Bayesian Analysis (2024) **19**, Number 2, pp. 439–463

Coarsened Mixtures of Hierarchical Skew Normal Kernels for Flow and Mass Cytometry Analyses[*]

Shai Gorsky[†], Cliburn Chan[‡] and Li Ma[§]

Abstract. Cytometry is the standard multi-parameter assay for measuring single cell phenotype and functionality. It is commonly used for quantifying the relative frequencies of cell subsets in blood and disaggregated tissues. A typical analysis of cytometry data involves cell classification—that is, the identification of cell subgroups in the sample—and comparisons of the cell subgroups across samples or conditions. While modern experiments often necessitate the collection and processing of samples in multiple batches, analysis of cytometry data across batches is challenging because differences across samples may occur due to either true biological variation or technical reasons such as antibody lot effects or instrument optics across batches. Thus a critical step in comparative analyses of multi-sample cytometry data—yet missing in existing automated methods for analyzing such data—is cross-sample calibration, whose goal is to align corresponding cell subsets across multiple samples in the presence of technical variations, so that biological variations can be meaningfully compared. We introduce a Bayesian nonparametric hierarchical modeling approach for accomplishing both calibration and cell classification simultaneously in a unified probabilistic manner. Three important features of our method make it particularly effective for analyzing multi-sample cytometry data: a nonparametric mixture avoids prespecifying the number of cell clusters; a hierarchical skew normal kernel that allows flexibility in the shapes of the cell subsets and cross-sample variation in their locations; and finally the "coarsening" strategy makes inference robust to departures from the model not captured by the skew normal kernels. We demonstrate the merits of our approach in simulated examples and carry out a case study in the analysis of a multi-sample cytometry data set. We provide an R package for our method.

Keywords: nonparametric estimation, mixture models, cytometry.

1 Introduction

Cytometry is a standard biological assay for measuring single cell features, referred to as "markers", and is commonly used for quantifying the relative frequencies of cell subsets

[*] Li Ma's research is partly supported by National Science Foundation grants DMS-1749789 and DMS-2013930. This work has also been supported in part through an EQAPOL collaboration with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Contract Number HHSN272201700061C and the Duke University Center for AIDS Research (CFAR), an NIH funded program (5P30 AI064518).

[†] Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003, sgorsky@umass.edu

[‡] Center for Human Systems Immunology, Duke University School of Medicine, Durham, NC 27708, cliburn.chan@duke.edu

[§] Department of Statistical Science, Duke University, Durham, NC 27708, li.ma@duke.edu

https://doi.org/10.1214/22-BA1356

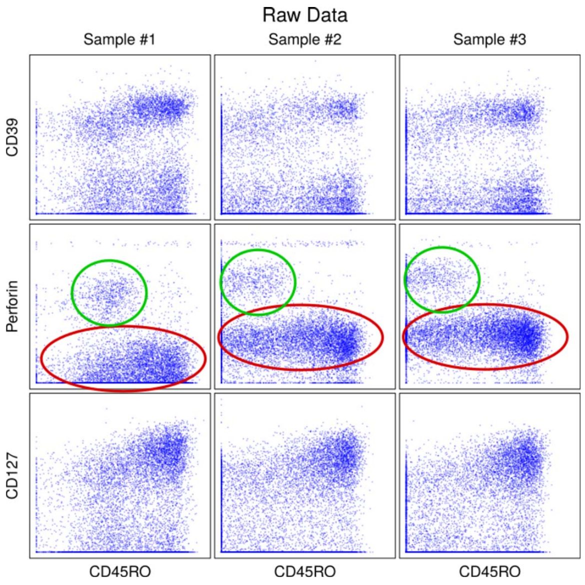
in blood and disaggregated tissues. In this assay, individual cells are first incubated with monoclonal antibodies linked to fluorescent molecules (flow cytometry) or heavy metal isotopes (mass cytometry), then passed through a detection system that quantifies the relative amount of each marker (a macromolecule recognized by the monoclonal antibody) on every cell. Flow and mass cytometry can scan millions of cells in solution in minutes and provide information on the distribution of cell types in a sample, most typically reporting on the distribution of immune cell subsets in a blood sample. This information has multiple applications in clinical research, including determining the immune response to infection or vaccine challenge. As the data generated from flow and mass cytometry are very similar in nature, we treat them interchangeably in our analysis.

Cytometry data can generally be presented as an $n \times p$ matrix with n being the number of cells and p the number of markers. Each row (observation) represents an individual cell, and each column (variable) represents a parameter (e.g., marker) value of the cells. In multiple-sample studies, each observation is also associated with a sample (e.g. with an additional column denoting a sample number.)

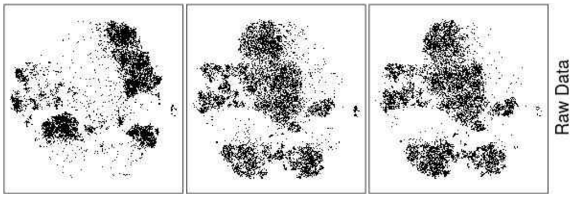
Traditionally, cytometry data is analyzed manually by visual demarcation of cell subsets on a sequence of 2D projections, a process known as gating. Manual gating becomes unwieldy as the data dimensionality grows, and automated methods for cell subset identification from cytometry data are becoming increasingly necessary.

However, variations in cell subset locations (in the marker space) frequently occur due to uncontrolled technical reasons unrelated to the underlying biological differences. In particular, technical differences often make it extremely challenging to compare samples processed in different batches and/or laboratories. Within a single cytometry laboratory, batch differences may occur because of instrument and reagent variability over time. In Figure 1 we provide an example of a data set arising from measuring the same biological specimen in three different samples. Such “negative controls”—as there are no underlying biological difference—are referred to as “batch controls”, whose main purpose is to allow the quantification of cross-sample technical variation. These sample differences are compounded when multiple laboratories are involved in processing the data, for example, in large multi-center vaccine trials, with additional variability introduced by center-specific instruments and sample processing protocols. Thus a prerequisite for proper analysis of such data is cross-sample calibration—aligning cell subsets across multiple samples—so that cell subset properties can be meaningfully compared.

While we present a first unified framework for achieving automated classification and calibration simultaneously, it is worth noting that a number of previous works in the literature for cytometry data have considered cross-sample variability in the context of cell classification, without explicitly addressing the calibration as a standalone task. For example, Cron et al. (2013) developed an extension to the Dirichlet Process Mixture (DPM) of Gaussian kernels, in which each sample has its own set of weights, thus accounting for sample variability in subset sizes. Dundar et al. (2014) further expanded the approach to also model cross-sample variability in kernel parameters by adding a DPM prior on the means of the Gaussian kernels. Pyne et al. (2014) and Lee et al. (2015) proposed a two-layer model where each sample is a finite mixture of multivariate



(a)



(b)

Figure 1: Plots of three samples of 19-dimensional, uncalibrated batch control cytometry data. (a) Three pairs of margins from the dataset. In circles are the misaligned clusters. (b) t-SNE plots of the raw data. The t-SNE method (van der Maaten and Hinton, 2008) is a nonlinear dimensionality reduction method that projects high-dimensional data to two dimensions. In the two-dimensional representation, points that are near one another are also more similar in the high dimensional space. Thus, clusters of points in the two-dimensional representation reflect clusters of points in the high-dimensional space. Here, we apply the t-SNE algorithm to the full data set (all samples) at once and present the results separately for each sample. The left sample is clearly misaligned with the two right samples. The goal of calibration in the context of batch control is to bring the different samples to look as similar as possible. In Section 4 we demonstrate how our method successfully performs this task for these data and compare it to other state-of-the-art methods for classification. There are many zero values in the data, likely arising from the number of metal isotopes being below the limit of detection when the cell expression of a marker is zero or very low. While this has been addressed using approaches such as zero-inflated models (Minoura et al., 2020), we have found that the “coarsening” strategy (Section 2.2) we utilize in our method provides a very effective and robust approach to addressing this challenge.

skew-t distributions with an embedded linear random effects model to allow cross-sample variation in the cluster, estimated using Expectation-Maximization (EM). Instead of linear mixed-effects models, Soriano and Ma (2019) suggested using a (nonparametric) mixture with a hierarchical kernel to allow for variability in cell subset locations across samples.

We present a model that adopts and enriches key features of these previous methods for characterizing cross-sample variation. Our approach incorporates all sources of uncertainty in a joint hierarchical model and can be applied either as a joint clustering method over related samples that adjust for cross-sample heterogeneity or as a standalone tool for cross-sample calibration that aligns the samples for downstream analysis. Specifically, we explicitly model cross-sample variability in cluster weights by allowing each sample its own set of weights. Second, we incorporate hierarchical multivariate skew normal kernels to characterize both the flexible shapes of the cell subsets as well as their variation across samples. This hierarchical kernel gives rise to a natural scheme for performing calibration, as explained in Section 2.3. Finally, we incorporate a recently introduced model-robustification strategy called "coarsening" (Miller and Dunson, 2018).

We have found that robustifying generative models for cytometry data is necessitated by the complex shapes of cell clusters and the presence of outliers. With this strategy, our method becomes robust to limitations of the skew Gaussian kernels and mitigates the undesirable issue of DPMs in producing a diverging number of clusters as sample size grows, thereby increasing the efficacy of our model in classification and calibration. A final contribution of our framework is computational. When fitting mixtures with multivariate skew kernels, maximum-likelihood estimation with expectation-maximization (EM) or Bayesian approaches with conjugate priors and Gibbs sampling face difficulties in overcoming the multimodality of the likelihood function of those kernels. We introduce a hybrid sampler that embeds a Population Monte Carlo (PMC) step into a Gibbs sampler, which reduces the risk of the sampler to be caught in a local mode and allows us to adopt non-conjugate priors effectively. We call our method COMIX, for COarsened MIXtures of hierarchical skew normal kernels, and provide an R package of the same name that implements our method.

Before describing our method in detail, we provide a quick review of some other previous statistical approaches to analyzing cytometry data in terms of classification only, without attempting to calibrate misaligned data. Murphy (1985) applied K-means cluster analysis to cytometry data. Generally, K-means clustering can be highly dependent on the initialization and centroid calculations, and lacks statistical interpretation. Bakker Schut et al. (1993) performed cluster analysis using K-means initialized with a large number of seed points, followed by a modified nearest neighbor technique to reduce the large number of subclusters. This method caters to symmetric clusters. Boedigheimer and Ferbas (2008) applied finite Gaussian mixtures to cytometry data, and adopted a frequentist estimation strategy based on the EM algorithm. Chan et al. (2008) fit cytometry data with a finite mixture of multivariate Gaussians using standard conjugate Bayesian analysis and Gibbs sampling for inference. Methods that model cytometry data with vanilla Gaussian kernels suffer from the obvious weakness that cell clusters are typically asymmetric. Malsiner-Walli et al. (2017) offered a Bayesian model that allows a finite mixture of mixtures of Gaussian kernels and demonstrate this approach on cytometry data. This method automates the selection of the number of clusters, and allows asymmetric clusters. In the frequentist literature, Pyne et al. (2010) formulated the `FLAME` framework that models cytometry data with a finite mixture model of skew-t distributions. O'Hagan et al. (2016) suggested using the multivariate normal inverse Gaussian distribution kernels in the context of finite mixture models. Lo et al. (2008) proposed a finite mixture of t distributions with a Box-Cox transformation in order to reduce asymmetry. This method depends on expert opinion or the utilization of additional methods to determine initial conditions that will increase the chances of convergence of the EM algorithm, which means that full automation of the process is difficult. Arellano-Valle et al. (2009) discussed Bayesian mixtures of multivariate skew normal kernels, but did not provide a unified approach that handles all three parameters of the skew normal distribution. Frühwirth-Schnatter and Pyne (2010) developed a Bayesian, fully conjugate multivariate finite mixture model with multivariate skew normal and skew-t distributions. As is further discussed in Section 2.4 this formulation is constrained by the strong conjugate prior structure and does not allow for intuitive treatment of calibration for cytometry data. Hejblum et al. (2017) used the Bayesian formulation of Frühwirth-Schnatter and Pyne (2010) that also accommodates dependencies within the data using a sampler that sequentially uses information from previous time points or previous samples as priors for successive estimations of (approximate) posterior distributions. Other methods for classification that do not involve explicit statistical modeling are commonly used for flow cytometry data. Two of the most popular are `FLowSOM` (Van Gassen et al., 2015) and `PhenoGraph` (Levine et al., 2015). `FLowSOM` consists of a workflow of four steps: reading the data, building a self-organizing map, building a minimal spanning tree and computing a meta-clustering result. `PhenoGraph` constructs a nearest-neighbor graph to capture the phenotypic relatedness of high-dimensional data points and then applies the Louvain graph partition algorithm to dissect the nearest-neighbor graph into phenotypically coherent subpopulations.

The rest of the paper is organized as follows. Section 2 provides the details of our Bayesian hierarchical model, the coarsening strategy, as well as the inference recipe for achieving classification and calibration jointly. In Section 3 we provide numerical examples on simulated data that demonstrate the efficacy of our model. In Section 4 we carry out a case study on a 19-dimensional mass cytometry data set.

2 Method

2.1 A Bayesian nonparametric hierarchical model

The complexity of the observed structures in cytometry data requires flexible statistical models to characterize their key features. Our model captures those by a DPM (Ferguson, 1983) with several modeling choices that allow sample variability in weights and kernel locations as well as flexibility in the kernel shapes. Since cytometry data rarely involve symmetric clusters we choose to work with skew normal (SN) kernels. These are

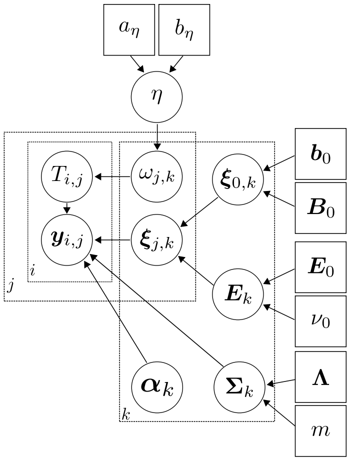


Figure 2: A graphical representation of our hierarchical model. j indicates the sample number; i the observation number within a sample j; k indexes cluster labels. $\boldsymbol{y}_{i,j}$ is the ith observation in sample j; $T_{i,j}$ the cluster assignment random variable specifying which cluster $\boldsymbol{y}_{i,j}$ belongs to. $\boldsymbol{\xi}_{j,k}$ is the sample and cluster-specific location parameter. $\omega_{j,k}$s are cluster and sample specific DPM weights. $\boldsymbol{\alpha}_k$ and $\boldsymbol{\Sigma}_k$ are, respectively, the skewness vector and scale matrix for each cluster, they are shared among all samples. $\boldsymbol{\xi}_{0,k}$ and $\boldsymbol{E}_k$ are, respectively, the mean and the covariance of the multivariate normal distribution from which the sample specific location parameters are drawn.

characterized by a location parameter $\boldsymbol{\xi}$, a scale parameter $\boldsymbol{\Sigma}$, and a skew parameter $\boldsymbol{\alpha}$. (See Supplementary Materials 1 (Gorsky et al., 2023) for further details about the multivariate SN distribution.)

To allow the clusters to vary in sizes across samples, we endow each sample with its own set of mixing weights. In order to allow differences in the location of the clusters, we posit a hierarchical structure for the SN kernels. Figure 2 presents a full graphical view of our model. We next describe the model components in detail.

In the following, we assume that p-dimensional data are generated in J samples:

$$\boldsymbol{y}_{i,j} \overset{iid}{\sim} F_j, \;\; i = 1, \dots, n_j \text{ and } j = 1, \dots, J$$

such that $\boldsymbol{y}_{i,j}$ is the ith observation in sample j, n_j is the number of observations in sample $j \in \{1, \dots, J\}$ and $n = \sum_{j=1}^{J} n_j$ is the total number of observations across all samples.

Cluster assignment and weights. Let $\mathcal{K}$ be a countable set of cluster labels shared over all samples. For each $i = 1, \dots, n_j$ and $j = 1, \dots, J$, let $T_{i,j}$ be the latent cluster assignment variable for observation i in sample j, such that $T_{i,j} = k$ if and only if $\boldsymbol{y}_{i,j}$ belongs to cluster k for each $k \in \mathcal{K}$. As discussed above, we allow each sample to have its own set of weights (i.e., cluster sizes), $\omega_{j,k} := \mathrm{P}(T_{i,j} = k)$. The subscript j indicates that cluster sizes may vary across samples.

To form a DPM, we assign a Griffiths-Engen-McCloskey (GEM) prior with parameter η (Ewens, 1990; Sethuraman, 1994) on the sample specific weights. That is, for each $j \in \{1, \ldots, J\}$, let

$$\{\omega_{j,k}\}_{k \in \mathcal{K}} \overset{iid}{\sim} \text{GEM}(\eta).$$

The weights $\{\omega_{j,k}\}_{k \in \mathcal{K}}$ are all positive. Nonetheless, these weights may be very close to zero in some sample but non-negligible in other samples. In such cases, a finite data set generated from the model may have certain clusters that include points in the sample with the non-negligible weights but no data points in the samples with the near zero weights. In order to learn η, we also assign it a hyperprior, $\eta \sim \text{Gamma}(a_\eta, b_\eta)$.

Hierarchical multivariate skew normal kernels. Assume now that each F_j has the density

$$f_j(\cdot) = \sum_{k \in \mathcal{K}} \omega_{j,k} \cdot g(\cdot \mid \boldsymbol{\lambda}_{j,k})$$

where g is the density function of the multivariate SN distribution, and each $\boldsymbol{\lambda}_{j,k}$ is comprised of the three parameters of the SN distribution: $\boldsymbol{\lambda}_{j,k} = \{\boldsymbol{\xi}_{j,k}, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k\}$. In particular, for each cluster k we assume that the scale parameter $\boldsymbol{\Sigma}_k$ and the skew parameter $\boldsymbol{\alpha}_k$ are the same across samples. We allow sample variability in the sample-specific location parameter $\boldsymbol{\xi}_{j,k}$ by assuming that they are distributed around a grand "centroid" cluster location $\boldsymbol{\xi}_{0,k}$ following a multivariate Gaussian with covariance $\boldsymbol{E}_k$. Thus we have the following hierarchical kernel for generating an observation in each cluster

$$[\boldsymbol{\xi}_{j,k} \mid \boldsymbol{\xi}_{0,k}, \boldsymbol{E}_k] \overset{iid}{\sim} \text{N}(\boldsymbol{\xi}_{0,k}, \boldsymbol{E}_k)$$
$$[\boldsymbol{y}_{i,j} \mid T_{i,j} = k, \boldsymbol{\xi}_{j,k}, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k] \overset{iid}{\sim} \text{SN}_p(\boldsymbol{\xi}_{j,k}, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k).$$

We will further discuss the justification for only letting the location parameter vary between samples in Section 2.2.

We assign multivariate Gaussian and inverse-Wishart priors for the means and covariances of the kernels:

$$\boldsymbol{\xi}_{0,k} \overset{iid}{\sim} \text{N}(\boldsymbol{b}_0, \boldsymbol{B}_0)$$
$$\boldsymbol{\Sigma}_k \overset{iid}{\sim} \mathcal{W}^{-1}(m, \boldsymbol{\Lambda})$$

and follow Liseo and Parisi (2013) and Parisi and Liseo (2018) in the prior assignment for the skewness parameter:

$$p(\boldsymbol{\delta}_k, \boldsymbol{\Sigma}_k) = p(\boldsymbol{\delta}_k \mid \boldsymbol{\Sigma}_k) p(\boldsymbol{\Sigma}_k)$$

$$p(\boldsymbol{\delta}_k \mid \boldsymbol{\Sigma}_k) = \left(\frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} \sqrt{|\boldsymbol{\Omega}_k|} \right)^{-1} \cdot \text{I}(\boldsymbol{\delta}_k^T \boldsymbol{\Omega}_k^{-1} \boldsymbol{\delta}_k < 1),$$

where $\text{I}(\cdot)$ is the indicator function, $\boldsymbol{\delta}_k$ is a transformed version of $\boldsymbol{\alpha}_k$ as explained in Supplementary Materials 1 (Gorsky et al., 2023), and $\boldsymbol{\Omega}_k$ is the correlation matrix corresponding to the covariance $\boldsymbol{\Sigma}_k$. That is, $\boldsymbol{\Omega}_k = \text{diag}(\Sigma_{k,1,1}^{-1/2}, \ldots, \Sigma_{k,p,p}^{-1/2}) \cdot \boldsymbol{\Sigma}_k \cdot$

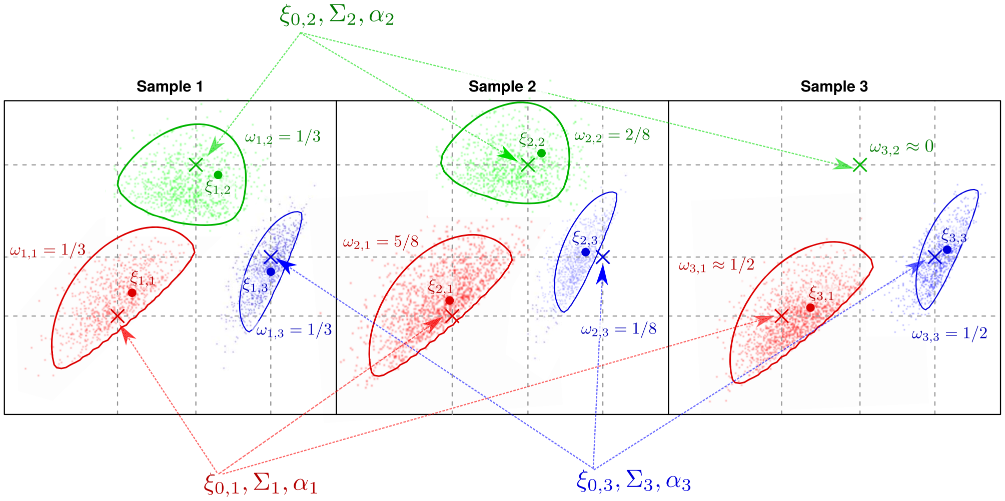


Figure 3: An illustration of a 3-sample data set generated from our model, limited to three clusters. The "centroid" cluster parameters $\boldsymbol{\xi}_{0,k}$, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\alpha}_k$ are all shared across the samples. Each sample has its own set of weights $\boldsymbol{\omega}_{j,k}$ and, e.g., cluster 2 in sample 3 is empty with $\omega_{3,2} \approx 0$ so that $T_{i,3} \neq 2$ for all $i \in \{1, \dots, n_3\}$. The location parameters of the first cluster $\boldsymbol{\xi}_{1,1}, \boldsymbol{\xi}_{2,1}$ and $\boldsymbol{\xi}_{3,1}$ are spread around $\boldsymbol{\xi}_{0,1}$. Similarly, $\boldsymbol{\xi}_{1,2}, \boldsymbol{\xi}_{2,2}, \boldsymbol{\xi}_{3,2}$ are spread around $\boldsymbol{\xi}_{0,2}$ and $\boldsymbol{\xi}_{1,3}, \boldsymbol{\xi}_{2,3}$ and $\boldsymbol{\xi}_{3,3}$ are spread around $\boldsymbol{\xi}_{0,3}$.

$\mathrm{diag}(\Sigma_{k,1,1}^{-1/2}, \dots, \Sigma_{k,p,p}^{-1/2})$ where $\mathrm{diag}(v_1, \dots, v_p)$ is the diagonal matrix whose diagonal elements are $(v_1, \dots, v_p)$. The priors for $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ induce priors on an alternative equivalent parametrization in terms of $\boldsymbol{\psi}$ and $\boldsymbol{G}$, which are derived from $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ by multiplying the following Jacobian term, defined separately for each cluster $k \in \mathcal{K}$:

$$|\mathcal{J}_k[(\boldsymbol{\xi}_{0,k}, \boldsymbol{\Sigma}_k, \boldsymbol{\delta}_k) \to (\boldsymbol{\xi}_{0,k}, \boldsymbol{G}_k, \boldsymbol{\psi}_k)]| = \prod_{j=1}^{p} (\boldsymbol{G}_k(j,j) + \boldsymbol{\psi}_k(j)^2)^{-\frac{1}{2}}.$$

In practice, we will infer on $\boldsymbol{\psi}_k$ and $\boldsymbol{G}_k$ and then transform them back to the original parameters. (See Supplementary Materials 1 for further details on the alternative parameterizations of the multivariate SN distribution and its computational advantages.)

We further assign an inverse-Wishart prior to the covariance of the normal distribution of the cluster locations around the grand location:

$$\boldsymbol{E}_k \overset{iid}{\sim} \mathcal{W}^{-1}(\nu_0, \boldsymbol{E}_0).$$

This completes the specification of our hierarchical model. Figure 2 provides a graphical model representation of the full hierarchical model while Figure 3 illustrates the structure of the data generated from this model.

2.2 Model robustification through coarsening

We use the "coarsening" strategy introduced in Miller and Dunson (2018) to make inference robust to model misspecifications. This is particularly relevant in the context of cytometry data, where clusters with parametric distributions usually do not adequately fit the actual shape of cell subsets and the massive sample sizes of cytometry data makes the resulting inference particularly sensitive to such model misspecifications.

Generally, for a model family indexed by some parameter θ, $\{P_\theta : \theta \in \Theta\}$, one defines an "idealized distribution" as a member in the model family that we use to represent the data generative mechanism. (Here it is the full hierarchical model presented above.) The coarsening approach assumes the existence of unobserved "ideal data", denoted by $Y_1, \ldots, Y_n$, from this distribution. The observed data, denoted by $y_1, \ldots, y_n$, are assumed to be drawn from a true distribution which is in an R-neighborhood (under some discrepancy measure d) of the idealized distribution. When the observed and ideal distributions are the same, Bayesian inference is conducted on the standard posterior distribution, $\pi(\theta \mid Y_1 = y_1, \ldots, Y_n = y_n)$. When they differ, Bayesian inference is performed on the "coarsened posterior": $\pi(\theta \mid d(\{Y_1, \ldots, Y_n\}, \{y_1, \ldots, y_n\}) < R)$. That is, the "posterior" is computed conditional on the event that the empirical distributions of the observed and ideal data are within an R-ball defined by a discrepancy measure d. R can be taken as a random variable, and assigned a prior. When $R \sim \exp(\gamma)$ and d is the Kullback-Leibler (KL) divergence (or d_n a consistent estimator of it), the coarsened posterior is approximated by:

$$\pi(\theta \mid d_n(\{Y_1, \ldots, Y_n\}, \{y_1, \ldots, y_n\}) < R) \underset{\sim}{\propto} p(\theta) \prod_{i=1}^{n} p_\theta(y_i)^\zeta$$

where $\underset{\sim}{\propto}$ means "approximately proportional to", $\zeta = \frac{\gamma}{\gamma+n}$, $p(\theta)$ is the prior on θ and p_θ is the density function of P_θ. Given some $\zeta \in [0,1]$, the quantity $\prod_{i=1}^{n} p_\theta(y_i)^\zeta$ is referred to as the "power likelihood" while $p(\theta) \prod_{i=1}^{n} p_\theta(y_i)^\zeta$ is referred to as the "power posterior". This form is easily implemented within the context of Gibbs sampling for mixture models, and we do so for our model. In our context, we consider the entire hierarchical DPM model as the idealized distribution. The types of deviations that the coarsening procedure tolerates depends on the discrepancy measure. In the current context, our hierarchical kernel robustifies inference with respect to cross-sample deviations in the cluster locations. Adopting the KL divergence in coarsening complements our model by robustifying inference with respect to variations in cluster shapes because the KL divergence tends to be more sensitive to changes in locations than shape variation.

2.3 Cross-sample calibration through posterior prediction

To calibrate multiple samples, we aim to shift each observation in a cluster in each sample by the estimated difference between the grand "centroid" mean and the sample-specific mean of that cluster. That is, when $T_{i,j} = k$ we compute a corrected value for each observation by adjusting for the shift in the mean (which is equal to the shift in the location parameters):

$$\tilde{\boldsymbol{y}}_{i,j} = \boldsymbol{y}_{i,j} - ((\boldsymbol{\xi}_{j,k} + \boldsymbol{\omega}_k \boldsymbol{\delta}_k \sqrt{2/\pi}) - (\boldsymbol{\xi}_{0,k} + \boldsymbol{\omega}_k \boldsymbol{\delta}_k \sqrt{2/\pi})) = \boldsymbol{y}_{i,j} - (\boldsymbol{\xi}_{j,k} - \boldsymbol{\xi}_{0,k})$$

where $\tilde{\boldsymbol{y}}_{i,j}$ is the shifted observation corresponding to the original observation $\boldsymbol{y}_{i,j}$. In the above, $\boldsymbol{\omega}$ and $\boldsymbol{\delta}$ are alternative parameterizations to $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$. See Supplementary Materials 1 (Gorsky et al., 2023) for further details on the multivariate SN distribution. To incorporate the posterior uncertainty in the cluster assignment, we follow the technique suggested in Soriano and Ma (2017) for calibration by integrating out the cluster assignment variables $T_{i,j}$:

$$\mathrm{E}[\tilde{\boldsymbol{y}}_{i,j} \mid \{\boldsymbol{y}_{i,j}\}] = \boldsymbol{y}_{i,j} - \mathrm{E}[\boldsymbol{\xi}_{j,k} - \boldsymbol{\xi}_{0,k} \mid \{\boldsymbol{y}_{i,j}\}] \approx \boldsymbol{y}_{i,j} - \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\xi}^{(n)}_{j,T^{(n)}_{i,j}} - \boldsymbol{\xi}^{(n)}_{0,T^{(n)}_{i,j}}).$$

where $\{\boldsymbol{y}_{i,j}\}$ represents the totality of all observations, N is the number of steps saved in the MCMC chain, $T^{(n)}_{i,j}$ is the assigned label to observation i in sample j at the nth step in the saved MCMC chain, $\boldsymbol{\xi}^{(n)}_{j,T^{(n)}_{i,j}}$ the estimated location parameter for sample j and cluster $T^{(n)}_{i,j}$ at the nth step in the saved MCMC chain and $\boldsymbol{\xi}^{(n)}_{0,T^{(n)}_{i,j}}$ the estimated "grand location" parameter for cluster $T^{(n)}_{i,j}$ at the nth step in the saved MCMC chain. A desirable byproduct of this integrating-out strategy is that through it we also bypass potential label switching issues, which will be discussed further in the next subsection.

2.4 Posterior computation by hybrid Gibbs-PMC sampling

The multimodality of the SN likelihood poses difficulties to EM estimation in frequentist settings and to Gibbs samplers in Bayesian settings. For example, Frühwirth-Schnatter and Pyne (2010) offered a conjugate structure and a Gibbs sampler to perform Bayesian estimation of the parameters of the SN distribution. Liseo and Parisi (2013) demonstrated how this approach may fail when multimodality arises in the likelihood. In addition, our experimentation in simulation studies of the conjugate prior approach also suggests that the posterior estimates are very close to the prior values, regardless of the values used to generate the data. One would often need hundreds of thousands of observations to allow correct inference for highly skewed kernels. Furthermore, the conjugate prior structure entails the elicitation of a joint prior for the location parameter $\boldsymbol{\xi}$ and the skewness parameter $\boldsymbol{\psi}$, which in the context of cytometry data will make tasks such as cross-sample calibration difficult. For these reasons, it is important to find a reasonable non-conjugate prior on the SN kernel parameters along with an efficient computational strategy in the context of cytometry data.

Liseo and Parisi (2013) suggested utilizing the Population Monte Carlo (PMC) approach to tackle the problem of multimodality while allowing prior modeling on the location parameter independently from those of the other SN parameters—that is, $\boldsymbol{\xi} \perp\!\!\!\perp (\boldsymbol{\Sigma}, \boldsymbol{\delta})$ *a priori*. This strategy allows us to select flexible and intuitive priors that offer a simple hierarchical structure for the location parameter. Because the resulting priors are not conjugate, we cannot use a vanilla Gibbs sampler. (In contrast, Hejblum et al. (2017), Dundar et al. (2014), and Soriano and Ma (2019) all constrained their models so that the prior structure will be conjugate and allow blocked Gibbs sampling.) We thus construct a hybrid "Gibbs-PMC" sampler.

A "Gibbs-PMC" hybrid sampler. Our sampler uses PMC moves for the SN parameters $\boldsymbol{\xi}_{j,k}$, $\boldsymbol{G}_k$, $\boldsymbol{\psi}_k$, $\boldsymbol{\xi}_{0,k}$, $\boldsymbol{E}_k$ and $z_{i,j}$. Given summary statistics (mean) of all particles for these parameters, it uses Gibbs moves for the weights $\omega_{j,k}$ and latent cluster assignment $T_{i,j}$ along with Metropolis-Hastings moves for the DPM concentration parameter η. The full sampling algorithm is as follows.

- Step 0: Initialize $T_{i,j}$ and a population of M particles $\boldsymbol{\xi}_{j,k}^{1:M}$, $\boldsymbol{G}_k^{1:M}$, $\boldsymbol{\psi}_k^{1:M}$, $\boldsymbol{\xi}_{0,k}^{1:M}$, $\boldsymbol{E}_k^{1:M}$ and $z_{i,j}^{1:M}$ for $j \in \{1, \dots, J\}$, $i \in \{1, \dots, n_j\}$ and $k \in \mathcal{K}$, the set of all cluster labels.

- Step $t > 0$:
  - Update the DPM concentration parameter η using a Metropolis-Hastings step with the proposal:

    $$\eta^* \mid \eta \sim \text{Gamma}(\eta^2 \cdot \eta_0, \eta \cdot \eta_0)$$

    where η_0 is calibrated during the burn-in iterations.
  - Sample from the full conditional distribution (FCD) of the mixture weights, $\omega_{j,k}$.
  - For $k \in \mathcal{K}$:

    If the number of observations in cluster k is 0, sample particles from priors. Else, for cluster k, follow the PMC sampling scheme suggested by Liseo and Parisi (2013):
    - For every $j \in \{1, \dots, J\}$, sample M particles $z_{1,j}^{1:M}, \dots, z_{n_{j,k},j}^{1:M}$ from the proposal $q_z^{(m)}$ as the FCD of $z_{i,j}$ (i.e. $z_{i,j}^{(m)}$ depends via $q_z^{(m)}$ on $\boldsymbol{\psi}_k^{(m)}$, $\boldsymbol{G}_k^{(m)}$, $\boldsymbol{\xi}_{j,k}^{(m)}$ and $\boldsymbol{\xi}_{0,k}^{(m)}$ for $m = 1, \dots, M$).
    - In a random order, perform the next 5 updating steps:
      1. For every $j \in \{1, \dots, J\}$, sample M particles $\boldsymbol{\xi}_{j,k}^{1:M}$ from the proposal $q_{\boldsymbol{\xi}_{j,k}}^{(m)}$ as the FCD of $\boldsymbol{\xi}_{j,k}$.
      2. Sample M particles $\boldsymbol{G}_k^{1:M}$ from the proposal $q_{\boldsymbol{G}_k}^{(m)}$ as the inverse-Wishart part of the FCD of $\boldsymbol{G}_k$.
      3. Sample M particles $\boldsymbol{\psi}_k^{1:M}$ from the proposal $q_{\boldsymbol{\psi}_k}^{(m)}$ as the p-dimensional multivariate normal part of the FCD of $\boldsymbol{\psi}_{j,k}$.
      4. Sample M particles $\boldsymbol{\xi}_{0,k}^{1:M}$ from the proposal $q_{\boldsymbol{\xi}_{0,k}}^{(m)}$ as the FCD of $\boldsymbol{\xi}_{0,k}$.
      5. Sample M particles $\boldsymbol{E}_k^{1:M}$ from the proposal $q_{\boldsymbol{E}_k}^{(m)}$ as the FCD of $\boldsymbol{E}_k$.
    - Compute the ratios

      $$\varrho^{(m)} \propto \frac{\pi\left(\{\boldsymbol{\xi}_{j,k}^{(m)}\}_j, \boldsymbol{G}_k^{(m)}, \boldsymbol{\psi}_k^{(m)}, \boldsymbol{\xi}_{0,k}^{(m)}, \boldsymbol{E}_k^{(m)}, \{z_{i,j}^{(m)}\}_{j,i} \mid \{\boldsymbol{y}_{i,j}\}_{j,i}\right)}{q^{(m)}\left(\boldsymbol{\xi}_{j,k}^{(m)}, \boldsymbol{G}_k^{(m)}, \boldsymbol{\psi}_k^{(m)}, \boldsymbol{\xi}_{0,k}^{(m)}, \boldsymbol{E}_k^{(m)}, \{z_{i,j}^{(m)}\}_{j,i}\right)}$$

      where $\{\cdot\}_j$ is a shorthand notation for $\{\cdot\}_{j\in\{1,\dots,J\}}$, $\{\cdot\}_{j,i}$ is a shorthand notation for $\{\cdot\}_{j\in\{1,\dots,J\}, i\in\{1,\dots,n_{j,k}\}}$ and $q^{(m)}$ is the joint proposal for each particle.

    - ∗ Scale the $\{\varrho^{(m)}\}_{m\in\{1,\dots,M\}}$ to sum to 1.
    - ∗ Resample $\left\{\{\boldsymbol{\xi}_{j,k}^{(m)}\}_j, \boldsymbol{G}_k^{(m)}, \boldsymbol{\psi}_k^{(m)}, \boldsymbol{\xi}_{0,k}^{(m)}, \boldsymbol{E}_k^{(m)}\right\}_{m\in\{1,\dots,M\}}$ according to the weights $\{\varrho^{(m)}\}_{m\in\{1,\dots,M\}}$.
    - ∗ Compute (mean over M index) $\left\{\overline{\boldsymbol{\xi}_{j,k}^{1:M}}\right\}_j$, $\overline{\boldsymbol{G}_k^{1:M}}$, $\overline{\boldsymbol{\psi}_k^{1:M}}$, $\overline{\boldsymbol{\xi}_{0,k}^{1:M}}$, $\overline{\boldsymbol{E}_k^{1:M}}$ and $\left\{\overline{z_{i,j}^{1:M}}\right\}_{j,i}$.
      (This is the last per-cluster PMC step)

- Sample from the FCD of $T_{i,j}$, based on the values $\left\{\overline{\boldsymbol{\xi}_{j,k}^{1:M}}\right\}_{j\in\{1,\dots,J\},k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{G}_k^{1:M}}\right\}_{k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{\psi}_k^{1:M}}\right\}_{k\in\mathcal{K}}$ and $\{\omega_{j,k}\}_{j\in\{1,\dots,J\},k\in\mathcal{K}}$
- (For each t) store: $\left\{\overline{\boldsymbol{\xi}_{j,k}^{1:M}}\right\}_{j\in\{1,\dots,J\},k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{G}_k^{1:M}}\right\}_{k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{\psi}_k^{1:M}}\right\}_{k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{\xi}_{0,k}^{1:M}}\right\}_{k\in\mathcal{K}}$, $\left\{\overline{\boldsymbol{E}_k^{1:M}}\right\}_{k\in\mathcal{K}}$ and $\left\{\overline{z_{i,j}^{1:M}}\right\}_{j,i}$, $\{T_{i,j}\}_{j,i}$, η and $\{\omega_{j,k}\}_{j\in\{1,\dots,J\},k\in\mathcal{K}}$.
- Merge clusters for which the Kullback-Leibler divergence is smaller than a preset threshold.

The full conditional and proposal distributions used in the sampler are described in Supplementary Materials 2 (Gorsky et al., 2023).

In implementing the sampler, we utilize the finite-dimensional symmetric Dirichlet distribution approximation (Ishwaran and James, 2001) to the Dirichlet process mixture. With this approximation, the J infinite sequences of mixture weights $\{\omega_{j,k}\}_{k\in\mathcal{K}}$ are replaced for each $j = 1,\dots,J$ by:

$$\{\omega_{j,k}\}_{k\in\{1,\dots,K\}} \overset{iid}{\sim} \text{Dirichlet}(\eta/K,\dots,\eta/K)$$

where we need to specify the maximal number of clusters K, which should be much larger than the actual number of clusters to provide adequate approximation to the DPM. For cytometry data, we found that a value of $K \geq 40$ to 50 is generally adequate. This approximation is a special case of a more general class of models called sparse finite mixtures (Malsiner-Walli et al., 2016). A known issue of this approximation for DPMs when applied to large data sets is that the posterior will concentrate on a diverging number of clusters and K is often the *de facto* number of estimated clusters. In our approach, however, this issue is addressed with the coarsening strategy—we can set K to be very large while the estimated number of clusters will still be much smaller than K even for large and noisy data sets such as those from flow or mass cytometry.

A common phenomenon in inference algorithms for mixture models that utilize a cluster assignment variable is known as label switching. Since the values the cluster assignment variable assigns (the labels) to the different clusters are exchangeable, the prior and posterior distributions for the parameters of the mixture components are symmetric with respect to permutations of the labels. This causes no concern in our framework when calibration is the sole purpose as our calibration strategy integrates out the different labeling scenarios. For cell classification, on the other hand, label switching needs to be addressed and many possible strategies are available. Our software implementation handles this issue *post hoc* using the method of Cron and West (2011). A coherent classification is maintained by choosing a reference classification, which we take from the last iteration of the MCMC chain. Then, for each saved classification a cost matrix is computed (based only on the values of the cluster assignment variable) and minimized by selecting a permutation on its columns using the Hungarian algorithm of Munkres (1957). The resulting minimizing permutation is then applied to the cluster labels. Individual labels for each observation are set to the mode from the set of all labels generated in the saved MCMC chain for an observation.

3 Simulation study

We demonstrate our method using 18-dimensional simulated data consisting of 10 samples with a total of 64,622 observations across all samples. To examine the robustness of our method to model misspecification in kernel shapes and the effects of coarsening, we apply our method to two different data sets. The first is simulated under a finite mixture model with 7 multivariate skew-normal components, with the same hierarchical structure for the kernels as in our model. This represents a case in which our model is correctly specified even without coarsening. The second data set is generated by "distorting" the first, "ideal" data set to induce model misspecification. In the distorted samples three out of the seven clusters are narrowed asymmetrically. Figure 4 presents two margins from both the "ideal" data (top row) and the "distorted" data (bottom row).

To each data set, we apply COMIX with $K = 40$, five PMC particles, and a burn-in period of 3,500 samples, and save 500 samples from the chain. To assess convergence, trace plots of the MCMC chain for several parameters are shown in Supplementary Materials 5 (Gorsky et al., 2023). All other parameters and hyperpriors are set to

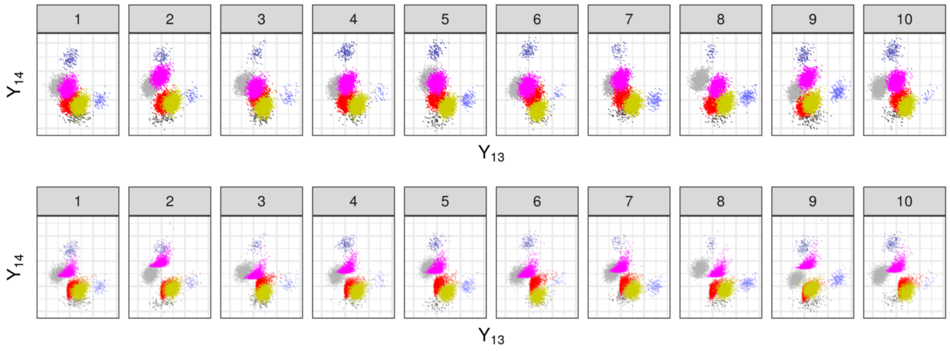


Figure 4: Two-dimensional scatter plots of the two 18-dimensional simulated data sets. Top row: an "ideal" data set with 7 clusters in 10 samples. Bottom row: a distorted version of the "ideal" data. The colors correspond to true cluster labels.

their default values for our implementation, as specified in Supplementary Materials 3 (Gorsky et al., 2023). The total run time for each model fit on a laptop computer utilizing four 3.00GHz Intel® Xeon(R) E3-1505M v6 CPU cores is approximately 1 hour and 40 minutes, utilizing a total of 1.3GB of RAM. We have found in many numerical studies that $\zeta = 0.2$ provides robust results for classification and calibration in many contexts where the observed data deviate from the theoretical skew kernels. As such we adopt 0.2 as a default value in our software. In order to demonstrate the effects of coarsening, we show model estimates and calibration results with our recommended level of coarsening ($\zeta = 0.2$), a suboptimal level of coarsening ($\zeta = 0.5$) and without coarsening ($\zeta = 1$). In applications we recommend the user to still carry out a context-specific sensitivity analysis for ζ to ensure that an appropriate value is selected. We carry out a sensitivity analysis on the value of ζ and provide the details in Supplementary Materials 4 (Gorsky et al., 2023). Later we will demonstrate how the recommended level 0.2 produces robust analysis in the case studies.

The classification and calibration results for the data drawn from the "ideal" kernels are shown in Figures 5 and 6. The results for the distorted data are shown in Figures 7 and 8. In Figure 5, where the clusters are drawn from the same distribution as in our generative model, each true cluster is identified as such both with and without coarsening. However, in Figure 7, where the distorted data are not exactly from SN kernels as our generative model prescribes, the true clusters are broken into several clusters

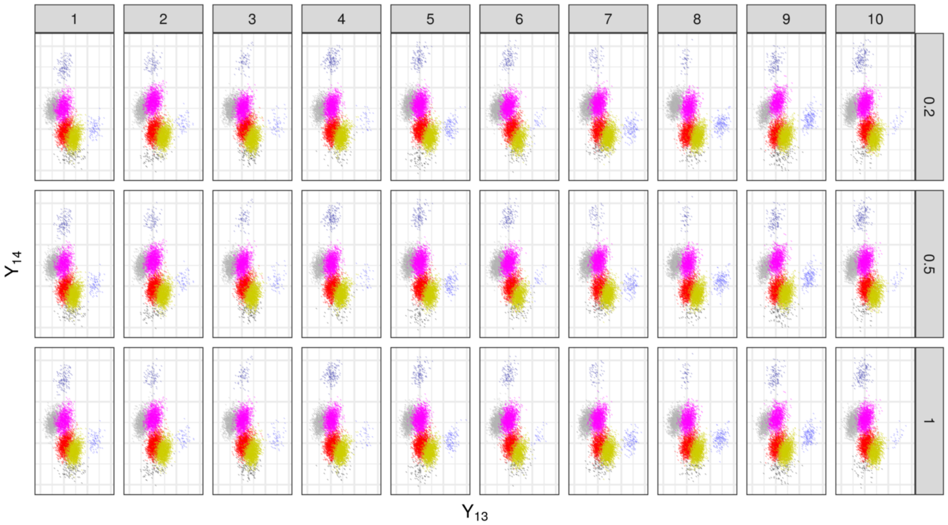


Figure 5: Calibrated samples for the "ideal" data with our recommended level of coarsening (top row, $\zeta = 0.2$), a suboptimal level of coarsening (middle row, $\zeta = 0.5$) and without coarsening (bottom row, $\zeta = 1$). The colors indicate estimated cluster label assignment. In this case, calibration results are similar across the different levels of coarsening.

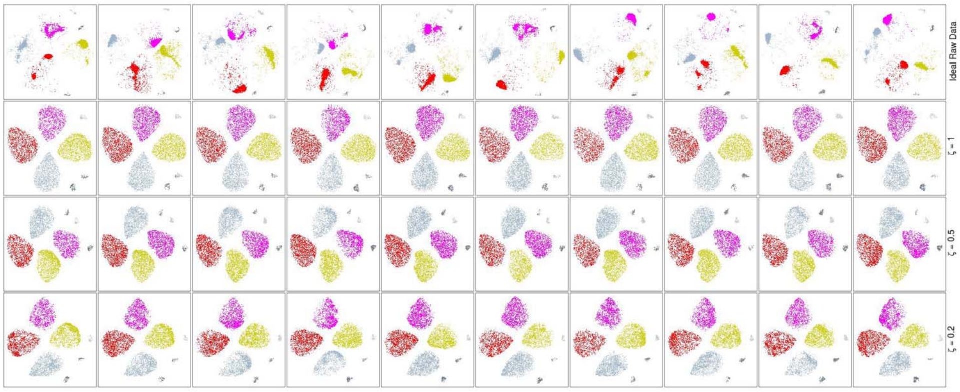


Figure 6: t-SNE plots showing the raw “ideal” data (top row) and calibration results for three different levels of coarsening: $\zeta = 1$ (top row), $\zeta = 0.5$ (middle row) and $\zeta = 0.2$ (bottom row). The colors indicate the true cluster labels in the top row, and the estimated cluster labels in the bottom three rows. Calibration results are similar across the different levels of coarsening.

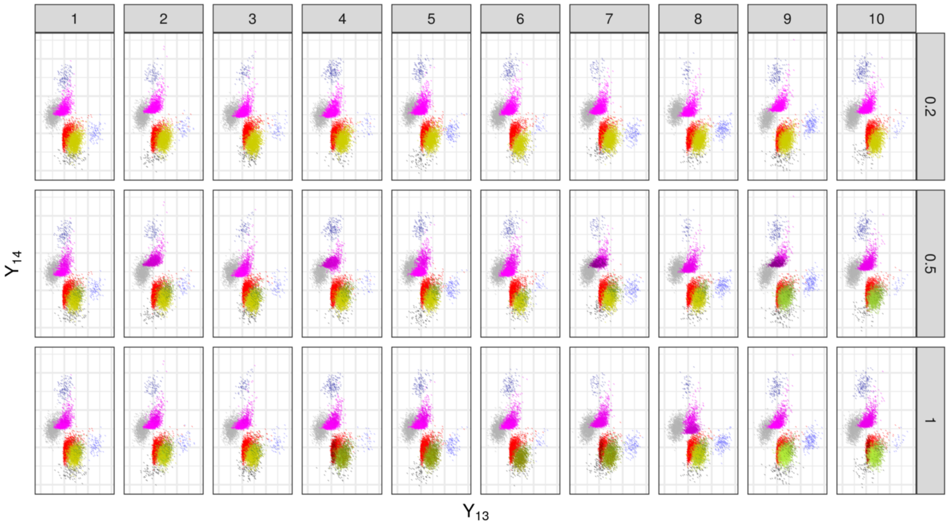


Figure 7: Calibrated samples for the distorted data with our recommended level of coarsening (top row, $\zeta = 0.2$), a suboptimal level of coarsening (middle row, $\zeta = 0.5$) and without coarsening (bottom row, $\zeta = 1$). The colors indicate estimated cluster label assignment. Only at $\zeta = 0.2$ do we get satisfactory results in calibration (as well as classification).

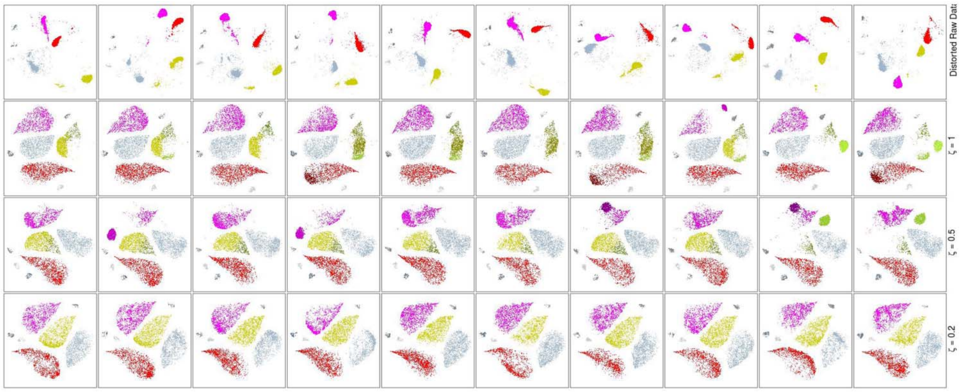


Figure 8: t-SNE plots showing the raw distorted data (top row) and calibration results for three different levels of coarsening: $\zeta = 1$ (top row), $\zeta = 0.5$ (middle row) and $\zeta = 0.2$ (bottom row). The colors indicate the true cluster labels in the top row, and the estimated cluster labels in the bottom three rows. Only at $\zeta = 0.2$ do we get satisfactory results in calibration (as well as classification).

without coarsening ($\zeta = 1$). Coarsening helps identify and align the distorted clusters correctly even though they are misspeciefied by our generative model. A comparison to a model fit with a simplified version of COMIX, with Gaussian kernels and no coarsening, is discussed in Supplementary Materials 8 (Gorsky et al., 2023).

Figures 3–22 in Supplementary Materials 5 (Gorsky et al., 2023) show trace plots and effective sample sizes of two key parameters, the weights $\omega_{j,k}$ and the grand means $\boldsymbol{\xi}_{0,k}$. The plots show that we achieve MCMC convergence with good effective sample sizes for all but the smallest clusters.

4 Case study: mass cytometry data

4.1 Fitting the COMIX model

We further apply our method to a publicly available 19-dimensional data set with 30,000 observations across three samples collected using a mass cytometer. The mass cytometry data was published in Kleinsteuber et al. (2016), and consists of seven Flow Cytometry Standard (FCS) files downloaded from FlowRepository (`http://flowrepository.org/id/FR-FCM-ZZTY`). The data from the FCS file is a matrix where the rows (observations) represent individual cells and the columns (variables) represent parameters of the cells. Generally, these parameters, also known as markers, represent counts for the specific type of cell surface protein (e.g. CD3, a component of the T cell receptor) that the isotope-labeled antibody is specific for. The distribution of these variables is highly right skewed, and part of the standard preprocessing for mass cytometry data is to apply an inverse hyperbolic sine or arcsinh transformation (see Figure 9). The arcsinh

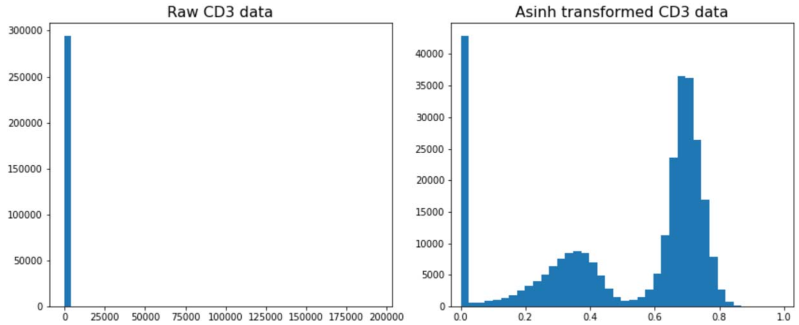


Figure 9: Data before and after arcsinh transformation.

transformation is applied to each of the variables independently. After the transform, each variable typically has a multimodal distribution. For example, there are 3 peaks in the histogram for CD3 – a zero peak representing cells that do not express CD3, and two peaks that represent cells expressing intermediate and high levels of CD3 respectively.

Kleinsteuber et al. (2016) described the use of spiked-in batch control samples to enable manual calibration of mass cytometry data for comparative analysis. By design, the test samples in this study were spiked with CD45-barcoded peripheral blood mononuclear cells (PBMC) from the same healthy donor as an internal control. Specifically, each sample contains 2×10^6 CD8 T cells from HIV-infected patients labeled with anti-CD45 antibodies conjugated to 141Pr, together with 4×10^5 spiked PBMC from the same healthy donor labeled with anti-CD45 antibodies conjugated to 89Y. Using the CD45 barcode, the spiked cells were retrieved from each sample and treated as batch controls for the evaluation of COMIX calibration. We show that COMIX can align the data using these spiked-in controls, providing a proof in principle approach to the automated calibration of mass cytometry data with spiked-in batch controls.

To each data set, we apply COMIX with $K = 50$, 50 PMC particles, and a burn-in period of 3,500 samples, and save 500 samples from the chain. All other parameters and hyperpriors are set to their default values for our implementation, as specified in Supplementary Materials 3 (Gorsky et al., 2023). The total run time for each model fit on a laptop computer utilizing four 3.00GHz Intel® Xeon(R) E3-1505M v6 CPU cores is approximately 6 hours and 30 minutes, utilizing a total of 900MB.

Figure 10 shows t-SNE plots for the raw and calibrated data of the three samples for $\zeta = 0.2$, $\zeta = 0.5$ and $\zeta = 1$. The calibrated data is very similar across the samples for all levels of coarsening, attaining our goal. In this case, we get that for $\zeta = 0.5$ and $\zeta = 1$ the number of estimated clusters is 12 across all samples, whereas for $\zeta = 0.2$ it is 11. Figure 11 shows some marginal scatter plots of the raw and calibrated data for the three settings. Although the results are largely similar, here too the smaller number of estimated clusters due to coarsening is beneficial, as the smallest cluster in the CD45RO-Perforin plot of the left sample for $\zeta = 0.5$ and $\zeta = 1$ appears to be artificial and uninformative.

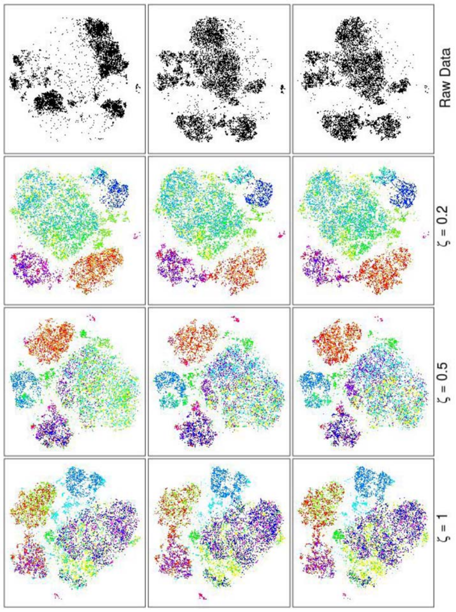


Figure 10: t-SNE plots for the 19-dimensional data. Top row: raw data. Second row from top: calibrated data for 3 samples with $\zeta = 0.2$. Third row from top: calibrated data for 3 samples with $\zeta = 0.5$. Bottom row: calibrated data for 3 samples with $\zeta = 1$. Color coding corresponds to estimated cluster assignment label.

Figure 12 presents the marginal densities of the 19 markers before and after calibration.

Figures 23–32 in Supplementary Materials 5 (Gorsky et al., 2023) show trace plots and effective sample sizes of two key parameters, the weights $\omega_{j,k}$ and the grand means $\boldsymbol{\xi}_{0,k}$. The plots show that we achieve MCMC convergence with good effective sample sizes for all but the smallest clusters.

Posterior mean estimates for skewness of the non-empty clusters when $\zeta = 1$ are shown in Tables 2 and 3 in Supplementary Materials 7 (Gorsky et al., 2023). (Here we focus on $\zeta = 1$ to eliminate coarsening thereby demonstrating the extent to which the generative model alone captures the skewness in the data.) The high skewness values indicate that there are significant shape deviations from Gaussian kernels. A comparison to a model fit with a simplified version of COMIX, with Gaussian kernels and no coarsening, is discussed in Supplementary Materials 9 (Gorsky et al., 2023).

The results show that COMIX works well for calibration of mass cytometry data with 19 dimensions, which span the range of dimensions used in the vast majority of such experimental data sets. Calibration is broadly useful not just for multi-center studies, but also for studies across batches of data as cytometer performance characteristics, antibody lots, and sample preparation often vary over time, necessitating time-consuming

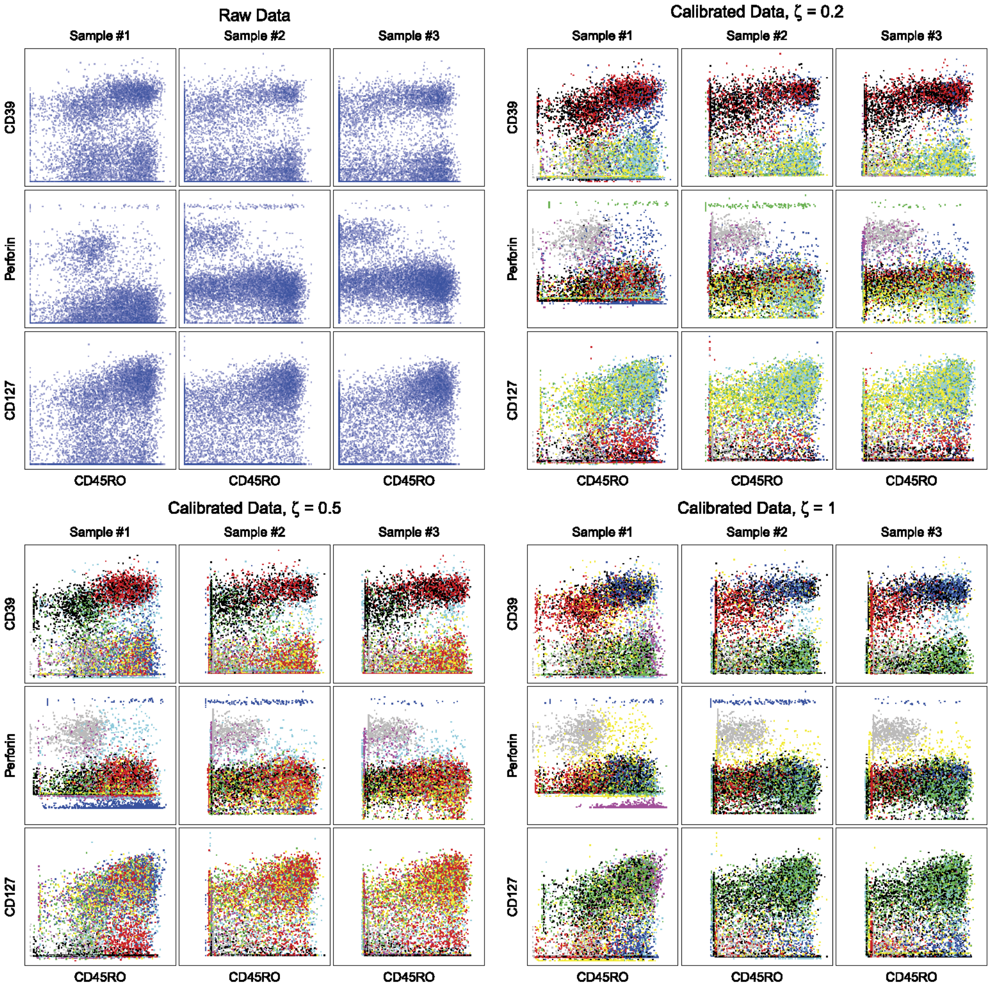


Figure 11: Marginal scatter plots for some markers and all samples in high dimensional mass cytometry data. Raw data (top left); calibrated data for $\zeta = 0.2$ (top right), $\zeta = 0.5$ (bottom left) and $\zeta = 1$ (bottom right). Color coding reflects inferred cluster labels.

and error-prone manual adjustment of gates across batches in manual analysis and reducing the robustness of automated methods that ignore the need for calibration.

4.2 Comparing COMIX to other state-of-the-art methods

To evaluate the contribution of cross-sample calibration to classification robustness, we compare the performance of COMIX to two state-of-the-art methods for cytometric cell subset classification, `FlowSOM` (Van Gassen et al., 2015) and `PhenoGraph` (Levine et al.,

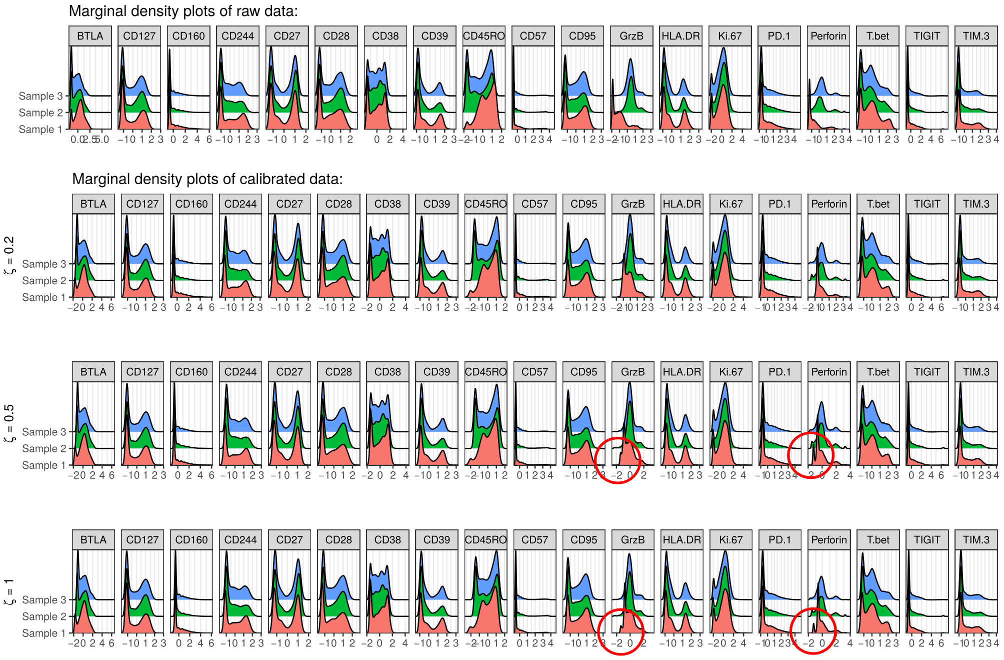


Figure 12: Marginal density plots for all markers and samples in the 19-dimensional mass cytometry data set. In red circles we highlight areas in the calibrated samples for $\zeta = 0.5$ and $\zeta = 1$ that are not as well aligned compared to the sample calibrated with $\zeta = 0.2$.

2015). `FlowSOM` first builds linked clusters using self-organizing maps and a minimal spanning tree, then applies a hierarchical consensus clustering (Wilkerson and Hayes, 2010) for meta-clustering across samples. The main parameter for the meta-clustering step is the number of clusters, which has to be set in advance. It is worth noting that unlike the maximal number of clusters K in COMIX, which only needs to be sufficiently large to provide adequate approximation to the DPM and may be very large without harming the analysis, the preset number of clusters in `FlowSOM` needs to be sufficiently close to the actual number of cell clusters to render reliable analysis. `PhenoGraph` identifies subpopulations, which are equivalent to clusters. The main tuning parameter of `PhenoGraph` is the number of nearest neighbors to be used for the nearest-neighbor graph to capture the phenotypic relatedness of data points. A higher tuning parameter prefers fewer, larger clusters. Meta-clustering across samples is performed on cluster centroids with partitioning across medoids (PAM) using the average silhouette width to determine the number of meta-clusters. Figure 13 shows the results of the classification provided by the two methods when applied to the mass cytometry data from Section 4.1, thereby allowing us to directly compare the performance of those to that of COMIX.

`FlowSOM`, when asked to classify the data to 12 clusters (top row in Figure 13)

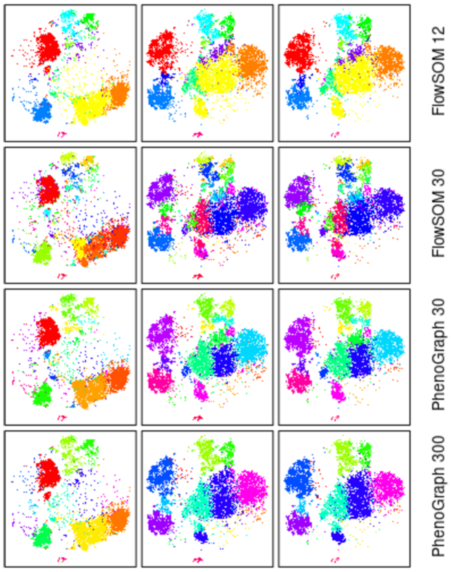


Figure 13: t-SNE plots for the 19-dimensional mass cytometry data. Top row: classification results for FlowSOM when 12 clusters are requested. Second row from top: classification results for FlowSOM when 30 clusters are requested. Third row from top: classification results for PhenoGraph with the tuning parameter set to 30 (the default value). Bottom row: classification results for PhenoGraph with the tuning parameter set to 300. Color coding corresponds to estimated cluster assignment label.

correctly identifies the misaligned cluster (yellow and orange cluster). However, in all other settings that cluster is identified as separate. This kind of result will be pervasive in the presence of misalignment with algorithms that do not explicitly account for it. When the correct number of clusters is correctly identified *a priori*, which is uncommon in practice, such algorithms could sometimes succeed in identifying the misalignment, but in general they tend to break larger clusters into an increasing number of smaller clusters as the number of specified clusters grows, and the first “victims” for this misclassification will be the misaligned clusters.

5 Discussion

We have presented a principled probabilistic approach for calibrating and classifying multi-sample cytometry data. Our approach utilizes a flexible Bayesian nonparametric mixture model with multivariate SN kernels to accommodate the key features of cytometry data and incorporate the “coarsening” strategy to make inference robust to model misspecification. Moreover, we constructed a Gibbs-PMC hybrid sampler, which embeds PMC moves for the SN parameters into a Gibbs sampler, thereby addressing the multimodality of the posterior on the kernel parameters.

It is possible to extend the COMIX model to incorporate additional experimental design features. For example, when the samples fall into multiple batches and the batch information is available, one could incorporate cross-batch variability by inserting an additional layer of hierarchical modeling into the kernel parameters and/or the weight generation.

While our method is motivated by and developed for the purpose of analyzing multi-sample cytometry data, the modeling and inference techniques used are generally applicable to other multi-sample data that can be effectively modeled by mixtures. In particular, the idea of adopting a flexible kernel, allowing hierarchical structure on the kernel, and incorporating coarsening to further robustify the method can all be readily applied to other types of data.

Software

Our R package `COMIX` is available at https://CRAN.R-project.org/package=COMIX. Code for the numerical examples is available at https://github.com/MaStatLab/COMIX_Numerical_Examples.

Supplementary Material

Supplementary material for coarsened mixtures of hierarchical skew normal kernels for flow and mass cytometry analyses (DOI: 10.1214/22-BA1356SUPP; .pdf). Includes the following sections: The multivariate skew normal distribution; Full conditionals and MCMC proposals; COMIX: default hyperpriors and parameters; Case study: sensitivity analysis; Simulation study: MCMC trace plots and diagnostics for main estimated parameters; Case study: MCMC trace plots and diagnostics for main estimated parameters; Case study: estimates of skewness parameters; Simulation study: model fit with Gaussian kernels; Case study: model fit with Gaussian kernels; Data generating mechanism for the simulation study of Section 3.

References

Arellano-Valle, R. B., Genton, M. G., and Loschi, R. H. (2009). "Shape mixtures of multivariate skew-normal distributions." *Journal of Multivariate Analysis*, 100(1): 91–101. MR2460479. doi: https://doi.org/10.1016/j.jmva.2008.03.009. 443

Bakker Schut, T., de Grooth, B., and Greve, J. (1993). "Cluster Analysis of Flow Cytometric List Mode Data on a Personal Computer." *Cytometry*, 14(1): 649–659. 442

Boedigheimer, M. J. and Ferbas, J. (2008). "Mixture modeling approach to flow cytometry data." *Cytometry Part A*, 73(5): 421–429. 442

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. B. (2008). "Statistical mixture modeling for cell subtype identification in flow cytometry." *Cytometry Part A*, 73(8): 693–701. 442

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J. P., van der Burg, S. H., West, M., and Chan, C. (2013). "Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples." *PLOS Computational Biology*, 9(7): 1–14. 440

Cron, A. J. and West, M. (2011). "Efficient Classification-Based Relabeling in Mixture Models." *The American Statistician*, 65(1): 16–20. MR2899648. doi: https://doi.org/10.1198/tast.2011.10170. 451

Dundar, M., Akova, F., Yerebakan, H. Z., and Rajwa, B. (2014). "A non-parametric Bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects." *BMC Bioinformatics*, 15(1): 314. 440, 448

Ewens, W. J. (1990). *Population Genetics Theory - The Past and the Future*, 177–227. Dordrecht: Springer Netherlands. MR1108002. 445

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In Rizvi, M. H., Rustagi, J. S., and Siegmund, D. (eds.), *Recent Advances in Statistics*, 287–302. Academic Press. MR0736538. 443

Frühwirth-Schnatter, S. and Pyne, S. (2010). "Bayesian Inference for finite mixtures of univariate and multivariate skew normal and skew-t distributions." *Biostatistics*, 11: 317–36. 443, 448

Gorsky, S., Chan, C., and Ma, L. (2023). "Supplementary Material for "Coarsened Mixtures of Hierarchical Skew Normal Kernels for Flow and Mass Cytometry Analyses"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1356SUPP. 444, 445, 448, 450, 451, 452, 454, 455, 456

Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F., and Thiébaut, R. (2017). "Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data." Preprint. MR3937443. doi: https://doi.org/10.1214/18-AOAS1209. 443, 448

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 450

Kleinsteuber, K., Corleis, B., Rashidi, N., Nchinda, N., Lisanti, A., Cho, J. L., Medoff, B. D., Kwon, D., and Walker, B. D. (2016). "Standardization and quality control for high-dimensional mass cytometry studies of human samples." *Cytometry Part A*, 89(10): 903–913. 454, 455

Lee, S. X., McLachlan, G. J., and Pyne, S. (2015). "Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure." *Cytometry Part A*, 89(1): 30–43. 440

Levine, J., Simonds, E., Bendall, S., Davis, K., ad D. Amir, E., Tadmor, M., Litvin, O., Fienberg, H., Jager, A., Zunder, E., Finck, R., Gedman, A., Radtke, I., Downing, J., Pe'er, D., and Nolan, G. (2015). "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis." *Cell*, 162(1): 184–197. 443, 457

Liseo, B. and Parisi, A. (2013). "Bayesian Inference for the Multivariate Skew-normal Model: A Population Monte Carlo Approach." *Computational Statistics & Data Analysis*, 63: 125–138. MR3040255. doi: https://doi.org/10.1016/j.csda.2013.02.007. 445, 448, 449

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). "Automated gating of flow cytometry data via robust model-based clustering." *Cytometry Part A*, 73(4): 321–332. 443

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and Computing*, 26(1): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 450

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). "Identifying Mixtures of Mixtures Using Bayesian Estimation." *Journal of Computational and Graphical Statistics*, 26(2): 285–295. MR3640186. doi: https://doi.org/10.1080/10618600.2016.1200472. 443

Miller, J. W. and Dunson, D. B. (2018). "Robust Bayesian Inference via Coarsening." *Journal of the American Statistical Association*, 0(0): 1–13. MR4011766. doi: https://doi.org/10.1080/01621459.2018.1469995. 442, 447

Minoura, K., Abe, K., Maeda, Y., Nishikawa, H., and Shimamura, T. (2020). "CYBERTRACK2.0: zero-inflated model-based cell clustering and population tracking method for longitudinal mass cytometry data." *Bioinformatics*, 37(11): 1632–1634. 441

Munkres, J. (1957). "Algorithms for the Assignment and Transportation Problems." *Journal of the Society for Industrial and Applied Mathematics*, 5(1): 32–38. MR0093429. 451

Murphy, R. (1985). "Automated identification of subpopulations in flow cytometric list mode data using cluster analysis." *Cytometry*, 6: 302–9. 442

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., and Karlis, D. (2016). "Clustering with the multivariate normal inverse Gaussian distribution." *Computational Statistics & Data Analysis*, 93: 18–30. MR3406193. doi: https://doi.org/10.1016/j.csda.2014.09.006. 443

Parisi, A. and Liseo, B. (2018). "Objective Bayesian analysis for the multivariate skew-t model." *Statistical Methods & Applications*, 27(2): 277–295. MR3807370. doi: https://doi.org/10.1007/s10260-017-0404-0. 445

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L., Baecher-Allan, C., McLachlan, G., Tamayo, P., Hafler, D., De Jager, P., and Mesirov, J. (2010). "Automated High-Dimensional Flow Cytometric Data Analysis." In Berger, B. (ed.), *Research in Computational Molecular Biology*, 577–577. Berlin, Heidelberg: Springer Berlin Heidelberg. 443

Pyne, S., Lee, S. X., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, T., Ng, S.-K., Hafler, D., Levy, R., Nolan, G. P., Mesirov, J., and McLachlan, G. J. (2014). "Joint Modeling and Registration of Cell Populations in Cohorts of High-Dimensional Flow Cytometric Data." *PLOS ONE*, 9(7): 1–11. 440

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 639–650. MR1309433. 445

Soriano, J. and Ma, L. (2017). "Mixture modeling on related samples by ψ-stick breaking and kernel perturbation." *arXiv e-prints*, arXiv:1704.04839. 448

Soriano, J. and Ma, L. (2019). "Mixture Modeling on Related Samples by ψ -Stick Breaking and Kernel Perturbation." *Bayesian Anal.*, 14(1): 161–180. MR3910042. doi: `https://doi.org/10.1214/18-BA1106`. 442, 448

van der Maaten, L. and Hinton, G. (2008). "Visualizing Data using t-SNE." *Journal of Machine Learning Research*, 9(86): 2579–2605. 441

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data." *Cytometry Part A*, 87(7): 636–645. 443, 457

Wilkerson, M. D. and Hayes, D. N. (2010). "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." *Bioinformatics*, 26(12): 1572–1573. 458

Acknowledgments

We thank two anonymous referees and an editor for their thoughtful and constructive comments that helped us improve the paper.