THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #237

Automated Gating of Flow Cytometry Data via Robust Model-based Clustering

Kenneth Lo, Ryan Remy Brinkman and Raphael Gottardo

December 2007

Automated Gating of Flow Cytometry Data via Robust Model-based Clustering[#]

Kenneth Lo[*a], Ryan Remy Brinkman[b] and Raphael Gottardo[a]

[a] Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, V6T 1Z2 Canada

[b] Terry Fox Laboratory, BC Cancer Research Center, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3 Canada

Email: Kenneth Lo[*] - c.lo@stat.ubc.ca; Ryan Remy Brinkman - rbrinkman@bccrc.ca; Raphael Gottardo - raph@stat.ubc.ca;

[*] Corresponding author. Tel.: 604 822 6911; Fax: 604 822 6960.

[#] This research was partially funded by the National Sciences and Engineering Research Council of Canada.

Abstract

Background: The capability of flow cytometry to offer rapid quantification of multidimensional characteristics for millions of cells has made this technology indispensable for health research, medical diagnosis, and treatment. However, the lack of statistical and bioinformatics tools to parallel recent high-throughput technological advancements has hindered this technology from reaching its full potential.

Methods: We propose a flexible statistical model-based clustering approach for identifying cell populations in flow cytometry data based on *t* mixture models with a Box-Cox transformation. This approach generalizes the popular Gaussian mixture models to account for outliers and allow for non-elliptical clusters. We describe an Expectation-Maximization (EM) algorithm to simultaneously handle parameter estimation and transformation selection.

Results: Using two publicly available datasets, we demonstrate that our proposed methodology provides enough flexibility and robustness to mimic manual gating results performed by an expert researcher. In addition, we present results from a simulation study, which show that this new clustering framework gives better results in terms of robustness to model mis-specification and estimation of the number of clusters, compared to the popular mixture models.

Conclusions: The proposed clustering methodology is well-adapted to automated analysis of flow cytometry data. It tends to give more reproducible results, and helps reduce the significant subjectivity and human time cost encountered in manual gating analysis.

Key terms: Box-Cox transformation; EM algorithm; mixture model; outliers; statistics; *t*-distribution; flow cytometry; gating; clustering

Introduction

Flow cytometry (FCM) can be applied to analyze thousands of samples per day. However, as each dataset typically consists of multiparametric descriptions of millions of individual cells, data analysis can present a significant challenge. As a result, despite its widespread use, FCM has not reached its full potential due to the lack of an automated analysis platform to parallel the high throughput data generation platform. As noted in a recent Communication to the Editor [1], in contrast to the tremendous interest in the FCM technology, there is a dearth of statistical and bioinformatics tools to manage, analyze, present, and disseminate FCM data. There is considerable demand for the development of appropriate software tools, as manual analysis of individual samples is error-prone, non-reproducible, non-standardized, not open to re-evaluation, and requires an inordinate amount of time, making it a limiting aspect of the technology [2–10].

One major component of FCM analysis involves gating, the process of identifying homogeneous groups of cells that display a particular function. This identification of cell populations currently relies on using software to apply a series of manually drawn gates (i.e., data filters) that select regions in 2-D graphical representations of the data. This process is based largely on intuition rather than standardized statistical inference [3, 11, 12]. It also ignores the high-dimensionality of FCM data, which may convey information that cannot be displayed in 1 or 2-D projections. This is illustrated in Supplementary Figure 1 with a synthetic dataset, consisting of two dimensions, generated from a *t* mixture model [13] with three components. While the three clusters can be identified using both dimensions, the structure is hardly recognized when the dataset is projected on either dimension. Such an example illustrates the potential loss of information if we disregard the multivariate nature of the data. The same problem occurs when projecting three (or more) dimensional data onto two dimensions.

Several attempts have been made to automate the gating process. Among those, the K-means algorithm [14] has found the most applications [15–18]. Demers et al. [17] have proposed an extension of K-means allowing for non-spherical clusters, but this algorithm has been shown to lead to performance inferior to fuzzy K-means clustering [18]. In fuzzy K-means [19], each cell can belong to several clusters with different association degrees, rather than belonging completely to only one cluster. Even though fuzzy K-means takes into consideration some form of classification uncertainty, it is a heuristic-based algorithm and lacks a formal statistical foundation. Other popular choices include hierarchical clustering algorithms (e.g., linkage or Pearson coefficients method). However,

these algorithms are not appropriate for FCM data since the size of the pairwise distance matrix increases in the order of n^2 with the number of cells, unless they are applied to some preliminary partition of the data [16], or they are used to cluster across samples, each of which is represented by a few statistics aggregating measurements of individual cells [20, 21]. Classification and regression trees (CART) [22], artificial neural networks (ANN) [23] and support vector machines (SVM) [24, 25] have also been used in the context of FCM analyses [26–29] but these supervised approaches require training data, which are not always available.

In statistics, the problem of finding homogeneous groups of observations is referred to as clustering. An increasingly popular choice is model-based clustering [13, 30–33] which has been shown to give good results in many applied fields involving high dimensions (greater than ten); see, for example, [33–35]. In this paper, we propose to apply an unsupervised model-based clustering approach to identify cell populations in FCM analysis. In contrast to previous unsupervised methods [6–8, 15–18], our approach provides a formal unified statistical framework to answer central questions such as: How many populations are there? Should we transform the data? What model should we use? How should we deal with outliers (aberrant observations)? These questions are fundamental to FCM analysis where one does not usually know the number of populations, and where outliers are frequent. By performing clustering using all variables consisting of fluorescent markers, the full multidimensionality of the data is exploited, leading to more accurate and more reproducible identification of cell populations.

The most commonly used model-based clustering approach is based on finite Gaussian mixture models [13, 31–33]. However, Gaussian mixture models rely heavily on the assumption that each component follows a Gaussian distribution, which is often unrealistic. A common approach is to look for transformations of the data that make the normality assumption more realistic. Box and Cox [36] discussed the power transformation in the context of linear regression, which has also been applied to Gaussian mixture models [37, 38]; see also [39] for a variant of Box-Cox transformation for FCM data. In addition to non-normality, there is also the problem of outlier identification in mixture modeling. Outliers can have a significant effect on the resulting clustering. For example, they will usually lead to overestimating the number of components to provide a good representation of the data. If a more robust model is used, fewer clusters may suffice. Outliers can be handled in the model-based clustering framework, by either replacing the Gaussian distribution with a more robust one (e.g., t [13, 40]) or adding an extra component to model the outliers (e.g., uniform [30]).

Transformation selection can be heavily influenced by the presence of outliers [41, 42]. To handle the issues of transformation selection and outlier identification simultaneously, we have developed an automated clustering approach based on *t* mixture models with Box-Cox transformation. The *t* distribution is similar in shape to the Gaussian distribution with heavier tails and thus provides a robust alternative [43]. The Box-Cox transformation is a type of power transformation, which can bring skewed data back to symmetry, a property of both the Gaussian and *t* distributions. In particular, the Box-Cox transformation is effective for data where the dispersion increases with the magnitude, a scenario not uncommon to FCM data.

Materials and Methods

Data Description

To demonstrate our proposed automated clustering we use the two publicly [53] available FCM datasets.

The Rituximab Dataset

Flow cytometric high-content screening (FC-HCS) [54] was applied in a drug-screening project to identify agents that would enhance the anti-lymphoma activity of Rituximab, a therapeutic monoclonal antibody [55]. 1600 different compounds were distributed into duplicate 96-well plates and then incubated overnight with the Daudi lymphoma cell line. Rituximab was then added to one of the duplicate plates and both plates were incubated for several more hours. In addition to cells treated with the compound alone, other controls included untreated cells and cells treated with Rituximab alone. During the entire culture period, cells were incubated with the thymidine analogue BrdU to label newly synthesized DNA. Following culture, cells were stained with anti-BrdU and the DNA binding dye 7-AAD. The proportion of cells in various phases of the cell cycle and undergoing apoptosis was measured with multiparameter FACS analysis.

The GvHD Dataset

Graft-versus-Host Disease (GvHD) occurs in allogeneic hematopoietic stem cell transplant recipients when donor-immune cells in the graft initiate an attack on the skin, gut, liver, and other tissues of the recipient. It is one of the most significant clinical problems in the field of allogeneic blood and marrow transplantation. FCM was used to collect data on patients subjected to bone marrow transplant with a goal of identifying biomarkers to predict the development of GvHD. The GvHD dataset is a collection of weekly peripheral blood samples obtained from 31 patients following allogeneic blood and marrow transplant [56]. Peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Hypaque and then cryopreserved for subsequent batch analysis. At the time of analysis, cells were thawed and aliquoted into 96-well plates at 1×10^4 to 1×10^5 cells per well. The 96-well plates were then stained with 10 different 4-color antibody combinations. All staining and analysis procedures were miniaturized so that small number of cells could be stained in 96-well plates with optimally diluted fluorescently conjugated antibodies.

Gaussian Mixture Models

The conventional model-based clustering approach is based on finite Gaussian mixture models [13, 31–33], where each cluster can be described by a separate Gaussian distribution. Formally, given data $\mathbf{y}$, with independent p-dimensional multivariate observations $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$, the likelihood for a mixture model with G components is

$$L(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G; w_1 \ldots, w_G|\mathbf{y}) = \prod_{i=1}^{n} \sum_{g=1}^{G} w_g \Phi_p(\mathbf{y}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \qquad \sum_{g=1}^{G} w_g = 1, \tag{1}$$

where $\Phi_p(\cdot|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and w_g is the probability that an observation belongs to the g-th component. Estimates of the unknown parameters $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, ..., \boldsymbol{\Psi}_G)$ where $\boldsymbol{\Psi}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, w_g)$ can be obtained conveniently using the Expectation-Maximization (EM) algorithm [32, 44, 45].

In EM, we first define the unobserved cluster membership associated with each observation $\mathbf{y}_i$ as $\mathbf{z}_i = (z_{i1}, ..., z_{iG})$ with

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to cluster } g \\ 0 & \text{otherwise.} \end{cases}$$

The E-step of the EM algorithm requires computing $\tilde{z}_{ig} \equiv E_{\boldsymbol{\Psi}}(Z_{ig} \mid \mathbf{y}_i)$, which is interpreted as the posterior probability that $\mathbf{y}_i$ belongs to cluster g:

$$\tilde{z}_{ig} \leftarrow \frac{w_g \Phi_p(\mathbf{y}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{j=1}^{G} w_j \Phi_p(\mathbf{y}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{2}$$

The M-step is filled by the following closed-form expressions for the unknown parameters:

$$\hat{w}_g \leftarrow \frac{n_g}{n}; \quad \hat{\boldsymbol{\mu}}_g \leftarrow \frac{\sum_{i=1}^{n} \tilde{z}_{ig} \mathbf{y}_i}{n_g}; \quad \hat{\boldsymbol{\Sigma}}_g \leftarrow \frac{\sum_{i=1}^{n} \tilde{z}_{ig} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_g)^T}{n_g} \tag{3}$$

where $n_g \equiv \sum_i \tilde{z}_{ig}$. The EM algorithm alternates between the E and M steps until convergence. Observation $\mathbf{y}_i$ may then be assigned to cluster g associated with the largest $\tilde{z}_{ig}$ value, which corresponds to the maximum a posteriori (MAP) classification.

t Mixture Models

The Multivariate *t* Distribution

In the presence of outliers, Gaussian distributions might give poor representations of clusters due to the large influence of outliers. One strategy is to replace the Gaussian distribution with a t distribution, of which the heavier

tail provides a mechanism to handle outliers. The *t* mixture likelihood can be written as in (1) where the Gaussian density is replaced by the *t* density with mean $\boldsymbol{\mu}$ ($v > 1$), covariance matrix $v\,(v-2)^{-1}\sum$ ($v > 2$) and v degrees of freedom, given by

$$\varphi(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma(\frac{\nu+p}{2})|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})\{1+(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})/\nu\}^{\frac{\nu+p}{2}}}. \tag{4}$$

As in the Gaussian case, estimates of the unknown parameters $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, ..., \boldsymbol{\Psi}_G, v)$ where $\boldsymbol{\Psi}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, w_g)$ can be obtained using the EM algorithm [40, 46, 47]. The algorithm uses the fact that we can parameterize a *t* distribution using a normal-gamma compound distribution.

Maximum Likelihood Estimation for a *t* Mixture Model

In EM for *t* mixture models, we define the unobserved cluster membership $\mathbf{z}_i = (z_{i1}, ..., z_{iG})$ as in the Gaussian case. To facilitate the formulation of the *t* distribution, we also define the weights u_i's, coming from the normal-gamma compound parameterization, with

$$\mathbf{Y}_i|u_i, z_{ig} = 1 \sim \mathrm{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g/u_i) \tag{5}$$

independently for $i = 1, ..., n$, and $U_i \sim \mathrm{Ga}(v/2, v/2)$. The advantage of writing the model in this way is that, conditional upon the U_i's, the sampling errors are again normal but with different variances, and estimation becomes a weighted least squares problem. Now the E-step requires computing $\tilde{z}_{ig} \equiv E_{\boldsymbol{\Psi}}(Z_{ig} \mid \mathbf{y}_i)$ and $\tilde{u}_{ig} \equiv E_{\boldsymbol{\Psi}}(U_i \mid \mathbf{y}_i, z_{ig}=1)$:

$$\tilde{z}_{ig} \leftarrow \frac{w_g\varphi_p(\mathbf{y}_i|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu)}{\sum_{j=1}^{G} w_j\varphi_p(\mathbf{y}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\nu)} \tag{6}$$

and

$$\tilde{u}_{ig} \leftarrow \frac{\nu+p}{\nu+(\mathbf{y}_i-\boldsymbol{\mu}_g)^T{\boldsymbol{\Sigma}_g}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_g)}, \tag{7}$$

which lead to the following closed form estimates for the unknown parameters during the M-step:

$$\hat{w}_g \leftarrow \frac{n_g}{n}; \quad \hat{\boldsymbol{\mu}}_g \leftarrow \frac{\sum_{i=1}^{n}\tilde{z}_{ig}\tilde{u}_{ig}\mathbf{y}_i}{\sum_{i=1}^{n}\tilde{z}_{ig}\tilde{u}_{ig}}; \quad \hat{\boldsymbol{\Sigma}}_g \leftarrow \frac{\sum_{i=1}^{n}\tilde{z}_{ig}\tilde{u}_{ig}(\mathbf{y}_{ig}-\hat{\boldsymbol{\mu}}_g)(\mathbf{y}_{ig}-\hat{\boldsymbol{\mu}}_g)^T}{n_g} \tag{8}$$

where $n_g \equiv \sum_i \tilde{z}_{ig}$. The EM algorithm alternates between the E and M steps until convergence.

Note that the $\tilde{u}_{ig}$'s as given by (7) can be interpreted as weights. This quantity holds a negative relationship with the Mahalanobis distance $(\mathbf{y}_i - \boldsymbol{\mu}_g)^T\,\boldsymbol{\Sigma}_g^{-1}\,(\mathbf{y}_i - \boldsymbol{\mu}_g)$ between $\mathbf{y}_i$ and $\boldsymbol{\mu}_g$. Hence, a small value would suggest that the corresponding observation is an outlier. Here we call all cells with $\tilde{u}_{ig}$ values less than 0.5 outliers. At the end of

the EM algorithm, the $\tilde{u}_{ig}$'s can be used to visualize which observations have been down-weighted. Since the $\tilde{u}_{ig}$'s may take any positive values, such a feature would let an outlier place little influence upon the estimation of the parameters of a *t* mixture model. In contrast, in the absence of such mechanism a Gaussian mixture model is not robust against outliers, as the constraint $\sum_g \tilde{z}_{ig} = 1$ imposed upon the $\tilde{z}_{ig}$'s forces all observations to make equal contributions towards parameter estimation overall.

While it is possible to estimate the degrees of freedom parameter v for each component of the mixture as part of the EM algorithm [40], fixing it to a reasonable predetermined value for all components reduces the computational burden while still providing robust results. A reasonable value for v is four, which leads to a distribution similar to the Gaussian distribution with slightly fatter tails accounting for outliers.

Box-Cox Transformation

To handle transformation and outlier identification simultaneously, we propose a *t* mixture model with Box-Cox transformation. The Box-Cox transformation [36] of an observation y is defined as follows:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}, \tag{9}$$

where λ is referred to as the Box-Cox parameter. The function stated in (9) above is defined for positive values of y only. In view of the occasional need to handle negative-valued data in FCM analysis, here we adopt a modified version [48] of the Box-Cox transformation which is also defined for negative values:

$$y^{(\lambda)} = \frac{\mathrm{sgn}(y)|y|^{\lambda} - 1}{\lambda}, \qquad \lambda > 0. \tag{10}$$

Note that the allowable range of λ in (10) is changed to be strictly positive to avoid discontinuity across zero, which would occur if a negative value for λ was used to transform data. When all data values are positive, this modified Box-Cox transformation is the same as the original version. In general, for multivariate data, we may specify a Box-Cox parameter for each dimension. However, in the context of FCM data, since different variables used in each stage of our sequential clustering (see below) share similar characteristics, it is reasonable to set the Box-Cox parameter common to all variables. When we allowed for different Box-Cox parameters for different variables, we found that the Box-Cox parameter estimates are of similar magnitudes, justifying the use of one Box-Cox parameter for all variables in each stage (data not shown).

While the E step remains basically the same, as given by (2), replacing $\mathbf{y}_i$ with $\mathbf{y}_i^{(\lambda)}$, the incorporation of the Box-Cox parameter slightly complicates the M step. No closed-form solution is available for λ, which needs to be estimated by some numerical optimization technique. Please see Section 1 of Supplementary Material for a detailed account of EM for Gaussian or *t* mixture models with transformation selection.

In each case, the EM algorithm needs to be initialized. Here we have chosen to use the algorithm of Fraley [49] for initialization; see Section 2 of Supplementary Material for details.

Density Estimation

To visualize FCM data, it may be convenient to project high-dimensional data on 1-D or 2-D density plots. One such application can be found in the analysis of the GvHD data, in which cells selected through the $CD3^+$ gate were projected on the CD4 and CD8β dimensions to produce contour plots (see Figure 1 and Supplementary Figure 4). Usually, nonparametric methods are applied to produce such plots. However, all nonparametric methods require a tuning parameter (e.g., bandwidth for kernel density estimation [50]) to be specified to control the smoothness of these plots, and different softwares have different default settings. In the model-based clustering framework, such plots can be easily generated at a very low computational cost once estimates of the model parameters are available. The degree of smoothness is controlled by the number of components, which is chosen by the Bayesian Information Criterion (BIC) [51]. Please see Section 3 of Supplementary Material for more details.

Selecting the Number of Clusters

When the number of clusters is unknown, we use the BIC. For Gaussian mixture models, it is defined as

$$\mathrm{BIC}_G = 2\log \tilde{L}_G - K_G \log(n) \qquad (11)$$

where $\tilde{L}_G$ is the maximized likelihood value of (1) for a *G*-component Gaussian mixture model, *n* is the sample size, and K_G is the number of independent parameters. BIC would then be computed for a range of possible values for G and the one with the largest BIC (or relatively close to it) would be selected.

The BIC formula introduced in (11) can still be used for *t* mixture models even with Box-Cox transformation. Note that since we do not estimate the degrees of freedom parameter, a *t* mixture model has the same number of parameters, K_G, as a Gaussian mixture model. However, when the Box-Cox transformation is included in the model, we have one more parameter.

Sequential Approach to Clustering

In practice, gating is often done on a preselected subset of the data chosen by projecting the data on the forward light scatter (FSC) and sideward light scatter (SSC) dimensions. These two variables, which measure the relative morphological properties (corresponding roughly to cell size and shape) of the cells, are often used to distinguish basic cell types (e.g., monocytes and lymphocytes) and/or to remove dead cells and cell debris. As a consequence, similar to Hahne et al. [52], we have adopted a sequential approach to clustering. We first use the FSC and SSC variables to cluster the data and find basic cell populations, and then perform clustering on one or more populations of interest using all other variables consisting of fluorescent markers. However, our methodology could also be applied to any subset or the entire set of variables.

Results

Application to Real Datasets

The Rituximab Dataset

We have re-analyzed a part of the Rituximab dataset using our sequential clustering approach. This data contains 1545 cells and four variables: FSC, SSC and two fluorescent markers, namely, 7-AAD and anti-BrdU. We compared the different models described in the Materials and Methods section (*t* mixture with Box-Cox, *t* mixture, Gaussian mixture with Box-Cox, and Gaussian mixture) with the results obtained through expert manual analysis using the commercial gating software FlowJo (Tree star Inc., Ashland, Oregon) and the K-means clustering algorithm [14]. As mentioned in the Materials and Methods section, we use a sequential approach where we first cluster the FSC vs. SSC variables to select basic cell populations (first stage), and then cluster the selected population(s) using all remaining variables (second stage).

Figure 2(a) shows the initial gating performed by a researcher using FlowJo on the FSC and SSC variables. To facilitate the comparison of our clustering approach with manual analysis at the second stage, we tried to mimic this analysis. In order to do so, we used a *t* mixture model with Box-Cox transformation, fixing the number of components at one, and removed points with weights (as given by (7)) less than 0.5, corresponding to outliers (see Materials and Methods for details). As shown in Figure 2, the selected cells are not exactly the same but close enough to allow us to compare our clustering approach to manual gating results when using the two fluorescent markers.

At the second stage, we compare the different clustering models on the selected cells. Since the number of clusters is unknown in advance, we make use of the BIC. The BIC curves shown in Supplementary Figure 2, corresponding to the different models, peak around three to four clusters, motivating us to examine the results obtained using three (Figure 3) and four (Supplementary Figure 3) clusters respectively. As expected, K-means performs poorly as spherical clusters do not provide a good fit. Similarly, untransformed mixture models (*t* and Gaussian), constrained by the assumption of elliptical clusters, are not flexible enough to capture the top cluster. Furthermore, Gaussian mixture models (even with Box-Cox transformation) are very sensitive to outliers, which can result in poor classification. For example, when four clusters are used, the Gaussian mixture model breaks the larger cluster into two to accommodate outliers, while the Gaussian mixture model with Box-Cox transformation also has a

large spread out cluster to accommodate outliers. Finally, Figure 3(b) and Supplementary Figure 3(b) show that our t mixture model-based clustering approach with Box-Cox transformation can provide comparable results with the manual gating analysis by identifying three of the four clusters with well-fit boundaries. Note, however, that none of the four clustering methods detect the left rectangular gate seen on Figure 3(a), which is most likely due to its lower cell density compared to the other gates and the lack of clear separation along the "7 AAD" dimension. This gate, which corresponds to apoptotic cells [55], contains a loose assemblage of cells located at the left of the three far right gates. Our methodology permits the identification of the three right clusters with well-fit boundaries, and thus could be combined with expert knowledge in order to identify apoptotic cells. For example, one could compute a one-dimensional boundary at the left-end border of the two largest clusters, and automatically label cells on the left of that line apoptotic.

Having shown the superiority of our clustering framework in terms of flexibility and robustness compared to common approaches, we now turn to a larger dataset to demonstrate further its capability.

The GvHD Dataset

Two samples of the GvHD dataset [56] have been re-analyzed, one from a patient who eventually developed acute GvHD, and one from a control. Both datasets consist of more than 12,000 cells and four markers, namely, anti-CD4, anti-CD8β , anti-CD3 and anti-CD8, in addition to the FSC and SSC variables. One objective of the analysis is to look for the $CD3^{+}CD4^{+}CD8\beta^{+}$ cells [56]. To demonstrate the capability of our proposed automated clustering approach, we try to mimic the gating strategy stated in [56]. Figure 1(a-c) and Supplementary Figure 4(a-c) shows the gating performed by an expert researcher using FlowJo.

In the initial gating, we first extracted the lymphocyte population using the FSC and SSC variables by applying a t mixture model with Box-Cox transformation, fixing the number of clusters from one to eight in turn. Supplementary Figure 5(a) shows that the BIC for the positive sample has a large increase from three to four clusters and remains relatively constant afterwards, suggesting a model fit using four clusters is appropriate. Supplementary Figure 5(b) is the corresponding scatterplot showing the cluster assignment of the points on removing those with weights less than 0.5, regarded as outliers. It is clear that the region combining three of the clusters formed matches closely with the gate drawn by the researcher as shown in Figure 1(a), corresponding to the lymphocyte population.

The next two stages in the manual gating strategy consist of locating the $CD3^+$ cells by placing a range gate in the CD3 density plot (Figure 1(b)), and then identifying the $CD3^+CD4^+CD8\beta^+$ cells through the upper right gate in the CD4 vs CD8β contour plot (Figure 1(c)). When applying our proposed clustering approach, we can combine these two stages by handling all the variables consisting of fluorescent markers at once, fully utilizing the multidimensionality of FCM data.

The fitted model with 12 clusters seems to provide a good fit as suggested by the BIC (Figure 4(a)). We compared our results to those obtained through the manual gating approach by first examining the estimated density projected on the CD3 dimension. The unimodal, yet skewed, density curve suggests that it is composed of two populations with substantially different proportions superimposed on each other (Figure 1(e)). At a level of around 280, we can well separate the 12 cluster means along the CD3 dimension into two groups, and use the group with high cluster means in the CD3 dimension to represent the $CD3^+$ population. The unimodal nature of the density curve (Figure 1(b,e)) implies that the two underlying populations somewhat mix together, and therefore setting a fixed cutoff to classify the cells is likely inappropriate. The merit of our automated clustering approach is shown here, that, instead of setting a cutoff, it makes use of the information provided by the other dimensions to help classify the cells into $CD3^+/CD3^-$ populations. The group with high cluster means in the CD3 dimension consists of five clusters, and among these five clusters, we can easily identify the two clusters at the upper right in the CD4 vs CD8β scatterplot (Figure 4(b)) as the $CD3^+CD4^+CD8\beta^+$ population.

We have applied the same strategy to the control sample; see Supplementary Figure 4 and Figure 4(c-d). Figure 4(c) suggests that, this time, only seven clusters are necessary as the BIC is relatively flat after that. The associated gating results for the control sample is characterized by an absence of the $CD3^+CD4^+CD8\beta^+$ cells, a distinct difference from the positive sample. This feature is also captured using our automated clustering approach; the fitted model contains no clusters at the upper right of the $CD4^+$ vs $CD8\beta^+$ scatterplot (Figure 4(d)). This cell population was of specific interest as it was identified as one possibly predictive of GvHD, based on the manual gating analysis [56].

Simulation studies

We have conducted a series of simulations to study the performance of different model-based clustering approaches under different model specifications. Model performance is compared using the following two criteria: (a) the

accuracy in cluster assignment; (b) the accuracy in selecting the number of clusters. We performed two simulation studies, one where we set the dimension to two resembling the Rituximab dataset, and one where the dimension was set to four resembling the GvHD dataset. In each case, we generated data from each of the following models: *t* mixture with Box-Cox, *t* mixture, Gaussian mixture with Box-Cox, and Gaussian mixture, using the parameter estimates obtained at the second stage in the Rituximab and GvHD (positive sample) analyses. For the GvHD, to reduce computational burden, we only selected the five clusters with the largest means in the CD3 dimension, corresponding to the $CD3^+$ population. We refer to the simulation experiments as the Rituximab and the GvHD settings, respectively. We fixed the number of cells at 500 and generated 1000 datasets under each of the above models. To study the accuracy in selecting the number of clusters using BIC, we generated 100 datasets from the same GvHD setting with 1000 cells. Here, we used 1000 cells to avoid numerical problems with small clusters when the number of clusters used is significantly larger than the true number, while we decreased the number of datasets to 100 due to the increase in computation when estimating the number of clusters.

Classification results: The four clustering methods in comparison were applied to each of the 1000 datasets generated from each model. Model fitting was done by presuming that the number of clusters is known, i.e. four clusters for the Rituximab setting and five for GvHD. We compared the models via misclassification rates, i.e. the proportions of cells assigned to incorrect clusters. When computing the misclassification rates, all permutations of the cluster labels were considered, and the lowest misclassification rate was determined.

The scatterplot of one of the datasets (GvHD setting) generated from the *t* mixture model with Box-Cox transformation can be found in Supplementary Figure 7. Overall results are shown in Table 1. As expected, the Gaussian mixture models perform poorly when data were generated from the *t* mixture models due to a lack of mechanisms to handle outliers. When a transformation was applied during data generation, the mixture models without Box-Cox transformation fail to perform well. On the contrary, the flexibility of the *t* mixture model with Box-Cox transformation does not penalize too much for model mis-specification. This is illustrated by the results from the GvHD setting: the *t* mixture model with Box-Cox transformation gives the lowest misclassification rates when the true model is instead the *t* mixture model without transformation or the Gaussian mixture model with Box-Cox transformation.

Selecting the number of clusters: In this part of the study, the four models in comparison were applied to each of the 100 datasets generated, setting the number of clusters from one to ten in turn. The number of clusters which delivered the highest BIC was selected. We compared the models via the mode and the 80% coverage interval of the number of clusters selected out of the 100 repetitions. As shown in Table 2, the *t* mixture models can select the correct number of clusters in the majority of repetitions, even in case of model mis-specification. In addition, they deliver the same 80% coverage intervals as the Gaussian mixture models do when data were generated from Gaussian mixtures, suggesting that the robustness against outliers of the *t* mixture models provides satisfactory protection against model mis-specification. On the contrary, the Gaussian mixture models tend to overestimate the number of clusters when an excess of outliers is present in the data generated from *t* mixtures, and in most instances in which overestimation happens, six clusters are selected.

Discussion

The experimental data and the simulation studies have demonstrated the importance of handling transformation selection, outlier identification and clustering simultaneously. While a stepwise approach in which transformation is preselected ahead of outlier detection (or vice versa) may be considered, it is unlikely to tackle the problem well in general as the preselected transformation may be influenced by the presence of outliers. This is shown in the analysis of the Rituximab dataset, without outlier removal the use of Gaussian mixture models led to inappropriate transformation and poor classification in order to accommodate outliers (Figure 3(d) and Supplementary Figure 3(d)). Conversely, without transformation, the *t* mixture model could not model the shape of the top cluster well (Figure 3(c) and Supplementary Figure 3(c)). Similarly, it is necessary to perform transformation selection and clustering simultaneously [37, 38] as opposed to a stepwise approach. It is difficult to know what transformation to select beforehand as one only observes the mixture distribution, and the classification labels are unknown. A skewed distribution could be the result of one dominant cluster and one (or more) smaller clusters. As shown by our analysis with the experimental data and the simulation studies, our proposed approach based on *t* mixture models with Box-Cox transformation benefits from handling these issues, which have mutual influence, simultaneously. Furthermore, confirmed by results of our simulation studies, our proposed approach is robust against model mis-specification and can avoid the problem of Gaussian mixture models that excessive clusters are often needed to provide a reasonable fit in case of model mis-specification [34].

One of the benefits of model-based clustering is that it provides mechanism for both "hard" clustering (i.e., the partitioning of the whole data into separate clusters) and fuzzy clustering (i.e., a "soft" clustering approach in which each event may be associated with more than one cluster). The latter approach is in line with the rationale that there exists uncertainty about to which cluster an event should be assigned. The overlaps between clusters as seen in Figure 3 and 4 reveal such uncertainty in the cluster assignment. When fuzzy clustering is considered, the posterior probability $\tilde{z}_{ig}$ can be interpreted as the evidence of the association of y_i with cluster g; when a partition of data is desired, we may assign each observation y_i to cluster g associated with the largest $\tilde{z}_{ig}$ value.

In many FCM clustering applications the number of clusters is usually unknown and requires estimation. There are several approaches for choosing the number of clusters in model-based clustering including resampling, cross validation, and various information criteria [57]. Our approach to the problem is based on the BIC, which gives

good results in the context of mixture models [33, 58]. BIC is computationally cheap to compute once maximum likelihood estimation for the model parameters has been completed, an advantage over other approaches, especially in the context of FCM where datasets tend to be very large. While computationally cheap, BIC relies heavily on an approximation of marginal likelihoods, which might not be very accurate for some data. Currently, we are looking for alternatives, for example, the Integrated Completed Likelihood (ICL) [59], to improve the estimation of the number of clusters. Nevertheless, combined with expert knowledge, we view BIC as a useful tool that can provide guidance on choosing a reasonable value, as supported by our simulation study of assessing the accuracy in selecting the number of clusters.

There exist several modified versions of the Box-Cox transformation to handle negative-valued data, for example, the log-shift transformation, which was also proposed in the paper for the original Box-Cox transformation [36]. The advantage of our choice, given by (10), is that while continuity is maintained across the whole range of the data, it retains the simplicity of the form of the transformation without introducing any additional parameters; when all data are positive, it reduces to the same form of the original Box-Cox transformation.

It is well known that the convergence of the EM algorithm depends on the initial conditions used. A bad initialization may incur slow convergence, and/or convergence to a local minimum. In the real-data examples and the simulation studies, we used a deterministic approach called hierarchical clustering [30, 49] for initialization. We have found this approach to perform well in the datasets explored here. However, better initialization, perhaps incorporating expert knowledge, might be needed for more complex datasets. For example, if there is a high level of noise in the data, it might be necessary to use an initialization method that accounts for such outliers; see [33] for an example.

To estimate how long it takes to analyze a sample of size typical for an FCM dataset, we have carried out a test run on a synthetic dataset, which consists of one million events and 10 dimensions. To complete an analysis with 10 clusters, it took about 20 minutes on a 3GHz Intel Xeon processor with 2GB of RAM. This illustrates that the algorithm should be quick enough for analyzing a large flow dataset. In general, the computational time increases linearly with the number of events and increases in the order of p^2 with the number of variables, p, per EM iteration. This is an advantage over hierarchical clustering in which the computational time and memory space required increase in the order of n^2 with the number of events, making a hierarchical approach impractical when a

sample of a moderate size, say, >5000, is investigated. Meanwhile, we realize the need of high computational speed in FCM analysis, and are currently investigating means to speed up the EM algorithm for parameter estimation.

Like all clustering approaches, the methodology we have developed includes assumptions, which may limit the applicability of this approach, and it will not identify every cell population in every sample. If the distribution of the underlying population is highly sparse without a well-defined core, our approach may not properly identify all sub-populations. This is illustrated in the Rituximab analysis where the loosely structured group of apoptotic cells was left undetected. This in turn has hindered the capability of the approach from giving satisfactory estimates of the G1 and S frequencies for the identified clusters that would be desired for normal analysis of a 7-AAD DNA distribution for cultured cells. On the other hand identification of every cluster may not always be important. The Rituximab study was designed as a high throughput drug screen to identify compounds that caused a >50% reduction in S-phase cells [55], as would be captured by both the manual gates and our automated analysis should it occur. Furthermore, the exact identification of every cluster through careful manual analysis may not always be possible, especially in high throughput experiments. For instance, in the manual analysis of the GvHD dataset, a quadrant gate was set in Figure 1(c) in order to identify the $CD3^{+}CD4^{+}CD8\beta^{+}$ population, which was of primary interest. For convenience sake, this gate was set at the same level across all the samples being investigated. While five clusters can be clearly identified on the graph, it would be time consuming to manually adjust the positions of each of the gates for all the samples in a high throughput environment, as well as identify all novel populations. Contrariwise, our automated approach can identify these clusters in short order without the need for manual adjustment. To complete the analysis of the GvHD dataset (>12,000 cells, six dimensions) to identify the $CD3^{+}CD4^{+}CD8\beta^{+}$ population (Figure 1), it took less than 5 minutes, using the aforementioned sequential approach to clustering, on an Intel Core 2 Duo with 2GB of RAM running Mac OS X 10.4.10.

A rigorous quantitative assessment is important before implementing this, or any approach, as a replacement for expert manual analysis. The availability of a wide variety of example data would aid in the development and evaluation of automated analysis methodologies. We are therefore developing such a public resource, and would welcome contributions from the wider FCM community.

An **R** [60] package called flowClust is being developed to implement the clustering methodology proposed in this paper. The source code is built in **C** for optimal utilization of system resources and makes use of the BLAS

library [61], which enables multi-threaded processes. The **R** package will be available from Bioconductor [62] at http://www.bioconductor.org.

Acknowledgements

This research was partially funded by the Natural Sciences and Engineering Research Council of Canada. Datasets were kindly provided by Maura Gasparetto and Clayton Smith. We would also like to thank Maura Gasparetto for her assistance on the FlowJo plots. RRB is an International Society for Analytical Cytology Scholar and a Michael Smith Foundation for Health Research Scholar.

References

1. Lizard G. Flow Cytometry Analyses and Bioinformatics: Interest in New Softwares to Optimize Novel Technologies and to Favor the Emergence of Innovative Concepts in Cell Research. Cytometry A 2007;71A:646–647.

2. Braylan RC. Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias. Cytometry A 2004;58A:57–61.

3. Bagwell CB. DNA histogram analysis for node-negative breast cancer. Cytometry A 2004;58A:76–78.

4. De Rosa SC, Brenchley JM, Roederer M. Beyond six colors: a new era in flow cytometry. Nat. Med. 2003;9:112–117.

5. Redelman D. CytometryML. Cytometry A 2004;62A:70–73.

6. Roederer M, Hardy R. Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56–64.

7. Roederer M, Treister A, Moore W, Herzenberg LA. Probability Binning Comparison: A Metric for Quantitating Univariate Distribution Differences. Cytometry 2001;45:37–46.

8. Roederer M, Moore W, Treister A, Hardy RR, Herzenberg LA. Probability Binning Comparison: A Metric for Quantitating Multivariate Distribution Differences. Cytometry 2001;45:47–55.

9. Tzircotis G, Thorne RF, Isacke CM. A new spreadsheet method for the analysis of bivariate flow cytometric data. BMC Cell Biol. 2004;5:10.

10. Spidlen J, Gentleman RC, Haaland PD, et al. Data standards for flow cytometry. OMICS 2006;10(2):209–214.

11. Suni MA, Dunn HS, Orr PL, et al. Performance of plate-based cytokine flow cytometry with automated data analysis. BMC. Immunol. 2003;4:9.

12. Parks DR. Data Processing and Analysis: Data Management. In: Current Protocols in Cytometry. New York: John Wiley & Sons Inc.; 1997. chap. 10:10.1.1–10.1.6.

13. McLachlan G, Peel D. Finite mixture models. New York: Wiley-Interscience; 2000.

14. MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. Berkeley: University of California Press; 1967. p 281–297.

15. Murphy RF. Automated Identification of Subpopulations in Flow Cytometric List Mode Data Using Cluster Analysis. Cytometry 1985;6:302–309.

16. Bakker Schut TC, Grooth BGD, Greve J. Cluster Analysis of Flow Cytometric List Mode Data on a Personal Computer. Cytometry 1993;14:649–659.

17. Demers S, Kim J, Legendre P, Legendre L. Analyzing Multivariate Flow Cytometric Data in Aquatic Sciences. Cytometry 1992;13:291–298.

18. Wilkins MF, Hardy SA, Boddy L, Morris CW. Comparison of Five Clustering Algorithms to Classify Phytoplankton From Flow Cytometry Data. Cytometry 2001;44:210–217.

19. Rousseeuw PJ, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices. Comput. Statist. Data Anal. 1996;23:135–151.

20. Maynadié M, Picard F, Husson B, Chatelain B, Cornet Y, Le Roux G, Campos L, Dromelet A, Lepelley P, Jouault H, Imbert M, Rosenwadj M, Vergé V, Bissières P, Raphaël M, Béné MC, Feuillard J, GEIL. Immunophenotypic clustering of myelodysplastic syndromes. Blood 2002;100(7):2349–2356.

21. Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, Salvioli G, Patsekin V, Robinson JP, Durante C, Cocchi M, Cossarizza A. Subject Classification Obtained by Cluster Analysis and Principal Component Analysis Applied to Flow Cytometric Data. Cytometry A 2007;71A:334–344.

22. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks; 1984.

23. Boddy L, Morris CW. Artificial neural networks for pattern recognition. In: Fielding AH, editor. Machine learning methods for ecological applications. Boston: Kluwer; 1999. p 37–87.

24. Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, Massachusetts: The MIT Press; 2002.

25. Burges CJC. A tutorial on support vector machines for pattern recognition. Boston: Kluwer; 1998.

26. Beckman RJ, Salzman GC, Stewart CC. Classification and Regression Trees for Bone Marrow Immunophenotyping. Cytometry 1995;20:210–217.

27. Kothari R, Cualing H, Balachander T. Neural Network Analysis of Flow Cytometry Immunophenotype Data. IEEE Trans. Biomed. Eng. 1996;43(8):803–810.

28. Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Mar. Ecol. Prog. Ser. 2000;195:47–59.

29. Morris CW, Autret A, Boddy L. Support vector machines for identifying organisms – a comparison with strongly partitioned radial basis function networks. Ecol. Model. 2001;146:57–67.

30. Banfield JD, Raftery AE. Model-based Gaussian and Non-Gaussian Clustering. Biometrics 1993;49:803–821.

31. Titterington DM, Smith AFM, Makov UE. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons Inc.; 1985.

32. McLachlan GJ, Basford KE. Mixture Models: Inference and Applications to Clustering. Marcel Dekker Inc.; 1988.

33. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 2002;97(458):611–631.

34. Yeung KY, Fraley C, Murua A, et al. Model-Based Clustering and Data Transformations for Gene Expression Data. Bioinformatics 2001;17:977–987.

35. Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. Genome Biol. 2002;3(2):R9.

36. Box GEP, Cox DR. An analysis of transformations. J. R. Statist. Soc. B 1964;26:211–252.

37. Schork NJ, Schork MA. Skewness and Mixtures of Normal Distributions. Comm. Statist. Theory Methods 1988;17:3951–3969.

38. Gutierrez RG, Carroll RJ, Wang N, et al. Analysis of Tomato Root Initiation Using a Normal Mixture Distribution. Biometrics 1995;51:1461–1468.

39. Dvorak JA, Banks SM. Modified Box-Cox Transform for Modulating the Dynamic Range of Flow Cytometry Data. Cytometry 1989;10:811–813.

40. Peel D, McLachlan GJ. Robust Mixture Modelling Using the *t* Distribution. Stat. Comput. 2000;10(4):339–348.

41. Carroll RJ. Prediction and power transformation when the choice of Power is restricted to a finite set. J. Amer. Statist. Assoc. 1982;77:908–915.

42. Atkinson AC. Transformations Unmasked. Technometrics 1988;30:311–318.

43. Lange KL, Little RJA, Taylor JMG. Robust Statistical Modeling Using the *t*-distribution. J. Amer. Statist. Assoc. 1989;84:881–896.

44. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Statist. Soc. B 1977;39:1–22.

45. Celeux G, Govaert G. Gaussian parsimonious clustering models. Pattern Recognit. 1995;28(5):781–793.

46. Liu C, Rubin D. ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. Statist. Sinica 1995;5:19–39.

47. Liu C. ML estimation of the multivariate *t* distribution and the EM algorithm. J. Multivariate Anal. 1997;63:296–312.

48. Bickel PJ, Doksum KA. An analysis of transformations revisited. J. Amer. Statist. Assoc. 1981;76(374):296–311.

49. Fraley C. Algorithms for Model-based Gaussian Hierarchical Clustering. SIAM J. Sci. Comput. 1998;20: 270–281.

50. Silverman B. Density Estimation for Statistics and Data Analysis. New York: Chapman-Hall; 1986.

51. Schwarz G. Estimating the Dimension of a Model. Ann. Statist. 1978;6:461–464.

52. Hahne F, Arlt D, Sauermann M, Majety M, Poustka A, Wiemann S, Huber W. Statistical methods and software for the analysis of high throughput reverse genetic assays using flow cytometry readouts. Genome Biol. 2006;7:R77.

53. Flow Informatics and Computational Cytometry Society [http://www.ficcs.org/software.html#Data_Files].

54. Abraham VC, Taylor DL, Haskins JR. High content screening applied to large-scale cell biology. Trends Biotechnol. 2004;22:15.

55. Gasparetto M, Gentry T, Sebti S, O'Bryan E, Nimmanapalli R, Blaskovich MA, Bhalla K, Rizzieri D, Haaland P, Dunne J, Smith C. Identification of compounds that enhance the anti-lymphoma activity of rituximab using flow cytometric high-content screening. J. Immunol. Methods 2004;292:59–71.

56. Brinkman RR, Gasparetto M, Lee SJJ, Ribickas A, Perkins J, Janssen W, Smiley R, Smith C. High-Content Flow Cytometry and Temporal Data Analysis for Defining a Cellular Signature of Graft-Versus-Host Disease. Biol. Blood Marrow Transplant. 2007;13(6):691–700.

57. Biernacki C, Govaert G. Choosing Models in Model-based Clustering and Discriminant Analysis. J. Stat. Comput. Simul. 1999;64:49–71.

58. Fraley C, Raftery AE. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Comput. J. 1998;41(8):578–588.

59. Biernacki C, Celeux G, Govaert G. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 2000;22(7):719–725.

60. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J. Comput. Graph. Statist. 1996;5:299–314.

61. Lawson CL, Hanson RJ, Kincaid DR, Krogh FT. Algorithm 539: Basic Linear Algebra Subprograms for Fortran Usage [F1]. ACM Transactions on Mathematical Software 1979;5(3):324–325.

62. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.

63. Celeux G, Govaert G. A classification EM algorithm for clustering and two stochastic versions. Comput. Statist. Data Anal. 1992;14(3):315–332.

Tables

Table 1. Misclassification rates for different models applied on data generated under the Rituximab or GvHD setting

		Model used to fit data			
		t + Box-Cox	*t*	Gaussian + Box-Cox	Gaussian
Model used to generate data under the Rituximab setting	*t* + Box-Cox	**0.187**	0.211	0.279	0.251
	t	**0.255**	0.263	0.339	0.315
	Gaussian + Box-Cox	0.321	0.400	**0.251**	0.352
	Gaussian	0.344	0.329	0.317	**0.301**
Model used to generate data under the GvHD setting	*t* + Box-Cox	**0.112**	0.116	0.205	0.230
	t	**0.107**	0.111	0.191	0.221
	Gaussian + Box-Cox	**0.135**	0.143	0.139	0.152
	Gaussian	0.134	0.132	0.132	**0.126**

The best results are shown in bold.

Table 2. Modes and 80% coverage intervals of the number of clusters selected using different models on data generated under the GvHD setting

		Model used to fit data							
		t + Box-Cox		*T*		Gaussian + Box-Cox		Gaussian	
		Mode	Interval	Mode	Interval	Mode	Interval	Mode	Interval
Model used to generate data	*t* + Box-Cox	5	(5, 6)	5	(5, 6)	6	(6, 7)	6	(6, 8)
	t	5	(5, 7)	5	(5, 6)	6	(6, 7)	6	(6, 8)
	Gauss. + Box-Cox	5	(5, 6)	5	(5, 6)	5	(5, 6)	5	(5, 6)
	Gaussian	5	(5, 6)	5	(5, 6)	5	(5, 6)	5	(5, 6)

Figures

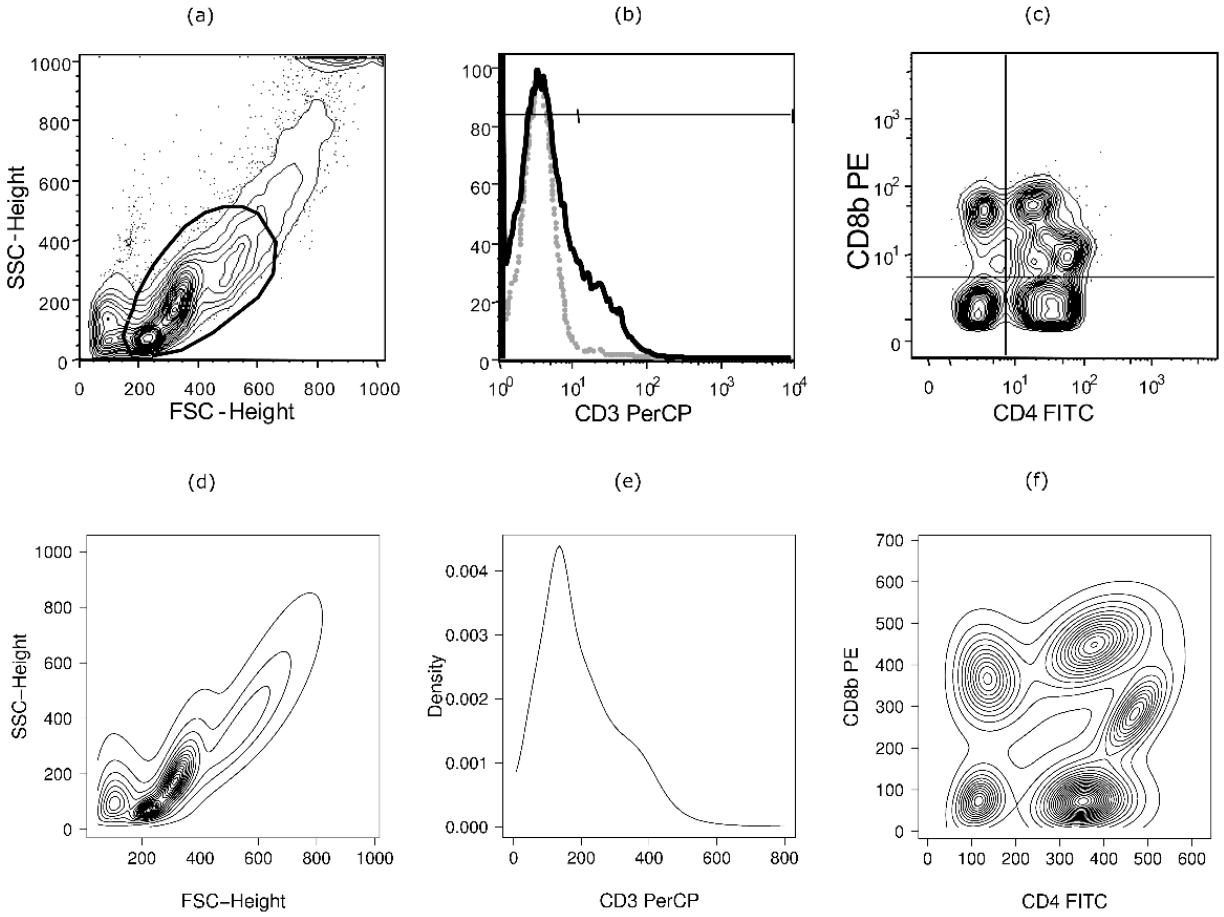

Figure 1. Strategy for clustering the GvHD positive sample to look for $CD3^{+}CD4^{+}CD8\beta^{+}$cells. The manual gating strategy is shown in (a-c). (a) Using FlowJo, a gate was drawn by an expert researcher to define the lymphocyte population. (b) The selected cells were projected on the CD3 dimension, and CD3 cells were defined through setting a cutoff at around 15. (c) Cells within the upper right gate were referred to as $CD3^{+}CD4^{+}CD8\beta^{+}$. (d-f) A *t* mixture model with Box-Cox transformation was used to mimic this manual selection process; here we display the corresponding density estimates. For FlowJo, the density estimates correspond to kernel estimates, while for our gating strategy, the density estimates are obtained from the estimated mixture models.

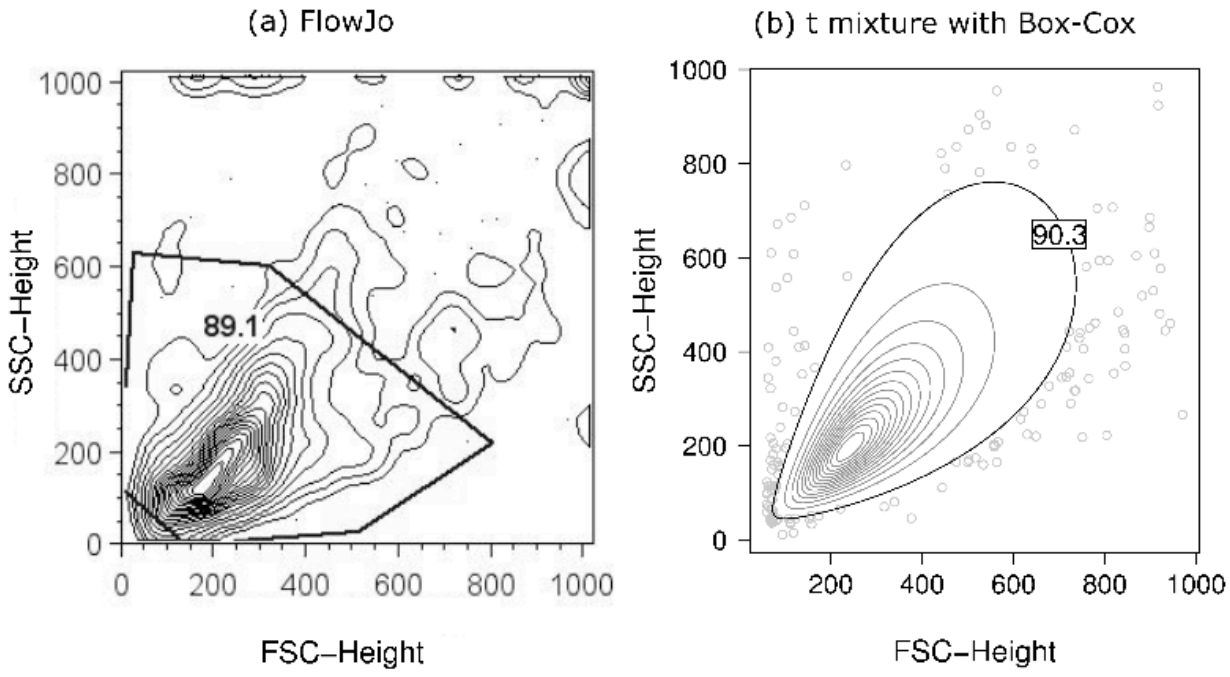

Figure 2. Initial clustering of the Rituximab data using the FSC and SSC variables. (a) In typical analysis a gate was manually drawn to select a group of cells for further investigation. (b) A *t* mixture model with Box-Cox transformation was used to mimic this manual selection process. In (b) points (shown in gray) outside the boundary drawn in black have weights ($\tilde{u}_{ig}$'s) less than 0.5 and will be removed from the next stage. It can be shown that this boundary corresponds approximately to the 90th percentile (a conventional choice) region for the *t* distribution transformed back on the original scale using the Box-Cox parameter. The numbers shown in both plots are the percentages of points within the boundaries which are extracted for the next stage. Both gates capture the highest density region, as shown by the two density estimates. For FlowJo, the density estimate corresponds to a kernel estimate, while for our gating strategy, the density estimate is obtained from the estimated mixture model.

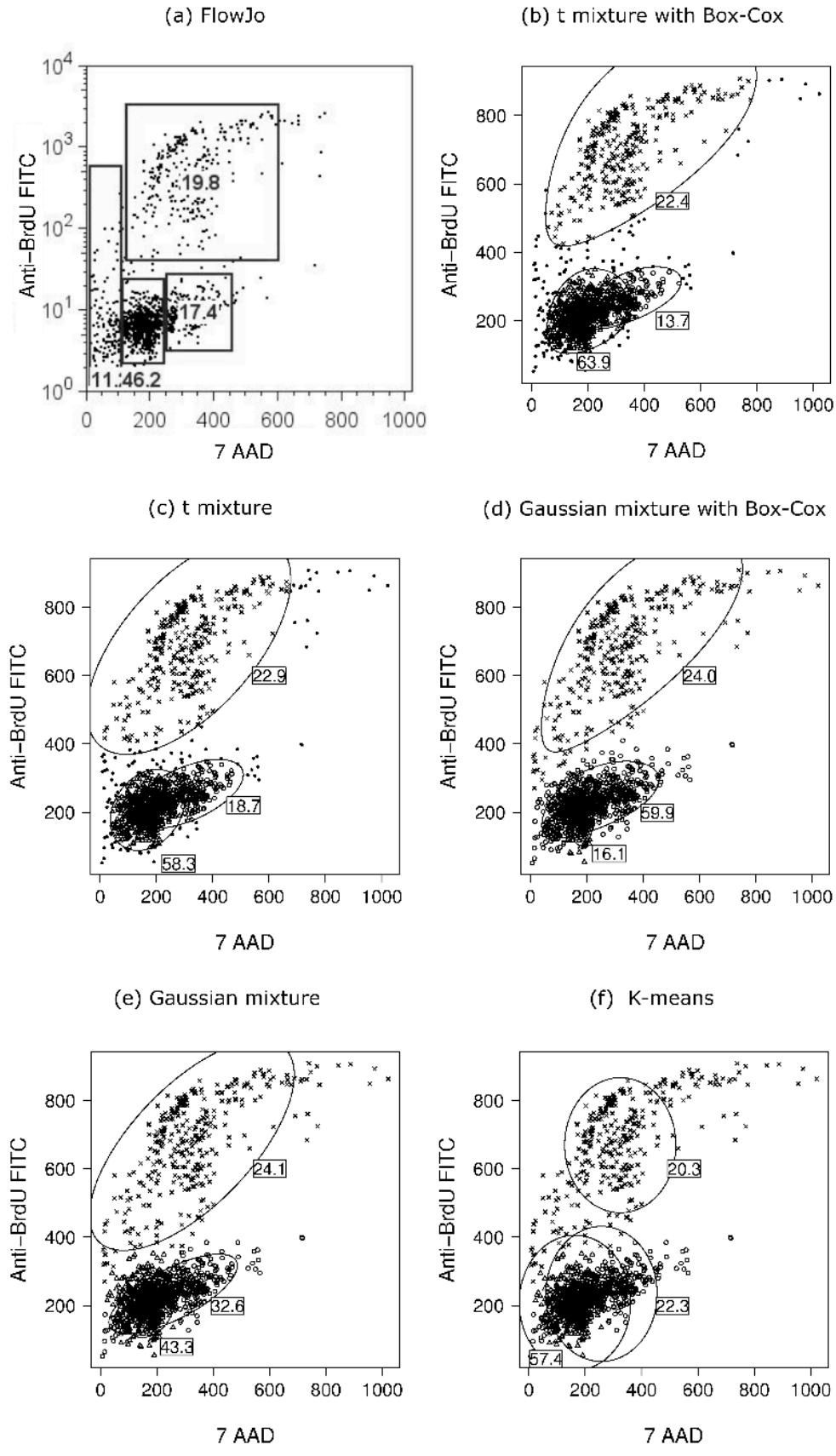
(a) FlowJo
(b) t mixture with Box-Cox
(c) t mixture
(d) Gaussian mixture with Box-Cox
(e) Gaussian mixture
(f) K-means
Anti−BrdU FITC
7 AAD
19.8
17.4
11.4
6.2
22.4
13.7
63.9
22.9
18.7
58.3
24.0
59.9
16.1
24.1
32.6
43.3
20.3
22.3
57.4

Figure 3. Second-stage clustering of the Rituximab data using all the fluorescent markers (3 clusters). Clustering was performed on the cells preselected from the first stage as shown in Figure 2. In (b-f) the number of clusters was set to be three. (b-c) Points outside the boundary drawn in black have weights less than 0.5 and are labeled with "•" when *t* distributions were used. (d-f) For clustering performed without using *t* distributions, for comparison sake, boundaries are drawn in a way such that they correspond to the region of the same percentile which the boundaries drawn in (b-c) represent. Different symbols are used for the different clusters. The numbers shown in all plots are the percentages of cells assigned to each cluster. The K-means algorithm is equivalent to the classification EM algorithm [45, 63] for a Gaussian mixture model assuming equal proportions and a common covariance matrix being a scalar multiple of the identity matrix. The spherical clusters with equal volumes drawn in (f) correspond to such a constrained model.

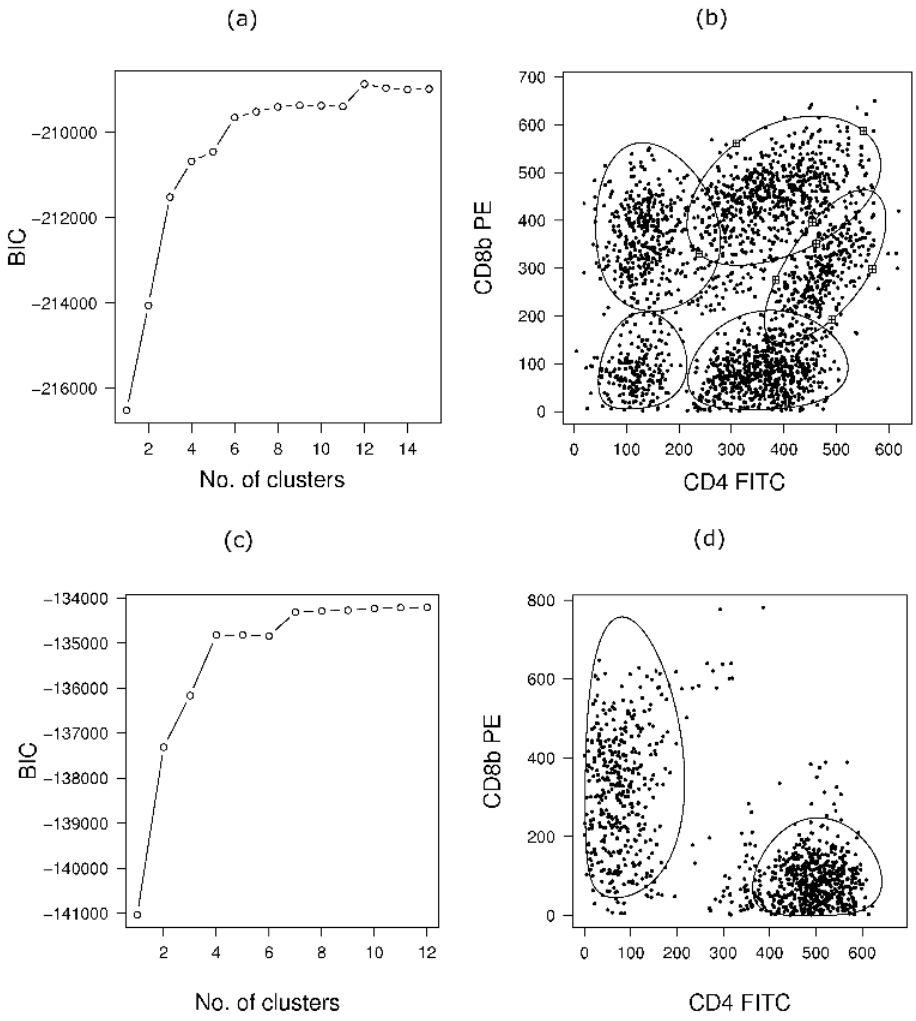

Figure 4. Second-stage clustering of the GvHD positive sample (a-b) and control sample (c-d) using all the fluorescent markers. Clustering was performed on the cells preselected from the first stage. For the positive sample, (a) the BIC reaches a maximum at 12 clusters; (b) the scatterplot reveals the cluster assignment of the cells. Points which are assigned to the five clusters with high CD3 means are classified as $CD3^+$ cells. The five regions drawn in solid lines form the CD3 population. The two regions in the upper right marked with the ⊞ symbols are identified as the $CD3^+CD4^+CD8\beta^+$ population. For the control sample, (c) little increment is observed in the BIC beyond seven clusters, suggesting that seven clusters, much fewer than for the positive sample, are enough to model the data in the second stage; (d) the scatterplot reveals the cluster assignment of the cells. Only two clusters have been used to model the $CD3^+$ population. Please refer to Supplementary Figure 6 for scatterplots showing additional information about the $CD3^-$ population.