Bayesian Feature Allocation Models for Natural Killer Cell Repertoire Studies Using Mass Cytometry Data

Arthur Lui
Department of Applied Mathematics and Statistics, UC Santa Cruz

May 25, 2018

Abstract

Bayesian feature allocation models (FAMs) embedded with clustering capabilities are developed to analyze mass cytometry data, so as to characterize underlying cell repertoire structures. Cell repertoires in samples are heterogeneous. Each repertoire consists of a collection of cells possessing different phenotypes that can be characterized by differences in expression levels of cell surface markers. In particular, mass cytometry data collected to study the clinical efficacy of natural killer (NK) cells as immunotherapeutic agents against leukemia are considered. NK cells play a critical role in cancer immune surveillance and are the first line of defense against viruses and transformed tumor cells. The data of interest includes expression levels of 32 surface markers on each of thousands of cells from multiple samples. NK cell repertoires may affect both NK cell function and immune surveillance. We present a key conceptual shift from existing approaches by explicitly characterizing latent cell phenotypes through a FAM. The models simultaneously (1) characterize NK cell phenotypes based on expression / non-expression of surface markers, (2) estimate compositions of the samples based on the identified phenotypes, and (3) infer associations between subject-covariates and the composition of the identified phenotypes in the samples. The conventional Indian buffet process (IBP), one of the most popular FAMs, is first utilized to model cell phenotypes. Non-ignorable missing data that is present due to technical artifacts in mass cytometry instruments are accounted for by using an informed prior missing mechanism. The repulsive FAM (rep-FAM) is next proposed. In contrast to the IBP, the rep-FAM produces a parsimonious representation of the latent phenotypes by discouraging the creation of redundant phenotypes, and can thus improve inference on phenotypes. Further extensions to incorporate subject based covariates are discussed to provide inferences on phenotypes potentially associated with positive clinical outcomes.

Contents

1 Introduction

Clinical application of natural killer (NK) cells has recently emerged as a powerful treatment modality for advanced cancers refractory to conventional therapies [Rezvani and Rouce, 2015]. NK cells play a critical role in cancer immune surveillance and are the first line of defense against viruses and transformed tumor cells. They have the intrinsic ability to infiltrate cancer tissue and their presence in tumors is reported to be associated with better clinical outcomes [Suck et al., 2016]. Drs. Thall and Rezvani, collaborators at UT MD Anderson Cancer Center, have conducted clinical trials to study the potential clinical efficacy of umbilical cord blood (UCB) transplantation as a therapy for leukemia. UCB has become an established source of hematopoietic stem cells for transplantation. UCB NK cell therapy has the advantage of low risk of viral transmission from donor to recipient [Sarvaria et al., 2017]. In the trials, leukemia patients received UCB cell transplants. During follow-ups, samples were taken at multiple time points from each patient. Samples from healthy subjects and cord blood samples also were collected for comparison to leukemia patient samples. The samples were processed and expression levels of 32 NK-cell-associated cell surface protein markers were measured for individual cells in the samples using mass cytometry. Their primary research goal is to understand phenotypes and functions structured across heterogeneous NK cells. Better understanding of the characteristics of NK cells is crucial to estimating the true potential of NK cell therapies against cancer.

Advances in cytometry have led to more research and greater understanding of NK cells and how their diversity impacts immunity against the development of tumors and other viral diseases. Flow cytometry has been routinely used for single-cell analysis in cellular and clinical immunology. A primary interest in cytometry data analysis is to identify different cell types in a heterogeneous population and measure their abundance. Cell-types are characterized by their distinct expression patterns of multivariate protein markers, and expression patterns can be affected by cell composition and biological function. In recent years, a new cytometry technique, Cytometry at time-of-flight (CyTOF, also known as mass cytometry) has surfaced. It makes use of time-of-flight mass spectrometry, where sophisticated devices are used to accelerate, separate, and identify ions by mass. This new method enables the detection of a greater number of parameters (biological, phenotypic, or functional markers), up to 40 parameters for a cell, in less time and at a higher resolution [Cheung and Utz, 2011]. Efficient inferential frameworks are required to understand complex cytometry data. Manual "gating" is a traditional method in this realm in which homogeneous cell-clusters are sequentially identified and refined using a set of markers. However, it has several serious shortcomings including its inherent subjectivity as it requires manual analysis, and

being unscalable for multiparameter data. While manual gating is still popular in practice, many computational methods that automatically identify cell clusters have been proposed to analyze high-dimensional cytometry data. Many existing automated analysis methods use dimension reduction techniques and/or clustering methods such as density-based clustering methods, model-based clustering methods, and self-organizing maps. For example, FlowSOM in Van Gassen et al. [2015] uses a self-organizing map (SOM), an unsupervised neural network technique, for clustering and dimension reduction. A low-dimensional representation of the input space is obtained using unsupervised neural networks for easy visualization in a graph called a map. FlowSOM is fast and can be used as a starting point for a manual gating or as a visualization tool after a gating. Other common approaches are density-based clustering methods including DBSCAN [Ester et al., 1996] and ClusterX [Chen et al., 2016], and model-based clustering methods including flowClust [Lo et al., 2009] and BayesFlow [Johnsson et al., 2016]. Density-based clustering methods like DBSCAN and ClusterX form clusters in data when regions are densely occupied by observations. Observations from dense clusters that are near-together enough according to some predetermined proximity metric and threshold are grouped, while observations that are considered far from dense regions are classified as outliers. Density-based methods can produce clusters that take on flexible shapes. FlowClust clusters Box-Cox transformed data using a mixture of t-distributions and provides parameter estimates through expectation-maximization (EM) algorithm. BayesFlow developed a Bayesian hierarchical model to cluster cells through a Gaussian mixture model. Weber and Robinson [2016] performed a study to compare freely available clustering methods for high-dimensional cytometry data. They analyzed six publicly available cytometry datasets and compared identified cell subpopulations to cell population identities known from expert manual gating.

Existing methods, while promising, have fallen short at providing comprehensive inference on the underlying cell phenotypes in cell populations. Many of them do not properly handle large sample-to-sample variations and abnormalities due to technical artifacts in cytometry data, and do not provide uncertainty measurements for resulting inferences. Expression levels in different samples can significantly vary due to technical variation and most existing methods often analyze samples separately. Moreover, observations can be missing due to technical limitations when markers are not expressed. Existing methods often ignore missing data or use pre-imputed data. More importantly, existing methods do not directly model the underlying cell types. Rather, they identify cell clusters based on similarities in expression patterns and determine cell types based on identified clusters. When cells have the same expression pattern but different expression levels due to experimental noise, they can be grouped into different clusters although they are likely to be of the same cell type. We will

utilize Bayesian feature allocation models (FAMs) embedded with clustering capabilities as our main tools to directly model cell types. FAMs will provide a solid foundation to an effective way of revealing the underlying cell phenotypes. We also propose mechanisms for imputing missing values within the models. The proposed models will efficiently provide a full model-based and probabilistic inference with honest uncertainty quantification.

1.1 Review: Feature Allocation Models

One of the main inferential goals in analyzing the motivating dataset is to learn a latent structure of predominant NK cell phenotypes, where cell phenotype are defined based on distinct expression combinations of the markers. Specifically, phenotype k is represented by a J-dimensional binary vector, $\boldsymbol{z}_k = (z_{1k}, \ldots, z_{Jk})$, with J denoting the number of markers, where z_{jk} equals 1 if marker j is expressed in phenotype k, and 0 otherwise. Let $\boldsymbol{Z}$ denote a $J \times K$ binary matrix by letting columns represent K different phenotypes. Then 2^J possible distinct phenotypes can be constructed for J markers. One may consider a $J \times 2^J$ binary matrix that includes all possible phenotypes generated by the J markers. But this is computationally infeasible even when J is moderately large. Taking a Bayesian approach, we consider a prior probability model over binary matrices, which in this case represents a library of phenotypes, to learn predominant phenotypes from the observed data a posteriori. These models are called latent feature allocation models (FAMs). In FAMs, rows and columns correspond to objects and features, respectively (in the culinary metaphor of FAMs, customers and dishes, respectively; and in our applications, markers and phenotypes, respectively). Similar to our construction of phenotypes, $z_{jk} = 1$ corresponds to object j possessing feature k. Conversely, $z_{jk} = 0$ corresponds to object j not possessing feature k. One popular model for binary feature matrices of this type is the Indian buffet process (IBP), a Bayesian nonparametric distribution over $\boldsymbol{Z}$ with an unbounded number of latent features, proposed by Griffiths and Ghahramani [2011]. They construct the IBP by considering the finite feature allocation model and taking the limit with respect to the number of features. Concretely, for a given K,

$$\begin{aligned} v_k \mid \alpha &\sim \text{Beta}(\alpha/K, 1), \;\; k = 1, \ldots, K \\ z_{jk} \mid \pi_k &\sim \text{Bernoulli}(v_k), \;\; k = 1, \ldots, K \;\; \text{ and } j = 1, \ldots, J. \end{aligned} \tag{1}$$

The marginal limiting distribution of $\boldsymbol{Z}$ defines an IBP as $K \to \infty$ and after dropping all columns with all 0's. That is, $Z \sim \text{IBP}(\alpha)$, for a positive real α. Under the IBP, each row has an expected row sum of α. It can be shown that the number of non-zero columns in Z

has a Poisson distribution with mean $\alpha \sum_{j=1}^{J} j^{-1}$. A prior distribution can be placed on α to reflect uncertainty of the number of latent features. A gamma prior is popular for α due to its conjugacy.

Much theoretical work for the IBP has been generated in recent years. Teh et al. [2007] represented the IBP using the stick-breaking construction similar to the stick-breaking representation of the Dirichlet process (DP). Williamson et al. [2010] developed a dependent IBP (dIBP) to induce correlations between objects and to model multiple possibly dependent $\boldsymbol{Z}$'s. Broderick et al. [2015] first coined the term "feature allocation model". They developed theory for an exchangeable feature probability function for certain feature allocation models, just as the class of probability distributions over partitions of a dataset has been characterized through exchangeable partition probability functions. Under their framework, many other extensions of the IBP can be proposed. Broderick et al. [2013] developed the beta-negative binomial process for admixtures, where observations are represented multiple times across several latent features. This is an extension of the IBP, where infinite-dimensional priors are proposed for vectors of counts. They develop some computationally efficient algorithms that rely on Gibbs sampling for posterior sampling. They applied their methods in some simulation studies involving topic modeling and computer vision.

The IBP as a prior distribution in FAMs has been successfully applied in diverse areas. Lee et al. [2015] modeled tumor-sample heterogeneity through a finite IBP using DNA sequencing data. They used $\boldsymbol{Z}$ to describe latent haplotypes. A prior distribution is placed on the number of subclones. These parameters are learned jointly through MCMC. Building on this work, Xu et al. [2015] proposed a general class of feature allocation model for exponential family sampling distributions for modeling tumor heterogeneity. They note that a computational challenge in sampling from the joint posterior in such models involves implementing reversible jump MCMC [Green, 1995] for transdimensional moves. They avoid this by proposing an MAP-based small-variance asymptotic approximation for any exponential family likelihood with an IBP feature allocation prior to learn the feature allocation matrix. The learned feature allocation matrix is then used to fix the dimensions of a subsequent MCMC conditioned on the number of features estimated. This method is orders magnitudes faster than reversible-jump implementations. Sengupta et al. [2014] and Lee et al. [2016] proposed a categorical variant of the IBP for tumor-heterogeneity modeling so that the feature allocation matrix may contain integer values beyond binary values. They demonstrated how to obtain posterior estimates for the feature allocation matrix. They also developed a computational method to model a random number of features so as to avoid reversible-jump.

1.2 Proposed Projects

We propose the following projects to study NK cell phenotypes using mass cytometry data.

- Project 1: We first develop a model to study the composition of cell populations in multiple samples. The model directly characterizes latent cell phenotypes with an IBP prior and clusters individual cells based on identified cell types. The model also includes a mechanism for imputing missing observations to account for missing marker-expression data missing not at random. We demonstrate the developed model with simulation studies and an analysis of a real cord-blood dataset.
- Project 2: Building upon Project 1, we extend our model so as to compare the composition of the NK cell populations present in cord blood samples and samples taken from healthy subjects. We develop a repulsive FAM (rep-FAM) that discourages cell phenotypes that are similar, and replace the IBP prior distribution with a rep-FAM prior to obtain a parsimonious representation of the underlying cell population structures in different samples. We examine the properties of the proposed rep-FAM and compare them to those of the IBP. Results from a simulation study highlighting the key differences between an IBP and a rep-FAM are presented. An analysis based on a real dataset will be performed.
- Project 3: We desire to study the change in the abundance of cell phenotypes in patients' blood samples collected over time. To this end, we will propose a FAM with a regression to model associations between time and cell type abundances. Simulation studies and an analysis based on longitudinal data from patients will be conducted.

The remainder of the document is organized as follows. § 2 to § 4 describe the proposed projects. § 5 discusses a plan to progress the projects. Appendix A contains details for posterior computations in Project 1.

2 Project 1: Bayesian Feature Allocation Model for Heterogeneous Cell Populations

2.1 Introduction

In this project, we develop a Bayesian FAM, embedded with clustering capabilities to characterize underlying cell repertoire structures in mass cytometry data. Samples in cytometry

data consist of tens of thousands of cells and expression levels of a set of markers are recorded for individual cells. Large and small observed expression levels may imply expression and non-expression of the markers, respectively. A phenotype is defined by its unique subset of expressed markers and a repertoire in a sample can be described as a collection of cells possessing different phenotypes. One primary research goal is to characterize underlying cell phenotypes in samples based on observed expression levels and differentiate the samples based on the identified phenotypes. Many of the existing methods proposed to analyze cytometry data use clustering approaches to identify cell subpopulations based on their marker-expression levels. They often focus on producing point estimates of inferred subpopulations, convenient two-dimensional visualizations of multivariate expression data, and computational efficiency. While encouraging, many existing methods fail to provide direct inference on latent cell types. Under the clustering methods, cells having the same set of expressed markers, but different observed expression levels due to technical variability in an experiment, can be grouped into different clusters. That is, cells in different clusters can be characterized as one cell phenotype. Most of them are also algorithmic methods and do not produce uncertainty quantification for the learned clusters and their abundances in samples. To address this challenge in characterizing underlying cell repertoire structures, we use a IBP, one of the most popular FAMs and directly model expression patterns of latent cell types. Individual cells in a sample possess cell types and the distribution of cell types differ in samples. In other words, individual cells will be connected to identified phenotypes via clustering, with cluster probabilities varying between samples. We further model observed marker expression levels with flexible mixture models to effectively accommodate variability in expression levels within cells having a cell type. We will also address the challenge of analyzing cytometry data containing data missing not at random. When a marker in a cell is not expressed, cytometry devices may not register a signal and fail to record expression levels, yielding missing values. Franks et al. [2016] provide a brief overview of typical approaches to imputing data missing not at random, including a computationally efficient method and intuitive model representation (Tukey's representation) based only on the observed data. In our application, the parameters are entangled with the missingness of the data. We model missing data through a selection factorization representation [Rubin, 1974] and incorporate missing data into the inference on latent cell repertoire structures.

In the remainder of the section, we will present the proposed statistical model in § 2.2, simulation studies in § 2.3, an analysis of real mass cytometry data in § 2.4, and some concluding remarks in § 2.5.

2.2 Probability Model

2.2.1 Sampling Model

I samples are taken from subjects, $i = 1, 2, ..., I$. Sample i consists of N_i cells, $n = 1, \ldots, N_i$ and for each cell, expression levels of J markers are measured. Let $\tilde{y}_{inj} \in \mathbb{R}^+$ represent the raw measurement of an expression level of marker j of cell n in sample i. Let c_{ij} denote the "cutoff" for marker j in sample i. A marker of a cell is likely to be expressed if its observed expression level is greater than the cutoff. A value of $\tilde{y}_{inj}$ below the cutoff may imply that marker j is not expressed in cell n of sample i. These cutoff values are computed by cytometry devices and may vary by sample due to noise in the environment. We consider the logarithm transformation after scaling $\tilde{y}_{inj}$ by c_{ij},

$$y_{inj} = \log\left(\frac{\tilde{y}_{inj}}{c_{ij}}\right) \in \mathbb{R}.$$

Due to the transformation, a value above (below) 0 is likely to represent (non-) expression. For some (i, n, j), $\tilde{y}_{inj}$ is missing due to experimental artifacts and we introduce a binary indicator,

$$m_{inj} = \begin{cases} 0, & \text{if } \tilde{y}_{inj} \text{ is observed,} \\ 1, & \text{if } \tilde{y}_{inj} \text{ is missing.} \end{cases}$$

We assume that a sample has heterogeneous cell populations having K different phenotypes. The phenotypes are not directly observable and we introduce latent phenotype indicators $\lambda_{in} \in \{1, \ldots, K\}$, for cell n in sample i, $i = 1, \ldots, I$ and $n = 1, \ldots, N_i$. The event $\lambda_{in} = k$, for $k = 1, \ldots, K$, represents that cell n in sample i possesses phenotype k. The cell phenotypes are defined by columns of a $J \times K$ binary matrix $\boldsymbol{Z}$. The element $z_{j,k} \in \{0, 1\}$ indicates whether marker j is expressed in cell phenotype k or not. The event $z_{jk} = 0$ represents that marker j is not expressed for phenotype k, and $z_{jk} = 1$ for expression. We let $\boldsymbol{Z}$ and λ_{in} be random quantities. Modeling details will be discussed later. Given $z_{j,\lambda_{in}} \in \{0, 1\}$, we assume a mixture of normals for y_{inj},

$$y_{inj} \mid \boldsymbol{\eta}_{ij}, \boldsymbol{\mu}^\star, \boldsymbol{\sigma}_i^{2\star} \overset{ind}{\sim} \begin{cases} \sum_{\ell=1}^{L^0} \eta_{ij\ell}^0 \text{ Normal}(\mu_{0\ell}^\star, \sigma_{0i\ell}^{2\star}), & \text{if } z_{j,\lambda_{in}} = 0, \\ \sum_{\ell=1}^{L^1} \eta_{ij\ell}^1 \text{ Normal}(\mu_{1\ell}^\star, \sigma_{1i\ell}^{2\star}), & \text{if } z_{j,\lambda_{in}} = 1, \end{cases} \tag{2}$$

where the number of mixture components L^0 and L^1 are fixed. The vectors $\boldsymbol{\eta}_{ij}^0$ and $\boldsymbol{\eta}_{ij}^1$ are mixture weights with $\sum_{\ell=1}^{L^0} \eta_{ij\ell}^0 = \sum_{\ell=1}^{L^1} \eta_{ij\ell}^1 = 1$, where $0 < \eta_{ij\ell}^1 < 1$ and $0 < \eta_{ij\ell}^0 < 1$. In

(2), $\boldsymbol{\mu}_0^\star$ and $\boldsymbol{\mu}_1^\star$ are common for all samples and markers but $\boldsymbol{\sigma}_0^{2\star}$ and $\boldsymbol{\sigma}_1^{2\star}$ are additionally indexed by sample i to account for sample specific variability. Sample and marker specific mixture weight vectors $\boldsymbol{\eta}_{ij}^0$ and $\boldsymbol{\eta}_{ij}^1$ allow markers in samples to have different distributions. The mixture model can thus flexibly capture various features in data. For computational convenience, we introduce mixture component indicators γ_{inj} for y_{inj}. Given $\lambda_{in} = k$, we define γ_{inj} for $i = 1, \ldots, I$, $n = 1, \ldots, N_i$ and $j = 1, \ldots, J$, by

$$P(\gamma_{inj} = \ell \mid \lambda_{in} = k) = \eta_{ij\ell}^{z_{jk}}, \text{ where } \ell \in \{1, \ldots, L^{z_{jk}}\}. \tag{3}$$

Given $\lambda_{in} = k$ and $\gamma_{inj} = \ell$, we assume a normal distribution for y_{inj}; for $i = 1, \ldots, I$, $n = 1, \ldots, N_i$ and $j = 1, \ldots, J$,

$$y_{inj} \mid \mu_{inj}, \sigma_{inj}^2 \overset{ind}{\sim} \text{Normal}(\mu_{inj}, \sigma_{inj}^2), \tag{4}$$

where $\mu_{inj} = \mu_{z_{j,k},\ell}^\star$ and $\sigma_{inj}^2 = {\sigma^{2\star}}_{z_{j,k}i\ell}$. After marginalizing over γ_{inj}, the model in (4) and (3) is equivalent to the model in (2).

We next build a model for the missingness mechanism. To build the mechanism, we incorporate information provided by a subject scientist that a marker expression level is recorded as "missing" when a maker in a cell has a very weak signal, strongly implying that the marker is not expressed. We take an empirical approach by assuming that the distribution of the values with $m_{inj} = 1$ is similar to the distribution of observed y below 0. We then let the probability of y being missing depend on its unobserved value of y. Given y_{inj}, we consider a selection function for m_{inj} for $i = 1, \ldots, I$, $n = 1, \ldots, N_i$ and $j = 1, \ldots, J$,

$$\begin{aligned} &m_{inj} \mid p_{inj} \overset{ind}{\sim} \text{Bernoulli}(p_{inj}) \\ &\text{logit}(p_{inj}) = \begin{cases} \beta_{0i} - \beta_{1i}(y_{inj} - c_0)^2, & \text{if } y_{inj} < c_0, \\ \beta_{0i} - \beta_{1i} c_1 \left(y_{inj} - c_0\right)^{1/2}, & \text{otherwise,} \end{cases} \end{aligned} \tag{5}$$

where c_0 and c_1 are real constants, $\beta_{0i} \in \mathbb{R}$, and $\beta_{1i} > 0$. Figure 1 shows an example missing mechanism. We design the missing mechanism to have a peak in the middle to discourage missing values from being imputed as either extremely large or extremely small values. This helps to control the size of the scale parameter in the mixture of Normals used in the data density. Note that the assumptions for the distribution of the unobserved data are untestable. However, we can incorporate input from biologists through informed prior specifications. Prior uncertainty can then be directly propagated to the posterior distribution for the missing mechanism.

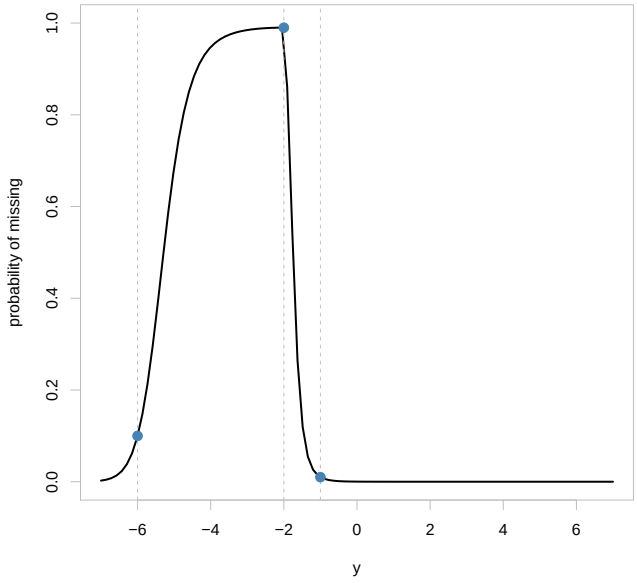


Figure 1: Example missing mechanism. The blue points serve as guides in determining a missing mechanism. Values for β and c can be solved for through a system of equations.

2.2.2 Priors

Latent cell phenotypes Recall that we characterize cell phenotypes with a $J \times K$ binary matrix $\boldsymbol{Z} = \{z_{jk}\}$. Following Williamson et al. [2010], we assume

$$\begin{aligned} v_k \mid \alpha &\overset{iid}{\sim} \text{Beta}(\alpha/K, 1),\ k = 1, \ldots, K, \\ \boldsymbol{h}_k &\overset{iid}{\sim} \text{Normal}_J(\mathbf{0}, \Gamma), \\ z_{jk} \mid h_{jk}, v_k &= \mathbb{I}\left\{\Phi(h_{jk} \mid 0, \Gamma_{jj}) < v_k\right\}, \end{aligned}$$

where $\Phi(h \mid m, s)$ is the cumulative distribution function of the normal distribution with mean m and variance s, and $\mathbb{I}(\cdot)$ is an indicator function having a value of 1 if $\Phi(h_{jk} \mid 0, \Gamma_{jj}) < v_k$, and 0, otherwise. As $K \to \infty$, the limiting distribution of Z is the IBP [Griffiths and Ghahramani, 2011]. Interactions between J markers in phenotypes can be modeled through Γ. Due to the multivariate probit construction for $\boldsymbol{Z}$, Γ is not identifiable and it is common to restrict Γ to be a correlation matrix. Prior distributions for correlation matrices have been proposed to handle such cases. Jointly uniform and marginally uniform prior distributions have been identified by Barnard et al. [2000] for correlation matrices. Box and Tiao [2011] have noted that the Jeffreys' prior for correlation matrices is $p(\Gamma) \propto |\Gamma|^{-(J+1)/2}$. Zhang et al. [2006] presents more generalized and flexible priors which can be reduced to the two previous

priors as special cases. We let $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ with mean a_α/b_α.

The K cell phenotypes are common in all samples but the relative weights vary across samples. Let w_{ik} denote an abundance level of phenotype k in sample i. We assume independent Dirichlet priors for $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iK})$ given K, $\boldsymbol{w}_i \mid K \overset{iid}{\sim} \text{Dirichlet}_K(d/K)$. For the latent cell phenotype indicators, we let $p(\lambda_{in} = k \mid \boldsymbol{w}_i) = w_{ik}$.

Parameters in the Mixture for y In (2), normal mixture models are assumed for y_{inj}. The mean expression level of marker j of sample i in cell n is determined by its phenotype λ_{in}. In particular, if the marker is not expressed in the cell type (i.e. $z_{j\lambda_{in}} = 0$), its mean expression level is below the cutoff, that is, a negative value. If the marker is expressed (i.e. $z_{j\lambda_{in}} = 1$), the expression level of marker j takes a positive value. Recall that $\mu^\star_{0\ell}$, $\ell = 1, \ldots, L^0$ are mixture locations for $z_{j\lambda_{in}} = 0$ and $\mu^\star_{1\ell}$, $\ell = 1, \ldots, L^1$ for $z_{j\lambda_{in}} = 1$. We assume

$$\mu^\star_{0\ell} \mid \psi_0, \tau_0^2 \overset{iid}{\sim} \text{N}_-(\psi_0, \tau_0^2), \quad \ell \in \left\{1, ..., L^0\right\},$$
$$\mu^\star_{1\ell} \mid \psi_1, \tau_1^2 \overset{iid}{\sim} \text{N}_+(\psi_1, \tau_1^2), \quad \ell \in \left\{1, ..., L^1\right\},$$

where $\text{N}_-(m, s^2)$ and $\text{N}_+(m, s^2)$ denote the normal distribution with mean m and variance s^2, truncated to take only negative values and positive values, respectively. The variances $\boldsymbol{\sigma}_0^{2\star}$ and $\boldsymbol{\sigma}_1^{2\star}$ in the mixture components differ by the value of $z_{j\lambda_{in}}$ and also vary across samples. We let, for $i = 1, \ldots, I$,

$$\sigma^2_{0i\ell} \mid s_i \overset{ind}{\sim} \text{Inverse-Gamma}(a_\sigma, s_i), \quad \ell \in \left\{1, ..., L^0\right\},$$
$$\sigma^2_{1i\ell} \mid s_i \overset{ind}{\sim} \text{Inverse-Gamma}(a_\sigma, s_i), \quad \ell \in \left\{1, ..., L^1\right\}.$$

We also assume $s_i \overset{iid}{\sim} \text{Gamma}(a_s, b_s)$, $i \in \{1, ..., I\}$, with mean a_s/b_s. Lastly, we consider a model for the mixture weights $\boldsymbol{\eta}^0_{ij}$ and $\boldsymbol{\eta}^1_{ij}$. To flexibly model the distribution of y_{inj}, we assume for a marker j in a sample i, that y_{inj} has two sets of weights – one for each value of $z \in \{0, 1\}$. That is, $\boldsymbol{\eta}^0_{ij}$ and $\boldsymbol{\eta}^1_{ij}$, for each (i, j). So for $i = 1, \ldots, I$, $n = 1, \ldots, N_i$ and $j = 1, \ldots, J$,

$$\boldsymbol{\eta}^0_{ij} \overset{iid}{\sim} \text{Dirichlet}_{L^0}(a_{\eta^0}/L^0),$$
$$\boldsymbol{\eta}^1_{ij} \overset{iid}{\sim} \text{Dirichlet}_{L^1}(a_{\eta^1}/L^1).$$

Parameters for Missingness Mechanism A prior distribution over the missing mechanism can be specified through placing priors on the parameters β_{0i} and β_{1i}. We assume that $\beta_{0i} \overset{iid}{\sim} \mathrm{N}(m_{\beta_0}, s^2_{\beta_0})$ and $\beta_{1i} \overset{iid}{\sim} \mathrm{N}_+(m_{\beta_1}, s^2_{\beta_1})$, $i = 1, \ldots, I$. We use data to specify the values of the fixed hyperparameters, m_{β_0} and m_{β_1}. We let $s^2_{\beta_0}$ and $s^2_{\beta_1}$ be small to induce an informative prior for β_{0i} and β_{1i}. One way of determining priors for the parameters in the missing mechanism is described in detail in the derivation of the full conditionals for β in Appendix A.

2.2.3 Posterior Computation

Let $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\mu}_0^\star, \boldsymbol{\mu}_1^\star, \boldsymbol{\sigma}^2_{0i}, \boldsymbol{\sigma}^2_{1i}, \boldsymbol{\eta}^0, \boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \alpha\}$ represent all random parameters. Let $\boldsymbol{y}$ and $\boldsymbol{m}$ denote y_{inj} and m_{inj} for all (i, n, j), respectively. The joint posterior distribution is

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{m}) &\propto p(\boldsymbol{\theta}) \prod_{i,n,j} p(m_{inj} \mid y_{inj}, \boldsymbol{\theta}) p(y_{inj} \mid \boldsymbol{\theta}) \\
&= p(\boldsymbol{\theta}) \prod_{i,n,j} \left[p_{inj}^{m_{inj}} (1 - p_{inj})^{1 - m_{inj}} \times \frac{1}{\sqrt{2\pi\sigma^2_{inj}}} \exp\left\{ -\frac{(y_{inj} - \mu_{inj})^2}{2\sigma^2_{inj}} \right\} \right].
\end{aligned}
$$

Posterior simulation can be done via Gibbs sampling by repeatedly and sequentially updating each parameter until convergence. Parameter updates are made by sampling from its full conditional distribution. Where this cannot be done conveniently, a Metropolis step can be used. Details for the posterior simulation are omitted due to the space limit.

Summarizing the joint posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{m})$ is challenging, especially for $\boldsymbol{Z}$, which may be susceptible to label switching problems common in mixture models. Note also that the posterior distributions of $\boldsymbol{Z}$ is dependent on those of $\boldsymbol{w}$ and $\boldsymbol{\lambda}$. To summarize the posterior distribution of $(\boldsymbol{Z}, \boldsymbol{W}, \lambda)$ with point estimates, we use a method based on sequentially-allocated latent structure optimization (SALSO) [Dahl and Müller, 2017] for summarizing random samples over partitions. To summarize random feature allocation matrices, SALSO first constructs $A(\boldsymbol{Z}) = \{A_{j,j'}\}$, the $J \times J$ pairwise allocation matrix corresponding to a binary matrix $\boldsymbol{Z}$, where

$$
A_{j,j'} = \sum_{k=1}^{K} \mathbb{I}(Z_{j,k} = 1) \mathbb{I}(Z_{j',k} = 1), \quad \text{for } 1 \leq j, j' \leq J,
$$

is the number of features that markers j and j' share. It then uses constrained optimization

to find a point estimate $\hat{\boldsymbol{Z}}$ that minimizes the sum of the element-wise squared distances,

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

where $\hat{A}$ is the pairwise allocation matrix averaged over all posterior samples of Z. We extend SALSO to find point estimates for each sample i, $\hat{\boldsymbol{Z}}_i$ by incorporating $\boldsymbol{w}_i$. Specifically, we consider the sum of element-wise squared distances weighted by $\boldsymbol{w}_i$. We use posterior Monte Carlo samples to obtain posterior point estimates $(\hat{\boldsymbol{Z}}_i, \hat{\boldsymbol{W}}_i)$ and $\hat{\lambda}_{in}$, for $i = 1, \ldots, I$ and $n = 1, \ldots, N_i$ as follows. Suppose we obtain B posterior samples simulated from the posterior distribution of $\boldsymbol{\theta}$. For each posterior sample of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$, we compute a $J \times J$ adjacency matrix, $\boldsymbol{A}_i^{(b)} = \{A_{i,j,j'}^{(b)}\}$, where

$$A_{i,j,j'}^{(b)} = \sum_{k=1}^{K} w_{ik}^{(b)} \mathbb{I}\left(z_{jk}^{(b)} = 1\right) \mathbb{I}\left(z_{j'k}^{(b)} = 1\right), b \in \{1, ..., B\}.$$

We then compute the mean adjacency matrix $\bar{A}_i = \sum_{b=1}^{B} A_i^{(b)}/B$. We report a posterior point estimate of $\boldsymbol{Z}_i$ by choosing

$$\hat{\boldsymbol{Z}}_i = \text{argmin}_{\boldsymbol{Z}} \sum_{j,j'} (A_{i,j,j'}^{(b)} - \bar{A}_{i,j,j'})^2). \tag{6}$$

If $\hat{\boldsymbol{Z}}_i = Z^{(b)}$, then we report the posterior point estimates $\hat{\boldsymbol{w}}_i = \boldsymbol{w}_i^{(b)}$ and $\hat{\lambda}_{in} = \lambda_{in}^{(b)}$. Equation (6) places greater weight on phenotypes that are more prevalent in samples, and down-weights phenotypes having small w_{ik} for $\hat{\boldsymbol{Z}}_i$.

2.3 Simulation Study

We conducted a simulation study to evaluate the performance of the proposed model and compare our model to existing methods. To simulate data, we assume the number of markers (J) to be 32, the number of samples (I) to be 3, and the number of cells in the samples (N) to be $(300, 200, 100)$. We let the true number of latent cell types $K^{\text{TR}} = 10$. We specified $\boldsymbol{Z}^{\text{TR}}$ and $\boldsymbol{w}_i^{\text{TR}}$, for $i = 1, 2, 3$ as follows: We first simulated $\boldsymbol{Z}^{\text{TR}}$ by setting $\boldsymbol{Z}_{jk}^{\text{TR}} = 1$ with probability 0.6 for $j = 1, ..., J$ and $k = 1, ..., K$. If any column or row in $\boldsymbol{Z}^{\text{TR}}$ is a column or row of only 0's, the entire matrix is re-sampled. We then simulated $\boldsymbol{w}_i^{\text{TR}}$ from a Dirichlet distribution with parameters being some permutation of $(1, \ldots, K)$. This encourages the simulated values in $\boldsymbol{w}_i^{\text{TR}}$ to contain large as well as small values. Figure 2 shows the transpose

Figure 2: The transpose of $\boldsymbol{Z}^{\mathrm{TR}}$ with markers in columns and latent phenotypes in rows. Black and white represents $z_{jk}^{\mathrm{TR}} = 1$ and 0, respectively. The phenotypes and $\boldsymbol{w}_i^{\mathrm{TR}}$ are shown on the left and right sides of each panel, respectively. The samples share the same $\boldsymbol{Z}^{\mathrm{TR}}$ and the phenotypes are arranged in order of w_{ik}^{TR} within each sample.

of $\boldsymbol{Z}^{\mathrm{TR}}$ and $\boldsymbol{w}_i^{\mathrm{TR}}$ for the samples. In each panel, the cell types ($\boldsymbol{Z}^{\mathrm{TR}}$) are sorted by w_{ik}^{TR}. Three and four mixture components were used for non-expressed and expressed marker expression levels, respectively. That is, $L^{0,\mathrm{TR}} = 3$ and $L^{1,\mathrm{TR}} = 4$. We set $\mu_0^{\star,\mathrm{TR}} = (-5, -2, -1)$ and $\mu_1^{\star,\mathrm{TR}} = (1, 2, 4, 5)$. Values of $\sigma_{0i\ell}^{\star 2,\mathrm{TR}}$ and $\sigma_{1i\ell}^{\star 2,\mathrm{TR}}$ drawn from a Uniform(0, 0.3) distribution for all i and ℓ. We simulated $\boldsymbol{\eta}_{0ij}^{\mathrm{TR}}$ from a Dirichlet distribution with parameters being some permutation of $(1, ..., L^{0,\mathrm{TR}})$. Similarly, $\boldsymbol{\eta}_{1ij}^{\mathrm{TR}}$ was simulated from a Dirichlet distribution with parameters being some permutation of $(1, ..., L^{1,\mathrm{TR}})$. We then simulated latent phenotype indicators $\lambda_{in}^{\mathrm{TR}}$ using $\boldsymbol{w}_i^{\mathrm{TR}}$, and conditionally on $z_{j,\lambda_{in}^{\mathrm{TR}}}^{\mathrm{TR}}$ generated y_{inj} from the mixture model,

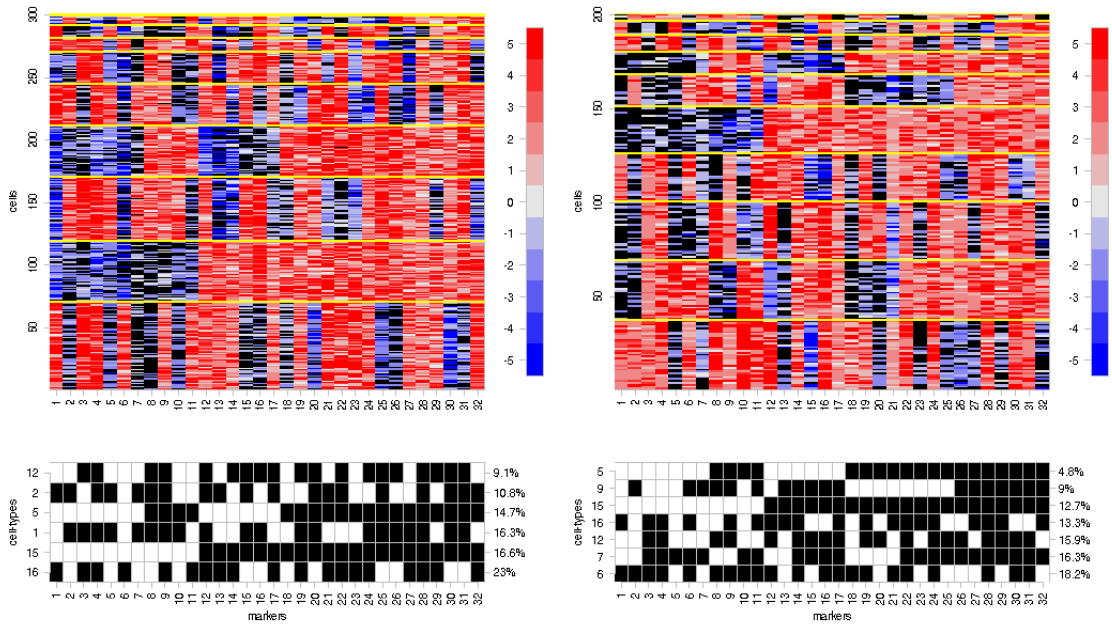
$\sum_{\ell=1}^{L^{0,\mathrm{TR}}} \eta_{0ij}^{\mathrm{TR}} \cdot \mathrm{N}(\mu_{0\ell}^{\star,\mathrm{TR}}, \sigma_{0i\ell}^{\star 2,\mathrm{TR}})$ or $\sum_{\ell=1}^{L^{1,\mathrm{TR}}} \eta_{1ij}^{\mathrm{TR}} \cdot \mathrm{N}(\mu_{1\ell}^{\star,\mathrm{TR}}, \sigma_{1i\ell}^{\star 2,\mathrm{TR}})$. Finally, let some of the y_{inj} be missing as follows. Simulate a proportion (p_{ij}) of values to be missing for marker j in sample i, from a Uniform$\big(0, \sum_k w_{ik}^{\mathrm{TR}}(1 - z_{jk}^{\mathrm{TR}})\big)$ distribution. Sample $p_{ij} \times N_i$ cells without replacement with probability proportional to

$$\begin{cases} \text{logistic}\left(4.6 - 0.42(y_{inj} + 3)^2\right), & \text{if } y_{inj} < -3 \\ \text{logistic}\left(4.6 - 0.42\sqrt{y_{inj} + 3}\right), & \text{otherwise,} \end{cases}$$

where $\text{logistic}(x) = (1 + \exp\{-x\})^{-1}$. Under the true missingness mechanism, y taking a negative value has a larger chance to be missing, while y with a positive value has only slight chance of being missing. Note that the true mechanism is different from that assumed in the proposed model. Heatmaps of the simulated $\boldsymbol{y}$ are shown in the heatmaps on the top of each panel in Figure 3. The y_{inj}'s are sorted within a sample according to their posterior phenotype estimates (will be discussed later). Red, blue and black colors represent high expression levels, low expression levels, and missing values, respectively.

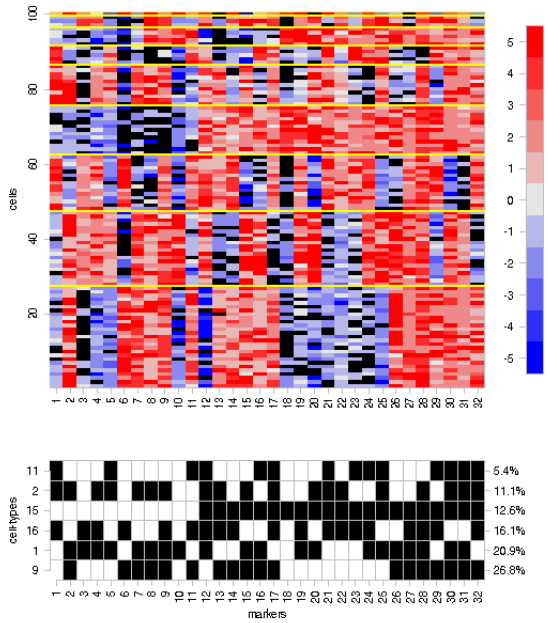
To fit the proposed model, we fixed $(K = 12, L^0 = 5, L^1 = 5)$. We used the mean of y_{inj} having negative values to specify c_0 and let $c_0 = -2.5$ and $c_1 = 10.43$. We fixed $\Gamma = \boldsymbol{I}_J$, $J \times J$ identity matrix for simplification. We specified the remaining fixed hyperparameters as follows: $a_\alpha = 3$ $b_\alpha = 2$, $\psi_0 = -2$, $\tau_0^2 = 0.09$, $\psi_1 = 2$, $\tau_1^2 = 0.09$, $a_\sigma = 6$, $a_s = 0.25$, $b_s = 0.5$, $a_{\eta^0} = 0.2$, $a_{\eta^1} = 0.2$, $m_{\beta_0} = 1.37$, $s_{\beta_0}^2 = 1$, $m_{\beta_1} = 0.57$, and $s_{\beta_1}^2 = 0.01$. To run the MCMC simulation, we initialized the parameters as follows; The missing values of y_{inj} are initialized at c_0. λ_{in} is sampled uniformly from $\{1, ..., K\}$. γ_{inj} is sampled uniformly from $\{1, ..., \min(L^0, L^1)\}$. v_k is initialized at $1/K$. h_{jk} is randomly sampled from the standard Normal distribution. $\boldsymbol{Z}$ is computed according to the initialized h and v. All other parameters are initialized at their prior means. Initial values of λ_{in} are used to initialize $\mu_{0\ell}^\star$, $\mu_{1\ell}^\star$, $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. We then implemented posterior inference using MCMC simulation over 3,000 iterations, discarding the first 1,000 iterations as burn-in. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots. We found no evidence of practical convergence problems.

Figure 3 summarizes the posterior inference for the simulated data. The posterior point estimates $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are obtained using the method described in § 2.2.3. The bottom of each panel illustrates $\hat{\boldsymbol{Z}}_i$ with $\hat{\boldsymbol{w}}_i$ in percentages on the right side. Among $K = 12$ phenotypes, phenotypes having the largest $\hat{w}_{ik}$'s that make up more than 90% of cells are included in the plots of $\hat{\boldsymbol{Z}}_i$. Compared to their truth in Figure 2, $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are very close to $\boldsymbol{Z}^{\mathrm{TR}}$ and $\boldsymbol{w}_i^{\mathrm{TR}}$ for all samples. Note that the phenotype labels do not match in the figures due to the



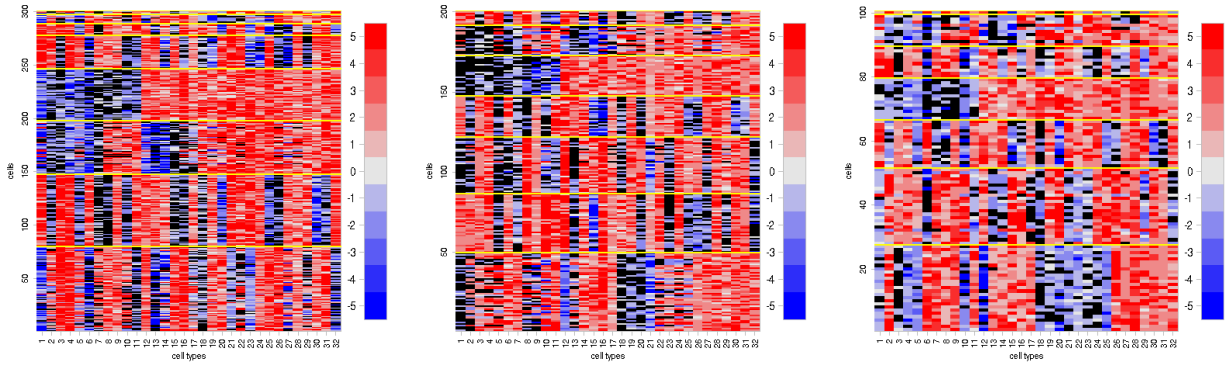
(a) Sample 1

(b) Sample 2



(c) Sample 3

Figure 3: [Simulation] Heatmaps of y for simulated data. Cells and markers are in rows and columns, respectively. Each column contains the expression levels of a marker for all cells in the sample. High expression levels are red, low expression levels are blue, missing values are black. Cells are rearranged by the corresponding posterior estimate of their phenotype indicator, $\hat{\lambda}_{in}$. Yellow horizontal lines separate cells by different phenotypes. At the bottom of each panel, the transpose of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are provided for each sample. We include phenotypes having largest $\hat{w}_{ik}$ to explain at least 90% of the cells in a sample.

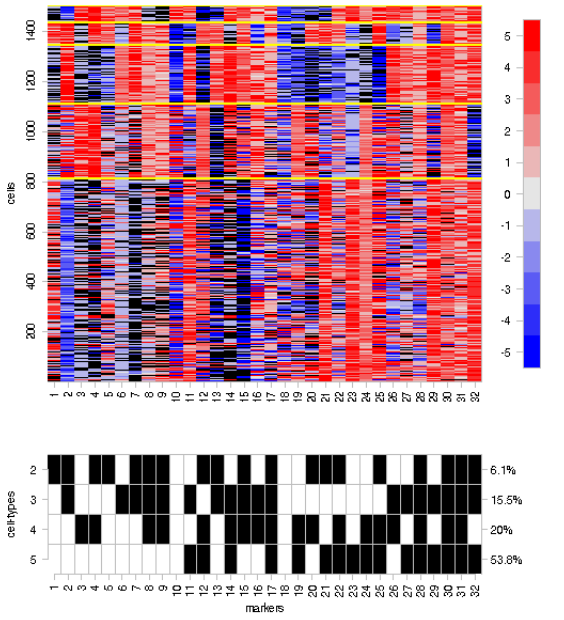


(a) Sample 1 (b) Sample 2 (c) Sample 3

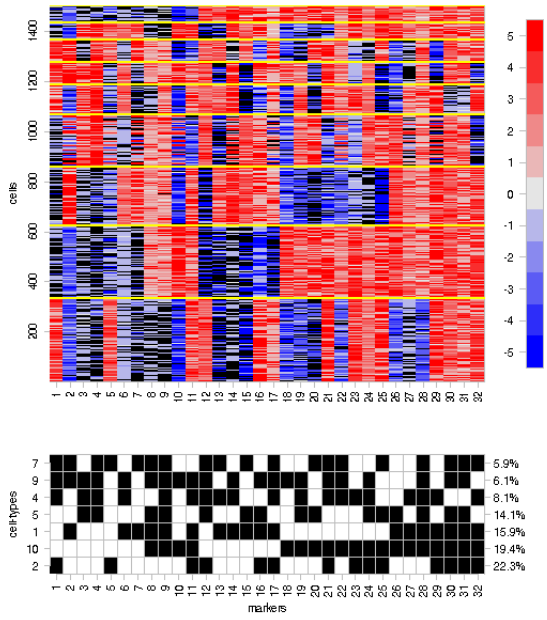
Figure 4: [FlowSOM for Simulated Data] Heatmaps of y_{inj} sorted by the cluster labels estimated by FlowSOM.

fact that the model for $\boldsymbol{Z}$ is invariant under relabelling of the phenotypes. For example, phenotype 3 in $\boldsymbol{Z}^{\text{TR}}$ has $w_{ik}^{\text{TR}} = 23.9\%$ for sample 1. That phenotype is shown as phenotype 16 in the very bottom of $\hat{\boldsymbol{Z}}_i$ with $\hat{w}_{ik} = 20.1\%$. The heatmaps of observed y arranged according to cell phenotype estimates $\hat{\lambda}_{in}$ show that the expression patterns in y are well explained by $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$. The top of each panel has a heatmap of y rearranged by their $\hat{\lambda}_{in}$, with the colors red, blue, and black for large, small, and missing values, respectively. The horizontal yellow lines separate cells based on $\hat{\lambda}_{in}$. It shows that the estimated phenotypes capture the expression patterns of y well. Phenotypes with reasonably large w_{ik}^{TR} in at least sample are not well estimated.

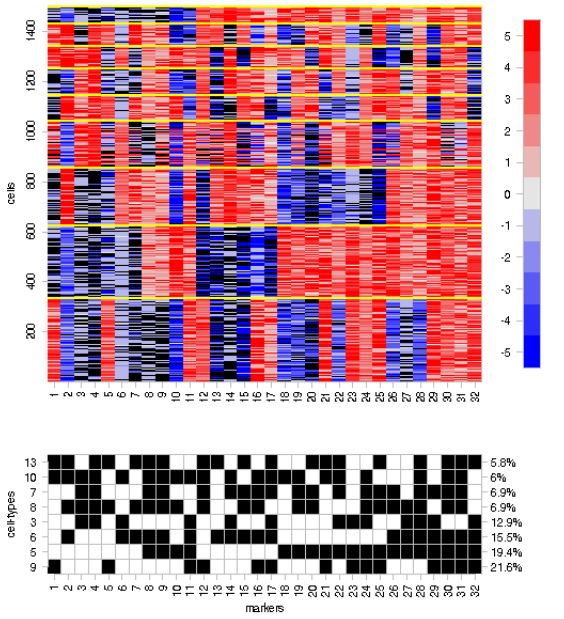
We compared our model to FlowSOM as it tends to be fast and performant in a variety of situations [Weber and Robinson, 2016]. Since FlowSOM does not impute missing values, the missing values were first set to be the minimum value of the observed values in the data. FlowSOM does not account for variability in samples and the three samples were combined for analysis. Eight cell clusters are produced and clustering of cells estimated by FlowSOM is summarized in Figure 4. Recall that FlowSOM uses similarities in expression levels for clustering instead of combinations of expression/no expression of the markers. For example, the largest cell cluster in sample 1 (very bottom of the heatmap) has large variability in y. In particular, cells in the cluster have missing values (black) and large expression values (red) for some markers. The cells having the same true cell phenotypes tend to be in a cluster but their cell clusters do not exactly correspond to any of the true phenotypes in $\boldsymbol{Z}^{\text{TR}}$. In some of our exploratory simulations, FlowSOM and the proposed FAM produce the same clusters when expression levels of the phenotypes in the simulated data are equidistant. This reveals a fundamental difference in our proposed FAM. We will conduct more simulation studies to



(a) $K = 3$



(b) $K = 10$



(c) $K = 20$

Figure 5: [Sensitivity analysis for K] Posterior estimates of $\hat{\boldsymbol{Z}}_i$ and heatmaps of y_{inj} rearranged by latent cell phenotype estimates for sample 1 ($i = 1$). Data is simulated with $K^{\text{TR}} = 10$ and the model was fit with three different values of K, $K = 3$, 10 and 20 in (a)-(c), respectively. For $\hat{\boldsymbol{Z}}$ we include phenotypes having the largest $\hat{w}_{ik}$ to explain at least 90% of the cells in a sample.

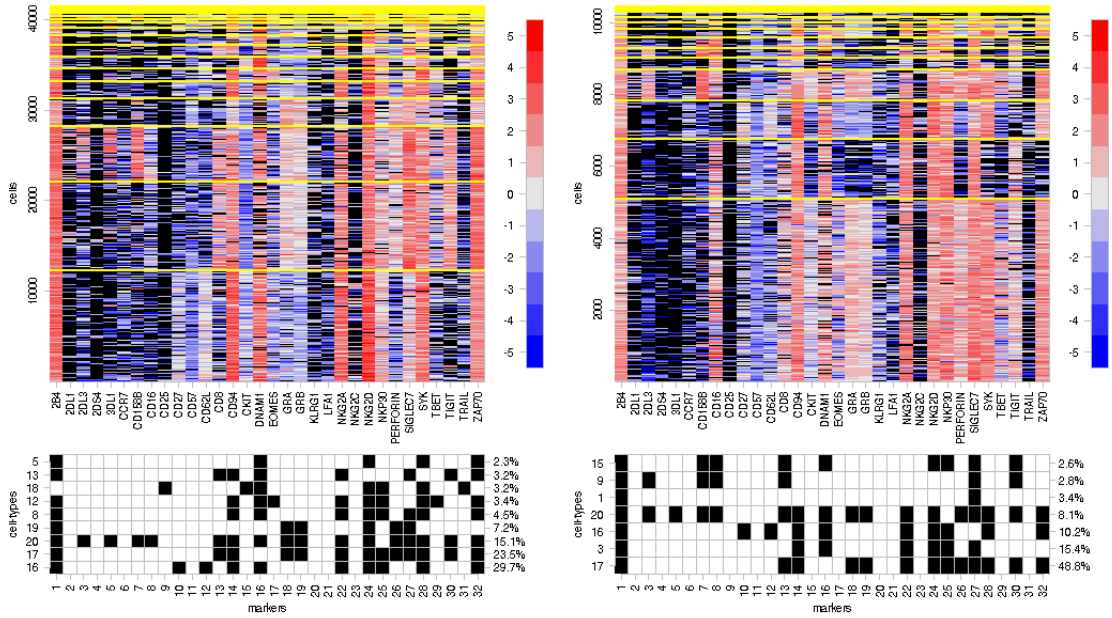
further investigate the performance of the proposed FAM and comparison to the existing models.

Sensitivity to Specified K The value of K determines the dimensions of the latent feature matrix $\boldsymbol{Z}$ and cell phenotype abundance vectors $\boldsymbol{w}_i$. To assess the FAMs sensitivity to the specification of K, we performed a simulation study keeping most of the simulation set-up described previously, except with the number of cells being $N = (1500, 600, 300)$. We fit the proposed model with different specifications of K, with $K = 5, 10,$ and 20. Recall that $K^{\text{TR}} = 10$ is used to simulate data. Figure 5 summarizes the posterior estimates of $\boldsymbol{Z}_i$ and displays y_{inj} rearranged by their estimated phenotypes for sample 1 ($i = 1$) with the three values of K. It is clear from panel (a) that when $K = 3$ (less than K^{TR}), the model compromises the true cell population structure and cells of different true phenotypes are forced to have those phenotype estimates. In particular, cells in estimated phenotype 5 show large variability in y_{inj}. On the other hand, from panels (b) where $K = K^{\text{TR}} = 10$ and (c) where $K = 20 \geq K^{\text{TR}}$, the predominant phenotypes are well recovered. When K is larger than K^{TR}, not only are the predominant cell phenotypes recoverable in $\boldsymbol{Z}$, but the model performance does not suffer critically either. We therefore recommend using a reasonably large K in a real data analysis.

2.4 Cord Blood Data

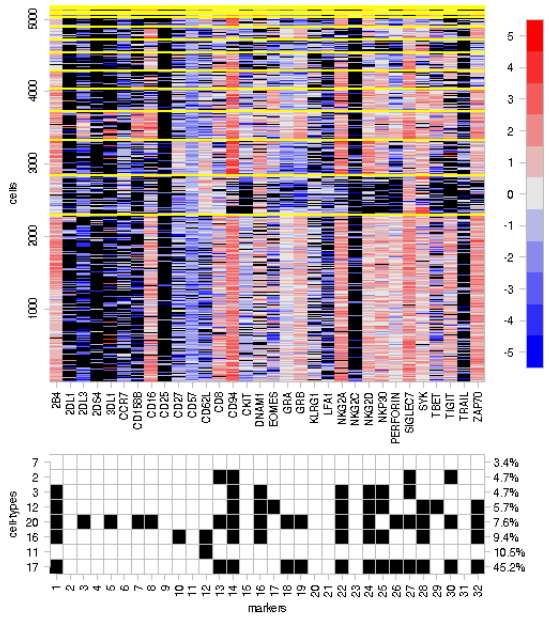
We fit the proposed model to a real data set comprising cord blood samples from three patients ($I = 3$), with the number of cells $N = (41474, 10454, 5177)$, for $J = 32$ markers. The heatmaps in Figure 6 illustrate the observed expression levels y_{inj}. Markers and cells are in columns and rows, respectively. Red, blue and black colors represent high expression levels, low expression levels and missing values, respectively. From the figure, the expression levels of some markers are missing in most cells. For instance, marker CD25 (column 9) is missing in more than 80% of cells in all samples. In each sample, the proportion of missing values in each marker can be as low as 0.1% and as high as 80%. Median proportions of missing values across all markers are 22%, 17% and 18% in the samples, respectively.

To carry out posterior inference, we specified prior distributions using hyperparameters similar to those in the simulation studies. As a preliminary analysis, we fixed β_{0i} and β_{1i}, parameters for the imputation of missing values. 2000 samples from the posterior distribution were obtained after a burn-in period of 1000 iterations. The results are summarized in Figure 6. Point estimates for $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are obtained using the method in § 2.2.3. Plots of $\hat{\boldsymbol{Z}}_i$ are given with the corresponding $\hat{\boldsymbol{w}}_i$ at the bottom of the panels. From the estimated weights ($\hat{\boldsymbol{w}}_i$), the samples have some common phenotypes such as phenotypes 17, 15 and 18. $\hat{\boldsymbol{w}}_i$ also shows heterogeneity across samples. For example, phenotype 17 comprises 29.7%,



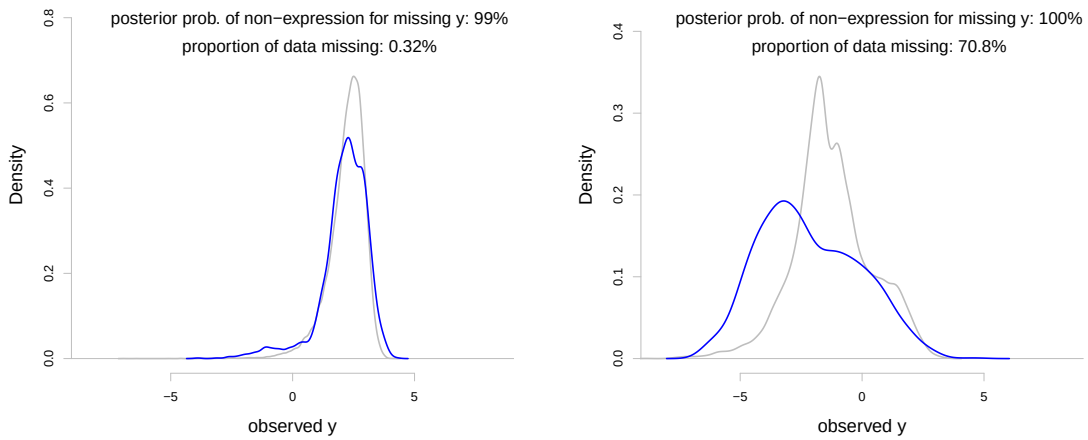
(a) Sample 1

(b) Sample 2



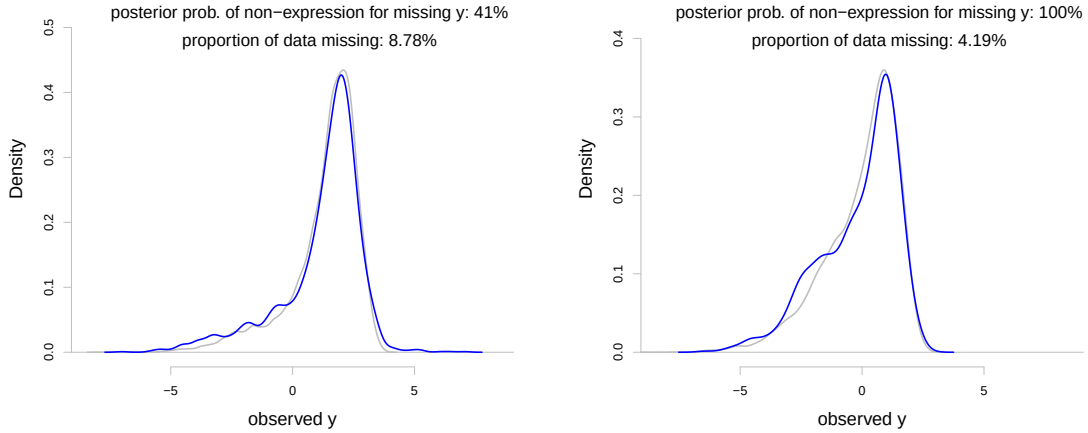
(c) Sample 3

Figure 6: [CB Data] Heatmaps of y for the samples. Cells and markers are in rows and columns, respectively. Each column contains the expression levels of a marker for all cells in the sample. High expression levels are red, low expression levels are blue, missing values are black. Cells are rearranged by their posterior estimate of phenotype indicator, $\hat{\lambda}_{in}$. Horizontal lines separate cells in different estimated phenotypes. At the bottom of each panel, the transpose of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are provided for each sample. We include phenotypes having largest $\hat{w}_{ik}$ to explain at least 90% of the cells in a sample.



(a) Sample 1, marker 1

(b) Sample 1, marker 2



(c) Sample 2, marker 22

(d) Sample 3, marker 19

Figure 7: [CB Data] Comparison of posterior predictive distribution of y_{inj} with $m_{inj} = 0$ (observed data, blue) to their empirical distributions of observed data (grey) for some selected (i, j). An estimate of posterior probabilities of non-expression $\Pr(z_{j,\lambda_{in}} = 0 \mid \boldsymbol{y},)$ averaged in a sample for missing y is shown in each panel.

48.8%, and 45.2% of cells for samples 1, 2 and 3, respectively. Similarly, phenotype 16 is most prevalent in sample 1. $\hat{w}_{ik}$ for the phenotype is 29.7%, 10.2%, and 9.4% in samples 1, 2 and 3, respectively. Also, note that some phenotypes are similar. For example, phenotypes 8 and 12 in panel (a). A possible reason for this is that independence across columns under the prior model for $\boldsymbol{Z}$ allows identical columns with positive probability. For the heatmaps on the top of the panels, y_{inj} are sorted by a posterior estimate $\hat{\lambda}_{in}$ of their cell phenotypes. The horizontal dotted lines separate cells using $\hat{\lambda}_{in}$. From the heatmaps, the cells having

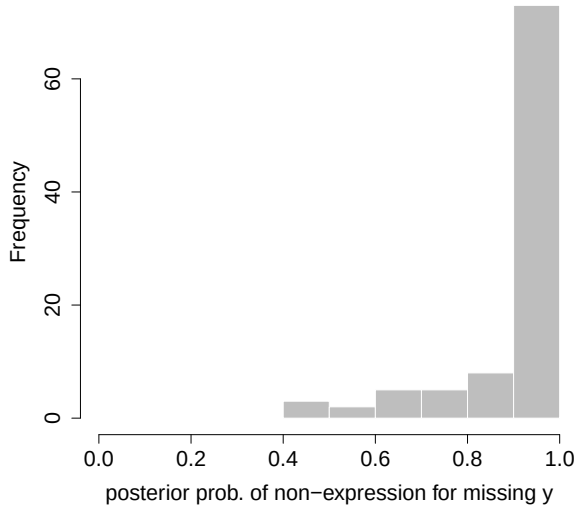
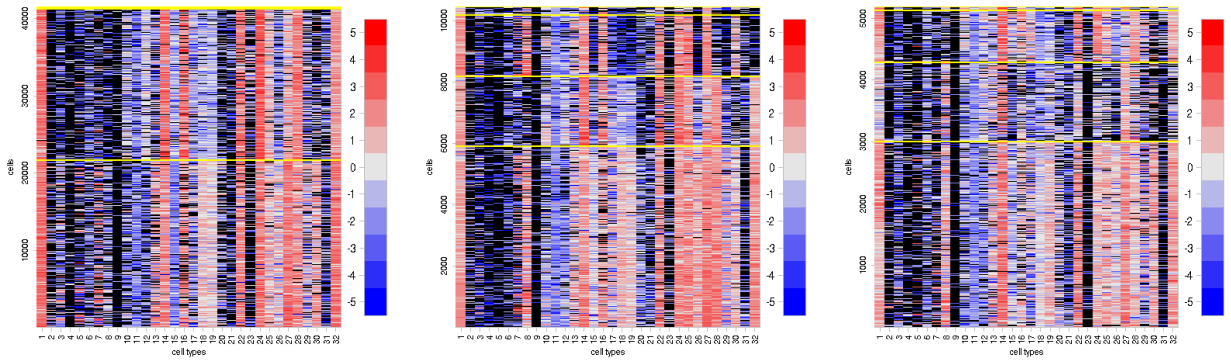


Figure 8: Histogram of posterior probabilities $\hat{q}_{ij}$ of non-expression for missing for all (i, j). The peak at the value of 1 suggests that most of the time, a marker is estimated as no expression if its expression level is missing.

a phenotype have similar expression patterns, implying that the model provides reasonable estimates of underlying cell subpopulations.

To assess model fit, we compare the posterior predictive distribution of y_{inj} with $m_{inj} = 0$ to the distribution of observed data. Figure 7 illustrates the posterior predictive distributions (blue) and empirical distribution estimates of the observed data y_{inj} (grey) for some selected (i, j). For the four cases, 0.32%, 70.8%, 8.78% and 4.19% of y_{inj} are missing, respectively. The model reasonably well fits for the observed data, especially for the three cases in panels (a), (c), and (d). For the case in panel (b), the fit is deteriorated. Some bigger discrepancy between the posterior predictive estimates and the empirical estimate is observed for negative values of y_{inj}. It is possibly because the proportion of missing data is extremely high (70%). We next check the model fit for missing data. We compute $\hat{q}_{ij} = \sum_{n=1}^{N_i} \mathbb{I}(m_{inj} = 1)\hat{\Pr}(z_{j\lambda_{in}} = 0 \mid \boldsymbol{y}, \mathbf{m}) / \sum_{n=1}^{N_i} \mathbb{I}(m_{inj} = 1)$ for (i, j), averaged posterior probability that a cell with missing value for marker j has a phenotype for which the marker is not expressed in sample i. That is, we examine how often a maker with expression level not observed was estimated as no expression. Values close to 1 for $\hat{q}$, implying that a maker is not expressed with high probability if its expression level is not observed, comply with our subject information. For the four cases in the figure, $\hat{q}$ are 99%, 100%, 41% and 100%, respectively. While $\hat{q}$ is close to 1 for the three cases in (a), (b) and (d), $\hat{q}$ is small for the case of $i = 2$ and $j = 22$ in (c), possibly because many of observed values are around zero and the missing proportion is not



(a) Sample 1 (b) Sample 2 (c) Sample 3

Figure 9: [CB data analyzed using FlowSOM] Heatmaps of y for the samples sorted by their cluster labels. Cells and markers are in rows and columns, respectively. High expression levels are red, low expression levels are blue, missing values are white. Yellow horizontal lines divide estimated cell clusters.

large. Figure 8 illustrates a histogram of $\hat{q}$ for all (i, j). The histogram is skewed such that there is a spike at 1, and most of the time our model learns that a marker is not expressed for missing observations.

Comparison of FlowSOM and FAM for CB Data For comparison, we fit FlowSOM to the CB data. Since FlowSOM does not account for different samples and missing values, we combined all samples into one sample, and set the missing values to be the minimum value of observed y_{inj}, as in the analysis of the simulated data in § 2.3. Cell clustering results under FlowSOM are summarized in Figure 9. FlowSOM identified four cell clusters. The proportions of cells assigned to the clusters are (0.522, 0.47, 0.001, 0.006), (0.565, 0.218, 0.192, 0.025) and (0.579, 0.247, 0.162, 0.012) for the three samples, respectively. Heterogeneity between cells in an estimated cluster under FlowSOM is greater than that within an inferred cell phenotype under the proposed model in Figure 6. In particular, in Figure 6, phenotype 20 has a very distinct expression pattern compared to the other phenotypes and its abundance estimates are $\hat{w}_{ik} = 15.1\%$, 6.1% and 7.6% in the samples. On the other hand, those cells do not form a separate cell cluster under FlowSOM.

2.5 Conclusions

We have proposed a Bayesian FAM to study NK-cell diversity from mass cytometry data in the presence of missing data. We used a simulation study to show that our model is able to recover the true latent feature allocation matrix generating the observed data. We are also

able to learn abundances of cell phenotypes with high accuracy, especially for prevalent phenotypes. Assumptions about the missing mechanism are untestable. But through providing an informed prior distribution on the missing mechanism, we are able to impute data missing not at random such that most of the imputed values correspond to the non-expression of markers. This is consistent with scientists' understanding of CyTOF instruments. While not explicitly shown, the posterior variance for $\boldsymbol{Z}$ can be computed easily from the posterior samples of $\boldsymbol{Z}$ when the number of cell types K is fixed. This will help in quantifying uncertainty about the learned NK-cell subpopulations. A simulation study comparing our model to FlowSOM shows that in some scenarios, FlowSOM and our FAM can retrieve the same clusters, and have similar performance. FlowSOM is orders of magnitudes faster than our model. But we are able to model the latent phenotype structure directly, whereas FlowSOM only implicitly models the latent phenotypes. Our proposed FAM has may be more effective at discovering latent phenotypes in data that comprise highly irregularly spaced clusters of cells, at a computational cost. A sensitivity analysis for selecting the upper-bound for the number of latent features K in the model suggests that it is preferable to make K large enough to potentially accommodate more cell types.

When applied to real cord blood data, we observed some redundancy in phenotypes estimates. This can be remedied using a repulsive FAM and will be included in Project 2.

3 Project 2: Repulsive Feature Allocation Model

3.1 Introduction

In project 2, we propose a repulsive FAM (rep-FAM), where repulsion discourages similar features. The traditional IBP assumes a priori independence between features and may yield redundant features. Thus, the independence assumption may not be desirable in many applications. The concept of repulsion is introduced to penalize creating similar features, resulting in a more parsimonious representation of the underlying structures. The property of inducing parsimony can be more critical for analyzing heterogeneous samples collected from different backgrounds, for example, a joint analysis of cord blood samples and samples collected from healthy subjects. Different approaches for repulsive models have been developed mostly in the context of mixture models [Petralia et al., 2012, Quinlan et al., 2017b, Xie and Xu, 2017, Quinlan et al., 2017a]. Independent priors for component specific parameters in a mixture are commonly assumed. For example, in a Dirichlet process mixture model, the atoms are iid draws from the baseline distribution and mixture weights are

constructed through stick-breaking. The independence assumption can produce redundant mixture components located close together, resulting in over-fitting. To separate mixture components, the repulsive models include a repulsion function and assume a joint model for all component specific parameters. The models smoothly push the components apart based on pairwise distances through some repulsion parameters, resulting in well separated clusters. Xu et al. [2016] innovatively used the detrimental point process (DPP) for repulsive mixture models and feature allocation model. We take a different approach to create repulsive features by exploiting repulsive mixture model developed in Quinlan et al. [2017a]. The rep-FAM explicitly incorporates a model for the repulsion that penalizes the inclusion of similar features, while DDP uses the determinant of a matrix as a repulsiveness metric. Properties of the proposed rep-FAM are explored through simulation studies and compared to the IBP. In addition, the model is further extended to let samples possess a subset of cell phenotypes. In our applications, some cell phenotypes can only be present in cord blood samples or healthy-subject samples, while some cell types are shared by both. By letting abundances exactly be zero for some phenotypes, different sets of phenotypes can be used to describe samples. The remainder of this section will outline the proposed rep-FAM in § 3.2 and present simulation studies to compare the rep-FAM and to the IBP in § 3.3.

3.2 Probability Model

Similar to the notation in § 2.2, suppose that I samples are taken from subjects, $i = 1, ..., I$. Sample i consists of N_i cells, $n = 1, \ldots, N_i$ and for each cell, expression levels of J markers are measured. We introduce x_i to denote covariates for sample i. In our data of cord blood samples and healthy subject samples, let $x_i = 0$ or 1 if sample i is a cord blood sample or a healthy subject sample, respectively. We also let y_{inj} represent the transformed observed expression levels of marker j in cell n for sample i and let m_{inj} represent its binary missingness indicator, where $m_{inj} = 0$ if y_{inj} is observed and $m_{inj} = 1$ otherwise. We assume the sampling distributions in (2) and (5) for y_{inj} and m_{inj}.

Repulsive Feature Allocation Model: Recall that a $J \times K$ binary matrix $\boldsymbol{Z}$ characterizes K different cell phenotypes. Let $v_k \mid \alpha \overset{iid}{\sim} \text{Be}(\alpha, 1)$, $k = 1, \ldots, K$. We define a joint distribution of $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, z_K]$ as

$$P(\boldsymbol{Z} \mid v, C_\phi) \propto \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} v_k^{z_{jk}} (1 - v_k)^{1 - z_{jk}} \right\} \times \prod_{k_1=1}^{K-1} \prod_{k_2=k_1+1}^{K} \left\{ 1 - C_\phi \left(\rho(\boldsymbol{z}_{k_1}, \boldsymbol{z}_{k_2}) \right) \right\}, \tag{7}$$

where $\rho(\boldsymbol{z}_{k_1}, \boldsymbol{z}_{k_2})$ measure distance between columns k_1 and k_2, for $k_1 \neq k_2$, and $C_\phi(\cdot)$ is a continuous decreasing function in distance with $C_\phi(0) = 1$ and $\lim_{d\to\infty} C_\phi(d) = 0$. For a distance metric, we use $\rho(\boldsymbol{z}_{k_1}, \boldsymbol{z}_{k_2}) = \sum_{j=1}^{J} |z_{jk_1} - z_{jk_2}|$, the number of discordance between columns k_1 and k_2. The function $C_\phi(\cdot)$ can be interpreted as a proximity function. A suitable form is $C_\phi(d) = \exp(-d/\phi)$. Quinlan et al. [2017b] showed that the model in (7) has a finite normalizing constant and the distribution is proper. Under the model in (7), probability 0 is assigned to $\boldsymbol{Z}$ having identical columns. For matrices $\boldsymbol{Z}$ that have the same number of 0's and 1's, $\boldsymbol{Z}$ with similar columns has a smaller probability since $C_\phi(\cdot)$ is decreasing in distance. $C_\phi(\cdot)$ smoothly penalizes any $\boldsymbol{Z}$ having similar columns in the prior and can remove redundant columns in posterior inference. Note that different from the IBP, under the model in (7), $\Pr(z_{jk} = 1) \neq v_k$ due to the repulsive function. We place a prior on α, such that $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$.

Feature Selection We assume that cord blood samples ($x_i = 0$) and healthy subject samples ($x_i = 1$) can have distinct sets of cell phenotypes. We introduce binary indicators, $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \ldots, K$, to indicate whether samples from x possess phenotype k ($\delta_{xk} = 1$), or do not possess phenotype k ($\delta_{xk} = 0$). Assume $\delta_{xk} \overset{ind}{\sim} \text{Ber}(p_x)$ and $p_x \overset{iid}{\sim} \text{Be}(a_p, b_p)$. Let the unnormalized cell phenotype abundances be $\tilde{w}_{ik} \overset{ind}{\sim} \text{Ga}(a_W/K, 1)$, $i = 1, \ldots, I$, and $k = 1, \ldots, K$. We define relative abundances in sample i from x_i as $w_{ik} = \tilde{w}_{ik}\delta_{x_ik} / \sum_{\ell=1}^{K} \tilde{w}_{i\ell}\delta_{x_i\ell}$. Relative abundance w_{ik} is exactly zero for $\delta_{x_ik} = 0$. Samples from x have the same subset of phenotypes but can have different relative abundances over the selected phenotypes. Phenotypes with $\delta_{0k} = \delta_{1k} = 1$ appear in all samples, while some are present in only one type of samples. Feature-selection by δ_{xk} efficiently facilitates joint analysis of samples obtained from different sources. The model also allows phenotypes to absent in all samples, implying that some phenotypes are not used to describe any cells, providing a more parsimonious representation of cell-populations in samples. The probability models for the other parameters remain unchanged as in § 2.2.

3.3 Simulation Studies

To better understand the behavior of the rep-FAM as a prior distribution for the feature allocation matrix in a FAM, we performed a small-scale simulation study. For preliminary study, we did not include δ_{xk} and kept the same model in § 2.2 for $\boldsymbol{w}_i$. Data were simulated as outlined in 2.3, but with some differences as follows: The number of observations for each sample are $N = (300, 200, 100)$, the true number of latent features is $K^{\text{TR}} = 4$, the number

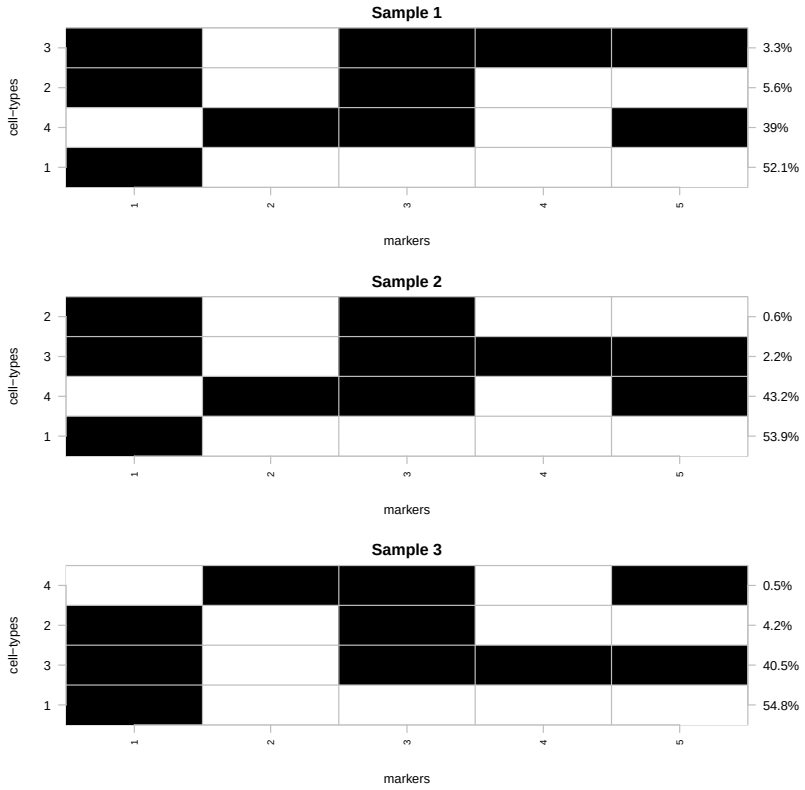
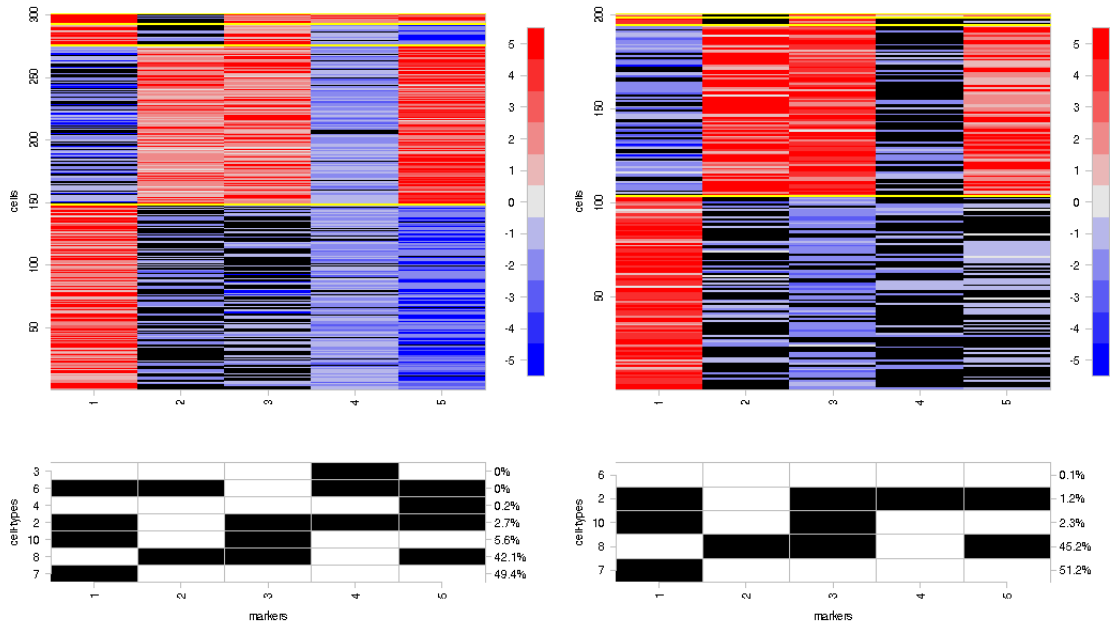


Figure 10: The transpose of $\boldsymbol{Z}^{\mathrm{TR}}$ with markers in columns and latent phenotypes in rows. Black and white represents $z_{jk}^{\mathrm{TR}} = 1$ and 0, respectively. The phenotypes and $\boldsymbol{w}_i^{\mathrm{TR}}$ are shown on the left and right sides of each panel. All samples share the same $\boldsymbol{Z}^{\mathrm{TR}}$ and the phenotypes are arranged in order of w_{ik}^{TR} within each sample.

of markers is $J = 5$, a hand-picked $\boldsymbol{Z}^{\mathrm{TR}}$ was used, and the feature abundances $\boldsymbol{w}_i^{\mathrm{TR}}$ were simulated from a Dirichlet$(15, 15, 1, 1)$, for each sample i. The simulated $\boldsymbol{Z}^{\mathrm{TR}}$ and $\boldsymbol{W}^{\mathrm{TR}}$ are shown in Figure 10.

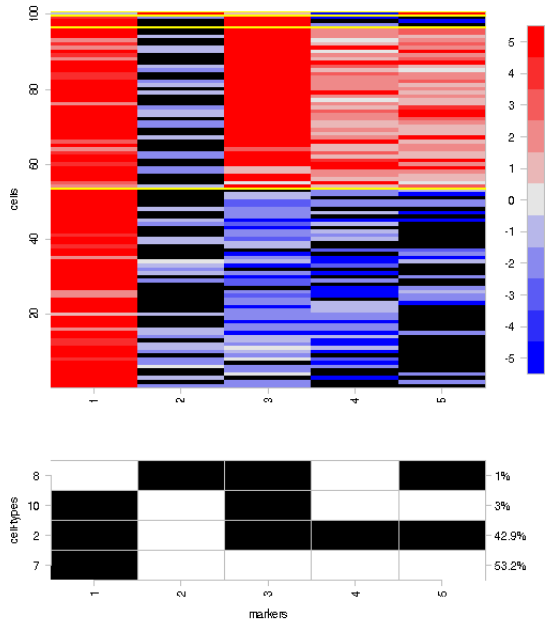
We fit the FAM with the IBP in Project 1 and the rep-FAM prior. We used the same hyperparameters as in § 2.3 but with $K = 10$. We initialized the parameters for the MCMC simulation as outlined in § 2.3. We then implemented posterior inference using MCMC simulation over 3,000 iterations, discarding the first 1,000 iterations as burn-in. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots. We found no evidence of practical convergence problems.

The posterior estimates for $\boldsymbol{Z}$ and $\boldsymbol{w}_i$ for the rep-FAM model and the FAM with an IBP prior are summarized in Figures 11 and 12, respectively. The figures also include the heatmaps of y for each sample, with cells rearranged by $\hat{\lambda}_{in}$. The posterior estimates for $\boldsymbol{Z}$ and $\boldsymbol{w}_i$ were similar to the simulation truth in both cases. However, in the posterior estimate of $\boldsymbol{Z}$



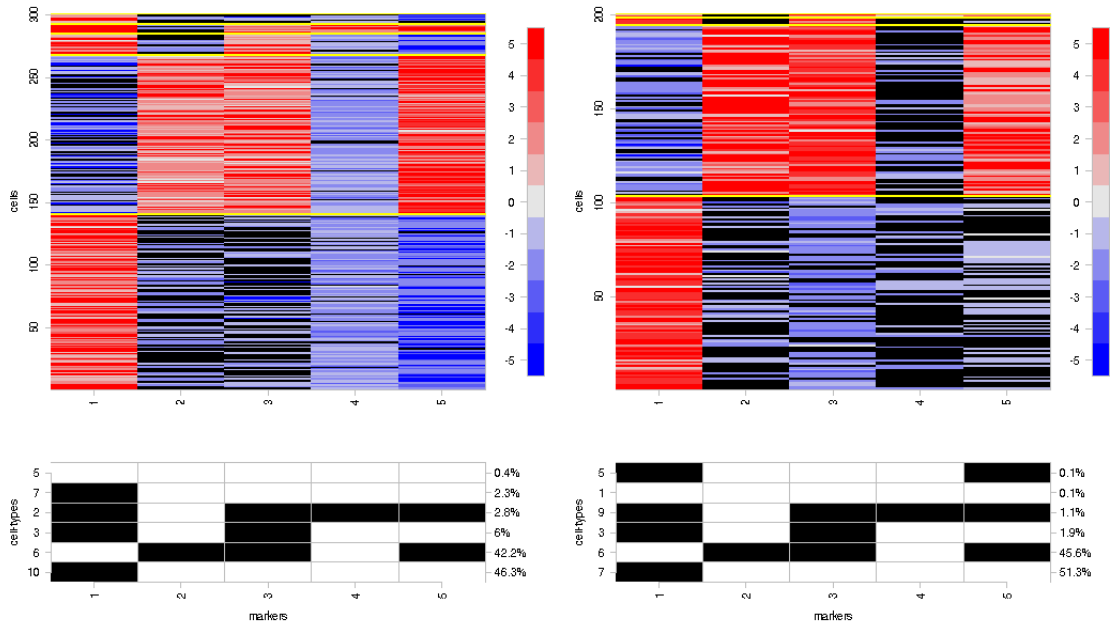
(a) Sample 1

(b) Sample 2



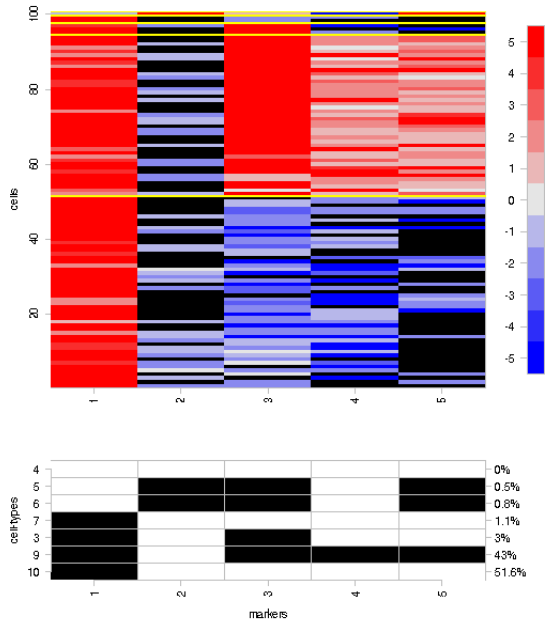
(c) Sample 3

Figure 11: [Rep-FAM Simulation Study (rep-FAM prior)] Heatmaps of y for simulated data. Cells and markers are in rows and columns, respectively. Each column contains the expression levels of a marker for all cells in the sample. High expression levels are red, low expression levels are blue, missing values are black. Cells are rearranged by the corresponding posterior estimate of their phenotype indicator, $\hat{\lambda}_{in}$. Yellow horizontal lines separate cells by different phenotypes. At the bottom of each panel, the transpose of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are provided for each sample. We include phenotypes having largest $\hat{w}_{ik}$ to explain at least 99.9% of the cells in a sample.



(a) Sample 1

(b) Sample 2



(c) Sample 3

Figure 12: [FAM with IBP] Heatmaps of y for simulated data. Cells and markers are in rows and columns, respectively. Each column contains the expression levels of a marker for all cells in the sample. High expression levels are red, low expression levels are blue, missing values are black. Cells are rearranged by the corresponding posterior estimate of their phenotype indicator, $\hat{\lambda}_{in}$. Yellow horizontal lines separate cells by different phenotypes. At the bottom of each panel, the transpose of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are provided for each sample. We include phenotypes having largest $\hat{w}_{ik}$ to explain at least 99.9% of the cells in a sample.

for the regular FAM, features are sometimes repeated. For instance, features 10 and 7 for sample 1 in panel (a) are identical and compose 46.3% and 2.3% of the cells, respectively. Similarly, features 5 and 6 are duplicates for sample 3 in panel (c). Feature 10 compose 46.3% and 51.6% of the data in samples 1 and 3 respectively, while feature 7 composes 2.3% and 1.1%, respectively in samples 1 and 3 respectively. These are cases where one feature has a strong presence in the sample, but is duplicated. In sample 3, features 5 and 6 are also duplicates, but their presence is small in the sample, composing only 0.8% and 0.5% respectively. Note that $\hat{\boldsymbol{Z}}$ in the rep-FAM does not contain duplicated features. This provides more natural interpretation of $\boldsymbol{Z}_i$ and $\boldsymbol{w}_i$, and a more parsimonious representation of the cell types. Specifically, posterior estimate $\hat{\lambda}_{in}$ contains only four cell-types in the rep-FAM, and the estimated cell types are the true cell types. In contrast, seven cell types are estimated in the FAM, but some cell types are repeated.

We will further investigate the rep-FAM through intensive simulation studies for comprehensive understanding. Also, we will analyze CB and healthy subject samples jointly with the proposed rep-FAM with feature selection to provide sample-specific sets of cell phenotypes.

4 Project 3: Feature Allocation Model with Regression for Abundances of Features

In Project 3, we further extend the proposed FAM to analyze samples taken at multiple time points from a patient after NK cell infusion. Suppose I samples are taken at time points $t_1, \ldots, t_I$. As an NK cell population evolves over time, samples may have different sets of cell phenotypes with different abundances. We model this process by letting phenotype abundances $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tK})$ be a function of time (t) after treatment. Inferred $\boldsymbol{w}(t)$ reflects the evolutionary process of cell subpopulation expansion over time. $\boldsymbol{Z}$ includes all phenotypes that can be possessed in a sample, and different compositions of NK cell populations in different samples is modeled through phenotype abundances $w_{t_i,k}$, $i = 1, \ldots, I$, which change over time in a time-dependent manner.

Similar to the model in § 3, let $\xi_{t_1,k}$ represent unnormalized abundance of phenotype k in sample 1 collected at time t_1. We obtain relative abundances by rescaling $w_{t_1,k} = \xi_{t_1,k} / \sum_{\ell=1}^{K} \xi_{t_1,\ell}$. We fix $\xi_{t_1,1} = a$, an arbitrary positive number, to avoid potential identifiability issues, and let $\xi_{t_1,k} = \max(\xi'_{t_1,k}, 0)$, for $k \geq 2$, where $\xi'_{t_1,k} \overset{iid}{\sim} \mathrm{N}(0, s_1^2)$. Since $\xi_{t_1,1}$ is fixed at $a > 0$, phenotype 1 is present in sample 1 and its relative abundance is determined by $\xi_{t_1,2}, \ldots, \xi_{t_1,K}$. For any phenotype with $\xi'_{t_1,k} < 0$, $k = 2, \ldots, K$, its relative abundance is zero and the phenotype is

absent. That is, sample 1 can have a subset of K phenotypes and if the cells in sample 1 have the same phenotype, the phenotype is set to be phenotype 1. Values of a and s_1^2 need to be jointly calibrated. For the remaining samples, we assume $\xi_{t_i,k} = \max(\xi'_{t_i,k}, 0)$, $i = 2, \ldots, I$ and $k = 1, \ldots, K$ with $\xi'_{t_i,k} = \xi'_{t_1,k} + f_k(t_i)$. $f_k(t)$ is a phenotype-specific function of time. $\xi'_{t_1,k}$ serves as a baseline abundance of phenotype k and $f_k(t)$ explains how abundance of phenotype k changes over time. Depending on $f_k(t)$, samples may have different subsets of phenotypes. Various functions can be used for $f_k(t)$. One choice for $f_k(t_i)$ is a quadratic function in time t, $f_k(t) = \xi'_{t_1,k} + \beta_{k1}t + \beta_{k2}t^2$, possibly with some constraints on β_{k1} and β_{k2}. For example, if $\beta_{k2} \leq 0$, the mean abundance of a phenotype cannot increase. We will further investigate this model so as to accommodate biological knowledge on dynamics of cell populations in the model for $\boldsymbol{w}$, and potentially in the model for $\boldsymbol{Z}$.

5 Timeline

Based on the projects listed above, we propose the following timeline.

Project	Academic Quarter
Project 1	Fall 17 - Fall 18
Project 2	Fall 18 - Spring 19
Project 3	Winter 19 - Fall 19
Thesis	Fall 19 - Winter 20

References

John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Tamara Broderick, Jim Pitman, Michael I Jordan, et al. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

Tamara Broderick, Lester Mackey, John Paisley, and Michael I Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.

Hao Chen, Mai Chan Lau, Michael Thomas Wong, Evan W Newell, Michael Poidinger, and Jinmiao Chen. Cytofkit: a bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS computational biology*, 12(9):e1005112, 2016.

Regina K Cheung and Paul J Utz. Screening: Cytof - the next generation of cell detection. *Nature Reviews Rheumatology*, 7(9):502, 2011.

David B. Dahl and Peter Müller. Summarizing distributions of latent structures. *Bayesian Nonparametric Inference: Dependence Structures & Applications Oaxaca, Mexico*, 2017.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Alexander M Franks, Edoardo M Airoldi, and Donald B Rubin. Non-standard conditionally specified models for non-ignorable missing data. *arXiv preprint arXiv:1603.06045*, 2016.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Thomas L Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224, 2011.

Kerstin Johnsson, Jonas Wallin, and Magnus Fontes. Bayesflow: latent modeling of flow cytometry cell populations. *BMC bioinformatics*, 17(1):25, 2016.

Juhee Lee, Peter Müller, Kamalakar Gulukota, Yuan Ji, et al. A bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics*, 9(2):621–639, 2015.

Juhee Lee, Peter Müller, Subhajit Sengupta, Kamalakar Gulukota, and Yuan Ji. 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, 2016.

Kenneth Lo, Florian Hahne, Ryan R Brinkman, and Raphael Gottardo. flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC bioinformatics*, 10(1): 145, 2009.

Francesca Petralia, Vinayak Rao, and David B Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.

José Quinlan, Fernando A Quintana, and Garritt L Page. Density regression using repulsive distributions. 2017a.

José J Quinlan, Fernando A Quintana, and Garritt L Page. Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*, 2017b.

Katayoun Rezvani and Rayne H Rouce. The application of natural killer cell immunotherapy for the treatment of cancer. *Frontiers in immunology*, 6, 2015.

Donald B Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69(346):467–474, 1974.

Anushruti Sarvaria, Dunia Jawdat, J Alejandro Madrigal, and Aurore Saudemont. Umbilical cord blood natural killer cells, their characteristics, and potential clinical applications. *Frontiers in Immunology*, 8, 2017.

Subhajit Sengupta, Jin Wang, Juhee Lee, Peter Müller, Kamalakar Gulukota, Arunava Banerjee, and Yuan Ji. Bayclone: Bayesian nonparametric inference of tumor subclones using ngs data. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 467–478. World Scientific, 2014.

Garnet Suck, Yeh Ching Linn, and Torsten Tonn. Natural killer cells for therapy of leukemia. *Transfusion Medicine and Hemotherapy*, 43(2):89–95, 2016.

Yee Whye Teh, Dilan Grür, and Zoubin Ghahramani. Stick-breaking construction for the indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563, 2007.

Sofie Van Gassen, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

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

Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent indian buffet processes. In *International Conference on Artificial Intelligence and Statistics*, pages 924–931, 2010.

Fangzheng Xie and Yanxun Xu. Bayesian repulsive gaussian mixture model. *arXiv preprint arXiv:1703.09061*, 2017.

Yanxun Xu, Peter Müller, Yuan Yuan, Kamalakar Gulukota, and Yuan Ji. Mad bayes for tumor heterogeneityfeature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514, 2015.

Yanxun Xu, Peter Müller, and Donatello Telesca. Bayesian inference for latent biologic structure with determinantal point processes (dpp). *Biometrics*, 72(3):955–964, 2016.

Xiao Zhang, W John Boscardin, and Thomas R Belin. Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896, 2006.