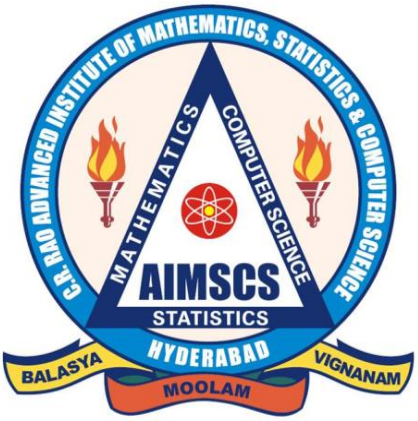
CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report



Author (s): **Sharon X. Lee, Geoffrey J. McLachlan & Saumyadipta Pyne**

Title of the Report: **Modelling of Inter-sample Variation in Flow Cytometric Data with the Joint Clustering and Matching (JCM) Procedure**

Research Report No.: **RR2015-06**

Date: **June 10, 2015**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Modelling of Inter-sample Variation in Flow Cytometric Data with the Joint Clustering and Matching (JCM) Procedure

Sharon X. Lee[1], Geoffrey J. McLachlan[1,⋆], Saumyadipta Pyne[2,3]

Draft: June 5, 2015

[1]Department of Mathematics, University of Queensland, St. Lucia, Queensland, 4072, Australia. [2]CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India. [3]Public Health Foundation of India.
⋆ E-mail: g.mclachlan@uq.edu.au

Abstract

We present an algorithm for modelling flow cytometry data in the presence of large inter-sample variation. Large-scale cytometry datasets often exhibit some within-class variation due to technical effects such as instrumental differences and variations in data acquisition, as well as subtle biological heterogeneity within the class of samples. Failure to account for such variations in the model may lead to inaccurate matching of populations across a batch of samples and poor performance in classification of unlabelled samples. In this paper, we describe the Joint Clustering and Matching (JCM) procedure for simultaneous segmentation and alignment of cell populations across multiple samples. Under the JCM framework, a multivariate mixture distribution is used to model the distribution of the expressions of a fixed set of markers for each cell in a sample such that the components in the mixture model may correspond to the various populations of cells, which have similar expressions of markers (that is, clusters), in the composition of the sample. For each class of samples, an overall class template is formed by the adoption of random-effects terms to model the inter-sample variation within a class. The construction of a parametric template for each class allows for direct quantification of the differences between the template and each sample, and also between each pair of samples, both within or between classes. The classification of a new unclassified sample is then undertaken by assigning the unclassified sample to the class that minimizes the distance between its fitted mixture density and each class density as provided by the class templates. We use a symmetric form of the Kullback-Leibler distance for this purpose. We show and demonstrate on four real datasets how the JCM procedure can be used to carry out the tasks of automated clustering and alignment of cell populations, and supervised classification of samples.

Key terms Flow cytometry, Classification, Class template, Inter-sample variation, Clustering, Matching, Skew mixture models, EM algorithm, JCM

1 Introduction

Flow cytometry (FCM) is a powerful tool in clinical diagnosis of health disorders, in particular, immunological diseases. It offers rapid high-throughput measurements of multiple characteristics on every cell in a sample, enabling biologists to study a variety of biological processes at

the cellular level. A critical component in the pipeline of flow cytometry data analysis is the identification of cell populations from the multidimensional FCM dataset, currently performed manually by visually separating regions or gates of interest on a series of sequential bivariate projections of the data, a process known as *gating*. However, this approach has many problems and limitations, including variability between different analysts, non-standardization and non-reproducibility of results, an unrealistic assumption that biological relationships between the markers exist only in the projected low-dimensional space and, most importantly, non-scalability to high-dimensional analysis, especially involving a large number of samples. With the advancement of technology, modern day flow cytometers allow simultaneous measurements of many markers on millions of cells, with some latest revolutionary mass cytometers capable of extending this up to 100 simultaneous parameters [1, 2]. As the number of markers increases, the number of bivariate projections increase rapidly. This renders conventional manual analysis practically infeasible for unbiased FCM analysis. Due to this and the subjective and time-consuming nature of this approach, recent efforts have turned to the development of computational methods for the analysis of high-dimensional flow cytometry data to automate the gating process; see [3] for a recent account.

Among these methods, mixture models have been widely employed as the underlying mechanism for characterizing the heterogeneous cell populations [4, 5, 6, 7, 8, 9, 10, 11], taking advantage of the convenient and formal framework offered by a model-based approach to modelling these complex and multimodal datasets. Using this approach, the FCM data can be conceptualized as a mixture of populations each of which consists of cells with similar expressions, the distribution of which can be characterized by a parametric density. The task of cell-population identification then translates directly to the classical problem of multivariate model-based clustering. It is well known that data generated from flow cytometric studies are often asymmetrically distributed, multimodal, as well as having longer and/or heavier tails than normal. To accommodate this, several methods use a mixture of mixtures approach [7, 12] where a final cluster may consist of more than one mixture component, while some others adopt mixture models with skew distributions as components [4, 5, 7] to enable a single component distribution to correspond to a cluster.

In addition to cell-population identification, there is also the challenging task of aligning or matching these populations across multiple samples. The classification of individual samples (for example, predicting the disease state of an unlabelled sample) presents another challenging task in the pipeline of flow cytometric data analysis. To this end, several techniques have been proposed recently in the literature, most of them based on extensions of algorithms for the cell-population identification mentioned above. In particular, many of these methods adopt a support vector machine (SVM) as their classifier, perhaps due to its simplicity, convenience, and computational efficiency. For example, the sample classifiers of flowBin [3], SWIFT [12], ASPIRE [8], and flowPeaks [13] are all based on a SVM, trained on some selected features from their respective model fitted to the data. Typically, the cluster proportions (size) are used as the feature vector under these settings. An allocation of a sample to one of the predefined classes of samples is then performed on the basis of this feature vector using a classifier formed from the available training data consisting of observations on the feature vector of known classification with respect to the classes. Sometimes different types of classifiers are used besides the SVM; for example, nearest neighbour in PBSC [3], and regularized regression in Citrus [14]. However, in all of these methods, only selected features are used in the post-hoc classification task. Hence their classifiers are based solely on the information in these features, ignoring a variety of other features of the underlying distributions such as their shapes or tails that may represent biologically interesting phenomena.

Here we present the JCM (Joint Clustering and Matching) procedure for the supervised classification of FCM samples with respect to a number of predefined classes. Unlike the above mentioned methods, JCM adopts a measure to quantify the similarity or dissimilarity between the fitted parametric models in terms of differences between the class templates representing the class densities of the markers being used. Based on the calculation of distance between a pair of samples modelled by JCM as mixture distributions, their high-dimensional, and potentially multi-modal, forms can be compared with precision and rigour. Indeed, the matrix of such pairwise distances for a given class of samples can be used to quantify the overall inter-sample variation, and we describe tools for visualizing the same. The distance-based approach provides a more complete assessment of the differences between the samples and the templates for different classes and so should lead to more objective discrimination among the classes.

In constructing a template for the density of the markers for a given class of samples, JCM models the inter-sample variation in a class through the adoption of a random-effects model. With the exception of ASPIRE, the methods described above do not allow for possible inter-sample variations. They either explicitly or implicitly assume the cell populations to have the same underlying distribution across all samples. In particular, the location (or mode) and the shape of the distribution of these populations are taken to be the same. These assumptions are not realistic given that the cell populations typically vary significantly across different samples; see, for example, the Lymphoma dataset described in Section 2.1.2. These assumptions are relaxed within the JCM framework by modelling each sample as an instance of a class template, possibly transformed with a flexible amount of variation. The latter is governed by a random-effects model, thus allowing one to establish a direct parametric correspondence between the cell populations in a sample and their corresponding components in the mixture model for the class template. This formulation also means that the cell populations are automatically matched across different samples without needing any further post-processing.

Previously in [7], the effectiveness of JCM in clustering cells within a given sample and aligning the cell populations across a batch of samples was illustrated in three experiments, where JCM provided excellent results. The first two of these experiments involved multiple samples and (multiplexed) staining panels, as well as multiple time-points and/or classes. In one of these two experiments, the B-cell receptor (BCR) one, which is also examined here (but reformulated as a classification problem), JCM identified a spatio-temporal signature of BCR signaling that improved the distinction between the two classes of patients previously reported in [15]. In the third illustration, it was demonstrated how JCM was able to identify cell populations in a batch of Diffuse Large B-Cell Lymphoma (DLBCL) samples from the FlowCAP-I challenge [3] that closely matched the gated populations identified by expert analysts. In this paper, we focus on the application of the JCM procedure in the subsequent stage of the flow cytometry pipeline – sample classification – and the modelling of the inter-sample variation within a batch of samples.

2 Materials and Methods

2.1 Overview of the Datasets

We shall demonstrate the effectiveness of JCM in automated gating and supervised classification of flow cytometry samples using four real benchmark datasets described here.

2.1.1 West Nile Virus (WNV) dataset

Thirteen blood samples were acquired from patients diagnosed with symptomatic West Nile Virus (WNV). These samples were stained to measure the expressions of CD4, IL17, and CFSE markers. Between 100,000 and 1,000,000 events were recorded for each sample. The samples were manually analyzed and gated. To demonstrate the ability of JCM in clustering samples to identify cell populations, we measure the level of agreement between the automated approach of JCM with manual gating in terms of the misclassification rate (MCR). Results are to be compared also with competing algorithms as mentioned above. For comparison purposes, we assume the partitions given by manual gating to be the ground truth.

2.1.2 Lymphoma dataset

Between 2003 and 2008, the British Columbia Cancer Agency collected samples from randomly selected lymph node biopsies from patients with diffuse large B-cell lymphoma (DLBCL). Each sample was stained for measuring the expressions of three surface markers, CD3, CD5, and CD19. This dataset contains 60 samples, with each sample containing between 3,000 and 100,000 events. This dataset is known to exhibit considerable inter-sample variation due to a voltage change in the instrument settings in 2005. We use this dataset to illustrate the usefulness of random effects modelling in handling datasets with large inter-sample variation.

2.1.3 B-cell receptor (BCR) dataset

The B-cell receptor (BCR) dataset contains flow cytometric measurements from 28 patients diagnosed with follicular lymphoma (FL). In brief, each sample was split into eight multiplexed panels for staining by F(ab')2 against the BCR heavy chain. For each panel, a pair of lineage markers (common to all panels) was used to label the B-cells, while another pair (different in all panels) of phospho-markers was used to measure BCR signalling characteristics, totalling 18 different markers across all panels. Measurements were recorded for all panels at basal (unstimulated) and again at 4 minutes after stimulation. Further technical details on this experiment and the processing of samples can be found in the Supplementary Methods of [15]. It was observed in [16] that the FL patients can be stratified into two classes that have distinctly different survival outcomes, which is linked to the presence or absence of a subpopulation of B-cells known as the Lymphoma Negative Prognostic (LNP) cells. For our illustration, the dataset is randomly partitioned into a training and test set with equal number of samples in each set. The task for the automated algorithms is to determine the class of FL for each patient, based on the training samples provided.

2.1.4 Acute Myeloid Leukemia (AML) dataset

As part of the FlowCAP-II Sample Classification challenge [3], the Acute Myeloid Leukemia (AML) dataset was split equally into a training set and a test set. Peripheral blood or bone marrow samples from a total of 359 patients were collected; 316 were healthy patients and 43 were diagnosed with AML [3]. Due to the large number of markers used, the sample collected from each patient was assayed into eight tubes for staining by different markers combinations, with five markers from each tube. In addition, the forward scatter (FS) and side scatter (SS) were also available for each tube, yielding a seven-dimensional dataset. This totalled to 2872 data files for analysis. The first and last tube (i.e., tubes 1 and 8) were controls and hence not used in our analysis. The data have been transformed logarithmically for the SS measurements and all fluorescent markers, while the FS measurements remained linear. All channels have also

been scaled to the unit interval during preprocessing. The dataset is available publicly from `http://flowrepository.org/id/FR-FCM-ZZYA`. For training, half of the dataset along with their known labels (156 healthy patients and 23 AML patients) was used. The challenge called for the construction of a classifier to label the samples in the test set (the other half consisting of data from 180 patients).

2.2 The JCM algorithm

The JCM methodology is based on a multi-level model-based clustering approach, where each sample is represented by a flexible finite mixture model. The latter is intrinsically linked to a class template through a random-effects model (REM). In brief, the JCM approach can be conceptualized as a two-level hierarchical model consisting of a lower sample-specific level and a higher batch-specific level. At the sample-specific level, each cell population in a sample is characterized by a parametric multivariate distribution. A sample with multiple populations can thus be viewed as having a mixture distribution. Note that each sample has its own mixture model, implying that all the component parameters are specific to that sample. It should be pointed out that this is significantly different to some of the other available approaches, such as SWIFT and HDPGMM, that require all samples to share the same component parameters (except the mixing proportions). Our approach gives JCM more flexibility in handling inter-sample variations. At the batch-specific level, an entire batch of samples can be modelled by a parametric multivariate template that describes the overall characteristics of the batch. To construct a representative template, JCM assumes that each sample-specific mixture model can be effectively modelled as an instance of a batch-specific (class) template with subtle inter-sample variations. This template is taken to be representative of the class from which the batch of samples is assumed to have been drawn. The class template is constructed by adopting a random-effects approach where individual mixture models are linked to the batch mixture models through a flexible affine transformation. A further advantage of this approach is that each cluster in a sample is automatically registered with respect to the corresponding cluster of the template. A third level can be envisaged when between-class comparisons in the case of multiple classes are made in a large cohort analysis. This additional higher level can be used to study intra-class relationships in situations where, for example, multiple disease types are present, individuals are measured at different time points, or multiple experiment conditions are analyzed. With the availability of an overall template for each class, classification of unlabelled samples can be easily performed by comparing its similarity with each class template. We describe here further technical details of each level.

2.2.1 Modelling of individual samples

As mentioned above, each sample is modelled by a finite mixture model. The JCM procedure provides two options for parametric densities, the default option using the multivariate skew t (MST) distribution and an option using its symmetric version, the multivariate t-distribution. The MST distribution has additional parameters for handling heavy tails and skewness, rendering it well suited for modelling non-elliptical and asymmetrical clusters that are typical in flow cytometry data. Moreover, it formally encompasses the multivariate normal, t, and skew normal distributions as special or limiting cases.

To establish notation, let the p-dimensional vector $\boldsymbol{y}_{jk} \in \mathbb{R}^p$ contain the measurements (on p markers) of cell j in sample k, where $j = 1, \ldots, n_k$ and $k = 1, \ldots, K$. Here n_k denotes the total number of cells in sample k, and K denotes the total number of samples. We let g denote the number of components in the mixture model. It is assumed that each cell in the jth sample

belongs to one of the g components. Then the density of $\boldsymbol{y}_{jk}$ can be expressed as

$$f(\boldsymbol{y}_{jk};\boldsymbol{\Psi}_k)=\sum_{h=1}^{g}\pi_{hk}f(\boldsymbol{y}_{jk};\boldsymbol{\theta}_{hk}), \tag{1}$$

where $\pi_{1k},\ldots,\pi_{gk}$ denote the mixing proportions which are non-negative and sum to one. Here, $f(\boldsymbol{y}_{jk};\boldsymbol{\theta}_{hk})$ denotes a component density with parameters specified by $\boldsymbol{\theta}_{hk}\,(h=1,\ldots,g)$. The vector $\boldsymbol{\Psi}_k$ consists of all the unknown parameters of the mixture model, and is given by $\boldsymbol{\Psi}_k=(\pi_{1k},\ldots,\pi_{g-1,k},\boldsymbol{\theta}_{1k}^T,\ldots,\boldsymbol{\theta}_{gk}^T)^T$, where the superscript T denotes vector or matrix transpose. In practice, the optimal value of g can either be specified *a priori*, or inferred from the data using some information criterion such the Bayesian Information Criterion (BIC).

Regarding the choice of the component densities in (1), JCM has options: (i) the multivariate t-distribution and (ii) a multivariate skew t-distribution. In the former case, the t-distribution allows for heavier tails than the normal distribution, thus providing a more robust approach to traditional Gaussian mixture modelling (GMM). The t-component density is given by

$$t_p(\boldsymbol{y}_{jk};\boldsymbol{\mu}_{jhk},\boldsymbol{\Sigma}_{hk},\nu_{hk})=\frac{\Gamma(\frac{\nu_{hk}+p}{2})}{(\pi\nu_{hk})^{\frac{p}{2}}|\boldsymbol{\Sigma}_{hk}|^{\frac{1}{2}}\Gamma(\frac{\nu_{hk}}{2})}\left(1+\frac{d_h(\boldsymbol{y}_{jk})}{\nu_{hk}}\right)^{-\frac{\nu_{hk}+p}{2}}, \tag{2}$$

where $\boldsymbol{\mu}_{jhk}$ is a (cell-specific) location parameter, $\boldsymbol{\Sigma}_{hk}$ is a positive-definite scale matrix, ν_{hk} is the degrees of freedom, $d(\boldsymbol{y}_{jk})=(\boldsymbol{y}_{jk}-\boldsymbol{\mu}_{jhk})^T\boldsymbol{\Sigma}_{hk}^{-1}(\boldsymbol{y}_{jk}-\boldsymbol{\mu}_{jhk})$ denotes the squared Mahalanobis distance between $\boldsymbol{y}_{jk}$ and $\boldsymbol{\mu}_{jhk}$ (with $\boldsymbol{\Sigma}_{hk}$ as the scale matrix), and $\Gamma(\cdot)$ the Gamma function. The degrees of freedom ν_{hk} acts as a tuning parameter for regulating the thickness of the tails of the t-distribution, which can be inferred from the data. The second option for $f(\cdot)$ provided by JCM is a skew version of (2), with an additional vector $\boldsymbol{\delta}_{hk}$ consisting of p skewness parameters for handling non-symmetric distributional shapes. In recent years, many different versions of the skew t-distribution have been proposed; see [17] for an overview on this topic. For our purposes, we shall adopt the popular version as proposed by [18]. Following the notation in [4], the skew t-density can be expressed as

$$\begin{aligned}&ST_p(\boldsymbol{y}_{jk};\boldsymbol{\mu}_{jhk},\boldsymbol{\Sigma}_{hk},\boldsymbol{\delta}_{hk},\nu_{hk})\\&\quad=2t_p(\boldsymbol{y}_{jk};\boldsymbol{\mu}_{jhk},\boldsymbol{\Sigma}_{hk},\nu_{hk})\\&\qquad T_1\left(\boldsymbol{\delta}_{hk}^T\boldsymbol{\Omega}_{hk}^{-1}(\boldsymbol{y}_{jk}-\boldsymbol{\mu}_{hk})\sqrt{\frac{(\nu_{hk}+p)}{\nu_{hk}+d_h(\boldsymbol{y}_{jk})}};0,1-\boldsymbol{\delta}_{hk}^T\boldsymbol{\Omega}_{hk}^{-1}\boldsymbol{\delta}_{hk},\nu_{hk}+p\right),\end{aligned} \tag{3}$$

where T_1 denotes the (scalar) t-distribution function, and $\boldsymbol{\Omega}_{hk}=\boldsymbol{\Sigma}_{hk}+\boldsymbol{\delta}_{hk}\boldsymbol{\delta}_{hk}^T$. The estimation of the parameters of a mixture model can be carried out using the expectation-maximization (EM) algorithm. Specific details for finite mixtures of (2) can be found in [19], and in [4] for finite mixtures of (3).

2.2.2 Parametric batch-specific (class) template

To form a batch-specific (class) template, inter-sample variations are modelled through the introduction of random-effects (RE) terms that specify how sample-specific component distributions may vary from an overall representative mixture model (that is, the template model). More specifically, these RE terms govern how the component-location parameters $\boldsymbol{\mu}_{hk}$ relate to the batch location parameter $\boldsymbol{\mu}_h$. We proceed by assuming that each $\boldsymbol{\mu}_{hk}$ is an affine transformation of $\boldsymbol{\mu}_h$, that is,

$$\boldsymbol{\mu}_{hk}=\boldsymbol{a}_{hk}\circ\boldsymbol{\mu}_h+b_{hk}\mathbf{1}_p, \tag{4}$$

where $\circ$ denotes the Hadamard (or elementwise) product, $\mathbf{1}_p$ denotes a p-dimensional vector with all elements being one, and the scaling and translation RE terms are independent and given by

$$\begin{aligned} \boldsymbol{a}_{hk} &\sim N_p(\mathbf{1}_p, \boldsymbol{\xi}_{1h}), \\ b_{hk} &\sim N_1(0, \xi_{2h}^2), \end{aligned} \tag{5}$$

respectively. Estimation of the parameters of the mixture model with (4) and (2) or (3) can be implemented using the EM algorithm. The technical details of the fitting algorithm for our JCM model have been described in the Supplementary Information of [7].

2.2.3 Classification of new samples

In the case of multiple classes, a template can be formed for each class with JCM, using the available training data (that is, the samples of known origin with respect to the classes). As can be observed from (1) and (4), a JCM template is characterized parametrically as a mixture model similar to the sample-specific model fitted to each sample. This facilitates the quantitative comparison between different templates. Once the class templates are constructed, they can be quantitatively compared using a range of information-based measures, such as the Kullback-Leibler (KL) distance. In this approach, a new sample is fitted with a mixture model and its KL divergence from each of the templates is calculated. The new sample is then classified to the class associated with the smallest KL distance. We use a symmetric combination of the KL information.

More specifically, the KL distance between two continuous densities $f_1(\boldsymbol{y})$ and $f_2(\boldsymbol{y})$ is defined by

$$KL(f_1, f_2) = \int_{-\boldsymbol{\infty}}^{\boldsymbol{\infty}} f_1(\boldsymbol{y}) \log \frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})} d\boldsymbol{y}. \tag{6}$$

As the KL distance is not a symmetric measure, that is $KL(f_1, f_2) \neq KL(f_2, f_1)$, we use the mean of $KL(f_1, f_2)$ and $KL(f_2, f_1)$ as our distance measure. When a new or unlabelled sample is presented to JCM, a mixture model is fitted to obtain a parametric representation of the sample.

Unlike ASPIRE and Citrus that use only certain features calculated from the predicted clusters for each sample, the classification approach of JCM is based on the fitted densities for the class templates rather than their selected features. More specifically, these other methods calculate features such as the proportions and the median of some or all of the markers for the identified clusters as a simplified representation of each sample. A classifier is then built on these features using a support vector machine (SVM) (as in ASPIRE) or based on regularized regression (as in Citrus). As such, the accuracy of these classifiers can be quite sensitive to the choice of features, and is critically dependent on the distinctiveness of the selected features across the different classes. In contrast, JCM does not rely on any particular selected features from the fitted model, but rather uses the entire fit provided by the model by proceeding on the basis of the relative size of the class densities provided by the templates. When an unlabelled sample is presented to JCM, an independently fitted model of the sample is directly compared to the estimate of the parametric form of the template density for each class. The KL distance provides a measure of this difference between the estimated density of the unclassified sample and the fitted template density for each class.

3 Results

The effectiveness of the JCM approach is illustrated on four separate datasets and compared to a number of other available methods. In the first two datasets, we are interested in the classification of flow cytometry samples into two classes. The first dataset was part of the FlowCAP-II competition, while the second has been previously analyzed in [7, 15]. For the final two datasets, we examine the performance of JCM and other competing algorithms in identifying cell populations in the presence of within-class inter-sample variations. In the WNV dataset, only the abundance of a particular population was observed to be varying greatly across the samples; whereas for the Lymphoma dataset, the variations were much more profound.

3.1 Performance evaluation measures

To evaluate the classification performance of JCM, we compute a number of measures as adopted in [3] namely sensitivity (or recall), specificity, accuracy, precision, and F-measure. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. These measures are defined as

$$\begin{aligned} sensitivity &= \frac{TP}{TP+FN}, && (7)\\ specificity &= \frac{TN}{TN+FP}, && (8)\\ accuracy &= \frac{TP+TN}{TP+TN+FP+FN}, && (9)\\ precision &= \frac{TP}{TP+FP}, && (10)\\ F\text{-measure} &= 2\times\frac{precision\times sensitivity}{precision+sensitivity}. && (11) \end{aligned}$$

In addition, we report in the second last column of Tables 3 and 2 the adjusted Rand index (ARI) [20], a popular measure of cluster agreement in the model-based clustering literature. The performance of the JCM model is also evaluated by the area under receiver operating characteristic (ROC) curve (AUC), reported in the last column of Table 3.

For the West Nile Virus dataset, the clustering performance of JCM is assessed using the misclassification rate (MCR), the standard measure used in statistics to evaluate the accuracy of clustering and/or classification algorithms against true labels. This MCR is based on the proportion of mislablled observations, calculated for each permutation of the predicted cluster labels against the labels given by manual gating and the rate reported is the minimum value over all such permutations.

3.2 Other methods

The performance of the JCM procedure is compared with some other available methods for automated clustering and sample classification, including HDPGMM [6], flowMatch[21], Citrus[14], and ASPIRE[8]. Like FLAME, JCM is based on finite mixtures of skew distributions. However, FLAME performs cluster alignment across samples in a post-hoc fashion using graph-matching techniques. In contrast, the multi-level modelling strategy of JCM allows for simultaneous clustering and matching cell populations across samples in a much more natural manner. Moreover, the availability of a parametric characterization of the class templates allows for the classification of unlabelled samples to be undertaken on the basis of differences between their densities

and the class densities provided by the templates. Following FLAME, Azad et al. [21] proposed flowMatch for classifying samples based on templates. The flowMatch procedure treats each sample as a leaf node of a template tree, then performs agglomerative hierarchical clustering of the nodes to obtain a template tree, using a similarity measure based on the KL distance. However, with flowMatch, each node in the template tree is characterized by a single normal mixture model. Also, cluster matching is performed between two nodes before calculating their similarity measure. In contrast, JCM automatically matches the clusters across the samples and/or templates in the fitting process. The KL distance can be calculated directly on the fitted models under the JCM framework.

Recently, Cron et al. [6] proposed a hierarchical version (HDPGMM) of the Dirichlet process Gaussian mixture model (DPGMM) for the automatic alignment of cell populations across a batch of samples. Their model has an additional layer above the DPGMM models fitted to each sample, placing a hierarchical prior over the base distribution to link these individual DPGMMs. The HDPGMM model assumes that all samples share the same component parameters with the exception of the mixing proportions. In a more recent contribution, Dundar et al. [8] extended the HDPGMM model to allow for inter-sample random-effects using a much similar conceptual strategy to JCM. This model, known as ASPIRE, relaxes the requirement for component means to be the same across samples by assuming that they deviate probabilistically from the corresponding template means under a random-effects model. Using this approach, a global template is readily constructed without the need for post-hoc mode clustering as in HDPGMM.

Citrus [14] is another computational tool for sample classification and cell-population identification. Unlike the above mentioned methods, Citrus proceeds by a hierarchical clustering of cells selected from each sample, and the classification of samples is based on regression methods on selected features. For scalability, Citrus applies hierarchical clustering to aggregated data consisting of randomly sampled cells from each sample rather than the data pooled from all samples. This approach removes the need for clustering individual samples, and automatically identifies local and global clusters, as well as matching them across different samples. From the hierarchical clustering of the aggregated data, the proportion of each sample in each cluster is calculated. For each sample, a feature vector is formed consisting of the cluster proportions along with other features of the sample such as the median value of each marker.

3.3 Experiment 1: Automated gating of WNV samples

We analyzed the 13 samples in the WNV dataset as a batch with JCM and assess its ability to reidentify the manually gated populations in each sample. With this dataset, experts identified four cell populations in manual analysis. Each sample was examined individually, and the populations were matched across the samples in a subsequent step. Four examples are shown in Figure 1, where the different populations as identified by manual gating are displayed in different colours. As can be observed in Figure 1(A), while the relative locations of these populations were quite similar across the samples, the abundance of the CD4+CFSE- populations varies considerably between samples, ranging from 0% (sample012) to 76% (sample013); see Figure 1(B). With the exception of sample013, the other three populations have very similar abundance across the samples. JCM and five competing algorithms (ASPIRE, HDPGMM, flowMatch, flowPeaks, and SWIFT) were applied to this dataset such that their templates have four global clusters. The template obtained by JCM as shown in Figure 1(C) suggests that the location of the four populations are well recovered. Inspection of the models fitted to each sample by JCM (not shown) reveals that they are very similar to the template, but subtle variations can be

observed in the mixing proportions. Indeed, the JCM model provides a quantitative measure of the inter-sample variations in the batch as part of its model fitting procedure. As described in Section 2.2.2, these are given in terms of the variance of the RE terms, namely $\boldsymbol{\xi}_{1h}$ and ξ_{2h} $(h = 1, \ldots, 4)$. The smaller the values of these parameters, the closer the location of the clusters in each sample is relative to that of the template. The values of $\boldsymbol{\xi}_{1h}$ are estimated to be

$$\boldsymbol{\xi}_{11} = \begin{bmatrix} 0.095 & 0.000 & 0.000 \\ 0.000 & 0.079 & 0.000 \\ 0.000 & 0.000 & 0.084 \end{bmatrix}, \boldsymbol{\xi}_{12} = \begin{bmatrix} 0.057 & 0.000 & 0.000 \\ 0.000 & 0.061 & 0.000 \\ 0.000 & 0.000 & 0.001 \end{bmatrix},$$
$$\boldsymbol{\xi}_{13} = \begin{bmatrix} 0.004 & 0.000 & 0.000 \\ 0.000 & 0.034 & 0.000 \\ 0.000 & 0.000 & 0.015 \end{bmatrix}, \boldsymbol{\xi}_{14} = \begin{bmatrix} 0.004 & 0.000 & 0.000 \\ 0.000 & 0.011 & 0.000 \\ 0.000 & 0.000 & 0.000 \end{bmatrix},$$

respectively, for the CD4+CFSE-, CD4-CFSE-, CD4-CFSE+, and CD4+CFSE+ populations in this dataset. Similarly, the value of ξ_{2h} are all small (around 0.1), giving an indication that the location of the clusters remains fairly constant across the samples.

On comparing the clustering results of these methods against manual analysis (Table 1), it can be observed that JCM had the lowest average MCR of 0.1243. In 8 of the 13 samples, JCM achieved the lowest MCR compared the other five algorithms. Note that as the location of populations remains similar across the sample, algorithms that employ pooling (such as flowPeaks and SWIFT) should not be disadvantaged. Indeed, algorithms such as HDPGMM are designed to adapt to this situation. This is reflected in the average MCR of these algorithms, where HDPGMM, flowPeaks, and SWIFT were quite similar and were reasonably close to the performance of JCM. The small difference in the performance of these algorithm may possibly be attributed to the slight variations in the shape of the CD4+CFSE- population across the samples. Looking at the case of sample012, the absence of the CD4+CFSE- population did not seem to have a great impact on the clustering performance of JCM, where it achieved a low MCR of 0.0742, being just under half the MCR of the next lowest MCR of 0.1815 (obtained by flowPeaks). With a relatively high average MCR above 0.2, the ASPIRE and flowMatch algorithms have poor performance compared to the other algorithms in this illustration.

3.4 Experiment 2: Identification and matching of cell populations in Lymphoma samples

To illustrate the ability of JCM to model and match clusters across samples in the presence of large inter-sample variations, we consider 30 random samples from the Lymphoma dataset. It is known that a subset of these samples was influenced by the voltage change in 2005, causing a shift in the sub-populations for the CD3 and CD5 channels. Examples from the the dataset are shown in the first column of Figure 2, where noticeable differences can be observed across the samples for the location of the two clusters. In addition, the distribution of the clusters varies considerably across the samples, and some clusters exhibit non-normal features. Of interest here is how well the automated algorithms can identify and discriminate the CD3+CD5+ and CD3-CD5- populations in each sample and correctly match them across the samples.

We applied five algorithms (JCM, flowMatch, flowPeaks, SWIFT, and HDPGMM) to this dataset and compared their predicted clustering. The algorithms JCM, SWIFT, and HDPMM identified two global clusters, whereas flowPeaks and flowMatch identified three clusters. A visual inspection of the results given by flowPeaks (column 4 of Figure 2) indicates that the pink cluster consist mainly of cells that expressed no or very low levels of CD5. If they can

be viewed as outlying observations of the CD3-CD5- cluster of cells, then flowPeaks can be considered to have only two global clusters. Regarding the matching of populations across samples, JCM and HDPGMM automatically register cell populations across samples. The flowMatch procedure perform population matching in a post-hoc step. The flowPeaks and SWIFT procedures employ data pooling and hence no matching is required.

Focusing first on samples with well-separated clusters, such as sample008 (row 2) and sample009 (row 3) of Figure 2, most of the algorithms can discriminate between the two clusters reasonably well. For sample008, it can be observed that with the exception of SWIFT and flowMatch, the other three algorithms can identify the CD3+CD5+ population (displayed as green dots in Figure 2). For sample009, flowPeaks and HDPGMM mislabeled a minor portion of CD3+CD5+ cells. In both these samples, flowMatch consistently mislabelled the tail portion of the CD3-CD5- population. Looking at a case where the two populations are close together (sample023, row 5, of Figure 2), flowPeaks and HDPGMM failed to separate these populations and labelled them as one cluster. SWIFT provided a partition into two clusters, but have clearly mislabelled a large portion of the CD3-CD5- population residing in upper tail region. The flowMatch algorithm modelled the CD3-CD5- populations with with two components. On merging these components to give two meta clusters, flowMatch provides a reasonable clustering of the data. There is, however, still a small portion of CD3-CD5- cells incorrectly labelled as CD3+CD5+. These are located near the tip of the tail of the CD3-CD5- population. As can be observed in Figure 2, JCM can capture both populations accurately and adapt more closely to the asymmetric shape of these clusters.

In the case where the CD3+CD5+ population is highly abundant compared to the CD3-CD5- population (sample013, row 4 of Figure 2), flowMatch can identify the abundant population as one cluster, but splits the less abundant population into two clusters. Similar results can be observed in cases where the CD3-CD5- population is relatively more abundant (sample001 and sample027 of Figure 2). In sample001, however, cells in the tail region of the CD3-CD5- population were incorrectly labelled by flowMatch. The flowPeaks and HDPGMM algorithms performed reasonably well in this sample, but are disappointing for sample013 and sample027. The poor performance in the latter two samples can be attributed to the large shift of the CD3+CD5+ population. In contrast, JCM was able to model the distribution of both populations with good precision in all three cases.

It is of interest to note that as flowPeaks and SWIFT perform clustering on the pooled data, all samples share the same classification boundaries. In effect, this is similar to drawing static gates on all samples, where the gates are based on the clustering of the pooled data. The HDPGMM model is designed to cater for some inter-sample variations. However, these are restricted to the mixing proportions only. Hence, the classification boundary remains the same for all samples. As can be observed in Figure 2, the static gates given by flowPeaks and HDPGMM failed to capture the CD3+CD5+ population in sample013, sample023, and sample027 due to the drastic shifts in the CD5 channel. For SWIFT, it resulted in the mislabelling of a large portion of the CD3-CD5- population in all samples.

Sample027 shows an example of a case where one of the population exhibit evident skewness. In this case, flowPeaks, SWIFT, and HDPGMM have difficulty identifying cells belonging to the CD3-CD5- population, especially those in the tail of the distribution. As mentioned above, with flowMatch, this population is partitioned into two components. However, even after merging, it cannot accurately separate the two populations around the tail region. The pink component contains both the CD3-CD5- cells that lies in the tail and also some cells from the more dispersed CD3+CD5+ population. In contrast, JCM provided a much more appropriate solution.

The results suggests that all algorithms but JCM failed to recover accurately the distribution of two populations in this dataset. Algorithms that rely on data pooling for handling a batch of samples (for example, flowPeaks and SWIFT) suffer significantly from the large variations between samples. These algorithms can be considered as applying a static gate to all samples. As a result, when the CD3+CD5+ population is located close to the CD3-CD5- population, SWIFT failed to identify a large portion of CD3-CD5- cells whereas flowPeaks can identify only a very small portion of CD3+CD5+ cells. While HDPGMM operates on individual samples and allows for some inter-sample variations, it failed to distinguish between the two populations in these cases due to its assumption that local clusters shares common location and scale parameters across all samples. Hence, this algorithm is effectively assuming static gates. The flowMatch algorithm also operates on each sample individually, but matching is performed in a post-hoc manner similar to FLAME. The allows for greater flexibility in terms of the location and shape of the local clusters, but the matching of local clusters show some limitations. To consider this further here, we examined the initial partitions given by the local clusters of flowMatch. It revealed that all samples have four local clusters. The extra cluster lies between the green and pink clusters and was merged with the green cluster on matching. Owing to the large inter-sample variation, this extra cluster consists of CD3-CD5- cells in some samples (for example, in the first three examples shown in Figure 2), and CD3+CD5+ cells in the other samples (due to the shift of this population in these samples). However, as this cluster have similar location across the sample and are closer to the green cluster, they were incorporated into the CD3+CD5+ meta-cluster during the matching step.

It is of interest to note that for these four algorithms, initial partitions of the data is based on normal mixture models and the final partitions are obtained by merging some of the components. Although this allows for some degree of flexibility in capturing non-normal clusters, it has significant limitations in the presence of random effects; see sample027 as an example. Unlike these algorithms, the skew t-mixture model adopted by JCM can naturally adapt to the asymmetric shape of the clusters and, by incorporating random-effects terms in the model specification, JCM can accommodate these samples in the presence of large inter-sample variations.

3.5 Experiment 3: Classification of FL patients

In [15], the FL patients were manually analyzed and classified into LNP^- and LNP^+ classes. Here, we applied seven algorithms (JCM, ASPIRE, Citrus, flowMatch, flowPeaks, HDPGMM, and SWIFT) to this dataset. For training, half of the data was provided to each algorithm, together with their class labels. The BCR dataset presents a more challenging case for the algorithms, as the distinction between the classes is noticeably less pronounced than the AML dataset. In particular, it can be observed from Figure 3 that the distribution of the markers on the cells appears to be similar between an LNP^- (Figure 3A) and LNP^+ (Figure 3B) sample. Perhaps the greatest difference occurs in the plot of the cells for the markers p-STAT5 and p.PLCg2 in Figures 3A3 and 3B3, where the shape of the distribution of the LNP^- samples is asymmetrical compared to the LNP^+ sample.

The procedures JCM, ASPIRE, flowMatch, flowPeaks, HDPGMM, and SWIFT identified 5, 3, 4, 22, 16, and 1 clusters, respectively, with their templates. Their performance measures are reported in Table 2. It can be observed that JCM achieved the highest F-measure value of 0.71. Citrus and HDPGMM produced results that are quite similar to each other, yielding the same specificity and AUC. However, HDPGMM had a lower sensitivity, accuracy, and precision. With a zero sensitivity and precision, the performance of flowMatch, ASPIRE, flowPeaks, and

SWIFT is relatively poor for this dataset. A closer look at the predicted classification labels given by ASPIRE, flowPeaks, and SWIFT reveals that they failed to discriminate between the two classes of patients, classifying all samples into one class.

3.6 Experiment 4: Classification of AML patients

We report in Table 3 the performance measures of JCM and other algorithms on the AML dataset. Since HDPGMM, flowPeaks, and SWIFT do not provide a strategy for sample classification, we follow a similar approach to [8]. For HDPGMM and flowPeaks, we extract the proportions of global clusters from each sample to form a feature vector to be used for training a SVM based on the labels from the training set. For SWIFT, a template or consensus model was computed based on the pooled samples. Subsequently, the cluster size for each sample were utilized as features for training a SVM. With ASPIRE, we adopted the traditional fully supervised classification setting as above, training a SVM on both the AML and normal cases in the training set.

Among these algorithms, JCM achieved the highest AUC value. Also, JCM along with Citrus achieved the highest F-measure value and ARI for this dataset. It can be observed from Table 3 that JCM and Citrus produced very similar results for the remaining performance measures. Indeed, both Citrus and JCM correctly predicted 179 of the 180 test-labels with Citrus incorrectly predicting an AML patient to be normal, whereas JCM mislabelled a normal patient to be AML. As ASPIRE did not produce any results (terminated itself with unknown reasons before completing the fitting procedure), we do not report the results for ASPIRE in Table 3. However, it is noted in Table 3 in [8] that ASPIRE achieved an overall AUC of 98.9, although using a different partition into training and test sets.

The remaining two methods, HDPGMM and SWIFT, achieved perfect specificity, but performed poorly in terms of sensitivity. In other words, they have correctly identified all normal patients, but mislabelled a number of AML patients as being normal. This is also reflected in their poor F-measure, ARI, and AUC values.

Inter-sample variation within a given class of samples can be studied in various ways. Such insights can lead to understanding of subtle subclass structures and allow for better classification. We provide a visual tool in the form of two-dimensional plots (see Figures 4 and 5) using Multi-dimensional Scaling based on the symmetrized KL distances between the JCM fitted multivariate mixture distributions for each pair of samples. For example, for the WNV dataset (Figure 5A), all samples (with the exception of the outlying sample) as modelled by JCM were very similar except for the mixing proportions, whereas much larger variation can be observed between samples in the CFSE dataset (Figure 5B).

4 Modifications for a large number of markers

The largest number p of markers considered simultaneously in the two examples above is seven. Our experience with the fitting of mixtures of restricted skew t (rMST) distributions to multivariate data consisting of p variables has shown that this mixture model is feasible in practice for p as large as at least 20. It would suggest that the JCM procedure is therefore also feasible for 20 or so markers, although it does require more fitting time than with just the fitting of mixtures of rMST distributions since it also has additional random-effects terms in the model to construct the class template.

One way to reduce the dimensionality when there is a very large number p of markers is to perform a principal component analysis (PCA) and work with the first q principal components,

where q is chosen to be appropriately small.

Another way of proceeding in the case of very large p is to use so-called matrix factorization. Let

$$\boldsymbol{X} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T \tag{12}$$

be the $n \times p$ data matrix. Then $\boldsymbol{X}$ can be expressed via matrix factorization as

$$\boldsymbol{X} = \boldsymbol{X}_1 \boldsymbol{X}_2, \tag{13}$$

where $\boldsymbol{X}_1$ is a $n \times q$ matrix and $\boldsymbol{X}_2$ is a $q \times p$ matrix and where q is chosen to be much smaller than p. For a specified value of q, the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are chosen to minimize

$$\|\boldsymbol{X} - \boldsymbol{X}_1 \boldsymbol{X}_2\|^2, \tag{14}$$

where $\|\cdot\|$ is the Frobenius norm (the sum of squared elements of the matrix). With this factorization, dimension reduction is effected by replacing the data matrix $\boldsymbol{X}$ by the solution $\hat{\boldsymbol{X}}_1$ for the factor matrix $\boldsymbol{X}_1$; the vth column of $\hat{\boldsymbol{X}}_1$ gives the values of the vth metavariable for the n observations in the sample. Thus the original p variables are replaced by q metavariables. When the elements of $\boldsymbol{X}$ are non-negative, we can restrict the elements of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to be non-negative. This approach is called non-negative matrix factorization (NMF) in the literature ([22]; [23]). We shall call the general approach where there are no constraints on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, general matrix factorization (GMF). Nikulin and McLachlan [24] have developed a very fast approach to the general matrix factorization (14), using a gradient-based algorithm that is applicable to an arbitrary (differentiable) loss function.

The dimension-reduction procedures PCA, GMF, or NMF can be undertaken either before or after the deletion from each sample of cells with outlying expression values for some of the markers. These outliers can be deleted first by implementing JCM in the first instance for one or more subsets of the markers of manageable dimension. Also, one might wish to use just the front and side scatter measurements to identify outliers.

In applications of normal or t-mixture models to high-dimensional datasets outside of flow cytometry such as in the clustering of microarray gene-expression data, mixtures of factor analyzers (MFA), or mixtures of common factor analysers (MCFA) have been applied; see, for example, Baek et al. [25]. With this factor-analytic approach, the correlations between the variables in the component-covariance matrices are explained by the variables depending linearly on a small number q of unobservable (latent) factors.

5 Discussion

We have considered the JCM procedure for the clustering of cells within a sample and the unsupervised and supervised classification of multiple samples of cells for multidimensional flow cytometric datasets. JCM addresses these two problems with a single multi-level framework. Firstly, JCM can perform cell-population identification and alignment across multiple samples in a fully automated manner. Secondly, the template approach of JCM provides a mathematically convenient way to classify new samples. For the former task, the effectiveness of JCM has been illustrated in [7] on three experiments, giving promising results. The focus of this paper is on the modelling of inter-sample variations and the latter task of sample classification.

For the problem of supervised classification of a new, unlabelled sample to one of a number of predefined classes, a template for each class is constructed by JCM. Unlike FLAME, the class template built by JCM is characterized parametrically. This facilitates the use of divergence measures such as the KL distance for quantitative comparison of two parametric distributions.

The use of the KL distance in classifying samples thus uses a different approach compared to methods such as HDPGMM, Citrus, and ASPIRE in which the classifier is built based on a limited number of features two derived from the fitted models. In contrast, JCM is based on the fitted densities for the class templates and the sample to be classified and not just a few features of these estimated densities such as the estimated mixing proportions and component/cluster means or medians of some of the markers.

The JCM procedure was compared to existing benchmark methods on four real FCM datasets. On the WNV and CFSE datasets where small and large inter-sample variations were observed, respectively, JCM was able to segment and model the populations identify by manual analysis with higher precision. In the AML classification challenge, JCM achieved an almost perfect AUC, the highest among the five methods (Citrus, flowMatch, SWIFT, and HDPGMM) considered. The JCM procedure was also able to correctly identify all AML samples, achieving a sensitivity of one, while the sensitivity of none of the other four methods exceeded 0.95. On the BCR dataset, JCM performed best on the basis of both the F-measure and AUC.

On the other hand, JCM's focus on modeling the overarching class structure could make it less suitable for certain datasets for which the classification is effectively driven by a specific feature or dimension of the samples rather than their overall high-dimensional form in terms of the mixture distribution. Comparison of classification results in Table 4 for two particular FlowCAP-II datasets illustrates this point.

6 Acknowledgement

The work of SXL and GJM is supported by a grant from the Australian Research Council. SP is supported by a Ramalingaswami Fellowship of DBT and grants from DRDE and MoS&PI, GOI.

7 References

[1] Tanner SD, Bandura DR, Ornatsky O, Baranov VI, Nitz M, Winnik MA. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. Pure Appl Chem 2008. 80:2627–2641.

[2] Bendall SC, Simonds EF, Qiu P, Amir EaD, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011. 332:687 – 696.

[3] Aghaeepour N, Finak G, The FLOWCAP Consortium, The DREAM Consortium, Hoos H, Mosmann T, Gottardo R, Brinkman RR, Scheuermann RH. Critical assessment of automated flow cytometry analysis techniques. Nature Methods 2013. 10:228–238.

[4] Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirow JP. Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences USA 2009. 106:8519–8524.

[5] Frühwirth-Schnatter S, Pyne S. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics 2010. 11:317–336.

[6] Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh SK, Britten CM, Welters MJ, van der Burg SH, West M, Chan C. Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. PLoS Computational Biology 2013. 9.

[7] Pyne S, Lee S, Wang K, Irish J, Tamayo P, Nazaire MD, Duong T, Ng SK, Hafler D, Levy R, Nolan GP, Mesirov J, McLachlan GJ. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. PLOS ONE 2014. 9(7):e100334.

[8] Dundar M, Akova F, Yerebakan HZ, Rajwa B. A non-parametric bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. BMC Bioinformatics 2014. 15(314).

[9] Rossin E, Lin TI, Ho HJ, Mentzer S, Pyne S. A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues. Bioinformatics 2011. 27:2746–2753.

[10] Ho HJ, Lin TI, Chang HH, Haase HB, Huang S, Pyne S. Parametric modeling of cellular state transitions as measured with flow cytometry different tissues. BMC Bioinformatics 2012. 13:(Suppl 5): S5.

[11] Ho HJ, Pyne S, Lin TI. Maximum likelihood inference for mixtures of skew student-t-normal distributions through practical EM-type algorithms. Statistics and Computing 2012. 22:287–299.

[12] Naim I, Datta S, Sharma G, Cavenaugh JS, Mosmann TR. Swift: Scalable weighted iterative sampling for flow cytometry clustering. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2010, pp. 509–512.

[13] Ge Y, Sealfon SC. flowpeaks: a fast unsupervised clustering for flow cytometry data via k-means and density peak finding. Bioinformatics 2012. 28:2052–2058.

[14] Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences 2014. 111(26):E2770–E2777.

[15] Irish JM, Myklebust JH, Alizadeh AA, Houot R, Sharman JP, Czerwinski DK, Nolan GP, Levy R. B-cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression. Proc Natl Acad Sci U S A 2010. 107:12747–12754.

[16] Irish JM, Kotecha N, Nolan GP. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. Nat Rev Cancer 2006. 6:146–155.

[17] Lee SX, McLachlan GJ. On mixtures of skew-normal and skew t-distributions. Advances in Data Analysis and Classification 2013. 7:241–266.

[18] Azzalini A, Capitanio A. Distribution generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. Journal of the Royal Statistical Society Series B 2003. 65(2):367–389.

[19] McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley Series in Probability and Statistics, 2000.

[20] Hubert L, Arabie P. Comparing partitions. Journal of Classification 1985. 2:193–218.

[21] Azad A, Rajwa B, Pothen A. Immunophenotypes of acute myeloid leukemia from flow cytometry data using templates. arXiv:1403.6358 [q-bio.QM] 2014.

[22] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature 1999. 401:788–791.

[23] Donoho D, Stodden V. When does non-negative matrix factorization give correct decomposition into parts? In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2004, vol. 16.

[24] Nikulin V, McLachlan G. On a general method for matrix factorisation applied to supervised classification. In: et al C, editor, Proceedings of 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, Washington, D.C. Los Alamitos, California: IEEE Computer Society, 2009, pp. 43–48.

[25] Baek J, McLachlan GJ. Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics 2011. 27:1269–1276.

Figures

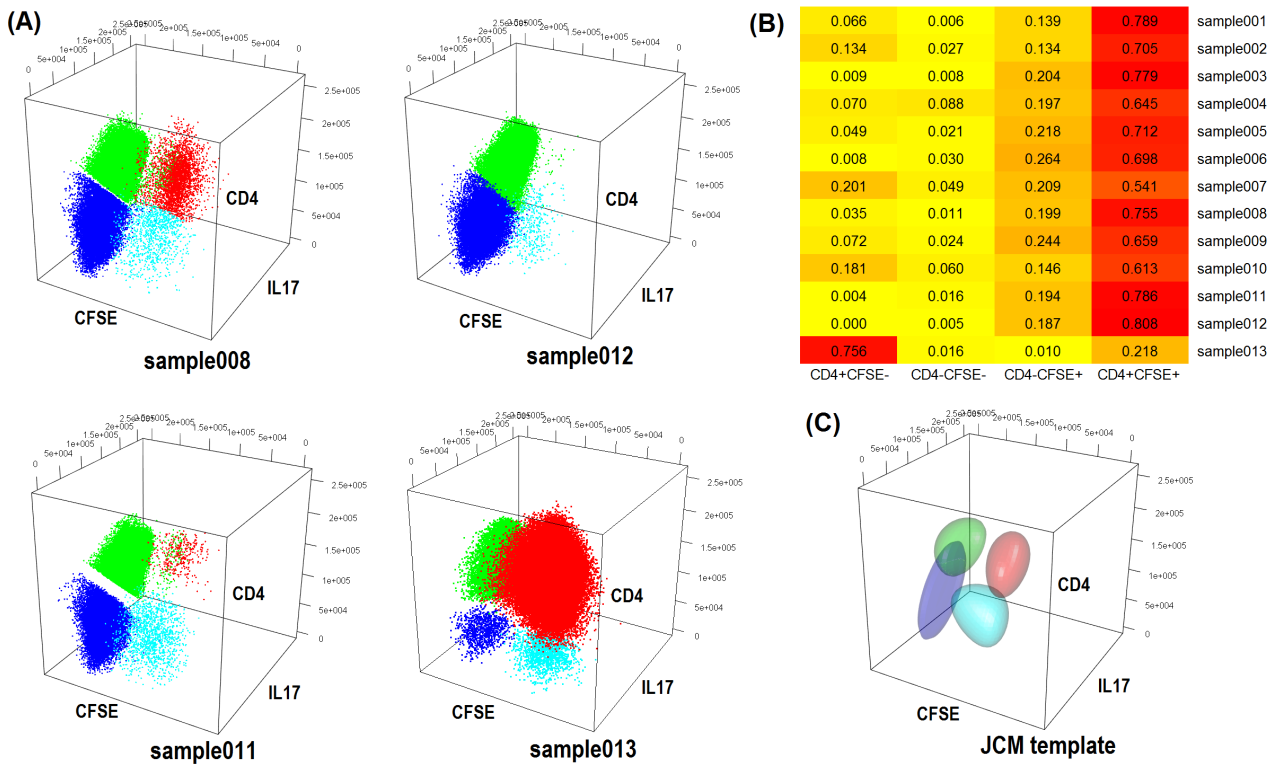


Figure 1: Comparison of JCM and manual analysis of samples from the WNV dataset. The four major populations identified manually (A) have similar locations across the samples, but the abundance of the CD4+CFSE- population varies greatly between samples. The proportions of the other three populations remain relatively similar as revealed by a heatmap of the proportions (B). The template by JCM (C) captures the all four populations with a parametric model.

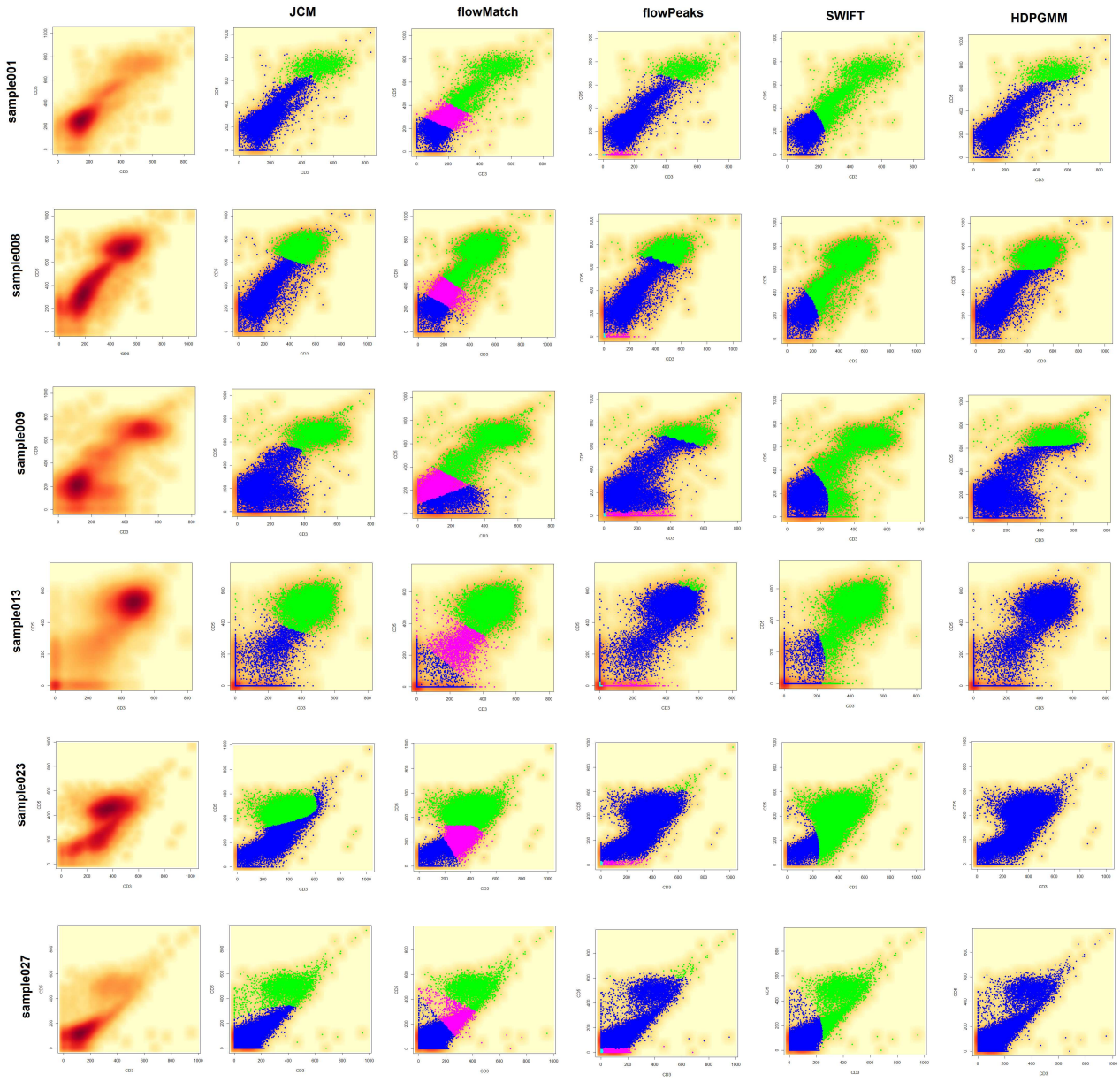


Figure 2: Clustering and matching of cell populations across DLBCL samples. The density plot of the raw data of selected samples are shown in the first column. The automated gating results of JCM, flowMatch, flowPeaks, SWIFT, and HDPGMM for each of these samples are given in columns two through five, respectively. As can be observed, with the exception of flowMatch and flowPeaks, two global clusters were identified by JCM, SWIFT, and HDPGMM.

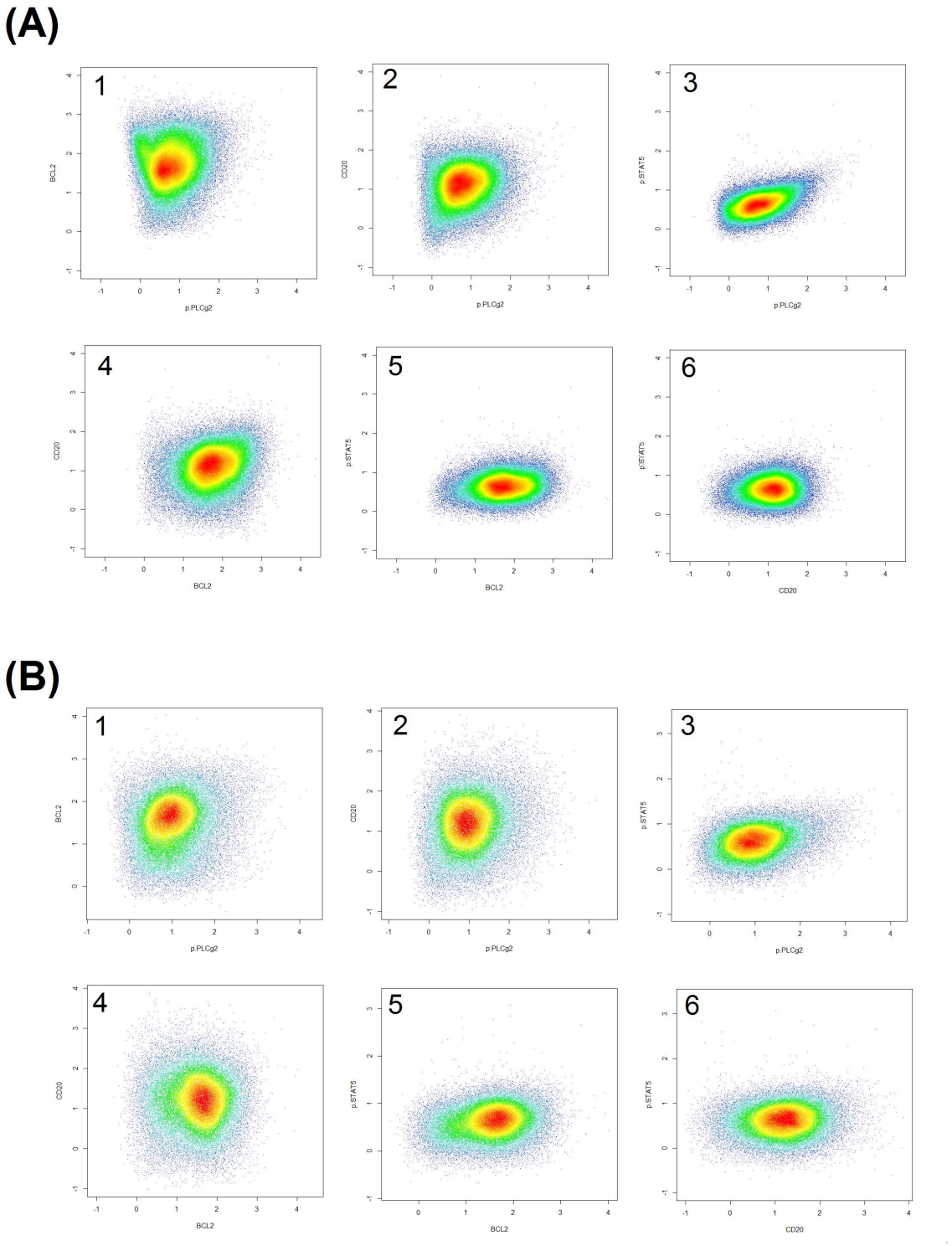


Figure 3: 2D scatter plots of markers from panel 4 for a sample from each of the LNP^- (A) and a LNP^+ (B) in the BCR dataset.

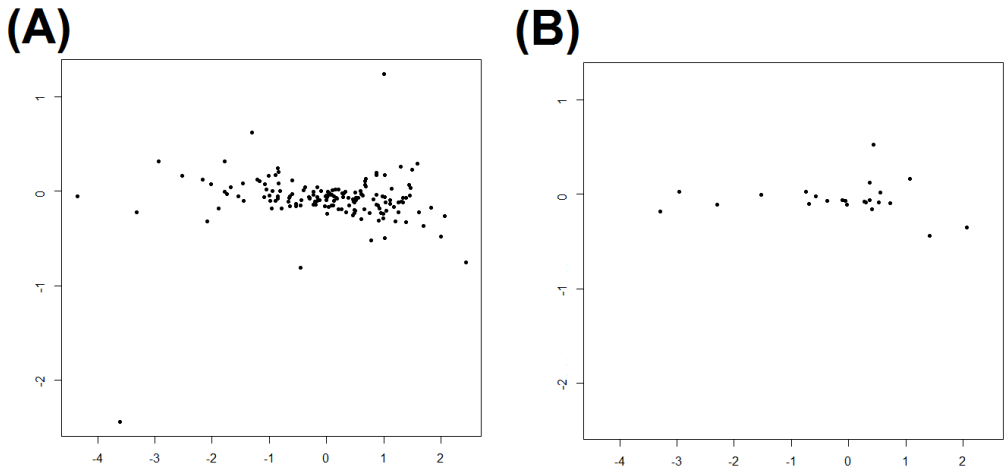


Figure 4: Inter-sample variation within (A) the normal and (B) the AML classes can be visualized in two-dimensions using Multi-dimensional Scaling (MDS) based on the symmetric version of KL distances between the JCM fitted distributions of each pair of samples.

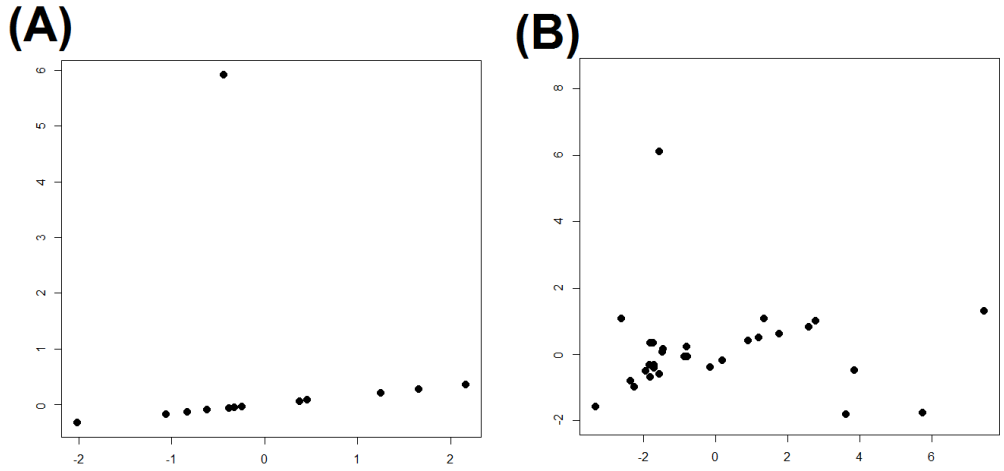


Figure 5: Visualization of inter-sample variation within the WNV and CFSE datasets through MDS reveals that (A) all but one sample (sample013) in the WNV dataset were quite similar and (B) the inter-sample variation is more profound in the CFSE dataset.

Tables

Table 1: **Misclassification rate (MCR) of automated gating by various algorithms against manual analysis on the CFSE dataset**

Sample	JCM	ASPIRE	HDPGMM	flowMatch	flowPeaks	SWIFT
001	0.0485	0.1880	0.1477	0.2767	0.1494	0.2459
002	0.1007	0.2026	0.1341	0.2651	0.1717	0.2300
003	0.0350	0.2568	0.2143	0.3391	0.1522	0.2172
004	0.2071	0.2945	0.2698	0.0095	0.2302	0.2121
005	0.0390	0.2806	0.1979	0.2349	0.1575	0.1913
006	0.2874	0.3116	0.2249	0.2122	0.1740	0.1242
007	0.2797	0.2794	0.2252	0.0071	0.2068	0.1656
008	0.0421	0.2535	0.2146	0.2750	0.1738	0.2121
009	0.1132	0.3196	0.2273	0.2792	0.1951	0.2258
010	0.0918	0.2373	0.2124	0.0456	0.1734	0.2338
011	0.0649	0.2344	0.1482	0.3083	0.1317	0.1612
012	0.0742	0.2076	0.1858	0.3141	0.1815	0.2556
013	0.2326	0.0437	0.0272	0.2316	0.1475	0.0935
AMCR	0.1243	0.2392	0.1869	0.2153	0.1727	0.1976

Table 2: **Classification results by each algorithm on the BCR dataset**

	Sensitivity	Specificity	Accuracy	Precision	F-measure	AUC
JCM	0.86	0.43	0.64	0.60	0.71	0.69
Citrus	0.29	0.71	0.50	0.41	0.36	0.59
HDPGMM	0.14	0.71	0.43	0.24	0.20	0.59
flowMatch	0.00	0.91	0.71	0.00	0.00	0.52
ASPIRE	0.00	1.00	0.50	0.00	0.00	0.50
flowPeaks	0.00	1.00	0.50	0.00	0.00	0.50
SWIFT	0.00	1.00	0.50	0.00	0.00	0.50

Table 3: **Classification results by each algorithm on the AML dataset**

	Sensitivity	Specificity	Accuracy	Precision	F-measure	ARI	AUC
JCM	1.00	0.99	0.99	0.95	0.97	0.97	0.997
Citrus	0.95	1.00	0.99	1.00	0.97	0.97	0.975
flowMatch	0.85	1.00	0.98	1.00	0.92	0.89	0.925
SWIFT	0.65	1.00	0.96	1.00	0.79	0.73	0.825
flowPeaks	0.45	1.00	0.94	0.62	0.62	0.55	0.725
HDPGMM	0.45	1.00	0.94	0.62	0.62	0.55	0.725

Table 4: **Comparative analysis of multiple methods on flowCAP-II datasets**

	HEU vs UE			HVTN		
	Accuracy	Precision	F-measure	Accuracy	Precision	F-measure
JCM	0.49	0.49	0.44	0.76	1.00	0.69
flowCore-flowStats	0.55	0.455	0.50	1.00	1.00	1.00
flowType-FeaLect	0.55	0.545	0.55	1.00	1.00	1.00
SWIFT	0.64	0.545	0.60	1.00	1.00	1.00
PBSC	0.273	0.36	0.30	0.95	0.95	0.95
PramSpheres	0.364	0.36	0.36	0.90	0.90	0.90