UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN MODELING FOR HETEROGENEOUS MULTIVARIATE DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Arthur Lui

March 2021

The Dissertation of Arthur Lui
is approved:

Associate Professor Juhee Lee, Chair

Professor Athanasios Kottas

Assistant Professor Zehang Richard Li

Dr. Peter F. Thall

Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bayesian Modeling for Heterogeneous Multivariate Data

by

Arthur Lui

This dissertation, comprising three projects, presents Bayesian statistical methods for analyzing heterogeneous multivariate data, with application to marker expression data obtained from cytometry at time-of-flight (CyTOF). In the first project, 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 CyTOF. Cell subpopulations are characterized by differences in expression patterns of markers, and individual cells are clustered into the subpopulations based on the patterns of their observed expression levels. A finite Indian buffet process is used to model subpopulations as latent features, and a model-based method based on these latent feature subpopulations is used to construct cell clusters within each sample. Non-ignorable missing data due to technical artifacts in mass cytometry instruments are accounted for by defining a static missingship mechanism. The second project builds upon the first by introducing a repulsive FAM (rep-FAM) which restructures the probability distribution of a traditional FAM to identify features more likely to be distinct from each other. The problem that a conventional FAM has a positive probability of repeating a feature is eliminated by the rep-FAM, which also increases the probability of larger differences between features. The rep-FAM thus yields clusters that are more biologically interpretable than those identified by a conventional FAM. The third project presents methods for differential distributions between two experimental conditions, in the context of CyTOF data. A zero-inflated mixture

of log-skew-t distributions is used to model the multi-modal, heavy tailed, and often highly skewed distributions that arise from these marker expression levels. A distance metric is proposed to quantify the degree of difference between distributions under various experimental conditions. In each chapter, we explore the performance and limitations of our proposed methodologies through simulation studies and real data analyses.

For Alex, Catherine, Makayla and Charlie.

Acknowledgments

I would like to thank my collaborators from M.D. Anderson Cancer Center, Katy Rezvani, May Daher, and Rafet Basar, for entrusting me with intriguing questions and data, from their field of work, which form the basis of my dissertation. I am very thankful to Peter Thall, whose contributions were critical to this work, for graciously sharing his expertise and knowledge in statistics and the biological sciences. I thank my past and present committee members, Professors Athanasios Kottas, Abel Rodriguez, and Zehang Li for their helpful feedback, which have strengthened my research. Finally, I thank my advisor, Professor Juhee Lee, for her mentorship, counsel, constant support, and thoroughness when reviewing my work. Her encouragement and teachings have impacted me tremendously as a statistician, for which I am extremely grateful.

Chapter 1

Introduction

1.1 Motivation and Background

A critical step in analyzing heterogeneous data is identifying meaningful subgroups, or clusters, of data which can explain the structure of the heterogeneity. Automating a clustering pipeline, however, is not trivial. Information about "true" clusterings of heterogeneous data is typically not available. Therefore, for a given application, domain experts are usually required to carefully inspect and validate these clusterings. Consequently, iterating through the model development cycle can be arduous for clustering applications, and good clustering strategies usually exploit domain expertise about known or hypothesized latent structures of data generating mechanisms per application. Yet, in many applications, clustering algorithms used do not incorporate knowledge of these latent structures.

Challenges in clustering multivariate data are compounded by the increase in dimensionality, such that visually inspecting data for clusterings is difficult or impractical. For example, cytometry at time-of-flight (CyTOF) data, which is used as a motivating example throughout this dissertation, typically contain expression levels for dozens of cell surface or intracellular markers for each of

the thousands of cells in a blood sample. Human cells are composed of many subpopulations (clusters) of cells with various functionality. Thus, different cells within a sample tend to have different marker expression patterns. To identify known subpopulations, scientists investigating the composition and heterogeneity of cells in samples frequently apply manual gating, in which homogeneous cell subpopulations are sequentially identified and refined by visually inspecting two-dimensional scatter plots of expression levels for a given set of 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.

Dimensionality reduction techniques are sometimes used to aid visually inspecting data for clusterings. The most well-known dimensionality reduction technique is perhaps principal components analysis (PCA) (Wold et al. 1987), in which (multivariate) data are linearly projected into lower dimensions via a change of basis. The t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008, Van Der Maaten 2014) is another dimensionality reduction technique, which, unlike PCA, transforms high-dimensional data to lower dimensions probabilistically and in a non-linear manner. After high-dimensional data are reduced in dimension (to usually 2 or 3), inspecting data for clusters can be much easier. While this is convenient, the amount of information that is lost when reducing the data can be great, and resulting clusterings can be coarse.

For CyTOF data, another task that requires special care is that of comparing marker expression levels across different experimental conditions. Heterogeneous data may not have unimodal or symmetric distributions. Thus, determining differences between data arising from such distributions by shifts in means alone may not be adequate. Again, in the context of cytometry data, heterogeneous

cell samples occasionally yield marker expression levels that have multimodal distributions. Researchers are often interested in the effect of a treatment on cell marker expressions. These effects may further complicate the shapes of the distributions. Statistical methods that comprehensively compare marker expressions and quantify the differences in expression levels are in need to obtain a better understanding of treatment effects.

1.2 Literature Review

1.2.1 Feature Allocation Models and Repulsive Clustering

Various clustering methods used for identifying cell subpopulations from CyTOF data have been proposed and subsequently compared by Weber and Robinson (2016). However, these existing methods cluster only on the observed expression levels, and do not explicitly provide the structure of the cell subpopulations. This dissertation considers a different approach of identifying cell subpopulations for CyTOF data. Cell subpopulations are characterized by latent binary marker "expression patterns". For a set of J markers, a cell expresses a subset of the analyzed markers. These expression patterns can be encoded in a J-dimensional binary vector that indicates which markers are expressed within a cell; 1 for expression of a marker, and 0 otherwise. The number of subpopulations within a sample is rarely known, but due to the diversity of human cells, we expect $K(\geq 1)$ cell subpopulations to exist within a sample. K binary vectors are collected in a $J \times K$ binary matrix $\boldsymbol{Z}$ and a feature allocation model (FAM) is used to model $\boldsymbol{Z}$. The most widely used FAM is the Indian buffet process (IBP) (Ghahramani and Griffiths 2006, Griffiths and Ghahramani 2011), which is the result of an infinite beta-bernoulli model. The IBP is a prior for binary matrices

having an infinite number of columns. In the following paragraphs, an overview of FAMs including the IBP is provided.

In a finite FAM, each of J objects possess a subset of K features. In the culinary analogy, objects and features are called customers and dishes, respectively. A $J \times K$ binary feature allocation matrix $\boldsymbol{Z}$ encodes the subset of features each object possesses – if element $z_{j,k} = 1$, then object j possess feature k; otherwise, object j does not possess feature k. The precise generative process of the finite FAM as presented by Ghahramani and Griffiths (2006) is given by

$$v_k \mid \alpha \overset{ind}{\sim} \text{Beta}(\alpha/K, 1) \quad \text{and} \quad z_{j,k} \mid v_k \overset{ind}{\sim} \text{Bernoulli}(v_k), \tag{1.1}$$

for $k \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, J\}$. In Equation (1.1), α is a mass parameter that determines the expected number of features each object possesses. As K approaches infinity, an infinite FAM known as the Indian buffet process (IBP) arises. The probability mass function (pmf) of the IBP is defined over an equivalence class left-ordered binary matrices, after dropping columns of zeros. This pmf is available in closed form given mass α. Under this equivalence class, though K is infinite, the total number of non-zero columns K_+ has a distribution of Poisson(αH_J), where $H_J = \sum_{j=1}^{J} 1/j$ is the J-th harmonic number. The IBP can be alternatively represented by a stick-breaking construction for the IBP (Teh et al. 2007), similar to that of the Dirichlet process. Teh et al. (2007) exploited the stick-breaking representation, and proposed an efficient variational inference sampling scheme. The first two parts of this dissertation investigate how the IBP can be utilized as a prior for cell subpopulations. In simple models, such as the linear Gaussian latent factor model (LGLFM) (Griffiths and Ghahramani 2011), efficient Gibbs sampling schemes can be developed for posterior inference of $\boldsymbol{Z}$ and other model parameters even though the effective dimensions of $\boldsymbol{Z}$ ($J \times K_+$)

is random. In more complex models, where the use of transdimensional Markov chain Monte Carlo (MCMC) schemes for posterior inference becomes computationally impractical, practitioners opt to use the simpler finite FAM with K set at a reasonably large value as determined per application. When K is to be learned, multiple models with various fixed K can be fit and then selected via a model selection criterion.

The IBP prior and its extensions in FAMs have been applied to a range of applications. Hai-son and Bar-Joseph (2011) proposed and applied an extension to the IBP that integrates prior information on interactions between objects to construct interaction networks for microRNA data. Sengupta et al. (2014) and Lee et al. (2015, 2016) proposed categorical IBP extensions to describe tumor heterogeneity (TH) using next-generation sequencing (NGS) data. Xu et al. (2015) proposed an efficient inference algorithm based on small-variance asymptotic approximations for a class models of models using IBP priors and exponential-family likelihoods. Their method was also applied to TH. Ni et al. (2019) analyzed electronic health records by developing a categorical matrix factorization method based on the IBP. Variants of the IBP that relax exchangeability assumptions have also been proposed. Williamson et al. (2010) proposed the dependent IBP, which induces dependence between observations through hierarchical Gaussian processes. Miller et al. (2012) proposed the phylogenetic Indian buffet process, which introduces dependencies between objects by conditioning on a dependency tree, and includes the regular IBP as a special case when all branches meet at the root. This model performs well for data which exhibit genealogical relationships and expresses prior object similarity through a tree. Gershman et al. (2014) proposed a distance dependent IBP in which objects that are similar in terms of some data-external criteria are more inclined to share features. The Indian buffet Hawkes process

(Tan et al. 2018) extended the IBP to capture latent temporal dynamics by incorporating ideas from the Hawkes process. Williamson et al. (2020) presented a class of nonexchangeable dynamic models constructed by adapting the IBP. These models are tailored to data that are assumed to be generated by latent features exhibiting temporal persistence.

A shortcoming of the finite FAM is that identical columns can appear in the feature allocation $\boldsymbol{Z}$. This is a nuisance when interpreting the feature allocation is of importance. In CyTOF applications, where columns of $\boldsymbol{Z}$ refer to cell subpopulations, each column should distinctly encode a unique subpopulation. To this end, a repulsive FAM (rep-FAM) is developed in Chapter 3. The idea of repulsive clustering is not new. Petralia et al. (2012), Quinlan et al. (2017, 2018), and Xie and Xu (2020) presented repulsive priors for mixtures of distributions. Concepts of repulsion are also used in determinantal point processes (DPP) and Gibbs point processes, which are able to describe spatial point patterns in data where nearby points repel each other. Lavancier et al. (2015) provided a comprehensive review of DPP for repulsion of spatial points and discussed computation methods for inference. Xu et al. (2016) introduced repulsiveness into mixture models and FAMs through determinantal point process priors as priors on latent mixture components and feature-specific parameters, respectively. When repulsiveness is imposed on components of a mixture model, the resulting mixture components are not only more distinct and less redundant, but they are also more interpretable. In a clustering application where mixture components are modeled independently, similar and superfluous components can be formed. This causes difficulty in interpreting the components as the superfluous components likely belong to the same group. Petralia et al. (2012) presented the following general class of repulsive priors to

smoothly repel components in mixture models:

$$\pi(\boldsymbol{\theta}) = C \prod_{k=1}^{K} g(\theta_k) f(\boldsymbol{\theta}), \tag{1.2}$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$ are mixture model component parameters (e.g. mixture component means in a Gaussian mixture model) for a mixture model with K components, and C is a normalizing constant. $f(\boldsymbol{\theta})$ is a repulsion function that may take, for instance, the form

$$f(\boldsymbol{\theta}) = \prod_{(i,j) \in A} \exp(-\tau / d(\theta_i, \theta_j)), \tag{1.3}$$

with $A = \{(i, j) : i = 1, \ldots, K; j < i\}$, a temperature parameter $\tau > 0$ which can accentuate repulsiveness between components, and a distance metric $d(\cdot, \cdot)$. Critically, as the distance d between a pair of components (θ_i, θ_j) approaches 0, $f(\boldsymbol{\theta})$ approaches 0, thereby prohibiting identical components. Moreover, as d approaches infinity, f approaches 1 such that no penalty is incurred when components are infinitely far apart. Thus, the repulsive function smoothly repels components based on their pairwise distance. Note that τ is doubly intractable as it appears in the normalizing constant C. Estimating τ or any other hyperparameter that may appear in the repulsive function will require special consideration in order to craft efficient algorithms for posterior inference. Alternatively, these hyperparameters are to be either carefully determined in light of expert knowledge per application, or calibrated as in Petralia et al. (2012), Quinlan et al. (2017), and Quinlan et al. (2018).

1.2.2 Finite Mixture Models and Statistical Divergences

Finite mixture models are used throughout this dissertation. Due to their flexibility in modeling various complex distributions, mixture models are often used to model the probability distribution of continuous measurements that cannot be easily represented by a single-component distribution. This makes them a suitable choice for modeling heterogeneous data. Mixture models with K components have the form

$$p(y \mid \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \cdot f(y \mid \theta_k), \tag{1.4}$$

with mixture component parameters θ_k, mixture weights w_k, and possibly multivariate data y. In many applications, the number of mixture components K is not known in advanced but can be estimated as shown by Richardson and Green (1997) using reversible-jumpMCMC (Green 1995), or selected via model selection criteria. A commonly used mixture model is the Gaussian mixture model (GMM) in which $f(\cdot \mid \theta_k)$ is the normal probability density function (pdf) with $\theta_k = (\mu_k, \Sigma_k)$, mean μ_k, and variance (or covariance) Σ_k. Oftentimes, the model is augmented by introducing data $\{y_1, \ldots, y_n\}$, component-membership indicators $\lambda_i \in \{1, \ldots, K\}$ for observation i to facilitate posterior sampling of the parameters μ_k, Σ_k, and w_k. For GMMs, posterior sampling can be performed using Gibbs sampling by sequentially sampling directly from the full conditional distributions of each model parameter under certain priors. Through the indicators λ_i, clusterings or partitions of the data are implicitly formed. Thus, GMMs and other finite mixture models are also used in clustering applications.

Finite mixtures of distributions other than the Gaussian are used in practice to achieve more flexibility. For example, (Peel and McLachlan 2000) presented

the use of t distributions for robust mixture modeling when data contains groups of observations with heavy tails. They demonstrated that in the presence of atypical observations or background noise, clustering using a mixture of normal distributions (even with an additional uniform component to model the atypical observations) can drastically misfit the data, while using a mixture of t distributions can lead to parsimonious and reasonable clusterings. Mixtures of skew-t distributions have also been extensively studied by Frühwirth-Schnatter and Pyne (2010) and reviewed by McLachlan et al. (2019), and demonstrated to be more flexible than mixtures of t and normal distributions as they also accommodate asymmetry in data. As demonstrated by Frühwirth-Schnatter and Pyne (2010), the skew-t distribution contains as special cases the normal, skew-normal, and t distributions and is governed by location, scale, degrees of freedom, and skewness parameters (Azzalini and Capitanio 2003). By varying the parameters, one is able to model skewed data with outliers where a GMM with several components may be needed. The skew-t pdf has the form

$$p(y \mid \mu, \sigma, \nu, \phi) = \frac{2}{\sigma} \cdot t_\nu(u) \cdot T_{\nu+1}\left(\phi \cdot u\sqrt{\frac{\nu+1}{\nu+u^2}}\right), \text{ for } y \in \mathbb{R}, \tag{1.5}$$

where $u = (y-\mu)/\sigma$, $t_\nu(\cdot)$ and $T_\nu(\cdot)$ denote, respectively, the pdf and cdf of a standard Student's t distribution with degrees of freedom ν, μ is the location, σ is the scale, and ϕ is a skew parameter. McLachlan et al. (2019) provided a review of some other non-normal mixture components for density estimation. An important class of such components is the family of generalized hyperbolic distributions (Barndorff-Nielsen 1977, Browne and McNicholas 2015), including the normal inverse Gaussian (Karlis and Santourian 2009) and the asymmetric Laplace distribution (Franczak et al. 2013). The generalized hyperbolic distributions contain the normal, t, and skew-t distributions as special or limiting cases,

and thus offer more modeling flexibility and lead to better fit when compared to their special or limiting cases. Inference algorithms for these generalized hyperbolic distributions, however, often are more complex to implement and require more computation time to run.

While mixture models provide estimates of distributions having various shapes, quantifying differences between distributions is not straightforward. One way to approach this is to measure the statistical distance or divergence between the distributions. A general of family statistical divergences is the f-divergences (Liese and Vajda 2006, Rényi et al. 1961) which have the from

$$D_f(P \,||\, Q) = \int_\Omega f\left(\frac{dP}{dQ}\right) dQ \tag{1.6}$$

where P and Q are two probability distributions over a space Ω such that P is absolutely continuous with respect to Q, and f is a convex function with $f(1) = 0$. By changing the form of f, the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951), squared Hellinger distance (Beran et al. 1977), total variation distance, and many other divergences can be obtained. These divergences can be used to measure differences between continuous parametric distributions. However, for mixtures of discrete and continuous distributions, which are typically referred to as "semicontinuous" distributions, computing these divergences is not appropriate. In addition to positive continuous expression levels, CyTOF data contain a substantial number of zeros due to experimental artifacts and weak levels of marker expression. Fitting zero-inflated mixture models to these semicontinuous data is possible, but computing KL divergence or Hellinger distance may not be appropriate as they are not clearly defined for semicontinuous parametric distributions. In addition, unlike the Hellinger distance, which is bounded between 0 and 1, the KL divergence is not bounded above, which makes interpreting it

difficult.

1.3 Contribution and Organization

The contribution for this work is the development of Bayesian methods for heterogeneous data with applications to CyTOF data. Specifically, CyTOF data obtained from natural killer (NK) cell studies conducted by collaborators at MD Anderson Cancer Center are used for illustrations of the developed methodologies.

In Chapter 2, a Bayesian FAM is presented for identifying cell subpopulations based on multiple samples of cell surface or intracellular marker expression level data obtained by CyTOF. Cell subpopulations are characterized by differences in expression patterns of makers, and individual cells are clustered into the subpopulations based on the patterns of their observed expression levels. A finite Indian buffet process is used to model subpopulations as latent features, and a model-based method based on these latent feature subpopulations is used to construct cell clusters within each sample. 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 based on observed marker expression levels that are applied separately to different samples, the FAM based method can be applied simultaneously to multiple samples, and also identify important cell subpopulations likely to be missed by conventional clustering methods. The proposed FAM-based method is applied to jointly analyze three datasets, generated by CyTOF, to study 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. Simulation studies of the pro-

posed method's behavior under two cases of known subpopulations are presented, followed by analysis of the CyTOF NK cell surface marker data.

Chapter 3 builds upon Chapter 2 by proposing a Bayesian repulsive feature allocation model (rep-FAM) that modifies a conventional FAM by restructuring the probability distribution to identify features more likely to be distinct from each other. The identified features are used to construct cell clusters. The problem that a conventional FAM has a positive probability of repeating features is eliminated by the rep-FAM, which also increases the probability of larger differences between features. The rep-FAM thus yields clusters that are more biologically interpretable than those identified by a FAM. The rep-FAM is applied to identify cell subpopulations based on CyTOF data. Binary features defined by cell surface marker expression patterns are used to define latent cell subpopulations that determine cell clusters, with each cluster characterized by a set of distinct features. Performance of the rep-FAM is examined by simulation, and the model is applied to analyze a CyTOF dataset. Comparisons to a conventional FAM and other existing clustering methods are included.

Chapter 4 presents a Bayesian statistical model that quantifies differential distributions across experimental conditions in marker expression level data obtained by CyTOF. As will be explained in Chapters 2 and 3, CyTOF data typically exhibits excess zeros, outliers, multimodality, or skewness. Thus, a zero-inflated mixture model is used to address excessive zeros and flexibly accommodate various patterns in data. Notably, skew-t distributions for the mixture components are used. Individual mixture components capture skewness and provide robustness to the presence of outliers. Mixture components are allowed to be shared across samples to facilitate borrowing information. While mixture weights are sample-specific, the number of mixture components is selected via calibration.

Distributional differences are quantified and summarized with a distance measure. Model performance is demonstrated via simulation studies and compared to zero-inflated mixture models of normal distributions. CyTOF NK cell surface marker data are analyzed to infer differential expressions of markers across two experimental conditions.

Chapter 5 provides a summary of main contributions from Chapters 2 to 4. Discussions and possible extensions of the methods presented will also be summarized.

Chapter 2

A Bayesian Feature Allocation Model for Identifying Cell Subpopulations Using Cytometry Data

2.1 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) is a decade-old 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 and 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 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 analyze 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 (Chen et al. 2016a), and model-based clustering methods, including flowClust (Lo et al. 2009) and BayesFlow (Johnsson et al. 2016). More sophisticated clustering methods based on Bayesian nonparametric models also have been proposed, see for example by Soriano et al. (2019). Weber and Robinson (2016) performed a study to compare

several 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. 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.

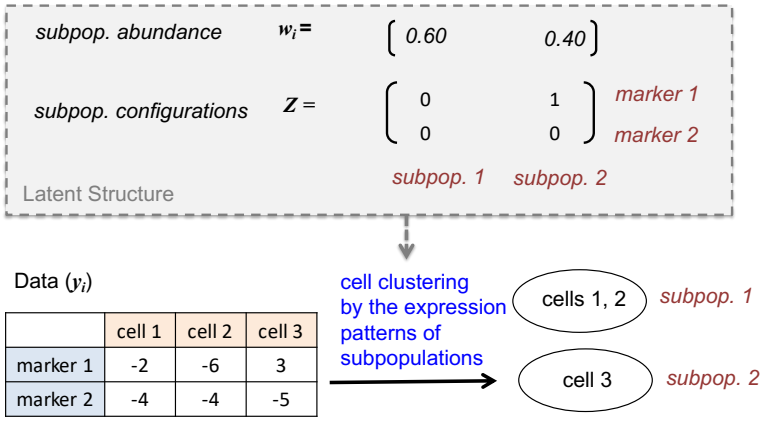


Figure 2.1: A stylized overview of the proposed feature allocation model. $\boldsymbol{Z}$ is a binary matrix whose columns define latent subpopulations, and $\boldsymbol{w}$ is a vector of cell subpopulation abundances. Two subpopulations are constructed in $\boldsymbol{Z}$ based on their marker expression patterns. Cells are clustered into the subpopulations based on their observed expression level patterns.

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 variabil-

ity, often due to technical variation in the cytometry measurement process, as well as variability in the expression measurement process. Figure 2.1 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 analyze different samples separately.

In this chapter, 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 2.1. 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 and Griffiths 2006). Applications of the IBP prior in FAMs for a range of biological applications are given by Hai-son and Bar-Joseph (2011), Chen et al. (2013), Xu et al. (2013), Sengupta et al. (2014), Xu et al. (2015), Lee et al. (2015, 2016), Ni et al. (2019). 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 to yield subpopulations that are characterized by binary indicators denoting subsets of expressed markers, 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 chapter 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 tumor cells. NK cells have the intrinsic ability to infiltrate cancer tissues. Recently, NK cells have been used therapeutically to treat a variety of diseases (Wu and Lanier 2003, Lanier 2008). In particular, NK cells have emerged as a potentially powerful treatment modality for advanced cancers refractory to conventional therapies (Rezvani and Rouce 2015, Suck et al. 2016, Shah et al. 2017, Miller et al. 2005, Lupo and Matosevic 2019, Liu et al. 2020). Because cell-surface protein expression levels are used as markers to describe the behavior 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 chapter is organized as follows. We present the proposed statistical model in § 2.2, simulation studies in § 2.3, and an analysis of the NK cell mass cytometry data in § 2.4. We close with concluding remarks in § 2.5.

2.2 Probability Model

2.2.1 Sampling Model

Index cell samples by $i = 1, 2, ..., I$. Suppose that N_i cells, indexed by $n = 1, \ldots, N_i$, are obtained from the i^{th} 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 cutoff 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, cutoffs obtained from such algorithms are crude, but useful as a starting point for our analysis. Let $c_{i,j}$ denote the cutoff 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 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 i^{th} sample, we assume conditional independence of the cell's J marker values given its latent subpopulation, formally $y_{i,n,1}, \cdots, y_{i,n,J} \mid \lambda_{i,n}$ are independent, and we write the joint model for $(y_{i,n,j}, m_{i,n.j})$ as follows;

$$y_{i,n,j} \mid \mu_{i,n,j}, s^2_{i,n}, \lambda_{i,n} \overset{ind}{\sim} \text{Normal}(\mu_{i,n,j}, s^2_{i,n}), \text{ and} \tag{2.1}$$

$$m_{i,n,j} \mid \rho_{i,n,j}(y_{i,n,j}), \lambda_{i,n} \overset{ind}{\sim} \text{Bernoulli}(1 - \rho_{i,n,j}(y_{i,n,j})). \tag{2.2}$$

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 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.

2.2.2 Priors

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 $j = 1, \dots J$ and $k = 1, \dots, K$,

$$z_{j,k} \mid v_k \overset{ind}{\sim} \text{Bernoulli}\,(v_k) \quad \text{and} \quad v_k \mid \alpha \overset{iid}{\sim} \text{Beta}(\alpha/K, 1). \tag{2.3}$$

As $K \to \infty$, the limiting distribution of $\boldsymbol{Z}$ in (2.3) is the IBP (Ghahramani and Griffiths 2006) with parameter α, after removing all columns that contain only zeros. We assume hyperprior $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ 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.

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 $\{1, \cdots, K\}$. We assume priors $\epsilon_i \overset{iid}{\sim} \text{Beta}(a_\epsilon, b_\epsilon)$ with fixed hyperparameters a_ϵ and b_ϵ, and $\boldsymbol{w}_i \mid K \overset{iid}{\sim} \text{Dirichlet}_K(d/K)$ 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 $i = 1, \cdots, I$, assigning cells to subpopulations using $\{\lambda_{i,n},\ i = 1, \cdots, N_i\}$ induces cell clusters. Thus, in contrast with clustering methods that infer only cell clusters in the i^{th} 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 a 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.

Priors for mean expression level The mean expression level $\mu_{i,n,j}$ of marker j for cell n in sample i in (2.2) 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 $\lambda_{i,n} \in \{1, \cdots, K\}$, 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 $\delta_{0,\ell} \overset{iid}{\sim} \text{TN}^-(\psi_0, \tau_0^2)$, $\ell = 1, \ldots, L_0$ 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 $0 > \mu^\star_{0,1} > \ldots > \mu^\star_{0,L_0}$. 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 cutoff $c_{i,j}$, 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 $\delta_{1,\ell} \overset{iid}{\sim} \text{TN}^+(\psi_1, \tau_1^2)$, $\ell = 1, \ldots, L_1$ 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 $0 < \mu^\star_{1,1} < \ldots < \mu^\star_{1,L_1}$. We then let $\mu_{i,n,j} = \mu^\star_{1,\ell}$ with probability $\eta^1_{i,j,\ell}$, and let $s^2_{i,n} = \sigma^2_i$ for $\lambda_{i,n} > 0$ and assume $\sigma^2_i \overset{ind}{\sim} \text{InverseGamma}(a_\sigma, b_\sigma)$. This leads to a mixture of normals for $y_{i,n,j}$ whose location parameters are determined by the cell's latent subpopulation,

$$\begin{aligned} y_{i,n,j} \mid z_{j,\lambda_{i,n}} = z, \boldsymbol{\mu}^\star_z, \boldsymbol{\eta}^z_{i,j}, \sigma^2_i &\overset{ind}{\sim} F^z_{i,j}, \text{ where} \\ F^z_{i,j} &= \sum_{\ell=1}^{L_z} \eta^z_{i,j,\ell} \cdot \text{Normal}(\mu^\star_{z,\ell}, \sigma^2_i), \\ &\quad \text{for } z \in \{0, 1\},\ k > 0. \end{aligned} \tag{2.4}$$

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

The mixture model in (2.4) 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 (2.4) 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. This provides a basis for obtaining a succinct representation of cell subpopulations. Because the locations $\mu_z^\star$ in (2.4) 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, Stephens 2000, Jasra et al. 2005, Frühwirth-Schnatter 2006).

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 (2.2), we model missingship $m_{i,n,j}$ conditional on $y_{i,n,j}$, i.e., $m_{i,n,j} \mid \rho_{i,n,j}(y_{i,n,j}) \overset{ind}{\sim} \text{Bernoulli}(1 - \rho_{i,n,j}(y_{i,n,j}))$. We assume a logit regression model for the probability $\rho_{i,n,j}$ that $m_{i,n,j} = 0$,

$$\text{logit}(\rho_{i,n,j}) = \beta_{0i} + \beta_{1i} y_{i,n,j} + \beta_{2i} y_{i,n,j}^2. \tag{2.5}$$

We take an empirical approach to specify values of $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{1i})$ in (2.5) for each sample $i = 1, \cdots, I$ 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 $\boldsymbol{\beta}_i$. 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 (2.5) to further examine robustness. § 2.3 and § 2.4 provide details of the sensitivity analyses. There is an extensive literature on analyzing data with observations missing not at random, including methods for Bayesian data imputation and frequentist multiple imputation (Rubin (1974, 1976), Allison (2001), Schafer and Graham (2002), Franks et al. (2016)). We refer to them for alternative models for the missingship mechanism.

Selection of K Instead of estimating K, we cast the problem of selecting a value for K as a model comparison problem. This reduces 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 and Eddy (1979), Gelfand and 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 Appendix §A.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. 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.

2.2.3 Posterior Computation

Let $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \boldsymbol{\sigma}^2, \boldsymbol{\eta}^0, \boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{\epsilon}, \alpha\}$ 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

$$
\begin{aligned}
& p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{m}, K) \\
\propto\ & p(\boldsymbol{\theta} \mid K) \prod_{i,n,j} p(m_{i,n,j} \mid y_{i,n,j}, \boldsymbol{\theta}, K) p(y_{i,n,j} \mid \boldsymbol{\theta}, K) \\
\propto\ & p(\boldsymbol{\theta} \mid K) \prod_{i,n} \left[\prod_{j} \rho_{i,n,j}^{1-m_{i,n,j}} \sum_{\ell=1}^{L_{z_{j,\lambda_{i,n}}}} \eta_{i,j,\ell}^{z_{j,\lambda_{i,n}}} \phi(y_{i,n,j} \mid \mu^{\star}_{z_{j,\lambda_{i,n}},\ell}, \sigma_i^2) \right]^{1(\lambda_{i,n}>0)} \\
& \qquad \times \left[\prod_{j} \rho_{i,n,j}^{1-m_{i,n,j}} \times \phi(y_{i,n,j} \mid 0, s_\epsilon^2) \right]^{1(\lambda_{i,n}=0)}, \qquad (2.6)
\end{aligned}
$$

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 (2.6). 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 Appendix §A.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 $\boldsymbol{A}_i = \{A_{i,(j,j')}(\boldsymbol{Z})\}$, the $J \times J$ pairwise allocation matrix corresponding to a binary matrix $\boldsymbol{Z}$, where

$$A_{i,(j,j')}(\boldsymbol{Z}) = \sum_{k=1}^{K} w_{i,k} \times 1(z_{j,k} = 1) \times 1(z_{j',k} = 1), \quad \text{for } 1 \leq j, j' \leq J, \tag{2.7}$$

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 (2.7) 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,

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

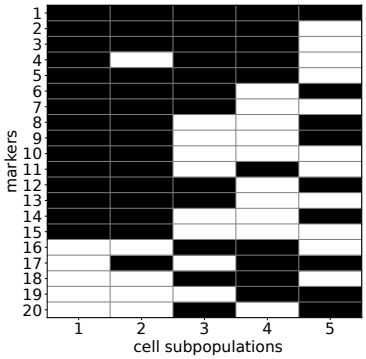
where $\bar{A}_{i,(j,j')}$ is the pairwise allocation matrix averaged by the posterior distribution of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. We use posterior Monte Carlo samples to obtain posterior

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 b^{th} posterior sample of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$, we compute the $J \times J$ adjacency matrix, $\boldsymbol{A}_i^{(b)} = \{A_{i,(j,j')}^{(b)}\}$, $b = 1, \ldots, B$ 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 mean 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 $\hat{\boldsymbol{Z}}_i \in \left\{ \boldsymbol{Z}^{(1)} \ldots \boldsymbol{Z}^{(B)} \right\}$. For $\hat{\boldsymbol{Z}}_i = \boldsymbol{Z}^{(b)}$, we report the posterior point estimates $\hat{\boldsymbol{w}}_i = \boldsymbol{w}_i^{(b)}$ and $\hat{\lambda}_{i,n} = \lambda_{i,n}^{(b)}$.

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 (Wainwright et al. 2008, Blei et al. 2017, 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 Appendix § A.1.2. A Julia package CytofResearch for implementing this methodology is available at `https://github.com/luiarthur/CytofResearch`.

2.3 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 $N_i = 4000$,



(a) $\boldsymbol{Z}^{\mathrm{TR}}$

	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

(b) $\boldsymbol{w}^{\mathrm{TR}}$

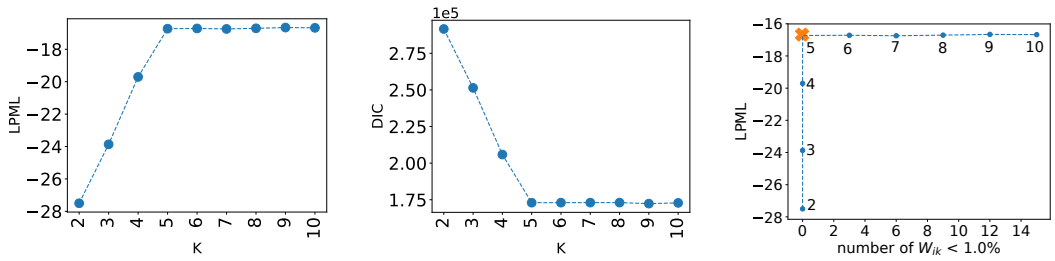
Table 2.1: Design of Simulation 1. $\boldsymbol{Z}^{\mathrm{TR}}$ and $\boldsymbol{w}^{\mathrm{TR}}$ are illustrated in (a) and (b), respectively. $K^{\mathrm{TR}} = 5$, $J = 20$, and $I = 3$ are assumed. In (a), black represents $z_{j,k}^{\mathrm{TR}} = 1$ (marker expression) and white represents $z_{j,k}^{\mathrm{TR}} = 0$ (marker non-expression).

500, and 1000 cells, for $i = 1$, 2, and 3, respectively. We set the true number of latent features (subpopulations) to be $K^{\mathrm{TR}} = 5$ and specified a $J \times 5$ binary feature-allocation matrix $\boldsymbol{Z}^{\mathrm{TR}}$ and 5-dimensional vectors $\boldsymbol{w}_i^{\mathrm{TR}}$ as follows: We first simulated $\boldsymbol{Z}^{\mathrm{TR}}$ by setting $z_{j,k}^{\mathrm{TR}} = 1$ with probability 0.6. If any column or row in $\boldsymbol{Z}^{\mathrm{TR}}$ consisted of all 0's, the entire matrix was re-sampled. We then simulated $\boldsymbol{w}_i^{\mathrm{TR}}$ 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 $\boldsymbol{w}_i^{\mathrm{TR}}$ would be likely to contain both large and small values. The assumed $\boldsymbol{Z}^{\mathrm{TR}}$ and $\boldsymbol{w}_i^{\mathrm{TR}}$ are given in Table 2.1. We set $\epsilon_i^{\mathrm{TR}} = 5\%$ of the cells to be noisy for all i. We specified the mixture models for the expression levels by setting $\boldsymbol{\mu}_0^{\star,\mathrm{TR}} = (-1, -2.3, -3.5)$ and $\boldsymbol{\mu}_1^{\star,\mathrm{TR}} = (1, 2, 3)$ with $L^{0,\mathrm{TR}} = L^{1,\mathrm{TR}} = 3$, and simulating mixture weights $\boldsymbol{\eta}_{i,j}^{z,\mathrm{TR}}$ from a Dirichlet distribution with parameters a random permutation of $(1, \ldots, L^{z,\mathrm{TR}})$, for $z \in \{0, 1\}$ and each (i, j). The values of $\sigma_i^{2,\mathrm{TR}}$ were set to 0.2, 0.1, and 0.3 for samples 1, 2, and 3, respectively. We then simulated latent subpopulation indicators $\lambda_{i,n}^{\mathrm{TR}}$ with probabilities $\Pr(\lambda_{i,n}^{\mathrm{TR}} = 0) = \epsilon_i^{\mathrm{TR}}$ and $\Pr(\lambda_{i,n}^{\mathrm{TR}} = k \mid \lambda_{i,n}^{\mathrm{TR}} \neq 0) = w_{i,k}^{\mathrm{TR}}$. We generated $y_{i,n,j} \overset{iid}{\sim} \mathrm{Normal}(0, 9)$ for all (i, n, j) with $\lambda_{i,n}^{\mathrm{TR}} = 0$. Otherwise,

we generated $y_{i,n,j}$ from a mixture of normals, $\sum_{\ell=1}^{L^{z,\mathrm{TR}}} \eta_{i,j}^{z,\mathrm{TR}} \cdot \mathrm{Normal}(\mu_{z\ell}^{\star,\mathrm{TR}}, \sigma_i^{2,\mathrm{TR}})$ given $z_{j\lambda_{i,n}^{\mathrm{TR}}}^{\mathrm{TR}} = z$ 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 Uniform $\left(0, 0.7 \cdot \sum_k w_{i,k}^{\mathrm{TR}}(1 - z_{j,k}^{\mathrm{TR}})\right)$ and sampled $p_{i,j} \times N_i$ cells without replacement with probability proportional to $\{1 + \exp(-9.2 - 2.3y_{i,n,j})\}^{-1}$. 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 (2.5). Heatmaps of the simulated $\boldsymbol{y}$ are shown in Figure 2.3(b), (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 colors 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 $s_\epsilon^2 = 10$ for each K. We specified the remaining fixed hyper-parameters as follows: $a_\alpha = b_\alpha = 0.1$ for α; $\psi_z = 1$ and $\tau_z^2 = 1$ for $\delta_{z,\ell}$; $a_\sigma = 3$ and $b_\sigma = 2$ for σ_i^2; $a_{\eta^z} = 1$ for $\boldsymbol{\eta}_{i,j}$; $d = 1$ for $\boldsymbol{w}_i$; $a_\epsilon = 1$ and $b_\epsilon = 99$ for ϵ_i. We used the empirical approach described in § 2.2 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 $\boldsymbol{\eta}_{i,j}^z$, we applied density-based clustering via finite Gaussian mixture models using the MClust package (Scrucca et al. 2016), and used the resulting clustering of $y_{i,n,j}$. 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 diagnosed convergence and mixing of the posterior MCMC simulations using trace plots, and found no evidence of convergence problems. Posterior inference for a model with $K = 5$

took 38 minutes per 1000 iterations on an interactive Linux server with four Intel Xeon E5-4650 processors and 512 GB of random access memory.



(a) LPML (b) DIC (c) Calibration of K

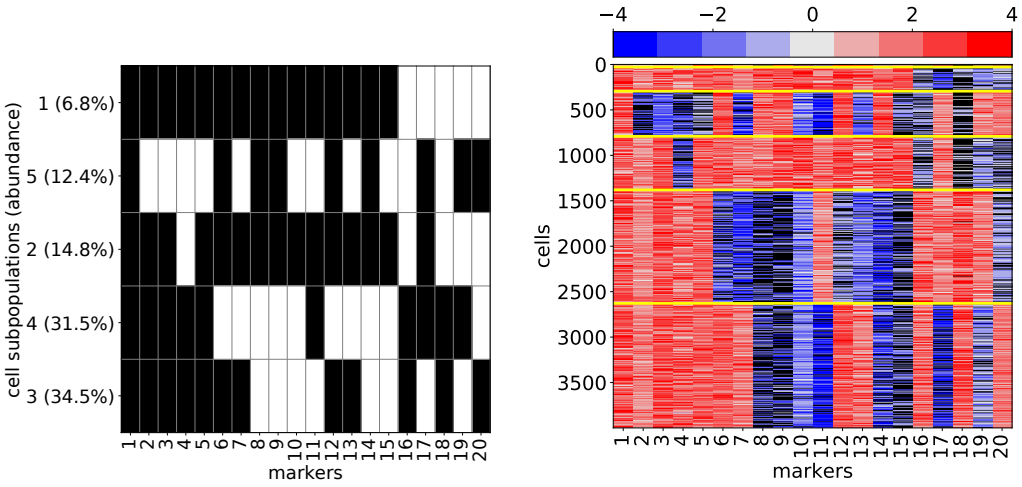
Figure 2.2: Results of Simulation 1. Plots of (a) LPML = log pseudo marginal likelihood, (b) DIC = deviance information criterion , and (c) calibration metric, for $K = 2, \ldots, 10$.

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 § 2.2.3. Figures 2.2(a) and (b) respectively show plots of LPML and DIC as functions of K. Figure 2.2(c) 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 2.3 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 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 Table 2.1 for all samples, implying that the true cell population structure is well recovered. 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 and 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

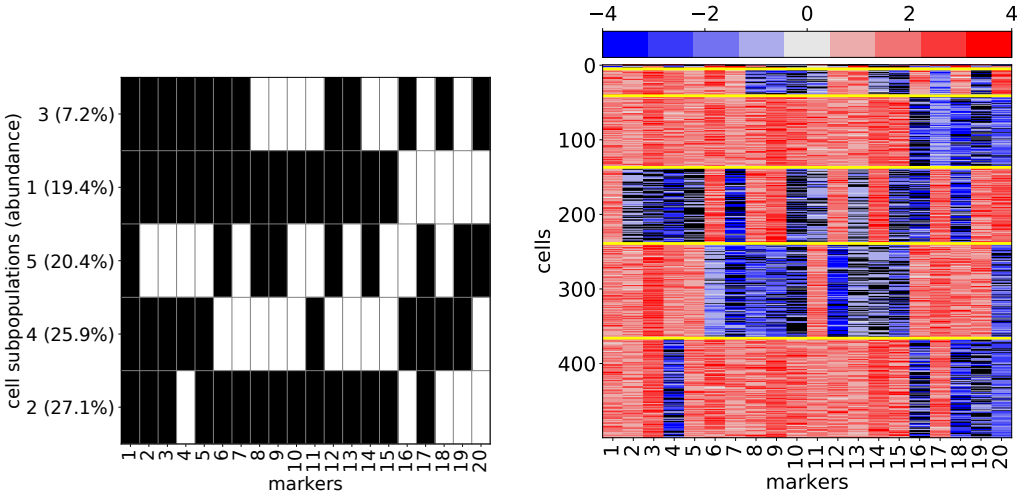
obtained ARIs are above 0.99 for all samples, indicating that the model recovers the true cell clusters very well. The heatmaps of y rearranged by cell clustering membership estimates $\hat{\lambda}_{i,n}$ are shown in panels (b), (d), and (f) of Figure 2.3, where the colors, red, blue, and black represent 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 2000, $K = 30$, and 20000 iterations. The time required to fit the model was approximately 18 minutes per 1000 iterations, which is substantially faster than that of the analogous MCMC method. Appendix Fig A.2 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 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 Appendix Table A.1. 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 Appendix Figures A.3 and A.4. 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. Appendix Table A.1 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 $\mu^{\star}_{0,L_0}$, 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 Appendix §A.4.



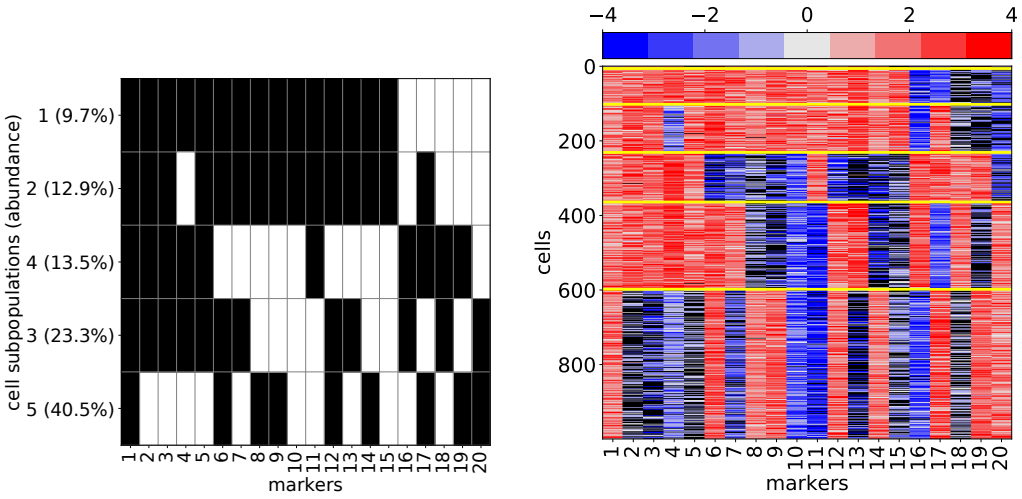
(a) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (b) heatmap of y_{1nj}



(c) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (d) heatmap of y_{2nj}

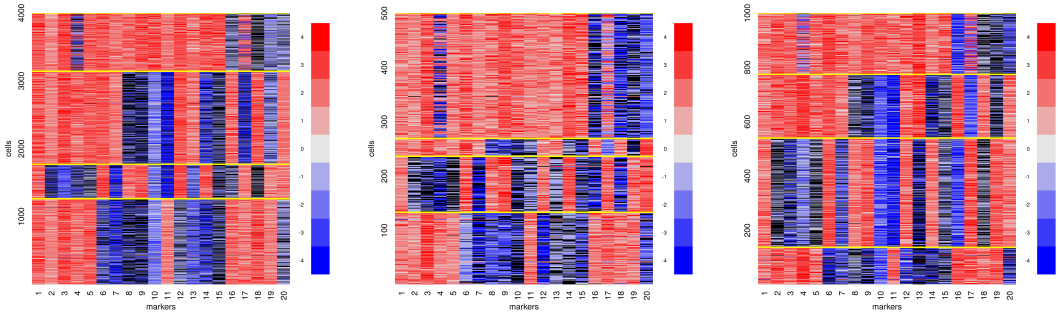
Figure 2.3: Results of Simulation 1. In (a) and (c), the transpose $\hat{\boldsymbol{Z}}_i'$ of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are shown for samples 1 and 2, respectively, with markers that are expressed denoted by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ 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.

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).



(e) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$ (f) heatmap of y_{3nj}

Figure 2.3 (continued): Results of Simulation 1 (continued). 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.



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

Figure 2.4: 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.

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 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 11 seconds on the simulated dataset. FlowSOM identified four cell clusters, as summarized in Figure 2.4, 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 2.2 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.

Method	Sample 1	Sample 2	Sample 3
FAM ($K = 5$)	0.999	0.993	0.999
FlowSOM	0.945	0.738	0.935

Table 2.2: Adjusted Rand index (ARI) for FAM and FlowSOM by sample for Simulation 1. Higher ARI is better, and values closer to 1 indicate that estimated clusters are closer to the truth.

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 $K^{\text{TR}} = 10$ and $N = (40000, 5000, 10000)$. $\boldsymbol{Z}^{\text{TR}}$ and $\boldsymbol{w}_i^{\text{TR}}$ are illustrated in Appendix Figure A.2. We considered ten models with $K = 2, 4, \cdots, 20$. 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 Appendix Figure A.6. 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 FlowSOM. Details of Simulation 2, including a sensitivity analysis for the data missingship mechanism and fast computation using ADVI, are given in Appendix § A.4.2.

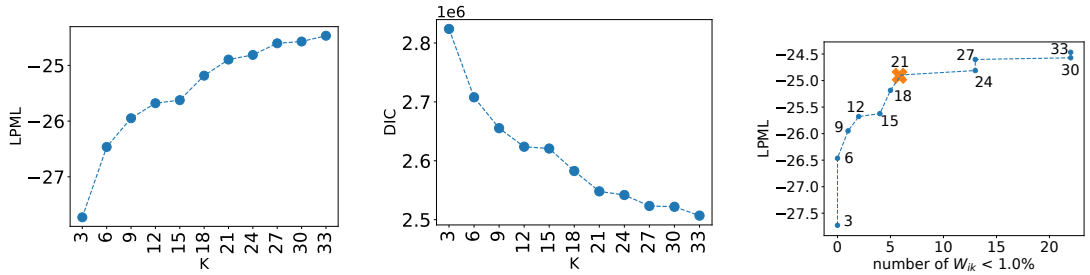
2.4 Analysis of Cord Blood Derived NK Cell Data

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 cutoff 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 § 2.2.1. We let $y_{i,n,j}$ =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. Appendix Table A.5 lists the markers included in the analysis. Figures 2.6(b), (d) and (e) show 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 Appendix A.5. We also plotted the data using the data visualization technique "t-SNE (t-Distributed Stochastic Neighbor Embedding)" in Appendix Figure A.11. t-SNE is a popular method for visualization of high-dimensional data in a two- or three-dimensional map through stochastic neighbor embedding (Maaten and 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. 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 § 2.3. Random parameters $\boldsymbol{\theta}$ also were initialized in a similar manner. 6000 samples from the posterior distribution of the model parameters were obtained after a burn-in of 10000 iterations. The posterior samples were thinned by selecting every other sample to yield a total of 3000 samples.

Figures 2.5 (a) 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 2.5 (c), our additional calibration method also indicates that the models with $K > 21$ include more cell subpopulations comprising less than one percent of a sample (i.e. $\sum_{i,k} \hat{w}_{i,k} < 1\%$ is larger), and improve fit only minimally.

Figure 2.6 summarizes posterior inference on the latent cell population structure with $\hat{K} = 21$. The cells are grouped by their estimated cell subpopulation

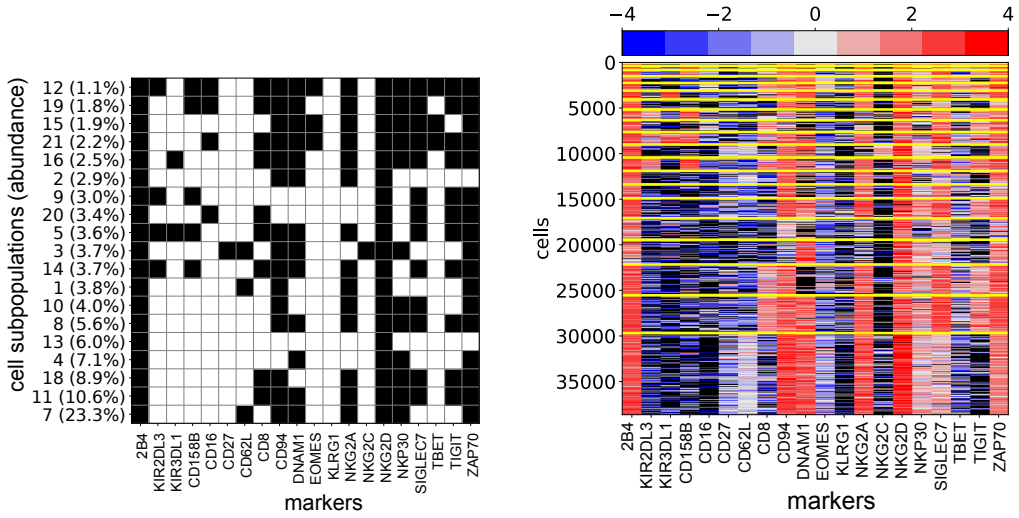


(a) LPML (b) DIC (c)

Figure 2.5: Analysis of UCB-derived NK cell data. Plots of (a) LPML, (b) DIC, and (c) calibration metric, for $K = 3, 6, \ldots, 33$.

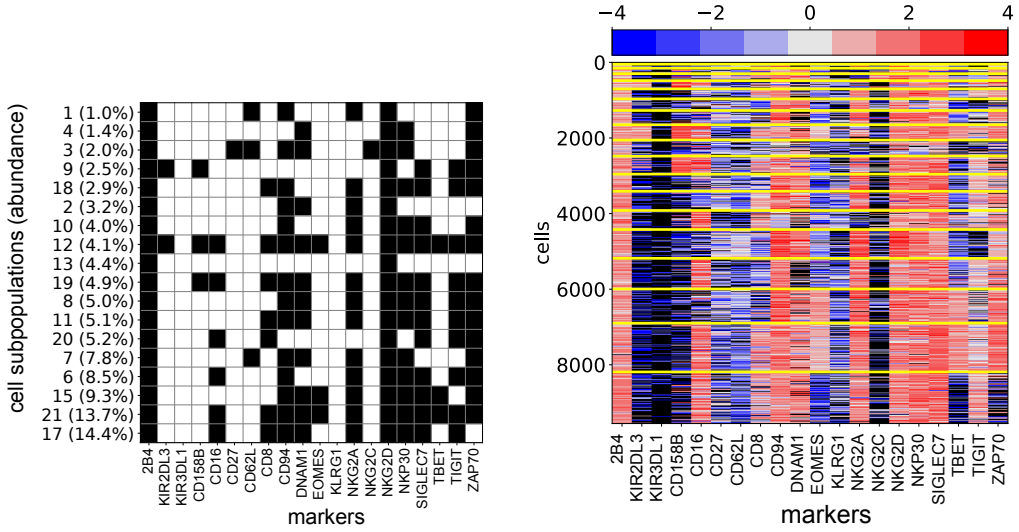
indicators $\hat{\lambda}_{i,n}$. The figure shows the estimated cell subpopulations $\hat{\boldsymbol{Z}}_i$ (in the left column) and clustered marker expression levels $\boldsymbol{y}_i$ (in the right column) for the samples. Cells having subpopulations with larger $\hat{w}_{i,k}$ are shown at the 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



(a) $\hat{\boldsymbol{Z}}_1'$ and $\hat{\boldsymbol{w}}_1$

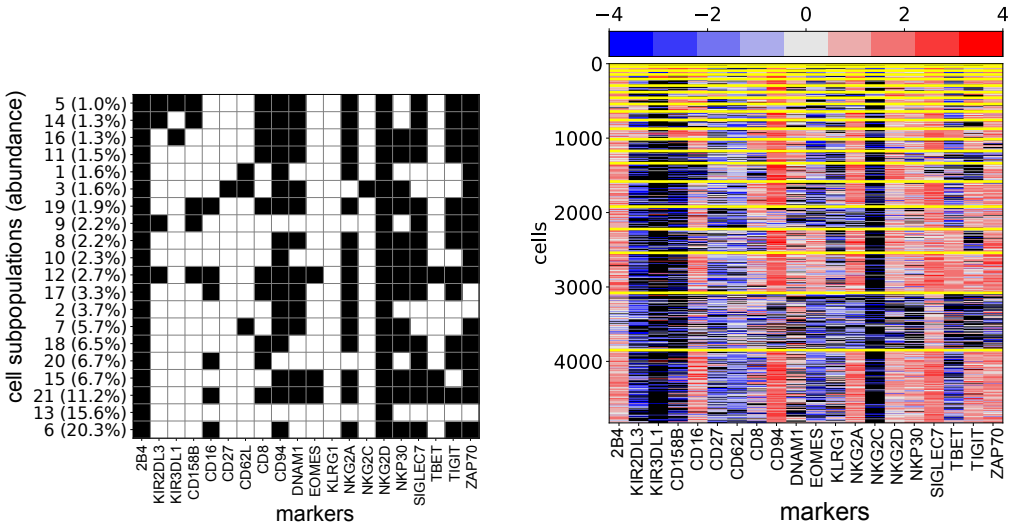
(b) Clustering of y_{1nj}



(c) $\hat{\boldsymbol{Z}}_2'$ and $\hat{\boldsymbol{w}}_2$

(d) Clustering of y_{2nj}

Figure 2.6: 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.



(e) $\hat{\boldsymbol{Z}}_3'$ and $\hat{\boldsymbol{w}}_3$ (f) Clustering of y_{3nj}

Figure 2.6 (continued): Analysis of the UCB-derived NK cell data (continued) $\hat{\boldsymbol{Z}}_i'$ 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.

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, Sarvaria et al. 2017), while a mature NK cell phenotype has been linked to superior cytotoxicity and better

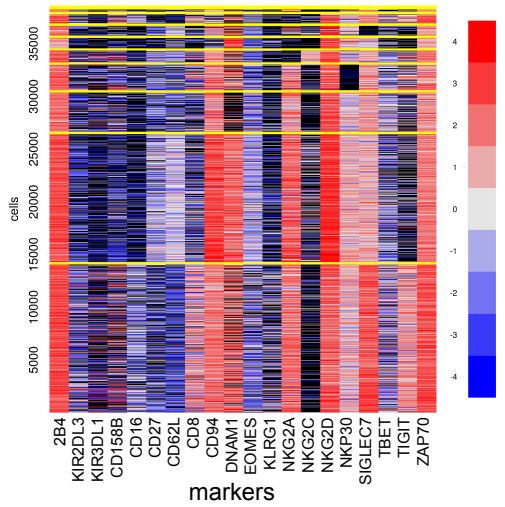
clinical outcomes in cancer patients (Ilander et al. 2017, Carlsten and Jaras 2019). These immature clusters of NK cells still retain expression of 2B4, CD94, NKG2A, NKG2D, and NKp30.

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.

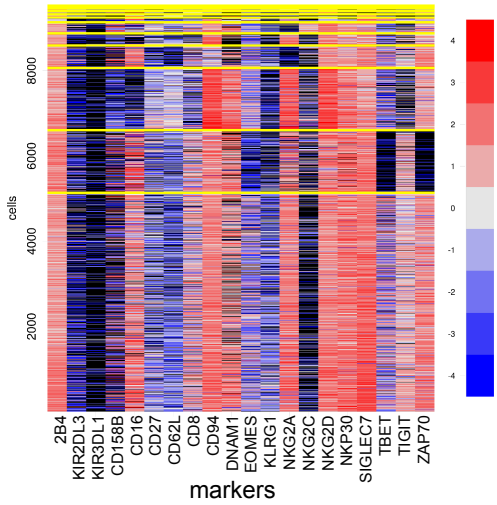
Model sensitivity to the specification of the data missingship mechanism in the NK cell data analysis was assessed by fitting the FAM under two additional specifications of $\boldsymbol{\beta}$, which we call data missingship mechanisms (MM) I and II. We will refer to the previous (default) missingship mechanism as MM-0. Appendix Tables A.6 and A.7 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 Appendix Figures A.12 and A.13. 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 2000 and $K = 30$ for 20000 iterations. The runtime was 6 hours on the previously described machine. Appendix Figure A.14 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.

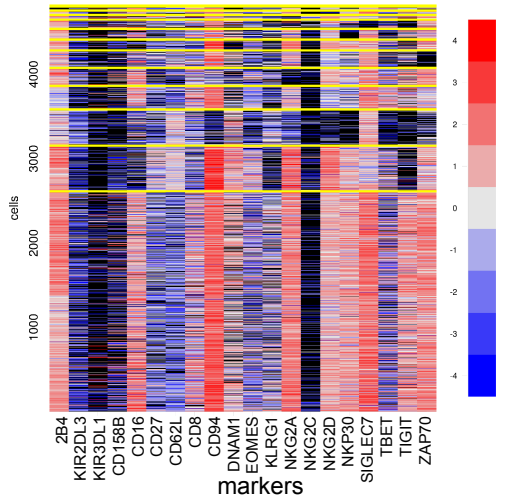
For comparison, we also applied FlowSOM to the UCB data. We fixed the missing values of $y_{i,n,j}$ at the minimum of the negative observed values of y for each (i, j) prior to analysis. FlowSOM identified 13 cell clusters in the samples. Heatmaps of $y_{i,n,j}$ rearranged by cell clustering estimates by FlowSOM are given in Figure 2.7 (a)-(c). Heterogeneity between cells within clusters estimated under FlowSOM is noticeably greater than that under the proposed FAM shown in Figure 2.6. For example, marker 10 shows a mix of red, blue, and black colors for cluster 1, the largest cluster. The proportions of cells assigned to the clusters are summarized in Figure 2.7(d). The clusters obtained by FlowSOM are much larger than those obtained by the FAM. In particular, cluster 1 under FlowSOM contains



(a) Clustering of y_{1nj}



(b) Clustering of y_{2nj}



(c) Clustering of y_{3nj}

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

(d) Proportions

Figure 2.7: [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).

36.7%, 53.8% and 54.1% of the cells in samples 1-3, respectively. Lastly, FlowSOM does not produce an explicit inference on the characterization of subpopulations.

2.5 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 and Xu (2020)). 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.

Chapter 3

A Bayesian Model for Identifying Distinct Features that Define Cell Subpopulations from Cytometry Data

3.1 Introduction

Feature allocation models (FAMs) have been used in many different settings to identify underlying latent structures, including a wide variety of biomedical applications given by Hai-son and Bar-Joseph (2011), Chen et al. (2013), Sengupta et al. (2014), Lee et al. (2015), Xu et al. (2016), Ni et al. (2019), Lui et al. (2020). In a FAM, a feature consists of a subset of experimental objects, such as biomarkers, and each object may belong to a finite number of features. A feature allocation is represented by a $J \times K$ binary matrix, $\boldsymbol{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_K]$, with the rows corresponding to objects and the columns to features. In $\boldsymbol{Z}$, the feature

membership indicator $z_{j,k} = 1$ if object j belongs to feature k, and $z_{j,k} = 0$ otherwise. Thus, the k^{th} column $\boldsymbol{z}_k$ of $\boldsymbol{Z}$ represents the set of objects having feature k. The FAM consists of independent priors $p(\boldsymbol{z}_1), \cdots, p(\boldsymbol{z}_K)$ on the features. Due to the assumption of independence, the objects are grouped into features without the restrictions that the features are either mutually exclusive or exhaustive subsets.

The assumptions underlying a conventional FAM may be at odds with the goal of obtaining features that have meaningful interpretations, however. In biomedical applications involving objects such as single nucleotide polymorphisms, messenger RNA (mRNA) strands, cell surface markers, or tumors, a major goal is to obtain features that can be interpreted as components of a meaningful biological structure. For example, Xu et al. (2016) applied a FAM to mRNA expression data consisting of tumor samples collected from breast cancer patients, with the goal to identify cancer subtypes. They assumed that a sample was composed of cells taking different latent cancer subtypes, and used features obtained from a FAM to represent the subtypes. In such applications, obtaining a parsimonious representation of the underlying structure, with fewer and more distinctive features, is critical for obtaining a biologically meaningful interpretation. A commonly assumed distribution for $\boldsymbol{Z}$, $p(\boldsymbol{Z})$, is the Indian buffet process (IBP) given by Griffiths and Ghahramani (2011), which can be defined as follows. Let $[v_k \mid \alpha] \sim \text{Beta}(\alpha/K, 1)$ and $z_{j,k} \mid v_k \overset{ind}{\sim} \text{Bernoulli}(v_k)$, for $j = 1, \ldots, J$, and $k = 1, \ldots, K$. By taking the limit as $K \to \infty$, and dropping any column having all entries 0, the remaining columns of $\boldsymbol{Z}$ can be rearranged into a left-ordered form (lof). A lof for $\boldsymbol{Z}$ is obtained by re-ordering the columns from left to right using the magnitudes of the binary number of units expressed in the columns. The resulting left-ordered $\boldsymbol{Z}$ has an IBP distribution. Under the IBP, the number of features K is random, and the columns of $\boldsymbol{Z}$ are independent. The IBP and its finite variants that fix

or place a prior on K instead of taking the limit are very commonly used as a prior distribution in Bayesian FAMs. Although these models are mathematically convenient, a potential drawback is that assuming independence across columns allows duplicated or very similar features, which make interpretation of the inferred features difficult.

In this chapter, we propose a repulsive-FAM (rep-FAM) that encourages distinct features *a priori* by parametrizing repulsion among features explicitly through a function, f_ϕ, incorporated into the prior, that controls the degree of dissimilarity between features. Under our rep-FAM, features are no longer independent, and are encouraged to be different. The shape of f_ϕ is determined by hyperparameters ϕ. We study how ϕ changes the shape of f_ϕ, and present a method for calibrating f_ϕ. Various approaches for using a repulsion function to model locations in a mixture model have been explored in the context of density estimation and clustering problems. For example, see Petralia et al. (2012), Quinlan et al. (2017, 2018), Xie and Xu (2020). Those papers showed that a repulsion function encourages well-separated mixture components, and that a mixture model with a repulsion function produces a parsimonious representation of the underlying structure that can be more scientifically meaningful. It can also yield better model fit and improved clustering. Other approaches for inducing dependence in FAMs have been proposed. For example, the distance-dependent IBP (Gershman et al. 2014) accounts for dissimilarities between rows, with experimental units tending to possess similar sets of features if their covariate vectors are close to each other. Williamson et al. (2010, 2019) developed a dependent IBP in which the elements of $\boldsymbol{Z}$ are allowed to persist or vary over time.

We build upon the FAM of Lui et al. (2020), and apply the proposed rep-FAM to construct a model that infers a latent population structure of cells in samples

using cytometry by time-of-flight (CyTOF) data.

CyTOF measures expression levels of cell surface biomarkers for individual cells in the data. Natural killer (NK) cells serve a critical role in cancer immune surveillance, and they have the intrinsic ability to infiltrate cancer tissues. A discussion of NK cell therapy for hematologic malignancies is given by Lui et al. (2020). CyTOF cell surface marker data obtained from one blood sample, typically from either patients or umbilical cord blood, are recorded for thousands of cells. High numerical values for a particular marker indicate high expression levels of the marker, and low numerical values indicate low expression levels. While marker expression levels obtained via CyTOF are non-negative real numbers, investigators who desire to identify subpopulations of cells in a sample often characterize expression using latent binary variables. This is done by defining marker expression vectors in each subpopulation, with entries 1 or 0 for each marker indicating expression or non-expression.

As done with a conventional FAM, we will characterize cell subpopulations using latent features that represent subpopulation-specific cell surface marker expression patterns. Using the posterior distributions of the features, the model clusters individual cells from one or more samples into subpopulations based on their markers' expression patterns. The rep-FAM obtains subpopulations that are more likely to be dissimilar. We fix K, but let each sample have its own set of features and abundance levels over the selected features. Thus, the k^{th} subpopulation, characterized by $\mathbf{z}_k$, may be present in only some of the samples, and its feature abundance levels may differ between samples. Unlike Lui et al. (2020), where a finite IBP with fixed K is used, the proposed model obtains a parsimonious and more clinically interpretable summary of the cell subpopulations. Also, the rep-FAM performs feature selection and enables inferences more tailored for

each of the samples. We conduct simulation studies and real data analyses and compare the rep-FAM to a FAM that assumes independent features (ind-FAM), and also to usual methods that cluster cells based on expression levels.

The remainder of this chapter proceeds as follows. In Section 3.2 we introduce the cytometry at time-of-flight (CyTOF) application and present our rep-FAM. Then, in Section 3.3, we compare the performance of the rep-FAM to those of the ind-FAM, and some existing clustering methods. Section 3.4 presents results of CyTOF data obtained from patients 30 days after cell infusion. We offer some concluding remarks in Section 3.5.

3.2 Probability Model

3.2.1 Repulsive Feature Allocation Model

In general, consider a dataset with I samples indexed $i = 1, \ldots, I$, and cells in the i^{th} sample indexed by $n = 1, \ldots, N_i$. Raw expression level measurements for marker $j = 1, \cdots, J$ on cell n in sample i are denoted by $\tilde{y}_{i,n,j}$. Let $c_{i,j}$ represent a threshold, calculated via a flow cytometry algorithm called *flowDensity* (Malek et al. 2014), which is used to normalize the raw measurements for marker j in sample i. These thresholds account for instrument fluctuations or signal crosstalk among channels designed for different markers. Though crude, these thresholds are needed as a starting point for our analyses. Let the normalized expression levels be $y_{i,n,j} = \log\left(\tilde{y}_{i,n,j}/c_{i,j}\right)$, so that raw measurements that are above (below) their respective thresholds yield positive (negative) normalized expression levels. In CyTOF analyses, marker expression levels are recorded as 0 due to experimental artifacts and weak levels of expression. When this is the case, the normalized values will be undefined. Thus, we treat zeros as missing values and impute them

to facilitate posterior computation.

We propose a rep-FAM with a prior on $\boldsymbol{Z}$ formulated to favor features that are more distinct from each other, as follows. Let $f_\phi(d)$ denote a repulsion function increasing in a distance metric, $d = d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2})$ that quantifies the dissimilarity between columns k_1 and k_2 of $\boldsymbol{Z}$, and takes on non-negative values, with $d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) = 0$ when $\mathbf{z}_{k_1} = \mathbf{z}_{k_2}$. We require that the repulsion function has the properties $f_\phi(0) = 0$ and $\lim_{d\to\infty} f_\phi(d) = 1$, where ϕ is a hyperparameter. An example of a function with these properties is $f_\phi(d) = \{1 - \exp(-\phi_1 d)\}^{\phi_2}$, with hyperparameters ϕ_1, $\phi_2 > 0$ that control the shape of f_ϕ. The function $f_\phi(d)$ will be incorporated into the prior to put probability mass on the columns of $\boldsymbol{Z}$ in such a way that more distinct columns are more likely. For our rep-FAM, we assume that the feature membership matrix $\boldsymbol{Z}$ follows the prior

$$p(\boldsymbol{Z} \mid \boldsymbol{v}, f_\phi) \propto \left\{ \prod_{k=1}^{K} \prod_{j=1}^{J} v_k^{z_{j,k}} (1 - v_k)^{1 - z_{j,k}} \right\} \times \left\{ \prod_{k_1=1}^{K-1} \prod_{k_2=k_1+1}^{K} f_\phi\left(d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2})\right) \right\}, \quad (3.1)$$

where $v_k \in (0, 1)$ for $k \in \{1, \ldots, K\}$. As in the IBP, we assume the independent level 2 priors $[v_k \mid \alpha] \overset{iid}{\sim} \text{Beta}(\alpha/K, 1)$, for $k = 1, \ldots, K$, and assume the level 3 prior $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$, which has mean a_α/b_α. For a distance metric, we use the L_1 norm, $d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) = \sum_{j=1}^{J} |z_{j,k_1} - z_{j,k_2}|$, so for $\boldsymbol{Z}$ with binary entries $d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2})$ is the number of discordances between columns $\mathbf{z}_{k_1}$ and $\mathbf{z}_{k_2}$, and it takes on values in $\{0, 1, \cdots, J\}$. If $\boldsymbol{Z}$ has any duplicated columns then $p(\boldsymbol{Z} \mid \boldsymbol{v}, f_\phi) = 0$, and $f_\phi(\cdot)$ smoothly penalizes $\boldsymbol{Z}$ having similar columns by making it less likely *a priori*. Figure 3.1 illustrates $f_\phi(d)$ as a function of d for different values of (ϕ_1, ϕ_2). In panel (a), we fix $\phi_1 = 1$ and vary the value of ϕ_2. Note that as $\phi_1 \to \infty$ or when $\phi_2 = 0$, $f_\phi(d) = 1$ for any $d \geq 0$, and the regular FAM is recovered. Figure 3.1(b) illustrates $f_\phi(d)$ as a function of ϕ_1 and ϕ_2, with d fixed at 1. For fixed (d, ϕ_2), as ϕ_1 increases, $f_\phi(d)$ increases and $p(\boldsymbol{Z})$ penalizes $\boldsymbol{Z}$ with similar features less.

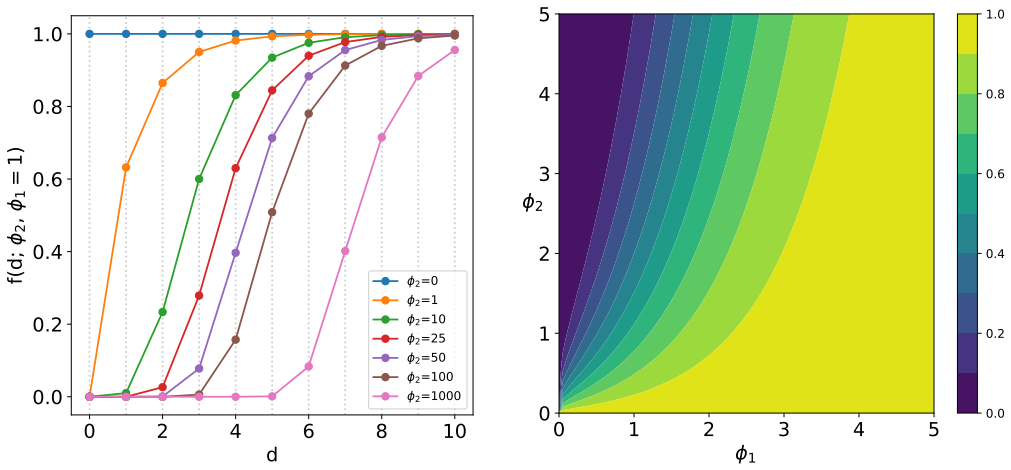


Figure 3.1: (a) Illustration of a repulsion function $f_\phi(d) = \{1 - \exp(-\phi_1 d)\}^{\phi_2}$ with $\phi_1 = 1$, $\phi_2 \in \{0, 1, 10, 25, 50, 100, 1000\}$. (b) Heatmap of $f(d; \phi_1, \phi_2)$ with $\phi_1 \in (0, 5)$, $\phi_2 \in (0, 5)$, and $d = 1$.

As ϕ_2 increases, $f_\phi(d)$ decreases and $p(\boldsymbol{Z})$ penalizes $\boldsymbol{Z}$ with similar features more heavily. Under the rep-FAM given in (3.1), the support of p_Z has ${}_{2^J}P_K$ values for any $J > 0$ and $K \leq 2^J$, and the normalizing constant is finite. Placing hyper-priors on $\boldsymbol{\phi} = (\phi_1, \phi_2)$ is possible, but it complicates posterior computation since the normalizing constant in (3.1) depends on $\boldsymbol{\phi}$, making posterior inference for $\boldsymbol{\phi}$ doubly intractable. Instead, we calibrate $\boldsymbol{\phi}$ using an *apriori* chosen separation level between features, similarly to the calibration method used by Petralia et al. (2012). We use the prior expected number of marker expression discordances between features to calibrate $\boldsymbol{\phi}$. For the simulation studies in § 3.3 and the data analysis in § 3.4, we impose the constraint $\Pr(\min_{1 \leq k_1 < k_2 \leq K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \geq \underline{d} \mid \boldsymbol{\phi}) > \underline{p}$, where $\underline{d}$ is a lower threshold for the minimum difference between a pair of features and $\underline{p}$ is a fixed lower probability cut-off. Given pre-specified values of $\underline{d}$ and $\underline{p}$, we solve numerically for values of $\boldsymbol{\phi}$. Additional details of the calibration of $\boldsymbol{\phi}$ are given in Appendix B.1. When no prior information is available, one may choose $\boldsymbol{\phi}$ using a model comparison statistic, such as the deviance information criterion

(DIC). For alternative choices of d and f_ϕ, see Petralia et al. (2012), Quinlan et al. (2017), Xie and Xu (2020). Since the IBP of (Ghahramani and Griffiths 2006) assumes independence across columns, it has a positive probability of allowing some columns to appear more than once in $\boldsymbol{Z}$. A key difference between the rep-FAM in (3.1) and the IBP is that the rep-FAM ensures that all features are distinct. This enhances interpretability, which is especially critical in our application since the features represent distinct subpopulations defined by unique subsets of expressed cell surface markers.

A possible alternative to our rep-FAM is the determinantal point process (DPP)-based FAM, which is a repulsive FAM given by Xu et al. (2016). Like our proposed rep-FAM, the DPP-based FAM defines a probability distribution over a $J \times K$ binary matrix, $\boldsymbol{Z}$. Xu et al. (2016) construct a $K \times K$ symmetric kernel matrix, C, having elements in [0, 1], quantifying closeness of feature pairs, with $C(k_1, k_2) = 1$ if the features (k_1, k_2) are identical. The prior $p(\boldsymbol{Z})$ is defined to be proportional to $\det(C)$. In the extreme case where $\boldsymbol{Z}$ has some identical columns, $\det(C) = 0$, so the DDP-based FAM prior $p(\boldsymbol{Z}) = 0$. More distinct feature pairs have off-diagonal elements of C close to 0, and make $\det(C)$ closer to 1. Thus, both the DDP-based FAM and our rep-FAM do not allow $\boldsymbol{Z}$ to have identical columns. However, because the DPP-based FAM first constructs C and uses $\det(C)$ to define $p(Z)$, it may be difficult to understand how repulsion is reflected in $p(\boldsymbol{Z})$, and posterior computation also is more difficult. In contrast, our rep-FAM explicitly parameterizes repulsion between features via the penalty function $f_\phi(d)$ and incorporates this directly into the prior, so it is straightforward to understand how $f_\phi(d)$ affects $p(\boldsymbol{Z})$.

The assumption that all features in $\boldsymbol{Z}$ may be present in all samples may be overly restrictive when each sample is composed of two or more subpopu-

lations. In our application, and in many other settings, a subpopulation may be present in one sample but not another. We thus construct the probability model to facilitate feature selection by identifying sample-specific feature configurations from the data. To do this, we introduce latent feature selection indicators, $\{r_{i,k}, i = 1, \ldots, I,\ k = 1, \ldots, K\}$, with $r_{i,k} = 1$ if sample i includes feature k, and $r_{i,k} = 0$ if not, and denote $\boldsymbol{r}_i = (r_{i,1}, \cdots, r_{i,K})$. We assume that $[r_{i,k} \mid p_i] \overset{ind}{\sim} \text{Bernoulli}(p_i)$ with each p_i fixed. Sample-specific feature selection using $r_{i,k}$ facilitates a joint analysis of multiple samples, possibly obtained from different sources. We denote the set of features selected in sample i by $R_i = \{k : r_{i,k} = 1, \text{ for } k = 1, \ldots, K\}$, which is the feature subpopulation for that cell sample. Because the latent feature selection indicators are random, each R_i and its cardinality, $|R_i|$, also are random. This implies that some features may not be chosen by any sample, in which case $\cup_{i=1}^{I} R_i$ is a proper subset of $\{1, \ldots, K\}$. We fix K to be a reasonably large value and, through feature selection, samples are allowed to select a subset of the K features, so the numbers of selected features, $|R_1|, \cdots, |R_I|$, vary randomly across samples.

3.2.2 Clustering by Latent Features

Recall that expression levels of J markers are recorded on each cell. The repFAM model sample i into $|R_i|$ cell subpopulations, with the k^{th} subpopulation characterized by the features in R_i. We next extend the probability model by introducing cluster membership indicators, $\lambda_{i,n} \in R_i$ for cell n in sample i. The event $(\lambda_{i,n} = k)$ implies that the marker expression pattern of cell n in sample i is the same as the pattern described by $\mathbf{z}_k$, so cells are clustered into subpopulations according to their marker expression patterns. This induces more biologically meaningful clustering of cells than cell clusterings based on their marker expression

levels alone (Lui et al. 2020). We let $p(\lambda_{i,n} = k \mid \boldsymbol{w}_i) = w_{i,k}$, for $k = 1, \ldots, K$, where $w_{i,k} = 0$ for $k \notin R_i$, $w_{i,k} > 0$ for $k \in R_i$, and $\sum_{k=1}^{K} w_{i,k} = 1$. We construct a Dirichlet distribution for the vector $\boldsymbol{w}_i = (w_{i,1}, \cdots, w_{i,K})$ using the random parameters $\boldsymbol{w}_i^\star$ and $\boldsymbol{r}_i$. We first assume $w_{i,k}^\star \overset{iid}{\sim} \text{Gamma}(a_w, 1)$, and define $w_{i,k} = (w_{i,k}^\star r_{i,k}) / \sum_{h=1}^{K} (w_{i,h}^\star r_{i,h})$, which implies that, $\boldsymbol{w}_i \mid \boldsymbol{r}_i \sim \text{Dirichlet}\ (a_w \boldsymbol{r}_i)$.

We assume conditional independence between the n^{th} cell's J marker values given $\lambda_{i,n}$, and we relate the distribution of $y_{i,n,j}$ to the marker expression configuration of cluster $\lambda_{i,n}$, i.e., $\mathbf{z}_{\lambda_{i,n}}$, as follows. Assume that each $y_{i,n,j}$ follows a mixture of normal distributions,

$$\left[y_{i,n,j} \mid z_{j,\lambda_{i,n}} = z, \boldsymbol{\mu}_z^\star, \boldsymbol{\eta}_{i,j}^z, \sigma_i^2\right] \overset{ind}{\sim} F_{i,j}^z = \sum_{\ell=1}^{L_z} \eta_{i,j,\ell}^z \cdot \text{Normal}(\mu_{z,\ell}^\star, \sigma_i^2), z \in \{0, 1\}, \quad (3.2)$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Recall that, if marker j is (is not) expressed in subpopulation k, with $z_{j,\lambda_{i,n}} = 1$ (0), then $y_{i,n,j}$ tends to be positive (negative). When a marker is not expressed, i.e., $z_{j,\lambda_{i,n}} = 0$, we let the mixture locations take negative values, $\mu_{0,\ell}^\star < 0$ with ordering constraint $0 > \mu_{0,1}^\star > \cdots > \mu_{0,L_0}^\star$ to avoid potential identifiability problems; we assume $\mu_{0,\ell}^\star = -\sum_{r=1}^{\ell} \delta_{0,r}$, where $\delta_{0,\ell} \overset{iid}{\sim} \text{TN}^+(\psi_0, s_0^2)$, $\ell = 1, \ldots, L_0$ with fixed L_0. Here $\text{TN}^+(\psi_0, s_0^2)$ denotes the normal distribution with mean ψ_0 and variance s_0^2 truncated below at zero. Similarly, we assume that the mixture locations take positive values, $\mu_{1,\ell}^\star > 0$, if the marker is expressed, $z_{j,\lambda_{i,n}} = 1$. We assume $\delta_{1,\ell} \overset{iid}{\sim} \text{TN}^+(\psi_1, s_1^2)$, $\ell = 1, \ldots, L_1$ with fixed L_1, and let $\mu_{1,\ell}^\star = \sum_{r=1}^{\ell} \delta_{1,r}$, so we have $0 < \mu_{1,1}^\star < \cdots < \mu_{1,L_1}^\star$. Finally, in (3.2) we let $\boldsymbol{\eta}_{i,j}^z \overset{iid}{\sim} \text{Dirichlet}_{L_z}(a_\eta)$, for $z = 0, 1$, $i = 1, \ldots, I$, and $j = 1, \ldots, J$, and $\sigma_i^2 \overset{iid}{\sim} \text{InverseGamma}(a_\sigma, b_\sigma)$, where $\text{Dirichlet}_L(a)$ denotes a Dirichlet distribution with L entries that all equal a. That is, we assume equal weights for the mixture components *a priori*. The mixture model in (3.2) is very flexible, and encompasses a wide class of distributions,

which may be multi-modal or skewed. The mixture locations $\boldsymbol{\mu}_z^\star$ are shared by all (i, j), and the model facilitates borrowing information across markers and samples. Thus, even when a marker is mostly expressed or not expressed, the model can obtain reliable estimates of $F_{i,j}^z$ for both cases of $z = 0$ and 1.

Expression levels that are missing are imputed using the same technique outlined in Lui et al. (2020). Subject experts provided information that expression levels of a marker are likely to be missing when the marker is not expressed. We build a joint model of expression levels and missingness indicators, and impute missing values of y during posterior simulation. Our approach is to craft a calibrated static missingship mechanism with a quadratic trend based on the experts' information, in order to limit imputed data to a range of negative expression levels consistent with the data. Our subpopulations are constructed by a pattern of expression or non-expression of the markers, but not directly by their expression level values, and our model is robust to the specification of the missingness mechanism. Appendix B.2.1 demonstrates how to specify a missing data mechanism for this model.

3.2.3 Posterior Computation

Posteriors for the rep-FAM can be computed via Gibbs sampling and other standard Markov chain Monte Carlo (MCMC) algorithms. Denote the vector of all random parameters and missing values by $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}, \boldsymbol{w}^\star, \boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{Z}, \boldsymbol{v}, \alpha, \boldsymbol{y}^{\text{miss}})$, where $\boldsymbol{\delta} = \{\delta_{z,\ell}, z = 0, 1, \text{ and } \ell = 1, \ldots, L_z\}$, $\boldsymbol{\sigma}^2 = \{\sigma_i^2, i = 1, \ldots, I\}$, $\boldsymbol{w}^\star = \{w_{i,k}^\star, i = 1, \ldots, I, \text{ and } k = 1, \ldots, K\}$, $\boldsymbol{r} = \{r_{i,k}, i = 1, \ldots, I, \text{ and } k = 1, \ldots, K\}$, $\boldsymbol{\lambda} = \{\lambda_{i,n}, i = 1, \ldots, I, \text{ and } n = 1, \ldots, N_i\}$, $\boldsymbol{v} = \{v_k, k = 1, \ldots, K\}$, and $\boldsymbol{y}^{\text{miss}}$ denotes missing values. For the priors described in § 3.2.1 and § 3.2.2 that are conjugate, their parameters can be updated easily. Due to the complexity of the

model and the size of the dataset, however, posterior simulation is computationally expensive and the Markov chain may converge slowly. Updating $\boldsymbol{Z}$ can be very slow, and involves several computational difficulties.

To make our algorithm more scalable, we exploit the idea of the intrinsic Bayes factor given by Berger and Pericchi (1996) and use a "minimally trained" prior, $p^{\star}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}' \mid \boldsymbol{\theta})$, where $\boldsymbol{y}'$ is a small subsample of the data, to replace the prior $p(\boldsymbol{\theta})$ and to generate proposals of $\boldsymbol{Z}$ in a Metropolis step. Because Bayes factors cannot be computed with improper priors, Berger and Pericchi (1996) first partitioned the data into a small, randomly chosen subset $\boldsymbol{y}'_{(1)}$ and its complement, $\boldsymbol{y}''_{(1)}$, computed $p^{\star}_1(\boldsymbol{\theta})$ using $\boldsymbol{y}'_{(1)}$, and computed the posterior $p^{\star}_{(1)}(\boldsymbol{\theta} \mid \boldsymbol{y}''_{(1)})$. To avoid dependence on the particular partition, they repeated this process for a sequence of partitions $\{\boldsymbol{y}'_{(h)}, \boldsymbol{y}''_{(h)}\}$ for $h = 1, \cdots, H$, then for each model being considered they computed the Bayes Factors, $\mathrm{BF}_1(\text{model}), \cdots, \mathrm{BF}_H(\text{model})$ and then averaged.

In the context of our posterior computations, we used this idea to facilitate updating $\boldsymbol{Z}$ in the MCMC. We did this by generating a prior $p^{\star}_{(h)}(\boldsymbol{\theta})$ from the h^{th} partition to replace the prior $p(\boldsymbol{\theta})$ and as a proposal distribution of $\boldsymbol{Z}$ in the MCMC. We updated $\boldsymbol{Z}$ by accepting $\boldsymbol{Z}'$ with probability $\min(1, \zeta)$, where $\zeta = p(\boldsymbol{y}''_{(h)} \mid \boldsymbol{Z}', \boldsymbol{\theta}_{-\boldsymbol{Z}})/p(\boldsymbol{y}''_{(h)} \mid \boldsymbol{Z}, \boldsymbol{\theta}_{-\boldsymbol{Z}})$ after some algebraic cancellations. Here, $\boldsymbol{\theta}_{-\boldsymbol{Z}}$ denotes all the subvector of $\boldsymbol{\theta}$ obtained by deleting $\boldsymbol{Z}$. This produces reasonable proposals of $\boldsymbol{Z}$, simplifies evaluation of the acceptance probability, and greatly speeds up the update for $\boldsymbol{Z}$. Because the posterior given the remainder of the data $p(\boldsymbol{\theta} \mid \boldsymbol{y}''_{(h)}) \propto p^{\star}_{(h)}(\boldsymbol{\theta})p(\boldsymbol{y}''_{(h)} \mid \boldsymbol{\theta}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})$, it follows that our sample is from the posterior based on the full data. For simulation studies and the CyTOF data analysis, for each partition, we used randomly chosen subsets comprising 10% of the data for the simulations and 5% for the Cytof data analysis as the training samples. We also modified this process by running the H Markov chains

in sequence, and using the final draw of $\boldsymbol{\theta}$ of the h^{th} chain as the initial value for the $(h+1)^{\text{th}}$ chain. We then pooled the H posterior samples of $\boldsymbol{\theta}$ obtained from $p(\boldsymbol{\theta} \mid \boldsymbol{y}^{''}_{(h)}), h = 1, \cdots, H$, to obtain a final posterior for inferences.

A similar approach of using the fractional Bayes factor is used in Lee et al. (2015, 2016). Our approach is different from mini-batch tempered MCMC (Li and Wong 2017), which uses a subset of the data to compute a tempered likelihood. Our method is also different from Chen et al. (2016b), where the Metropolis acceptance probability is computed based on a small subset of the data for computational efficiency; whereas we generate proposals from $p^{\star}_{(h)}(\boldsymbol{Z})$ and use the test set $\boldsymbol{y}^{''}_{(h)}$ to evaluate the Metropolis acceptance probability.

In addition, we improve mixing by using weight-preserving parallel tempering (WPPT) (Tawn et al. 2020). Parallel tempering (PT) (Earl and Deem 2005) is a general MCMC technique to increase the mixing rate of a model that suffers from poor mixing. For PT, multiple MCMC chains are run at various "temperatures" and updated in parallel for a given model. At regular iteration intervals, the entire states of pairs of chains are swapped with a positive probability, and different modes of the target distribution can be explored. Notably, we use a weight-stabilizing tempering scheme developed in Tawn et al. (2020). Tawn et al. (2020) showed that when tempering (simulated or parallel) is applied to mixture models, mixture components originally with smaller weights can have dominantly large mixture weights at high temperatures, and preserving the component weights can lead to inferences that are more sensible. Following their approach, we preserve the original weights $\eta^{z}_{i,j,\ell}$ of the mixture model for $\boldsymbol{y}$ by tempering only the kernels of the mixture components. Appendix B.2 provides details of PT, minimally trained priors, and posterior simulation.

As is common in most mixture models, $\boldsymbol{Z}$ suffers from label-switching issues

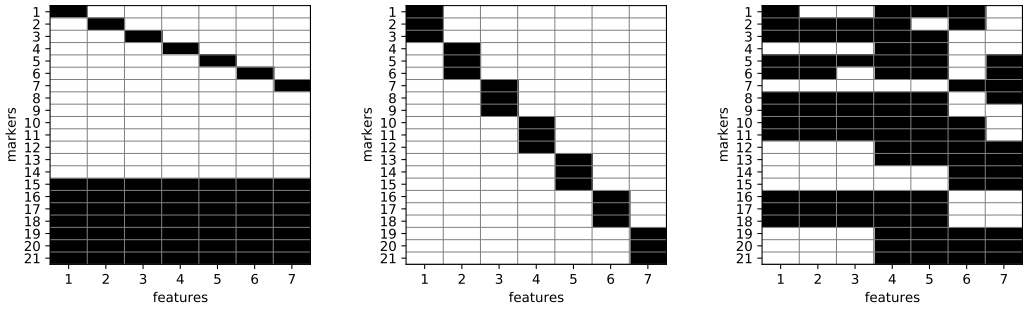
(Celeux et al. 2000, Stephens 2000, Jasra et al. 2005, Frühwirth-Schnatter 2006), and summarizing the posterior inference is challenging. Similar to the approach in Lui et al. (2020), we use a sequentially-allocated optimization (SALSO) method (Dahl and Müller 2017) with some modification to provide meaningful posterior point estimates of $\boldsymbol{\theta}$, especially $\boldsymbol{Z}$, $\boldsymbol{w}$ and $\boldsymbol{\lambda}$. For each sample i, we find point estimates, $\hat{\boldsymbol{Z}}_i$, $\hat{\boldsymbol{w}}_i$, and $\hat{\lambda}_{i,n}$ as follows; we first construct a $J \times J$ pairwise allocation matrix of $\boldsymbol{Z}$ for sample i, $A(\boldsymbol{Z})_i$ whose elements are $A(\boldsymbol{Z})_{i,(j,j')} = \sum_{k=1}^{K} w_{i,k} \times 1(z_{j,k} = 1) \times 1(z_{j',k} = 1)$, for $1 \leq j, j' \leq J$, i.e., the number of the features that markers j and j' have in common, weighted by $w_{i,k}$. Since $w_{i,k}$ can be exactly zero for some features, $A(\boldsymbol{Z})_i$ only accounts for selected features. We then find a point estimate $\hat{\boldsymbol{Z}}_i$ that minimizes the sum-of-squared differences

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

where $\bar{A}_{i,(j,j')}$ is the posterior mean of $A_{i,(j,j')}$ over $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. Since we compute $\hat{\boldsymbol{Z}}_i$ for each sample separately, inferred features can vary significantly between different samples. We use posterior Monte Carlo samples to obtain posterior point estimates $\hat{\boldsymbol{Z}}_i$ as

$$\hat{\boldsymbol{Z}}_i = \arg\min_{\boldsymbol{Z}} \int D(A(Z)_i, \bar{A}_i) dp(\boldsymbol{Z}, \boldsymbol{w}_i \mid \boldsymbol{y}) \approx \arg\min_{\boldsymbol{Z}^{(b)}} D(A(Z^{(b)})_i, \bar{A}_i),$$

for posterior samples $\{\boldsymbol{Z}^{(b)}, \boldsymbol{w}_i^{(b)}, b = 1, \ldots, B\}$. We report posterior point estimates, $\hat{\boldsymbol{w}}_i$ and $\hat{\lambda}_{i,n}$ conditional on $\hat{\boldsymbol{Z}}_i$. Additional details for implementing posterior inference are in Appendix B.2.



(a) $\boldsymbol{Z}^{1,\mathrm{TR}}$ (b) $\boldsymbol{Z}^{2,\mathrm{TR}}$ (c) $\boldsymbol{Z}^{3,\mathrm{TR}}$

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
Sample 1	0.04	0.05	0.39	0.00	0.06	0.43	0.03
Sample 2	0.00	0.05	0.54	0.20	0.14	0.07	0.00

(d) $\boldsymbol{w}_i^{\mathrm{TR}}$

Figure 3.2: Simulation truth: The true $\boldsymbol{Z}$ under three simulation scenarios are in (a)-(c). Each $\boldsymbol{Z}$ has $J = 21$ rows (markers) and $K = 7$ features (subpopulations). The true proportions of clusters $\boldsymbol{w}_i^{\mathrm{TR}}$ are in (d). The same $\boldsymbol{w}_i^{\mathrm{TR}}$ is used for all three scenarios.

3.3 Simulation Study

In this section, we examine the performance of the rep-FAM in § 3.2 through extensive simulation studies, and compare the model to a FAM that assumes independence between features (called the ind-FAM) and also to two existing clustering methods. We study three simulation scenarios (scenarios 1-3), each with different specifications of $\boldsymbol{Z}^{\mathrm{TR}}$, and evaluate the performance of the model for estimation of latent features and cell clustering. The specifications of $\boldsymbol{Z}^{\mathrm{TR}}$ are given in Figure 3.2(a)-(c). In all three scenarios, we assume $J = 21$ markers and $K^{\mathrm{TR}} = 7$ features. A superscript is used to denote the scenarios. In scenario 1, all features are similar to each other, and each pair of features differs by 2 markers in terms of whether a marker is expressed or not. In scenario 2, each pair of the features differs by 6 markers, and all features are fairly distinct. Scenario 3 has

groups of similar features and relatively dissimilar features. Features 1-3 are a group, differing by only one or two markers. Features 4 and 5 also are a group of similar features that differ by one marker, but they are very different from the features in the first group. The last two features 6 and 7 are each distinct from the rest of the features. We assume $I = 2$ samples, each with $N_i = 2000$ cells. The true proportions of cell clusters $\boldsymbol{w}_i^{\text{TR}}$ are given in Figure 3.2(d). The same $\boldsymbol{w}_i^{\text{TR}}$ is used for all scenarios. Each scenario contains features that are (1) rare in both samples, (2) very abundant in both samples, or (3) abundant in only one sample. Furthermore, features 1, 3, and 7 are present in one sample only, and $|R_i^{\text{TR}}| = 6$ and 5 for $i = 1, 2$, respectively. To simulate $\boldsymbol{y}_{i,n}$, we first sampled cell cluster membership labels $\lambda_{i,n}^{\text{TR}}$ according to $\boldsymbol{w}_i^{\text{TR}}$. For each (i, j), we simulated $\mu_{0,i,j}^{\text{TR}}$ from Uniform$(-1.1, -0.5)$ and $\mu_{1,i,j}^{\text{TR}}$ from Uniform$(0.7, 1.3)$. Given $z_{j,\lambda_{i,n}^{\text{TR}}} = z \in \{0, 1\}$, we generated $y_{i,n,j}$ from skewed normal distributions (Frühwirth-Schnatter and Pyne 2010) with location parameter $\mu_{z,i,j}^{\text{TR}}$, scale parameter $\sigma_i^{\text{TR}} = 1.0$, and skewness parameter $\zeta = -0.9$. The true distribution of $y_{i,n,j}$ is left skewed. When a marker is expressed, $y_{i,n,j}$ tends to have a value smaller than $\mu_{1,i,j}^{\text{TR}}$, and can be even negative with probability 0.475 when $\mu_{1,i,j}^{\text{TR}}$ is 0.7. Our ζ corresponds to $\delta \in (-1, 1)$ in Frühwirth-Schnatter and Pyne (2010). To make the simulated data more closely resemble our motivating CyTOF data, we randomly set 20% of the $y_{i,n,j}$ values to be missing when their corresponding $z_{j,\lambda_{i,n}^{\text{TR}}} = 0$. We imputed those missing values during inference, using the techniques outlined in § 3.2.3. We also did simulation studies without missing data and obtained similar inferences. We only report the results for the simulation studies with missing values.

Posterior samples were obtained via MCMC with a 6000-iteration burn-in period, and the subsequent 3000 samples kept for inference. We used weight-preserving parallel tempering with 4 chains and temperatures (1, 1.003, 1.006,

1.01) to accelerate mixing. To speed up the update of $\boldsymbol{Z}$, we constructed $p^{\star}(\boldsymbol{Z})$ using 200 cells for each sample and a thinning factor of $M = 5$ as descried in § 3.2.3. We fixed $K = 15$ and calibrated values of $\boldsymbol{\phi}$ as described in § 3.2.1; we fixed $\phi_1 = 1$ and selected the smallest value of ϕ_2 such that $\Pr(\min_{1\leq k_1<k_2\leq K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \geq \underline{d} \mid \boldsymbol{\phi}) > \underline{p}$ with $\underline{d} = 4$ and $\underline{p} = 0.95$. $\underline{d} = 4$ implies 20% of the markers are different between features in expression/no expression. We searched for such a value of ϕ_2 on a grid and chose $\phi_2 = 10$. In the Appendix, Figure B.1(a) illustrates $\Pr(\min_{1\leq k_1<k_2\leq K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \geq \underline{d} \mid \boldsymbol{\phi})$ for different values of $\underline{d}$ and ϕ_2. In addition, we performed sensitivity analyses to assess model sensitivity to the specification of ϕ_2. All computations were done on the Hummingbird Linux compute cluster at UC Santa Cruz, and the computation took approximately 0.78 hours per 1000 iterations.

Posterior inference under the rep-FAM with $\boldsymbol{\phi} = (1, 10)$ is illustrated in Figures 3.3(a)-(f), 3.4(a)-(c) and 3.5(a)-(f). In Figure 3.3(a)-(f), posterior point estimates, $\hat{\boldsymbol{Z}}_i$ (after transposing) are shown with $\hat{\boldsymbol{w}}_i$ in parentheses for each of the simulation scenarios. If inferred features are identical to a feature in the true $\boldsymbol{Z}$, they are labeled with the corresponding feature number in the true $\boldsymbol{Z}$; otherwise, they are given new feature labels, starting from $K^{\text{TR}} + 1$. In scenario 1, $\boldsymbol{Z}^{1,\text{TR}}$ and $\boldsymbol{w}_i^{\text{TR}}$ are well recovered for both samples. In scenario 2, two additional features, features 8 and 9, are included in $\hat{\boldsymbol{Z}}_i^2$. Those additional features are close to true feature 3 of $\boldsymbol{Z}^{2,\text{TR}}$, and markers 8 and 9 which are expressed in true feature 3 are inferred as not expressed for them, potentially due to the left skewness of the true distribution of y. We further checked $y_{i,n,j}$ of those markers for the cells with $\hat{\lambda}_{i,n} = 8$ or 9, and found that their observed expression levels are close to zero or negative. For scenario 3, almost all the true features in $\boldsymbol{Z}^{3,\text{TR}}$ are recovered for both samples, except for true feature 2 in sample 1. True feature 2 is similar

to true features 1 and 3, but is not abundant in the samples. Instead, true feature 1 is recovered and included in both $\hat{\boldsymbol{Z}}_i^2$, $i = 1, 2$, although the feature does not appear in sample 2 in the truth. On the other hand, true feature 7 is rare and present in sample 1 only, but it is well recovered in $\hat{\boldsymbol{Z}}_1^3$ and $\hat{\boldsymbol{w}}_1$ since it is very different from the other features. Figures 3.4(a)-(c) show the posterior distributions of the number of selected features, $|R_i|$ for each of the three scenarios. The red dashed vertical lines denote the simulation truth of $|R_i|$. The posterior distributions are highly peaked at their modes in all scenarios. As shown in the figures, the simulation truth is well recovered in both samples under scenario 1. For scenario 2, due to the presence of superfluous small clusters, the posterior mode differs from the truth by 2. For scenario 3, the posterior mode of $|R_2|$ is the same as its truth, but the posterior mode of $|R_1|$ is smaller than its truth by 1 since two similar features are merged. Figure 3.5(a)-(f) shows heatmaps of $\boldsymbol{y}_{i,n}$, where cells are in rows and markers in columns. The cells are rearranged by their cluster membership estimates $\hat{\lambda}_{i,n}$ within each sample. Red and blue colors represent positive and negative expression levels, respectively; while black represents missing values. The yellow horizontal lines divide inferred cell clusters. Cells within a cluster, characterized by $\hat{\mathbf{z}}_{i,k}$, tend to have homogeneous expression patterns, indicating good model fit. When the expression level of a marker is missing in a cell, the cell tends to be clustered with cells that have low expression levels for the same marker. We also compare clustering estimates $\hat{\lambda}_{i,n}$ to the truth by using the adjusted Rand index (ARI) (Hubert and Arabie 1985). An ARI of 1 indicates perfect clustering, the expected ARI of independent clusterings is 0, and clusterings that are worse than independent clusterings can yield a negative ARI. Table 3.1 shows that the ARI under the rep-FAM are large, indicating that the rep-FAM produces reasonable estimates of cell clustering.

Method	Sc. 1 Sample 1	Sc. 2 Sample 1	Sc. 3 Sample 1	Sc. 1 Sample 2	Sc. 2 Sample 2	Sc. 3 Sample 2
rep-FAM	0.811	**0.946**	**0.935**	**0.790**	**0.877**	0.834
ind-FAM	**0.816**	0.914	0.930	0.781	0.842	**0.846**
FlowSOM	0.239	0.928	0.801	0.466	0.853	0.751
MClust	0.562	0.839	0.865	0.616	0.750	0.540

Table 3.1: ARI (adjusted Rand index) for different methods under each of the three simulation scenarios. Each simulated data includes two samples. A larger value of ARI is better. The method with the highest ARI for each sample is in bold.

For comparison, we applied the ind-FAM to the simulated datasets. The ind-FAM assumes independence between features with fixed K similar to a conventional FAM based on the IBP. We built the ind-FAM by assuming $z_{j,k} \mid v_k \overset{ind}{\sim}$ Bernoulli(v_k) instead of (3.1), and preserved the other model components in the rep-FAM, such as feature selection by random $r_{i,k}$. The results under the ind-FAM are shown in Figures 3.3(g)-(ℓ), 3.4(d)-(f), and 3.5(g)-(ℓ). Overall, the ind-FAM recovers the simulation truth well for scenario 1, similar to the rep-FAM. However, it tends to yield more features in scenarios 2 and 3, as shown in Figures 3.3 (h), (k) and (ℓ), than the rep-FAM. Particularly, in scenario 2, true feature 3 is repeated in sample 1 (see panel (h) of the figure), and duplicate columns appear in $\hat{\boldsymbol{Z}}_1^2$. Features 8 and 9 also resemble true feature 3. That is, the ind-FAM uses four features to infer true feature 3. Figure 3.4(d)-(f) illustrates the posterior distributions of $|R_i|$. In contrast to those under the rep-FAM, the posterior distributions of the ind-FAM tend to be more disperse, especially in scenario 2. Figures 3.5(g)-(ℓ) illustrate clusterings of $y_{i,n}$ by $\hat{\lambda}_{i,n}$ under the ind-FAM. Similar to the clustering under the rep-FAM, the ind-FAM clusters cells based on expression patterns by $\hat{\mathbf{z}}_k$. Since the ind-FAM tends to produce more redundant features, it produces some trivial cell clusters. Also, the ARI in Table 3.1 indicates that the ind-FAM produces reasonable estimates of cell clustering, and its performance is close to

that of the rep-FAM. However, clustering by the ind-FAM is worse than that by the rep-FAM in scenario 2 possibly because the ind-FAM produces duplicated features.

As additional comparators, we applied some commonly used clustering methods to the simulated datasets. We combined all cells from the samples to obtain a joint clustering of the samples, and used MClust (Scrucca et al. 2016) and FlowSOM (Van Gassen et al. 2017). MClust performs model-based clustering by fitting Gaussian mixture models. It provides various options of estimating the covariance matrix, and estimates its parameters via the EM algorithm. FlowSOM is a common clustering method using self-organizing maps (SOM) for cytometry data. Weber and Robinson (2016) reported that FlowSOM outperforms its competing clustering methods in terms of providing fast and quality clustering of cytometry data. MClust and FlowSOM cluster cells based on marker expression values. They do not handle missing values; while we imputed the missing values by specifying the missing mechanism as outlined in Appendix B.2.1 prior to analysis. Thus, prior to analyzing the data via FlowSOM and MClust, we replaced the missing data with random samples of the negative portion of the observed data. Appendix Figures B.8-B.10 provide heatmaps of the data with cells arranged by clusters obtained by FlowSOM and MClust. Overall, FlowSOM collapses similar true clusters and produces a smaller number of large clusters. On the other hand, MClust tends to produce more clusters of smaller sizes. We also computed ARI to compare their clustering estimates to the truth. Table 3.1 shows that the performance of MClust and FlowSOM is poor in clustering cells, especially for scenarios 1 and 3, where the truth includes similar features. We also used t-SNE (Maaten and Hinton 2008), a stochastic non-linear dimensionality-reduction technique, to inspect a lower-dimensional summary of the simulated data. The t-SNE plots are

included in Appendix Figure B.11. Some subpopulations are well separated in the lower-dimensional space, but the populations having similar maker expression patterns are greatly overlapped, especially for scenario 1, where subpopulations differ by only two markers. A more detailed comparison is included in Appendix B.3.

In addition, we conducted a sensitivity analysis of the rep-FAM to the specification of $\boldsymbol{\phi}$. While increasing ϕ_2 encourages features to be more distinct, ϕ_2 cannot be made arbitrarily large. Figure 3.6 displays the posterior inference under the rep-FAM with $\boldsymbol{\phi} = (1, 100)$. The rep-FAM with $\boldsymbol{\phi} = (1, 100)$ implies $\Pr(\min_{1 \le k_1 \le k_2 \le K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \ge 5 \mid \boldsymbol{\phi}) \approx 0.95$ *a priori*. Due to the strong repulsion implied in the prior, the inference under the rep-FAM suffers in scenarios 1 and 3 which include similar features in the truth. On the other hand, the inference in scenario 2 is significantly improved since the true features are very distinct. We also tried $\boldsymbol{\phi} = (1, 1)$ and (1, 25). Appendix B.3 includes details of the sensitivity analyses.

1: 2B4	2: 3DL1	3: CD158B	4: CD8
5: CD94	6: CKIT	7: DNAM1	8: EOMES
9: NKG2A	10: NKG2D	11: NKP30	12: SIGLEC7
13: SYK	14: TBET	15: ZAP70	

Table 3.2: Index for markers referenced in data analysis.

3.4 Analysis of the CyTOF Data

We next report application of the proposed rep-FAM to analyze the CyTOF dataset, which consists of two samples taken from leukemia patients 30 days after natural killer cell infusion. As articulated in Chapter 2, 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 promote identifying NK cell subpopulations that are more distinct in terms of marker expression levels, we apply rep-FAM to the samples analyzed by CyTOF. The samples contain expression levels for 32 surface markers from 4677 and 1367 individual cells, respectively. We removed having markers whose transformed expression levels that were nearly all above zero or all below zero, because they do not contribute to differentiating cell subpopulations. After this preprocessing, 15 markers were included for analysis, and are listed in Table 3.2. Cells with extreme expression values also were removed, after which 4556 and 1308 cells remained in the samples.

Posterior inference was performed via MCMC, as outlined in § 3.2.3. A burn-in of 10000 was used and the subsequent 5000 posterior samples were kept for inference. To facilitate better mixing, we used WPPT with temperatures $(1, 1.003, 1.006, 1.01)$ and proposed swaps between every pair of chains at each MCMC step. For the intrinsic MCMC, we used 5% of the data for a minimal training sample and a thinning factor M of 5. We used cell clustering estimates obtained by MClust to initialize $\boldsymbol{\theta}$ and specify some fixed hyperparameters. For example, we obtained a crude preliminary estimate of the number of clusters, which was 5, and set $K = 25$ to accommodate potentially more clusters. We also used the calibration approach in § 3.2 to specify $\boldsymbol{\phi}$. We used $\underline{u} = 3$ and $\underline{p} = 0.95$ with $J = 15$ and $K = 25$, and chose $\phi_1 = 1$ and $\phi_2 = 25$ through the calibration process described in § 3.2. This reflects the prior belief that subpopulations are expected to differ in their expression pattern for 20% of the markers. For the remaining hyperparameters, we let $L_0 = 6$ and $L_1 = 3$ to accommodate left-skewness observed from empirical distributions of observed $y_{i,n,j}$, and set $\alpha \sim \text{Gamma}(1, 1)$, $\delta_{z,\ell} \overset{iid}{\sim} \text{TN}^+(1, 0.1)$, $\eta_{z,i,j,\ell} \sim \text{Dirichlet}_{L_z}(1)$, and $\sigma_i^2 \sim \text{InverseGamma}(3, 1)$. The computations took approximately 3.3 hours per 1000 iterations.

Figures 3.7(a)-(b) show posterior point estimates of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. Since features (cell subpopulations) can appear in both samples, features that appear in $\hat{\boldsymbol{Z}}_1$ first were ordered by $\hat{w}_{1,k}$, and labeled using integers starting from 1. Features in $\hat{\boldsymbol{Z}}_2$ that did not appear in $\hat{\boldsymbol{Z}}_1$ again were ordered and labeled by $\hat{w}_{2,k}$, starting from $|R_1| + 1$. Features included in both samples thus have the same labels. Together, the two $\hat{\boldsymbol{Z}}_i$'s have a total of 22 features, most of which are shared. Features 15 and 18 are unique in sample 1, and features 23 and 24 unique in sample 2. The abundances of those features are small, however. Although most features are shared by the samples, their abundances are very different between the samples. In particular, features 1 and 2 are very abundant in sample 1 and make up 23.6% and 16.6% of the sample, respectively, but they make up only 10.5% and 2.0% of sample 2. The most abundant feature in sample 2 is feature 13, which makes up about 22.6% of the sample feature 13 accounts for only 1.2% of sample 1. The posterior distributions of $|R_i|$ are shown in Figure 3.8(a). The number of selected features for each sample is centered around 22 for both samples, with $\Pr(|R_i| = 22 \mid \boldsymbol{y}) > 0.5$. Figures 3.9 (a) and (b) show heatmaps of the expression levels of cells after rearrangement by clustering estimates in each sample, with the most abundant clusters at the bottom. The yellow horizontal lines separate cell clusters by $\hat{\lambda}_{i,n}$. The heatmaps show that expression levels within a cluster are homogeneous.

For comparison, we also applied the ind-FAM to the CyTOF data. Figures 3.7(c)-(d) show posterior point estimates of $\boldsymbol{Z}$ and $\boldsymbol{w}_i$. Compared to the rep-FAM, the ind-FAM tends to produce more features with smaller abundances. Unlike the rep-FAM, the ind-FAM infers 24 features for each sample. Using the ind-FAM, the two samples share eight features, and together have a total of 35 features. $\hat{\boldsymbol{Z}}_i$ minimizes the sum-of-squared differences separately for each sample, and the total

number of features in $\{\hat{\boldsymbol{Z}}_i, i = 1, \ldots, I\}$ may exceed K. The features common in both samples account for 68.3% of the cells in sample 2, and the remaining features are specific to sample 2 and include 31.7% of the cells. Figure 3.8(b) shows the posterior distributions of the number of selected features in each of the samples, $|R_i|$. Compared to the figure in (a), they are concentrated at very large values, such as 24 and 25. Notably, K may need to be set at a larger value. We also compared pairwise distances between the inferred features in $\hat{\boldsymbol{Z}}_i$ under the two models, $d(\hat{\mathbf{z}}_{k_1}, \hat{\mathbf{z}}_{k_2})$. The histograms of the pairwise distances in Figure 3.8(c) and (d) show that overall the features inferred under the rep-FAM are more distinctive than those under the ind-FAM. For instance, in sample 1 (Figure 3.8(c)), feature-pairs with pairwise distances less than 4 are appear much less frequent in the rep-FAM; whereas pairwise distances of 5 to 9 appear with much greater frequency.

For comparison, we applied MClust and FlowSOM to the dataset to obtain clusterings of the samples, combining the two samples for these analyses. The number of clusters was chosen to be 7 for MClust based on BIC and 12 by FlowSOM. Figure B.23 has heatmaps of y after rearrangement by cluster membership estimates. These clustering methods appear to yield clusters that are less homogeneous. Particularly in (c), cells which having high and low expressions of the same markers are included in the bottom cluster.

We also performed sensitivity analyses on the specification of the fixed hyperparameters by varying the values of ϕ_2 and p_i. Appendix Figures B.18-B.20 illustrate results for $\phi_2 = 1$, 10, and 100. Results for p_i=0.2 and 0.3 are given in the Appendix Figures B.14 - B.17.

3.5 Conclusions

We have proposed a repulsive FAM which encourages features that are distinct from each other by including a repulsion function to limit the degree to which features in the (object, feature) identification matrix may be similar. The repulsion function can be calibrated using prior simulations. We demonstrated a way to calibrate hyperparameters of the repulsion function using prior information. Compared to conventional FAMs that assume independence between features, the rep-FAM yields more parsimonious results and potentially more biologically more meaningful inferences on the underlying structure. We applied the rep-FAM to the problem of inferring cell subpopulations using CyTOF data.

The proposed rep-FAM may replace conventional FAMs when it is desirable to identify distinct features. The rep-FAM can be extended in several ways. For example, when samples are recorded with covariates, $\mathbf{x}$, we may develop a regression model $p(\boldsymbol{Z} \mid \mathbf{x})$ that allows $\boldsymbol{Z}$ to be indexed by x. Similarly to Williamson et al. (2010, 2019), a Gaussian process may be assumed to induce explicit dependence of $\boldsymbol{Z}$ on $\mathbf{x}$. This creates more complexity, but may improve inferences on sample-specific structures when heterogeneity between samples can be explained by $\mathbf{x}$.

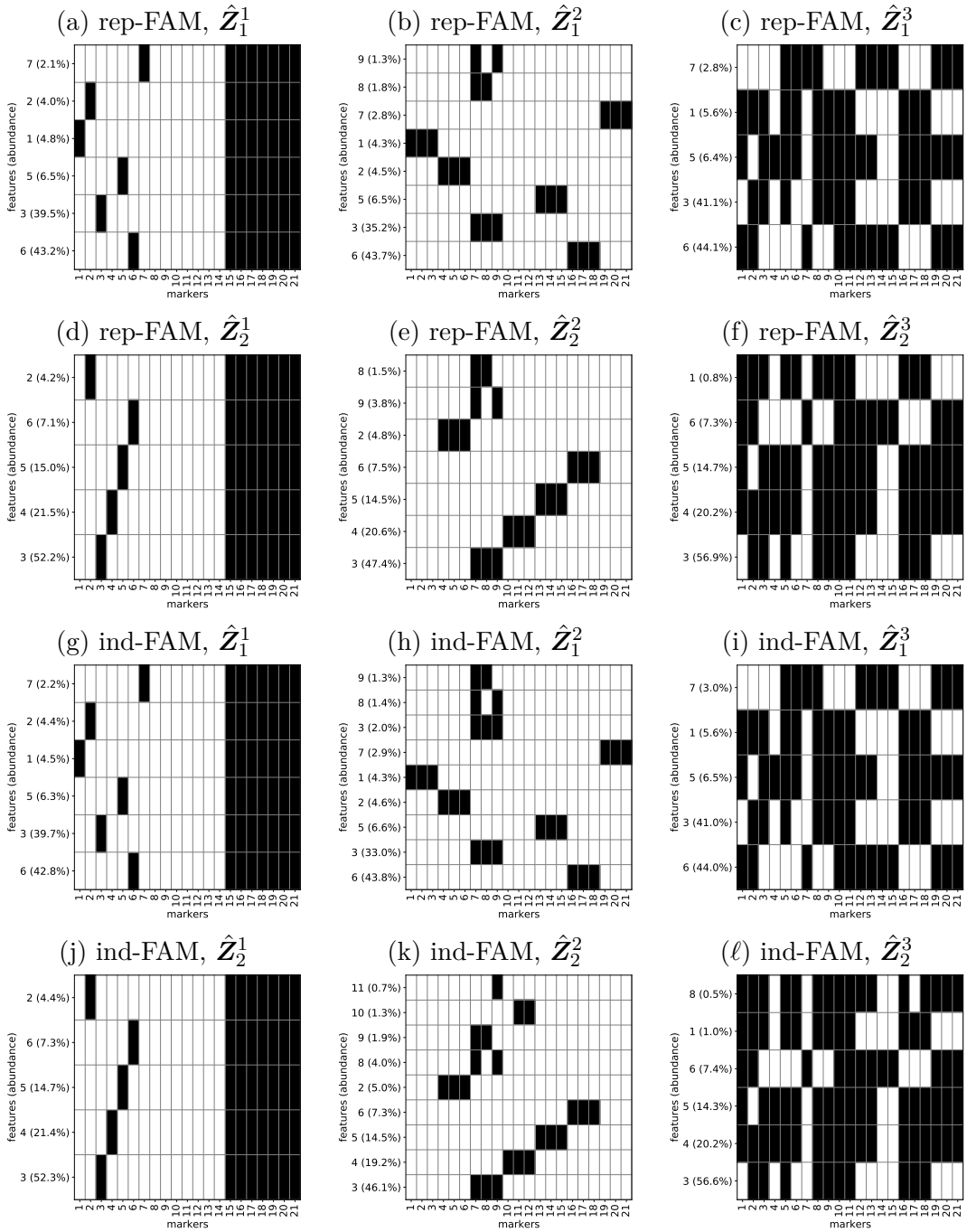


Figure 3.3: Posterior point estimates for transpose of $\boldsymbol{Z}$ and $\boldsymbol{w}$ for each sample ($i = 1, 2$) under the three scenarios. Panels (a)-(f) show $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ under the rep-FAM with $\boldsymbol{\phi} = (1, 10)$, and panels (g)-($\ell$) show $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ under the ind-FAM. The results under scenarios 1-3 are in columns 1-3, respectively. In $\hat{\boldsymbol{Z}}_i$, colors white and black represent 0 and 1, respectively, and $\hat{\boldsymbol{w}}_i$ is shown on the left.

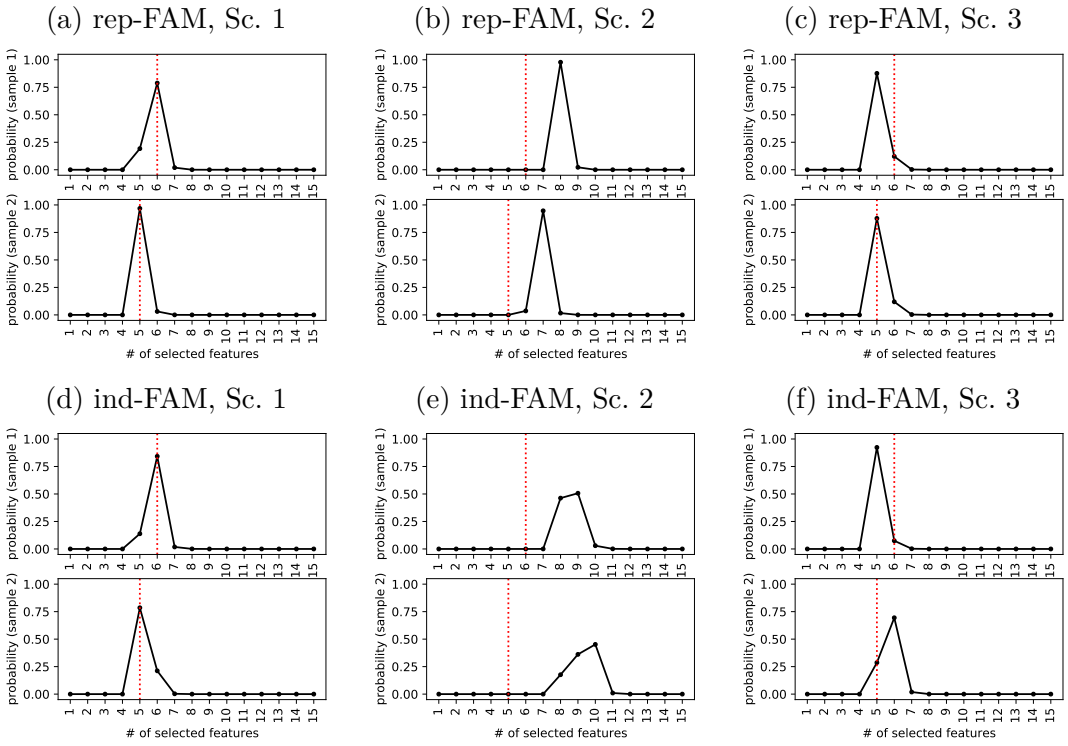


Figure 3.4: Posterior distributions of number of selected features, $|R_i|$ for each sample under the three simulation scenarios. Simulation truth for $|R_i$ are represented by the dashed vertical lines. The results under the rep-FAM and ind-FAM are in the top and bottom rows, respectively.

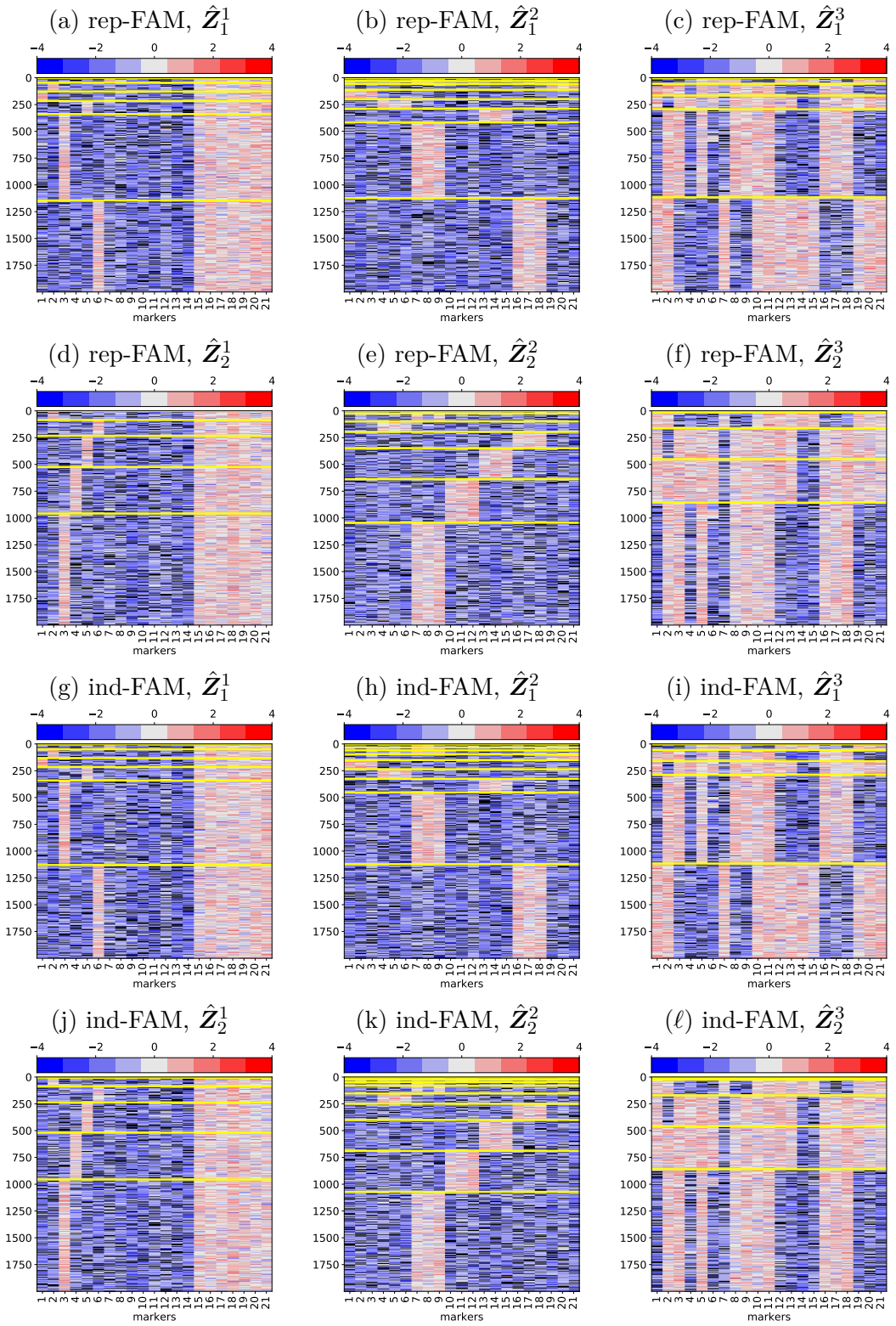


Figure 3.5: Clustering of $\boldsymbol{y}_{i,n}$. Heatmaps of $\boldsymbol{y}_{i,n}$ are shown in each panel after rearranged by posterior point estimate of clustering membership $\hat{\lambda}_{i,n}$ for each sample ($i = 1, 2$) under scenarios 1-3. Panels (a)-(f) show $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ under the rep-FAM with $\boldsymbol{\phi} = (1, 10)$, and panels (g)-(ℓ) show $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ under the ind-FAM.

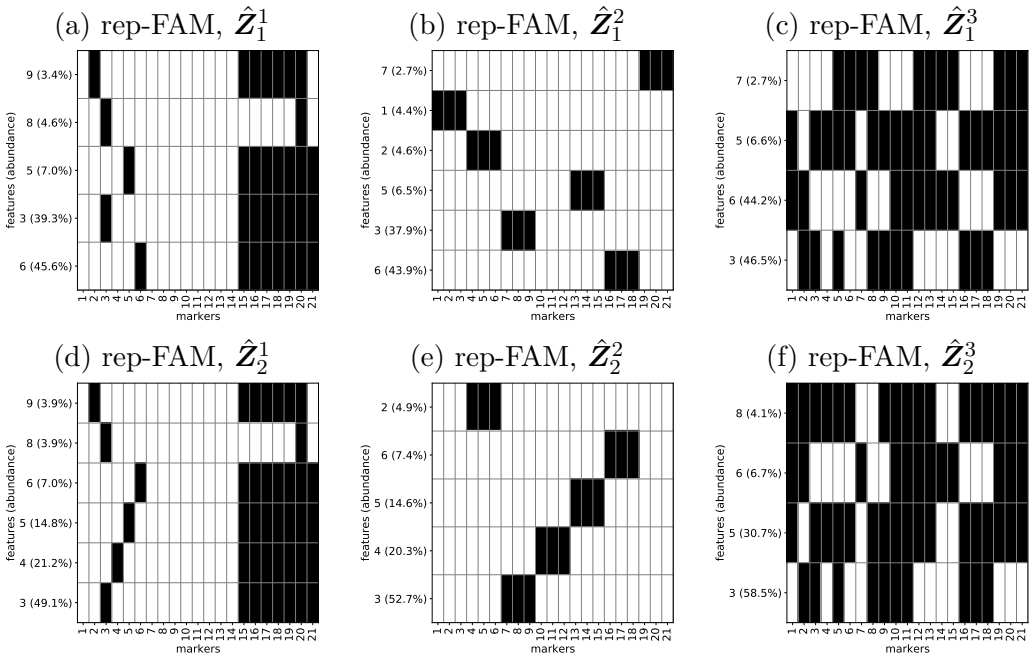


Figure 3.6: [Sensitivity Analysis for the Simulation Studies] Posterior point estimates under the rep-FAM with $\boldsymbol{\phi} = (1, 100)$ are illustrated. Transpose of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ for samples 1 and 2 under the three simulation scenarios are shown.

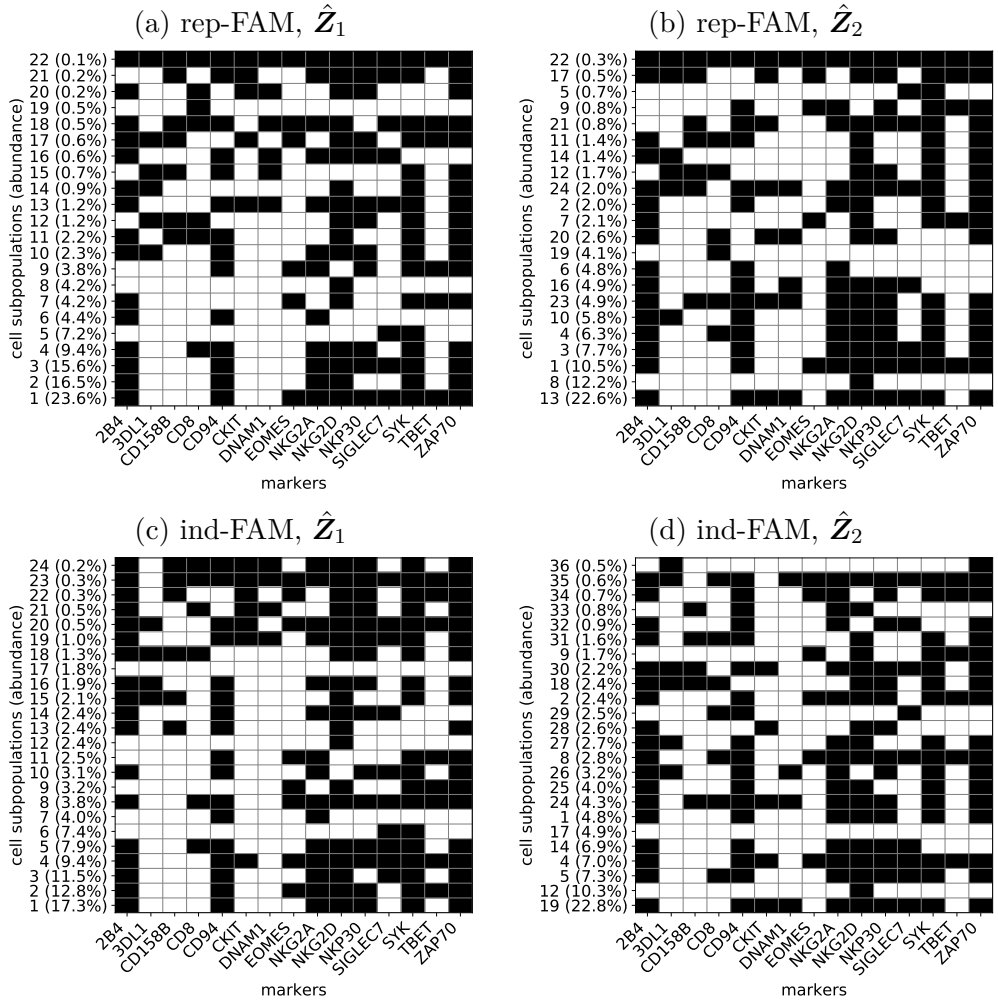


Figure 3.7: [CyTOF Data] Posterior point estimates, $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ under the rep-FAM are shown in panels (a) and (b) for samples 1 and 2, respectively, and those under the ind-FAM are shown in (c) and (d).

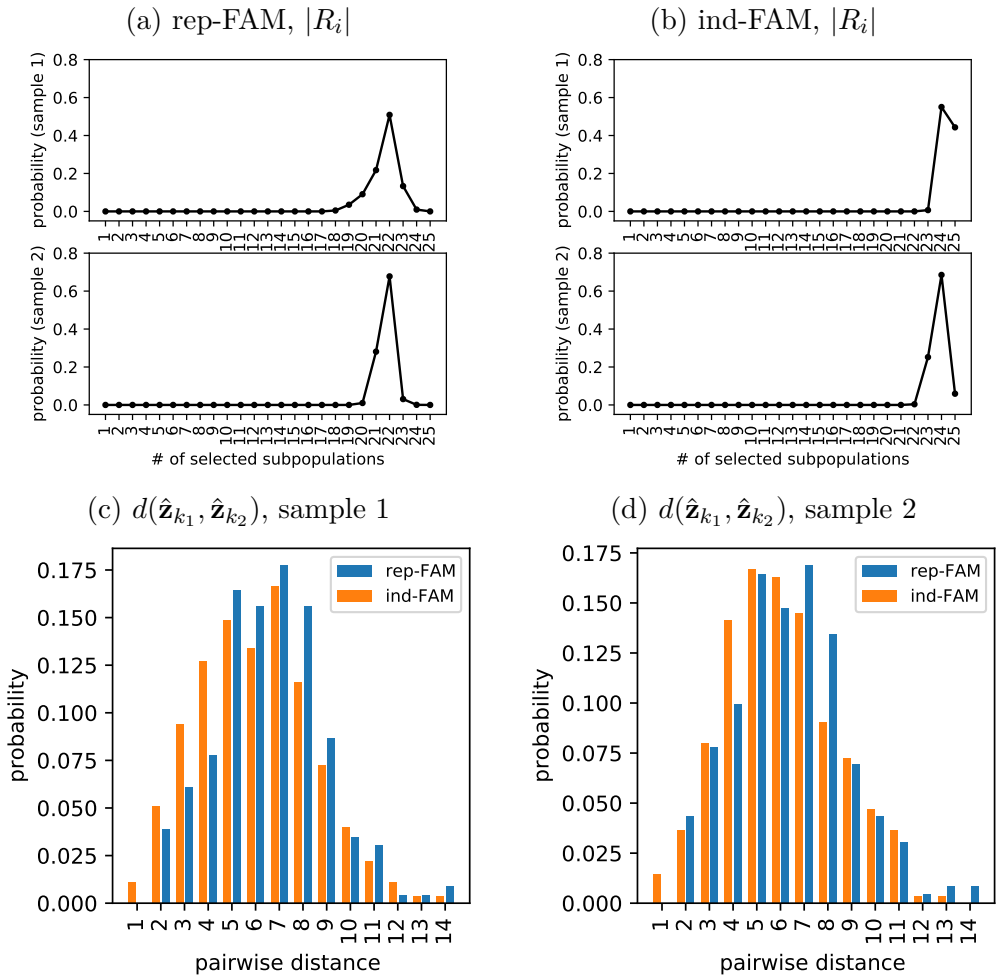


Figure 3.8: [CyTOF Data] Panels (a) and (b) have posterior distributions of the number of selected features within each sample $|R_i|$ under the rep-FAM and ind-FAM, respectively. Panels (c) and (d) shows histograms of $d(\hat{\mathbf{z}}_{k_1}, \hat{\mathbf{z}}_{k_2})$ for every pair of features in $\hat{\boldsymbol{Z}}_i$ under the two models.

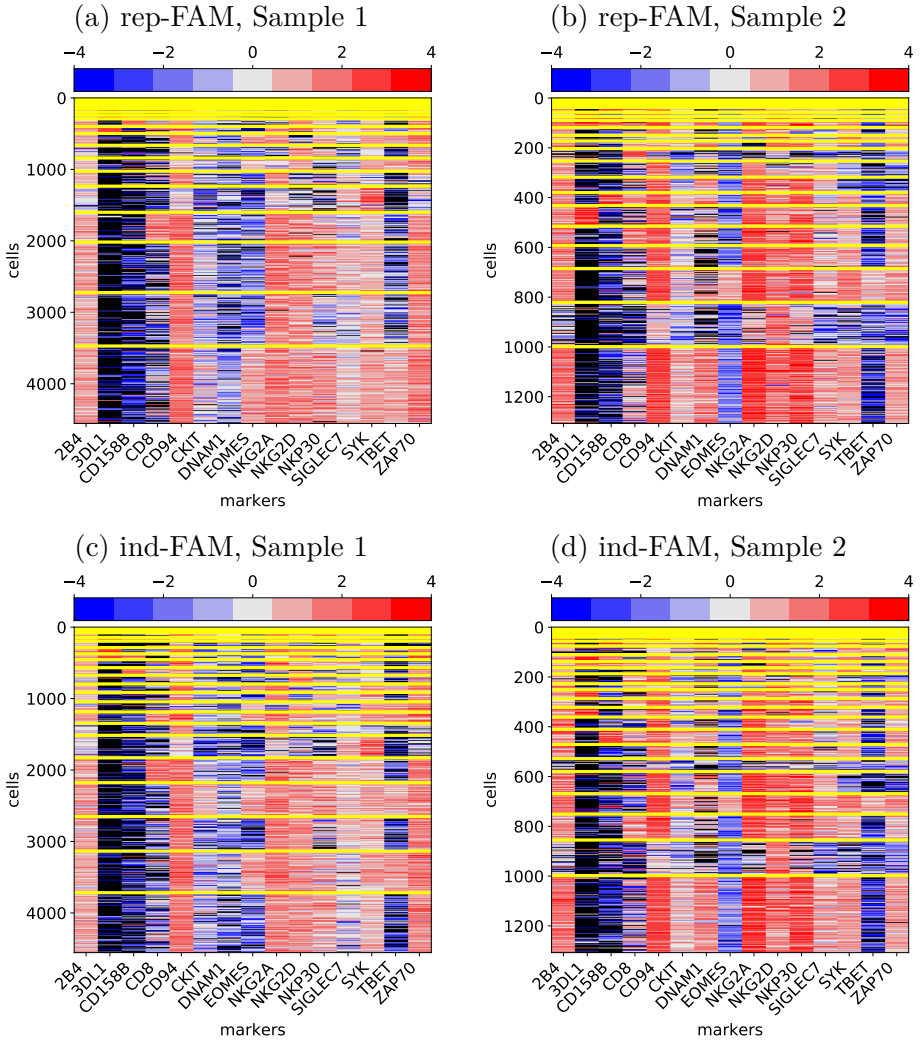


Figure 3.9: Marker expression levels y_i for each cell subpopulation, sorted by row according to posterior estimate of subpopulation membership labels $\lambda_{i,n}$, with the most abundant subpopulations at the bottom, for each sample ($i = 1, 2$), with $p_i = 0.2$ and $\phi_2 = 0,25$.

Chapter 4

A Bayesian Differential Distribution Approach for Zero-inflated Data with Applications to Cytometry Data

4.1 Introduction

In clinical applications, practitioners often attempt to compare the effects of various treatments. When responses are real values, testing for shifts in mean responses under various experimental conditions may be plausible if measurements fluctuate, due to biological and technical variabilities, around some latent (average) value. However, in many cases, tests of mean-shifts alone may not sufficiently describe meaningful distributional differences. For example, if the responses under a particular treatment are multimodal, then computing differences in mean responses under a treatment and control group can be misleading, as noted by

Korthauer et al. (2016) in single-cell RNA-seq experiments.

As a specific example, we consider cytometry by time-of-flight (CyTOF) data to evaluate changes in marker expression levels of natural killer (NK) cells by transforming growth factor beta (TGF-β). NK cells are lymphocytes of the innate immune system which are able to recognize and kill virally infected cells. They play a critical role in cancer immune surveillance. The use of NK cells for cancer therapy has been demonstrated by Wu and Lanier (2003) and Lanier (2008) and has the potential to become a powerful modality in cancer treatment. TGF-β is an immunosuppressive cytokine that severely affects the function of NK cells (Regis et al. 2020) and can be produced by tumor cells to attack NK cells. However, the way in which TGF-β alters the activity of NK cells has not been fully investigated. Understanding how TGF-β impacts the activities of NK cells is thus important in order to develop NK cell-based immunotherapies (Slattery and Gardiner 2019). In NK cell studies, expression levels for several NK cell markers are measured for each cell in a sample by CyTOF. In CyTOF data, multiple cell surface marker expression levels are recorded simultaneously on several thousand cells. Marker expression levels are non-negative values where small values represent low marker expression levels and larger values represent higher levels of marker expression. A variety of NK cell subpopulations coexist within a given sample and this heterogeneity is exhibited through different expression patterns of the NK cell markers. These NK cell subpopulations also vary in biological function and are correlated with expression patterns. The distribution of the marker expression levels are oftentimes multimodal because of the heterogeneous nature of NK cells. They also tend to be skewed even after log transformation and occasionally contain outliers. For example, Figure 4.1(a) shows the histograms of logarithm transformed expressions of marker CD3z for non-zero expression levels

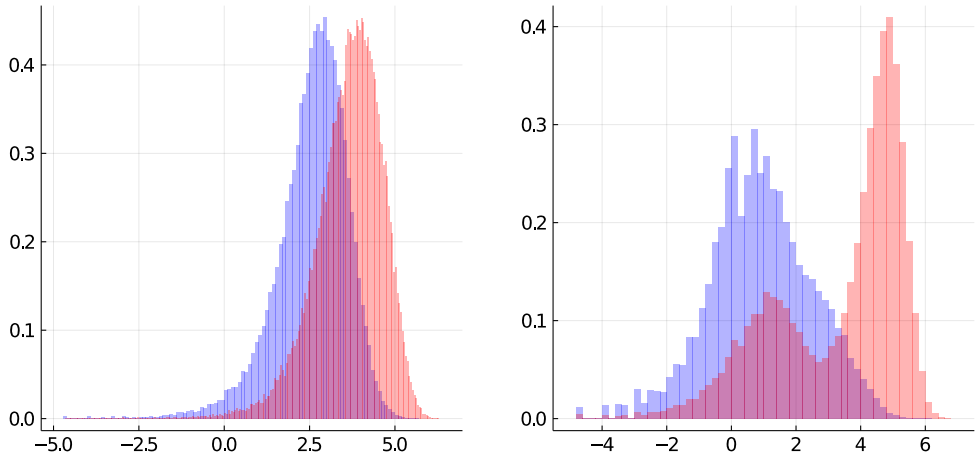


Figure 4.1: Histogram of $\tilde{y}_i$, for $i \in \{1, 2\}$ and markers CD3z and CD103. (a) Distribution of $\tilde{y}_1$ and $\tilde{y}_2$ in blue and red respectively for marker CD3z from a donor, with $N_1 = 86915$, $N_2 = 92468$, $Q_1 = 1361$, $Q_2 = 411$. (b) Distribution of $\tilde{y}_1$ and $\tilde{y}_2$ in blue and red respectively for marker CD103 from another donor, with $N_1 = 90067$, $N_2 = 92044$, $Q_1 = 49461$, $Q_2 = 9801$.

in two samples, where the blue and red colors denote the conditions, before and after treatment TGF-β, respectively. In the figure, the empirical distribution is left skewed and some outliers are present. Similarly, Figure 4.1(b) shows histograms of logarithm transformed expression levels CD103 in different samples. The distribution is apparently bimodal and slightly left-skewed. Critically, stochastic ordering constraints (Hanson et al. 2008, Kottas 2011) for these distributions cannot in general be imposed due to the potential multimodality of expression levels in a sample. Analyses of CyTOF data are further complicated by the presence of exact zeros which, when present in large quantities, indicate that a particular marker is likely not expressed. Papoutsoglou et al. (2019) recommend jittering the zeros by adding arbitrarily generated random values to aid visualization in manual gating tasks. While suitable for gating, jittering the zeros for modeling purposes introduces biases into the analysis. Constructing a flexible model that accommodates complexity in CyTOF data is critical to better understanding the

mechanism in NK cells induced by TGF-β.

In this chapter, we propose a zero-inflated mixture of skew-t distributions to quantify the effects of TFG-β on the distribution of NK marker expression levels, to model data that exhibit excess zeros, outliers, skewness, or multimodality. The skew-t distribution contains as special cases the skew-normal and t distributions and is governed by location, scale, degrees of freedom, and skewness parameters (Azzalini and Capitanio 2003). By varying the parameters, one is able to model skewed data with outliers where a Gaussian mixture model with several components may be needed. A skew-t mixture additionally allows us to model multimodality and shifts in distributions across samples (Frühwirth-Schnatter and Pyne 2010). We demonstrate how in the presence of outliers and highly skewed data, mixtures of skew-t distributions are able to more efficiently model complex distributions as seen in CyTOF data using fewer mixture components than their Gaussian counterparts. To avoid introducing biases by adding noise to the zeros, we model zeros by a zero-inflated mixture component. The resulting model enables coherent comparisons of distributions from various experimental conditions. We also propose a metric to quantify the difference between a pair of distributions. This metric computes a normalized area between estimated cumulative distribution functions over a majority of the support of the data to give an indication of the degree to which distributions are different. While CyTOF is able to produce analyses for multiple markers simultaneously, our proposed methods are applied to markers one at a time for computational savings.

This project will proceed as follows. In Section 4.2, we propose our probability model for this work. In Section 4.3 we demonstrate the effectiveness of our model via a simulation study intended to mimic real data which we will then analyze in Section 4.4. We present concluding remarks in Section 4.5.

4.2 Probability Model

Sampling Model We will proceed by establishing model notation. For a marker in samples from a particular donor, let $y_{i,n} \geq 0$ be the expression level of a marker in cell $n \in \{1, \ldots, N_i\}$ for sample $i \in \{1, \ldots, I\}$, where I denotes the number of samples, and N_i denotes the number of cells in sample i. In the simulation studies and the data analyses in sections 4.3 and 4.4, respectively, we have $I = 2$, where $i = 1$ and 2 denote control and treatment groups, respectively. Marker expression levels $y_{i,n}$ are recorded as non-negative real values. When the signals from the CyTOF instrument are weak for the marker in a cell, expression levels are recorded as 0. Throughout the real CyTOF data, 0's are observed frequently, and thus need to be accounted for in the modeling. We let $y_{i,n} \mid F_i \overset{ind}{\sim} F_i$, and assume that F_i is a zero-inflated mixture model

$$F_i(y) = \gamma_i \cdot \delta_0(y) + (1 - \gamma_i) \cdot G_i(y), \tag{4.1}$$

with $\delta_A(\cdot)$ denoting the Dirac measure at A, γ_i the probability of $y_{i,n}$ being zero, and G_i a probability distribution on $\mathbb{R}^+$. The model in (4.1) assumes that an observed expression level in sample i takes the value of zero with probability γ_i, and with the remaining probability $(1 - \gamma_i)$, $y_{i,n}$ follows G_i. We use a mixture model of log-skew-t distributions for G_i to capture various patterns including multi-modality and skewness,

$$G_i = \sum_{k=1}^{K} \eta_{i,k} \cdot \text{log-skew-}t(\mu_k, \sigma_k, \nu_k, \phi_k), \tag{4.2}$$

where K is a pre-specified number of mixture components and $\boldsymbol{\eta}_i$ are probability vectors of length K. In Equation (4.2), the samples share the mixture components, while the mixture weights $\boldsymbol{\eta}_i$ are indexed by i, as is the case with γ_i in

Equation (4.1). log-skew-$t(\mu_k, \sigma_k, \nu_k, \phi_k)$ denotes the log-skew-t distribution with location μ_k, scale σ_k, degrees of freedom ν_k, and shape ϕ_k, and the distributions is defined as follows; we first let $\tilde{y}_{i,n} = \log(y_{i,n})$ for $y_{i,n} > 0$ and assume that $\tilde{y}_{i,n}$ follows a skew-t mixture distribution (Frühwirth-Schnatter and Pyne 2010), that has a pdf of the form

$$p(\tilde{y} \mid \mu, \sigma, \nu, \phi) = \frac{2}{\sigma} \cdot t_\nu(u) \cdot T_{\nu+1}\left(\phi \cdot u\sqrt{\frac{\nu+1}{\nu+u^2}}\right), \text{ for } \tilde{y} \in \mathbb{R}, \tag{4.3}$$

where $u = (\tilde{y} - \mu)/\sigma$, and $t_\nu(\cdot)$ and $T_\nu(\cdot)$ denote, respectively, the pdf and cdf of a standard Student's t distribution with degrees of freedom ν. We will write $\tilde{y}_{i,n} \sim \tilde{G}_i$, where $\tilde{G}_i = \sum_{k=1}^{K} \eta_{i,k} \cdot \text{skew-}t(\mu_k, \sigma_k, \nu_k, \phi_k)$ for brevity where applicable, as this is equivalent to the model in (4.2). We let the parameters of the log-skew-t components be random, and will later discuss the prior specification of the parameters. In practice, a Gaussian mixture model is a common choice for G_i. However, skew-t mixture distributions more efficiently model complex data with multi-modality, skewness, or outliers than their Gaussian counterparts as the individual skew-t components are able to capture skewness and outliers (Frühwirth-Schnatter and Pyne 2010). Figures 4.2 and 4.3 compare the performance of a skew-t mixture model to those for mixtures of t, normal, and skew-normal distributions for skew data with outliers. Figure 4.2(a) shows a histogram of 1000 data points simulated from a skew-$t(2, 1, 7, -7)$ distribution, where the black dotted line represents the true density. Mixtures of normal, t, skew-normal, and skew-t were fit to the skew-t data, with $K = 1, \ldots, 5$, and the deviance information criterion (DIC) was computed (Spiegelhalter et al. 2002). DIC is a model comparison metric which penalizes for model complexity, commonly used for Bayesian model

selection problems. Models with lower DIC are favored. DIC can be computed as

$$\text{DIC} \quad = \quad \overline{D(\theta)} + \overline{\text{var}(D(\theta))}. \tag{4.4}$$

$D(\theta) = -2 \log p(\text{data} \mid \theta) + C$ is the deviance with log-likelihood $\log p(\text{data} \mid \theta)$ and model parameters θ, and a constant C that cancels out in computations that compare models; $\overline{D(\theta)}$ is the deviance averaged over posterior samples of θ; and $\overline{\text{var}(D(\theta))}$ is the variance of the deviance evaluated at posterior samples θ. The variance term is guaranteed to be positive and increases with model complexity. Figure 4.2(b) shows the deviance information criterion (DIC) for the various models. According to the figure, the skew-t mixture model yields the best model fit overall, followed by the skew-normal. More importantly, the DIC indicates that the model with $K = 1$ is the best among the skew-t mixture distributions. However, more than one component is needed to adequately model the data for the other mixtures. Figure 4.3 illustrates estimates with 95% pointwise credible intervals in blue under each model, with the pointwise posterior mean denoted by dotted lines, and the true underlying density in solid blue. The skew-t model captures the true density most closely with one component; whereas the other models show misfits, even with up to $K = 5$ mixture components.

Alternatively, the G_i's can be modeled using Bayesian nonparametric (BNP) approaches. The Dirichlet process (DP) (Ferguson 1973, Ferguson et al. 1974, Antoniak 1974, Sethuraman 1994) is a popular BNP prior for probability measures. In the context of mixture modeling and density estimation, DP mixture models (Escobar and West 1995, Gasparini 1996) elegantly provide the flexibility to model complex distributions without requiring the pre-specification of a fixed number of mixture components, as it can be estimated from the data when appropriate priors are specified (Escobar and West 1995, Lo 1984, Rasmussen et al.

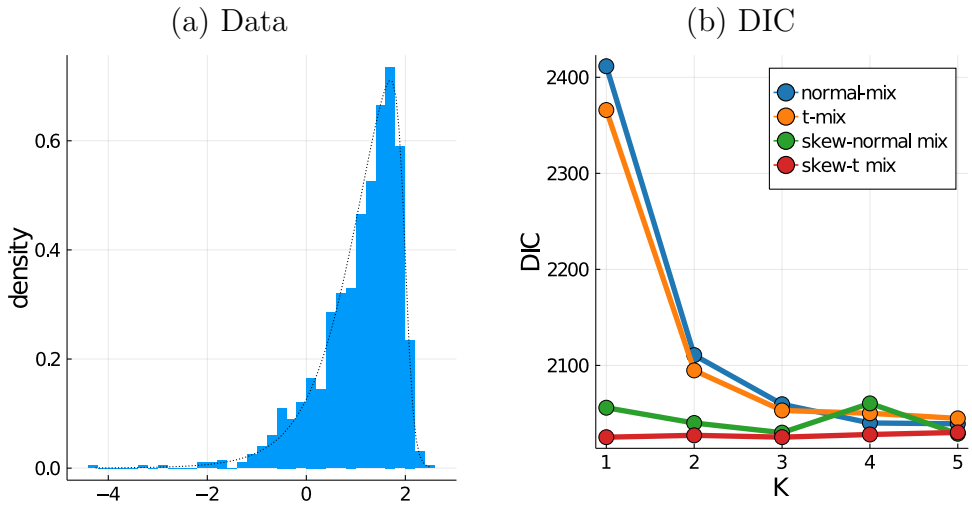


Figure 4.2: (a) Probability density of skew-t(location=2, scale=1, df=7, skew=-10) (dotted line) and histogram of 1000 realizations. (b) DIC of various mixture models (mixtures of Normal, skew-Normal, t, and skew-t) at different K.

1999, MacEachern 1994, Müller et al. 1996). A natural replacement for the model used in Equation (4.2) can be a model that uses dependent DP (DDP) (MacEachern 1999, 2000) priors for the G_i's. In general, dependent DPs are used to model a collection of distributions, where distributions are indexed by covariates such as sample, time, and spatial region. Jointly modeling the distributions through a BNP approach facilitates borrowing of information between groups of distributions more efficiently. Specifically, hierarchical DPs (HDPs) (Teh et al. 2006, Teh and Jordan 2010), which are a type of DDP where multiple DP mixtures are fit simultaneously to different samples with separate sets of weights for the mixtures in each sample, and common atoms for the mixture components across the samples, can be used to model the G_i's. Despite their greater flexibility and ability to estimate the number of required mixture components, the computational complexity for (dependent) DP models can be much greater, especially for large datasets that are commonly encountered in cytometry.

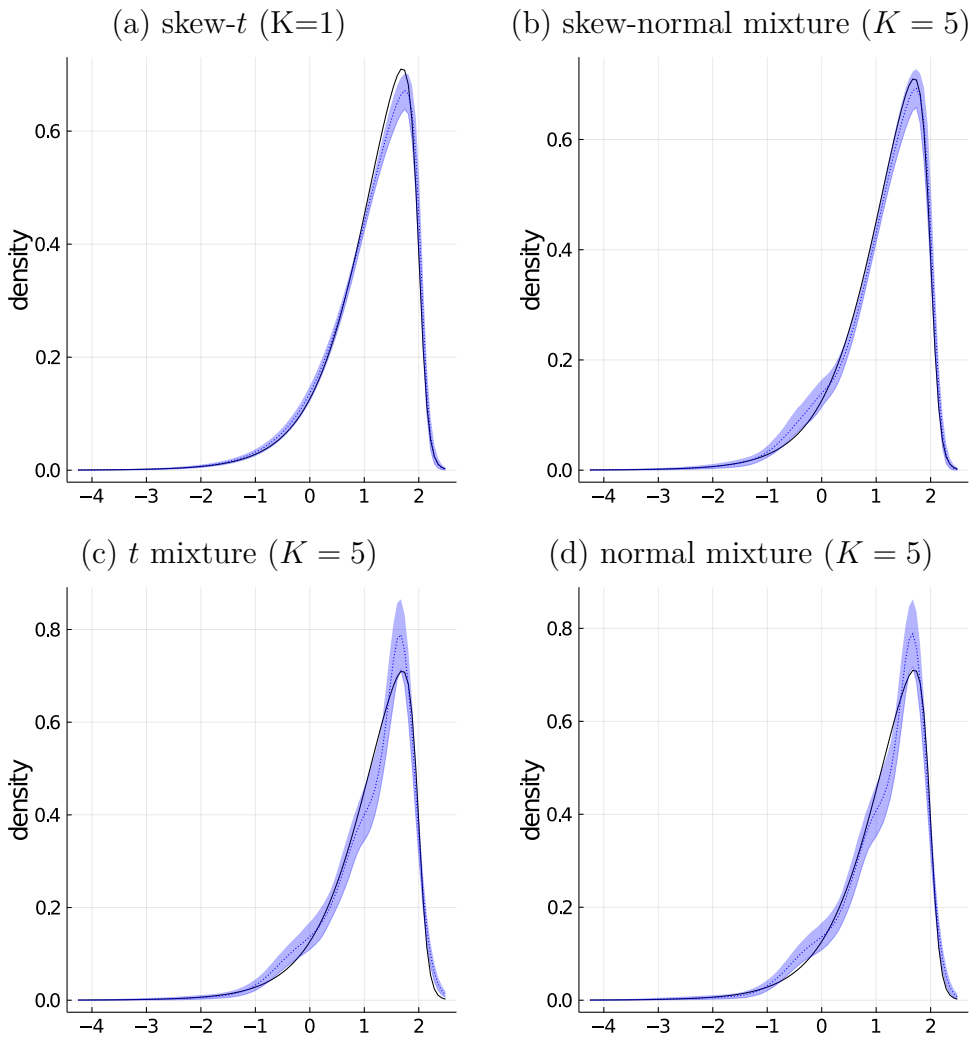


Figure 4.3: 95% pointwise credible intervals for density estimates of simulated skew-t data, under various models. The solid line is the true density. Dotted lines are the pointwise posterior mean density.

Prior Specification As proposed in Frühwirth-Schnatter and Pyne (2010), we introduce auxiliary variables to represent the mixture of skew-t distributions as

follows;

$$
\begin{aligned}
\lambda_{i,n} \mid \boldsymbol{\eta}_i &\sim \text{Categorical}(\boldsymbol{\eta}_i) \\
v_{i,n} \mid \nu_{\lambda_{i,n}} &\sim \text{Gamma}(\nu_{\lambda_{i,n}}/2, \nu_{\lambda_{i,n}}/2) \\
\zeta_{i,n} \mid v_{i,n} &\sim \text{TruncatedNormal}_{[0,\infty)}(0, 1/v_{i,n}) \\
y_{i,n} \mid \mu_{\lambda_{i,n}}, \psi_{\lambda_{i,n}}, \omega_{\lambda_{i,n}}, \zeta_{i,n}, v_{i,n} &\sim \text{Normal}(\mu_{\lambda_{i,n}} + \psi_{\lambda_{i,n}} \cdot \zeta_{i,n},\ \omega_{\lambda_{i,n}}/v_{i,n}),
\end{aligned}
$$

where $\text{Gamma}(s, r)$ denotes the gamma distribution with shape s and rate r, $\text{TruncatedNormal}_{[0,\infty]}(m, v)$ denotes the truncated normal distribution with mean m and variance v before truncation, and $\text{Normal}(m, v)$ denotes the normal distribution with mean m and variance v. After marginalizing over $v_{i,n}$ and $\zeta_{i,n}$, $[y_{i,n} \mid \lambda_{i,n} = k, \mu_k, \sigma_k, \nu_k, \phi_k] \sim \text{skew-}t(\mu_k, \sigma_k, \nu_k, \phi_k)$, where $\phi_k = \psi_k/\sqrt{\omega_k}$ and $\sigma_k^2 = \psi_k^2 + \omega_k$ under the hierarchical representation. For each mixture component, we place priors on $(\mu_k, \omega_k, \nu_k, \psi_k)$ instead of $(\mu_k, \sigma_k, \nu_k, \phi_k)$. We first let, $\mu_k(\boldsymbol{\iota}) = \sum_{\ell=1}^{k} \iota_\ell$ with $\iota_1 \sim \text{Normal}(m_{\iota_1}, s_\iota^2)$ and $\iota_k \overset{iid}{\sim} \text{TruncatedNormal}_{[0,\infty]}(m_{\iota_k}, s_\iota^2)$ for $k = 2, \ldots, K$ for the location. This induces an ordered prior for the μ_k such that $\mu_1 \leq \cdots \leq \mu_K$ so as to avoid identifiability issues common in mixture models. (Celeux et al. 2000, Stephens 2000, Jasra et al. 2005, Frühwirth-Schnatter 2006). We let $\omega_k \overset{iid}{\sim} \text{InverseGamma}(a_\sigma, b_\sigma)$ for the transformed squared scale, $\psi_k \overset{iid}{\sim} \text{Normal}(m_\phi, s_\phi^2)$ for the transformed skew parameter, and $\nu_k \overset{iid}{\sim} \text{LogNormal}(m_\nu, s_\nu)$ for the degrees of freedom. To complete the model specification, we place the priors $\gamma_i \sim \text{Beta}(a_\gamma, b_\gamma)$ and $\boldsymbol{\eta}_i \sim \text{Dirichlet}_K(1/K)$, for $i \in \{1, \ldots, I\}$. Note that γ_i and $\boldsymbol{\eta}_i$ are indexed by i, whereas $(\mu_k, \omega_k, \psi_k, \nu_k)$ are indexed only by k, making mixture components shared across samples. This facilitates borrowing of information across samples and yields a smaller number of used mixture components overall.

Selection of K Thus far, we have assumed a fixed K, which may be difficult to pre-specify. A value of K that is too small will lead to poor model fit, and a value that is too large will lead to high computational cost and unnecessary model complexity. We fit the model with different K within a reasonable range, and select a value for K via model selection. In particular, we select a "best value" for K using a calibration technique that considers the model fit and model complexity. The metric we use for model fit is the DIC; and the metric we use for complexity is the posterior mean number of superfluous mixture components with $\sum_{i=1}^{I} \eta_{i,k} < r$, where $r \in (0, 1)$. That is, $R_r = E\left[\sum_{k=1}^{K} \mathbb{1}\left\{\sum_{i=1}^{I} \eta_{i,k} < r\right\} \mid \text{data}\right]$, which can be estimated from posterior samples of $\eta_{i,k}$. We choose $\hat{K}$ to be the K that achieves the lowest estimated R_r (i.e. $\hat{R}_r$) among the models with the lowest DIC and will demonstrate how to do this graphically in Section 4.3. Often in practical applications of mixture models, small components continue to form as the number of mixture components K increases, while their contributions to model fit decreases. Thus, R_r, which represents the number of superfluous components, is used as a metric for model complexity. For the simulation studies and real data analyses, we use the threshold of $r = 0.01$, which is admittedly arbitrary. Other reasonable thresholds can be used subject to the study of interest. This calibration technique is similarly used by Miller and Dunson (2018) 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. Instead of selecting K via this calibration technique, K can be modeled as a random quantity by placing a prior on K. However, modeling K usually adds computational burden to the problem because the model dimension changes with K. A common transdimensional method is reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995). Alternatively, the unnormalized posterior probabilities $p(y \mid K)p(K)$ can

be used as illustrated in Frühwirth-Schnatter and Pyne (2010) to select K. But evaluating the marginal likelihood $p(y \mid K)$ is unstable for large K, even with the assistance of techniques such as bridge sampling (Meng and Wong 1996). To avoid the added computational burdens of RJMCMC and the numerical instabilities in methods that compute marginal likelihoods in order to compute model posterior probabilities, we opt to select K via calibration.

Posterior Computation Posterior inference for model parameters can be performed by a Metropolis-within-Gibbs algorithm. Each ν_k is updated via a Metropolis step as its full conditional is not available in closed form; while all other model parameters can be updated sequentially by sampling directly from their respective full conditional distributions. Appendix C.1 provides details for posterior simulation, including the full conditional distributions for all model parameters.

Computing Distance between Distributions When two distributions F_i and $F_{i'}$ are not symmetric and/or unimodal, a naive comparison of their means can be misleading (Korthauer et al. 2016). We quantify differential distribution functions F_i and $F_{i'}$, using the distance measure

$$\Delta_{i,i'} \quad = \frac{\int_{\underline{y}}^{\bar{y}} |F_i(y) - F_{i'}(y)|\, dy}{\bar{y} - \underline{y}}, \text{ for } i, i' \in \{1, \ldots, I\}, \tag{4.5}$$

where the interval $(\underline{y}, \bar{y})$ is chosen to cover ranges of y which make up a high probability region within the support. $\Delta_{i,i'}$ takes a value between 0 and 1. A value of $\Delta_{i,i'}$ close to 0 implies that F_i and $F_{i'}$ are similar; similarly if the distribution functions are different, $\Delta_{i,i'}$ has a value close to 1. In our CyTOF data application, $\Delta_{i,i'}$ can be used to identify markers that investigators need to investigate further.

To estimate $\Delta_{i,i'}$, we use pointwise posterior estimates, $\hat{F}_i$ and $\hat{F}_{i'}$, and compute $\hat{\Delta}_{i,i'} = \left(\int_{\underline{y}}^{\bar{y}} \left|\hat{F}_i(y) - \hat{F}_{i'}(y)\right| dy\right) / (\bar{y} - \underline{y})$. We let $\underline{y} = 0$ and $\bar{y}$ be the maximum of the 99-th percentiles of $\hat{F}_i$ and $\hat{F}_{i'}$. This metric provides a one-number summary of the differences between two univariate distributions, while accounting for distributions that may not be unimodal. Other types of statistical divergences were considered instead of Δ. For example, the family of f-divergences (Liese and Vajda 2006, Rényi et al. 1961) which include the Kullback Leibler (KL) divergence (Kullback and Leibler 1951), squared Hellinger distance (Beran et al. 1977), and total variation distance. KL divergence is asymmetric. That is, $D_{KL}(P \mid\mid Q) \neq D_{KL}(Q \mid\mid P)$. More importantly, it has no upper bound, making it difficult to interpret. The squared Hellinger distance is defined only for discrete distributions and absolutely continuous distributions. In our zero-inflated mixture model, the mixture model G_i is chosen to be absolutely continuous. However, the zero-inflated mixture component in F_i is a singular measure. Thus, using it to compare F_i's would be inappropriate. The total variation distance measures the largest possible difference between probabilities that two probability distributions can assign to the same event. Therefore, differences throughout two distributions may be undermined by large differences in probabilities within a particular region of the support of the distributions. For example, if the differences in the γ_i's are sufficiently large, then differences in the G_i's may be seemingly ignored in the total variation statistic. For these reasons, we propose Δ for quantifying distributional differences.

4.3 Simulation Study

Simulation Setup We assessed the performance of our model through the following simulation study. We assumed four different simulation scenarios, scenarios

I-IV, and generated a dataset from each scenario. Table 4.1 contains the simulation truth of the model parameters under the four scenarios. Figure 4.4 shows histograms of the simulated data and their true densities. In scenarios I and II, the true densities G_i^{TR} are visibly different across i, while the true proportions of zeros γ_i^{TR}, are the same in both samples. In scenario III, the G_i^{TR}'s are the same, but the γ_i^{TR}'s differ. In scenario IV, both the G_i^{TR}'s and γ_i^{TR}'s are identical in the simulation truth, resulting in the same F_i^{TR}. In each scenario, we assume $I = 2$ and $N_i = 100000$. We also assume $K^{\text{TR}} = 3$ for scenario I and $K^{\text{TR}} = 4$ for scenarios II-IV. For scenarios I and II, one sample has fewer components than the other. These scenarios were crafted to imitate the real data in section 4.4.

To fit the model, the priors were specified as follows. For a given $K \in \{2, \ldots, 9\}$, $\gamma_i \overset{iid}{\sim} \text{Beta}(1, 1)$, $\boldsymbol{\eta}_i \overset{iid}{\sim} \text{Dirichlet}_K(1/K)$, $\iota_1 \sim \text{Normal}(m_{\iota_1}, s_\iota^2)$, $\iota_k \overset{ind}{\sim} \text{TruncatedNormal}_{[0,\infty)}(m_{\iota_k}, s_\iota^2)$ for $k > 1$, $\tau \sim \text{Gamma}(0.5, 1)$, $\omega_k \mid \tau \overset{iid}{\sim} \text{InverseGamma}(2.5, \tau)$, $\nu_k \overset{iid}{\sim} \text{LogNormal}(3, 0.5)$, and $\psi_k \overset{iid}{\sim} \text{Normal}(-1, 9)$. Note that the prior specifications for τ and ω_k follow recommendations offered by Frühwirth-Schnatter and Pyne (2010). They note that posterior inference for ω_k can be substantially affected by the choice of prior ω_k (without τ) and thus recommend that ω_k be conditioned on τ for added flexibility and improved inference. We empirically specify the hyperparameters for ι_k (m_{ι_k} and s_ι^2) as follows. Let $\tilde{\boldsymbol{y}}$ be the logarithm transformed values of the positive values of $y_{i,n}$, and let q_p denote the p-th percentile of $\tilde{\boldsymbol{y}}$. Then, $m_{\iota_1} = q_{10}$, $m_{\iota_k} = (q_{90} - q_{10})/(K - 1)$, and s_ι is the empirical standard deviation of $\tilde{\boldsymbol{y}}$ divided by K. This prior specification strategy allows ι_1 to have prior mean at the lower quantiles of the data and encourages mixture component centers μ_k to be evenly spaced *a priori*. Posterior simulation is done by MCMC, with a 30000-iteration burn-in, and the next 8000 samples thinned by every other sample collected to yield 4000 samples for

	Scenario I	Scenario II	Scenario III	Scenario IV
K	3	4	4	4
γ_C	0.1	0.1	0.1	0.15
γ_T	0.1	0.1	0.2	0.15
$\boldsymbol{\eta}_C$	(0.25,0.75,0)	(0.1,0.1,0.5,0.3)	(0.05,0.05,0.5,0.4)	(0.05,0.05,0.5,0.4)
$\boldsymbol{\eta}_T$	(0.1,0.1,0.8)	(0.1,0.1,0.8,0)	(0.05,0.05,0.5,0.4)	(0.05,0.05,0.5,0.4)
$\boldsymbol{\mu}$	(-1.5,3.5,5.1,5)	(-1.5,3.5,1.5,4.3)	(-1.5,3.5,5.1,4.3)	(-1.5,3.5,5.1,4.3)
$\boldsymbol{\sigma}$	(1.6,1.76,1.76,1.6)	(1.6,1.76,1.76,1.6)	(1.6,1.76,1.76,1.6)	(1.6,1.76,1.76,1.6)
$\boldsymbol{\nu}$	(12,10,10,15)	(12,10,10,15)	(12,10,10,15)	(12,10,10,15)
$\boldsymbol{\phi}$	(0,-10,-10,0)	(12,10,10,-11)	(0,-10,-10,-11)	(0,-10,-10,-11)

Table 4.1: Simulation truth of model parameters under four simulated scenarios.

posterior inference. For $K = 2$, the inference speed was approximately 1 iteration per second; while for $K = 9$, the inference speed was approximately 0.5 iterations per second. All computations for this chapter were done on an interactive Linux server with four Intel Xeon E5-4650 processors (64 cores total) and 512 GB of random access memory. Algorithms for posterior inference were implemented[1] in the Julia programming language (Bezanson et al. 2017).

Comparison to Normal Mixture Models For comparison, we also used normal mixture models for $\tilde{G}_i$ to yield a zero-inflated mixture of normal distributions;

$$y_{i,n} \mid \gamma_i, G_i \sim \gamma_i \cdot \delta_0(y_{i,n}) + (1 - \gamma_i) \cdot G_i(y_{i,n}),$$

where $G_i = \sum_{k=1}^{K} \eta_{i,k} \cdot \text{Normal}(\mu_k, \omega_k)$. Prior specifications for model parameters, including $\gamma_i, \boldsymbol{\eta}_i, \mu_k$, and ω_k, remain unchanged from the inflated skew-t mixture model.

Simulation Results We obtained $\hat{K}$ for each scenario under the normal and skew-t mixture models using the calibration technique from Section 4.2. $\hat{K}$ is

[1]Source code for this project can be found at `https://github.com/luiarthur/CytofDiffDensity.jl`.

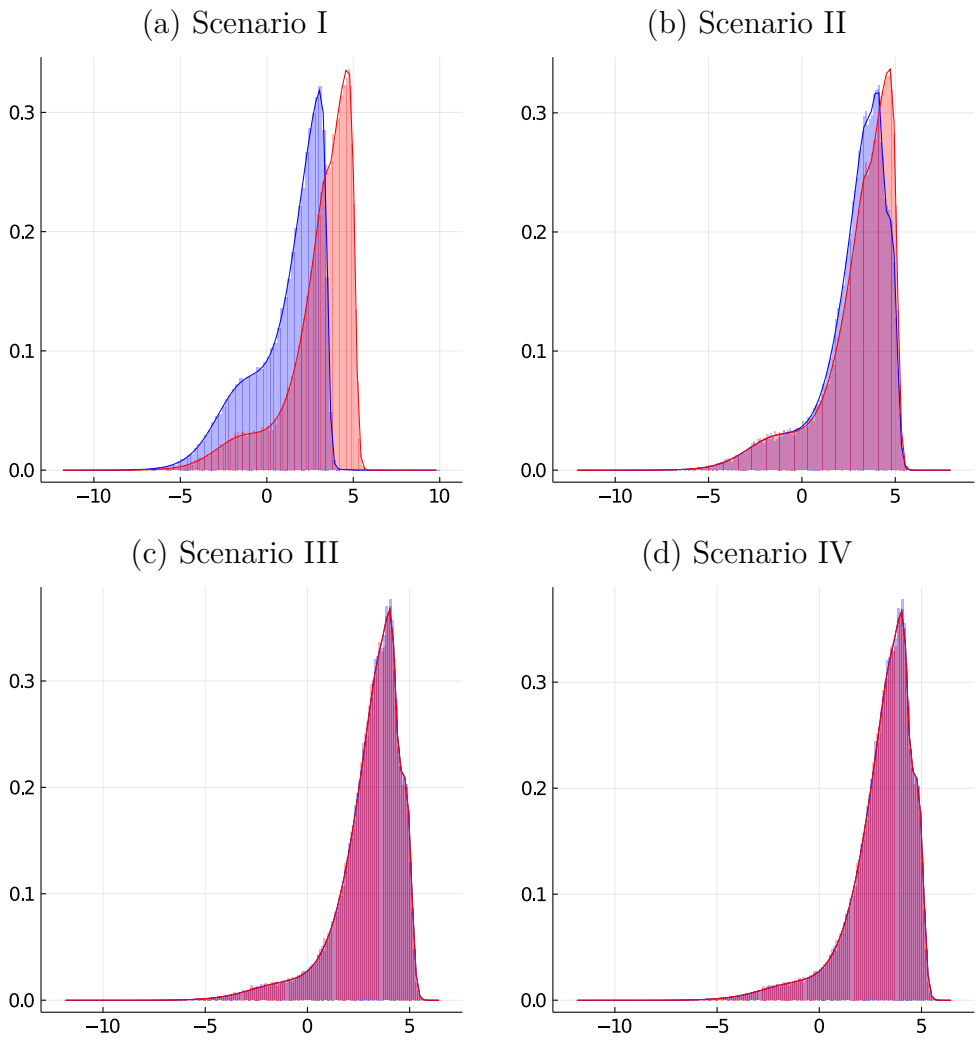


Figure 4.4: Histograms of logarithm of non-zero values of $y_{i,n}$ in the simulated data where blue and red represent samples 1 and 2, respectively. The density of the simulation truths are depicted by solid lines.

listed in Table 4.2. Figure 4.5 plots the DIC against $\hat{R}_{1\%}$ for $K = 2, \ldots, 9$. Among models with the lowest DIC, we select the model with the lowest $\hat{R}_{1\%}$ (which is usually where $\hat{R}_{1\%} = 0$). Thus, models that appear in the left bottom corner of each graph are chosen. Under the skew-t mixture, $\hat{K}$ for scenario I is equal to the simulation truth ($K^{\text{TR}} = 3$). For scenario II, $\hat{K} = 5$ is greater than $K^{\text{TR}} = 4$ by one. For scenarios III and IV, $\hat{K} = 6$ is greater than $K^{\text{TR}} = 4$ by two. Despite

	skew-t mixture			Normal mixture		
Scenario	$\hat{K}$	DIC	$\hat{\Delta}$	$\hat{K}$	DIC	$\hat{\Delta}$
I	3	672720	0.162	8	673176	0.161
II	5	660863	0.036	8	660942	0.036
III	6	578299	0.02	6	578051	0.02
IV	6	578486	0.0	6	578439	0.0

Table 4.2: [Simulation study] $\hat{\Delta}$, and DIC for best models (selected by the calibration method discussed) under various scenarios and models (skew-t or normal mixture).

the discrepancies, fewer components are required to adequately model the data using skew-t mixtures, as opposed to Gaussian mixtures in scenarios I and II, as shown in the Table 4.2. Under normal mixtures, 8 components are required to adequately model the data. Also, the normal mixture model with $\hat{K} = 8$ yields a greater DIC than the skew-t mixture with $\hat{K} = 3$ for scenario I and $\hat{K} = 5$ for scenario II. For scenarios III and IV, $\hat{K} = 6$ for the two mixture models and the normal mixture model yields smaller DIC. This may be due to the presence of fewer outliers and modes in the simulated data. From these results, we see that occasionally, normal mixtures can fit data as well as skew-t counterparts. However, for skewed data with outliers, normal mixtures may require more components to rival skew-t mixtures.

Figures 4.6-4.7 compare posterior estimates of the density $\tilde{g}_i$ (of $\tilde{G}_i$) to their truth under the model with $K = \hat{K}$. Figure 4.6 contains the results for the skew-t mixture model, while Figure 4.7 contains the results for the normal mixture model. From Figure 4.6(a) and Figure 4.7(a), better model fit under the skew-t model can be observed for scenario I, where the estimate of $\tilde{g}_i$ more closely follows the simulation truth in Figure 4.6(a). Table 4.3 includes posterior summaries of γ_i for each scenario. Posterior means and 95% credible intervals for γ_i are include for each i. The intervals contain the simulation truth in all cases. To quantify the

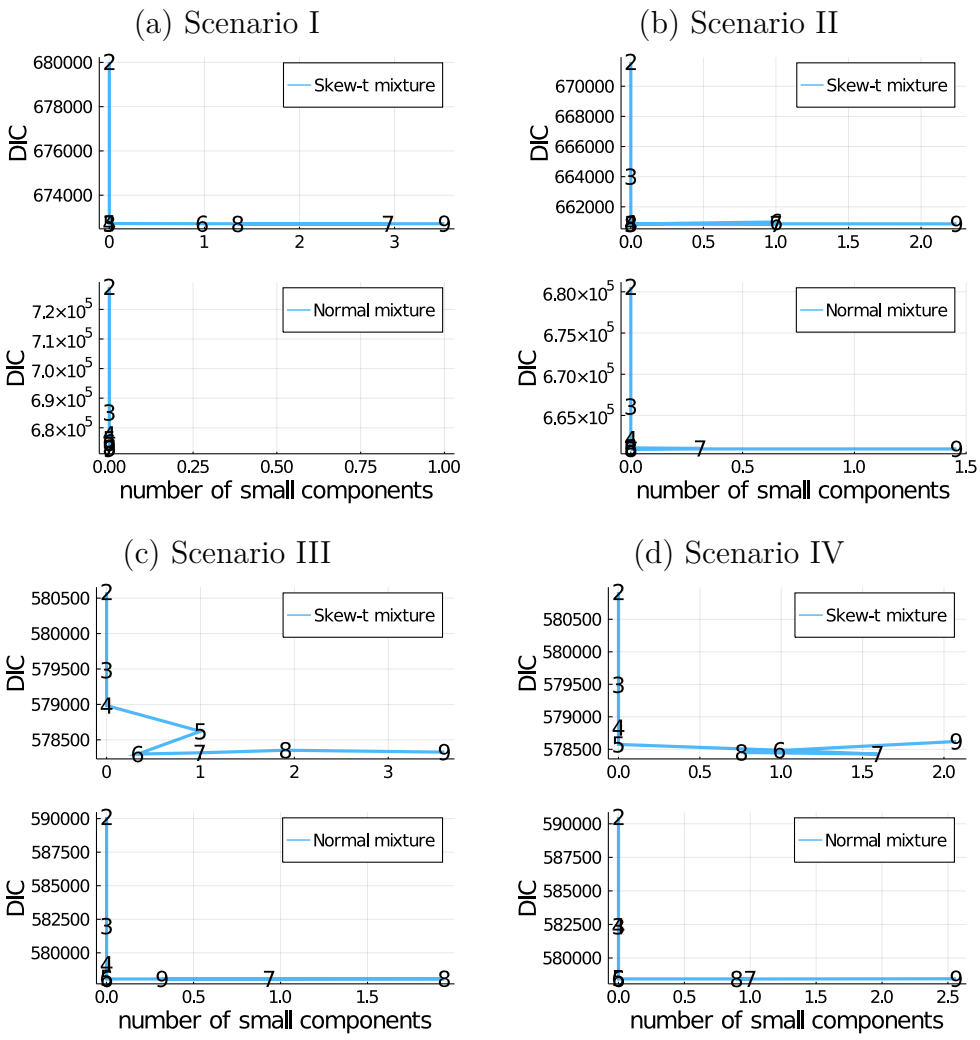


Figure 4.5: [Simulation Study] Plot of DIC against the number of small components, $\hat{R}_{1\%}$. The values that have low DIC and $\hat{R}_{1\%}$ are chosen. Each panel has two plots; the top for the zero-inflated skew-t mixture models, and the bottom for the zero-inflated normal mixture models.

difference between distributions across the two samples, we estimated Δ for each scenario. Table 4.2 lists $\hat{\Delta}$ under both models. In each scenario, $\hat{\Delta}$ is similar under the skew-t and normal mixtures. For scenario IV, where the two F_i's are identical in the simulation truth, $\hat{\Delta}$ is less than 0.001; whereas in scenario I, where the $\tilde{g}_i$ are visibly different, $\hat{\Delta}$ is 0.162. Scenario III is similar to scenario IV, except that

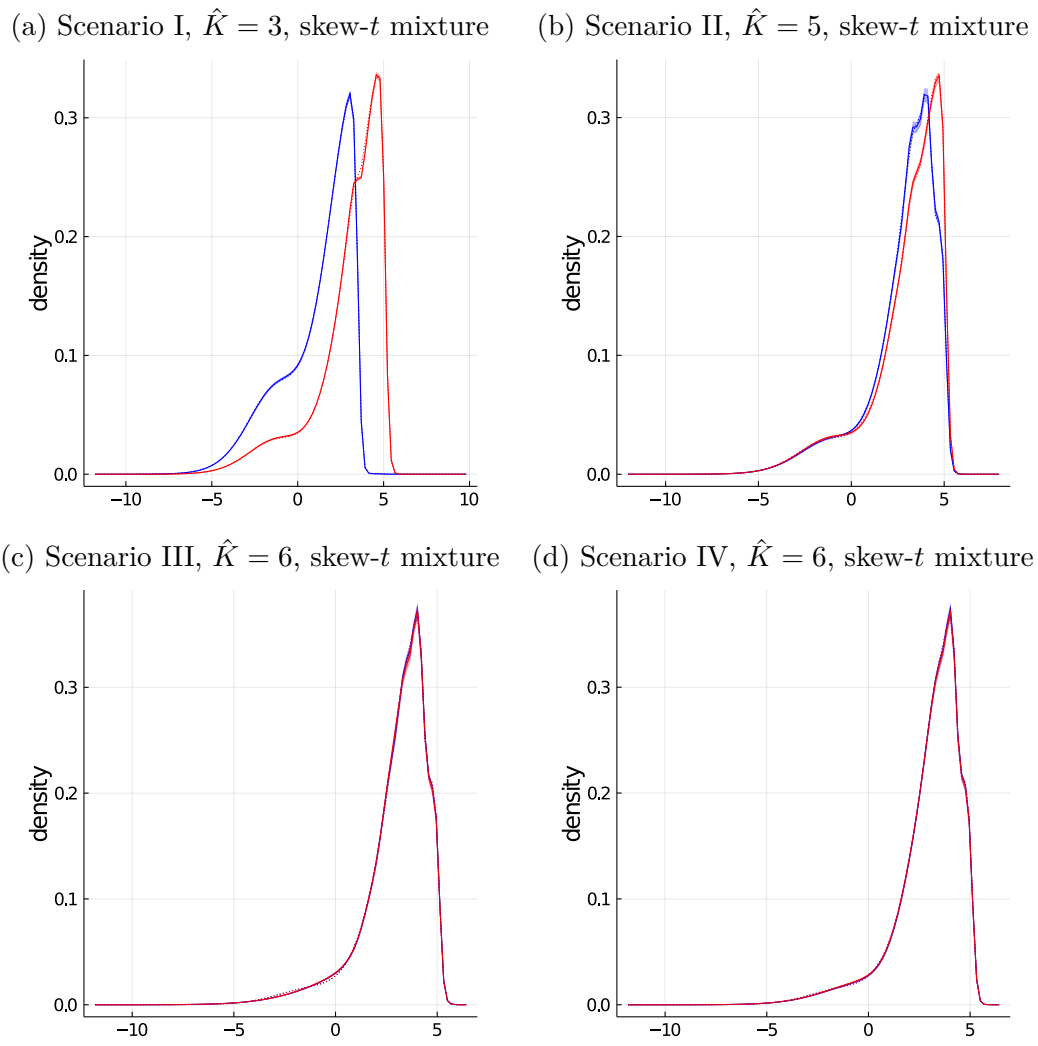
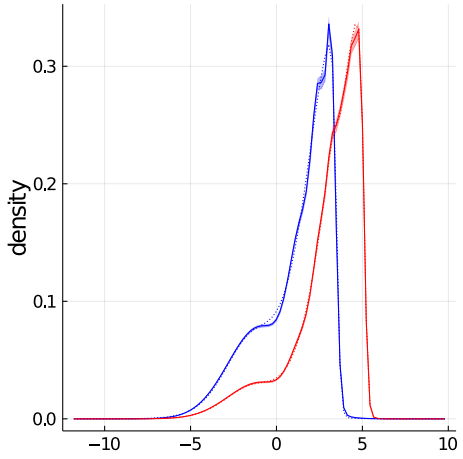


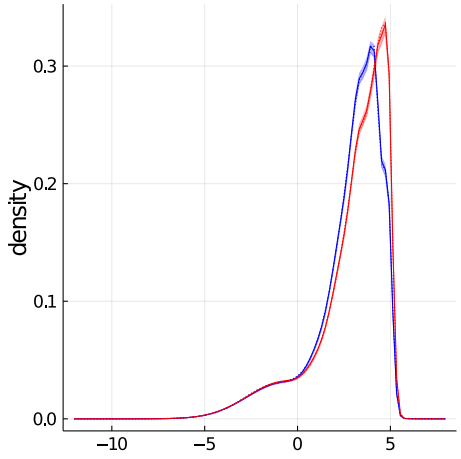
Figure 4.6: Posterior estimates of $\tilde{g}_i$ under the skew-t mixture model for $K = \hat{K}$ in the simulation study. The blue and red curves are the posterior densities of $i = 1$ and $i = 2$, respectively. The solid and dashed lines are the posterior means and simulation truths, respectively. The shaded regions are 95% credible intervals.

the γ_i's are different. Thus, $\hat{\Delta}$ is 0.02. In scenario II, the $\tilde{g}_i$'s are again visibly different, but much less so than in scenario I. This is manifested in $\hat{\Delta}$ for scenario II being 0.036, which is smaller than that in scenario I.

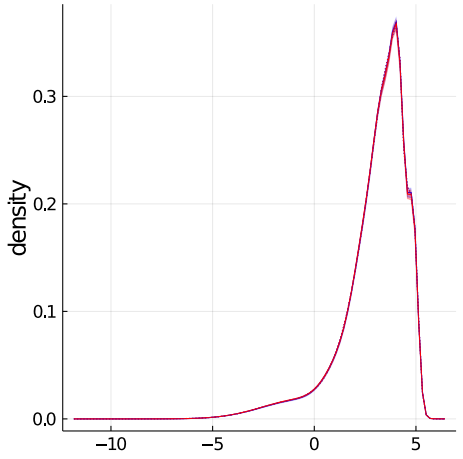
(a) Scenario I, $K = 8$, Normal mixture



(b) Scenario II, $K = 8$, Normal mixture



(c) Scenario III, $K = 6$, Normal mixture



(d) Scenario IV, $K = 6$, Normal mixture

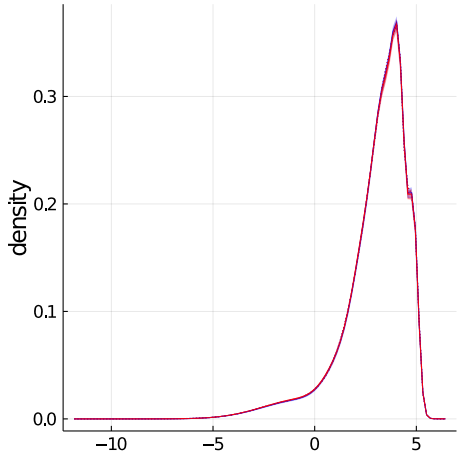


Figure 4.7: Posterior estimates of $\tilde{g}_i$ under the normal mixture model for $K = \hat{K}$ in the simulation study. The blue and red curves are the posterior densities of $i = 1$ and $i = 2$, respectively. The solid and dashed lines are the posterior means and simulation truths, respectively. The shaded regions are 95% credible intervals.

4.4 Analysis of CyTOF Data

In this section, we present real data analyses of marker expression data produced by CyTOF. NK cells harnessed from a healthy individual were expanded ex vivo and exposed to the immunosuppressive, TGF-β. CyTOF was then used to profile the impact of TGF-β on NK cell functions. In particular, the dataset con-

Scenario	γ_1^{TR}	$\hat{\gamma}_1$ (95% CI)	γ_2^{TR}	$\hat{\gamma}_2$ (95% CI)
I	0.10	0.1004 (0.0986, 0.1023)	0.10	0.0987 (0.0968, 0.1005)
II	0.10	0.1004 (0.0986, 0.1023)	0.10	0.0987 (0.0968, 0.1005)
III	0.10	0.1004 (0.0986, 0.1023)	0.20	0.1996 (0.1971, 0.2021)
IV	0.15	0.1508 (0.1486, 0.1530)	0.15	0.1482 (0.1460, 0.1504)

Table 4.3: Posterior summary of γ_i for the various simulated scenarios. Second and fourth columns contain simulation truth for γ_i. Third and fifth columns contain posterior mean ($\hat{\gamma}_i$) (and 95% credible intervals) for $i = 1$ and $i = 2$, respectively.

sists of the two samples of NK cells – one before and one after TGF-β exposure. We let $i = 1$ and 2 denote the samples before and after the exposure, respectively. Expression levels of cell surface markers were measured from individual cells in the samples by CyTOF. TGF-β may alter expression levels for some of the markers. Furthermore, the population of NK cells is heterogeneous, and TGF-β may influence only some subsets of NK cells. Those changes may yield differential distributions in the expression levels. The data includes expression levels of a total of 38 markers. Also, the samples contain $N_i = 86915$ and $N_2 = 92468$ cells. For the illustrations of the model, we present results for four selected markers, CD3z, EOMES, Granzyme A, and Siglec7. As previously described, CyTOF data occasionally includes a substantial number of zeros due to some experimental artifacts. For example, marker EOMES expression levels are zero in 9.63% and 7.97% of the cells in samples 1 and 2, respectively; while that of marker Granzyme A are less than 0.1% for each sample. The fractions of zeros (Z_i) in each sample (i) for each marker are listed in Table 4.4. The effects of TGF-β on the distribution of marker expression is of primary interest in this study. Thus, our distance metric for marker expression distributions will be used to quantify differences in distribution before and after treatment. Zero-inflated skew-t and normal mixtures were fit for each marker, for each $K \in \{2, 3, \ldots, 9\}$. Prior distributions for model

Marker	Z_1	$\hat{\gamma}_1$ (95% CI)	Z_2	$\hat{\gamma}_2$ (95% CI)
CD3z	0.0157	0.0157 (0.0149, 0.0165)	0.0044	0.0045 (0.0040, 0.0049)
EOMES	0.0963	0.0964 (0.0944, 0.0983)	0.0797	0.0797 (0.0780, 0.0815)
Granzyme A	0.0003	0.0003 (0.0002, 0.0005)	0.0009	0.0009 (0.0007, 0.0011)
Siglec7	0.0663	0.0663 (0.0646, 0.0679)	0.0498	0.0498 (0.0484, 0.0512)

Table 4.4: Posterior summary of γ_i for four NK cell markers. Second and fourth columns contain empirical fractions of zeros, Z_i, in sample (i). Third and fifth columns contain posterior mean $(\hat{\gamma}_i)$ (and 95% credible intervals) for $i = 1$ and $i = 2$, respectively. Number of cells in donor sample before (N_C) and after (N_T) treatment are 86915 and 92468, respectively.

	skew-t mixture			Normal mixture		
Marker	$\hat{K}$	DIC	$\hat{\Delta}$	$\hat{K}$	DIC	$\hat{\Delta}$
CD3z	5	488192	0.121	8	488198	0.12
EOMES	3	510144	0.103	8	510111	0.103
Granzyme A	4	438170	0.015	7	438176	0.015
Siglec7	5	527162	0.028	7	527151	0.028

Table 4.5: $\hat{\Delta}$, and DIC for best models (selected by the calibration method discussed) for various markers and models (skew-t or normal mixture), for the CyTOF data analysis.

parameters were specified in a manner similar to that in section 4.3, except that $\tau \sim \text{Gamma}(1, 1)$, to encourage σ to be further away from 0. Posterior inference was made via MCMC, with the first 40000 iterations discarded as burn in. The subsequent 8000 samples were thinned by every other sample to yield 4000 samples for inference.

Table 4.4 additionally shows posterior summaries for γ_i, including the posterior means $\hat{\gamma}_i$ and 95% credible intervals. For each marker, with the exception of marker Granzyme A, the estimated proportion of zeros $(\hat{\gamma}_i)$ is substantially greater in the sample prior to treatment $(i = 1)$. Figure 4.8 shows, for each marker, the DIC against $\hat{R}_{1\%}$ for $K = 2, \ldots, 9$ under the skew-t and normal mixture models. For each marker and model, $\hat{K}$ was selected using the calibration method in section 4.2, and are included in Table 4.5, along with the DIC for those models.

Under both mixture models, the DICs are similar for the model with $K = \hat{K}$. The selected values of K are noticeably smaller for the skew-t mixture model. For markers CD3z and Granzyme A, $\hat{K}$ is smaller by 3 components under the skew-t model; for marker EOMES, $\hat{K}$ is smaller by 5 components; and for marker Siglec7, $\hat{K}$ is smaller by 2 components. This indicates that in a normal mixture model with large K, many potentially superfluous components tend to form. $\hat{\Delta}$ is also illustrated in Table 4.5 and is similar for the skew-t and normal mixtures. $\hat{\Delta}$ for markers Granzyme A and Siglec7 is less than 0.03, indicating that the impact on the expression levels' distribution by TGF-β is small. $\hat{\Delta}$ for markers CD3z and EOMES are between 0.10 and 0.15, indicating that expression levels are more affected by TGF-β. This suggests further investigation on those markers. Figures 4.9 and 4.10 show the posterior estimates of $\tilde{g}_i$ for each marker under skew-t and normal mixtures, for their $\hat{K}$. The blue and red curves are for $i = 1$ and $i = 2$ respectively. The histograms of the data are overlaid in grey. As can be seen from the figures, the model fits the data well for the selected K. Of note, a moderate proportion of cells have higher expression levels for markers CD3z and EOMES after TGB-β exposure; whereas for markers Granzyme A and Siglec7, a small proportion of cells have lower expression levels after treatment. The fit between the two mixture models is similar. But, as an important illustration, $\hat{K}$ tends to be greater under the normal mixture models. For example, $K = 3$ under the normal mixture would lead to poor fit for EOMES due to the skewness in both samples and the bimodality in sample 2.

4.5 Discussion

We have proposed a method for modeling marker expression levels obtained from CyTOF via a zero-inflated skew-t mixture model. The skew-t mixture is

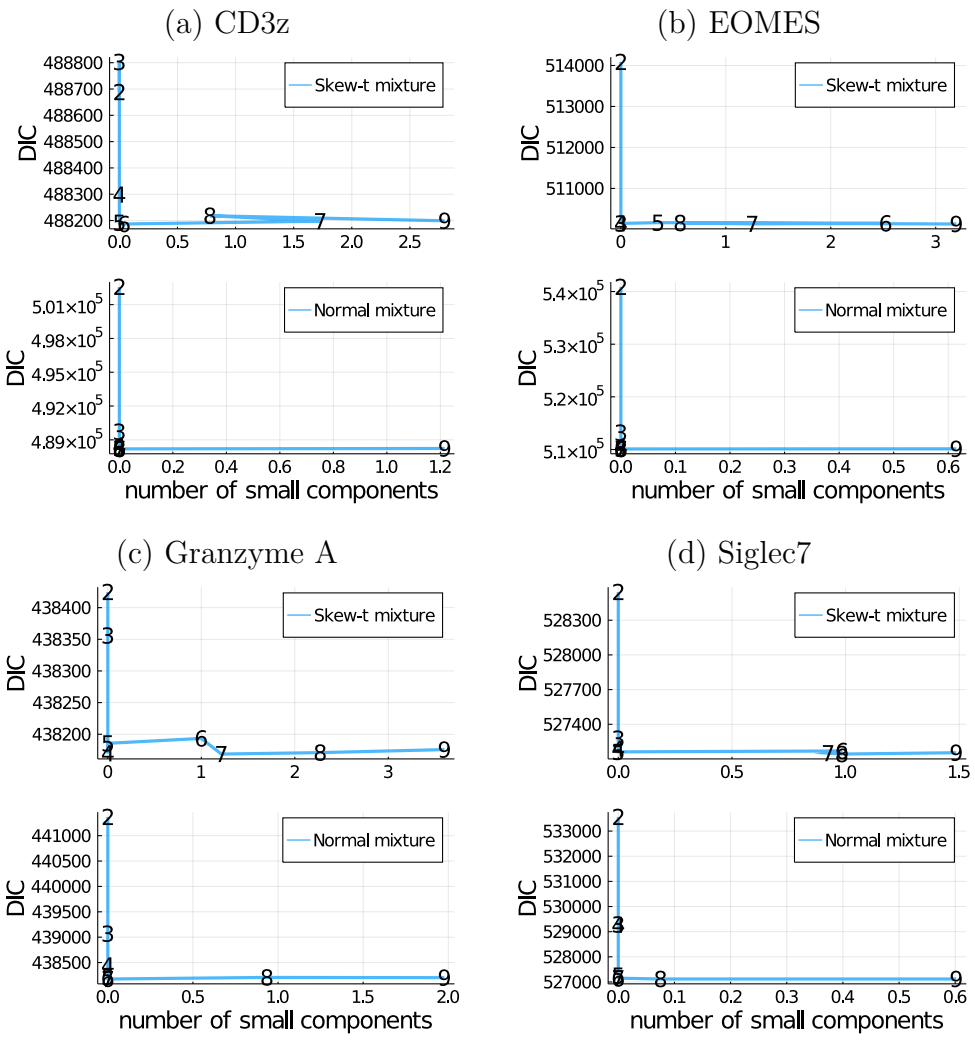


Figure 4.8: Plots of DIC against the $\hat{R}_{1\%}$ for various markers and various K, used for selecting K, for the CyTOF data analysis. Plots for the skew-t and normal mixtures are included.

able to model skewed distributions with outliers using fewer mixture components than Gaussian mixtures. We provide a calibration method to select the number of mixture components, and a metric for quantifying the difference between distributions due to experimental conditions. We demonstrated the performance of our method through simulation studies and applied our method to CyTOF data for four NK cell markers. While our examples include two experimental conditions

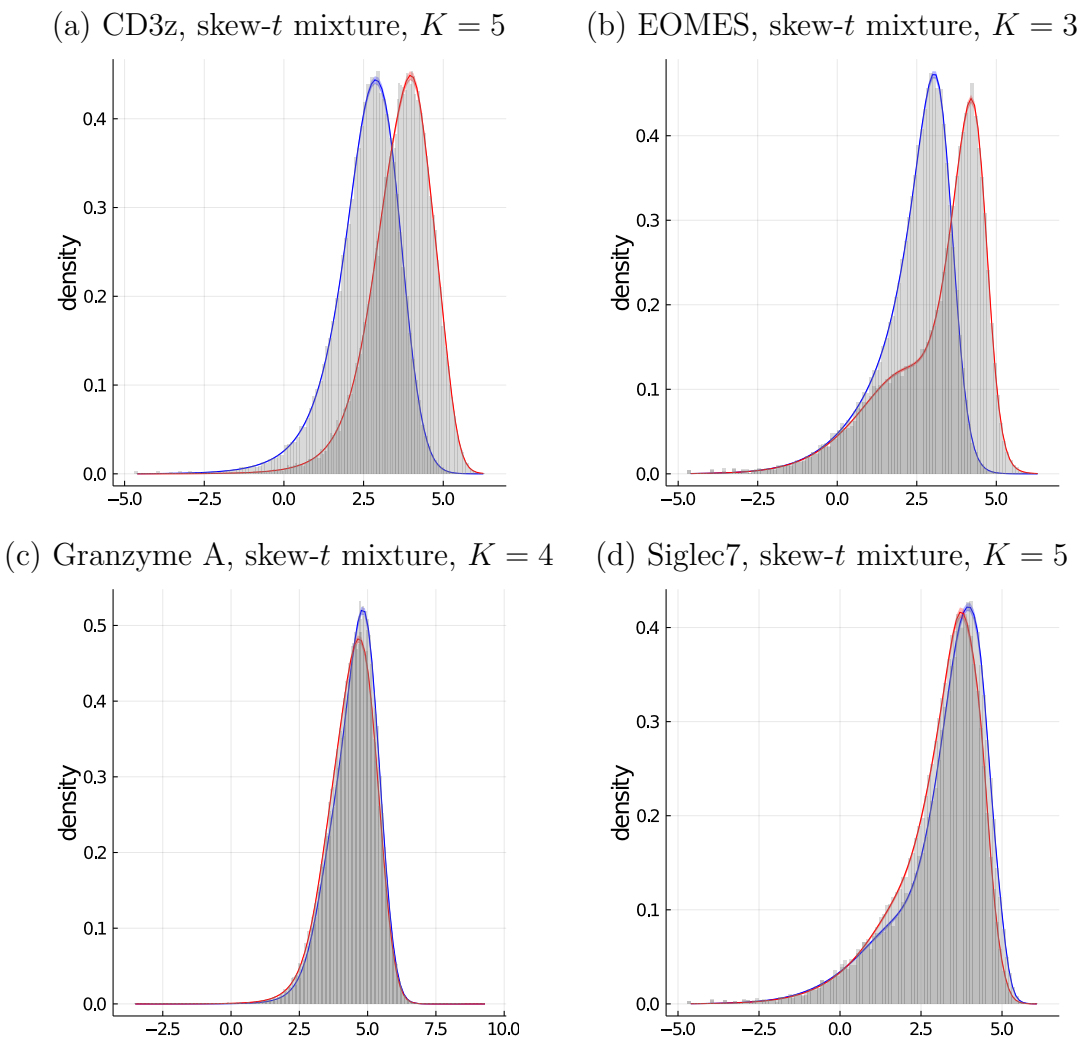


Figure 4.9: Estimates of density of $\tilde{G}_i$ (blue for $\tilde{G}_1$ and red for $\tilde{G}_2$) for skew-t mixtures, for the CyTOF data analysis. Histogram of data in grey.

at a time, it can be applied to data with multiple condition.

In this chapter, we considered one marker at a time. A natural extension is to analyze multiple markers jointly using a multivariate zero-inflated mixture of skew-t distributions and quantifying differences between the multivariate distributions. For example, a multivariate skew-t distribution can be used as a mixture component; assume $z_{i,n,j} \mid \gamma_{i,j} \overset{ind}{\sim} \text{Bernoulli}(\gamma_{i,j})$ with $\lambda_{i,n} \mid \boldsymbol{\eta}_i \sim \text{Categorical}(\boldsymbol{\eta}_i)$, where $z_{i,n,j} = \mathbb{1}\{y_{i,n,j} = 0\}$ indicates whether the expression level is 0 in cell n of

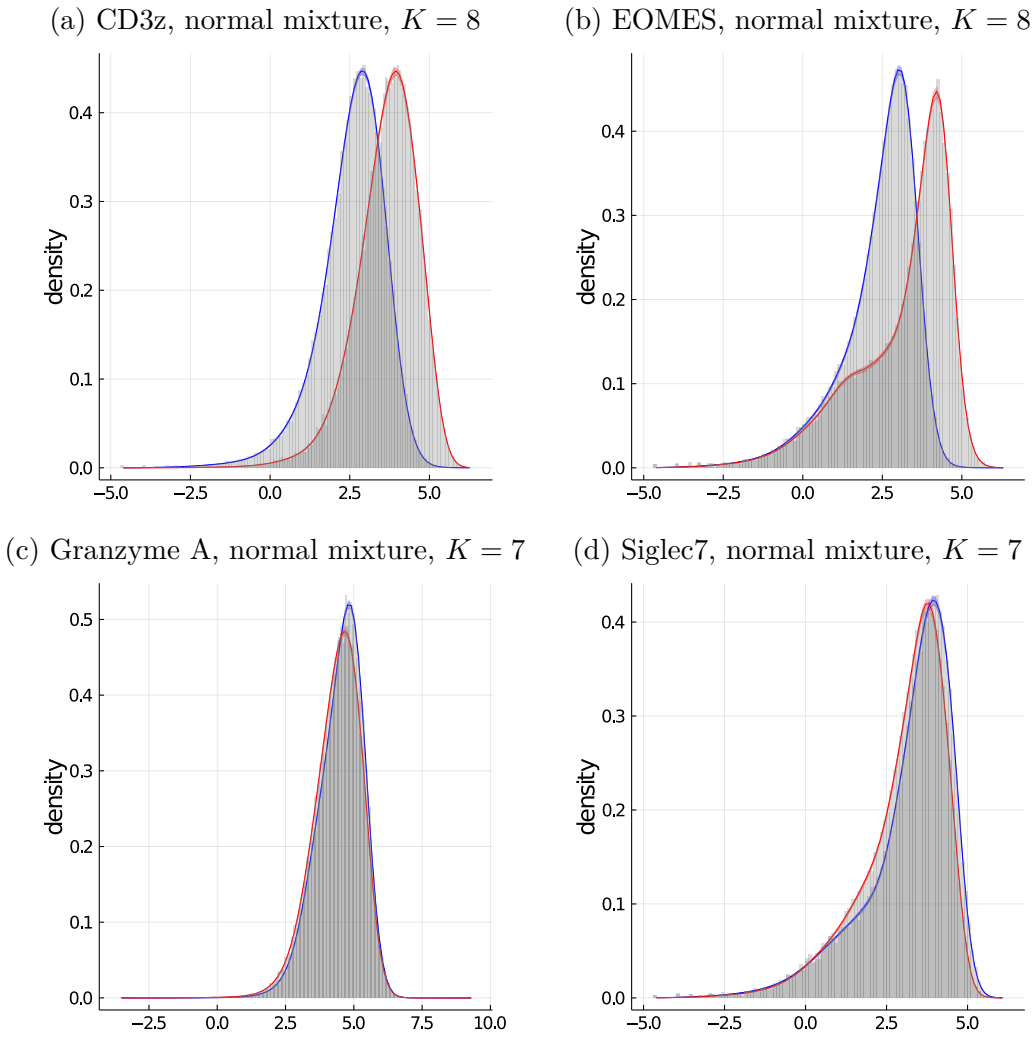


Figure 4.10: Estimates of density of $\tilde{G}_i$ (blue for $\tilde{G}_1$ and red for $\tilde{G}_2$) for normal mixtures, for the CyTOF data analysis. Histogram of data in grey.

sample i for marker j. Given $\lambda_{i,n} = k$, we let $y_{i,n,j} = \exp(\mu_{j,k} + \psi_{j,k} \cdot \zeta_{i,n} + \epsilon_{i,n,j})$, if $z_{i,n,j} = 1$, and otherwise $y_{i,n,j} = 0$, where $y_{i,n,j}$ is the expression level in cell n for marker j (of J markers) in sample i, and $\boldsymbol{\epsilon}_{i,n} \mid \boldsymbol{\Omega}_k, v_{i,n}, \lambda_{i,n} = k \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}_k / v_{i,n})$, where Ω_k is a $J \times J$ covariance matrix which captures the correlation in expression levels between markers. Under this approach, mixture components are shared across samples to facilitate borrowing of information across all markers. Frühwirth-Schnatter and Pyne (2010) have discussed in depth how

to implement a posterior simulation scheme for multivariate skew-t mixture models but without zero-inflation. However, in the above zero-inflated model, zeros and positive values may occur in $\boldsymbol{y}_{i,n}$, which complicates posterior computation. Careful considerations will be required to develop an efficient sampling scheme. Differences between distributions from different experimental conditions may be quantified for each marker or for a subset of the markers using an estimate of the joint cdf and an appropriate scaling factor to ensure the metric is between 0 and 1.

Chapter 5

Conclusion

This work presented Bayesian methods for the analysis of NK cell marker expression data obtained from CyTOF. Novel strategies for identifying (in particular, NK) cell subpopulations from CyTOF data were proposed. A zero-inflated mixture model was introduced to model zero-inflated, semicontinuous CyTOF data and a method to quantify differences between distributions that result from various experimental conditions was presented.

Chapter 2 presented a novel method for identifying cell subpopulations from multiple samples of CyTOF marker expression level data. Subpopulations were characterized by latent expression patterns which were modeled via a FAM, which induces clusterings that form subpopulations. Zeros throughout the marker expression data represent unlikely expression of certain markers within cells. They were treated as missing and imputed via a static data missingship mechanism to account for uncertainty and facilitate inference. Compared to established clustering methods, the proposed FAM is more effective at discovering latent subpopulations when the underlying subpopulations are similar. Biologically recognized NK cell subpopulations were identified by the FAM and identification of novel subpopulations is possible.

Chapter 3 built upon the work of Chapter 2 and presents a novel rep-FAM which encourages latent subpopulations, i.e. columns of the feature allocation matrix, to be distinct. The degree to which columns are different can be calibrated via hyperparameters in the repulsive function and expert knowledge. WPPT and intrinsic MCMC were used to more efficiently sample from the joint posterior distribution of the model parameters. Feature allocations resulting from the rep-FAM were more parsimonious than those resulting form the regular FAM.

Chapter 4 presented methods for jointly modeling marker expression CyTOF data due to multiple experimental conditions. A zero-inflated mixture of log-skew-t distributions were used to model the expressions levels. A metric was proposed to quantify distributional differences between pairs of experimental conditions.

A calibration technique for selecting the number of features in the FAMs and the number of skew-t mixture components was presented and used throughout Chapters 2-4. This was done to circumvent learning the number of components via transdimensional MCMC methods, such as RJMCMC, which would add computational complexity to an already complex problem. The core idea common to each chapter is to select the model with the least complexity (according the number of superfluous or negligible components) among models with the best fit (according to some goodness-of-fit metric).

One way to extend the FAMs is to include covariate information into the models. Samples with similar covariates may have similar subpopulation structures. The proposed FAMs can incorporate such information by incorporating appropriate regression submodels to enhance inferences and study how the structures may change with covariates.

In Chapter 4, analysis of multiple markers simultaneously is a possible exten-

sion using a multivariate zero-inflated mixture of skew-t distributions. Computational challenges are likely to arise as a result of the added complexity. Thus, careful considerations will be required to ensure computational tractability since CyTOF datasets tend to be large, with up to 40 markers and marker expression levels for tens of thousands of cells.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Paul D Allison. *Missing data*, volume 136. Sage publications, 2001.

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003.

Ole Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 353(1674):401–419, 1977.

Matthew James Beal et al. *Variational algorithms for approximate Bayesian inference*. university of London London, 2003.

Rudolf Beran et al. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

James O Berger and Luis R Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL `https://doi.org/10.1137/141000671`.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Ryan P Browne and Paul D McNicholas. A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2):176–198, 2015.

M Carlsten and M Jaras. Natural killer cells in myeloid malignancies: Immune surveillance, nk cell dysfunction, and pharmacological ppportunities to bolster the endogenous nk cells. *Front Immunol*, 10:2357, 2019.

Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

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, 2016a.

Haoyu Chen, Daniel Seita, Xinlei Pan, and John Canny. An efficient minibatch acceptance test for metropolis-hastings. *arXiv preprint arXiv:1610.06848*, 2016b.

Mengjie Chen, Chao Gao, and Hongyu Zhao. Phylogenetic indian buffet process: Theory and applications in integrative analysis of cancer genomics. *arXiv preprint arXiv:1307.8229*, 2013.

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.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

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

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.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Thomas S Ferguson et al. Prior distributions on spaces of probability measures. *Annals of Statistics*, 2(4):615–629, 1974.

Brian C Franczak, Ryan P Browne, and Paul D McNicholas. Mixtures of shifted asymmetriclaplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1149–1157, 2013.

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.

Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

Sylvia Frühwirth-Schnatter and Saumyadipta Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336, 2010.

Mauro Gasparini. Bayesian density estimation via dirichlet density processes. *Journal of Nonparametric Statistics*, 6(4):355–366, 1996.

Seymour Geisser and William F Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

Alan E Gelfand, DK Dey, and Hong Chang. Bayesian statistics. *Bernardo, JM*, pages 147–159, 1992.

Samuel J Gershman, Peter I Frazier, and David M Blei. Distance dependent infinite latent feature models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):334–345, 2014.

Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482, 2006.

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.

P Le Hai-son and Ziv Bar-Joseph. Inferring interaction networks using the ibp applied to microrna target prediction. In *Advances in Neural Information Processing Systems*, pages 235–243, 2011.

Timothy E Hanson, Athanasios Kottas, and Adam J Branscum. Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian nonparametric approaches. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(2):207–225, 2008.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

M Ilander, U Olsson-Stromberg, H Schlums, and et al. Increased proportion of mature nk cells is associated with successful imatinib discontinuation in chronic myeloid leukemia. *Leukemia*, 31(5):1106–1116, 2017.

Mike Innes. Flux: Elegant machine learning with julia. *Journal of Open Source Software*, 2018. doi: 10.21105/joss.00602.

Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

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

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233, 1999.

Dimitris Karlis and Anais Santourian. Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 19(1):73–83, 2009.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17(1): 222, 2016.

Athanasios Kottas. Bayesian semiparametric modeling for stochastic precedence, with applications in epidemiology and survival analysis. *Lifetime data analysis*, 17(1): 135–155, 2011.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. URL `http://jmlr.org/papers/v18/16-107.html`.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Lewis L Lanier. Up on the tightrope: natural killer cell activation and inhibition. *Nature immunology*, 9(5):495, 2008.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 853–877, 2015.

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.

Dangna Li and Wing H Wong. Mini-batch tempered mcmc. *arXiv preprint arXiv:1707.09705*, 2017.

L Li, H Cheen, D Marin, and et al. A novel immature natural killer cell subpopulation predicts relapse after cord blood transplantation. *Blood Adv*, 3(23):4117–4130, 2019.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

E Liu, D Marin, P Banerjee, H Macapinlac, PF Thall, K Rezvani, and et al. Il-15 armored car-transduced nk cells against cd19 positive b cell tumors. *New England Journal of Medicine*, 382(1):545–553, 2020.

Albert Y Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357, 1984.

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.

Arthur Lui, Juhee Lee, Peter F Thall, May Daher, Katy Rezvani, and Rafet Basar. A bayesian feature allocation model for identification of cell subpopulations using cytometry data. *arXiv preprint arXiv:2002.08609*, 2020.

Kyle B Lupo and Sandro Matosevic. Natural killer cells as allogeneic effectors in adoptive cancer immunotherapy. *Cancers*, 11(6):769, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

SN MacEachern. Dependent dirichlet processes. ohio state university, department of statistics. Technical report, Technical report, 2000.

Steven N MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

Steven N MacEachern. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999, 1999.

Mehrnoush Malek, Mohammad Jafar Taghiyar, Lauren Chong, Greg Finak, Raphael Gottardo, and Ryan R Brinkman. flowdensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, 31(4):606–607, 2014.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Jeffrey S Miller, Yvette Soignier, Angela Panoskaltsis-Mortari, Sarah A McNearney, Gong H Yun, Susan K Fautsch, David McKenna, Chap Le, Todd E Defor, Linda J Burns, et al. Successful adoptive transfer and in vivo expansion of human haploidentical nk cells in patients with cancer. *Blood*, 105(8):3051–3057, 2005.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13, 2018.

Kurt T Miller, Thomas Griffiths, and Michael I Jordan. The phylogenetic indian buffet process: A non-exchangeable nonparametric prior for latent features. *arXiv preprint arXiv:1206.3279*, 2012.

Peter Müller, Alaattin Erkanli, and Mike West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996.

Yang Ni, Peter Müller, and Yuan Ji. Bayesian double feature allocation for phenotyping with electronic health records. *Journal of the American Statistical Association*, pages 1–15, 2019.

Georgios Papoutsoglou, Vincenzo Lagani, Angelika Schmidt, Konstantinos Tsirlis, David-Gómez Cabrero, Jesper Tegnér, and Ioannis Tsamardinos. Challenges in the multivariate analysis of mass cytometry data: the effect of randomization. *Cytometry Part A*, 95(11):1178–1190, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

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

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

José J Quinlan, Garritt L Page, and Fernando A Quintana. Density regression using repulsive distributions. *Journal of Statistical Computation and Simulation*, 88(15): 2931–2947, 2018.

Carl Edward Rasmussen et al. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

Stefano Regis, Alessandra Dondero, Fabio Caliendo, Cristina Bottino, and Roberta Castriconi. Nk cell function regulation by tgf-β-induced epigenetic mechanisms. *Frontiers in Immunology*, 11, 2020.

Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

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

Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

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

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

A Sarvaria, D Jawdat, JA Madrigal, and A Saudemont. Umbilical cord blood natural killer cells, their characteristics, and potential clinical applications. *Front Immunol*, 8:329, 2017.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016. URL `https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf`.

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.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Nina Shah, Li Li, Jessica McCarty, Indreshpal Kaur, Eric Yvon, Hila Shaim, Muharrem Muftuoglu, Enli Liu, Robert Z Orlowski, Laurence Cooper, et al. 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, 2017.

Karen Slattery and Clair M Gardiner. Nk cell metabolism and tgfβ–implications for immunotherapy. *Frontiers in immunology*, 10:2915, 2019.

Jacopo Soriano, Li Ma, et al. Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis*, 14(1):161–180, 2019.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

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

Xi Tan, Vinayak A Rao, and Jennifer Neville. The indian buffet hawkes process to model evolving latent influences. In *UAI*, pages 795–804, 2018.

Nicholas G Tawn, Gareth O Roberts, and Jeffrey S Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, 30(1):27–41, 2020.

Stan Development Team et al. Stan modeling language users guide and reference manual. *Technical report*, 2016.

Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207, 2010.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

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

Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

Sofie Van Gassen, Britt Callebaut, and Yvan Saeys. *FlowSOM: Using self-organizing maps for visualization and interpretation cytometry data*, 2017. http://www.r-project.org, http://dambi.ugent.be.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305, 2008.

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.

Sinead A Williamson, Michael Minyi Zhang, and Paul Damien. A new class of time dependent latent factor models with applications. *arXiv preprint arXiv:1904.08548*, 2019.

Sinead A Williamson, Michael Minyi Zhang, and Paul Damien. A new class of time dependent latent factor models with applications. *Journal of Machine Learning Research*, 21(27):1–24, 2020.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Jun Wu and Lewis L Lanier. Natural killer cells and cancer. *Advances in cancer research*, 90(1):127–56, 2003.

Fangzheng Xie and Yanxun Xu. Bayesian repulsive gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203, 2020.

Yanxun Xu, Juhee Lee, Yuan Yuan, Riten Mitra, Shoudan Liang, Peter Müller, and Yuan Ji. Nonparametric bayesian bi-clustering for next generation sequencing count data. *Bayesian analysis (Online)*, 8(4):759, 2013.

Yanxun Xu, Peter Müller, Yuan Yuan, Kamalakar Gulukota, and Yuan Ji. Mad bayes for tumor heterogeneity—feature 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.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Appendix A

A Bayesian Feature Allocation Model for Identifying Cell Subpopulations Using Cytometry Data

A.1 Posterior Computation

A.1.1 MCMC Simulation

Recall that $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \boldsymbol{\sigma}^2, \boldsymbol{\eta}^0, \boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{\epsilon}, \alpha\}$ denotes all random parameters. We let expression levels $\boldsymbol{y}$ and binary indicators $\boldsymbol{m}$ denote $y_{i,n,j}$ and $m_{i,n,j}$, respectively, for all (i, n, j). To facilitate the posterior sampling of $\delta_{z,\ell}$, we introduce auxiliary indicators for normal mixture components $\gamma_{i,n,j} \in \{1, \ldots, L_{z_{j,\lambda_{i,n}}}\}$ for each $y_{i,n,j}$ when $\lambda_{i,n} \neq 0$. That is, $\Pr(\gamma_{i,n,j} = \ell \mid z_{j,\lambda_{i,n}} = z, \eta^z_{i,j,\ell}, \lambda_{i,n} \neq 0) = \eta^z_{i,j,\ell}$, where $\ell \in \{1, \ldots, L_{z_{j,\lambda_{i,n}}}\}$, and let $\mu_{i,n,j} = \mu^{\star}_{z_{j,\lambda_{i,n}},\gamma_{i,n,j}}$. We extend the vector of

random parameters, $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \{\gamma_{i,n,j}\})$ by including $\gamma_{i,n,j}$ for more convenient posterior simulation. Similar to the joint posterior distribution of $\boldsymbol{\theta}$ in (2.6) of the main text, the joint posterior probability model of $\widetilde{\boldsymbol{\theta}}$ under our Bayesian FAM model is

$$\begin{aligned} p(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{y}, \boldsymbol{m}, K) \;\; \propto \;\; & p(\widetilde{\boldsymbol{\theta}} \mid K) \times \\ & \prod_{i,n} \left[\prod_j \rho_{i,n,j}^{1-m_{i,n,j}} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{(y_{i,n,j} - \mu_{i,n,j})^2}{2\sigma_i^2} \right\} \right]^{1(\lambda_{i,n} \neq 0)} \times \\ & \left[\prod_j \rho_{i,n,j}^{1-m_{i,n,j}} \frac{1}{\sqrt{2\pi s_\epsilon^2}} \exp\left\{ -\frac{y_{i,n,j}^2}{2s_\epsilon^2} \right\} \right]^{1(\lambda_{i,n}=0)} . \end{aligned} \tag{A.1}$$

Posterior samples of $\widetilde{\boldsymbol{\theta}}$ are obtained by iteratively drawing samples from each of the full conditionals using the most recent estimate of the parameters and the data. For the parameters whose conditional distributions are known and are easy to sample from, we used Gibbs sampling. To sample from full conditionals which are otherwise difficult to sample from, the Metropolis-Hastings algorithm was used.

1. Full Conditional for v_k

   Recall that the prior distribution for v_k is $v_k \mid \alpha \overset{ind}{\sim} \text{Beta}(\alpha/K, 1)$, for $k = 1, ..., K$, that is, $p(v_k \mid \alpha) = \frac{\alpha}{K} v_k^{\alpha/K - 1}$.

$$\begin{aligned} p(v_k \mid \boldsymbol{y}, \text{rest}) &\propto p(v_k) \prod_{j=1}^{J} p(z_{j,k} \mid v_k) \\ &\propto \frac{\alpha}{K} v_k^{\alpha/K-1} \prod_{j=1}^{J} v_k^{z_{j,k}} (1 - v_k)^{1 - z_{j,k}} \\ &\propto v_k^{\alpha/K + \sum_{j=1}^{J} z_{j,k} - 1} (1 - v_k)^{J - \sum_{j=1}^{J} z_{j,k}} \end{aligned}$$

$$\Rightarrow v_k \mid \boldsymbol{y}, \text{rest} \sim \text{Be}\left(\alpha/K + \sum_{j=1}^{J} z_{j,k},\;\; J + 1 - \sum_{j=1}^{J} z_{j,k} \right).$$

We use "rest" to denote all parameters except the parameter(s) that we sample. For example, "rest" implies $\widetilde{\boldsymbol{\theta}} \backslash \{v_k\}$ for updating v_k.

2. Full Conditional for $z_{j,k}$

   Let $S_k = \{(i,n) : \lambda_{i,n} = k\}$, the set of cells in samples taking cell subpopulation k.

$$
\begin{aligned}
p(z_{j,k} = 1 \mid \boldsymbol{y}, \text{rest}) &\propto p(z_{j,k} = 1 \mid v_k) \prod_{(i,n) \in S_k} p(y_{i,n,j} \mid \boldsymbol{\mu}_1^\star, \boldsymbol{\eta}_{i,j}^1, \sigma_i^2) \\
&\propto v_k \prod_{(i,n) \in S_k} \sum_{\ell=1}^{L} \eta_{i,j,\ell}^1 \cdot \phi(y_{i,n,j} \mid \mu_{1,\ell}^\star, \sigma_i^2), \\
p(z_{j,k} = 0 \mid \boldsymbol{y}, \text{rest}) &\propto p(z_{j,k} = 0 \mid v_k) \prod_{(i,n) \in S_k} p(y_{i,n,j} \mid \boldsymbol{\mu}_0^\star, \boldsymbol{\eta}_{i,j}^0, \sigma_i^2) \\
&\propto (1 - v_k) \prod_{(i,n) \in S_k} \sum_{\ell=1}^{L} \eta_{i,j,\ell}^0 \cdot \phi(y_{i,n,j} \mid \mu_{0,\ell}^\star, \sigma_i^2),
\end{aligned}
$$

   where $\phi(y \mid m, s^2)$ denotes the probability density function of the normal distribution with mean m and variance s^2, evaluated at y.

$$
\begin{aligned}
\Rightarrow z_{j,k} \mid \boldsymbol{y}, \text{rest} &\sim \text{Bernoulli}(p_{j,k}), \text{ where} \\
p_{j,k} &= \left[1 + \frac{(1 - v_k) \prod_{(i,n) \in S_k} \sum_{\ell=1}^{L} \eta_{i,j,\ell}^0 \cdot \phi(y_{i,n,j} \mid \mu_{0,\ell}^\star, \sigma_i^2)}{v_k \prod_{(i,n) \in S_k} \sum_{\ell=1}^{L} \eta_{i,j,\ell}^1 \cdot \phi(y_{i,n,j} \mid \mu_{1,\ell}^\star, \sigma_i^2)} \right]^{-1}.
\end{aligned}
$$

3. Full Conditional for α

$$
\begin{aligned}
p(\alpha \mid \boldsymbol{y}, \text{rest}) &\propto p(\alpha) \times \prod_{k=1}^{K} p(v_k \mid \alpha) \\
&\propto \alpha^{a_\alpha - 1} \exp\{-b_\alpha \alpha\} \times \prod_{k=1}^{K} \alpha \, v_k^{\alpha/K} \\
&\propto \alpha^{a_\alpha + K - 1} \exp\left\{ -\alpha \left(b_\alpha - \sum_{k=1}^{K} \log v_k / K \right) \right\}
\end{aligned}
$$

$$\Rightarrow \alpha \mid \boldsymbol{y}, \text{rest} \sim \text{Gamma}\left(a_\alpha + K,\ b_\alpha - \sum_{k=1}^{K} \log v_k / K\right).$$

4. Full Conditional for $\lambda_{i,n}$

   The prior for $\lambda_{i,n}$ is

$$p(\lambda_{i,n} = k \mid \boldsymbol{w}_i, \epsilon_i) = \begin{cases} \epsilon_i, & \text{if } k = 0 \\ (1 - \epsilon_i) \cdot w_{i,k}, & \text{if } k \in \{1, \ldots, K\}. \end{cases}$$

   We thus have

$$\begin{aligned} p(\lambda_{i,n} = 0 \mid \boldsymbol{y}, \text{rest}) &\propto p(\lambda_{i,n} = 0)\ p(\boldsymbol{y} \mid \lambda_{i,n} = 0, \text{rest}) \\ &\propto \epsilon_i \prod_{j=1}^{J} \phi(y_{i,n,j} \mid 0, s_\epsilon^2), \\ p(\lambda_{i,n} = k \mid \boldsymbol{y}, \text{rest}) &\propto p(\lambda_{i,n} = k)\ p(\boldsymbol{y} \mid \lambda_{i,n} = k, \text{rest}) \\ &\propto (1 - \epsilon_i) w_{ik} \prod_{j=1}^{J} \left(\sum_{\ell=1}^{L} \eta_{i,j,\ell}^{z_{j,k}} \cdot \phi(y_{i,n,j} \mid \mu^\star_{z_{j,k},\ell}, \sigma_i^2) \right), \\ &\qquad \text{for } k = 1, \ldots, K. \end{aligned}$$

   We sample $\lambda_{i,n}$ with probabilities proportional to $p(\lambda_{i,n} = k \mid \boldsymbol{y}, \text{rest})$ for $k \in \{0, \ldots, K\}$.

5. Full Conditional for $\boldsymbol{w}_i$

   The prior for $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,K})$ is $\boldsymbol{w}_i \sim \text{Dirichlet}(d/K, \cdots, d/K)$. The

full conditional for $\boldsymbol{w}_i$ is:

$$\begin{aligned} p(\boldsymbol{w}_i \mid \text{rest}) \propto\ & p(\boldsymbol{w}_i) \times \prod_{n=1}^{N_i} p(\lambda_{i,n} \mid \boldsymbol{w}_i) \\ \propto\ & \prod_{k=1}^{K} w_{ik}^{\left(d/K + \sum_{n=1}^{N_i} 1(\lambda_{i,n}=k)\right)-1}. \end{aligned}$$

Therefore,

$$\boldsymbol{w}_i \mid \boldsymbol{y}, \text{rest} \ \sim\ \text{Dirichlet}\left(d/K + \sum_{n=1}^{N_i} 1(\lambda_{i,n}=1), \ldots, d/K + \sum_{n=1}^{N_i} 1(\lambda_{i,n}=K)\right)$$

6. Full Conditional for $\gamma_{i,n,j}$

   For the cells with $\lambda_{i,n} > 0$,

$$\begin{aligned} p(\gamma_{i,n,j} = \ell \mid \boldsymbol{y}, z_{j\lambda_{i,n}} = z, \text{rest}) &\propto p(\gamma_{i,n,j} = \ell) \times p(y_{i,n,j} \mid \gamma_{i,n,j} = \ell, \text{rest}) \\ &= \eta^{z}_{ij\ell} \times \phi(y_{i,n,j} \mid \mu^{\star}_{z\ell}, \sigma_i^2). \end{aligned}$$

   Therefore, sample $\gamma_{i,n,j}$ with probabilities proportional to $p(\gamma_{i,n,j} = \ell \mid \boldsymbol{y}, \text{rest})$ for $\ell = 1, ..., L^{z_{j,\lambda_{i,n}}}$.

7. Full Conditional for $\delta_{z,\ell}$

   For $\delta_{1,\ell}$, let $S_{1,i,\ell} = \left\{(i,n,j) : \left(z_{j,\lambda_{i,n}} = 1 \ \cap\ \gamma_{i,n,j} \geq \ell\right)\right\}$ and $|S_{1,i,\ell}|$ the car-

dinality of $S_{1,i,\ell}$.

$$
\begin{aligned}
p(\delta_{1,\ell} \mid \boldsymbol{y}, \text{rest}) &\propto p(\delta_{1,\ell} \mid \psi_1, \tau_1^2) \times p(\boldsymbol{y} \mid \delta_{1,\ell}, \text{rest}) \\
&\propto 1(\delta_{1,\ell} \geq 0) \times \exp\left\{\frac{-(\delta_{1,\ell} - \psi_1)^2}{2\tau_1^2}\right\} \\
&\qquad \times \prod_{i=1}^{I} \prod_{(i,n,j) \in S_{1i\ell}} \exp\left\{-\left(y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} \delta_{1r}\right)^2 \Big/ 2\sigma_i^2\right\} \\
&\propto \exp\left\{-\frac{(\delta_{1,\ell})^2}{2}\left(\frac{1}{\tau_1^2} + \sum_{i=1}^{I} \frac{|S_{1,i,\ell}|}{\sigma_i^2}\right) + \delta_{1,\ell}\left(\frac{\psi_1}{\tau_1^2} + \sum_{i=1}^{I} \sum_{S_{1,i,\ell}} \frac{g_{i,n,j}}{\sigma_i^2}\right)\right\} \\
&\qquad \times 1(\delta_{1,i,\ell} \geq 0),
\end{aligned}
$$

where $g_{i,n,j} = y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} (\delta_{1,r})^{1(r \neq \ell)}$.

$$
\Rightarrow \delta_{1,\ell} \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim} \text{TN}^{+}\left(\frac{\psi_1 + \tau_1^2 \sum_{i=1}^{I} \sum_{S_{1,i,\ell}} (g_{i,n,j}/\sigma_i^2)}{1 + \tau_1^2 \sum_{i=1}^{I} (|S_{1,i,\ell}|/\sigma_i^2)}, \frac{\tau_1^2}{1 + \tau_1^2 \sum_{i=1}^{I} (|S_{1,i,\ell}|/\sigma_i^2)}\right).
$$

Similarly, for $\delta_{0,\ell}$, let $S_{0,i,\ell} = \left\{(i,n,j) : \left(Z_{j,\lambda_{i,n}} = 0 \ \cap \ \gamma_{i,n,j} \geq \ell\right)\right\}$ and $|S_{0,i,\ell}|$ be the cardinality of $S_{0,i,\ell}$.

$$
\Rightarrow \delta_{0,l} \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim} \text{TN}^{+}\left(\frac{\psi_0 + \tau_0^2 \sum_{i=1}^{I} \sum_{S_{1,i,\ell}} (g_{i,n,j}/\sigma_i^2)}{1 + \tau_0^2 \sum_{i=1}^{I} (|S_{0,i,\ell}|/\sigma_i^2)}, \frac{\tau_0^2}{1 + \tau_0^2 \sum_{i=1}^{I} (|S_{0,i,\ell}|/\sigma_i^2)}\right),
$$

where $g_{i,n,j} = -y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} (\delta_{0,r})^{1(r \neq \ell)}$.

8. Full Conditional for σ_i^2

Let $r_{i,n,j} = 1(\lambda_{i,n} > 0)$, and let $R_i = \sum_{n=1}^{N_i} \sum_{j=1}^{J} r_{i,n,j}$. We then have

$$\begin{aligned}
p(\sigma_i^2 \mid \boldsymbol{y}, \text{rest}) &\propto p(\sigma_i^2) \times p(\boldsymbol{y} \mid \sigma_i^2, \text{rest}) \\
&\propto (\sigma_i^2)^{-a_\sigma - 1} \exp\left\{-\frac{b_\sigma}{\sigma_i^2}\right\} \prod_{j=1}^{J} \prod_{n=1}^{N_i} \left\{\frac{1}{\sqrt{2\sigma_i^2}} \exp\left\{\frac{-(y_{i,n,j} - \mu_{i,n,j})^2}{2\sigma_i^2}\right\}\right\} \\
&\propto (\sigma_i^2)^{-\left(a_\sigma + \frac{R_i}{2}\right) - 1} \exp\left\{-\frac{1}{\sigma_i^2}\left(b_\sigma + \sum_{j=1}^{J}\sum_{n=1}^{N_i} r_{i,n,j} \cdot \frac{(y_{i,n,j} - \mu_{i,n,j})^2}{2}\right)\right\}.
\end{aligned}$$

$$\Rightarrow \sigma_i^2 \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim} \text{InverseGamma}\left(a_\sigma + \frac{R_i}{2}, \quad b_\sigma + \sum_{j=1}^{J}\sum_{n=1}^{N_i} r_{i,n,j} \cdot \frac{(y_{i,n,j} - \mu_{i,n,j})^2}{2}\right).$$

9. Full Conditional for $\eta_{i,j}^z$

The prior for $\boldsymbol{\eta}_{i,j}^z$ is $\boldsymbol{\eta}_{i,j}^z \sim \text{Dirichlet}_{L_z}(a_{\eta^z})$, for $z \in \{0, 1\}$. So the full conditional for $\boldsymbol{\eta}_{i,j}^z$ is:

$$\begin{aligned}
p(\boldsymbol{\eta}_{i,j}^z \mid \text{rest}) \propto \;& p(\boldsymbol{\eta}_{i,j}^z) \times \prod_{n=1}^{N_i} p(\gamma_{i,n,j} \mid \boldsymbol{\eta}_{i,j}^z) \\
\propto \;& \prod_{\ell=1}^{L_z} \left(\eta_{i,j,\ell}^z\right)^{a_{\eta^z} - 1} \times \prod_{\ell=1}^{L_z} \prod_{n=1}^{N_i} \left(\eta_{i,j,\ell}^z\right)^{1\{(\gamma_{i,n,j} = \ell) \;\&\; (z_{j,\lambda_{i,n}} = z) \;\&\; (\lambda_{i,n} > 0)\}} \\
\propto \;& \prod_{\ell=1}^{L_z} \left(\eta_{i,j,\ell}^z\right)^{\left(a_{\eta^z} + \sum_{n=1}^{N_i} 1\{(\gamma_{i,n,j} = \ell) \;\&\; (z_{j,\lambda_{i,n}} = z) \;\&\; (\lambda_{i,n} > 0)\}\right) - 1}.
\end{aligned}$$

$$\Rightarrow \boldsymbol{\eta}_{i,j}^z \mid \boldsymbol{y}, \text{rest} \;\sim\; \text{Dirichlet}_{L_z}\left(a_1^*, ..., a_{L_z}^*\right),$$

where $a_\ell^* = a_{\eta^z} + \sum_{n=1}^{N_i} 1\{(\gamma_{i,n,j} = \ell) \;\&\; (z_{j,\lambda_{i,n}} = z) \;\&\; (\lambda_{i,n} > 0)\}$.

10. Full Conditional for ϵ_i

$$
\begin{aligned}
p(\epsilon_i \mid y, \text{rest}) &\propto p(\epsilon_i) \prod_{n=1}^{N_i} \epsilon_i^{1(\lambda_{i,n}=0)} (1-\epsilon_i)^{1(\lambda_{i,n}>0)} \\
&\propto \epsilon_i^{a_\epsilon - 1} (1-\epsilon_i)^{b_\epsilon - 1} \epsilon_i^{\sum_{n=1}^{N_i} 1(\lambda_{i,n}=0)} (1-\epsilon_i)^{\sum_{n=1}^{N_i} 1(\lambda_{i,n}>0)} \\
&\propto \epsilon_i^{a_\epsilon + \sum_{n=1}^{N_i} 1(\lambda_{i,n}=0) - 1} (1-\epsilon_i)^{b_\epsilon + \sum_{n=1}^{N_i} 1(\lambda_{i,n}>0) - 1}.
\end{aligned}
$$

$$
\Rightarrow \epsilon_i \mid y, \text{rest} \sim \text{Beta}\left(a_\epsilon + \sum_{n=1}^{N_i} 1(\lambda_{i,n}=0), b_\epsilon + \sum_{n=1}^{N_i} 1(\lambda_{i,n}>0) \right).
$$

11. Full Conditional for Missing $y_{i,n,j}$

$$
\begin{aligned}
p(y_{i,n,j} \mid m_{i,n,j} = 1, \text{rest}) &\propto p(m_{i,n,j} = 1 \mid y_{i,n,j}, \text{rest})\ p(y_{i,n,j} \mid \text{rest}) \\
&\propto \rho_{i,n,j} \sum_{\ell=1}^{L} \eta_{i,j,\ell}^{z_{j,\lambda_{i,n}}} \cdot \phi(y_{i,n,j} \mid \mu^\star_{z_{j,k},\ell}, \sigma_i^2).
\end{aligned}
$$

Direct sampling from the full conditional of $y_{i,n,j}$ is difficult, so we use a Metropolis step with a normal proposal distribution to sample from the full conditional instead.

A.1.2 Variational Inference Implementation Details

Variational inference (VI) is a popular alternative for fitting Bayesian models (Jordan et al. 1999, Beal et al. 2003, Wainwright et al. 2008, Blei et al. 2017). VI tends to be faster and more scalable with data size than the traditional MCMC method. In particular, we utilize automatic differentiation variational inference (ADVI), (Kucukelbir et al. 2017), a derivation-free method. It is a gradient-based stochastic optimization method and is amenable to common machine learning techniques, such as stochastic gradient descent, which makes inference for large

datasets more tractable. For a comprehensive review of recent advances in VI, see Blei et al. (2017) and Zhang et al. (2018).

In VI, parameters of a tractable approximating "variational" distribution are iteratively optimized until it "sufficiently" resembles the target (posterior) distribution. The most common metric for measuring the "closeness" of the target distribution to the variational distribution is the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951). For our Bayesian feature allocation model (FAM), minimizing the KL divergence between the variational distribution and the posterior distribution is equivalent to maximizing the following evidence lower bound (ELBO)

$$\begin{aligned}
\text{ELBO} &= \mathrm{E}_Q \left[\log p(\boldsymbol{m}, \boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) - \log q(\boldsymbol{y}^{\text{missing}}) \right] \\
&= \mathrm{E}_Q \left[\log p(\boldsymbol{m} \mid \boldsymbol{y}, \boldsymbol{\theta}) + \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) - \log q(\boldsymbol{y}^{\text{missing}}) \right] \\
&= \mathrm{E}_Q \left[\log p(\boldsymbol{m} \mid \boldsymbol{y}) + \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) - \log q(\boldsymbol{y}^{\text{missing}}) \right].
\end{aligned} \tag{A.2}$$

$p(\boldsymbol{m} \mid \boldsymbol{y})$ and $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ are the sampling distributions of $m_{i,n,j}$ and $y_{i,n,j}$, and $p(\boldsymbol{\theta})$ is the prior distribution for all model parameters. $q(\boldsymbol{\theta})$ is the mean-field variational distribution for model parameters. For $q(\boldsymbol{\theta})$, each model parameter is transformed to the unconstrained space (Kucukelbir et al. 2017) and is assumed to have a normal distribution (Kucukelbir et al. 2017). $q(\boldsymbol{y}^{\text{missing}}) = \prod_{i,n,j} q(y_{i,n,j})^{1-m_{i,n,j}}$ is an amortized variational distribution for the missing values(Kingma and Welling 2013). Specifically, $q(y_{i,n,j}^{\text{missing}})$ is a normal probability density function with mean $r_{i,j}$ and standard deviation $s_{i,j}$. This simplification for the missing $y_{i,n,j}$ will produce imputed values different from those under our Bayesian FAM, but yields acceptable performance in our simulation studies at greatly reduced computational

cost. Computing the gradient (in gradient descent) requires the computation of the ELBO using the entire dataset. This can be computationally prohibitive for large datasets. Instead, stochastic gradient descent (SGD) is used. A mini-batch of size B (much less than the size of the full data set N) can be sampled at each iteration of the SGD to compute the ELBO. The ELBO should be appropriately scaled by N/B. This works well in practice provided that the size of the mini-batch is sufficiently large.

In our model, parameters of primary interest $\boldsymbol{Z}$ and $\boldsymbol{\lambda}$ are discrete. Since ADVI is only valid for continuous parameters in differentiable models, we let $z_{j,k} = 1(v_k > h_{j,k})$, where $v_k \mid \alpha \sim \text{Beta}(\alpha/K, 1)$, and $h_{j,k} \sim \text{Uniform}(0, 1)$, similar to the construction of the dependent IBP in Williamson et al. (2010). We approximate the gradient of the indicator function with the gradient of $\text{logistic}\left((\text{logit}(v_k) - \text{logit}(h_{j,k})) \cdot 1000\right)$, which is smooth. We marginalize over $\boldsymbol{\lambda}$ for VI, and then sample from their full conditionals using the parameters estimated from the variational distributions.

For completeness, we have included key terms in the computation of the ELBO using SGD. $p(\boldsymbol{m} \mid \boldsymbol{y})$ is defined as

$$
\begin{aligned}
p(\boldsymbol{m} \mid \boldsymbol{y}) &= \prod_{i=1}^{I} \prod_{n=1}^{N_i} p(\boldsymbol{m}_{i,n} \mid \boldsymbol{y}_{i,n}) \\
&= \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} \left(1 - \rho_{i,n,j}\right)^{m_{i,n,j}} \\
&= \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} c_{i,n,j} \\
&= \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} c_{i,n,j} \\
&= C \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}},
\end{aligned}
$$

where $\rho_{i,n,j} = \text{logistic}(\beta_{0,i} + \beta_{1,i} y_{i,n,j} + \beta_{2,i} y_{i,n,j}^2)$, and $C = \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} c_{i,n,j}$ is a constant. Evaluating $p(\boldsymbol{m} \mid \boldsymbol{y})$ is computationally expensive when N_i is large. Hence, we can approximate it by only iterating through a subset of the data, and scaling the relevant terms. The log of the resulting expression is:

$$\begin{aligned}
\log p(\boldsymbol{m} \mid \boldsymbol{y}) &= \log C + \sum_{i=1}^{I} \sum_{n=1}^{N_i} \sum_{j=1}^{J} (1 - m_{i,n,j}) \log \rho_{i,n,j} \\
&\approx \log C + \sum_{i=1}^{I} \frac{N_i}{|S_i|} \sum_{n \in S_i} \sum_{j=1}^{J} (1 - m_{i,n,j}) \log \rho_{i,n,j}
\end{aligned}$$

where S_i is a subset of $\{1, \ldots, N_i\}$.

The likelihood term $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is defined as

$$\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{I} \prod_{n=1}^{N_i} A_{i,n}, \text{ where} \\
A_{i,n} &= \epsilon_i \prod_{j=1}^{J} \text{Normal}(0, s_\epsilon^2) + \\
&\qquad (1 - \epsilon_i) \sum_{k=1}^{K} w_{i,k} \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \text{Normal}(y_{i,n,j} \mid \mu^\star_{z_{j,k},\ell}, \sigma_i^2).
\end{aligned}$$

We thus have

$$\begin{aligned}
\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) &= \sum_{i=1}^{I} \sum_{n=1}^{N_i} \log A_{i,n} \\
&\approx \sum_{i=1}^{I} \frac{N_i}{|S_i|} \sum_{n \in S_i} \log A_{i,n} \quad \text{(if using mini-batches)}
\end{aligned}$$

Finally, the variational distribution for the missing values in $\boldsymbol{y}$ is defined as

$$
\begin{aligned}
q(\boldsymbol{y}) &= \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} q(y_{i,n,j} \mid r_{i,j}, s_{i,j})^{m_{i,n,j}} \\
\Rightarrow \log q(\boldsymbol{y}) &= \sum_{i=1}^{I} \sum_{n=1}^{N_i} \sum_{j=1}^{J} m_{i,n,j} \log q(y_{i,n,j} \mid r_{i,j}, s_{i,j}) \\
&\approx \sum_{i=1}^{I} \frac{N_i}{|S_i|} \sum_{n \in S_i} \sum_{j=1}^{J} m_{i,n,j} \log q(y_{i,n,j} \mid r_{i,j}, s_{i,j}) \quad \text{(if using mini-batches)}
\end{aligned}
$$

As previously noted, independent Gaussian variational distributions were placed on all other model parameters $\boldsymbol{\theta}$ after they were transformed to have support on $\mathbb{R}^{\dim(\boldsymbol{\theta})}$. Notably, the parameters with support on simplexes (i.e. $\boldsymbol{\eta}$ and $\boldsymbol{w}$) were transformed using the stick breaking transformation (Team et al. 2016).

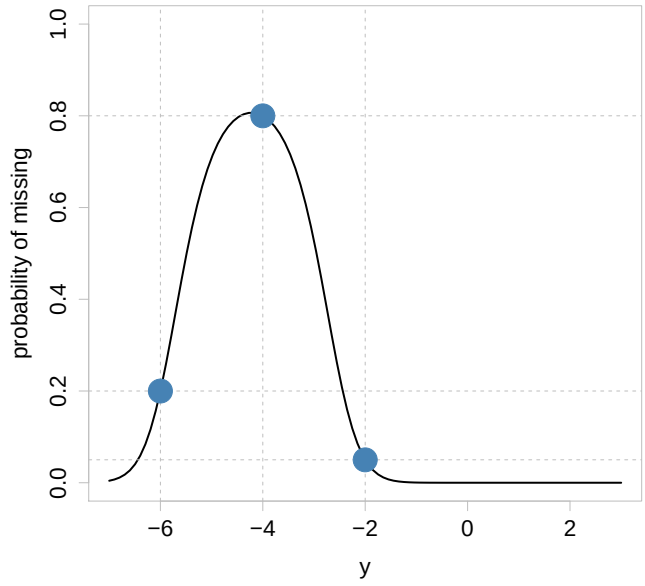


Figure A.1: A quadratic data missingship mechanism for imputing missing data that passes through the points $(y_1 = -6.0, p_1 = 0.2)$, $(y_2 = -4.0, p_2 = 0.8)$, and $(y_3 = -2.0, p_3 = 0.05)$.

A.2 Specification of Data Missingship Mechanism

We now discuss the approach used to specify the data missingship mechanism. Recall that we assume a logit regression model for the probability $\rho_{i,n,j}$ for the missing $y_{i,n,j}$ in (2.5) of the main text, $\text{logit}(\rho_{i,n,j}) = \beta_{0,i} + \beta_{1,i}y_{i,n,j} + \beta_{2,i}y_{i,n,j}^2$, with $\beta_{p,i} \in \mathbb{R}$, $p \in \{0, 1, 2\}$. To specify values of $\beta_{p,i}$, we first select three points of $(\tilde{y}, \tilde{\rho})$ for each sample, $(\tilde{y}_1, \tilde{\rho}_1)$, $(\tilde{y}_2, \tilde{\rho}_2)$, and $(\tilde{y}_3, \tilde{\rho}_3)$. We let $\text{logit}(\tilde{\rho}) = \beta_{0,i} + \beta_{1,i}\tilde{y} + \beta_{2,i}\tilde{y}^2$ and solve for $\beta_{i,p}$. We accommodate the subject knowledge that missing $y_{i,n,j}$ strongly indicates that the marker is not expressed in the selection of three points of $(\tilde{y}, \tilde{\rho})$, and the mechanism encourages imputed values to take on negative values. For instance, Figure A.1 shows an example of data missingship mechanism specified by selecting $(-6.0, 0.2)$, $(-4.0, 0.8)$, and $(-2.0, 0.05)$ of $(\tilde{y}, \tilde{\rho})$. This specification imputes values between -2 and -6 with large probability. The mechanism thus strongly implies that the marker is not expressed. We used empirical quantiles of negative values of observed y to specify $\tilde{y}$.

A.3 Computation of LPML and DIC

We used the log pseudo marginal likelihood (LPML) and deviance criterion information (DIC) to select the number of cell subpopulations (K) as discussed in §2.2 of the main text. LPML (Gelfand and Dey 1994, Gelfand et al. 1992)) is defined as $\text{LPML} = \sum_{i=1}^{n} \log \text{CPO}_i$, where $\text{CPO}_i = \int f(\text{data}_i \mid \text{data}_{-i}, \theta)p(\theta \mid \text{data}_{-i})d\theta \approx \left[\frac{1}{B}\sum_{b=1}^{B} \frac{1}{f(\text{data}_i|\theta^{(b)})}\right]^{-1}$, where $f(\text{data}_i \mid \theta^{(b)})$ is the likelihood evaluated at Monte Carlo sample b of B samples for observation i, and CPO_i is the

conditional predictive ordinates. The likelihood of cell n in sample i is

$$f(\boldsymbol{m}_{i,n}, \boldsymbol{y}_{i,n} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} (1-\rho_{i,n,j})^{m_{i,n,j}} \cdot \phi(y_{i,n,j} \mid \mu_{i,n,j}, \sigma_i^2)$$
$$\propto \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} \cdot \phi(y_{i,n,j} \mid \mu_{i,n,j}, \sigma_i^2), \tag{A.3}$$

where $\phi(y \mid m, s^2)$ denotes the probability density function of the normal distribution with mean m and variance s^2, evaluated at y. Note that $(1-\rho_{i,n,j})^{m_{i,n,j}}$ in (A.3) is dropped since it remains constant for observed $y_{i,n,j}$. We then compute LPML as

$$\begin{aligned}
\text{LPML} &= \sum_{i=1}^{I} \sum_{n=1}^{N_i} \log \text{CPO}_{i,n} \\
&\approx \sum_{i=1}^{I} \sum_{n=1}^{N_i} \log \left\{ \frac{1}{B} \sum_{b=1}^{B} \frac{1}{f(\boldsymbol{m}_{i,n}, \boldsymbol{y}_{i,n} \mid \boldsymbol{\theta}^{(b)})} \right\}^{-1} \\
&\propto \sum_{i=1}^{I} \sum_{n=1}^{N_i} \log \left\{ \frac{1}{B} \sum_{b=1}^{B} \frac{1}{\prod_{j=1}^{J} (\rho_{i,n,j}^{(b)})^{m_{inj}} \cdot \phi(y_{i,n,j} \mid \mu_{i,n,j}^{(b)}, \sigma_i^{2,(b)})} \right\}^{-1}.
\end{aligned}$$

Deviance is defined as as $D = -2 \log f(\boldsymbol{m}, \boldsymbol{y} \mid \boldsymbol{\theta})$, where $f(\boldsymbol{m}, \boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood. The deviance criterion information (DIC) (Spiegelhalter et al. 2002) is computed as $\text{DIC} = \bar{D} - D(\bar{\theta})$, where $\bar{D} = \text{E}[D]$ is the posterior mean of the deviance, and $\bar{\boldsymbol{\theta}}$ is the posterior mean of the parameters $\boldsymbol{\theta}$. We compute the likelihood as

$$f(\boldsymbol{m}, \boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-m_{i,n,j}} \cdot \phi(y_{i,n,j} \mid \mu_{i,n,j}, \sigma_i^2). \tag{A.4}$$

The parameters that appear in the likelihood include $\mu_{i,n,j}, \sigma_i^2$, and the missing values $y_{i,n,j}^{\star}$. So $\bar{\boldsymbol{\theta}}$ can be obtained by computing the posterior means of $\mu_{i,n,j}$, σ_i^2, and $y_{i,n,j}^{\star}$.

A.4 Simulation Study

A.4.1 Additional Results for Simulation 1

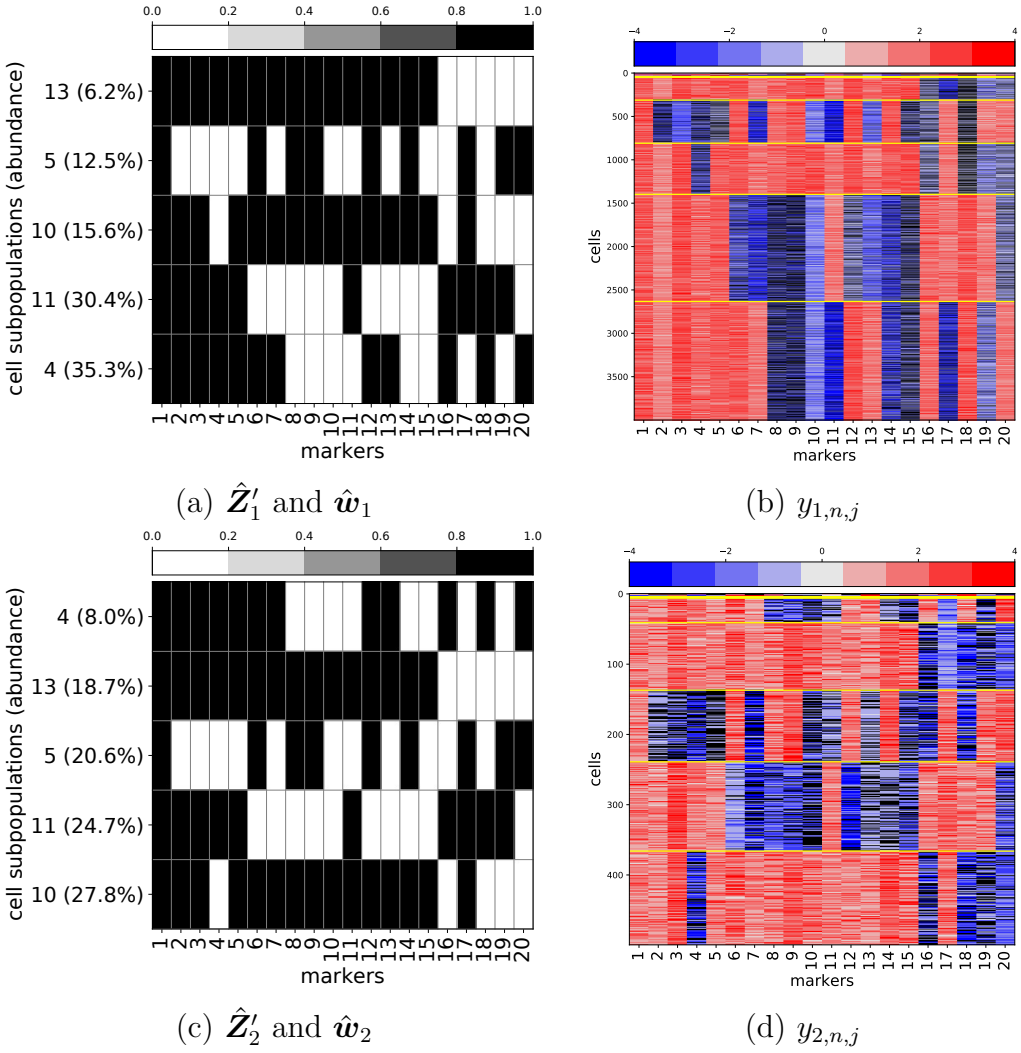
Here we present additional figures and tables for Simulation 1. Figure A.2 summarizes the results from the analysis of Simulation 1 via ADVI. It contains the element-wise posterior means of $\boldsymbol{Z}$ and the posterior means of $\boldsymbol{w}_i$ (panels (a), (c), and (e)), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior mode of the cell subpopulation indicators $\hat{\lambda}_{i,n}$ (panels (b), (d), and (f)).

Table A.1 contains the three data missingship mechanisms (MM) used in Simulation 1. MM0 is the default mechanism. Recall that we used empirical $\tilde{\boldsymbol{q}}$-quantiles to specify $\tilde{\boldsymbol{y}}$. Different $\tilde{\boldsymbol{q}}$ yields different values of $\boldsymbol{\beta}$. Three different sets of $\tilde{\boldsymbol{q}}$ are used for the sensitivity analysis, while fixing $\tilde{\boldsymbol{\rho}}$. For each mechanism, the LPML and DIC are shown in the last two columns of the table.

MM	$\tilde{\boldsymbol{q}}$	Probability of Missing ($\tilde{\boldsymbol{\rho}}$)	LPML	DIC
0	(0%, 25%, 50%)	(5%, 80%, 5%)	-16.728	172989
I	(0%, 20%, 40%)	(5%, 80%, 5%)	-16.681	172914
II	(0%, 15%, 30%)	(5%, 80%, 5%)	-16.462	170971

Table A.1: Data missingship mechanisms (MM) used for Simulation 1. $\tilde{\boldsymbol{q}}$-quantiles of the negative observed values in each sample are used to specify $\tilde{\boldsymbol{y}}$, and $\tilde{\boldsymbol{\rho}}$ are the probability of missing at those $\tilde{\boldsymbol{y}}$. Three different sets of $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{\rho}}$ are used to examine the sensitivity to the missingship mechanism specification. LPML and DIC are shown in the last two columns under each of the specification.

Figures A.3 and A.4 respectively summarize the results for the analysis of Simulation 1 under data missingship mechanism I and II, done via MCMC. The figures contain the posterior estimate of $\boldsymbol{Z}$ and $\boldsymbol{w}$ in panels (a), (c), and (e), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior estimate of the cell subpopulation indicator $\hat{\lambda}_{i,n}$ in panels (b), (d), and (f).



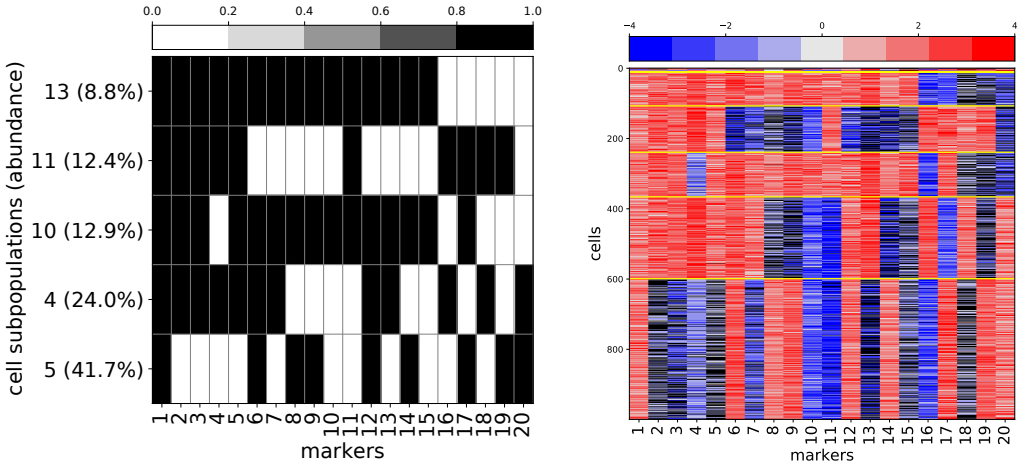
(a) $\hat{\boldsymbol{Z}}_1'$ and $\hat{\boldsymbol{w}}_1$

(b) $y_{1,n,j}$

(c) $\hat{\boldsymbol{Z}}_2'$ and $\hat{\boldsymbol{w}}_2$

(d) $y_{2,n,j}$

Figure A.2: [ADVI for Simulation 1] In (a) and (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 dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ are shown for sample 1 in (b) and sample 2 in (d). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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. Posterior estimates are obtained via ADVI.

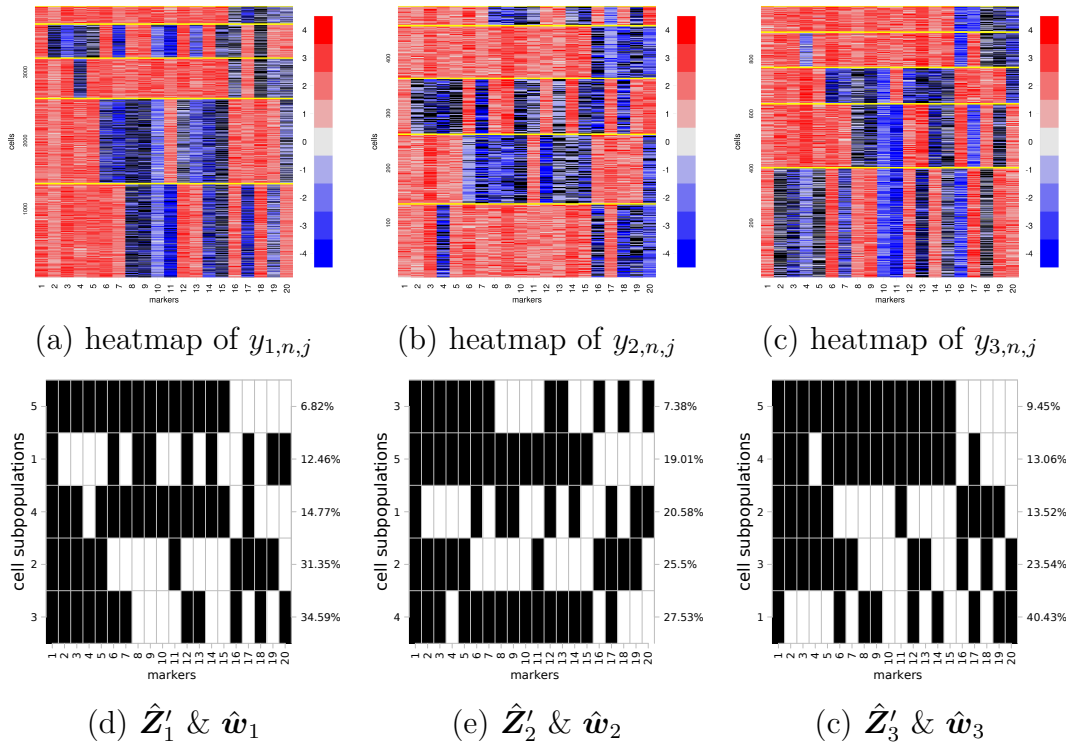


(e) $\hat{\boldsymbol{Z}}_3'$ and $\hat{\boldsymbol{w}}_3$ (f) $y_{3,n,j}$

Figure A.2 (continued): 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 dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ for sample 3 is shown in (f). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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. Posterior estimates are obtained via ADVI.

A.4.2 Simulation 2

An additional simulation study, Simulation 2, that assumes a larger simulated dataset and a more complex cell subpopulation structure, was performed. The dataset was simulated in a manner similar to Simulation 1 in § 2.3 of the main text, but the data size is larger with $N = (40000, 5000, 10000)$, and has more cell subpopulations with $K^{\text{TR}} = 10$. We first specify $\boldsymbol{Z}^{\text{TR}}$ and simulated $\boldsymbol{w}_i^{\text{TR}}$ from a Dirichlet distribution with parameters being some random permutation of $(1, \ldots, K)$. Table A.2 illustrates $\boldsymbol{Z}^{\text{TR}}$ and $\boldsymbol{w}^{\text{TR}}$. Parameters $\mu_0^{\star,\text{TR}}$, $\mu_1^{\star,\text{TR}}$, and $\sigma_i^{2,\text{TR}}$ are set in the same way as Simulation 1. We fit the model over a grid for K, for K from 2 to 20 in increments of 2. For all models, we fixed $L_0 = 5$ and $L_1 = 5$. Recall that $L_0^{\text{TR}} = L_1^{\text{TR}} = 3$. All other parameter specifications, MCMC



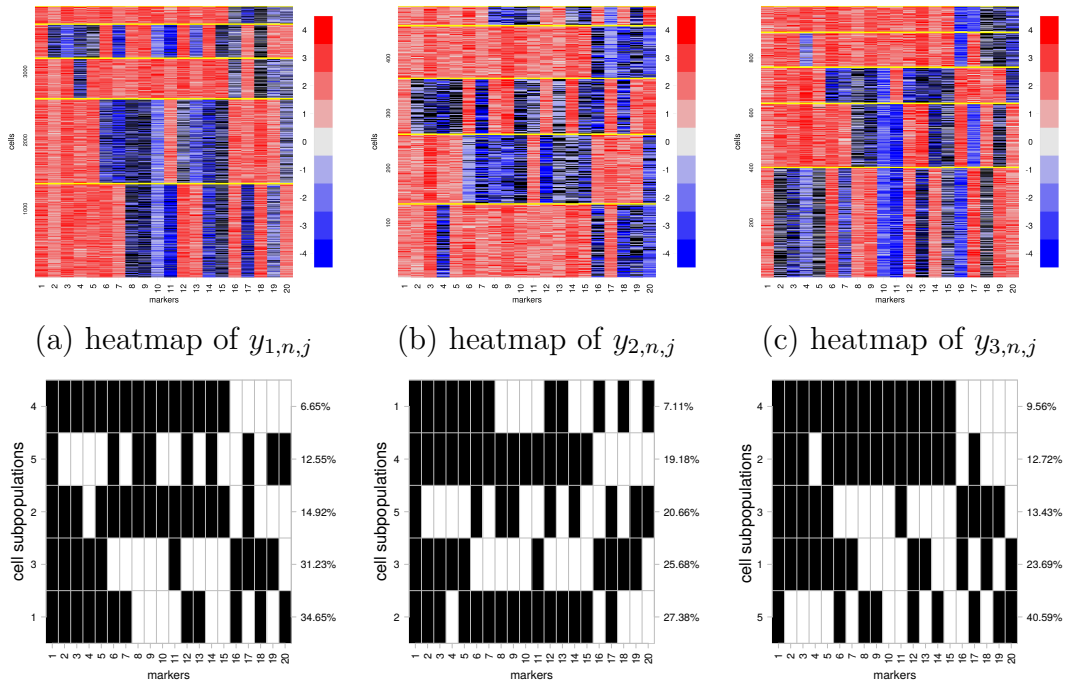
(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (c) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

Figure A.3: Data missingship mechanism sensitivity analysis for Simulation 1. Specification I is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_i$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of cell clustering, $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f). We include only subpopulations with $\hat{w}_{i,k} > 1\%$.

initialization, and MCMC specifications were done in the same way as Simulation 1.

The LPML, DIC, and calibration metric for K are presented in Figure A.5. The metrics indicate that the model with $\hat{K} = 10$ fits the data best and achieves a balance between good model fit and low model complexity. Figure A.6 shows posterior estimates of the clusterings for each sample for the large simulated dataset, along with posterior estimates of the subpopulations present ($\hat{\boldsymbol{Z}}_i$) and their abun-



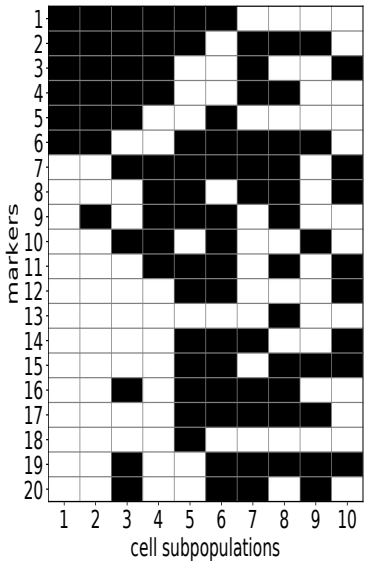
(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (f) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

Figure A.4: Data missingship mechanism sensitivity analysis for Simulation 1. Specification II is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_i$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of cell clustering, $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f).We include only subpopulations with $\hat{w}_{i,k} > 1\%$.

dances ($\hat{\boldsymbol{w}}_i$) in each sample. The red, blue, and black cells represent high, low, and non-observed expression levels, respectively. Horizontal yellow lines separate cells into clusters. The simulation truth for the cell subpopulations in $\boldsymbol{Z}^{\text{TR}}$ is recovered by $\hat{\boldsymbol{Z}}$, and $\hat{\boldsymbol{w}}_i$ is close to $\boldsymbol{w}^{\text{TR}}$.

Figure A.7 shows estimated clusterings for each sample $\boldsymbol{y}_i$ using FlowSOM. The largest cluster in sample 1 shown in panel (a) contains a mixture of high and low expression levels for marker 9, resulting in poor performance of clustering cells.

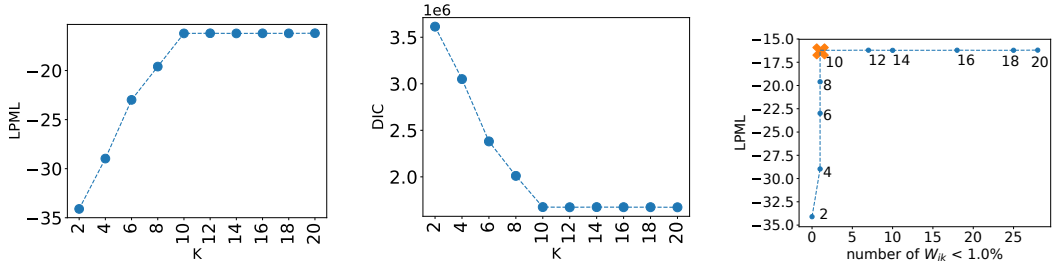


(a) $\boldsymbol{Z}^{\text{TR}}$

subpopulations	sample 1	sample 2	sample 3
$k = 1$	0.136	0.160	0.033
$k = 2$	0.132	0.021	0.128
$k = 3$	0.111	0.037	0.257
$k = 4$	0.157	0.084	0.110
$k = 5$	0.044	0.183	0.049
$k = 6$	0.046	0.111	0.142
$k = 7$	0.215	0.045	0.142
$k = 8$	0.072	0.109	0.001
$k = 9$	0.018	0.109	0.099
$k = 10$	0.065	0.135	0.035

(b) $\boldsymbol{w}^{\text{TR}}$

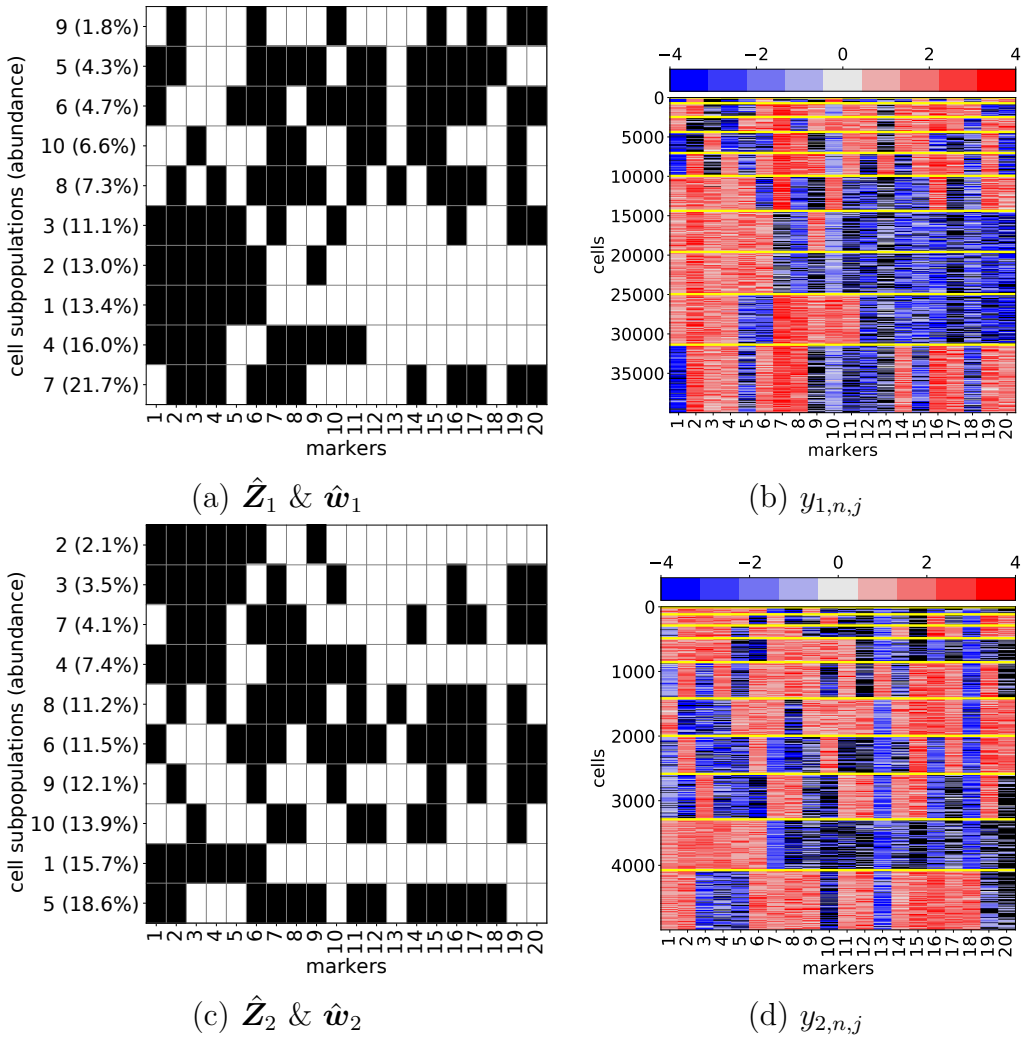
Table A.2: [Simulation 2] $\boldsymbol{Z}^{\text{TR}}$ and $\boldsymbol{w}^{\text{TR}}$ are illustrated in (a) and (b), respectively. $K^{\text{TR}} = 10$, $J = 20$, $I = 3$ and $N = (40000, 5000, 10000)$ are assumed. Black and white in (a) represents $z_{j,k}^{\text{TR}} = 1$ and 0, respectively.



(a) LPML (b) DIC (c) Calibration of K

Figure A.5: [Simulation 2] Plots of (a) LPML, (b) DIC, and (c) calibration metric, for $K = 2, 4, \ldots, 20$, for large simulated data suggest that $\hat{K} = 10$ is sufficient to explain the latent cell subpopulations.

This undesired behavior is not observed in the FAM. We again used the adjusted Rand index (ARI) to assess the accuracy of cluster assignments produced by the FAM and FlowSOM by comparing them to the true clustering. Table A.3 shows the ARI by sample for each method. Our method produced higher ARIs for all samples. The ARI in sample 1 is especially low for FlowSOM, as the two similar



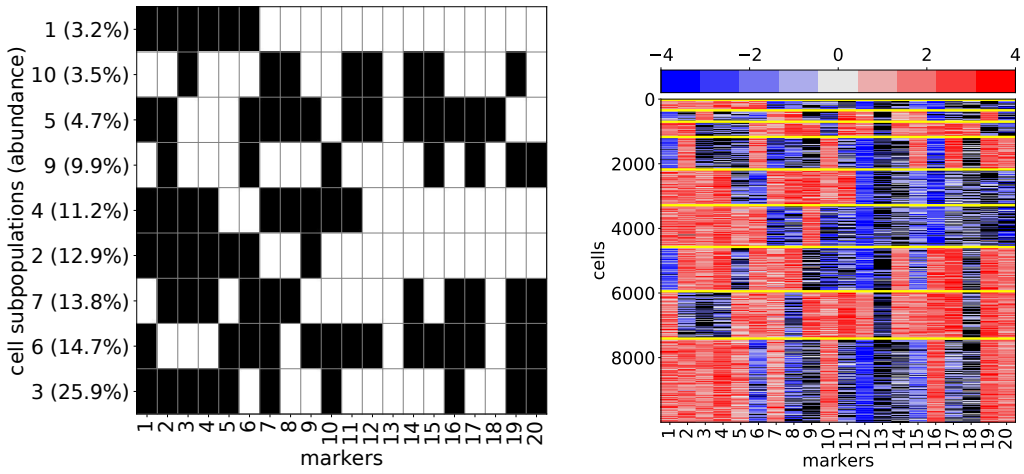
(a) $\hat{\boldsymbol{Z}}_1$ & $\hat{\boldsymbol{w}}_1$ (b) $y_{1,n,j}$

(c) $\hat{\boldsymbol{Z}}_2$ & $\hat{\boldsymbol{w}}_2$ (d) $y_{2,n,j}$

Figure A.6: [Simulation 2]. In (a) and (c), $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for samples 1 and 2, respectively, with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ are shown for sample 1 in (b) and sample 2 in (d). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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.

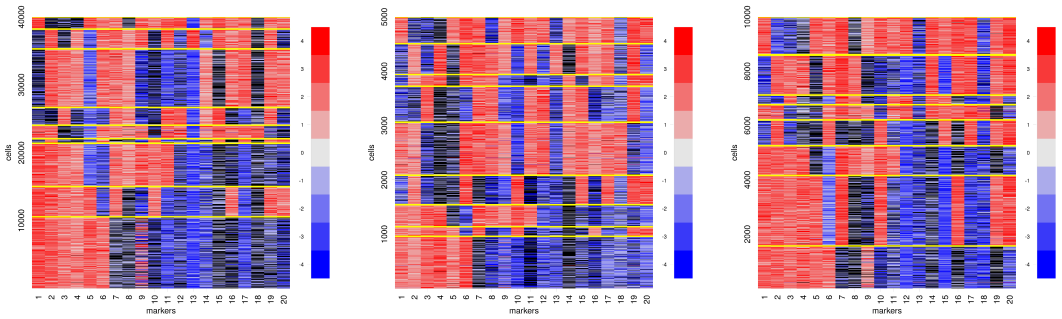
subpopulations that were grouped together make up a large portion of the cells in that sample.

Figure A.8 summarizes the posterior inference obtained via ADVI. The poste-



(e) $\hat{\boldsymbol{Z}}_3$ & $\hat{\boldsymbol{w}}_3$ (f) $y_{3,n,j}$

Figure A.6 (continued): Results of Simulation 2. In (e), $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for sample 3, with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ for sample 3 is shown in (f). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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.



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

Figure A.7: [FlowSOM for Simulation 2] Heatmaps of $\boldsymbol{y}_i$ for Simulation 2. Samples 1-3 are in (a)-(c), respectively. The cells are sorted by the cluster labels $\lambda_{i,n}$ for each sample, estimated by FlowSOM.

rior mean of $\boldsymbol{Z}$ and the posterior mean of $\boldsymbol{w}_i$ are in panels (a), (c), and (e), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior mode of

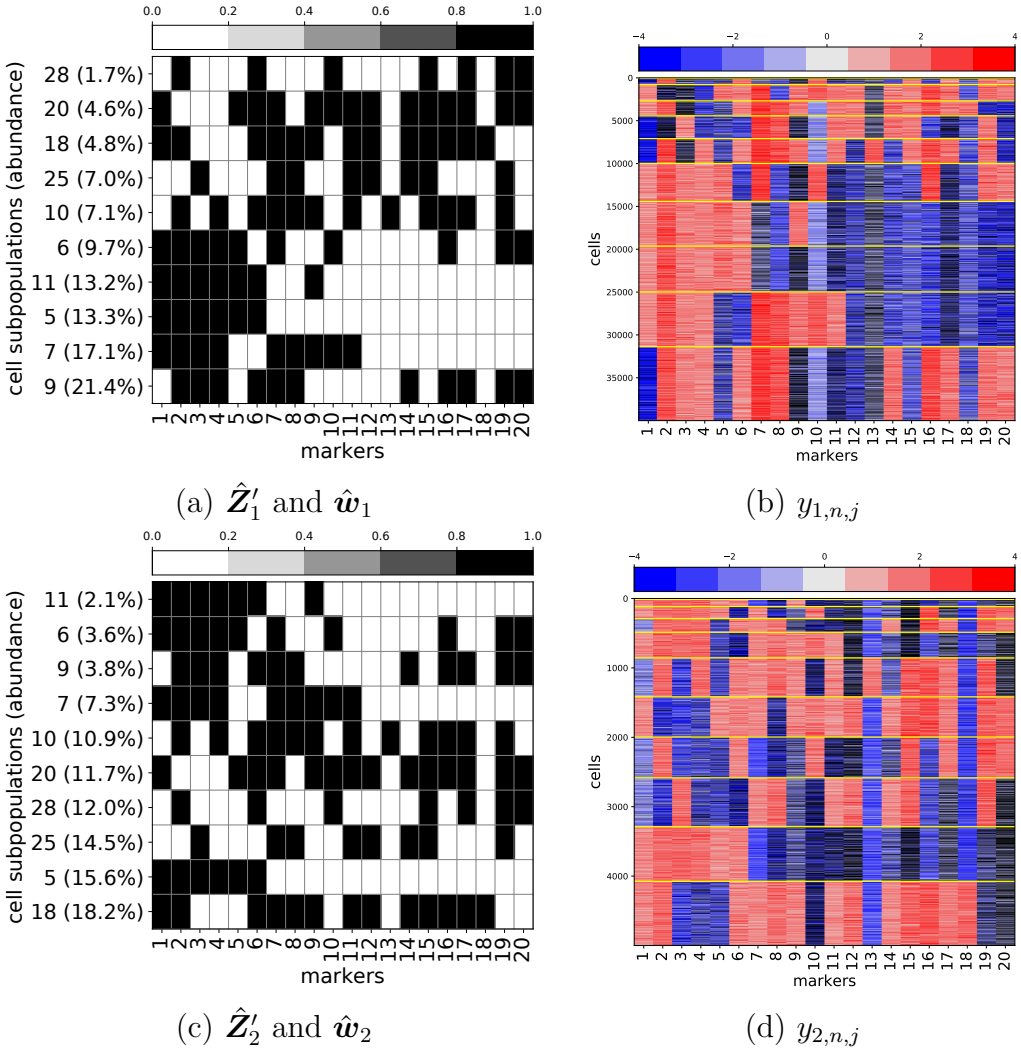
Method	Sample 1 ARI	Sample 2 ARI	Sample 3 ARI
FAM ($K = 10$)	0.999	0.996	0.999
FlowSOM	0.858	0.940	0.959

Table A.3: Adjusted Rand index (ARI) for FAM and FlowSOM by sample for Simulation 2. Higher ARI is better, and values closer to 1 indicate that estimated clusters are closer to the truth.

the cell subpopulations $\hat{\lambda}_{i,n}$ in panels (b), (d), and (f). The posterior inference covers the simulation truth well.

We performed a sensitivity analysis to the specification of the data missingship mechanism after selecting $K = 10$ via DIC and LPML. Table A.4 summarizes the missingship mechanisms used in the sensitivity analysis. Again, we note that inference on $\boldsymbol{Z}$ and $\boldsymbol{w}$ do not change significantly across the various missing mechanisms. However, the fit (in terms of LPML and DIC) on the observed data was highest for missingship mechanism II, which encourages imputing values that are more negative, as it best matched the simulation truth.

Figures A.9 and A.10 respectively summarize the results for the analysis of Simulation 1 under data missingship mechanism I and II, done via MCMC. The figures contain the posterior estimate of $\boldsymbol{Z}$ and $\boldsymbol{w}$ in panels (a), (c), and (e), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior estimate of the cell subpopulation indicators $\hat{\lambda}_{i,n}$ in panels (b), (d), and (f).



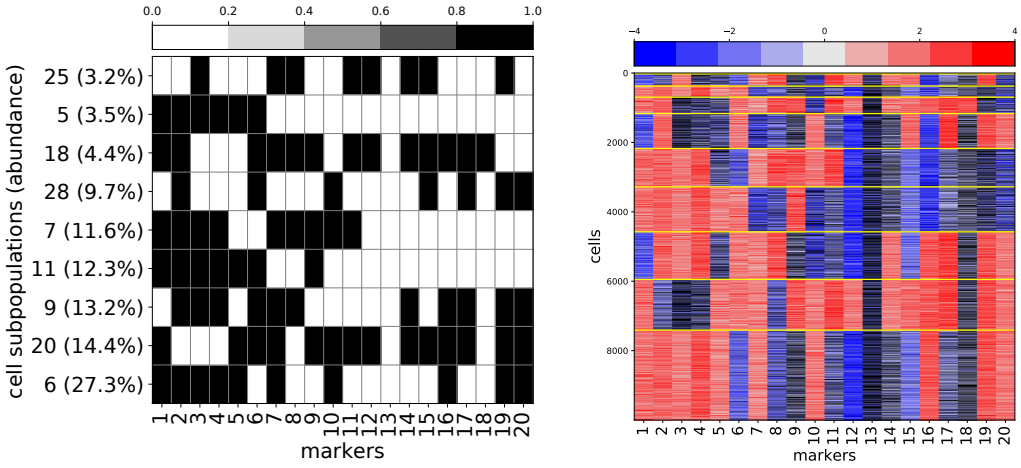
(a) $\hat{\boldsymbol{Z}}_1'$ and $\hat{\boldsymbol{w}}_1$

(b) $y_{1,n,j}$

(c) $\hat{\boldsymbol{Z}}_2'$ and $\hat{\boldsymbol{w}}_2$

(d) $y_{2,n,j}$

Figure A.8: [ADVI for Simulation 2] In (a) and (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 dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ are shown for sample 1 in (b) and sample 2 in (d). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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. Posterior estimates are obtained via ADVI.



(e) $\hat{\boldsymbol{Z}}_3'$ and $\hat{\boldsymbol{w}}_3$ (f) $y_{3,n,j}$

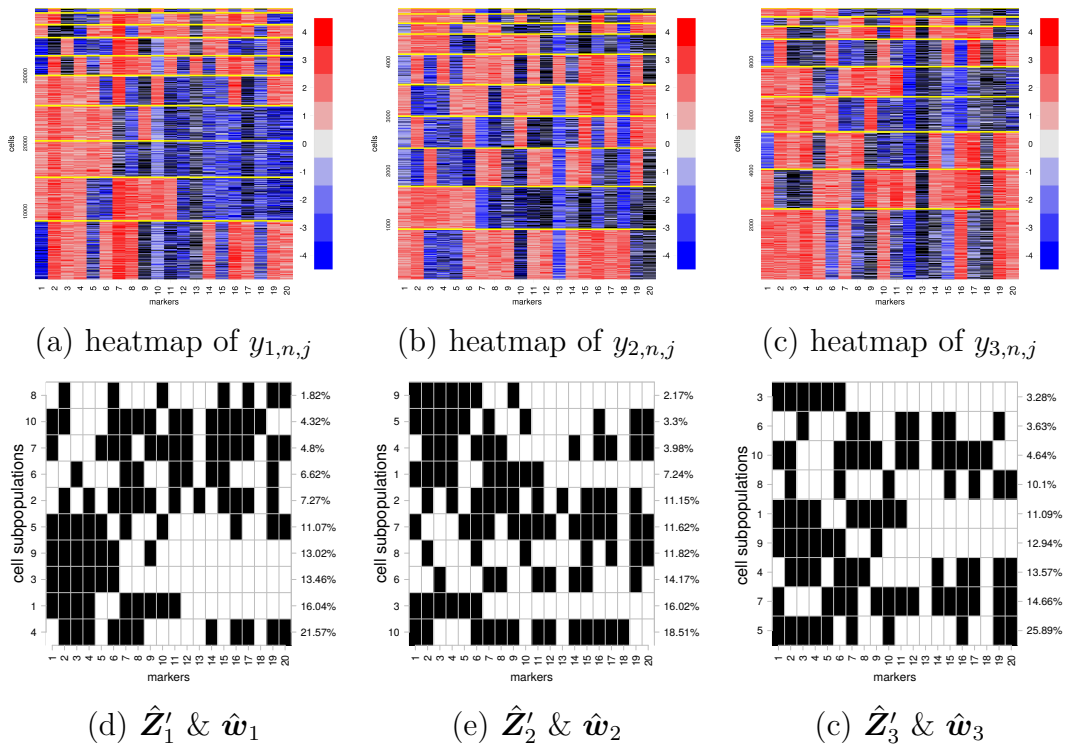
Figure A.8 (continued): [ADVI for Simulation 2] In (e), the transpose $\hat{\boldsymbol{Z}}_i'$ of $\hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{w}}_i$ are shown for sample 3, with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ for sample 3 is shown in (f). Cells are ordered by posterior point estimates of their subpopulations, $\hat{\lambda}_{i,n}$. Cells are given in rows and markers are given in columns. 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. Posterior estimates are obtained via ADVI.

MM	$\tilde{\boldsymbol{q}}$	Probability of Missing ($\boldsymbol{\rho}$)	LPML	DIC
0	(0%, 25%, 50%)	(5%, 80%, 5%)	-16.215	1675117
I	(0%, 20%, 40%)	(5%, 80%, 5%)	-16.052	1662834
II	(0%, 15%, 30%)	(5%, 80%, 5%)	-15.771	1640255

Table A.4: Missingness mechanisms used for Simulation 2. $\tilde{\boldsymbol{q}}$-quantiles of the negative observed values in each sample are used to specify $\tilde{\boldsymbol{y}}$, and $\boldsymbol{\rho}$ are the probability of missing at $\tilde{\boldsymbol{y}}$. Three different sets of $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{\rho}}$ are used to examine the sensitivity to the missingship mechanism specification. LPML and DIC are shown in the last two columns under each of the specification.

A.5 Additional Results for Analysis of Cord Blood Derived NK Cell Data

This section contains additional figures and tables for the CB NK cell data analysis presented in § 2.4 of the main text. Table A.5 lists the marker names

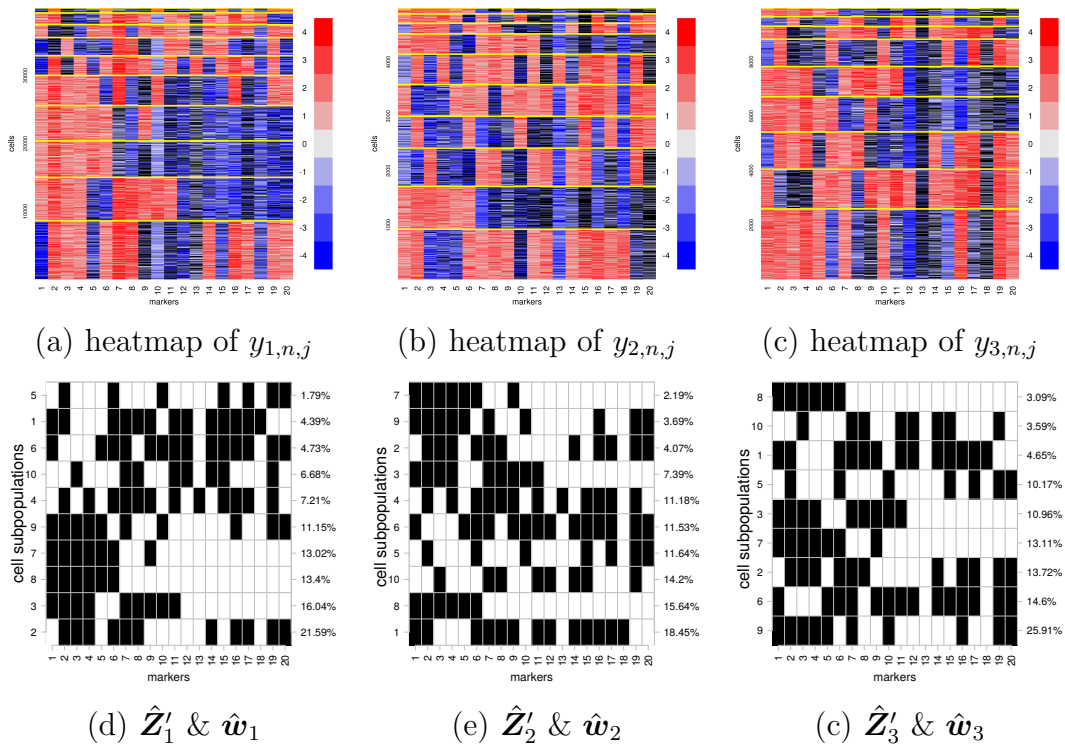


(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (c) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

Figure A.9: Data missingship mechanism sensitivity analysis for Simulation 2. Specification I is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_i$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of cell clustering, $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f). We include only subpopulations with $\hat{w}_{i,k} > 1\%$.

and numbers for each marker included in the CB derived NK data analysis. Figure A.11 visualizes the CB NK cell data in a two-dimensional space using a data visualization technique "t-SNE (t-Distributed Stochastic Neighbor Embedding)" (Maaten and Hinton 2008, Van Der Maaten 2014). The two dimensional embeddings are learned separately for each sample. Cells are represented with different symbols and colors by their posterior estimate $\hat{\lambda}_{in}$ of the cell clustering. All cells in the samples are used to obtain the embeddings, but only cells in the subpopu-



(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (c) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

Figure A.10: Data missingship mechanism sensitivity analysis for Simulation 2. Specification II is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_i$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of cell clustering, $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f). We include only subpopulations with $\hat{w}_{i,k} > 1\%$.

lations with $\hat{w}_{ik} \geq 0.05$ are included in the plots for better illustration.

Table A.6 contains the three data missingship mechanisms (MM) used in analyzing the CB derived NK data. MM0 is the default mechanism. Each mechanism is defines the parameters $\boldsymbol{\beta}$ through the quantiles of the negative observed values in each sample $\tilde{\boldsymbol{q}}$, and probability that a record is missing at those quantiles $\tilde{\boldsymbol{\rho}}$. For each mechanism, the LPML and DIC are shown. Table A.7 list the implied $\boldsymbol{\beta}$ for each data missingship mechanism.

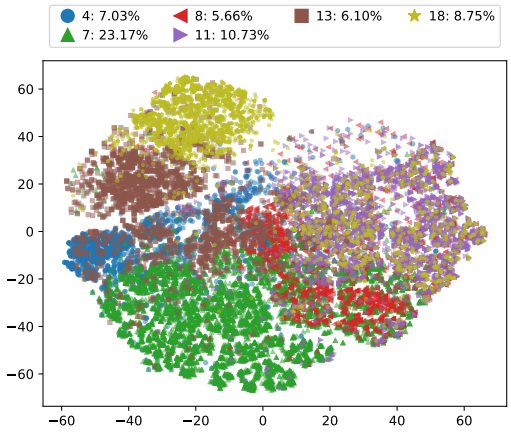
Marker Number	Marker Name
1	2B4
2	KIR2DL3
3	KIR3DL1
4	CD158B
5	CD16
6	CD27
7	CD62L
8	CD8
9	CD94
10	DNAM1
11	EOMES
12	KLRG1
13	NKG2A
14	NKG2C
15	NKG2D
16	NKP30
17	SIGLEC7
18	TBET
19	TIGIT
20	ZAP70

Table A.5: Marker names and numbers for each marker referenced in the CB NK cell data.

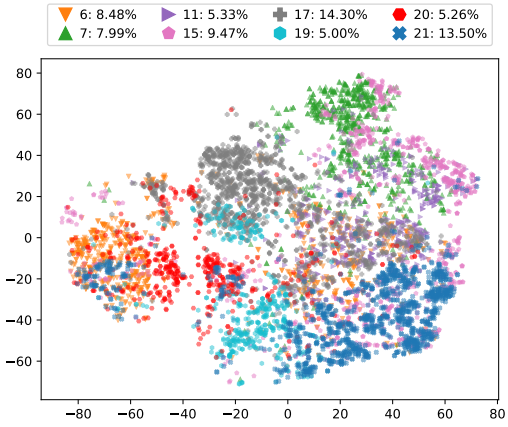
MM	$\tilde{\boldsymbol{q}}$	Probability of Missing ($\boldsymbol{\rho}$)	LPML	DIC
0	(0%, 25%, 50%)	(5%, 80%, 5%)	-24.90	2569097
I	(0%, 20%, 40%)	(5%, 80%, 5%)	-24.93	2569098
II	(0%, 15%, 30%)	(5%, 80%, 5%)	-24.98	2569098

Table A.6: $\tilde{\boldsymbol{q}}$-quantiles of the negative observed values in each sample are used to specify $\tilde{\boldsymbol{y}}$, and $\boldsymbol{\rho}$ are the probability of missing at $\tilde{\boldsymbol{y}}$. Three different sets of $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{\rho}}$ are used to examine the sensitivity to the missingship mechanism specification. LPML and DIC are shown in the last two columns under each of the specification.

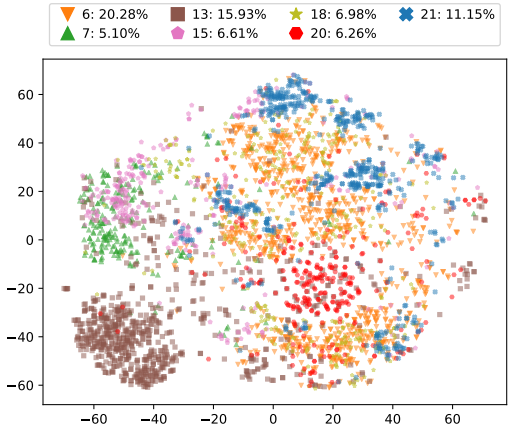
Figures A.12 and A.13 respectively summarize the results for the analysis of the CB NK cell data under data missingship mechanism I and II, done via MCMC. The posterior estimate of $\boldsymbol{Z}$ and $\boldsymbol{w}$ are shown in panels (a), (c), and (e), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior estimate



(a) Sample 1



(b) Sample 2



(c) Sample 3

Figure A.11: [Plots of t-SNE's for the CB data] The CB data is visualized using two-dimensional t-SNE's that are learned separately on each sample, where each point represents a cell. Cells in different subpopluations estimated by the FAM are marked by different symbols and colors. On the top of the scatterplots, the subpopulation numbers are listed with their corresponding symbols and colors. All cells are used to obtain t-SNE embeddings, but only cell subpopulations belonging to subpopluations with $\hat{w}_{ik} \geq 0.05$ are included in the plots for better illustration.

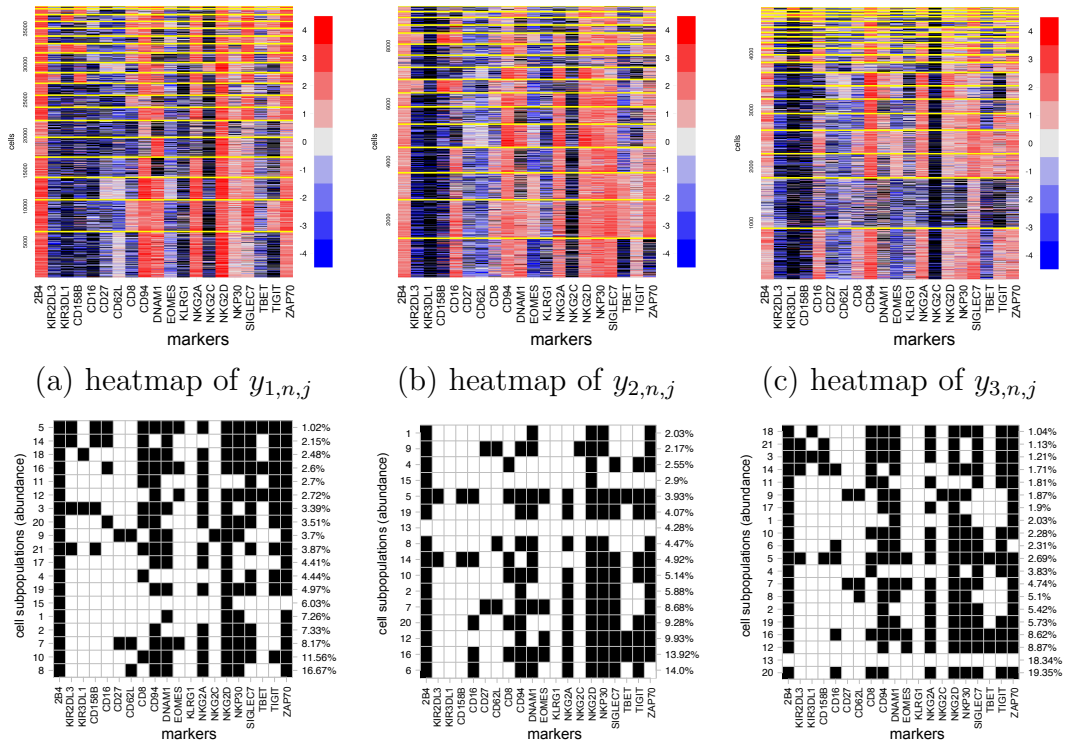
of the cell subpopulations $\hat{\lambda}_{i,n}$ in panels (b), (d), and (f)).

Figure A.14 summarizes the results from the analysis of the UCB NK cell data via ADVI. The posterior mean of $\boldsymbol{Z}$ and the posterior mean of $\boldsymbol{w}_i$ are in panels (a), (c), and (e)), and heatmaps of the simulated data $y_{i,n,j}$ sorted according to the posterior mode of the cell subpopulations $\hat{\lambda}_{i,n}$ in panels (b), (d), and (f).

Data Missingship Mechanism	β	Sample 1	Sample 2	Sample 3
0	β_0	-15.35	-15.73	-13.66
	β_1	-10.39	-10.20	-9.60
	β_2	-1.38	-1.34	-1.30
I	β_0	-20.40	-21.50	-18.21
	β_1	-12.60	-12.76	-11.62
	β_2	-1.61	-1.61	-1.51
II	β_0	-27.43	-29.21	-25.26
	β_1	-15.52	-15.86	-14.62
	β_2	-1.90	-1.91	-1.81

Table A.7: Values for β used for the sensitivity analysis to the missingship mechanism in CB NK cell data analysis.

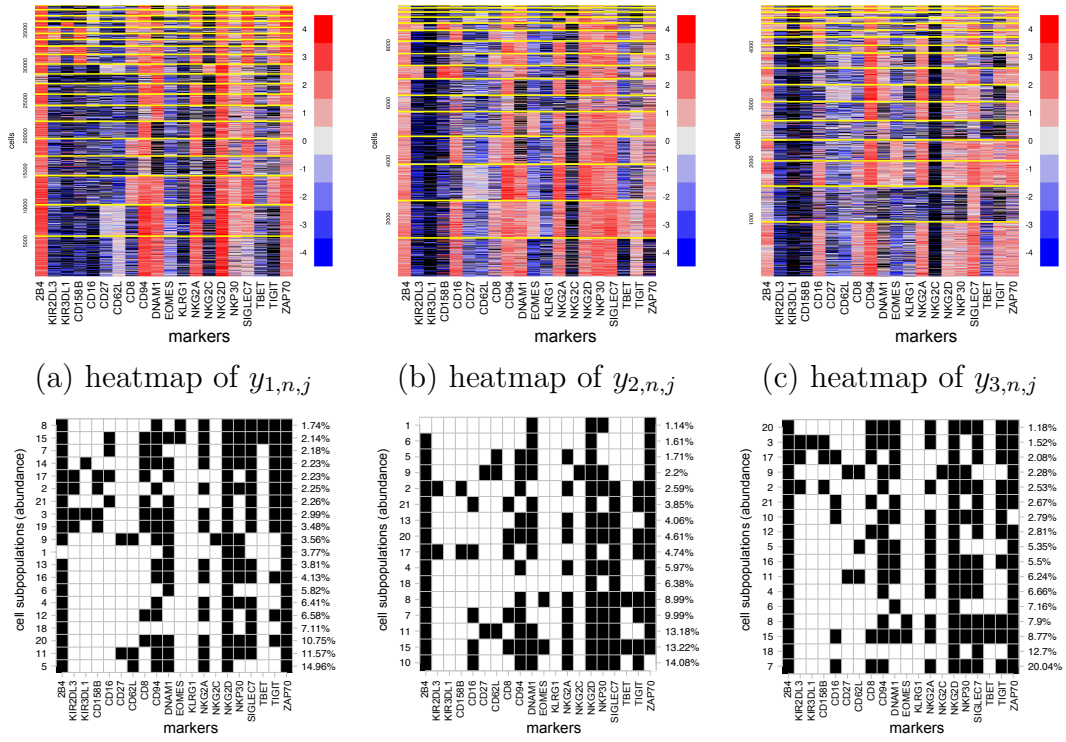
We repeated the analysis with different preprocessing rules for removing markers. We removed makers that have either negative or positive expression levels in more than $p \times 100\%$ cells, and we varied $p \in \{0.85, 0.90, 0.95\}$. Different thresholds yield different sets of markers. The markers included for analysis with $p = 0.9$ is listed in Tab A.5. Table A.8 additionally lists, for the various p, markers included in the analysis and the reasons for their exclusions, if applicable. Figures A.15-A.16, respectively, provide the estimates for $\boldsymbol{Z}_i$ and the heatmaps of y_i with cells sorted by subpopulation membership. We demonstrate that the choice of p does not heavily affect the estimation of $\boldsymbol{Z}_i$ for large subpopulations. See, for example, Figure A.15 where the middle column are the results from ADVI with $p = 0.9$. For sample 1 (first row of images), the largest subpopulations in (a) and (b) are the same for markers that are common between the two datasets. Between (b) and (c) the largest subpopulations are nearly the same, differing by only markers CD62L and SIGLEC7. Minor differences likewise appear in the other larger subpopulations.



(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (f) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

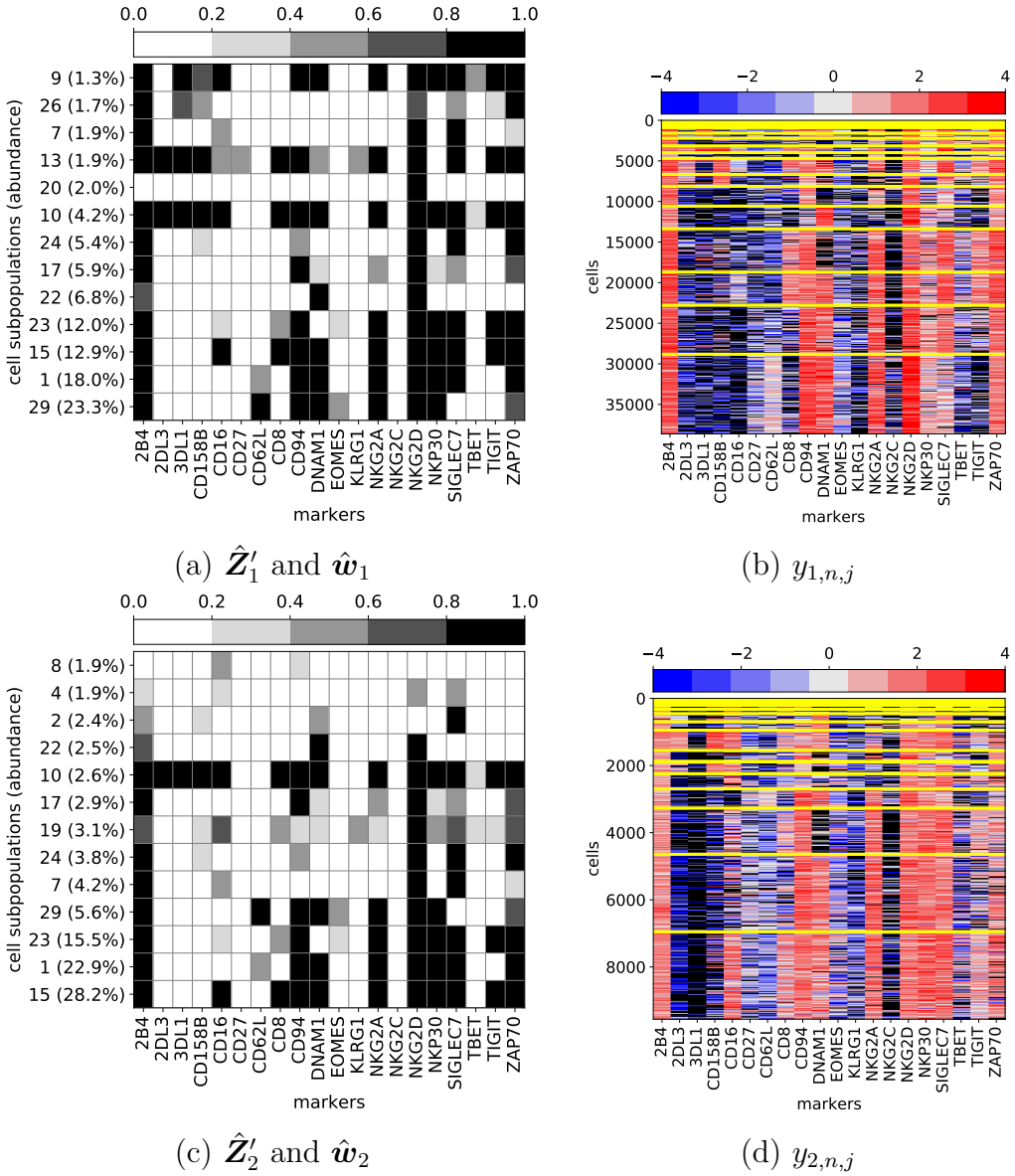
Figure A.12: Data missingship mechanism sensitivity analysis for CB NK cell data analysis. Specification I is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_u$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of the cell clusterings $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f). We include only subpopulations with $\hat{w}_{i,k} > 1\%$.



(a) heatmap of $y_{1,n,j}$ (b) heatmap of $y_{2,n,j}$ (c) heatmap of $y_{3,n,j}$

(d) $\hat{\boldsymbol{Z}}_1'$ & $\hat{\boldsymbol{w}}_1$ (e) $\hat{\boldsymbol{Z}}_2'$ & $\hat{\boldsymbol{w}}_2$ (f) $\hat{\boldsymbol{Z}}_3'$ & $\hat{\boldsymbol{w}}_3$

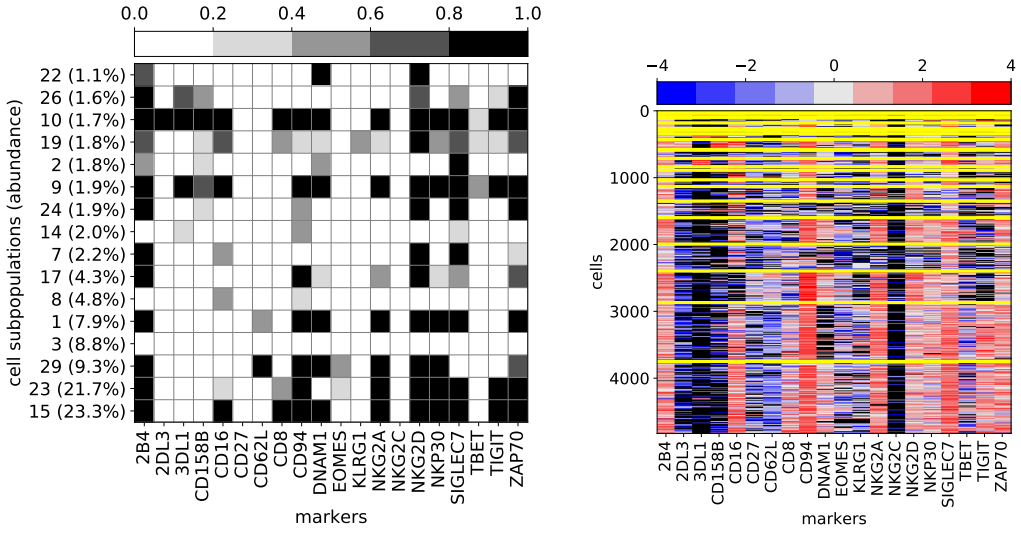
Figure A.13: Data missingship mechanism sensitivity analysis for CB NK cell data analysis. Specification II is used for $\boldsymbol{\beta}$. Heatmaps of $\boldsymbol{y}_i$ are shown in (a)-(c) for samples 1-3, respectively. Cells are rearranged by the posterior point estimate of the cell clusterings $\hat{\lambda}_{i,n}$. Cells and markers are in rows and columns, respectively. High and low expression levels are in red and blue, respectively, and black is used for missing values. Yellow horizontal lines separate cells by different subpopulations. $\hat{\boldsymbol{Z}}_i'$ and $\hat{\boldsymbol{w}}_i$ are shown for each of the samples in (d)-(f). We include only subpopulations with $\hat{w}_{i,k} > 1\%$.



(a) $\hat{\boldsymbol{Z}}'_1$ and $\hat{\boldsymbol{w}}_1$ (b) $y_{1,n,j}$

(c) $\hat{\boldsymbol{Z}}'_2$ and $\hat{\boldsymbol{w}}_2$ (d) $y_{2,n,j}$

Figure A.14: [CB NK cell data] Inference obtained by VI is illustrated. $\hat{\boldsymbol{Z}}'_i$ and $\hat{\boldsymbol{w}}_i$ of samples 1 and 2 are illustrated in panels (a) and (c), respectively, with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ are shown in panels (b) and (d) for samples 1 and 2, respectively. 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 and low expression levels are red are blue, respectively. Missing values are black. Cells are rearranged by the corresponding posterior estimate of their subpopulation indicator, $\hat{\lambda}_{i,n}$. Yellow horizontal lines separate cells by different subpopulations.



(e) $\hat{\boldsymbol{Z}}'_3$ and $\hat{\boldsymbol{w}}_3$ (f) $y_{3,n,j}$

Figure A.14 (continued): [CB NK cell data] Inference obtained by VI is illustrated. $\hat{\boldsymbol{Z}}'_i$ and $\hat{\boldsymbol{w}}_i$ of sample 3 illustrated in panel (e), with markers that are expressed dented by black and not expressed by white. Only subpopulations with $\hat{w}_{i,k} > 1\%$ are included. Heatmaps of $\boldsymbol{y}_i$ are shown in panels (b) and (d) for samples 1 and 2, respectively. 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 and low expression levels are red are blue, respectively. Missing values are black. Cells are rearranged by the corresponding posterior estimate of their subpopulation indicator, $\hat{\lambda}_{i,n}$. Yellow horizontal lines separate cells by different subpopulations.

Marker	p=0.85	p=0.90	p=0.95	Reason for exclusion
2B4	included	included	included	NA
2DL1	excluded	excluded	included	(-)
2DL3	excluded	included	included	(-)
2DS4	excluded	excluded	excluded	(-)
3DL1	excluded	included	included	(-)
CCR7	excluded	excluded	included	(-)
CD158B	included	included	included	NA
CD16	included	included	included	NA
CD25	excluded	excluded	excluded	(-)
CD27	included	included	included	NA
CD57	excluded	excluded	excluded	(-)
CD62L	included	included	included	NA
CD8	included	included	included	NA
CD94	included	included	included	NA
CKIT	excluded	excluded	excluded	(-)
DNAM1	included	included	included	NA
EOMES	included	included	included	NA
GRA	excluded	excluded	excluded	(0)
GRB	excluded	excluded	excluded	(0)
KLRG1	included	included	included	NA
LFA1	excluded	excluded	excluded	(-)
NKG2A	included	included	included	NA
NKG2C	excluded	included	included	(-)
NKG2D	included	included	included	NA
NKP30	included	included	included	NA
PERFORIN	excluded	excluded	excluded	(0)
SIGLEC7	included	included	included	NA
SYK	excluded	excluded	excluded	(+)
TBET	included	included	included	NA
TIGIT	included	included	included	NA
TRAIL	excluded	excluded	excluded	(-)
ZAP70	included	included	included	NA

Table A.8: Inclusion of markers in the analysis for various preprocessing threshold p, and the reasons for exclusion, if applicable. (-) denotes that expression levels were mostly negative or missing and (+) denotes that expression levels were mostly positive, (0) denotes that expressions were mostly around 0.

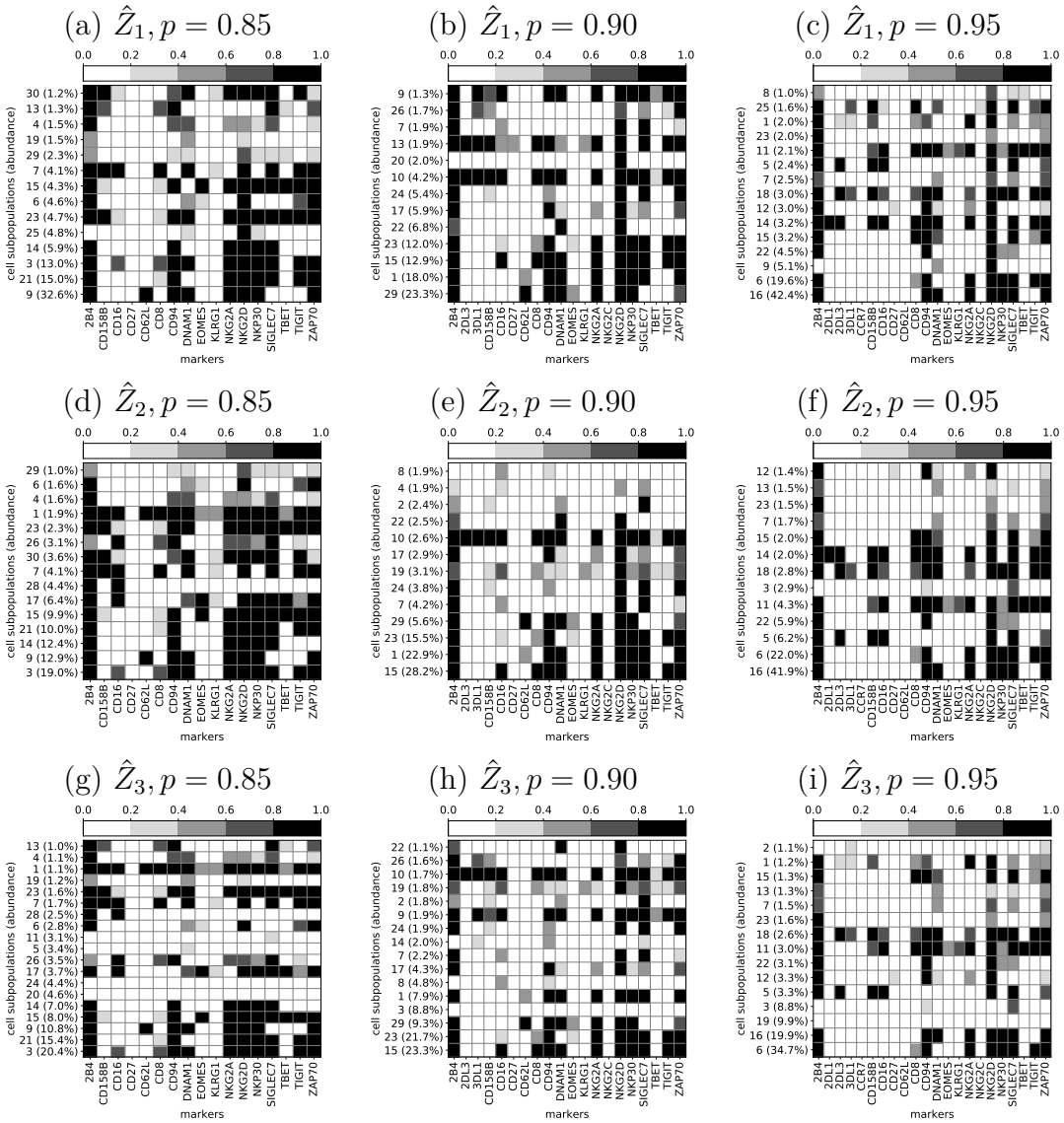


Figure A.15: Sensitivity of estimates $\hat{Z}_i$ to specification of p in preprocessing, for $i = 1, 2, 3$ and $p = 0.85, 0.90, 0.95$, using ADVI.

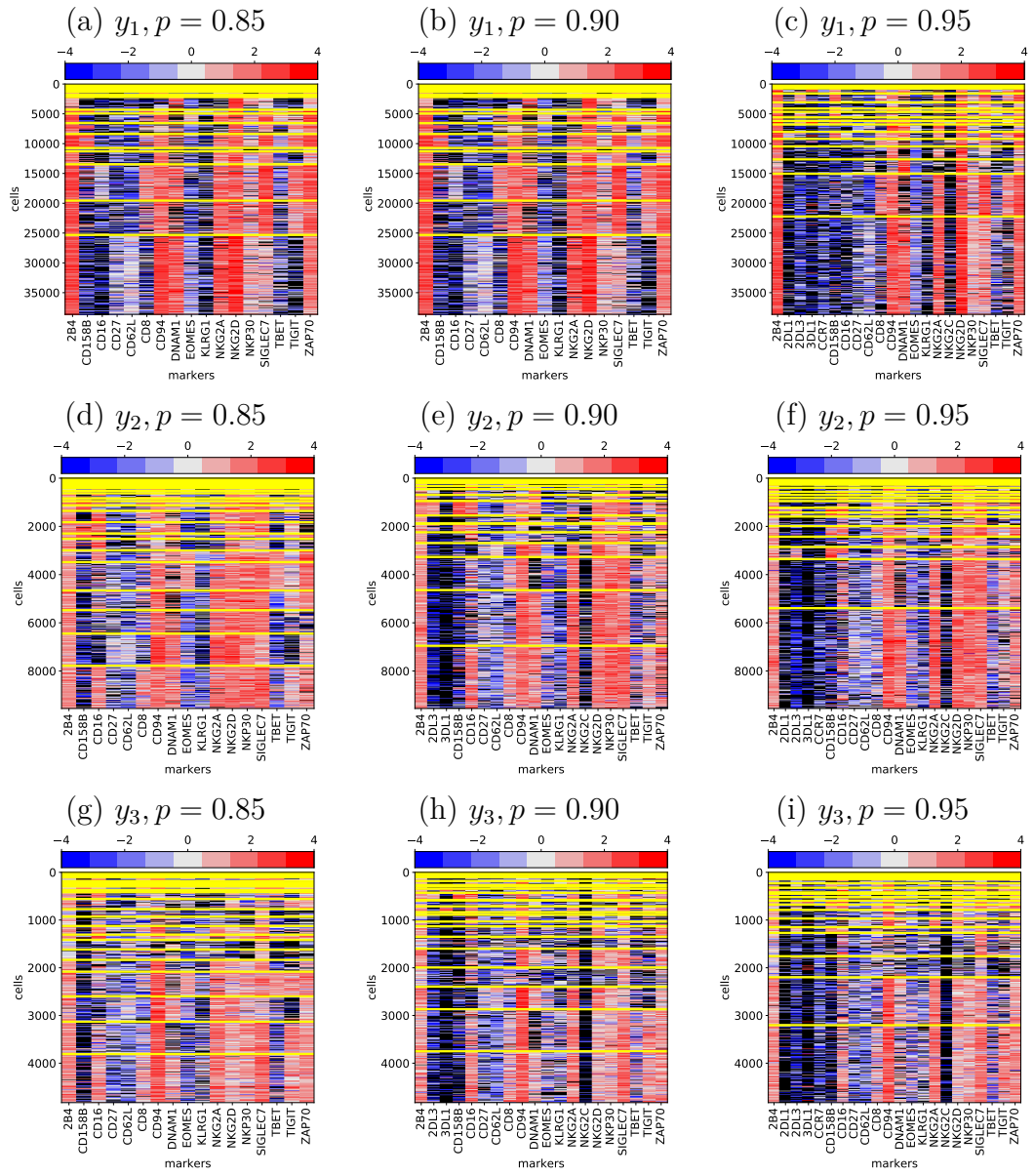


Figure A.16: Heatmaps of y_i with cells sorted by subpopulation membership for each specification of p in preprocessing, for $i = 1, 2, 3$ and $p = 0.85, 0.90, 0.95$, using ADVI.

Appendix B

A Bayesian Model for Identifying Distinct Features that Define Cell Subpopulations from Cytometry Data

B.1 Prior Calibration

Recall that $\boldsymbol{\phi} = (\phi_1, \phi_2)$ are the hyperparameters of the repulsion function $f_\phi(d)$. We establish numerical values of $\boldsymbol{\phi}$ as follows; we find the values of $\boldsymbol{\phi}$ such that

$$\Pr\left(\min_{1\leq k_1<k_2\leq K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \geq \underline{d} \mid \boldsymbol{\phi}\right) > \underline{p}, \tag{B.1}$$

where $\underline{d}$ and $\underline{p}$ are thresholds of the minimum difference of a pair of features and the probability, respectively. To simplify the search, we fix ϕ_1 and find the smallest

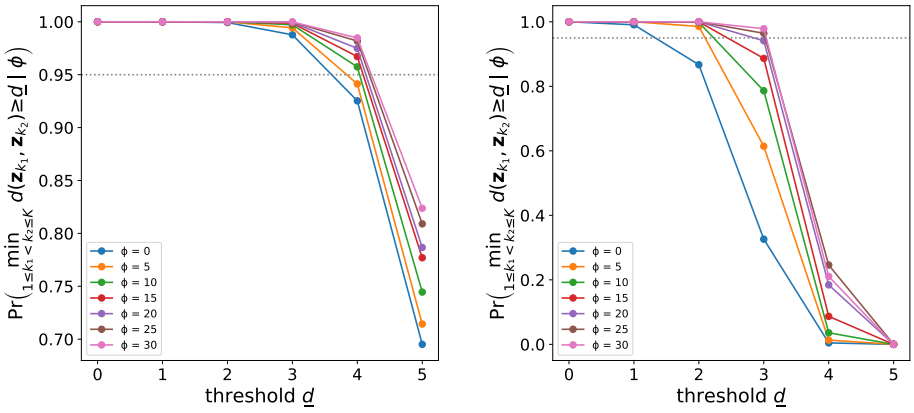
value of ϕ_2 that satisfies (B.1). Since it is very hard to analytically evaluate (B.1), we used the importance sampling scheme and numerically evaluate it for different values of $\boldsymbol{\phi}$ and $\underline{d}$ as follows.

1. Simulate a sample of $\boldsymbol{Z}$ of size B, $\{\boldsymbol{Z}^{(b)}, b = 1, \ldots, B\}$, from the distribution $p'(\boldsymbol{Z} \mid \boldsymbol{v}) = \prod_{j=1}^{J} \prod_{k=1}^{K} v_k^{z_{j,k}} (1 - v_k)^{1 - z_{j,k}}$, where B is a sufficiently large number.

2. Let $g(\boldsymbol{Z}) = \mathrm{I}(\min_{1 \le k_1 < k_2 \le K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \ge \underline{d})$, where $\mathrm{I}(A)$ is a binary indicator function, and approximate (B.1) using $\{\boldsymbol{Z}^{(b)}, b = 1, \ldots, B\}$;

$$
\begin{aligned}
&\Pr\left(\min_{1 \le k_1 < k_2 \le K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \ge \underline{d} \mid \boldsymbol{\phi}\right) \\
=&\mathrm{E}\left[g(\boldsymbol{Z}) \mid \boldsymbol{v}, \boldsymbol{\phi}\right] \\
=&\sum_{Z \in \mathcal{Z}} g(Z) \cdot p(\boldsymbol{Z} \mid \boldsymbol{v}, \boldsymbol{\phi}) \\
\approx&\frac{\sum_{b=1}^{B} g(\boldsymbol{Z}^{(b)}) \cdot \prod_{k_2=2}^{K} \prod_{k_1=1}^{k_2-1} f_\phi(d(\mathbf{z}_{k_1}^{(b)}, \mathbf{z}_{k_2}^{(b)}))}{\sum_{b=1}^{B} \prod_{k_2=2}^{K} \prod_{k_1=1}^{k_2-1} f_\phi(d(\mathbf{z}_{k_1}^{(b)}, \mathbf{z}_{k_2}^{(b)}))}.
\end{aligned}
$$

For our calibration, we used $v_k = 0.5$ for all k. The plots in Figure B.1 illustrate $\Pr(\min_{1 \le k_1 < k_2 \le K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \ge \underline{d} \mid \boldsymbol{\phi})$ as a function of $\underline{d}$. We also vary the value of ϕ_2, while fixing $\phi_1 = 1$ for simplicity. For the simulation studies, we used $K = 15$, $\underline{d} = 4$ and $\underline{p} = 0.95$. As shown in panel (a) of the figure, we let $\phi_1 = 1$ and set $\phi_2 = 10$, the smallest value of ϕ_2 satisfying (B.1). Similarly, for the real data analysis, we used $K = 25$, $\underline{d} = 3$ and $\underline{p} = 0.95$. As shown in panel (b) of the figure, we let $\phi_1 = 1$ and chose $\phi_2 = 25$ for the real data analysis. $\underline{d} = 4$ and 3 implies that any pair of features are expected to differ in approximately 20% of the markers.



(a) $K = 15$ (b) $K = 25$

Figure B.1: Plot of $\Pr(\min_{1\le k_1<k_2\le K} d(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) \ge d \mid \boldsymbol{\phi})$ as a function of d. ϕ_2 is also varied while fixing $\phi_1 = 1$. $K = 15$ and 25 are in panels (a) and (b), respectively.

B.2 Supplementary Posterior Computation

B.2.1 Supplementary Material for Missing Data Mechanism

When the expression level of a marker is very weak, its expression level is not recorded in a cytometry experiment due to some experimental artifacts. Missing values in CyTOF data are missing not at random. Following the expert knowledge that the expression value is not recorded when its potentially observable value is low, we build a missingness mechanism and impute missing expression levels during the posterior simulation to accurately account for uncertainty. Similar to the missingness mechanism used in Lui et al. (2020), we consider a static logistic regression model as follows; Let $o_{i,n,j} = 1$ if $y_{i,n,j}$ is observed; and 0, otherwise. Then, let $\rho_{i,n,j}(y_{i,n,j})$ be the probability that the expression level of marker j in cell n of sample i is missing given its potentially observed numerical value, $y_{i,n,j}$, i.e., $\Pr(o_{i,n,j} = 0 \mid y_{i,n,j}) = \rho_{i,n,j}(y_{i,n,j})$, and assume a Bernoulli distribution for $o_{i,n,j}$,

$o_{i,n,j} \mid \rho_{i,n,j}(y_{i,n,j} \overset{ind}{\sim} \text{Bernoulli}(1 - \rho_{i,n,j}(y_{i,n,j}))$. Combining with the marginal distribution of $y_{i,n,j}$ in (2) of the main text, we have a joint model of $y_{i,n,j}$ and $o_{i,n,j}$. We assume the following quadratic model for $\rho_{y_{i,n,j}}$;

$$\rho_{i,n,j}(y_{i,n,j}) = \text{logistic}\left(\beta_{0,i} + \beta_{1,i} \cdot y_{i,n,j} + \beta_{2,i} \cdot y_{i,n,j}^2\right),$$

where $(\beta_{0,i}, \beta_{1,i}, \beta_{2,i})$ are fixed. We calibrate $\boldsymbol{\beta}_i$ using the expert knowledge such that imputed values of the missing expression levels take negative values with large probabilities. Specifically, we (empirically) used the minimum, first quantile, and median of negative $y_{i,n,j}$ values, and set their $\rho_{i,n,j}$ values to 0.05, 0.80, and 0.50 respectively, to solve for β_i. For more details, see Lui et al. (2020). We let $\boldsymbol{\theta}$ represents all model parameters and write the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{o}$

$$\begin{aligned} p\left(\boldsymbol{y}, \boldsymbol{o} \mid \theta\right) = & \prod_{i=1}^{I} \prod_{n=1}^{N_i} \sum_{k=1}^{K} w_{i,k} \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^{\star}_{z_{j,k},\ell}\right)^2}{2\sigma_i^2} \right\} \right) \\ & \times \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-o_{i,n,j}} (1 - \rho_{i,n,j})^{o_{i,n,j}}, \end{aligned} \tag{B.2}$$

Since $\rho_{i,n,j}$ is a constant, it can be dropped for observed $y_{i,n,j}$ from (B.2). Thus, we have

$$\begin{aligned} p\left(\boldsymbol{y}, \boldsymbol{o} \mid \boldsymbol{\theta}\right) \propto & \prod_{i=1}^{I} \prod_{n=1}^{N_i} \sum_{k=1}^{K} w_{i,k} \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^{\star}_{z_{j,k},\ell}\right)^2}{2\sigma_i^2} \right\} \right) \\ & \times \prod_{i=1}^{I} \prod_{n=1}^{N_i} \prod_{j=1}^{J} \rho_{i,n,j}^{1-o_{i,n,j}}. \end{aligned} \tag{B.3}$$

B.2.2 Supplementary Material for Parallel Tempering

Parallel tempering (PT) (Earl and Deem 2005) is a general MCMC technique to increase the mixing rate of a model that suffers from poor mixing. In posterior inference using PT, multiple MCMC chains of various "temperatures" are modeled and updated in parallel for a given model. Typically, each chain samples from the posterior distribution $\pi_\tau(\theta) \propto \mathcal{L}(\theta)^{1/\tau} p(\theta)$, where $\tau \geq 1$ is called the temperature, θ are model parameters, $p(\theta)$ is the prior, $\mathcal{L}(\theta)$ is the likelihood, and $\pi_\tau(\theta)$ is the posterior of θ under temperature τ. At higher temperatures, a greater area of the sample space can be explored. When the temperature is 1, the original distribution is recovered. T chains are run in parallel, for a sequence of increasing temperatures $\tau_1, \tau_2, \ldots, \tau_T$ where $T \in \mathbb{N}$. At regular intervals, the entire states in pairs of chains are swapped with some probability. For example, a possible swapping scheme is that at every M iterations of the MCMC, starting from the hottest (highest-temperature) chain, the states θ_t and θ_{t-1} at adjacent temperatures τ_t and τ_{t-1} are swapped with probability

$$\begin{aligned}
\alpha_{ij} &= \min\left\{1, \frac{\pi_{\tau_t}(\theta_{t-1}) \cdot \pi_{\tau_{t-1}}(\theta_t)}{\pi_{\tau_t}(\theta_t) \cdot \pi_{\tau_{t-1}}(\theta_{t-1})}\right\} \\
&= \min\left\{1, \frac{\mathcal{L}(\theta_{t-1})^{1/\tau_t} p(\theta_{t-1}) \cdot \mathcal{L}(\theta_t)^{1/\tau_{t-1}} p(\theta_t)}{\mathcal{L}(\theta_t)^{1/\tau_t} p(\theta_t) \cdot \mathcal{L}(\theta_{t-1})^{1/\tau_{t-1}} p(\theta_{t-1})}\right\} \\
&= \min\left\{1, \frac{\mathcal{L}(\theta_{t-1})^{1/\tau_t} \cdot \mathcal{L}(\theta_t)^{1/\tau_{t-1}}}{\mathcal{L}(\theta_t)^{1/\tau_t} \cdot \mathcal{L}(\theta_{t-1})^{1/\tau_{t-1}}}\right\}.
\end{aligned}$$

Swapping with this probability preserves the detailed balance of the underlying simulation.

Tawn et al. (2020) showed that when tempering (simulated or parallel) is applied to mixture models, it is conceivable that mixture components originally with substantially smaller weights can have an dominantly large mixture weight at

high temperatures. To preserve the original weights of these mixture components, they proposed a weight-stabilizing tempering scheme in which only the kernels of the mixture components are tempered. Applied to our model, accounting for possibly missing data, the likelihood $\mathcal{L}_\tau(\boldsymbol{\theta})$ at temperature τ is

$$\mathcal{L}_\tau(\boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{n=1}^{N_i}\sum_{k=1}^{K} w_{i,k} \prod_{j=1}^{J}\sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{\frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,k},\ell}\right)^2}{2\sigma_i^2}\right\}\right)^{1/\tau} \times \prod_{i=1}^{I}\prod_{n=1}^{N_i}\prod_{j=1}^{J} \rho_{i,n,j}^{1-o_{i,n,j}}. \tag{B.4}$$

B.2.3 Full Conditional Distributions of Model Parameters

This section presents detailed derivations of full conditional distributions for each model parameter.

- Full Conditional for $z_{j,k}$

  Each $z_{j,k}$ can be updated sequentially, as follows:

$$\begin{aligned}
& p_\tau(z_{j,k} = z \mid \lambda, \boldsymbol{Z}_{-(j,k)}, \text{rest}) \\
\propto\ & p(z_{j,k} = z \mid \boldsymbol{v}, f_\phi) \cdot p_\tau(\boldsymbol{y} \mid \boldsymbol{Z}, \lambda, \text{rest}) \\
\propto\ & p(z_{j,k} = z \mid \boldsymbol{v}, f_\phi) \cdot \prod_{i=1}^{I}\prod_{n=1}^{N_i}\left(\sum_{\ell=1}^{L_z} \eta_{i,j,\ell}^z \cdot p(y_{i,n,j} \mid \mu^\star_{z,\ell}, \sigma_i^2)^{1/\tau}\right)^{\mathbb{1}\{\lambda_{i,n}=k\}} \\
\propto\ & p(z_{j,k} = z \mid \boldsymbol{v}, f_\phi) \cdot \prod_{i=1}^{I}\prod_{n=1}^{N_i}\left(\sum_{\ell=1}^{L_z} \eta_{i,j,\ell}^z \cdot p(y_{i,n,j} \mid \mu^\star_{z,\ell}, \sigma_i^2\tau)\right)^{\mathbb{1}\{\lambda_{i,n}=k\}}
\end{aligned}$$

  where $\boldsymbol{Z}_{-(j,k)}$ refers to all elements in $\boldsymbol{Z}$ except element $z_{j,k}$, and $z \in \{0, 1\}$. Note that γ is marginalized over because proposing a new value value of $z_{j,k}$ is not sensible when conditioned on γ.

We can further marginalize over $\lambda_{i,n}$ to achieve better mixing. Hence,

$$\begin{aligned}
p_\tau(z_{j,k} = z \mid \boldsymbol{Z}_{-(j,k)}, \text{rest}) &\propto p(z_{j,k} = z \mid \boldsymbol{v}, f_\phi) \cdot p_\tau(\boldsymbol{y} \mid \boldsymbol{Z}, \text{rest}) \\
&\propto v_k^{z_{j,k}}(1 - v_k)^{1-z_{j,k}} \cdot \prod_{h \neq k} f_\phi(d(\mathbf{z}_h, \mathbf{z}_k)) \times \\
&\quad \prod_{i=1}^{I} \prod_{n=1}^{N_i} \sum_{h=1}^{K} W_{i,h} \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,h}}} \eta_{i,j,\ell}^{z_{j,h}} \cdot p(y_{i,n,j} \mid \mu^\star_{z,\ell}, \sigma_i^2)^{1/\tau} \\
&\propto v_k^{z_{j,k}}(1 - v_k)^{1-z_{j,k}} \cdot \prod_{h \neq k} f_\phi(d(\mathbf{z}_h, \mathbf{z}_k)) \times \\
&\quad \prod_{i=1}^{I} \prod_{n=1}^{N_i} \sum_{h=1}^{K} W_{i,h} \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,h}}} \eta_{i,j,\ell}^{z_{j,h}} \cdot p(y_{i,n,j} \mid \mu^\star_{z,\ell}, \sigma_i^2 \tau).
\end{aligned}$$

Thus, $(\boldsymbol{Z}_{j,k} \mid \boldsymbol{Z}_{-(j,k)}, \text{rest})$ can be updated using a Gibbs step. The normalizing constant is simply the sum of the (un-normalized) term above evaluated at $z = \{0, 1\}$.

- Full Conditional for v_k

  Recall that the prior distribution for v_k is $v_k \mid \alpha \overset{ind}{\sim} \text{Beta}(\alpha/K, 1)$, for $k = 1, \ldots, K$. Thus, $p(v_k \mid \alpha) = \frac{\alpha}{K} v_k^{\alpha/K - 1}$.

$$\begin{aligned}
p(v_k \mid \boldsymbol{y}, \text{rest}) &\propto p(v_k) \prod_{j=1}^{J} p(z_{j,k} \mid v_k, f_\phi) \\
&\propto \frac{\alpha}{K} v_k^{\alpha/K - 1} \prod_{j=1}^{J} v_k^{z_{j,k}} (1 - v_k)^{1 - z_{j,k}} \\
&\propto v_k^{\alpha/K + \sum_{j=1}^{J} z_{j,k} - 1} (1 - v_k)^{J - \sum_{j=1}^{J} z_{j,k}}
\end{aligned}$$

$$\therefore v_k \mid \boldsymbol{y}, \text{rest} \sim \text{Beta}\left(\alpha/K + \sum_{j=1}^{J} z_{j,k},\ J + 1 - \sum_{j=1}^{J} z_{j,k}\right).$$

- Full Conditional for $r_{i,k}$ (marginalized over $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$)

  Recall that $r_{i,k} \mid p_i \sim \text{Bernoulli}(p_i)$. Since $\lambda_{i,n}$ and $r_{i,k}$ are highly dependent,

we will marginalize over $\lambda_{i,n}$ and use a Metropolis step to update $r_{i,k}$, where the proposal step flips $r_{i,k}$ (from 0 to 1 and vice versa).

$$\begin{aligned} p_\tau(r_{i,k} \mid \text{rest}) &\propto p(r_{i,k} \mid p_i) \cdot p_\tau(\boldsymbol{y} \mid \text{rest}) \\ &\propto p_i^{r_{i,k}}(1-p_i)^{1-r_{i,k}} \cdot \prod_{n=1}^{N_i} \sum_{h=1}^{K} w_{i,h} \cdot p_\tau(\boldsymbol{y}_{i,n} \mid \mathbf{z}_h, \text{rest}) \end{aligned}$$

where

$$\begin{aligned} p_\tau(\boldsymbol{y}_{i,n} \mid \mathbf{z}_h, \text{rest}) &\propto \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,h}}} \eta_{i,j,\ell}^{z_{j,h}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,h},\ell}\right)^2}{2\sigma_i^2} \right\} \right)^{1/\tau} \\ &\propto \prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,h}}} \eta_{i,j,\ell}^{z_{j,h}} \cdot \frac{1}{\sqrt{2\pi\sigma_i^2\tau}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,h},\ell}\right)^2}{2\sigma_i^2\tau} \right\}. \end{aligned}$$

Note that since $\{\boldsymbol{Z}, \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^2, \boldsymbol{\eta}\}$ are held constant throughout the updates of each $r_{i,k}$, the computation of $p(\boldsymbol{y}_{i,n} \mid \mathbf{z}_h, \text{rest})$ needs to be computed (and then cached) only once per update of the entire $\boldsymbol{r}$ matrix.

Let $w_{i,k}^{(z)}$ be computed as $w_{i,k}$ with the current $\boldsymbol{r}_i$ but replacing $r_{i,k}$ with $z \in \{0, 1\}$. Then for a current $r_{i,k}$, the acceptance ratio can be computed as

$$\min\left\{ 1, \left(\frac{p_i}{1-p_i}\right)^{1-r_{i,k}} \left(\frac{1-p_i}{p_i}\right)^{r_{i,k}} \cdot \frac{\prod_{n=1}^{N_i} \sum_{h=1}^{K} w_{i,h}^{(1-r_{i,k})} \cdot p_\tau(\boldsymbol{y}_{i,n} \mid \mathbf{z}_h, \text{rest})}{\prod_{n=1}^{N_i} \sum_{h=1}^{K} w_{i,h}^{(r_{i,k})} \cdot p_\tau(\boldsymbol{y}_{i,n} \mid \mathbf{z}_{k'}, \text{rest})} \right\}.$$

- Full Conditional for p_i

$$
\begin{aligned}
p(p_i \mid \text{rest}) &\propto p(p_i) \cdot \prod_{k=1}^{K} p(r_{i,k} \mid p_i) \\
&\propto p_i^{a_p - 1}(1-p_i)^{b_p - 1} \cdot p_i^{\sum_{k=1}^{K} r_{i,k}} (1-p_i)^{K - \sum_{k=1}^{K} r_{i,k}} \\
\therefore p_i \mid \boldsymbol{r} &\sim \text{Beta}\left(a_p + \sum_{k=1}^{K} r_{i,k},\ b_p + K - \sum_{k=1}^{K} r_{i,k}\right).
\end{aligned}
$$

- Full Conditional for $w^\star_{i,k}$ (conditioned on $\boldsymbol{\lambda}$)

  Recall that $w^\star_{i,k} \sim \text{Gamma}(a_w, 1)$, where $a_W = K^{-1}$. And $w_{i,k} = \dfrac{w^\star_{i,k} \cdot r_{i,k}}{\sum_{\ell=1}^{K} w^\star_{i,\ell} \cdot r_{i,\ell}}$.

$$
\begin{aligned}
p(w^\star_{i,k} \mid \text{rest}) &\propto p(w^\star_{i,k}) \cdot p(\boldsymbol{\lambda}_i \mid \boldsymbol{w}_i) \\
&\propto (w^\star_{i,k})^{a_w - 1} \exp(-w^\star_{i,k}) \cdot \left(\prod_{n=1}^{N_i} p(\lambda_{i,n} \mid \boldsymbol{w}_i)\right) \\
&\propto (w^\star_{i,k})^{a_w - 1} \exp(-w^\star_{i,k}) \cdot \left(\prod_{n=1}^{N_i} w_{i,\lambda_{i,n}}\right)^{r_{i,k}}
\end{aligned}
$$

Note that if $r_{i,k} = 0$, then $w^\star_{i,k}$ is independent of the data. So $w^\star_{i,k}$ will simply be sampled from the prior. If $r_{i,k} = 1$,

$$
\begin{aligned}
p(w^\star_{i,k} \mid \text{rest}) &\propto (w^\star_{i,k})^{a_w - 1} \exp(-w^\star_{i,k}) \cdot \left(\prod_{n=1}^{N_i} \frac{w^\star_{i,\lambda_{i,n}} \cdot r_{i,\lambda_{i,n}}}{w^\star_{i,k} \cdot r_{i,k} + \sum_{\ell \neq k} w^\star_{i,\ell} \cdot r_{i,\ell}}\right) \\
&\propto \frac{(w^\star_{i,k})^{a_W - 1} \exp(-w^\star_{i,k})}{\left(w^\star_{i,k} \cdot r_{i,k} + \sum_{\ell \neq k} w^\star_{i,\ell} \cdot r_{i,\ell}\right)^{N_i}} \cdot \prod_{n=1}^{N_i} (w^\star_{i,k})^{\mathbb{1}\{\lambda_{i,n} = k\}} \\
&\propto \frac{(w^\star_{i,k})^{a_W + (\sum_{n=1}^{N_i} \mathbb{1}\{\lambda_{i,n} = k\}) - 1} \exp(-w^\star_{i,k})}{\left(w^\star_{i,k} + \sum_{\ell \neq k} w^\star_{i,\ell} \cdot r_{i,\ell}\right)^{N_i}}.
\end{aligned}
$$

Since the full conditional distribution for $w^\star_{i,k}$ cannot be directly sampled

from, it may be sampled from by a Metropolis step with a Normal proposal distribution. The parameter first needs to be log-transformed. Let the full conditional of the transformed parameter be $p(\phi \mid \boldsymbol{y}, \text{rest}) = p_{w^{\star}_{i,k}}(\exp(\phi) \mid \boldsymbol{y}, \text{rest}) \exp(\phi)$. Then, the proposed state of the transformed parameter (ϕ) is accepted with probability

$$\min\left\{1, \frac{p(\tilde{\phi} \mid \boldsymbol{y}, \text{rest})}{p(\phi \mid \boldsymbol{y}, \text{rest})}\right\}.$$

Exponentiating the updated value for ϕ returns the updated value for $w^{\star}_{i,k}$.

- Full Conditional for α

  Recall that $v_k \mid \alpha \sim \text{Beta}(\alpha/K, 1)$ and $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ where $\text{Gamma}(a, b)$ denotes a Gamma distribution with mean a/b. Thus, the full conditional for α can be computed as:

$$\begin{aligned}
p(\alpha \mid \boldsymbol{y}, \text{rest}) &\propto p(\alpha) \times \prod_{k=1}^{K} p(v_k \mid \alpha) \\
&\propto \alpha^{a_\alpha - 1} \exp\{-b_\alpha \alpha\} \times \prod_{k=1}^{K} \alpha \; v_k^{\alpha/K} \\
&\propto \alpha^{a_\alpha + K - 1} \exp\left\{-\alpha\left(b_\alpha - \frac{\sum_{k=1}^{K} \log v_k}{K}\right)\right\} \\
\therefore \alpha \mid \boldsymbol{y}, \text{rest} &\sim \text{Gamma}\left(a_\alpha + K,\; b_\alpha - \frac{\sum_{k=1}^{K} \log v_k}{K}\right).
\end{aligned}$$

- Full Conditional for $\lambda_{i,n}$ (marginalized over $\boldsymbol{\gamma}$)

Recall that $P(\lambda_{i,n} = k \mid \boldsymbol{w}_i) = w_{i,k}$, for $k \in \{1, \ldots, K\}$.

$$
\begin{aligned}
& P_\tau(\lambda_{i,n} = k \mid \boldsymbol{y}, \text{rest}) \\
\propto\ & p(\lambda_{i,n} = k) \cdot p_\tau(\boldsymbol{y} \mid \lambda_{i,n} = k, \text{rest}) \\
\propto\ & w_{i,k} \cdot \left(\prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,k},\ell}\right)^2}{2\sigma_i^2} \right\} \right)^{1/\tau} \right) \\
\propto\ & W_{i,k} \cdot \left(\prod_{j=1}^{J} \sum_{\ell=1}^{L_{z_{j,k}}} \eta_{i,j,\ell}^{z_{j,k}} \cdot \frac{1}{\sqrt{2\pi\sigma_i^2\tau}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,k},\ell}\right)^2}{2\sigma_i^2\tau} \right\} \right).
\end{aligned}
$$

Thus, $P_\tau(\lambda_{i,n} = k \mid \boldsymbol{y}, \text{rest}) = \dfrac{P_\tau(\lambda_{i,n} = k \mid \boldsymbol{y}, \text{rest})}{\sum_{h=1}^{K} P_\tau(\lambda_{i,n} = h \mid \boldsymbol{y}, \text{rest})}$, for $k \in \{1, \ldots, K\}$.

- Full Conditional for $\delta_{z,\ell}$ (conditioned on $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$)

  For $\delta_{1,\ell}$, let $S_{1,i,\ell} = \left\{(i,n,j) : \left(z_{j,\lambda_{i,n}} = 1 \ \cap\ \gamma_{i,n,j} \geq \ell\right)\right\}$ and $|S_{1,i,\ell}|$ the cardinality of $S_{1,i,\ell}$.

$$
\begin{aligned}
& p_\tau(\delta_{1,\ell} \mid \boldsymbol{y}, \text{rest}) \\
\propto\ & p(\delta_{1,\ell} \mid \psi_1, \kappa_1^2) \times p_\tau(\boldsymbol{y} \mid \delta_{1,\ell}, \text{rest}) \\
\propto\ & 1(\delta_{1,\ell} \geq 0) \times \exp\left\{ \frac{-(\delta_{1,\ell} - \psi_1)^2}{2\kappa_1^2} \right\} \times \\
& \left(\prod_{i=1}^{I} \prod_{(i,n,j) \in S_{1,i,\ell}} \exp\left\{ -\left(y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} \delta_{1,r} \right)^2 \Big/ 2\sigma_i^2\tau \right\} \right) \\
\propto\ & \exp\left\{ -\frac{(\delta_{1,\ell})^2}{2} \left(\frac{1}{\kappa_1^2} + \sum_{i=1}^{I} \frac{|S_{1,i,\ell}|}{\sigma_i^2\tau} \right) + \delta_{1,\ell} \left(\frac{\psi_1}{\kappa_1^2} + \sum_{i=1}^{I} \sum_{S_{1,i,\ell}} \frac{g_{i,n,j}}{\sigma_i^2\tau} \right) \right\} \times \\
& 1(\delta_{1i\ell} \geq 0),
\end{aligned}
$$

where $g_{i,n,j} = y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} (\delta_{1r})^{1(r\neq\ell)}$. Therefore, $\delta_{1,\ell} \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim}$

$$\text{TN}^{+}\left(\frac{\psi_1 + \kappa_1^2 \sum_{i=1}^{I} \sum_{S_{1,i,\ell}} (g_{i,n,j}/(\sigma_i^2\tau))}{1 + \kappa_1^2 \sum_{i=1}^{I} (|S_{1,i,\ell}|/(\sigma_i^2\tau))}, \frac{\kappa_1^2}{1 + \kappa_1^2 \sum_{i=1}^{I} (|S_{1,i,\ell}|/(\sigma_i^2\tau))}\right).$$

Similarly for $\delta_{0,\ell}$, let $S_{0,i,\ell} = \left\{(i,n,j) : \left(z_{j,\lambda_{i,n}} = 0 \ \cap \ \gamma_{i,n,j} \geq \ell\right)\right\}$ and $|S_{0,i,\ell}|$ be the cardinality of $S_{0,i,\ell}$. Therefore, $\delta_{0l} \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim}$

$$\text{TN}^{+}\left(\frac{\psi_0 + \kappa_0^2 \sum_{i=1}^{I} \sum_{S_{0,i,\ell}} (g_{i,n,j}/(\sigma_i^2\tau))}{1 + \kappa_0^2 \sum_{i=1}^{I} (|S_{0,i,\ell}|/(\sigma_i^2\tau))}, \frac{\kappa_0^2}{1 + \kappa_0^2 \sum_{i=1}^{I} (|S_{0,i,\ell}|/(\sigma_i^2\tau))}\right),$$

where $g_{i,n,j} = -y_{i,n,j} - \sum_{r=1}^{\gamma_{i,n,j}} (\delta_{0r})^{1(r\neq\ell)}$.

- Full Conditional for σ_i^2 (conditioned on $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$)

$$\begin{aligned}
&p_\tau(\sigma_i^2 \mid \boldsymbol{y}, \text{rest}) \\
\propto &p(\sigma_i^2) \times p_\tau(\boldsymbol{y} \mid \sigma_i^2, \text{rest}) \\
\propto &(\sigma_i^2)^{-a_\sigma - 1} \exp\left\{-\frac{b_\sigma}{\sigma_i^2}\right\} \left(\prod_{j=1}^{J} \prod_{n=1}^{N_i} \left\{\frac{1}{\sqrt{2\sigma_i^2}} \exp\left\{\frac{-(y_{i,n,j} - \mu_{i,n,j})^2}{2\sigma_i^2}\right\}\right\}\right)^{1/\tau} \\
\propto &(\sigma_i^2)^{-a_\sigma - 1} \exp\left\{-\frac{b_\sigma}{\sigma_i^2}\right\} \left(\prod_{j=1}^{J} \prod_{n=1}^{N_i} \left\{(\sigma_i^2)^{-1/2\tau} \exp\left\{\frac{-(y_{i,n,j} - \mu_{i,n,j})^2}{2\sigma_i^2\tau}\right\}\right\}\right) \\
\propto &(\sigma_i^2)^{-\left(a_\sigma + \frac{N_i J}{2\tau}\right) - 1} \exp\left\{-\left(\frac{1}{\sigma_i^2}\right)\left(b_\sigma + \sum_{j=1}^{J} \sum_{n=1}^{N_i} \frac{(y_{i,n,j} - \mu_{i,n,j})^2}{2\tau}\right)\right\}.
\end{aligned}$$

$$\therefore \sigma_i^2 \mid \boldsymbol{y}, \text{rest} \overset{ind}{\sim} \text{InverseGamma}\left(a_\sigma + \frac{N_i J}{2\tau}, \ \ b_\sigma + \sum_{j=1}^{J} \sum_{n=1}^{N_i} \frac{(y_{i,n,j} - \mu_{i,n,j})^2}{2\tau}\right),$$

where $\mu_{i,n,j} = \mu^{\star}_{z_{j,\lambda_{i,n}},\gamma_{i,n,j}}$.

- Full Conditional for $\boldsymbol{\gamma}$ (conditioned on $\boldsymbol{\lambda}$) The prior for $\gamma_{i,n,j}$ is $p(\gamma_{i,n,j} = \ell \mid z_{j,\lambda_{i,n}} = z, \eta^{z}_{i,j,\ell}) = \eta^{z}_{i,j,\ell}$, where $\ell \in \{1, ..., L\}$.

$$
\begin{aligned}
p_{\tau}(\gamma_{i,n,j} = \ell \mid \boldsymbol{y}, z_{j,\lambda_{i,n}} = z, \text{rest}) &\propto p(\gamma_{i,n,j} = \ell) \cdot p(y_{i,n,j} \mid \gamma_{i,n,j} = \ell, \text{rest})^{1/\tau} \\
&\propto p(\gamma_{i,n,j} = \ell) \cdot p(y_{i,n,j} \mid \mu^{\star}_{z,\ell}, \sigma^2_i, \text{rest})^{1/\tau} \\
&\propto \eta^{z}_{i,j,\ell} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2_i}} \exp\left\{ \frac{-\left(y_{i,n,j} - \mu^{\star}_{z,\ell}\right)^2}{2\sigma^2_i} \right\} \right)^{1/\tau} \\
&\propto \eta^{z}_{i,j,\ell} \cdot \exp\left\{ -\frac{(y_{i,n,j} - \mu^{\star}_{z,\ell})^2}{2\sigma^2_i \tau} \right\}
\end{aligned}
$$

The normalizing constant is obtained by summing the last expression over $\ell = 1, ..., L^{z_{j,\lambda_{i,n}}}$. Moreover, since ℓ is discrete, a Gibbs update can be done on $\gamma_{i,n,j}$.

- Full Conditional for $\boldsymbol{\eta}^{z}_{i,j}$ (conditioned on $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$)

The prior for $\boldsymbol{\eta}^{z}_{i,j}$ is $\boldsymbol{\eta}^{z}_{i,j} \sim \text{Dirichlet}_{L_z}(a_{\eta^z})$, for $z \in \{0, 1\}$. Thus, the full conditional for $\boldsymbol{\eta}^{z}_{i,j}$ is:

$$
\begin{aligned}
p(\boldsymbol{\eta}^{z}_{i,j} \mid \text{rest}) &\propto p(\boldsymbol{\eta}^{z}_{i,j}) \times \prod_{n=1}^{N_i} p(\gamma_{i,n,j} \mid \boldsymbol{\eta}^{z}_{i,j}) \\
&\propto p(\boldsymbol{\eta}^{z}_{i,j}) \times \prod_{n=1}^{N_i} \prod_{\ell=1}^{L_z} \left(\eta^{z}_{i,j,\ell}\right)^{\mathbb{1}\left\{\gamma_{i,n,j} = \ell \,\cap\, z_{j,\lambda_{i,n}} = z\right\}} \\
&\propto \prod_{\ell=1}^{L_z} \left(\eta^{z}_{i,j,\ell}\right)^{a_{\eta^z} - 1} \times \prod_{n=1}^{N_i} \prod_{\ell=1}^{L_z} \left(\eta^{z}_{i,j,\ell}\right)^{\mathbb{1}\left\{\gamma_{i,n,j} = \ell \,\cap\, z_{j,\lambda_{i,n}} = z\right\}} \\
&\propto \prod_{\ell=1}^{L_z} \left(\eta^{z}_{i,j,\ell}\right)^{\left(a_{\eta^z} + \sum_{n=1}^{N_i} \mathbb{1}\left\{\gamma_{i,n,j} = \ell \,\cap\, z_{j,\lambda_{i,n}} = z\right\}\right) - 1}
\end{aligned}
$$

Therefore,

$$\boldsymbol{\eta}^z_{i,j} \mid \boldsymbol{y}, \text{rest} \;\sim\; \text{Dirichlet}_{L_z}\left(a^\star_1, ..., a^\star_{L_z}\right)$$

where $a^\star_\ell = a_{\eta^z} + \sum_{n=1}^{N_i} \mathbb{1}\left\{\gamma_{i,n,j} = \ell \;\cap\; z_{j,\lambda_{i,n}} = z\right\}$.

- Full Conditional for missing $y_{i,n,j}$ (conditioned on $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$)

$$p_\tau(y_{i,n,j} \mid o_{i,n,j} = 0, \text{rest}) \propto p(o_{i,n,j} = 0 \mid y_{i,n,j}, \text{rest}) \times \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{\frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,\lambda_{i,n}},\gamma_{i,n,j}}\right)^2}{2\sigma_i^2}\right\}\right)^{1/\tau}$$

$$\propto \rho_{i,n,j} \times \exp\left\{\frac{-\left(y_{i,n,j} - \mu^\star_{z_{j,\lambda_{i,n}},\gamma_{i,n,j}}\right)^2}{2\sigma_i^2\tau}\right\}$$

Since the full conditional distribution cannot be directly sampled from, it may be sampled from by a Metropolis step with a Normal proposal distribution. The proposed state is accepted with probability

$$\min\left\{1, \frac{p_\tau(\tilde{y}_{i,n,j} \mid \boldsymbol{y}, \text{rest})}{p_\tau(y_{i,n,j} \mid \boldsymbol{y}, \text{rest})}\right\}.$$

B.2.4 Intrinsic MCMC

Posterior simulation is computationally expensive when N_i is large. Particularly, updating $\boldsymbol{Z}$ takes substantially longer than updating other parameters. To speed up our posterior simulation, we exploit the idea of "intrinsic Bayes factor" (Berger and Pericchi 1996), and propose a sampling scheme that uses a minimal subsample of data to generate a proposal of $\boldsymbol{Z}$ and replace the prior of $\boldsymbol{\theta}$. Specifically, we partition the data $\boldsymbol{y}$ (which includes the current imputed missing data)

into $\boldsymbol{y}'$ and $\boldsymbol{y}''$, where $\boldsymbol{y}'$ is a "small" training sample of $\boldsymbol{y}$, and $\boldsymbol{y}''$ is the complement of $\boldsymbol{y}'$, and build a "minimally" trained prior using $\boldsymbol{y}'$, $p^\star(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}' \mid \boldsymbol{\theta})$. We replace the prior with $p^\star(\boldsymbol{\theta})$, and execute the posterior simulation with the remainder of the data, i.e., sample $\boldsymbol{\theta}$ from $p^\star(\boldsymbol{\theta} \mid \boldsymbol{y}'') \propto p^\star(\boldsymbol{\theta})p(\boldsymbol{y}'' \mid \boldsymbol{\theta})$. Also, we use $p^\star(\boldsymbol{\theta})$ to generate a proposal of $\boldsymbol{Z}$ in updating $\boldsymbol{Z}$ as follows;

1. Simulate a proposal of $\boldsymbol{Z}$ from $p^\star(\boldsymbol{Z})$, say $\boldsymbol{Z}'$. Specifically, we use the full conditional of $z_{j,k}$ given above and sample $\boldsymbol{Z}$ from $p^\star(\boldsymbol{Z})$ via MCMC using the current $\boldsymbol{Z}$ as an initial value, while fixing the other parameters at their current value. We repeat this for a small number of times (M) and generate a proposal of $\boldsymbol{Z}$ approximately independent of the current $\boldsymbol{Z}$.

2. Compute the metropolis acceptance ratio for the proposal, $\boldsymbol{Z}'$ as

$$\begin{aligned}\zeta = \frac{\pi(\boldsymbol{Z}' \mid \boldsymbol{y}, \boldsymbol{\theta}_{-Z})}{\pi(\boldsymbol{Z} \mid \boldsymbol{y}, \boldsymbol{\theta}_{-Z})} \cdot \frac{q(\boldsymbol{Z})}{q(\boldsymbol{Z}')} &= \frac{p(\boldsymbol{y}'' \mid \boldsymbol{Z}', \boldsymbol{\theta}_{-Z}) \cdot p^\star(\boldsymbol{Z}')}{p(\boldsymbol{y}'' \mid \boldsymbol{Z}, \boldsymbol{\theta}_{-Z}) \cdot p^\star(\boldsymbol{Z})} \cdot \frac{p^\star(\boldsymbol{Z})}{p^\star(\boldsymbol{Z}')} \\ &= \frac{p(\boldsymbol{y}'' \mid \boldsymbol{Z}', \boldsymbol{\theta}_{-Z})}{p(\boldsymbol{y}'' \mid \boldsymbol{Z}, \boldsymbol{\theta}_{-Z})},\end{aligned}$$

   where $\boldsymbol{\theta}_{-Z}$ denotes all random parameters except $\boldsymbol{Z}$. We accept $\boldsymbol{Z}'$ with probability $\min\{1, \zeta\}$.

3. Update all other parameters in $\boldsymbol{\theta}$ by sequentially sampling from their full conditionals.

Since $p^\star(\boldsymbol{\theta} \mid \boldsymbol{y}'') \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})$, updating all other parameters in $\boldsymbol{\theta}$ remains the same. Samples of $\boldsymbol{Z}$ under our current method can be highly correlated when M is too small. As M increases, the correlation will decrease, but the computation time will increase. Similarly, a small size for $\boldsymbol{y}'$ will increase the speed of sampling, but at the expense of obtaining highly variable proposals for $\boldsymbol{Z}'$, which may lower the

acceptance rate for $\boldsymbol{Z}$. In our simulation studies and data analysis, we consider hardware and time constraints, and select M and the size of $\boldsymbol{y}'$ primarily through preliminary simulation studies.

B.3 Additional Results for Simulation Studies

We conducted sensitivity analyses of the model by changing values of ϕ_2. We fitted the rep-FAM with different values of ϕ_2, $\phi_2 \in \{1, 10, 25, 100\}$ while fixing $\phi_1 = 1$. The results with $\phi_2 = 10$ and 100 are presented in the main text. Figures B.2 and B.3 show the transpose of the posterior estimates of $\boldsymbol{Z}$ and $\boldsymbol{w}$ with $\phi_2 = 1$ and 25, respectively. For a larger value of ϕ_2, $\hat{\boldsymbol{Z}}_i$ includes features that are more distinctive. Figures B.4-B.6 show heatmaps of the simulated expression levels of a sample for each of the three scenarios. For the three plots, we vary the value of ϕ_2 by letting $\phi_2 \in \{1, 25, 100\}$. The cells (rows) are rearranged according to their posterior cluster membership estimates, $\hat{\lambda}_{i,n}$. High expression levels (in red) are likely to correspond to marker expressions in the corresponding features in $\hat{\boldsymbol{Z}}_i$; and low expression levels (in blue) to non-expression of markers in corresponding features in $\hat{\boldsymbol{Z}}_i$. Missing values, artificially included to emulate real data, are indicated by black cells. Figure B.7 presents the posterior distribution for the number of selected features $|R_i|$ for each sample ($i = 1, 2$), scenario (1,2,3), and $\phi_2 \in \{1, 25, 100\}$. The red line indicates the true value of $|R_i|$. As ϕ_2 increases, R_i tends to decrease as well.

Figures B.8-B.10 show heatmaps of $\boldsymbol{y}$, where cells in rows are arranged by cluster membership labels estimated from FlowSOM and MClust. The figures are for each of the three simulation scenarios. Expression levels are more heterogeneous within some clusters. For example, as shown in Figures B.8 (a) and (c),

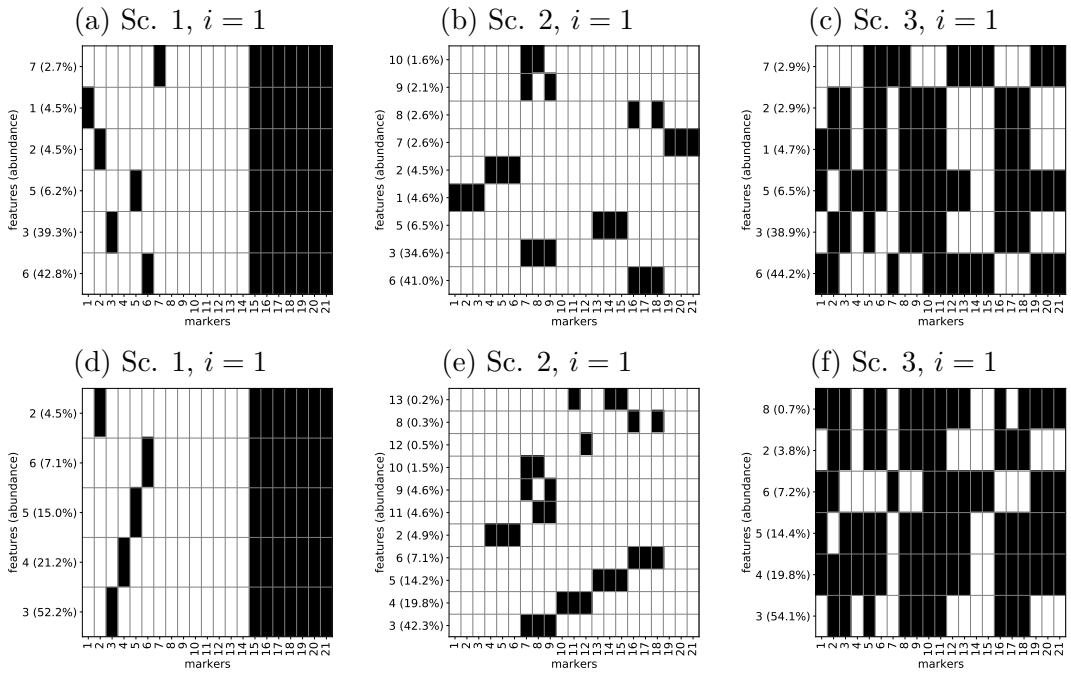


Figure B.2: [Simulation Study] Posterior estimate for transpose of $\boldsymbol{Z}$ and $\boldsymbol{w}$ in simulation studies for each sample ($i = 1, 2$) and scenarios (1,2,3). $\phi_2 = 1$ is used. The binary matrices are the estimates of $\boldsymbol{Z}$ and the numbers on the left axes are the feature number, and their abundance in parentheses.

FlowSOM and MClust collapsed some true clusters into a cluster and the most abundant (bottom) cluster contains a mix of high and low expression for marker 6

Figures B.11- B.13 show 2-dimensional t-SNE of the simulated data from each scenario. In each scenario, the embeddings were computed jointly for both samples. Points are color-coded by their true cluster labels. When a feature is very different from the others, the cells in the corresponding cluster have embeddings well separated from the other embeddings, e.g., Scenario 2 shown in Figure B.12. However, when a feature is similar to some other features, the embeddings of the cells in the corresponding cluster are greatly overlapped with those of the cells in other clusters, as in Scenarios 1 and 3.

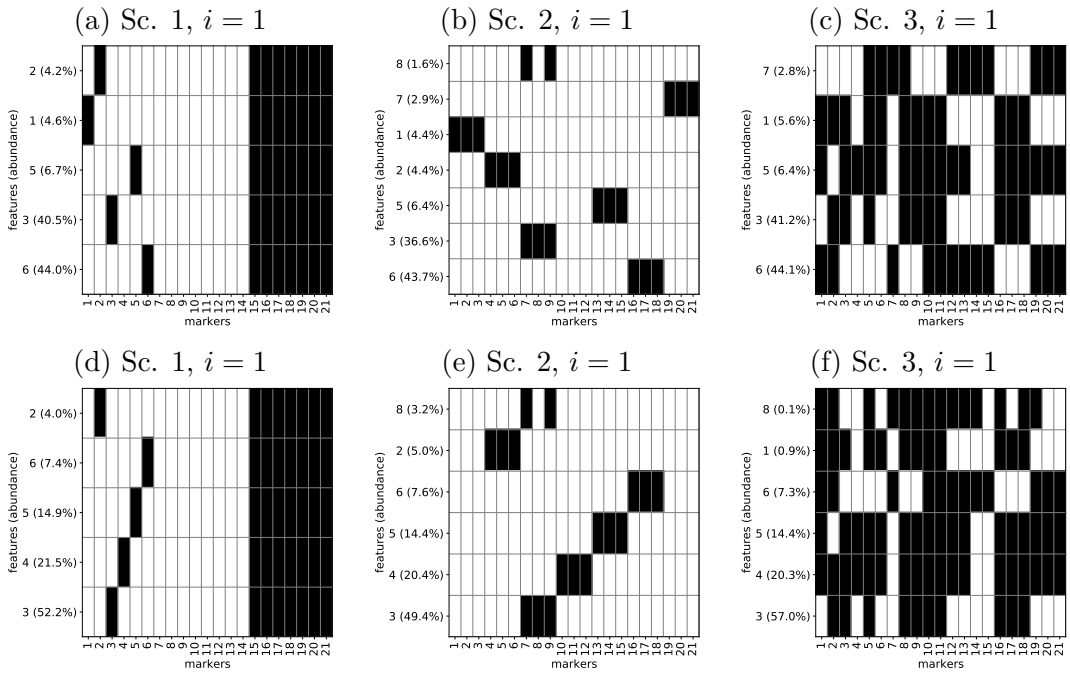


Figure B.3: [Simulation Study] Posterior estimate for transpose of $\boldsymbol{Z}$ and $\boldsymbol{w}$ in simulation studies for each sample ($i = 1, 2$) and scenarios (1,2,3). $\phi_2 = 25$ is used. The binary matrices are the estimates of $\boldsymbol{Z}$ and the numbers on the left axes are the feature number, and their abundance in parentheses.

B.4 Additional Results for Data Analysis

Here we provide additional results from the sensitivity analyses to the specification of $\boldsymbol{\phi}$ and p_i. As mentioned in the main text, we calibrated the model such that, *a priori*, 5 subpopulations on average are selected for each sample. For $K = 25$, $p_i = 0.2$ corresponds to a prior mean of 5 subpopulations per sample. We additionally fit the model with different values of p_i, $p_i = 0.1$ or 0.3 to assess model sensitivity to the specification of p_i. Figures B.14-B.16 show that the specification of $p_i \in \{0.1, 0.3\}$ only weakly influence the inference. From Figures B.14, we observe that the most abundant subpopulation in the estimated $\hat{\boldsymbol{Z}}_i$ for sample 1 is the same, with similar abundances for the different values of p_i. In addition, the four most abundant features for the different p_i are similar for a given sample. For

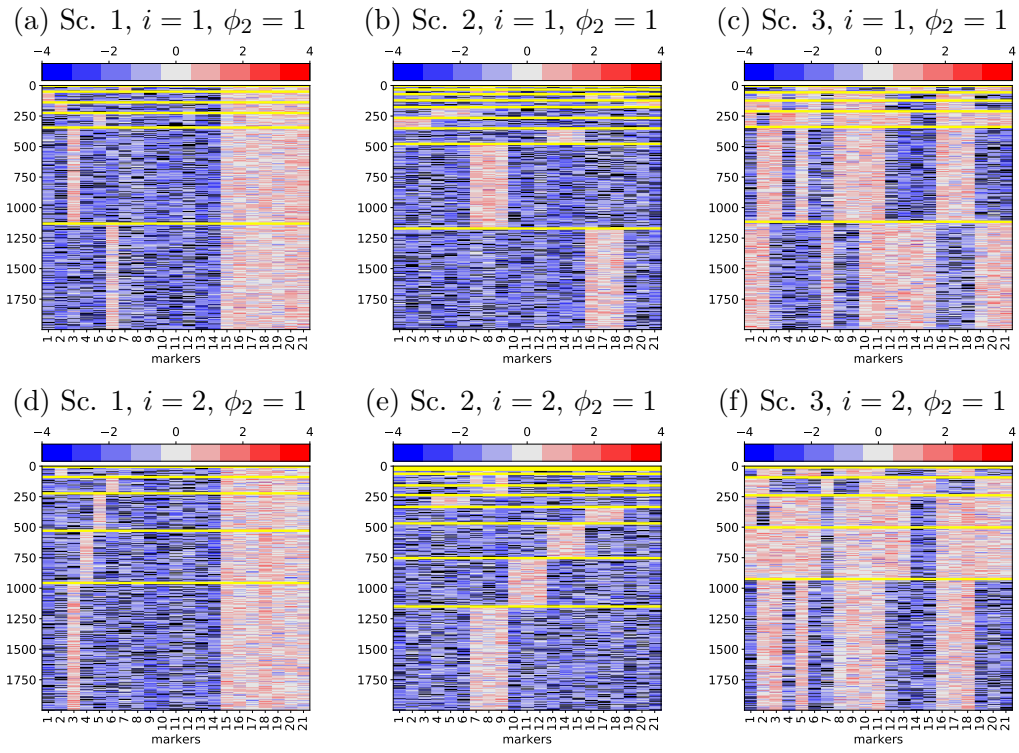


Figure B.4: Marker expression levels $\boldsymbol{y}_i$ sorted by row according to posterior estimate of feature membership labels $\lambda_{i,n}$, for each sample ($i = 1, 2$), scenario (1,2,3). $\phi_2 = 1$ is used.

reference, heatmaps of the data with cells (rows) sorted according to the estimated subpopulation ($\hat{\lambda}_{i,n}$) are included in Figure B.15. Figure B.16 shows the posterior distribution of $|R_i|$ for $\phi_2 = 25$, each sample, and $p_i = 0.1$ and 0.3. Note that with $p_i = 0.1$, the number of selected subpopulations is smaller as subpopulations with smaller abundances are not selected. Figure B.17 shows the distribution of the pairwise-column distances between subpopulation estimates $\hat{\boldsymbol{Z}}_i$.

The calibration of ϕ_2 was discussed in the main document. Here, we provide supporting figures to assess model sensitivity to the specification of $\phi_2 \in \{1, 10, 100\}$. Figure B.18 show the estimates $\hat{\boldsymbol{Z}}_i$ for the various ϕ_2. Note that larger subpopulations have a tendency to be recovered for the various ϕ_2. For example, subpopulations 1 for $\phi_2 = 1, 100$ and subpopulation 3 for $\phi_2 = 2$ are

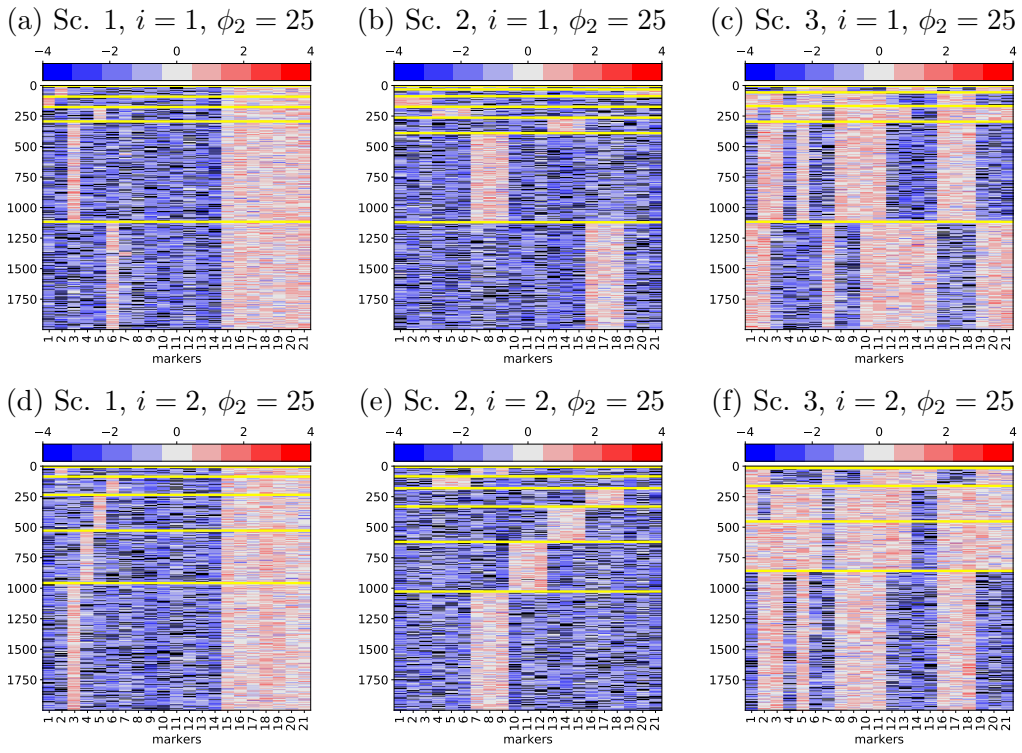


Figure B.5: Marker expression levels $\boldsymbol{y}_i$ sorted by row according to posterior estimate of feature membership labels $\lambda_{i,n}$, for each sample ($i = 1, 2$), scenario (1,2,3). $\phi_2 = 25$ is used.

identical. Figure B.19 also shows that the number of selected subpopulations decreases as ϕ_2 increases. Figure B.20 shows the distribution of pairwise-column distances between subpopulation estimates $\hat{\boldsymbol{Z}}_i$ for each sample and the various ϕ_2. Note the tendency for subpopulations to be more varied as ϕ_2 increases. For example, when $\phi_2 = 100$, we see that all subpopulations are different (from other subpopulations) by at least 3 markers; whereas at $\phi_2 = 1$, about 5% and 4% of the subpopulations are different by 1 marker in samples 1 and 2, respectively. Thus, ϕ_2 should not be made arbitrarily large, but should be reasonably calibrated as suggested in the main text. For reference, Figure B.21 presents heatmaps of the data with cells (rows) sorted according to the estimated subpopulation ($\hat{\lambda}_{i,n}$).

Figure B.22 shows plots of 2-dimensional t-SNE embeddings for our CyTOF

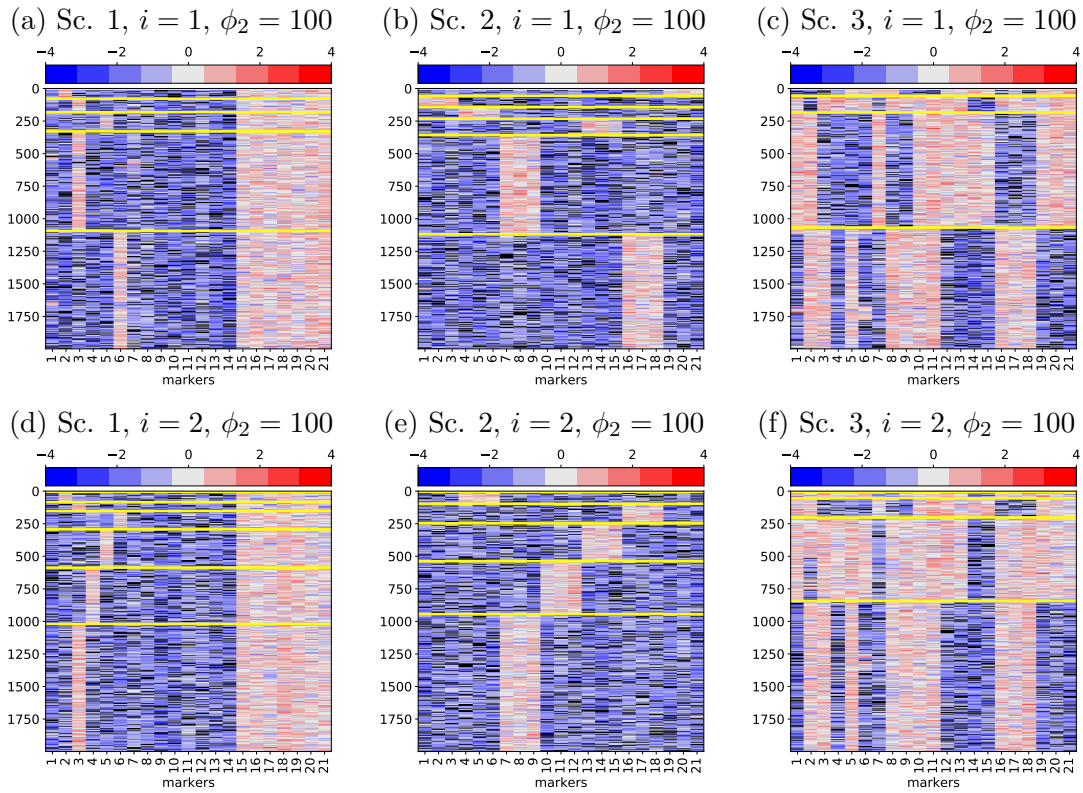


Figure B.6: Marker expression levels $\boldsymbol{y}_i$ sorted by row according to posterior estimate of feature membership labels $\lambda_{i,n}$, for each sample ($i = 1, 2$), scenario (1,2,3). $\phi_2 = 100$ is used.

data. The embeddings are colored by subgroup label estimates from the rep-FAM and from the ind-FAM. For both, $p_i = 0.2$ is assumed, and for the rep-FAM, $\phi_2 = 25$ is set.

Figure B.23 shows heatmaps of the patients dataset with cells (rows) arranged by cluster memberships estimates by FlowSOM and MClust. The expression levels in clusters estimated by those methods are more different than those in the clusters estimated by our rep-FAM. The clusters by FlowSOM and MClust includes both high and low expression levels, e.g., see the second largest subpopulation in sample 2 in panel (b) and the largest subpopulation in sample 2 in panel (d).

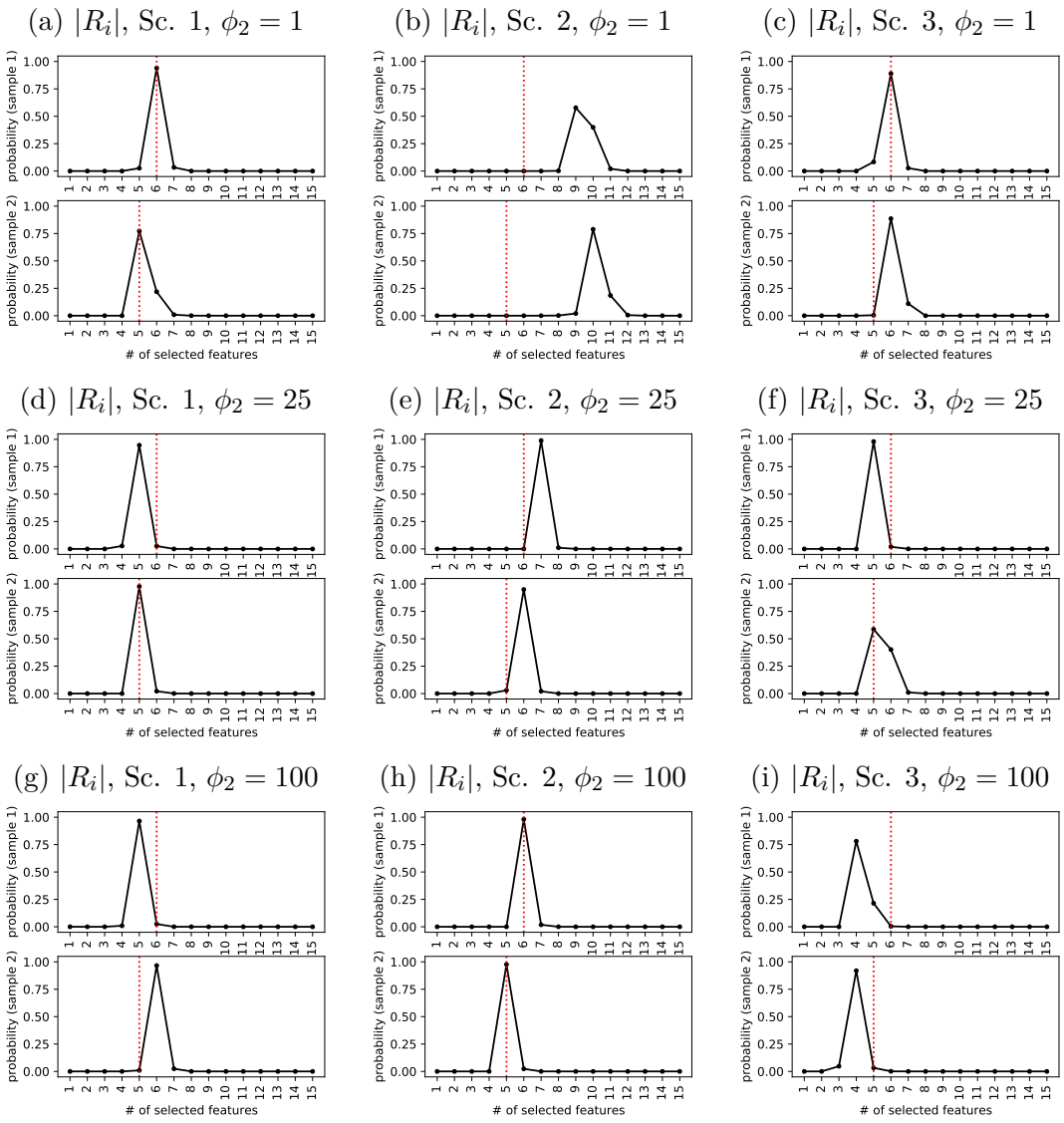


Figure B.7: Posterior distribution of number of selected features $|R_i|$ for each sample (i =1,2), scenario (1,2,3), and $\phi_2 \in (1, 25, 100)$.

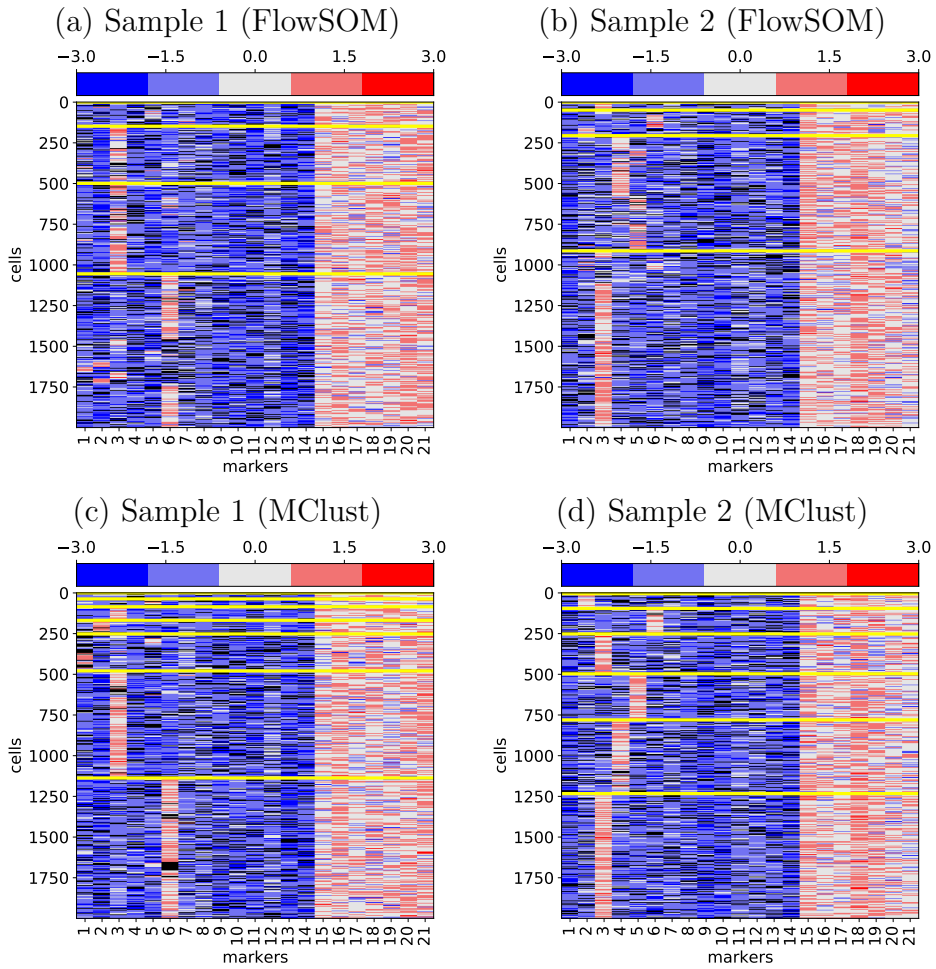


Figure B.8: [Simulation Scenario 1] Heatmap of the data $\boldsymbol{y}_i$ in simulation scenario 1 for each sample ($i = 1, 2$). Cells in rows are arranged by their cluster membership estimates. Clustering method, FlowSOM are used for (a) and (b), and MClust for (c) and (d).

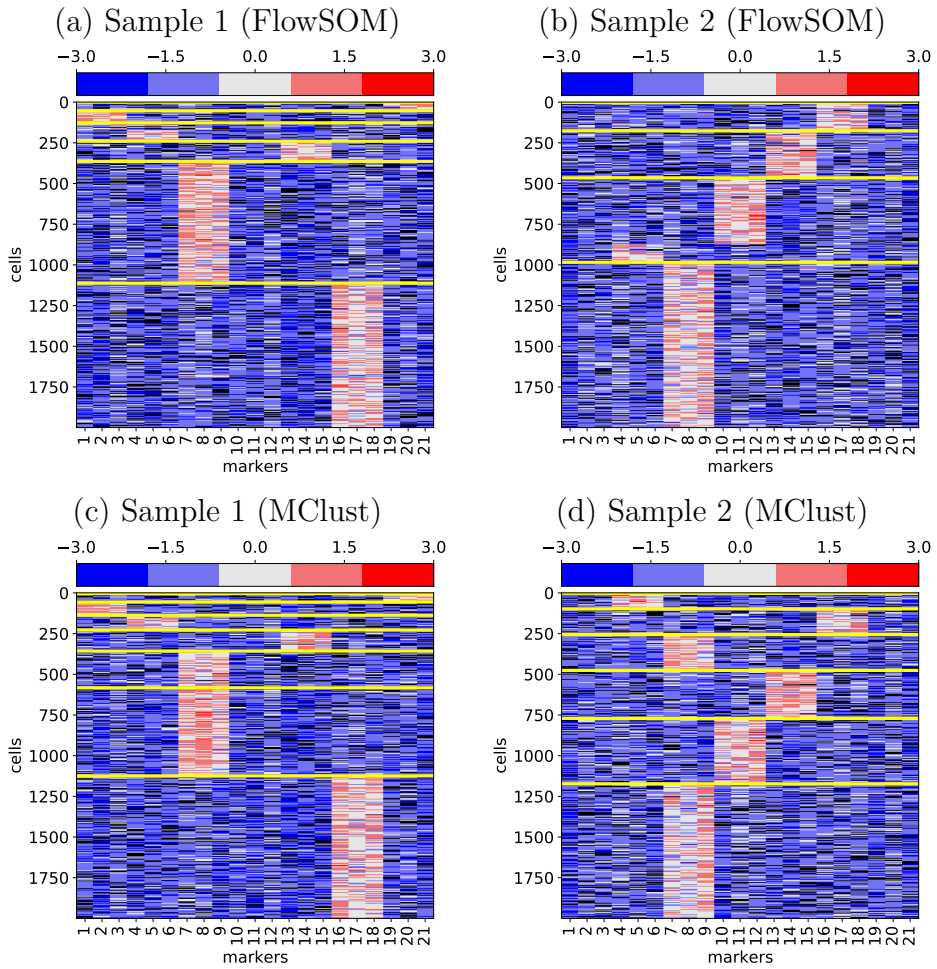


Figure B.9: [Simulation Scenario 2] Heatmap of the data $\boldsymbol{y}_i$ in simulation scenario 2 for each sample ($i = 1, 2$). Cells in rows are arranged by their cluster membership estimates. Clustering method, FlowSOM are used for (a) and (b), and MClust for (c) and (d).

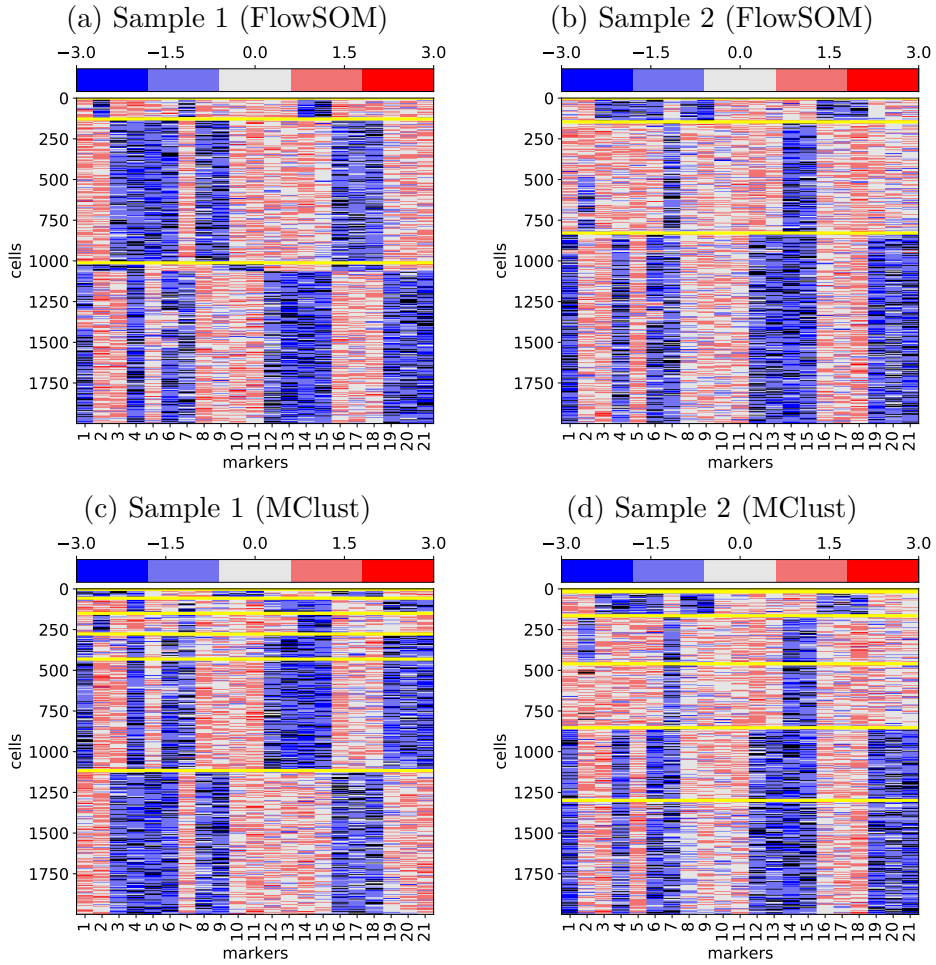


Figure B.10: [Simulation Scenario 3] Heatmap of the data $\boldsymbol{y}_i$ in simulation scenario 3 for each sample ($i = 1, 2$). Cells in rows are arranged by their cluster membership estimates. Clustering method, FlowSOM are used for (a) and (b), and MClust for (c) and (d).

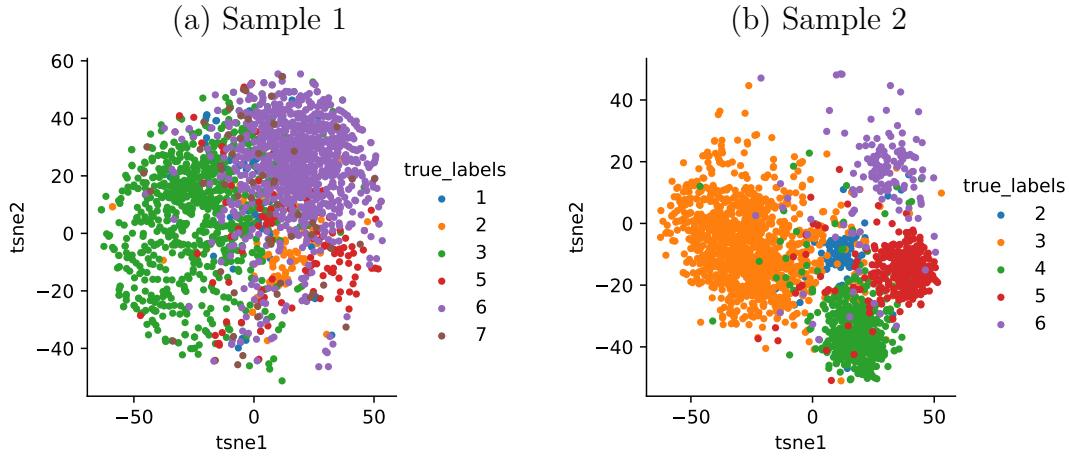


Figure B.11: t-SNE for Scenario 1. The embeddings of the cells are colored by their true cluster labels.

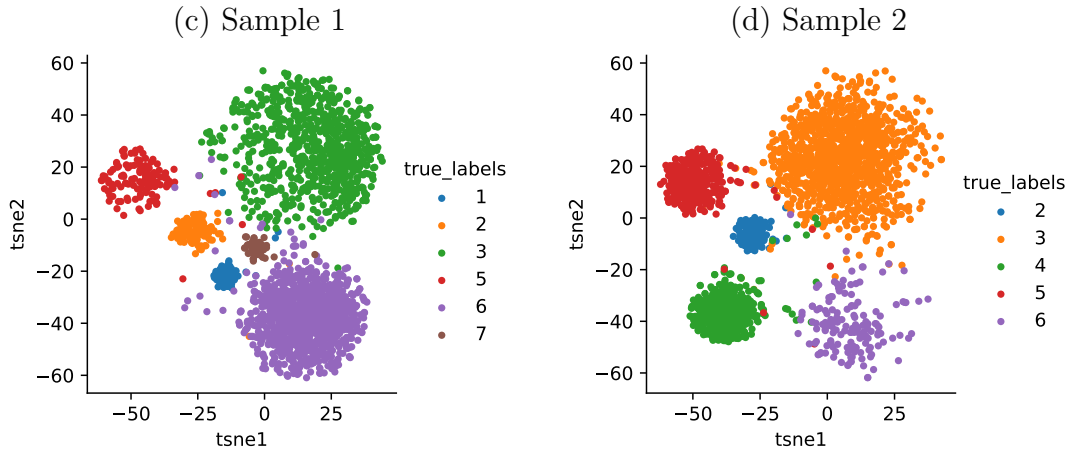


Figure B.12: Plots of t-SNE for Scenario 2. The embeddings of the cells are colored by their true cluster labels.

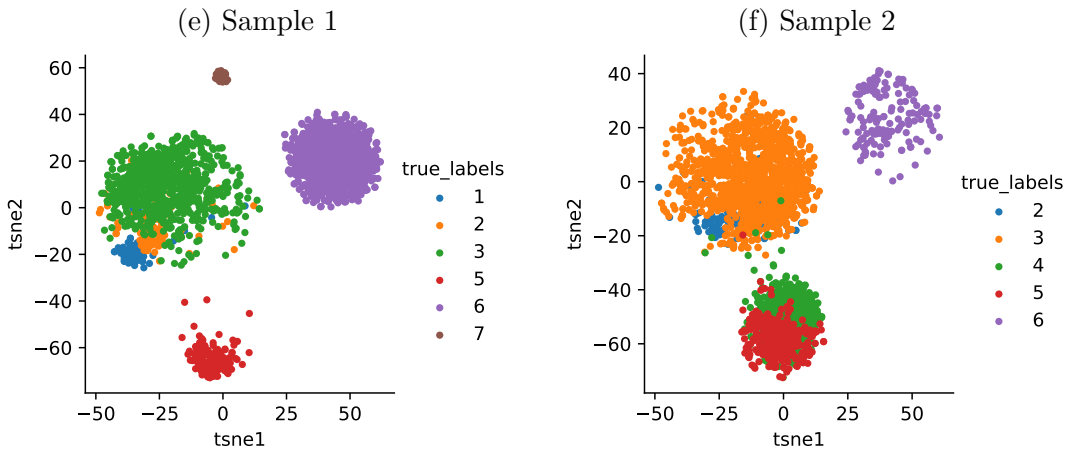


Figure B.13: Plots of t-SNE for Scenario 3. In this dataset, distinct and similar features are present. TheThe embeddings of the cells are colored by their true cluster labels.

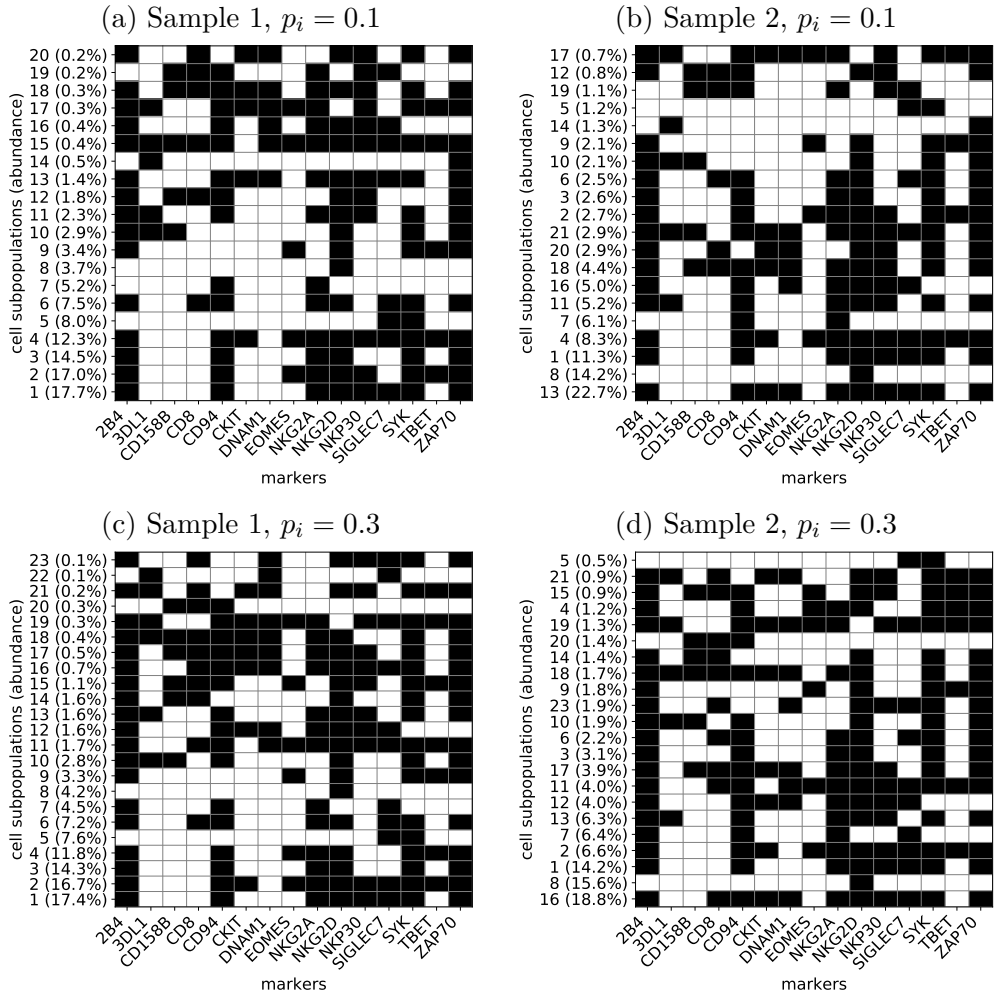


Figure B.14: Point estimates of NK cell subpopulations Z in cytometry samples taken from 2 subjects, for each sample ($i = 1, 2$), with p_i fixed at 0.1 and 0.3, and $\phi_2 = 25$

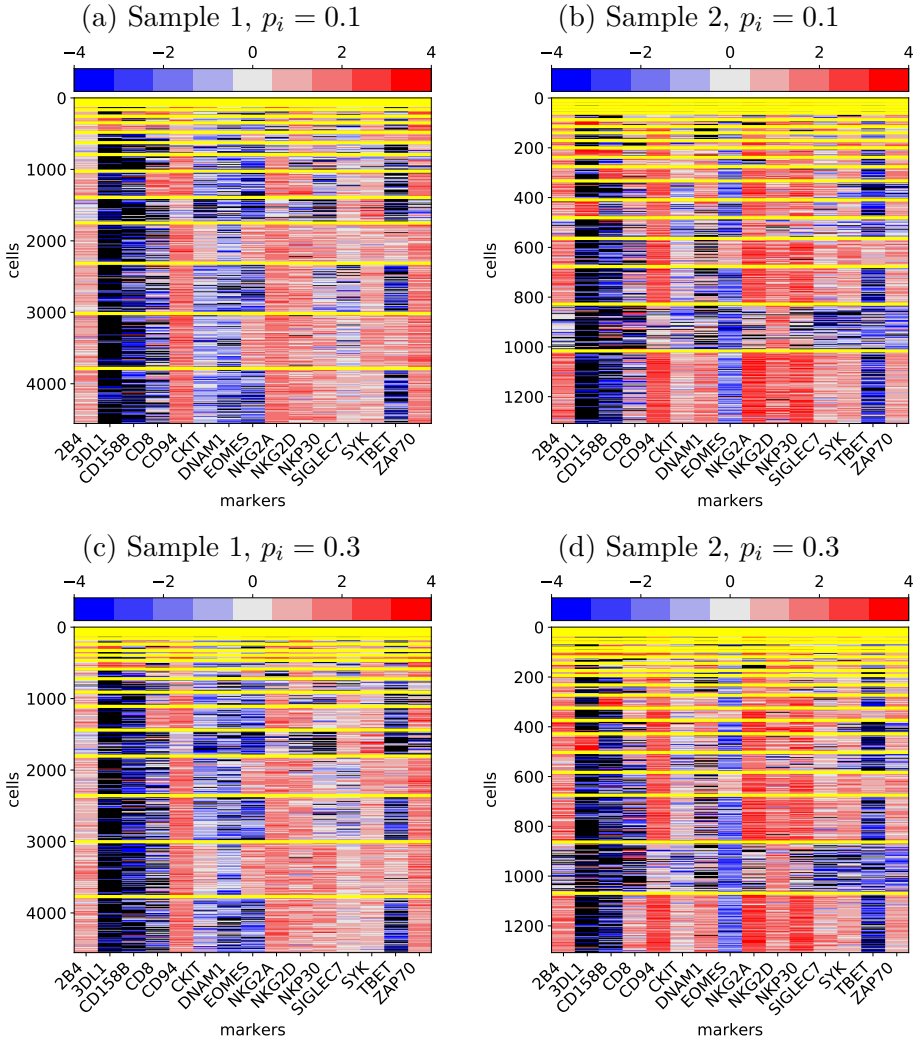


Figure B.15: Marker expression levels y_i for each cell subpopulation, sorted by row according to posterior estimate of subpopulation membership labels $\lambda_{i,n}$, with the most abundant subpopulations at the bottom, for each sample ($i = 1, 2$), with $p_i = 0.1, 0.3$ and $\phi_2 = 0, 25$.

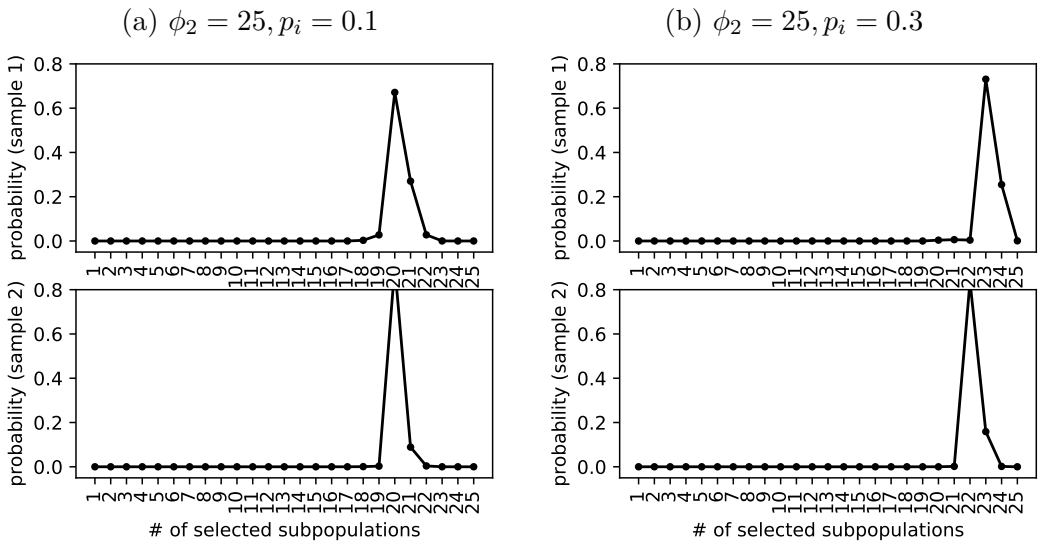


Figure B.16: Posterior distribution of the number of selected subpopulations within each sample, for p_i fixed at 0.1 and 0.3, and $\phi_2 = 25$.

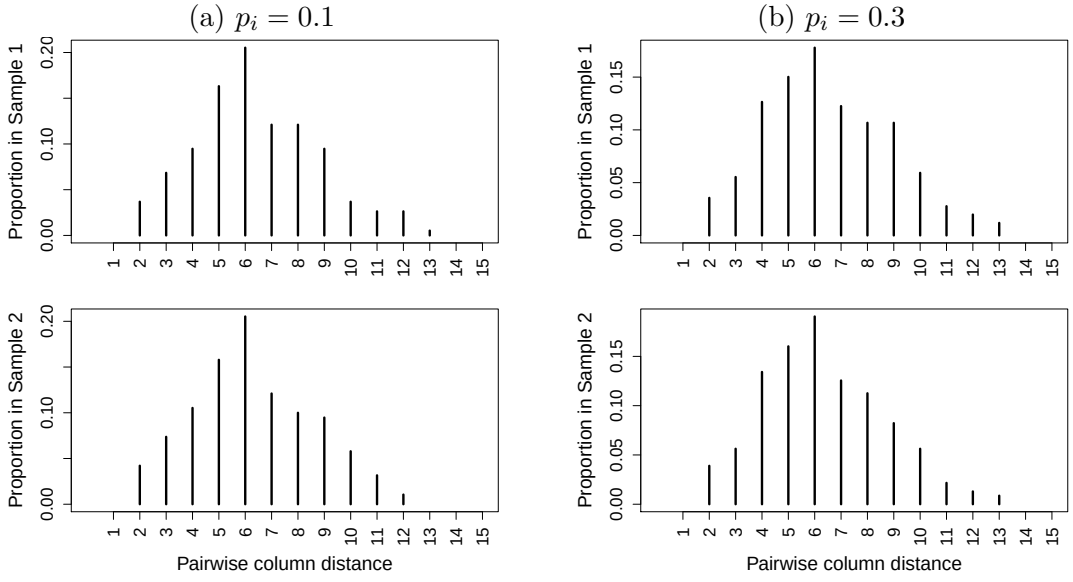


Figure B.17: Distribution of the pairwise-column distances between subpopulation estimates $\hat{\boldsymbol{Z}}_i$ for each sample, for p_i fixed at 0.1 and 0.3, and $\phi_2 = 25$.

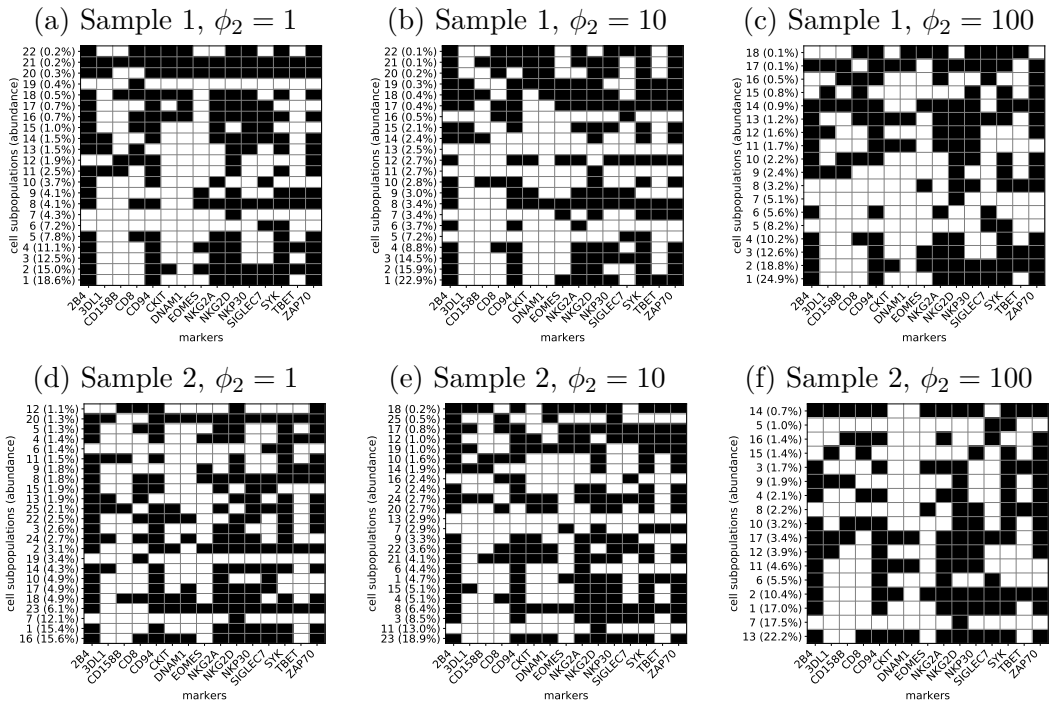


Figure B.18: Point estimates of NK cell subpopulations Z in cytometry samples taken from 2 subjects, for each sample ($i = 1, 2$), with $p_i = 0.2$ and $\phi_2 = 1, 10, 100$.

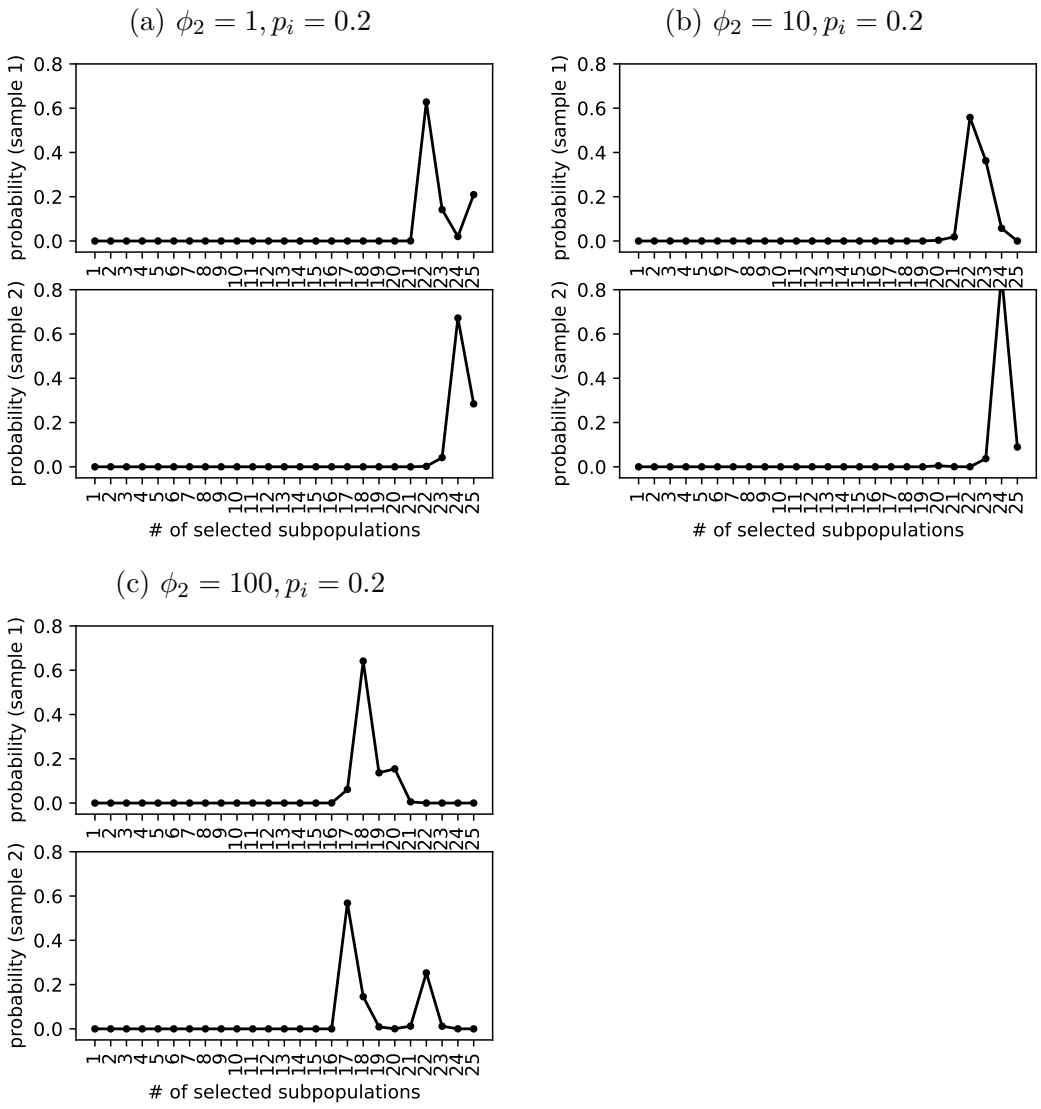


Figure B.19: Posterior distribution of the number of selected subpopulations within each sample, for $\phi = 1, 10, 100$, and $p_i = 0.2$.

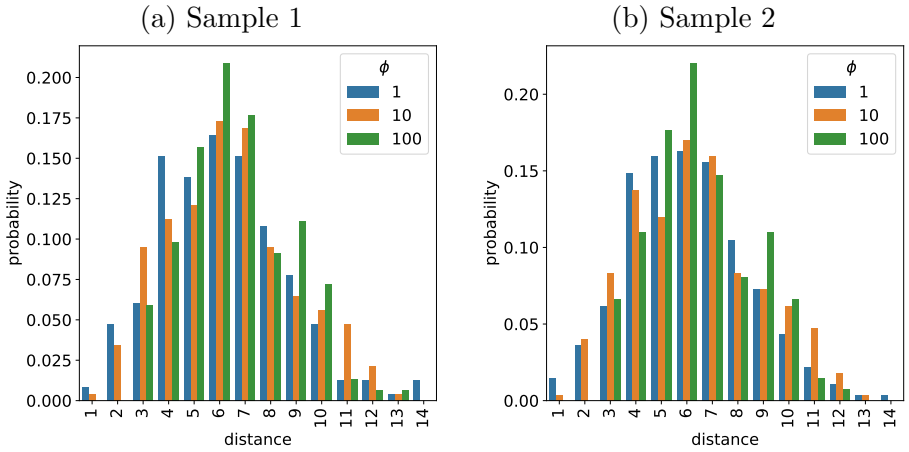


Figure B.20: Distribution of the pairwise-column distances between subpopulation estimates $\hat{\boldsymbol{Z}}_i$ for each sample, for $\phi = 1, 10, 100$, and $p_i = 0.2$.

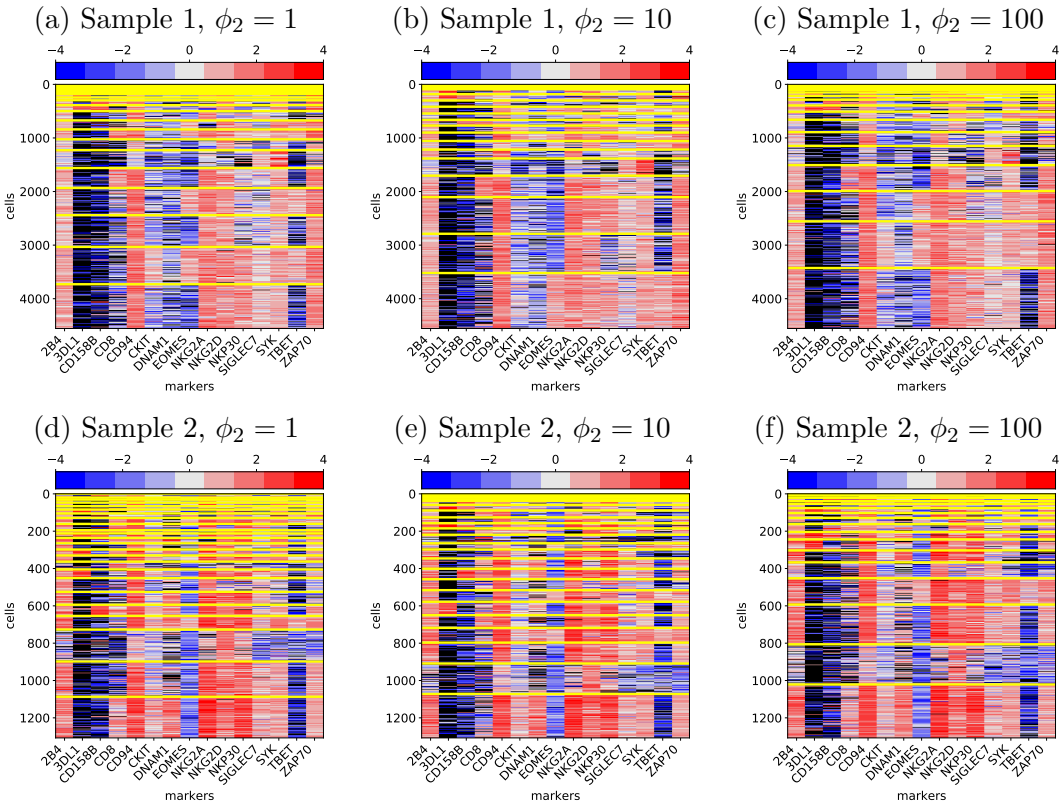


Figure B.21: Marker expression levels y_i for each cell subpopulation, sorted by row according to posterior estimate of subpopulation membership labels $\lambda_{i,n}$, with the most abundant subpopulations at the bottom, for each sample ($i = 1, 2$), with $p_i = 0.2$ and $\phi_2 = 1, 10, 100$.

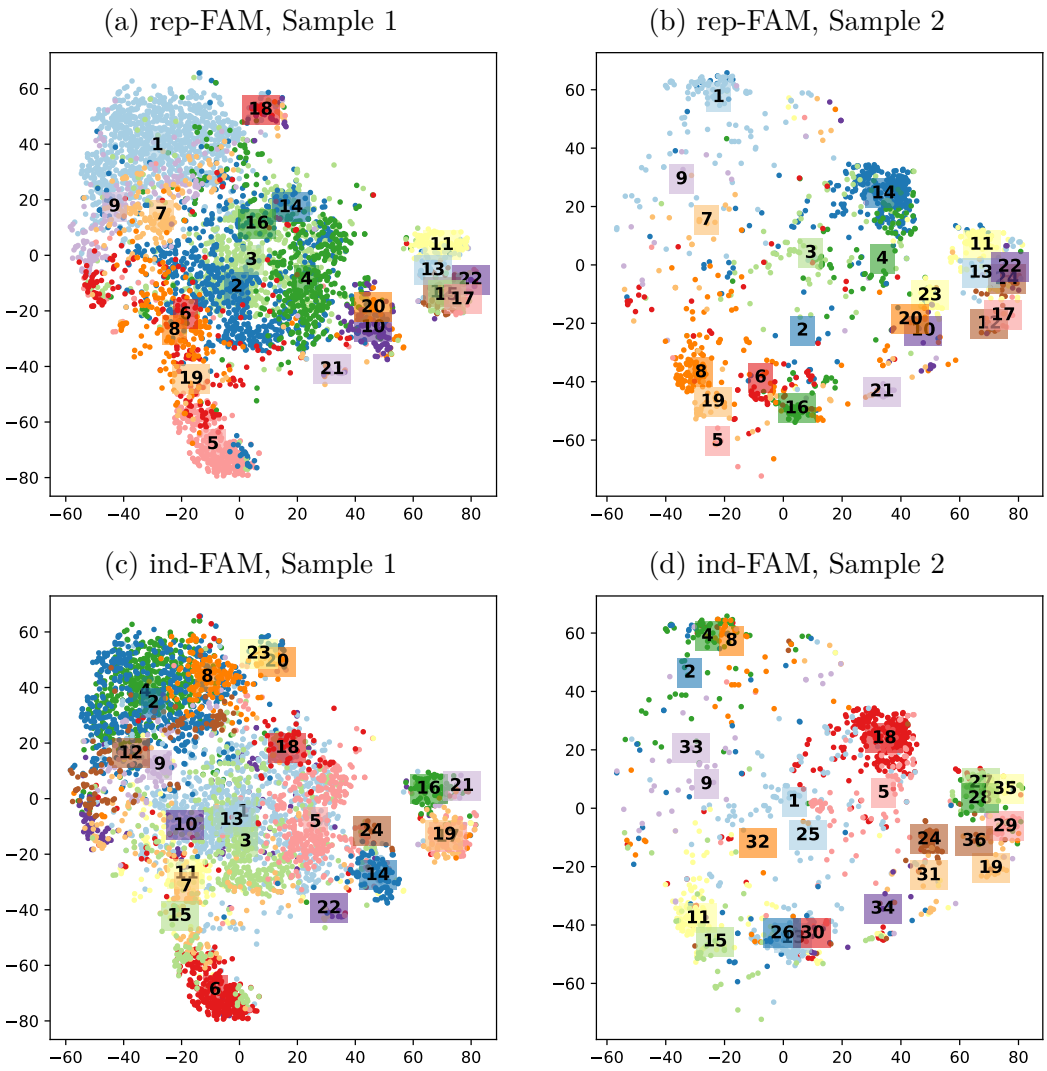


Figure B.22: t-SNE of patients data set computed jointly for both samples. t-SNE are color-coded according to the induced clusterings in rep-FAM ($\phi_2 = 25$) and ind-FAM.

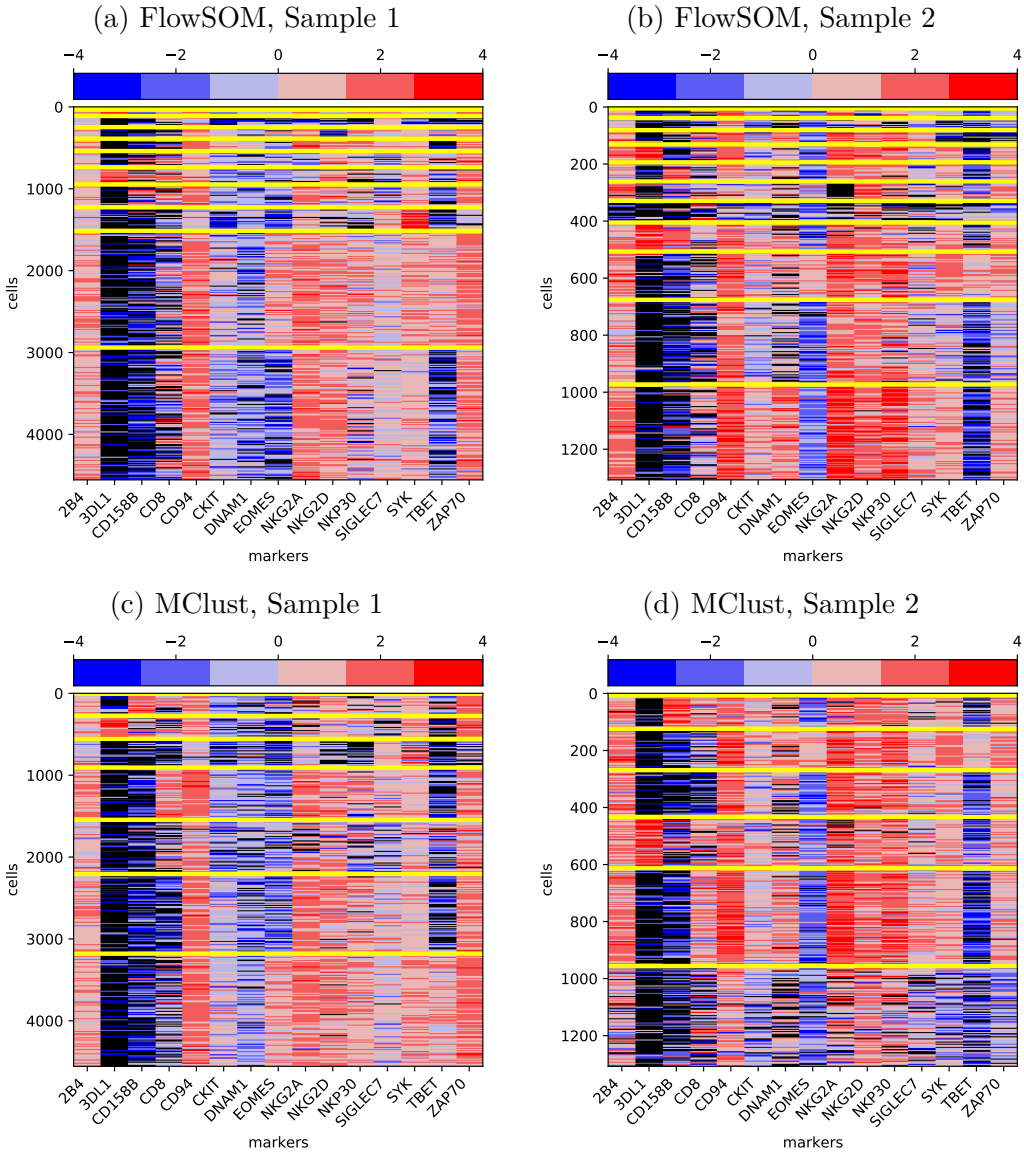


Figure B.23: Heatmap of patients dataset with cells arranged by cluster membership for FlowSOM and MClust.

Appendix C

A Bayesian Differential Distribution Approach for Zero-inflated Data with Applications to Cytometry Data

C.1 Full Conditionals for Model Parameters

To better facilitate posterior sampling from a mixture model, we introduce an auxiliary parameter $\lambda_{i,n} \in \{1, \ldots, K\}$ such that $\lambda_{i,n} \mid \boldsymbol{\eta}_i \sim \text{Categorical}(\boldsymbol{\eta}_i)$ represents the mixture component taken by cell n in sample i, if $y_{i,n} > 0$. With

$\lambda_{i,n}$, the joint posterior for this model is

$$
\begin{aligned}
&p(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\nu}, \boldsymbol{\omega}, \tau, \boldsymbol{\zeta}, \boldsymbol{v}, \boldsymbol{\lambda} \mid \boldsymbol{y}) \propto \\
&\quad \left\{ \prod_{i=1}^{I} \prod_{n=1}^{N_i} p_y(\tilde{y}_{i,n} \mid \mu(\boldsymbol{\iota})_{\lambda_{i,n}} + \psi_{\lambda_{i,n}} \zeta_{i,n}, \text{var} = \omega_{\lambda_{i,n}} / v_{i,n})^{1-z_{i,n}} \right\} \times \\
&\quad \left\{ \prod_{i=1}^{I} \gamma_i^{Q_i} (1-\gamma_i)^{N_i - Q_i} \times p_\gamma(\gamma_i) \times p_\eta(\boldsymbol{\eta}_i) \right\} \times \\
&\quad \left\{ \prod_{i=1}^{I} \prod_{n=1}^{N_i} p_\zeta(\zeta_{i,n} \mid v_{i,n}) p_v(v_{i,n} \mid \nu_{\lambda_{i,n}}) p_\lambda(\lambda_{i,n} \mid \boldsymbol{\eta}_i) \right\} \times \\
&\quad \left\{ \prod_{k=1}^{K} p_\psi(\psi_k) p_\omega(\omega_k \mid \tau) p_\nu(\nu_k) p_\iota(\iota_k) \right\} \times p_\tau(\tau)
\end{aligned}
$$

The model parameters can be updated sequentially via Gibbs sampling. Except for ν_k, the full conditional distributions for all model parameters are available in closed form. For convenience, let $\boldsymbol{N}_i^+$ denote $n \in \{n : y_{i,n} > 0\}$.

$$
\begin{aligned}
\gamma_i \mid \text{data}, \text{rest} &\sim \text{Beta}(a_\gamma + Q_i, b_\gamma + N_i - Q_i), \text{ for } i \in \{1, \ldots, I\} \\
\boldsymbol{\eta}_i \mid \text{data}, \text{rest} &\sim \text{Dirichlet}(\boldsymbol{a}_i^\star), \text{ for } i \in \{1, \ldots, I\}, \text{ where} \\
a_{i,k}^\star &= a_{\eta,k} + \sum_{n \in \boldsymbol{N}_i^+} \mathbb{1}\{\lambda_{i,n} = k\} \\
\Pr(\lambda_{i,n} = k \mid \text{data}, \text{rest}) &\propto \eta_{i,k} \cdot \text{normal}(\tilde{y}_{i,n} \mid \mu_k + \psi_k \cdot \zeta_{i,n}, \text{var} = \omega_k / v_{i,n}) \cdot \\
&\quad \text{gamma}(v_{i,n} \mid \nu_k/2, \nu_k/2) \\
\iota_1 \mid \text{data}, \text{rest} &\sim \text{Normal}(m_{\iota_1}^\star, v_{\iota_1}^\star) \\
\iota_k \mid \text{data}, \text{rest} &\sim \text{TruncatedNormal}_{(0,\infty)}(m_{\iota_k}^\star, v_{\iota_k}^\star), \text{ for } k = 2, \ldots, K, \text{ where} \\
v_{\iota_k}^\star &= \left(\frac{1}{s_\mu^2} + \frac{1}{\omega_k} \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} v_{i,n} \cdot \mathbb{1}\{\lambda_{i,n} \geq k\} \right)^{-1} \\
m_{\iota_k}^\star &= v_{\mu_k}^\star \cdot \left(\frac{m_\mu}{s_\mu^2} + \frac{1}{\omega_k} \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} g_{i,n} \cdot v_{i,n} \cdot \mathbb{1}\{\lambda_{i,n} \geq k\} \right), \text{ with}
\end{aligned}
$$

$$
\begin{aligned}
g_{i,n} &= \tilde{y}_{i,n} - \psi_k \cdot \zeta_{i,n} - \sum_{\ell=1}^{\lambda_{i,n}} \iota_\ell \cdot \mathbb{1}\{\ell \neq k\} \\
p(\nu_k \mid \text{data}, \text{rest}) &\propto \text{normal}(\log \nu_k \mid m_\nu, s_\nu^2) \times \\
&\quad \prod_{i=1}^{I} \prod_{n \in \boldsymbol{N}_i^+} \text{gamma}(v_{i,n} \mid \nu_k/2, \text{rate} = \nu_k/2)^{\mathbb{1}\{\lambda_{i,n}=k\}} \\
&\quad \text{(update } \log(\nu_k) \text{ with a Metropolis step.)} \\
\omega_k \mid \text{data}, \text{rest} &\sim \text{InverseGamma}(a^\star_{\omega_k}, b^\star_{\omega_k}), \text{where} \\
a^\star_{\omega_k} &= a_\omega + \frac{1}{2} \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} \mathbb{1}\{\lambda_{i,n} = k\} \\
b^\star_{\omega_k} &= \tau + \frac{1}{2} \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} \mathbb{1}\{\lambda_{i,n} = k\}\, v_{i,n} (\tilde{y}_{i,n} - \mu_k - \psi_k \cdot \zeta_{i,n})^2 \\
\tau \mid \text{data}, \text{rest} &\sim \text{Gamma}(a_\tau + K \cdot a_\omega,\ \text{rate} = b_\tau + \sum_{k=1}^{K} \omega_k^{-1}) \\
v_{i,n} \mid \text{data}, \lambda_{i,n} = k, \text{rest} &\sim \text{Gamma}\left(\frac{\nu_k}{2} + 1, \frac{\nu_k + \zeta_{i,n}^2 + \omega_k^{-1}(\tilde{y}_{i,n} - \mu_k - \psi_k \cdot \zeta_{i,n})^2}{2}\right), \\
&\quad \text{if } y_{i,n} > 0 \\
\zeta_{i,n} \mid \text{data}, \lambda_{i,n} = k, \text{rest} &\sim \text{TruncatedNormal}_{[0,\infty)}(m^\star, v^\star),\ \text{if } y_{i,n} > 0,\ \text{where} \\
v^\star_{\zeta_{i,n}} &= \left(v_{i,n} + \frac{v_{i,n}\psi_k^2}{\omega_k}\right)^{-1} \\
m^\star_{\zeta_{i,n}} &= v^\star_{\zeta_{i,n}} \cdot \left(\frac{v_{i,n}\psi_k(\tilde{y}_{i,n} - \mu_k)}{\omega_k}\right) \\
\psi_k \mid \text{data}, \text{rest} &\sim \text{Normal}(m^\star_{\psi_k}, v^\star_{\psi_k}),\ \text{where} \\
v^\star_{\psi_k} &= \left(\frac{1}{s_\psi^2} + \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} \mathbb{1}\{\lambda_{i,n} = k\} \cdot \frac{\zeta_{i,n}^2 \cdot v_{i,n}}{\omega_k}\right)^{-1} \\
m^\star_{\psi_k} &= v^\star_{\psi_k} \cdot \left(\frac{m_\psi}{s_\psi^2} + \sum_{i=1}^{I} \sum_{n \in \boldsymbol{N}_i^+} \mathbb{1}\{\lambda_{i,n} = k\} \cdot \frac{\zeta_{i,n}(y_{i,n} - \mu_k) \cdot v_{i,n}}{\omega_k}\right)
\end{aligned}
$$