 5.

2 Probability model

2.1 Sampling model

Index cell samples by $i = 1, 2, \ldots, I$. Suppose that N_i cells, indexed by $n = 1, \ldots, N_i$, are obtained from the ith sample, and the expression levels of J markers on each cell within each sample are measured. Let $\tilde{y}_{i,n,j} \in \mathbb{R}^+$ denote the raw measurement of the expression level of marker j on cell n in sample i. While raw measurement values reflect actual expression or non-expression of markers on cells, they also vary between cells and between samples for several reasons, including biological heterogeneity in the range of expression among different populations, as well as experimental artifacts or batch effects, such as instrument fluctuations or signal crosstalk among channels designed for

different markers. While, compared to conventional flow cytometry and the use of fluorescent antibodies, the use of pure metal isotopes minimizes spectral overlap among measurement channels in CyTOF, crosstalk still may be observed due to the presence of isotopic impurity, oxide formation, and properties related to the mass cytometer. Raw measurements are normalized using cut-off values computed by a flow (rather than mass) cytometry algorithm called flowDensity (Malek et al., 2014), which aims to gate predefined cell populations of interest, in settings where the gating strategy is known. This frees practitioners from the need to manually gate analysis results, but it relies substantially on user-provided information to produce good results. Consequently, cut-offs obtained from such algorithms are crude, but useful as a starting point for our analysis. Let $c_{i,j}$ denote the cut-off for marker j in sample i obtained by flowDensity. A marker of a cell is likely to be expressed if its observed expression level $\tilde{y}_{i,n,j} > c_{i,j}$, while a value $\tilde{y}_{i,n,j} < c_{i,j}$ may imply that marker j is not expressed on cell n in sample i. To reduce the skewness of the marker distributions, we will base our model on the log-transformed values $y_{i,n,j} = \log(\tilde{y}_{i,n,j}/c_{i,j}) \in \mathbb{R}$. This transformation makes 0 the reference point for dichotomizing marker expression and non-expression. To account for the fact that some $y_{i,n,j}$ may be missing due to experimental artifacts, we define the binary indicator $m_{i,n,j} = 1$ if $y_{i,n,j}$ is observed, and $m_{i,n,j} = 0$ if missing. Denote the probability that $y_{i,n,j}$ is missing by $\Pr(m_{i,n,j} = 0 \mid y_{i,n,j}) = \rho_{i,n,j}(y_{i,n,j})$, so $1 - \rho_{i,n,j}(y_{i,n,j})$ is the probability that $y_{i,n,j}$ is observed. Below, we will define the latent subpopulation membership indicator, $\lambda_{i,n}$, of cell n in sample i. For each cell in the ith sample, we assume conditional independence of the cell's J marker values given its latent subpopulation, formally $y_{i,n,1}, \ldots, y_{i,n,J} \mid \lambda_{i,n}$ are independent, and we write the joint model for $(y_{i,n,j}, m_{i,n,j})$ as follows:

$$y_{i,n,j} \mid \mu_{i,n,j}, s^2_{i,n}, \lambda_{i,n} \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{i,n,j}, s^2_{i,n}) \tag{1}$$

and

$$m_{i,n,j} \mid \rho_{i,n,j}(y_{i,n,j}), \lambda_{i,n} \overset{\text{ind}}{\sim} \mathrm{Ber}(1 - \rho_{i,n,j}(y_{i,n,j})). \tag{2}$$

This joint model provides a basis for imputing missing expression levels by drawing $y_{i,n,j}$ from $p(y_{i,n,j} \mid m_{i,n,j})$ if $m_{i,n,j} = 0$, and it also facilitates posterior simulation. Below, we will relate the mean expression $\mu_{i,n,j}$ to the configuration of cell subpopulation $\lambda_{i,n}$. To reflect the expert biological knowledge of the investigators, a model for $\rho_{i,n,j}$ as a function of $y_{i,n,j}$ will be given in the following section.

2.2 Priors

2.2.1 Priors for latent cell subpopulation

We assume that each sample consists of a heterogeneous cell population, and denote the number of different latent subpopulations by K. The cell subpopulations are defined by the columns of the $J \times K$ (marker, subpopulation) stochastic binary matrix $\boldsymbol{Z}$. The element $z_{j,k} \in \{0, 1\}$ of $\boldsymbol{Z}$ determines marker expression by subpopulation, with $z_{j,k} = 0$ if marker j is not expressed and $z_{j,k} = 1$ if it is expressed for subpopulation k. We construct a *feature allocation prior* for $\boldsymbol{Z}$ as follows: For $j = 1, \ldots J$ and $k = 1, \ldots, K$,

$$z_{j,k} \mid v_k \overset{\text{ind}}{\sim} \mathrm{Ber}(v_k) \quad \text{and} \quad v_k \mid \alpha \overset{\text{iid}}{\sim} \mathrm{Be}(\alpha/K, 1). \tag{3}$$

As $K \to \infty$, the limiting distribution of $\boldsymbol{Z}$ in (3) is the IBP (Ghahramani & Griffiths, 2006) with parameter α, after removing all columns that contain only zeros. We assume hyperprior $\alpha \sim \mathrm{Gamma}(a_\alpha, b_\alpha)$ with mean a_α/b_α. The IBP, which is one of the most popular FAMs, thus defines a distribution over binary matrices having an unbounded number of columns (features). For our purposes, the simpler version of the IBP with finite K provides a very useful statistical tool for identifying marker expression patterns to define latent cell subpopulations from CyTOF surface marker data. While $z_{j,k}$ in (3) can be 0 or 1 for non-expression/expression, the model can be further extended to accommodate ordered categories of a marker expression, such as no/moderate/high expressions. For example, we may let $z_{j,k} = 0$, 1, or 2 for no/moderate/high expressions, and consider the categorical IBP in Sengupta et al. (2014) and Lee et al. (2016) as a prior for such a $\boldsymbol{Z}$. This extended model may be preferred when a finer categorization of expression level is more desirable.

We assume that each of the K cell subpopulations may potentially appear in each sample, but allow their cellular fractions to differ between samples. In addition, we include a special, $(K+1)$st 'noisy' cell type, indexed by $k=0$, to address the problem that some cells do not belong to any of the K cell subpopulations. In sample i, let $0<\epsilon_i<1$ denote the proportion of noisy cells and $(1-\epsilon_i)w_{ik}$ the proportion of cells belonging to subpopulation k, where $\boldsymbol{w}_i=(w_{i,1},\ \ldots,\ w_{i,K})$ with $\sum_{k=1}^{K} w_{i,k}=1$ and $w_{i,k}>0$, is a probability distribution on $\{1,\ \ldots,\ K\}$. We assume priors $\epsilon_i \overset{\text{iid}}{\sim} \text{Be}(a_\epsilon, b_\epsilon)$ with fixed hyperparameters a_ϵ and b_ϵ, and $\boldsymbol{w}_i \mid K \overset{\text{iid}}{\sim} \text{Dir}_K(d/K)$ with fixed hyperparameter d. For cell $n=1,\ \ldots,\ N_i$ in sample $i=1,\ \ldots,\ I$, we introduce stochastic *latent subpopulation indicators* (equivalently, cell cluster membership labels) $\lambda_{i,n}\in\{0,\ 1,\ \ldots,\ K\}$. We set $\lambda_{i,n}=0$ if cell n in sample i does not belong to any of the cell subpopulations in $\boldsymbol{Z}$, and set $\lambda_{i,n}=k>0$ if cell n in sample i belongs to subpopulation k. For the latent subpopulation indicators, we assume $\Pr(\lambda_{i,n}=0\mid\epsilon_i)=\epsilon_i$ to account for noisy cells, and $\Pr(\lambda_{i,n}=k\mid\lambda_{i,n}\neq 0,\boldsymbol{w}_i)=w_{ik}$. Within each sample $i=1,\ \ldots,\ I$, assigning cells to subpopulations using $\{\lambda_{i,n},\ i=1,\ \ldots,\ N_i\}$ induces cell clusters. Thus, in contrast with clustering methods that infer only cell clusters in the ith sample based on $\{y_{i,n,j}\}$, our proposed method produces direct inferences on both characterization of cell subpopulations and cell clusters simultaneously for all samples. This is highly desirable because the primary aim is to identify and make inferences about cell subpopulations.

Since the number of columns containing non-zero entries under the IBP is random, the dimensions of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$ may vary during posterior computation. Because this dimension change may cause a high computational cost, especially for big datasets such as those obtained by CyTOF, we use a finite version of the IBP with fixed K. Because the number of latent subpopulations is not known *a priori*, we consider a set of different values for K, from which we select one value of K using Bayesian model selection criteria. We will discuss this selection process below.

2.2.2 *Priors for mean expression level*

The mean expression level $\mu_{i,n,j}$ of marker j for cell n in sample i in (2) is determined by the cell's latent subpopulation. For cells in the noisy cell subpopulation where $\lambda_{i,n}=0$, we fix $\mu_{i,n,j}=0$ for all j and $s^2_{i,n}=s^2_\epsilon$, where s^2_ϵ is a large constant. For a cell with $\lambda_{i,n}\in\{1,\ \ldots,\ K\}$, if the marker is not expressed in cell subpopulation $\lambda_{i,n}$ (i.e., $z_{j,\lambda_{i,n}}=0$), we restrict its mean expression level to be a negative value, $\mu_{i,n,j}<0$. Specifically, for $(i,\ n,\ j)$ with $z_{j,\lambda_{i,n}}=0$, we introduce a set of means for expression levels of markers not expressed, $\mu^\star_{0,\ell}=\sum_{r=1}^{\ell}\delta_{0,r}$, where $\delta_{0,\ell}\overset{\text{iid}}{\sim}\text{TN}^-(\psi_0,\tau_0^2)$, $\ell=1,\ \ldots,\ L_0$ with fixed L_0. Here, $\text{TN}^-(\psi,\tau^2)$ denotes the normal distribution with mean ψ and variance τ^2 truncated above at zero. This construction induces the ordering $0>\mu^\star_{0,1}>\cdots>\mu^\star_{0,L_0}$. We then let $\mu_{i,n,j}=\mu^\star_{0,\ell}$ with probability $\eta^0_{i,j,\ell}$. Note that even for a marker not expressed, positive $y_{i,n,j}$ can be observed due to measurement error or estimation error in the cut-off $c_{i,j}$, and the model accounts for such cases through $s^2_{i,n}$. Similarly, we assume that the mean expression level of marker j takes a positive value $(\mu_{i,n,j}>0)$ if the marker is expressed $(z_{j,\lambda_{i,n}}=1)$. For cases with $z_{j,\lambda_{i,n}}=1$, we construct another set of δ, with distribution $\delta_{1,\ell}\overset{\text{iid}}{\sim}\text{TN}^+(\psi_1,\tau_1^2)$, $\ell=1,\ \ldots,\ L_1$ for fixed L_1, where $\text{TN}^+(\psi,\tau^2)$ denotes the normal distribution truncated below at zero with mean ψ and variance τ^2. We let $\mu^\star_{1,\ell}=\sum_{r=1}^{\ell}\delta_{1,r}$, so $0<\mu^\star_{1,1}<\cdots<\mu^\star_{1,L_1}$. We then let $\mu_{i,n,j}=\mu^\star_{1,\ell}$ with probability $\eta^1_{i,j,\ell}$, and let $s^2_{i,n}=\sigma_i^2$ for $\lambda_{i,n}>0$ and assume $\sigma_i^2\overset{\text{ind}}{\sim}\text{IG}(a_\sigma,b_\sigma)$. This leads to a mixture of normals for $y_{i,n,j}$ whose location parameters are determined by the cell's latent subpopulation,

$$y_{i,n,j}\mid z_{j,\lambda_{i,n}}=z,\boldsymbol{\mu}^\star_z,\eta^z_{i,j},\sigma_i^2\overset{\text{ind}}{\sim}F^z_{i,j}=\sum_{\ell=1}^{L_z}\eta^z_{i,j,\ell}\cdot\text{N}(\mu^\star_{z,\ell},\sigma_i^2),\quad z\in\{0,1\},\ k>0. \tag{4}$$

Finally, we let $\boldsymbol{\eta}^z_{i,j}\overset{\text{iid}}{\sim}\text{Dir}_{L_z}(a_{\eta^z}/L_z)$, for $z=0,\ 1$, $i=1,\ \ldots,\ I$, and $j=1,\ \ldots,\ J$.

The mixture model in (4) encompasses a wide class of distributions, which may be multi-modal or skewed. It captures virtually any departure from a conventional distribution, such as a parametric exponential family model, that may appear to give a good fit to the log-transformed expression

values. A key property of (4) is that it allows cells with very different numerical expression values to have the same subpopulation if their marker expression/non-expression pattern is the same. The mixture model can also account for batch effects through model-based centring and scaling of observed expression levels, in addition to sample and marker-specific cut-off values $c_{i,j}$. If considered more desirable, other batch adjustment approaches can be applied prior to analyses. For example, Schuyler et al. (2019) estimates batch effect explicitly and adjusts samples within a batch for datasets including technical replicates. This provides a basis for obtaining a succinct representation of cell subpopulations. Because the locations $\mu_z^\star$ in (4) are shared for all (i, n, j), the model borrows strength across both samples and markers, while $\boldsymbol{\eta}_{i,j}^z = (\eta_{i,j,1}^z, \ldots, \eta_{i,j,L^z}^z)$ allows the distribution of $y_{i,n,j}$ to vary across both samples and markers. The construction of $\mu_{z,\ell}^\star$ through $\delta_{z,\ell}$ also ensures ordering in $\mu_{z,\ell}^\star$ and circumvents potential identifiability and label-switching issues that may be present in conventional Bayesian mixture models (Celeux et al., 2000; Frühwirth-Schnatter, 2006; Jasra et al., 2005; Stephens, 2000).

2.2.3 *Model for the data missingship mechanism*

We next build a model for the data missingship distribution. To do this, we incorporate information provided by a subject area expert, that a marker expression level is recorded as 'missing' in a cell ($m_{i,n,j} = 0$) when the marker's signal is very weak, which strongly implies that the marker is not expressed on that cell. In (2), we model missingship $m_{i,n,j}$ conditional on $y_{i,n,j}$, i.e., $m_{i,n,j} \mid \rho_{i,n,j}(y_{i,n,j}) \overset{\text{ind}}{\sim} \text{Ber}(1 - \rho_{i,n,j}(y_{i,n,j}))$. We assume a logit regression model for the probability $\rho_{i,n,j}$ that $m_{i,n,j} = 0$,

$$\text{logit}(\rho_{i,n,j}) = \beta_{0i} + \beta_{1i} y_{i,n,j} + \beta_{2i} y_{i,n,j}^2. \tag{5}$$

We take an empirical approach to specify values of $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})$ in (5) for each sample $i = 1, \ldots, I$ by using the minimum, first quartile, and median of negative observed expression levels, setting their $\rho_{i,n,j}$ values to .05, .80, and .50, respectively, and solving for $\boldsymbol{\beta}_i$. More details of the specification of $\boldsymbol{\beta}_i$ are in Online Supplementary Material, Section 2. The proposed specification of $\boldsymbol{\beta}_i$ reflects the key fact that when $m_{i,n,j} = 0$, its potentially observed numerical value is very likely negative. The dataset does not contain information for inferring the missingness mechanism, and it cannot be anticipated that the imputed values are close to their potentially observed values. However, our construction of subpopulations is based on patterns of expression levels, not actual expression levels, and the task of recovering $\boldsymbol{Z}$, $\boldsymbol{w}$, and $\boldsymbol{\lambda}$, which is the primary aim of the analyses, is not affected by particular imputed values. We performed sensitivity analyses to the specification of the $\boldsymbol{\beta}_i$'s to examine robustness of the estimation of $\boldsymbol{Z}$, $\boldsymbol{w}$, and $\boldsymbol{\lambda}$. Additionally, in our simulation studies, missing values were generated under a mechanism different from that in (5). The underlying cell subpopulations were well recovered even with the misspecified missingness mechanism, indicating the model's robustness. Sections 3 and 4 provide details of the sensitivity analyses. There is an extensive literature on analysing data with observations missing not at random, including methods for Bayesian data imputation and frequentist multiple imputation (Allison, 2001; Franks et al., 2016; Rubin, 1974, 1976; Schafer & Graham, 2002). We refer to them for alternative models for the missingship mechanism.

2.2.4 *Selection of* K

Instead of estimating K, we cast the problem of selecting a value for K as a model comparison problem. This reduces the computational burden, especially for large datasets. To identify a value of K that optimizes model fit while penalizing for high model complexity, we choose K using the deviance criterion information (DIC, Spiegelhalter et al., 2002) and log pseudo marginal likelihood (LPML, Geisser & Eddy, 1979; Gelfand & Dey, 1994). The DIC and LPML are commonly used to quantify goodness-of-fit for comparing Bayesian models. The DIC measures posterior prediction error based on deviance penalized by model complexity, with lower values corresponding to a better fit. The LPML is a metric based on cross-validated posterior predictive probability, and is defined as the sum of the logarithms of conditional predictive ordinates (CPOs), with larger LPML corresponding to a better fit. Details of the computation of DIC and LPML are given in

Online Supplementary Material, Section 3. In addition, we count the number of subpopulations having negligible weights, $\sum_{i,k} \mathrm{I}(w_{i,k} < 1\%)$, for each value of K and plot the LPML against the number of such subpopulations. A model with larger K may produce cell subpopulations with very small $w_{i,k}$ that only make subtle contributions to model fit in terms of LPML. We thus search for a value of K, where the change rate of the increase in LPML drops. J. W. Miller and Dunson (2018) used a similar calibration method to tune a model hyperparameter that determines how much coarsening is required to obtain a model that maximizes model fit while maintaining low model complexity.

2.3 Posterior computation

Let $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \sigma^2, \boldsymbol{\eta}^0, \boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{\epsilon}, \boldsymbol{\alpha}\}$ denote all model parameters. Let $\boldsymbol{y}$ and $\boldsymbol{m}$ denote the vectors of $y_{i,n,j}$ and $m_{i,n,j}$ values for all (i, n, j). The posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{m}, K) &\propto p(\boldsymbol{\theta} \mid K) \prod_{i,n,j} p(m_{i,n,j} \mid y_{i,n,j}, \boldsymbol{\theta}, K) p(y_{i,n,j} \mid \boldsymbol{\theta}, K) \\ &\propto p(\boldsymbol{\theta} \mid K) \prod_{i,n} \left[\prod_j \rho_{i,n,j}^{1-m_{i,n,j}} \sum_{\ell=1}^{L_{z_{j,\lambda_{i,n}}}} \eta_{i,j,\ell}^{z_{j,\lambda_{i,n}}} \phi(y_{i,n,j} \mid \mu^{\star}_{z_{j,\lambda_{i,n}},\ell}, \sigma_i^2) \right]^{1(\lambda_{i,n}>0)} \\ &\quad \times \left[\prod_j \rho_{i,n,j}^{1-m_{i,n,j}} \times \phi(y_{i,n,j} \mid 0, s_\epsilon^2) \right]^{1(\lambda_{i,n}=0)}, \end{aligned} \tag{6}$$

where $\phi(y \mid \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2 evaluated at y. Since $\rho_{i,n,j}$ is a constant for a given y with fixed β's, the terms $p(m_{i,n,j} = 1) = (1 - \rho_{i,n,j})^{m_{i,n,j}}$ for observed $y_{i,n,j}$ do not appear in (6). Posterior simulation can be done via standard Markov chain Monte Carlo (MCMC) methods with Gibbs and Metropolis steps. Each parameter is updated sequentially by sampling from its full conditional distribution. Details of the posterior simulation are described in Online Supplementary Material, Section 1.

Summarizing the joint posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{m}, K)$ is challenging, especially for $\boldsymbol{Z}$, which may be susceptible to label-switching problems common in mixture models. Moreover, the distributions of $\boldsymbol{w}_i$ and $\boldsymbol{\lambda}_i$ depend on $\boldsymbol{Z}$. To summarize the posterior distribution of $(\boldsymbol{Z}, \boldsymbol{w}_i, \boldsymbol{\lambda}_i)$ with point estimates, we extend the sequentially allocated latent structure optimization (SALSO) method in Dahl and Müller (2017) to incorporate $\boldsymbol{w}_i$. To summarize random feature allocation matrices, we first construct $\boldsymbol{A_i} = \{\boldsymbol{A}_{i,(j,j')}(\boldsymbol{Z})\}$, the $J \times J$ pairwise allocation matrix corresponding to a binary matrix $\boldsymbol{Z}$, where

$$A_{i,(j,j')}(\boldsymbol{Z}) = \sum_{k=1}^{K} w_{i,k} \times 1(z_{j,k} = 1) \times 1(z_{j',k} = 1), \quad \text{for } 1 \le j, j' \le J, \tag{7}$$

is the number of active features that markers j and j' have in common in sample i, weighted by $w_{i,k}$. The form of (7) encourages selection of entries in $\boldsymbol{Z}$ based on subpopulations that are prevalent in the samples. We find a point estimate $\hat{\boldsymbol{Z}}_i$ for sample i that minimizes the sum of the element-wise squared distances

$$\underset{Z}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{j'=1}^{J} (A(\boldsymbol{Z})_{i,(j,j')} - \bar{A}_{i,(j,j')})^2,$$

where $\bar{A}_{i,(j,j')}$ is the pairwise allocation matrix averaged by the posterior distribution of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. We use posterior Monte Carlo samples to obtain posterior point estimates $\hat{\boldsymbol{Z}}_i$ as follows. Suppose that we obtain B posterior samples simulated from the posterior distribution of $\boldsymbol{\theta}$. For the bth posterior sample of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$, we compute the $J \times J$ adjacency matrix, $\boldsymbol{A}_i^{(b)} = \{\boldsymbol{A}_{i,(j,j')}^{(b)}\}$, $b = 1, \ldots, B$ and

then the mean adjacency matrix $\bar{A}_i = \sum_{b=1}^{B} A_i^{(b)}/B$. We determine a posterior point estimate of $\boldsymbol{Z}$ for sample i by minimizing the sum of squared deviation, $\hat{\boldsymbol{Z}}_i = \text{argmin}_{\boldsymbol{Z}} \sum_{j,j'} (A_{i,j,j'}^{(b)} - \bar{A}_{i,j,j'})^2$, where $\hat{\boldsymbol{Z}}_i \in \{\boldsymbol{Z}^{(1)} \ldots \boldsymbol{Z}^{(B)}\}$. For $\hat{\boldsymbol{Z}}_i = \boldsymbol{Z}^{(b)}$, we report the posterior point estimates $\hat{\boldsymbol{w}}_i = \boldsymbol{w}_i^{(b)}$ and $\hat{\lambda}_{i,n} = \lambda_{i,n}^{(b)}$. Alternatively, we can find estimates $\hat{\boldsymbol{Z}}$ common for all samples by finding $\hat{\boldsymbol{Z}}$ such that

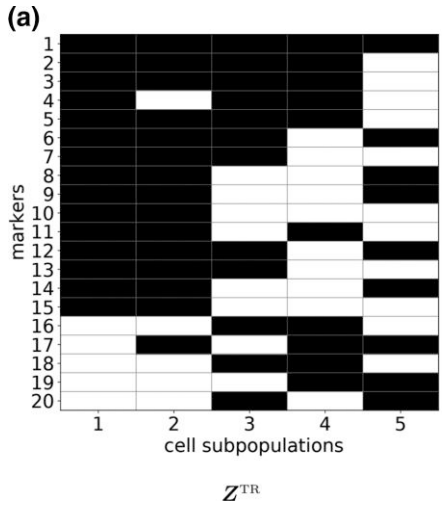
$$\hat{\boldsymbol{Z}} = \underset{\boldsymbol{Z}}{\text{argmin}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{j'=1}^{J} (A(\boldsymbol{Z})_{i,(j,j')} - \bar{A}_{i,(j,j')})^2.$$

Similar to $\hat{\boldsymbol{Z}}_i$, we use posterior samples obtained through MCMC simulation, and report posterior sample b' that achieves the minimum as point estimates common for all i, $\hat{\boldsymbol{Z}} = \boldsymbol{Z}^{(b')}$. We then let $\hat{\boldsymbol{w}}_i = \boldsymbol{w}_i^{(b')}$ and $\hat{\lambda}_{i,n} = \lambda_{i,n}^{(b')}$.

Because the model is complex and the dataset is large, as an alternative method for posterior computation we explored the use of variational inference (VI), which approximates the posterior distribution of $\boldsymbol{\theta}$ through optimization (Blei et al., 2017; Wainwright & Jordan, 2008; Zhang et al., 2018). Because VI tends to be faster than MCMC, it is a popular emerging alternative, especially for complex models and/or large datasets. We used automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) to simplify the process of implementing variational inference for differentiable models. ADVI requires no model-specific analytical derivations of derivatives, and it is relatively simple to implement using an automatic differentiation library such as PyTorch (Paszke et al., 2017), TensorFlow (Abadi et al., 2015), and Flux (Innes, 2018). Details of the VI implementation using ADVI are included in Online Supplementary Material, Section 1.2. A Julia package CytofResearch implementing this methodology is available at https://github.com/luiarthur/CytofResearch. The repository also includes a brief demonstration of how to use the software at https://github.com/luiarthur/CytofResearch/tree/master/demos/minimal-example.

3 Simulation studies

In this section, we present simulation studies to assess the performance of the proposed FAM-based method for identifying features and clustering cells within each sample, and we compare the FAM to an alternative model and method. We simulated data for $I = 3$ samples, each with 20 markers, consisting of $N_i = 4{,}000$, 500, and 1,000 cells, for $i = 1$, 2, and 3, respectively. We set the true number of latent features (subpopulations) to be $K^{\text{TR}} = 5$ and specified a $J \times 5$ binary feature-allocation matrix $\boldsymbol{Z}^{\text{TR}}$ and 5-dimensional vectors $\boldsymbol{w}_i^{\text{TR}}$ as follows: We first simulated $\boldsymbol{Z}^{\text{TR}}$ by setting $z_{j,k}^{\text{TR}} = 1$ with probability 0.6. If any column or row in $\boldsymbol{Z}^{\text{TR}}$ consisted of all 0's, the entire matrix was re-sampled. We then simulated $\boldsymbol{w}_i^{\text{TR}}$ from a Dirichlet distribution with parameters being random permutations of $(1, \ldots, 5)$ for each i. This was done so that the resulting elements of $\boldsymbol{w}_i^{\text{TR}}$ would be likely to contain both large and small values. The assumed $\boldsymbol{Z}^{\text{TR}}$ and $\boldsymbol{w}_i^{\text{TR}}$ are given in Figure 2. We set $\epsilon_i^{\text{TR}} = 5\%$ of the cells to be noisy for all i. We specified the mixture models for the expression levels by setting $\boldsymbol{\mu}_0^{\star,\text{TR}} = (-1, -2.3, -3.5)$ and $\boldsymbol{\mu}_1^{\star,\text{TR}} = (1, 2, 3)$ with $L^{0,\text{TR}} = L^{1,\text{TR}} = 3$, and simulating mixture weights $\boldsymbol{\eta}_{i,j}^{z,\text{TR}}$ from a Dirichlet distribution with parameters a random permutation of $(1, \ldots, L^{z,\text{TR}})$, for $z \in \{0, 1\}$ and each (i, j). The values of $\sigma_i^{2,\text{TR}}$ were set to 0.2, 0.1, and 0.3 for samples 1, 2, and 3, respectively. We then simulated latent subpopulation indicators $\lambda_{i,n}^{\text{TR}}$ with probabilities $\Pr(\lambda_{i,n}^{\text{TR}} = 0) = \epsilon_i^{\text{TR}}$ and $\Pr(\lambda_{i,n}^{\text{TR}} = k \mid \lambda_{i,n}^{\text{TR}} \neq 0) = w_{i,k}^{\text{TR}}$. We generated $y_{i,n,j} \overset{\text{iid}}{\sim} \text{N}(0, 9)$ for all (i, n, j) with $\lambda_{i,n}^{\text{TR}} = 0$. Otherwise, we generated $y_{i,n,j}$ from a mixture of normals, $\sum_{\ell=1}^{L^{z,\text{TR}}} \eta_{i,j}^{z,\text{TR}} \cdot \text{N}(\mu_{z\ell}^{\star,\text{TR}}, \sigma_i^{2,\text{TR}})$ given $z_{j,\lambda_{i,n}^{\text{TR}}}^{\text{TR}} = z$ for each (i, n, j). To simulate the missingship indicators, $m_{i,n,j}$, we first generated the proportions $p_{i,j}$ of missing values for each (i, j) from a Unif$(0, 0.7 \cdot \sum_k w_{i,k}^{\text{TR}}(1 - z_{j,k}^{\text{TR}}))$ and sampled $p_{i,j} \times N_i$ cells without replacement with probability proportional to $\{1 + \exp(-9.2 -$



(b)

	Cell Subpopulations				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
sample 1	0.068	0.163	0.351	0.297	0.118
sample 2	0.194	0.282	0.066	0.257	0.199
sample 3	0.112	0.141	0.224	0.119	0.402

w^{TR}

Figure 2. Design of Simulation 1. $\boldsymbol{Z}^{TR}$ and $\boldsymbol{w}^{TR}$ are illustrated in (a) and (b), respectively. $K^{TR}=5$, $J=20$, and $I=3$ are assumed. In (a), black represents $z^{TR}_{j,k}=1$ (marker expression) and white represents $z^{TR}_{j,k}=0$ (marker non-expression).

$2.3y_{i,n,j})\}^{-1}$. We let $y_{i,n,j}=$ NA if $m_{i,n,j}=0$. Under the true missingness mechanism, a marker having a lower expression level has a higher chance of being recorded as missing. Note that the true mechanism is different from that assumed in (5). As the results will show, the model's performance of recovering the true cell subpopulation structure is robust to misspecification of the data missingship mechanism model. Heatmaps of the simulated $\boldsymbol{y}$ are shown in Figure 4b, d and f. The $y_{i,n,j}$'s are sorted within a sample according to their posterior subpopulation indicator estimates $\hat{\lambda}_{i,n}$, which we explain below. The colours red, blue, and black represent high expression levels, low expression levels, and missing values, respectively.

We fit a separate model for each $K=2, 3, \ldots, 10$, fixing $L^0=L^1=5$ and $s^2_\epsilon=10$ for each K. We specified the remaining fixed hyper-parameters as follows: $a_\alpha=b_\alpha=0.1$ for α; $\psi_z=1$ and $\tau^2_z=1$ for $\delta_{z,\ell}$; $a_\sigma=3$ and $b_\sigma=2$ for σ^2_i; $a_{\eta^z}=1$ for $\boldsymbol{\eta}_{i,j}$; $d=1$ for $\boldsymbol{w}_i$; $a_\epsilon=1$ and $b_\epsilon=99$ for ϵ_i. The specification implies a weakly informative prior, except for ϵ_i. The values of a_ϵ and b_ϵ are used to strongly imply that only a small fraction of cells belongs to the noisy cell type, $k=0$. We used the empirical approach described in Section 2 to obtain values of $\boldsymbol{\beta}$ for the missingship mechanism. For each i, we initialized the missing values at $-\beta_{2i}/(2\beta_{1i})$, which corresponds to the largest missing probabilities *a priori*. To initialize $\lambda_{i,n}$, $\boldsymbol{w}_i$, $\boldsymbol{Z}$, α, and $\boldsymbol{\eta}^z_{i,j}$, we applied density-based clustering via finite Gaussian mixture models using the MClust package (Scrucca et al., 2016), and used the resulting clustering of $y_{i,n,j}$. Other reasonable methods can be used for the Markov chain initialization. We then drew samples of $\boldsymbol{\theta}$ and imputed missing values of $y_{i,n,j}$ using MCMC simulation based on 16,000 iterations, discarding the first 10,000 iterations as burn-in for each model, and then thinned by keeping every other draw. We monitored convergence and mixing of the MCMC posterior simulation by inspecting trace plots of the log-likelihood. Online Supplementary Material, Figure 2 shows trace plots of the log-likelihood from two independent chains with different initial values. The plots show only minor differences, indicating that the two chains traced out a common distribution. The burn-in period was chosen via visual inspection of the trace plots of the log-likelihoods. Posterior inference for a model with $K=5$ took 10 hr for 16,000 iterations on an interactive Linux server with four Intel Xeon E5-4650 processors and 512 GB of random access memory.

For each value of K, we computed the LPML and DIC, and obtained point estimates $\hat{\boldsymbol{Z}}_i$, $\hat{\boldsymbol{w}}_i$ and $\hat{\boldsymbol{\lambda}}_i$ using the method described in Section 2.3. Figure 3a,b, respectively, show plots of LPML and DIC as functions of K. Figure 3c plots LPML against the number of subpopulations with $\hat{w}_{i,k}<1\%$. The increase in LPML is very minimal, while negligible subpopulations are added for values of $K>5$. The plots clearly indicate that $\hat{K}=5$ yields a parsimonious model with good fit. Figure 4 illustrates $\hat{\boldsymbol{Z}}_i$, $\hat{\boldsymbol{w}}_i$ and $\hat{\lambda}_{i,n}$ for $\hat{K}=5$. Panels (a), (c) and (e) show $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ for

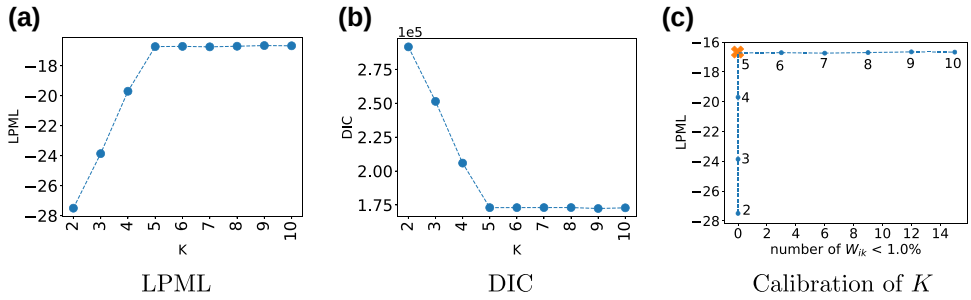


Figure 3. Results of Simulation 1. Plots of (a) LPML = log pseudo marginal likelihood, (b) DIC = deviance information criterion, and (c) calibration metric, for $K = 2, \ldots, 10$.

samples 1, 2, and 3, respectively. The subpopulations with $\hat{w}_{ik} > 1\%$ are included in the plots of $\hat{\boldsymbol{Z}}_i$. The estimates $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are close to their truth values in Figure 2 for all samples, implying that the true cell population structure is well recovered. We quantified the proximity between the point estimates, $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ and their truth using the metrics, d_i^Z and d_i^w defined in Online Supplementary Material, Section 4. The metrics also indicate that the point estimates are close to their truth. More details are in Online Supplementary Material, Section 4.

We compared the resulting clustering of the cells by $\hat{\lambda}_{i,n,j}$ to the truth. We used the adjusted Rand index (ARI) (Hubert & Arabie, 1985), which measures the agreement between two sets of clusterings. A larger value implies greater agreement, and in the case of random clusterings, ARI is expected to be 0. ARI can be negative in cases where the agreement between clusters is less than what is expected from random clusterings. The obtained ARIs are above 0.99 for all samples, indicating that the model recovers the true cell clusters very well. The heatmaps of y rearranged by cell clustering membership estimates $\hat{\lambda}_{i,n}$ are shown in panels (b), (d), and (f) of Figure 4, where the colours, red, blue, and black represent high, low, and missing expression levels, respectively. The horizontal yellow lines separate cells by $\hat{\lambda}_{i,n}$. The figures also show that the cell clustering based on the estimated subpopulations captures the true clustering of y quite well.

We also fit the model to the simulated data using ADVI, with a mini-batch size of 2,000, $K = 30$, and 20,000 iterations. The time required to fit the model was approximately 6 hr for 20,000 iterations, which is substantially faster than that of the analogous MCMC method. Online Supplementary Material, Figure 3 shows the posterior estimates of $\boldsymbol{Z}$, $\boldsymbol{w}$, and $\lambda_{i,n}$ obtained via ADVI. Inferences for model parameters using ADVI are similar to those using MCMC. The simulation truth for the model parameters $\boldsymbol{\theta}$ are well recovered, as in the MCMC implementation.

We assessed the sensitivity of the model to the data missingship mechanism by fitting the FAM using different specifications of $\boldsymbol{\beta}$ with $K = \hat{K}$, and comparing the inferences. The two different specifications of $\boldsymbol{\beta}$ are given in Online Supplementary Material, Table 3. The estimates of $\boldsymbol{\theta}$ do not change significantly across different specifications of $\boldsymbol{\beta}$. Point estimates of $\boldsymbol{Z}$, $\boldsymbol{w}_i$, and $\lambda_{i,n}$ are shown in Online Supplementary Material, Figures 4 and 5. The estimates $\hat{\boldsymbol{Z}}$ remain the same for all specifications of $\boldsymbol{\beta}$, and the $\hat{\boldsymbol{w}}_i$ values also are very similar. Online Supplementary Material, Table 3 shows that LPML and DIC are slightly better for the data missingship mechanisms that encourage imputing smaller missing values $y_{i,n,j}$. This results in $\mu^{\star}_{0,\mathrm{L}_0}$, the smallest of the mixture component locations for non-expressed markers, being smaller than that obtained under the other specifications, accidentally more closely resembling the simulation truth. Details of the sensitivity analysis are in Online Supplementary Material, Section 4.

We compared our model via simulation to FlowSOM in Van Gassen et al. (2015), which is implemented in the R package FlowSOM (Van Gassen et al., 2017). FlowSOM fits a model with a varying number of clusters and selects a value of K that minimizes the within-cluster variance while also minimizing the number of clusters via an 'elbow' criterion, an *ad hoc* graphical method that chooses K such that $K + 1$ does not substantially increase the percentage of variation explained. FlowSOM does not impute missing values, so we used all y assuming that there is no missing y. In practice, missing values could be pre-imputed, or multiple imputation could be employed. Note that FlowSOM does not account for variability between samples. We combined the samples

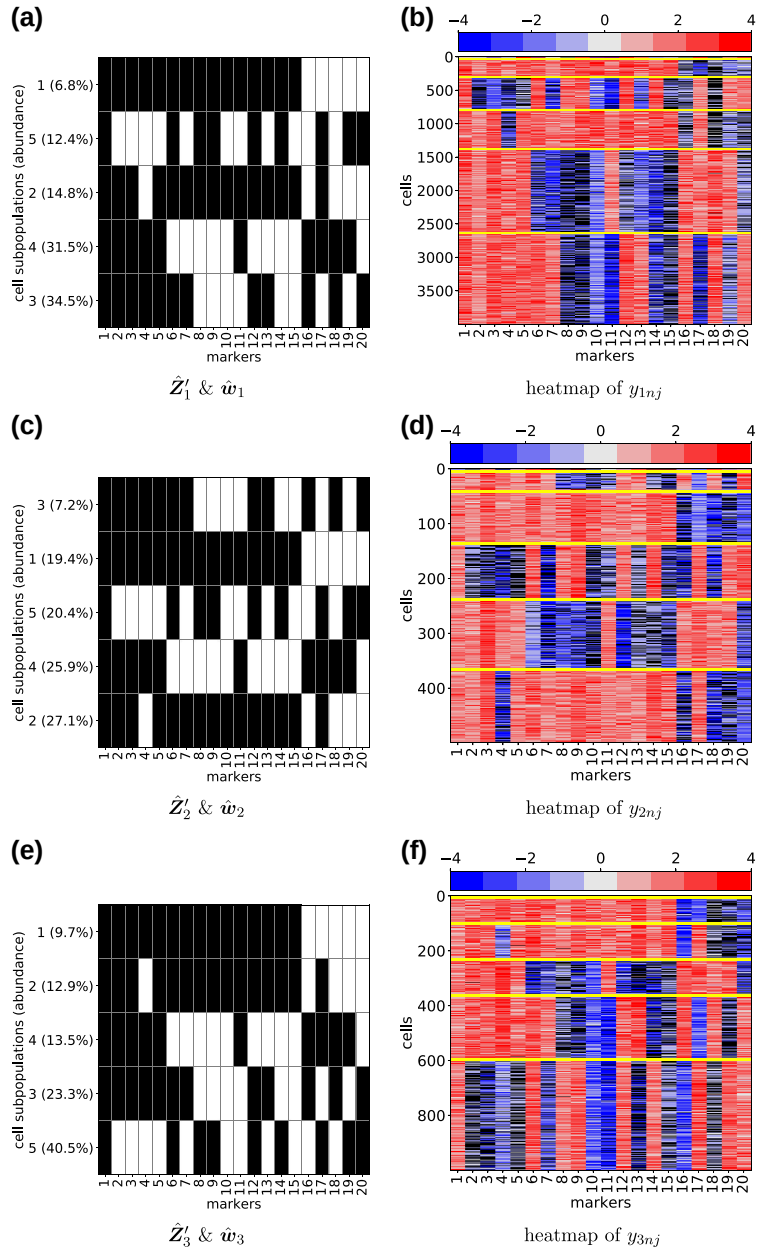


Figure 4. Results of Simulation 1. In (a,c), the transpose $\hat{\boldsymbol{Z}}_i'$ of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are shown for samples 1 and 2, respectively, with markers that are expressed denoted by black and not expressed by white. Only subpopulations with $\hat{w}_{ik} > 1\%$ are included. Heatmaps of $\boldsymbol{y_i}$ are shown for sample 1 in (b) and sample 2 in (d). Cells are given in rows and markers are given in columns, with cells ordered by posterior point estimates of their subpopulation indicators, $\hat{\lambda}_{i,n}$. High and low expression levels are represented by red and blue, respectively, and black represents missing values. Yellow horizontal lines separate cells into five subpopulations. In (e), the transpose $\hat{\boldsymbol{Z}}_i'$ of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are shown for sample 3, with markers that are expressed denoted by black and not expressed by white. Only subpopulations with $\hat{w}_{ik} > 1\%$ are included. Heatmaps of $\boldsymbol{y_i}$ for sample 3 is shown in (f). Cells are given in rows and markers are given in columns, with cells ordered by posterior point estimates of their subpopulation indicators, $\hat{\lambda}_{i,n}$. High and low expression levels are represented by red and blue, respectively, and black represents missing values. Yellow horizontal lines separate cells into five subpopulations.

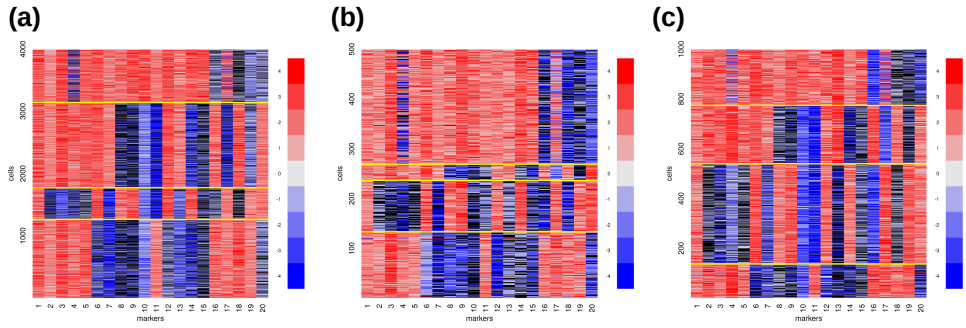


Figure 5. Results of Simulation 1 (continued). Heatmaps of $\boldsymbol{y_i}$ for clusters estimated by FlowSOM, with cells ordered by the cluster labels $\lambda_{i,n}$. Cells are in rows and markers are in columns. High, low, and missing expression levels are in red, blue, and black, respectively. Yellow horizontal lines separate the identified cell clusters. (a) Sample 1, (b) sample 2, and (c) sample 3.

Table 1. Adjusted rand index (ARI) for FAM and the comparators, FlowSOM, FlowMeans, and PhenoGraph, by sample for simulation 1

Method	Sample 1	Sample 2	Sample 3
FAM ($K = 5$)	0.999	0.993	0.999
FlowSOM	0.945	0.738	0.935
FlowMeans	0.949	0.732	0.938
PhenoGraph	0.977	0.912	0.968

Note. Higher ARI is better, and values closer to 1 indicate that estimated clusters are closer to the truth.

for analysis to avoid a further *ad-hoc* process of finding common clusters among the samples. If desired, one might do separate analyses for each of the samples. FlowSOM was considerably faster than our model, with a computation time of 6 s on the simulated dataset. FlowSOM identified four cell clusters, as summarized in Figure 5, where the cells are rearranged by their cluster membership estimates in each sample. The fourth cluster (shown near the top of the heatmaps) is a mix of the cells having the true subpopulations 1 and 2 that differ only by markers 4 and 17, and its performance of cell clustering deteriorates. We again computed the ARI to compare the clustering estimates obtained by FlowSOM to the truth. The ARIs obtained under FlowSOM are 0.945, 0.738, and 0.935 for samples 1, 2, and 3, respectively. The ARI in sample 2 is especially low for FlowSOM because the two cell subpopulations combined by FlowSOM have large abundances in the sample. Table 1 summarizes the ARIs from FAM with $K = 5$ and FlowSOM, and shows that our FAM outperforms FlowSOM in estimation of cell clustering. More importantly, FlowSOM does not provide a model or inferences for the latent structure of cell subpopulations. For this simulation scenario, the FAM easily recovers the truth, but a clustering-based method such as FlowSOM may perform poorly in cell clustering. In addition, we compared our FAM to FlowMeans (Aghaeepour et al., 2011) and PhenoGraph (Levine, Simonds, Bendall, Davis, Tadmor, et al., 2015; Levine, Simonds, Bendall, Davis, El-ad, et al., 2015). Similar to FlowSOM, they are cell clustering algorithms based on marker expression levels and available in R and Python, respectively. Specifically, FlowMeans is a *K*-means-based clustering algorithm and automatically selects the number of clusters using a change point detection algorithm. PhenoGraph constructs a nearest-neighbour graph of cells that represents the phenotypic relationships between cells and partitions the graph into subpopulations of similar cells. As summarized in Table 1, compared to the FAM, these methods yield lower ARI values for all samples. FlowMeans found four cell clusters by combining the true subpopulations 1 and 2, resulting in poor cell

clustering. On the other hand, PhenoGraph found seven clusters by including two redundant clusters. More discussion of the comparison is included in Online Supplementary Material, Section 4.1.

We further examined the performance of our FAM in an additional simulation study, Simulation 2, in which we kept most of the set-up used in Simulation 1, but assumed a more complex subpopulation structure with much larger numbers of cells, by assuming $K^{TR} = 10$ and $N = (40{,}000, 5{,}000, 10{,}000)$. $\boldsymbol{Z}^{TR}$ and $\boldsymbol{w}_i^{TR}$ are illustrated in Online Supplementary Material, Figure 8. We considered ten models with $K = 2, 4, \ldots, 20$. For the fixed hyperparameters, we let $L^0 = L^1 = 5$, and the remaining specifications for hyperparameters were the same as those in Simulation 1. The model comparison metrics strongly suggest $\hat{K} = 10$, for which the posterior point estimates of the underlying structure, including $\boldsymbol{Z}$, $\boldsymbol{w}$, and $\lambda_{i,n}$ recover the simulation truth quite well, as shown in Online Supplementary Material, Figure 11. In contrast, in this case, FlowSOM groups cells into two subpopulations that have similar configurations, similarly to Simulation 1, and estimates nine cell clusters. The FAM provides direct inference on cell subpopulations, and the cell clustering by subpopulations is superior to that obtained by the comparators. Details of Simulation 2, including a sensitivity analysis for the data missingship mechanism and fast computation using ADVI, are given in Online Supplementary Material, Section 4.2.

4 Analysis of cord blood-derived NK cell data

We next report an analysis of the CyTOF dataset of surface marker expression levels on UCB-derived NK cells. Identifying and characterizing NK cell subpopulations in terms of marker expression may serve as a critical step to identifying NK cell subpopulations to develop disease-specific therapies for a variety of severe hematologic malignancies. To gain insight into the phenotype of cord blood-derived NK cells, CyTOF was used with a customized panel including 32 antibodies against well-established inhibitory and activating receptors, as well as differentiation, homing, and cytotoxicity markers relevant to NK cell biology and function. Our NK cell dataset consists of three samples collected from different cord blood donors, containing 41,474, 10,454, and 5,177 cells, respectively. We first obtained the cut-off values $c_{i,j}$ using flowDensity and computed the transformed raw expression levels, $y_{i,n,j} = \log(\tilde{y}_{i,n,j}/c_{i,j})$ if $m_{i,n,j} = 1$ as explained in Section 2.1. We let $y_{i,n,j} = \text{NA}$ if $m_{i,n,j} = 0$. Because markers that are either expressed or not expressed in most of cells are not informative for constructing subpopulations under our FAM, we removed markers having positive values in more than 90% of the cells in all samples, or with missing or negative values in over 90% of the cells in all samples. We also removed all cells with an expression level $y_{i,n,j} < -6$ for any marker. This accounted for only a very small number of cells, and it encourages imputed marker expression levels to be in a reasonable range. Thus, we recommend removing outliers in this fashion. After this preprocessing, $J = 20$ markers remained and the numbers of cells in the samples were N_i= 38,636, 9,555, and 4,827 for subsequent analysis. Online Supplementary Material, Table 6 lists the markers included in the analysis. Figure 7b, d, and e shows heatmaps of $\boldsymbol{y}$ after rearranging the cells by posterior estimates $\hat{\lambda}_{in}$ of the cell clusterings for each sample. Using a threshold of 90% to remove some markers yields a reasonable set of markers, but may seem arbitrary. We performed the analyses with different choices of the threshold, such as 0.85 and 0.95. The results are presented in Online Supplementary Material, Section 5. We also plotted the data using the data visualization technique 't-SNE (t-Distributed Stochastic Neighbour Embedding)' in Online Supplementary Material, Figure 18a–c. t-SNE is a popular method for visualization of high-dimensional data in a two- or three-dimensional map through stochastic neighbour embedding (Maaten & Hinton, 2008; Van Der Maaten, 2014). It also is used for detecting clusters in data. We used Barnes-Hut-SNE implemented in the Python library *sklearn* to obtain two-dimensional t-SNE embeddings separately for each sample. For comparison, Online Supplementary Material, Figure 18d-(ℓ) plots the obtained two-dimensional t-SNEs colour-coded by the clusterings estimated by the comparators. We fit our FAM over a grid for K from 3 to 33 in increments of 3, as opposed to increments of 1, due to constraints on computational resources available to us. We set $L_0 = 5$ and $L_1 = 3$. We set priors and the data missingship mechanism as outlined in Section 3. The specified values of the fixed hyperparameters allow a reasonable amount of prior uncertainty, and with the large values of N_i, the prior has a small effect on the posterior inference. Also, as will be shown below, the model's performance is not

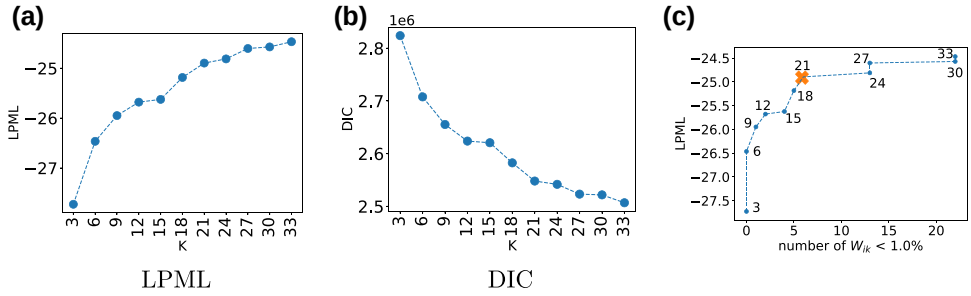


Figure 6. Analysis of UCB-derived NK cell data. Plots of (a) LPML, (b) DIC, and (c) calibration metric, for $K = 3, 6, \ldots, 33$.

sensitive to the specification of $\boldsymbol{\beta}_i$. Random parameters $\boldsymbol{\theta}$ also were initialized in a similar manner. 6,000 samples from the posterior distribution of the model parameters were obtained after a burn-in of 10,000 iterations. The posterior samples were thinned by selecting every other sample to yield a total of 3,000 samples. As done in the simulation study, we monitored convergence and mixing of the MCMC posterior simulation by inspecting trace plots of the log-likelihood. Online Supplementary Material, Figure 17 shows trace plots of the log-likelihood from two independent chains with different initial values. The plots show only minor differences, indicating that the two chains traced out a common distribution. The burn-in period was, again, chosen via visual inspection of the trace plots of the log-likelihoods.

Figure 6a and b display LPML and DIC as functions of K. The LPML changes sharply for small values of K, and tapers at $K = 21$, indicating that $\hat{K} = 21$. A similar pattern is seen for DIC. As depicted in Figure 6c, our additional calibration method also indicates that the models with $K > 21$ include more cell subpopulations comprising less than one per cent of a sample (i.e., $\sum_{i,k} \hat{w}_{i,k} < 1\%$ is larger), and improve fit only minimally.

Figure 7 summarizes posterior inference on the latent cell population structure with $\hat{K} = 21$. The cells are grouped by their estimated cell subpopulation indicators $\hat{\lambda}_{i,n}$. The figure shows the estimated cell subpopulations $\hat{\boldsymbol{Z}}_i$ (in the left column) and clustered marker expression levels $\boldsymbol{y}_i$ (in the right column) for the samples. Cells having subpopulations with larger $\hat{w}_{i,k}$ are shown at the bottom of the heatmaps. The subpopulations with the two largest $\hat{w}_{i,k}$ are different in the samples. The resulting inference indicates that the composition of the NK cell population varies across the samples, pointing to variations in the phenotype of NK cells among different cord blood donors. We observe similarities in the phenotypes of NK cells from samples 2 and 3, however, while sample 1 displays a different phenotype and a distinct distribution of cell subsets. NK cells from all three samples express 2B4, CD94, DNAM-1, NKG2A, NKG2D, Siglec-7, NKp30, and Zap70 in the majority of their identified subpopulations. These markers dictate NK cell functional status. While their interactions are very complicated, taken together they provide a basis for determining whether NK cells have a normal function, and whether they are mature or not.

Despite great variability between cord blood sample 1 and the other two cord blood samples, all three had a significant subset of cells with an immature phenotype. Cord blood 1 Cluster 7, cord blood 2 Cluster 17 and cord blood 3 Cluster 6 comprise the largest population of immature cells, defined as EOMES (−), TBET (−), and KIR (−). Markers KIR2DL3 and KIR3DL1 belong to killer-cell immunoglobulin-like receptors (KIRs). These immature clusters of NK cells still retain expression of 2B4, NKG2A, NKG2D, CD94 and NKp30. In particular, NKp30 is a natural cytotoxicity receptor, while KIR is not. This implies that, despite great variability between sample 1 and the other two samples, all three have a significant subset of cells with an immature phenotype. Markers EOMES, TBET, Zap70 and KIR are not expressed in the largest subpopulation of each sample, indicating that those are subsets of immature cells. An immature phenotype of NK cells usually is associated with low diversity and low effector function in the absence of exogenous cytokines, Li et al. (2019) and Sarvaria et al. (2017), while a mature NK cell phenotype has been linked to superior cytotoxicity and better clinical outcomes in cancer patients (Carlsten & Jaras,

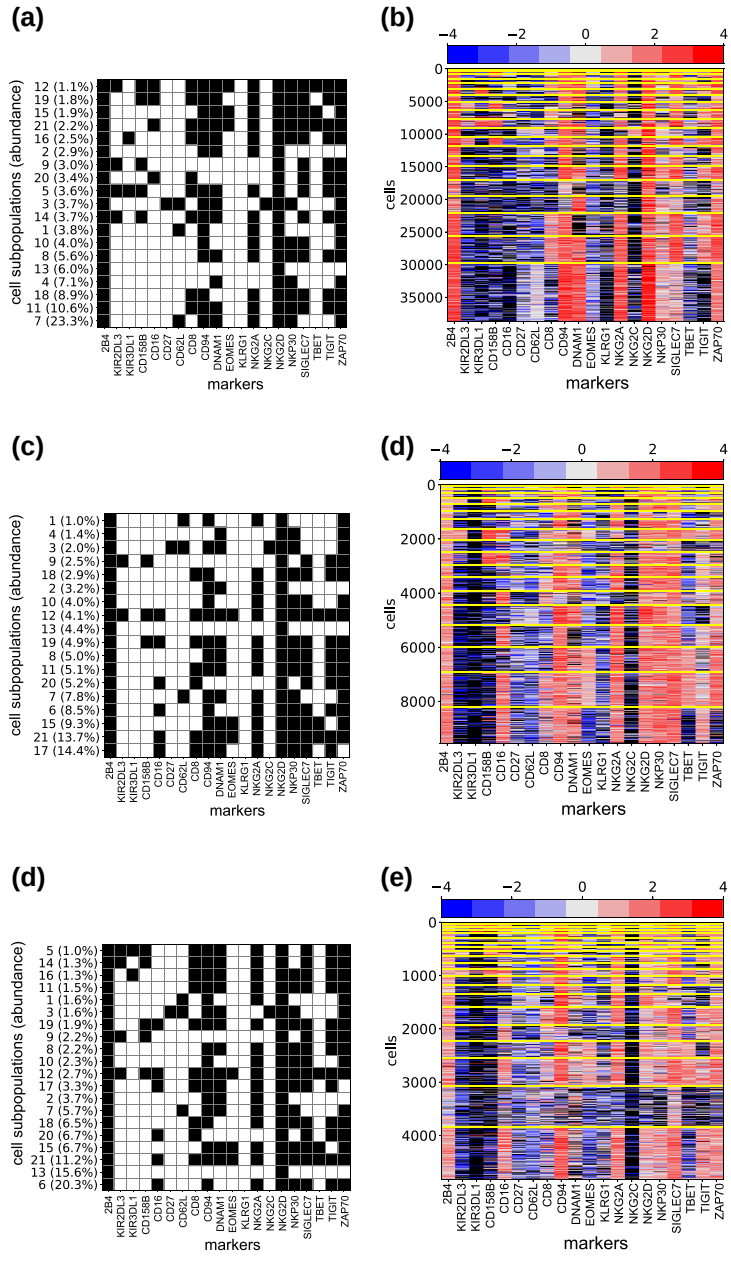


Figure 7. Analysis of the UCB-derived NK cell data. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ of samples $i = 1$ and 2 are illustrated in panels (a) and (c), respectively, with markers that are expressed denoted by black and not expressed by white. Only subpopulations with $\hat{w}_{ik} > 1\%$ are included. Heatmaps of expression level $\boldsymbol{y_i}$ are shown in panels (b) and (d) for samples 1 and 2, respectively, with cells in rows and markers columns. Each column thus contains the expression levels of one marker for all cells in a sample. High, low, and missing expression levels are red, blue, and black, respectively. Cells are ordered by the posterior estimates of their clustering memberships, $\hat{\lambda}_{i,n}$. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ of sample 3 are illustrated in panel (e), with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{ik} > 1\%$ are included. Heatmaps of $\boldsymbol{y_i}$ are shown in panel (f) for sample 3. Cells are in rows and markers in columns. Each column contains the expression levels of a marker for all cells in the sample. High, low, and missing expression levels are red, blue, and black, respectively. Cells are ordered by the posterior estimates of their clustering memberships, $\hat{\lambda}_{i,n}$. Yellow horizontal lines separate cells by different subpopulations. (a) $\hat{\boldsymbol{Z}}_1'$ and $\hat{\boldsymbol{w}}_1$. (b) Clustering of y_{1nj}. (c) $\hat{\boldsymbol{Z}}_2'$ and $\hat{\boldsymbol{w}}_2$. (d) Clustering of y_{2nj}. (e) $\hat{\boldsymbol{Z}}_3'$ and $\hat{\boldsymbol{w}}_3$ and (f) Clustering of y_{3nj}.

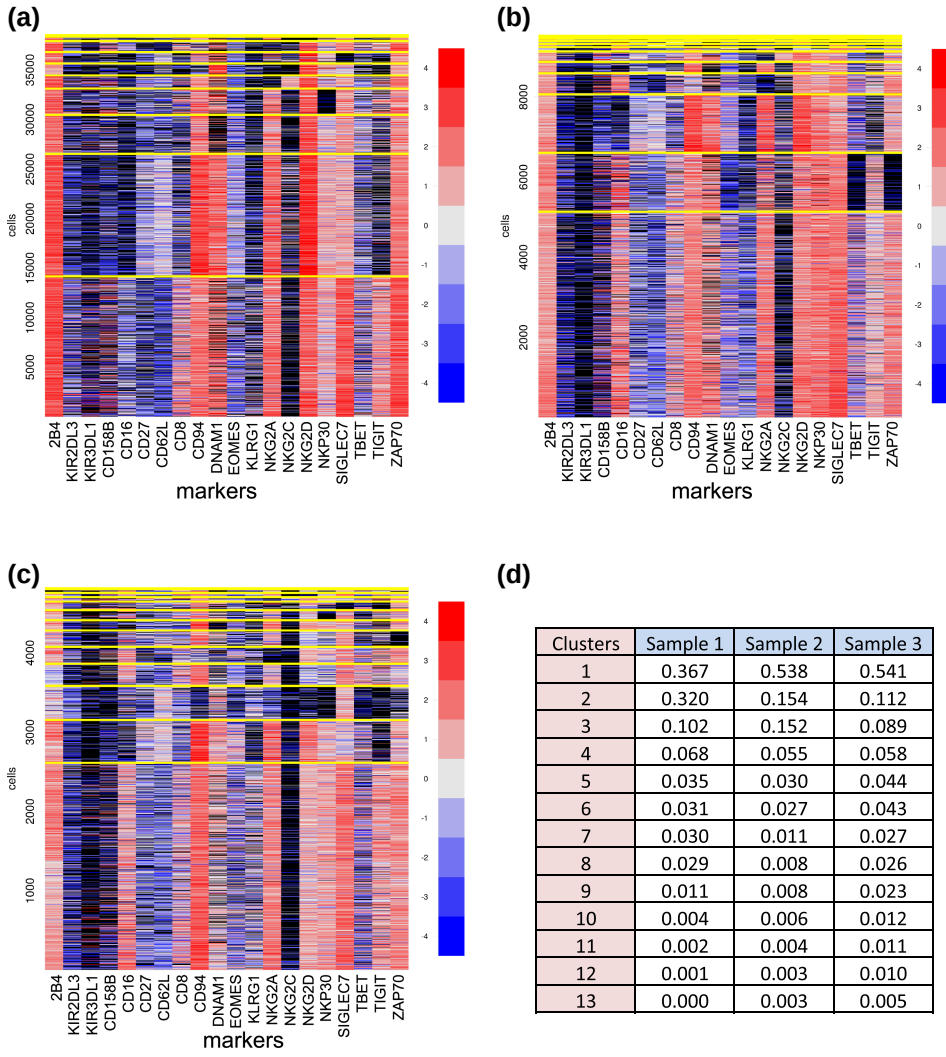
2019; Ilander et al., 2017). These immature clusters of NK cells still retain expression of 2B4, CD94, NKG2A, NKG2D, and NKp30.

In addition, we identified three subpopulations (12, 15, and 21) that are conserved among the three samples, although at lower percentages in sample 1. In these subpopulations, EOMES and TBET are expressed, indicating that they are a more mature phenotype. The subset with expression of EOMES and TBET could be further divided into three subpopulations based on the expressions of markers CD8, CD16, TIGIT, and KIR. Subpopulations 12 and 21 are very similar, sharing positivity for CD16, CD8 and TIGIT, and are differentiated by KIR expression, which are negative in subpopulation 21 and positive in subpopulation 12. Subpopulation 15, however, is negative for CD16, CD8, TIGIT and KIR, making EOMES and TBET its only differentiation markers. These novel subsets of cord blood NK cells have not been described in the literature previously, and may need to be further validated. We also identified cluster 3 as an important conserved cluster among all 3 samples, which is positive for NKG2C, CD62L and CD27, which could indicate a memory subset in cord blood NK cells which has not been well described previously. Taken together, these data indicate that the FAM allows not only the definition of biologically recognized subsets of NK cells, but also may be applied for the discovery of novel NK cell subpopulations.

Model sensitivity to the specification of the data missingship mechanism in the NK cell data analysis was assessed by fitting the FAM under two additional specifications of $\boldsymbol{\beta}$, which we call data missingship mechanisms (MM) I and II. We will refer to the previous (default) missingship mechanism as MM-0. Online Supplementary Material, Tables 7 and 8 list the different data missingship mechanism specifications and the corresponding $\boldsymbol{\beta}$ values, respectively. Under the different specifications of $\boldsymbol{\beta}$, the estimates $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are similar, as shown in Online Supplementary Material, Figures 20 and 21. The subpopulations estimated under MM-I and MM-II are identical to or differ by no more than three markers, when compared to those under MM-0.

We also fit the model to the UCB-derived NK cell data computing posteriors using ADVI with a mini-batch size of 2,000 and $K = 30$ for 20,000 iterations. The runtime was approximately 6 hr on the previously described machine. Online Supplementary Material, Figure 22 summarizes the posterior distribution of $\boldsymbol{Z}$ and the posterior mode of cell clusterings $\hat{\lambda}_{i,n}$. The cell subpopulations inferred by ADVI are similar to those obtained by MCMC, but the cell clustering estimates are quite different. Notably, subpopulations with large $\hat{w}_{ik}$ can be found in the estimates obtained by both methods, e.g., the subpopulations with the largest abundances in sample 1. For subpopulations with smaller $\hat{w}_{ik}$, we do not find clear matches. The cluster sizes obtained by ADVI are larger than those obtained from MCMC and cells in the clusters are less homogeneous. It thus appears that ADVI should be used very cautiously in this type of setting, and that its shorter runtime compared to MCMC may be a false economy.

For comparison, we also applied the comparators to the UCB data. We fixed the missing values of $y_{i,n,j}$ at the minimum of the negative observed values of y for each (i, j) prior to analysis. FlowSOM identified 13 cell clusters in the samples. Heatmaps of $y_{i,n,j}$ rearranged by cell clustering estimates by FlowSOM are given in Figure 8a–c. Heterogeneity between cells within clusters estimated under FlowSOM is noticeably greater than that under the proposed FAM shown in Figure 7. For example, marker 10 shows a mix of red, blue, and black colours for cluster 1, the largest cluster. The proportions of cells assigned to the clusters are summarized in Figure 8d. The clusters obtained by FlowSOM are much larger than those obtained by the FAM. In particular, cluster 1 under FlowSOM contains 36.7%, 53.8%, and 54.1% of the cells in samples 1–3, respectively. The cluster estimates by FlowMeans and PhenoGraph are presented in Online Supplementary Material, Figures 18 and 19. FlowMeans produces a cluster that contains 74.82%, 66.44%, and 74.77% of the cells in the samples, respectively. On the other hand, the clustering estimate by PhenoGraph has many small clusters, with the largest clusters of the samples containing 7.53%, 13.09%, and 15.02% of the cells. More details are presented in Online Supplementary Material, Section 5 Lastly, note that the comparators do not produce an explicit inference on the characterization of subpopulations. Online Supplementary Material, Table 1 summarizes the time required to fit each of the various models (FAM-MCMC, FAM-ADVI, PhenoGraph, FlowSOM, FlowMeans) to each of the two simulated data sets and the real UCB data.



Clusters	Sample 1	Sample 2	Sample 3
1	0.367	0.538	0.541
2	0.320	0.154	0.112
3	0.102	0.152	0.089
4	0.068	0.055	0.058
5	0.035	0.030	0.044
6	0.031	0.027	0.043
7	0.030	0.011	0.027
8	0.029	0.008	0.026
9	0.011	0.008	0.023
10	0.004	0.006	0.012
11	0.002	0.004	0.011
12	0.001	0.003	0.010
13	0.000	0.003	0.005

Figure 8. [CB Data: Comparison to FlowSOM] Heatmaps of cells in (a)–(c) for samples 1–3, respectively. Cells are arranged by the cluster membership estimates by FlowSOM. The clusters are separated by yellow horizontal lines, with the most abundant clusters in each sample closer to the bottom. High, low, and missing expression levels are red, blue, and black, respectively. The proportions of the cells in the estimated clusters are shown in (d). (a) Clustering of y_{1nj}, (b) clustering of y_{2nj}, (c) clustering of y_{3nj}, and (d) proportions.

5 Discussion

We have proposed a Bayesian FAM to identify and estimate cell subpopulations using CyTOF data. Our FAM identifies latent subpopulations, defined as functions of the marker expression levels, and fits the data in multiple samples simultaneously. The model accounts formally for missing values and between-sample variability. The fitted FAM assigns each cell in each sample to exactly one subpopulation, but each surface marker can belong to more than one subpopulation. The method also yields cell clusters within each sample that are defined in terms of the inferred subpopulations. We constructed a data missingship mechanism based on expert knowledge, and we examined the robustness of the model to the specification of the missingship mechanism through simulation. This showed that inferences were not sensitive to changes in the specification of the missingship mechanism. Compared to established clustering methods, including FlowSOM, the proposed FAM is more effective at discovering latent subpopulations when the underlying cell subpopulations are similar.

Our proposed FAM can be extended to accommodate similar but more complex data structures, in particular, data including covariates. For example, samples with similar covariates may also have similar cell subpopulation structures. The model can incorporate such information by incorporating appropriate regression submodels, to enhance inferences and study how the structures may change with covariates. One also may introduce the concept of 'repulsiveness' to latent features and obtain a more parsimonious representation of the latent subpopulations by discouraging the creation of redundant subpopulations. Repulsive models, which are more likely to produce features that differ from each other substantially, have been developed mostly in the context of mixture models (e.g., see Petralia et al., 2012; Quinlan et al., 2018; Xie & Xu, 2019). Xu et al. (2016) used the detrimental point process (DPP) for a repulsive FAM that uses the determinant of a matrix as a repulsiveness metric. A model that explicitly penalizes the inclusion of similar features also can be developed to replace the IBP in our model.

Conflict of interest: The authors have no conflicts of interest to declare.

Funding

This work was supported by the NIH 1 R01 CA211044-01, 5 P01CA148600-03, and P50CA100632-16 (K.R.), a grant (CA016672) to the M.D. Anderson Cancer Center from the NIH (K.R.) and the NSF grant DMS-1662427 (J.L.).

Data availability

The umbilical cord blood data used in this work is publicly available at https://github.com/luiarthur/cytof-data.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series C* online.

References

Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G. S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., & Mané D. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.

Aghaeepour N., Nikolic R., Hoos H. H., & Brinkman R. R. (2011). Rapid cell population identification in flow cytometry data. *Cytometry Part A*, *79*(1), 6–13. https://doi.org/10.1002/cyto.a.21007

Allison P. D. (2001). *Missing data (Vol. 136)*. Sage Publications.

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. https://doi.org/10.1080/01621459.2017.1285773

Carlsten M., & Jaras M. (2019). Natural killer cells in myeloid malignancies: Immune surveillance, NK cell dysfunction, and pharmacological opportunities to bolster the endogenous NK cells. *Frontiers in Immunology*, *10*, 2357. https://doi.org/10.3389/fimmu.2019.02357

Celeux G., Hurn M., & Robert C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957–970. https://doi.org/10.1080/01621459.2000.10474285

Chen H., Lau M. C., Wong M. T., Newell E. W., Poidinger M., & Chen J. (2016). Cytofkit: A bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS Computational Biology*, *12*(9), e1005112. https://doi.org/10.1371/journal.pcbi.1005112

Chen M., Gao C., & Zhao H. (2013). 'Phylogenetic Indian buffet process: Theory and applications in integrative analysis of cancer genomics', arXiv, arXiv:1307.8229, preprint: not peer reviewed.

Cheung R. K., & Utz P. J. (2011). Screening: CyTOF—The next generation of cell detection. *Nature Reviews Rheumatology*, *7*(9), 502–503 https://doi.org/10.1038/nrrheum.2011.110

Dahl D. B., & Müller P. (2017). Summarizing distributions of latent structures. *Bayesian Nonparametric Inference: Dependence Structures & Applications*, Oaxaca, Mexico.

Ester M., Kriegel H.-P., Sander J., & Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings for the second international conference on knowledge discovery and data mining*, *96*, 226–231.

Franks A. M., Airoldi E. M., & Rubin D. B. (2016). 'Non-standard conditionally specified models for non-ignorable missing data', arXiv, arXiv:1603.06045, preprint: not peer reviewed.
Frühwirth-Schnatter S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.
Geisser S., & Eddy W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160. https://doi.org/10.1080/01621459.1979.10481632
Gelfand A. E., & Dey D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 501–514. https://doi.org/10.1111/j.2517-6161.1994.tb01996.x
Ghahramani Z., & Griffiths T. L. (2006). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*(32), 1185–1224. http://jmlr.org/papers/v12/griffiths11a.html
Griffiths T. L., & Ghahramani Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.
Hai-son P. L., & Bar-Joseph Z. (2011). Inferring interaction networks using the IBP applied to microRNA target prediction. *Advances in Neural Information Processing Systems*, *24*.
Hubert L., & Arabie P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075
Ilander M., Olsson-Stromberg U., Schlums H., Guilhot J., Brück O., Lähteenmäki H., Kasanen T., Koskenvesa P., Söderlund S., Höglund M., & Markevärn B. (2017). Increased proportion of mature NK cells is associated with successful imatinib discontinuation in chronic myeloid leukemia. *Leukemia*, *31*(5), 1106–1116. https://doi.org/10.1038/leu.2016.360
Innes M. (2018). Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, *3*(25), 602. https://doi.org/10.21105/joss.00602
Jasra A., Holmes C. C., & Stephens D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67. https://doi.org/10.1214/088342305000000016
Johnsson K., Wallin J., & Fontes M. (2016). Bayesflow: Latent modeling of flow cytometry cell populations. *BMC Bioinformatics*, *17*(1), 25. https://doi.org/10.1186/s12859-015-0862-z
Kucukelbir A., Tran D., Ranganath R., Gelman A., & Blei D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(14), 1–45. http://jmlr.org/papers/v18/16-107.html
Lanier L. L. (2008). Up on the tightrope: Natural killer cell activation and inhibition. *Nature Immunology*, *9*(5), 495–502. https://doi.org/10.1038/ni1581
Lee J., Müller P., Gulukota K., & Ji Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics*, *9*(2), 621–639. https://doi.org/10.1214/15-AOAS817
Lee J., Müller P., Sengupta S., Gulukota K., & Ji Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *65*(4), 547–563. https://doi.org/10.1111/rssc.12136
Levine J. H., Simonds E. F., Bendall S. C., Davis K. L., El-ad D. A., Tadmor M. D., Litvin O., Fienberg H. G., Jager A., Zunder E. R., Finck R., Gedman A. L., Radtke I., Downing J. R., Pe'er D., & Nolan G. P. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, *162*(1), 184–197. https://doi.org/10.1016/j.cell.2015.05.047
Levine J. H., Simonds E. F., Bendall S. C., Davis K. L., Tadmor M. D., Litvin O., Fienberg H. G., Jager A., Zunder E. R., Finck R., Gedman A. L., Radtke I., Downing J. R., Pe'er D., & Nolan G. P. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, *162*(1), 184–197. https://doi.org//10.1016/j.cell.2015.05.047
Li L., Cheen H., Marin D., Xi Y., Miao Q., Lv J., Banerjee P. P., Shaim H., Daher M., Basar R., & Imahashi N. (2019). A novel immature natural killer cell subpopulation predicts relapse after cord blood transplantation. *Blood Advances*, *3*(23), 4117–4130. https://doi.org/10.1182/bloodadvances.2019000835
Liu E., Marin D., Banerjee P., Macapinlac H., Thall P., & Rezvani K. (2020). IL-15 armored car-transduced NK cells against CD19 positive B cell tumors. *New England Journal of Medicine*, *382*(6), 545–553. https://doi.org/10.1056/NEJMoa1910607
Lo K., Hahne F., Brinkman R. R., & Gottardo R. (2009). flowClust: A bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, *10*(1), 145. https://doi.org/10.1186/1471-2105-10-145
Lupo K. B., & Matosevic S. (2019). Natural killer cells as allogeneic effectors in adoptive cancer immunotherapy. *Cancers*, *11*(6), 769. https://doi.org/10.3390/cancers11060769
Maaten L. v. d., & Hinton G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*-(2605), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html
Malek M., Taghiyar M. J., Chong L., Finak G., Gottardo R., & Brinkman R. R. (2014). flowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, *31*(4), 606–607. https://doi.org/10.1093/bioinformatics/btu677
Miller J. S., Soignier Y., Panoskaltsis-Mortari A., McNearney S. A., Yun G. H., Fautsch S. K., McKenna D., Le C., Defor T. E., Burns L. J., & Orchard P. J. (2005). Successful adoptive transfer and in vivo expansion of human

haploidentical NK cells in patients with cancer. *Blood*, *105*(8), 3051–3057. https://doi.org/10.1182/blood-2004-07-2974

Miller J. W., & Dunson D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, *114*(527), 1113-1125. https://doi.org/10.1080/01621459.2018.1469995

Ni Y., Mueller P., & Ji Y. (2018). 'Bayesian double feature allocation for phenotyping with electronic health records', arXiv, arXiv:1809.08988, preprint: not peer reviewed.

Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L., & Lerer A. (2017). Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*: The Future of Gradient-based Machine Learning Software and Techniques.

Petralia F., Rao V., & Dunson D. B. (2012). Repulsive Mixtures. *Advances in Neural Information Processing Systems*, *25*, 1889–1897.

Quinlan J. J., Page G. L., & Quintana F. A. (2018). Density regression using repulsive distributions. *Journal of Statistical Computation and Simulation*, *88*(15), 2931–2947. https://doi.org/10.1080/00949655.2018.1491578

Rezvani K., & Rouce R. H. (2015). The application of natural killer cell immunotherapy for the treatment of cancer. *Frontiers in Immunology*, *6*, 578. https://doi.org/10.3389/fimmu.2015.00578

Rubin D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, *69*(346), 467–474. https://doi.org/10.1080/01621459.1974.10482976

Rubin D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sarvaria A., Jawdat D., Madrigal J., & Saudemont A. (2017). Umbilical cord blood natural killer cells, their characteristics, and potential clinical applications. *Frontiers in Immunology*, *8*(10), 329. https://doi.org/10.3389/fimmu.2017.00329

Schafer J. L., & Graham J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schuyler R. P., Jackson C., Garcia-Perez J. E., Baxter R. M., Ogolla S., Rochford R., Ghosh D., Rudra P., & Hsieh E. W. (2019). Minimizing batch effects in mass cytometry data. *Frontiers in Immunology*, *10*, 2367. https://doi.org/10.3389/fimmu.2019.02367

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), 205–233. https://doi.org/10.32614/RJ-2016-021

Sengupta S., Wang J., Lee J., Müller P., Gulukota K., Banerjee A., & Ji Y. (2014). Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Pacific symposium on biocomputing co-chairs* (pp. 467–478). World Scientific.

Shah N., Li L., McCarty J., Kaur I., Yvon E., Shaim H., Muftuoglu M., Liu E., Orlowski R. Z., Cooper L., & Lee D. (2017). Phase I study of cord blood-derived natural killer cells combined with autologous stem cell transplantation in multiple myeloma. *British Journal of Haematology*, *177*(3), 457–466. https://doi.org/10.1111/bjh.14570

Soriano J., & Ma L. (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis*, *14*(1), 161–180. https://doi.org/10.1214/18-BA1106

Spiegelhalter D. J., Best N. G., Carlin B. P., & Van Der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stephens M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809. https://doi.org/10.1111/1467-9868.00265

Suck G., Linn Y. C., & Tonn T. (2016). Natural killer cells for therapy of leukemia. *Transfusion Medicine and Hemotherapy*, *43*(2), 89–95. https://doi.org/10.1159/000445325

Van Der Maaten L. (2014). Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research*, *15*(93), 3221–3245. http://jmlr.org/papers/v15/vandermaaten14a.html

Van Gassen S., Callebaut B., & Saeys Y. (2017). FlowSOM: Using self-organizing maps for visualization and interpretation cytometry data. http://dambi.ugent.be

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. https://doi.org/10.1002/cyto.a.22625

Wainwright M. J., & Jordan M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1–305. http://dx.doi.org/10.1561/2200000001

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. https://doi.org/10.1002/cyto.a.23030

Wu J., & Lanier L. L. (2003). Natural killer cells and cancer. *Advances in Cancer Research*, *90*(1), 127–156. https://doi.org/10.1016/S0065-230X(03)90004-2

Xie F., & Xu Y. (2019). Bayesian repulsive gaussian mixture model. *Journal of the American Statistical Association*, *115*(529), 187–203. https://doi.org/10.1080/01621459.2018.1537918

Xu Y., Lee J., Yuan Y., Mitra R., Liang S., Müller P., & Ji Y. (2013). Nonparametric Bayesian bi-clustering for next generation sequencing count data. *Bayesian Analysis (Online)*, *8*(4), 759. https://doi.org/10.1214/13-BA822

Xu Y., Müller P., & Telesca D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, *72*(3), 955–964. https://doi.org/10.1111/biom.12482

Xu Y., Müller P., Yuan Y., Gulukota K., & Ji Y. (2015). Mad Bayes for tumor heterogeneity—Feature allocation with exponential family sampling. *Journal of the American Statistical Association*, *110*(510), 503–514. https://doi.org/10.1080/01621459.2014.995794

Zhang C., Butepage J., Kjellstrom H., & Mandt S. (2018). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774