Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

Mixture Modeling Approach to Flow Cytometry Data

Michael J. Boedigheimer,[1]* John Ferbas[2]

[1]Computational Biology, Amgen, Thousand Oaks, California

[2]Department of Clinical Immunology, Amgen, Thousand Oaks, California

Received 10 July 2007; Accepted 2 February 2008

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/1552-4922/suppmat.

Grant sponsor: Amgen Inc.

*Correspondence to: Michael J. Boedigheimer, One Amgen Center Drive, Thousand Oaks, CA 93130, USA.

Email: mboedigh@amgen.com

Published online 27 March 2008 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20553

Abstract

Flow Cytometry has become a mainstay technique for measuring fluorescent and physical attributes of single cells in a suspended mixture. These data are reduced during analysis using a manual or semiautomated process of gating. Despite the need to gate data for traditional analyses, it is well recognized that analyst-to-analyst variability can impact the dataset. Moreover, cells of interest can be inadvertently excluded from the gate, and relationships between collected variables may go unappreciated because they were not included in the original analysis plan. A multivariate non-gating technique was developed and implemented that accomplished the same goal as traditional gating while eliminating many weaknesses. The procedure was validated against traditional gating for analysis of circulating B cells in normal donors ($n = 20$) and persons with Systemic Lupus Erythematosus ($n = 42$). The method recapitulated relationships in the dataset while providing for an automated and objective assessment of the data. Flow cytometry analyses are amenable to automated analytical techniques that are not predicated on discrete operator-generated gates. Such alternative approaches can remove subjectivity in data analysis, improve efficiency and may ultimately enable construction of large bioinformatics data systems for more sophisticated approaches to hypothesis testing.

Key terms

flow cytometry; Gaussian mixture modeling; immunophenotyping; automated data analysis

THE use of flow cytometry to simultaneously measure multiple characteristics of thousands of individual cells has become widespread and techniques and protocols have been developed to study a variety of biological processes, including cell lineage, maturation and activation state, apoptosis, activation of signaling cascades ("phosflow"), and proliferation status (1). A fundamental component of flow cytometric analyses includes the application of binary thresholds, called gates, to some or all of the detection channels of the instrument. To date, the most popular gating strategies are manual, with variations and guidelines for primary gating strategies on (e.g.) whole blood specimens ranging from physical characteristics of the cells (only) versus the use of CD45 as a tool to assist the analyst (2,3). Despite the simplicity and long history of successful clinical applications for the manual gating approach, it remains a subjective and intensive manual effort. Many users have called attention to the need for more advanced analysis techniques (4).

When manual gating strategies are implemented, individual gates are usually limited to one, two, or rarely three dimensions. Rigid gating and sample analysis in higher dimensions is currently difficult because it is difficult to conceptualize and cannot be conducted with existing software. The practical implication is that it is difficult and time consuming to find optimal gates for various subsets of cells in a time-efficient manner and data analysis is by necessity restricted to the populations of cells that are of the most interest the individual investigator. Ultimately, the ability to exploit multiple parameters simultaneously depends on the use of more sophisticated multivariate methods. Although some publications show the advantage of such

ISAC
International Society for Advancement of Cytometry

methods (4), they do not seem to be widely used. This manuscript reports on our efforts to apply mixture modeling (MM) as a means to apply objective and reproducible gating strategies and analyses in a time efficient way.

In this method, a sample is assumed to consist of a mixture of components, each of which can be characterized by a multivariate distribution. The components in the model correspond to event types (e.g., lymphocyte) and the variables in the model correspond to attributes measured during flow cytometry. In a Gaussian mixture model, the components are multidimensional equivalents of ellipsoids. After the number of desired components is chosen, a search procedure is used that finds a location, size, and shape of the spheres that is maximally likely for the given data. MM is commonly used in other fields such as artificial intelligence (5), but it has not yet been used to evaluate flow cytometry data, a task to which it appears to be ideally suited.

In this article, we demonstrate the theoretical advantage of using a Gaussian mixture model and an expectation maximization (EM) algorithm relative to the traditional gating approach. The performance of this automated application is compared to a manual expert analysis. The data show that the automated method can recapitulate manual analyses with the added benefits of speed, objectivity, reproducibility, and the ability to evaluate several subpopulations in many dimensions simultaneously.

Materials and Methods

Study Participants and Flow Cytometry

Data for this analysis were list-mode files from a longitudinal study of 42 persons with systemic lupus erythematosus (SLE) and 20 non-SLE controls (6), who provided informed consent prior to participating in the study. Persons with SLE were categorized with mild or more severe disease activity according to their SLEDAI score (7), using an arbitrary cutoff. Their samples were run in parallel with the non-SLE controls with the aim to identify unique immunophenotypic features of B cell subsets in SLE patients; these data will be reported elsewhere (manuscript in preparation). For the current investigation—where our emphasis was placed on the analysis algorithm rather than unique cellular phenotypes among SLE patients—we chose list-mode data for CD19 and CD20 fluorescence intensity and percentages. It was already determined by traditional gating methods that differences existed between persons with SLE and controls, making this an appropriate dataset to test the hypothesis that the MM approach could recapitulate such trends in the data and offer additional analytical advantages. Moreover, inclusion of a diseased population was done to insure that the MM approach held value in instances where specimens may have unusual physical properties or staining characteristics as a function of disease or disease severity.

The blood specimens for this effort were collected in sodium heparin (Cedars Sinai Medical Center, Los Angeles, CA) and transported at ambient temperature to the Amgen Clinical Immunology Flow Cytometry Laboratory for next day processing. The analysis tube used to generate the CD20 data was whole blood labeled with the manufacturer's recommended amounts of anti-CD20 FITC, anti-CD14 PE, anti-CD19 PerCP, and anti-CD45 APC (Becton Dickinson Immunocytometry Systems, San Jose, CA). Prior to analysis on a FACSCalibur flow cytometer (Becton Dickinson Immunocytometry Systems), red blood cells were lysed with an ammonium chloride lysis buffer (8); cells were not fixed but rather analyzed immediately after processing. Data were analyzed as part of the study using standard manual CD45 (versus side scatter) gating procedures (2,3) on CellQuest version 3.3 software on a Macintosh computer running operating system version 9.0. Total B cells were derived as CD19 positive, CD20 positive cells in the CD45 gate, whereas CD19 and CD20 fluorescence intensities were derived from single parameter histograms with appropriate analysis regions set by the operator. Instrument setup included daily runs of BD Calibrite^{TM} beads with the lyse-wash setup selected in FACSComp^{TM} (Becton Dickinson Immunocytometry Systems); glutaraldehyde-fixed chicken red blood cells (Biosure, Grass Valley, CA) were used to adjust and standardize photomultiplier tube settings to control for fluorescence intensity measurements (9).

The Mixture Modeling Method

After the clinical study was concluded and the data were analyzed, the MM method was developed and the data were reanalyzed. The FCS files were imported into Matlab R2006a and were transformed to the same units used in the manual process according the specification in the FCS file (10). Matlab software implementation of the methods will be made available on the Mathworks File Exchange web site. The results of the new method are compared with results from our standard operating procedures for cell counting and fluorescence intensity reporting.

An EM algorithm was used to model flow cytometry data as a mixture of Gaussians. The model includes a discrete process that models a flow-cytometry event originating from one of K possible event-types and a continuous process that models the distribution of intensities in N channels or dimensions for each event-type. Each event type occurs with probability α_k, which is interpreted as the class frequency. It is assumed that observations of events from each class are sampled from independent normal distributions, x_k, although the algorithm can work with any continuous distribution. The normal density function is given by

$$p_k = (2\pi)^{-N/2}|\Sigma_k|^{-1/2}e^{-(1/2)(x_k-\mu_k)'\Sigma_k^{-1}(x_k-\mu_k)}.$$

For each event type or class, we are interested in estimating the class probabilities, α_k, a N-vector of mean intensities, μ_k, and the $N \times N$ covariance, σ_k. The EM algorithm finds locally optimal parameter estimates (11), θ, through the interactive application of a two step process. Given an initial set of starting conditions, θ_0, the first step of the algorithm is the expectation step, which estimates the expectation of the underlying distribution given a set of observations, $y_{j\{j=1\ldots M\}}$, and a set of parameter estimates, θ_t: $E(x_k|y_j, \theta_t) = \alpha_k p_k$.

The second step, the maximization step, estimates a new set of parameters, θ_{t+1}, given the estimated expectations of the previous step. The new estimates are given by:

$$\alpha_k = \frac{\sum_{j=1}^{M} E(x_k|y_j,\theta_t)}{\sum_{j=1}^{M}\sum_{k=1}^{N} E(x_k|y_j,\theta_t)}$$
$$\mu_k = \frac{\sum_{j=1}^{M} E(x_k|y_j,\theta_t) y_j}{\sum_{j=1}^{M} E(x_k|y_j,\theta_t)}$$
$$\sigma_k = \frac{\sum_{j=1}^{M} E(x_k|y_j,\theta_t)(y_j-\mu_k)'(y_j-\mu_k)}{\sum_{j=1}^{M} E(x_k|y_j,\theta_t)}$$

It can be seen that the new μ and σ are the weighted mean and weighted covariance, where the estimated expectation values of each observation is used for the weights. Each iteration of the algorithm is guaranteed not to decrease the likelihood and therefore continues to produces more refined estimates. The process is stopped when predefined criterion are reached. Our default is to stop when changes to the average log likelihood are less than 1e-8 or when the number of iterations exceeds 200 times the number of estimated parameters.

This model allows for some classes to share a common mean and covariance structure in some dimensions. This is done by applying constraints so that for some subset of dimensions, q, and subset of classes, s, the underlying populations assumed to be from the same distributions.

$$\mu_{(a,b)} = \mu_{(a,b)} \quad (\forall_{a\in q,}\forall_{b\in s})$$
$$\sigma_{(a,a,b)} = \sigma_{(a,a,b)} \quad (\forall_{a\in q,}\forall_{b\in s})$$

These constraints are enforced during the maximization step, using the combined weighted mean and pooled covariance.

$$\mu_{(a,b\in s)} = \frac{\sum_{a\in q}\sum_{j=1}^{M} E(x_{(a,b)}|y_j,\theta_t) y_j}{\sum_{a\in q}\sum_{j=1}^{M} E(x_{(a,b)}|y_j,\theta_t)}$$
$$\sigma_{(a,a,b\in s)} = \frac{\sum_{a\in q}\sum_{j=1}^{M} E(x_{(a,b)}|y_j,\theta_t)(y_j-\mu_k)'(y_j-\mu_k)}{\sum_{j=1}^{M} E(x_{(a,b)}|y_j,\theta_t)}$$

The covariance matrix for unconstrained dimensions is calculated as usual. It is not guaranteed the sigma calculated in this way is a positive definite matrix. Thus calculating the expectation using variance constraints can be problematic for some datasets.

Statistics

The subsets of B-cells were defined by a single Gaussian in the mixture. The estimated frequency of these cells along with the estimated fluorescence level of CD19 and CD20 was determined using the EM method. In each case the parameters were compared between normal, mild SLE (SLEDAI <= 7), and severe SLE (SLEDAI >= 8) using analysis of variance model with visit and disease status as factors. The model also includes an interaction term. A threshold of 0.05 was used for significance testing.

Results

Figure 1 shows the results of a simulation to illustrate the motivation behind applying the mixture model (MM) method to flow cytometry data. In each run, a sample of 10,000 points from two normal distributions with relative frequency of 0.1 and 0.9 was taken. The distance between populations was varied in each run and ranged from 1 to 2 standard deviations. These distances were specifically chosen to be small compared to those typically observed in flow cytometry data to show that the MM works reliably and without appreciable bias in its estimates of location or frequency. The top two panels show the range of separations used and the bottom panels show the performance of the two approaches. The standard rigid gating approach was simulated by application of a rigid gate half-way between the population means. This simulates a "perfect" gate in the sense that the gates are set with knowledge of the actual locations. The rigid gating shows a large negative bias in the estimated location of the major population (black) as the populations approach. The bias of the manual method is larger in the estimated frequency of the minor population, which is overestimated by more than 2-fold even when the clusters were separated by 2 standard deviations. The MM approach worked well throughout these ranges.

We added a feature that allows estimates of means and variances to be constrained in some attributes in order to implement the assumption that some subsets of cells are indistinguishable in some characteristics. Figure 2 shows the results of using these constraints. The plots contain the forward scatter (FSC) and side scatter (SSC) data collected from one blood sample. The model included four measured parameters: FSC, SSC, CD20, CD14 and six clusters (including monocytes, lymphocytes, CD20 positive lymphocytes and three clusters for debris and high SSC populations not of direct interest in this study). The ellipses are one standard deviation representations of the estimates for three important classes using (a) no constraints, (b) constrained means, and (c) constrained means and variances. Not surprisingly, applying constraints produces different results. If the assumptions that subpopulations share some common characteristics are valid, then the estimates with constrained means will by definition be more likely than those without constraints. In addition, the sample size used to estimate the means and variances in those shared dimensions is effectively larger because the two populations are pooled.

In any MM approach, it is necessary to determine the number of components needed to adequately describe the data. There are automated methods that could be applied (see for example (12)), but we chose empirically based on six samples from the experimental data set. We required at least three components to estimate the parameters for the three cell-populations of clinical interest. In addition, we needed one for the monocytes and granulocytes and debris fields that were frequently included in the data despite the threshold applied during data collection. We discovered that an additional com-

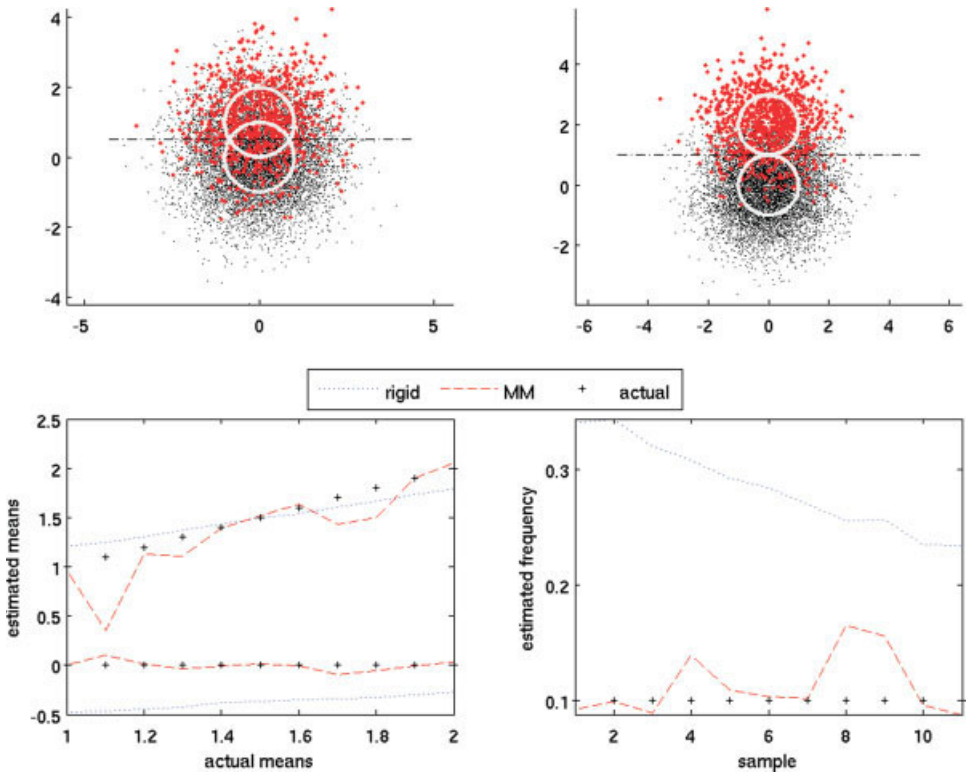

Figure 1. Motivation for using Mixture Model. A series of 11 random Gaussian mixtures of two populations shown as red and black dots (top two panels) with unequal frequency (red 10%, black 90%) were generated so that the distance between populations increased from 1 standard deviation in the first sample (top left) to 2 standard deviation in the last sample (top right); The nine intermediate samples are not shown. The dashed lines represent the ''perfect'' rigid gate half-way between the populations and one standard deviation circles represent the mixture model. The bottom panels show the performance of the mixture modeling (MM) approach and a rigid gating approach. On the left a comparison of the estimated means is made. The rigid gating approach (blue short-dashed lines) introduces systematic bias into the estimate (true means are marked by black ''+''). The MM approach (red long-dashed lines) accurately predicts the true means. The lower right panel shows estimated frequency of the red population in each sample. As the separation between populations increases the accuracy of rigid gating estimate improves (blue line approaches the true frequency), but a bias remains whereas the MM approach accurately predicts the true frequency.

ponent was useful to improve stability of the parameter estimates among the training set, although we later discovered that this did not typically have a large effect on the estimates of interest.

The initial parameter estimates was made using a *k*-means algorithm using the first sample. The optimized parameters were saved as the initial fit for the next five samples. These estimates were averaged to make new initial estimates that were then applied to the entire dataset. Our model also included a hierarchical constraint that ensured that the classes we call CD20 positive and CD20 negative lymphocytes were derived from an identical population as measured in the FSC and SSC.

Figure 3 shows the range of variability encountered between samples in this study. The location of lymphocytes (yellow) and monocytes (red) is shown in one standard deviation ellipses for FSC (*x*-axis) and SSC (*y*-axis). Panel A shows a typical light-scatter pattern of cells. Panels B and C show that some variability was observed in the abundance and relative abundance of certain subsets of cells. Panel C shows a large population of cells with SSC above 300 (granulocytes) that is virtually absent from the sample shown in panel B. Panel D shows an atypical light-scatter pattern from a specimen that would require reprocessing/analysis for clinical use. In all these cases, the placement of the automated gates was acceptable.

Comparison of Manual Gating Versus the Automated MM Approach

An existing set of 390 list-mode files from the blood from 20 normal and 42 SLE subjects taken at seven time points during a clinical study were evaluated (not all subjects completed every study visit). The list-mode file contained data for CD20, CD14, CD19, and CD45 expression as well as FSC and SSC of each cell. With respect to the traditional gating and analysis of these data, CD45, CD14 and FSC and SSC were collectively used by the operator to draw an appropriate lymphocyte gate during analysis (2,3). Of clinical interest was the fluorescence intensity and relative abundance of lymphocytes that expressed the B cell lineage markers CD19 and CD20. With

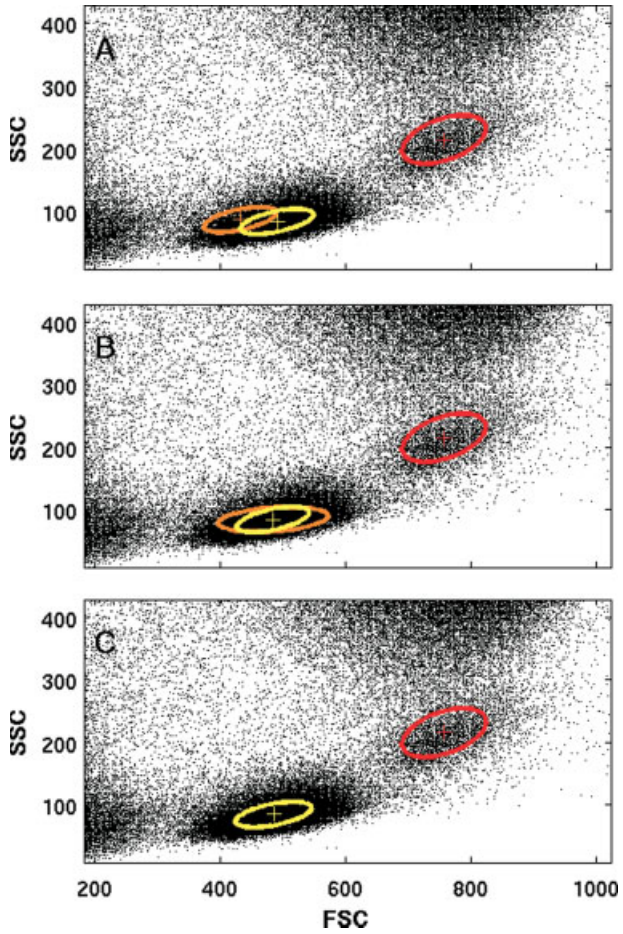

Figure 2. Hierarchically constrained class structure. One standard deviation ellipses indicate the EM solution for three classes using the same four dimensional models, starting points and data. In this figure, two dimensions are shown (FSC and SSC). The expected location for lymphocytes is shown in orange and yellow as points of reference, however, these are one standard deviation ellipses rather than analysis gates; likewise, monocytes can be identified by red ellipses. (**A**) Unconstrained class structure typically estimates different means and variances for subpopulation of lymphocytes. (**B**) Model constrained to have equal means for the two subpopulations. (**C**) Model constrained to have equal means and variances for the two subpopulations. Note that the *x* and *y*-axis (FSC and SSC, respectively) were truncated (do not extend from 0—1024) during acquisition by a primary gate to either exclude debris (FSC) or to limit file size (SSC).

respect to the automated MM analysis, the entire set of files was processed in a few hours. The exact amount of time for these analyses depends on the complexity of the model in terms of dimensions, number of classes, and the stopping criterion.

Mean fluorescence intensity of CD19 and CD20 and B-cell abundance were similar between the manual and automated approaches (Figs. 4 and 5). For example, both methods showed the same trends when comparing samples from subjects with SLE (Systemic Lupus Erythematosus) versus healthy volunteers and had similar variances. Both methods found a significant difference ($P < 0.05$) in CD19 levels between severe SLE and healthy subjects (Figs. 4C and 4D). Both methods also found that B-cell frequency was significantly higher ($P < 0.05$) in subjects with severe SLE (Figs. 4E and 4F). There were, however, some quantitative differences in these two approaches. Notably, only the automated method showed a difference in CD20 levels.

In a sample by sample comparison of CD20 MFI levels we found 10 samples that where clearly very different between the automated and manual processing. Inspection of these revealed that they were the result of convergence to a local minimum and that other starting conditions yielded more optimal fits that converged to a similar solution obtained by the expert. For the above analysis, the outliers were left in the analysis to represent a completely automated approach. However, each of these outliers is automatically identifiable from the data set itself by comparing the final configuration with the starting configuration of clusters and therefore could be identified without the need for an expert (Fig. 5). The Matlab application we developed compares the final solution to the starting point and identifies three types of outliers. The first represents a translation of the centroids. This happens when the final solution is similar to the initial conditions plus or minus a constant for each dimension. The second type of outlier happens when some subset of classes is shifted from its starting point, but each class is still nearest the corresponding class in the initial configuration. The final type of outlier happens when some classes the final configuration are closer to different classes in the initial configuration. In this case, the final configuration may be optimal, but the labels given for the population are inappropriate. In all cases where the initial and final configurations are significantly different a warning flag is set to allow the user to follow up.

Discussion

The ability to separate cells into subsets for further individual analysis is a powerful feature of flow cytometry. Subsets are traditionally defined by classifying cells into groups with detectable/high and nondetectable/low levels of expression or physical attributes. An arbitrary intensity threshold, or gate, is drawn and cells on one side of the gate are included while other are excluded (13). We call this approach rigid gating. Although this method can and does work in a variety of situation it places unnecessary demands on assay design and it has a number of drawbacks that have motivated the development of an alternative method presented here. The approach is called MM and represents a fundamentally different way to view flow cytometry data. Instead of viewing cells as either expressing or not expressing a particular marker, it formally recognizes that changes in expression are continuous and that different cell-types will have a characteristic level of expression. This allows estimates of the relative abundance of the event types, the mean fluorescence intensity and covariance to be made without the need for gating.

Recently an attempt to automate the analysis of multiparameter flow cytometry using a vector quantization algorithm (14). This approach is related to MM approach in that a quantitative description of the location of components

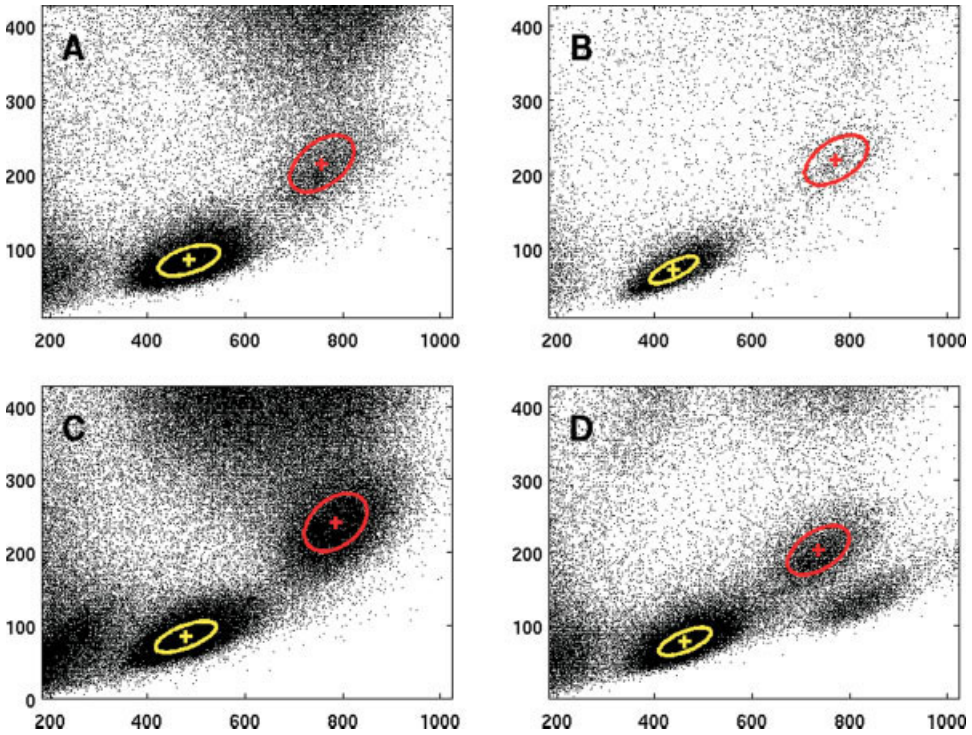

Figure 3. Forward and Side Scatter from four donors. Plot of FSC (*x*-axis) versus SSC (*y*-axis) from four widely different samples along with the final configuration found from the algorithm. As was the case in Figure 2, one standard deviation ellipses indicate the EM solution (yellow = lymphocytes; red = monocytes); they are not traditional analysis gates. Panel (**A**) represents typical data from a healthy volunteer. The remaining plots were selected to represent atypical specimens that likely reflect abnormalities in cell counts as a consequence of SLE or SLE therapeutic intervention (**B**, **C**) or potential technical error in processing (**D**). These latter examples illustrate that the EM solution was not impacted by variations in the proportions or light scatter profiles of the specimens. Note that the *x* and *y*-axis (FSC and SSC, respectively) were truncated (do not extend from 0—1024) during acquisition by a primary gate to either exclude debris (FSC) or to limit file size (SSC).

(vector) is used to classify new events. It differs primarily in that the configuration of components is not optimized in the sense of maximum likelihood. The author also use the vectors to assign events to components, which is not done here.

A major advantage of the MM approach is improved efficiency at distinguishing subsets of cells that are very close together. The efficiency stems from the simultaneous use of data from all dimensions. Subsets are modeled as a multivariate distributions and each cell is a point in multiple dimensions. The likelihood that each cell belongs to each subset is calculated given a the optimal mixture model. Each cell contributes to the estimated mean of each subset proportional to the likelihood of membership in each subset. In this way, cells contribute to varying degrees to all subsets. On the other hand, a rigid gating approach is inefficient because it calculates distance based on each channel separately. This results in the unnecessary exclusion of some desirable events and inclusion of undesirable events. For example, a cell that is outside a gate in one channel is excluded regardless of its position in other channels. This can remove cells that are relatively close to the centroid overall or could include cells far from the centroid if they are near the margins in all dimensions.

We demonstrated the theoretical advantage of the MM when two subsets are close enough that the joint density is difficult or impossible to separate by eye (Fig. 1). We generated a mixture of two classes of homoscedastic, normally distributed univariate data in which the population means where separated by one standard deviation. We also tested a peak-finding approach (not shown) based on scale-space filtering (15). This approach attempts to define regions in the distribution where the slope is unequal to zero or the slope is changing significantly. In other words, it tries to identify valleys and peaks. However, this approach only identified one of the two populations. This is because the approach starts from an inappropriate null hypothesis for our question. It asks whether the data is adequately explained without a new component. To accept the alternate hypothesis of two components we need to reject the null hypothesis with a large degree of certainty (e.g., $P < 0.05$). The question answered by the MM approach is far less demanding: what is the most likely location and size of the peak given that there are two peaks?

The MM approach is robust to certain types of contamination from nearby clusters (Fig. 1). When using rigid gating, the effects of contamination from other event types can create two types of bias. First, it can alter the apparent abundance of some classes of cells. For example, say there is a nearly uniform background of events that contaminate all populations and we are trying to calculate the relative abundance of one

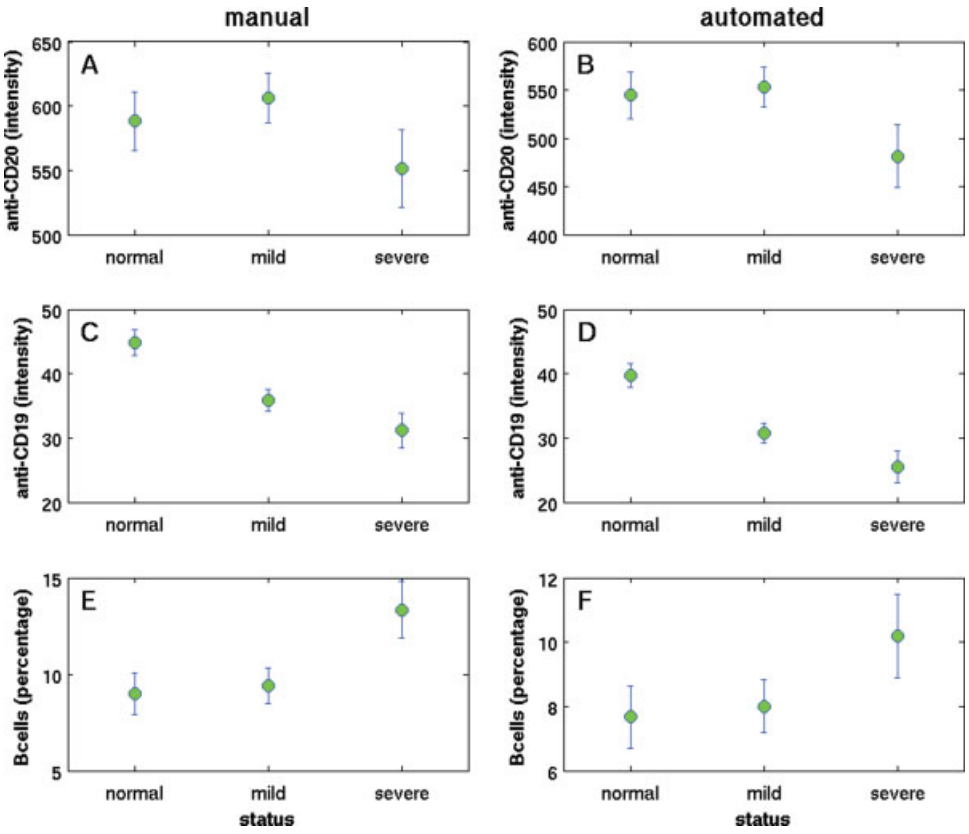

Figure 4. Manual versus automated gating. A comparison manual (left) and automated (right) of (from top to bottom) CD19 fluorescence intensity, CD20 fluorescence intensity and B-cell abundance. The error bars reflect 95% confidence intervals. The overall trends are similar in both methods. The automated method shows significant differences in CD20 and CD19 intensity ($P < 0.05$) between severe SLE and healthy volunteers, whereas the manual method shows significant differences only in CD19 levels. Significant differences in B-cell abundance were found using both methods ($P < 0.05$).

subset. A rigid gating approach would include a constant number of contaminating events in both the numerator and denominator of the estimated bringing the fraction toward one. This attenuation is emphasized when the subset is rare. Another drawback of rigid gates is that they can easily create biased estimates of the intensities by having nonsymmetric inclusion or exclusion properties. For example, drawing a gate that excludes one tail of a distribution of a desirable population in order to avoid another population will create bias. Changes in the position or frequency of the contaminating population will alter the amount of bias and could easily confound analysis of the desired subset. The MM approach can tolerate contaminating or atypical populations without large effects on the desired population. One such example appeared in our analysis of flow cytometry data from an SLE patient (Fig. 3D). Cells from the specimen from this individual exhibited an atypical light-scattering pattern, with a population of cells appearing below the expected position of monocytes. Although this light scatter pattern would trigger reprocessing and analysis for clinical purposes, the MM excluded the unexplained population of cells. Of note, the MM could be coded to exclude, include or separately analyze such events.

Because no gating is formally done, a direct comparison of how events are classified compared to a manual approach is not straightforward. However, we demonstrated that the algorithm can be used to automate analysis of flow cytometry data with a dramatic reduction in time and manual effort with results similar to those from an expert. To show the results were comparable, we looked at CD19 and CD20 in patients with SLE compared to controls. Both manual and automated processes found significantly ($P < 0.05$) more B cells in SLE subjects with SLEDAI >= 8. Both methods also found decreases in the fluorescence intensity of CD19. The automated method also found a significant ($P < 0.05$) reduction in the mean fluorescence levels of CD20 on B-cells in severe lupus subjects, whereas the manual process did not. The manual process did have a nonsignificant ($P > 0.05$) trend in the same direction, and it is likely that statistical significance was not demonstrated because the difference between means was not as large using the manual method (Fig. 5D). We are following up on these and other observations that point toward difference between B cell populations in SLE patients versus normal donors, some of which are already reported in the literature (16).

There are some points about MM in general and about mixtures of Gaussians in particular that are worth discussion. These questions are around the assumptions of normality, sensitivity to starting conditions and how to determine good starting conditions.

For our models we assumed a mixture of Gaussians. This is because the computations are easily done with Gaussians due to the closed form solutions for likelihood and parameter estimates. This may not be a large drawback because many types of data appear to be well approximated by a normal distribution or can be transformed to near normal using a power family of transformations (17). The MM method itself is ap-

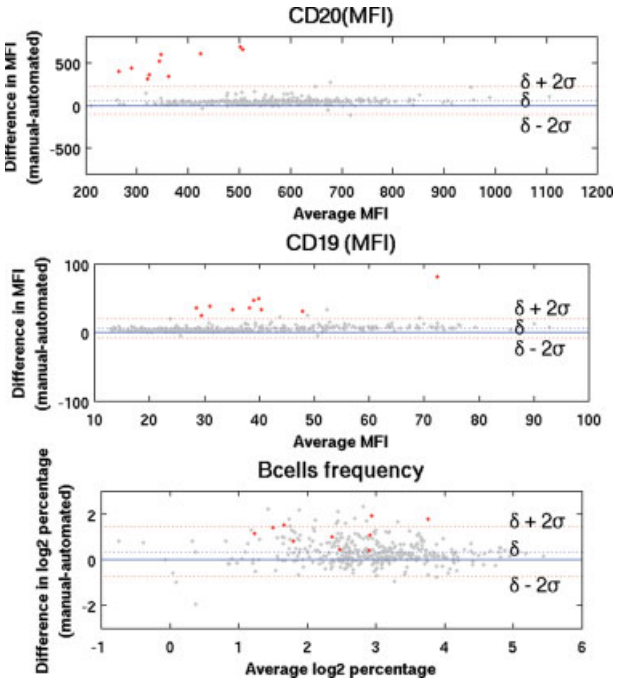

Figure 5. Direct comparison of population estimates. The plots show the estimated mean fluorescence intensity of CD20 and CD19 for the CD20 positive cell population and percentage of B-cells from automated and manual methods. Each point on the graph represents a single sample, the *x*-axis represents the average of the estimates and the *y*-axis shows the difference in the two methods. A dashed blue line represents the average difference, and two dashed red lines shows 2 standard deviations above and below the average difference. The points marked red were flagged automatically during the run as having abnormal MFI compared to expected CD20 positive cells.

plicable to any type of continuous distribution. Although we did not code it, other distributions, if necessary, could be modeled and accommodated with the method. Alternatively, other distributions can themselves be modeled by a mixture of sufficiently many Gaussian distributions. In fact, this occurred in our data. The debris fields and granulocytes were trimmed during data collection so that events from some parts of the intensity spectrum were removed leaving a partial Gaussian distribution. The addition of an extra component was able to model these well enough to produce reasonable estimates of the subset of clinical interest.

Another important concept in the MM approach is choosing the starting conditions. This is analogous in some ways to defining gates but it takes on a larger role in MM. There are two main topics. These are determining the number of populations to model and choosing initial conditions. Different algorithms are expected to vary with their sensitivity to starting conditions. Here we used an EM method, but other methods, such as gradient methods, could be used. Xu and Jordan (18) compared EM to the gradient methods and found that EM is expected to behave well when gradients behave well and may have some advantages in difficult cases where the Hessian matrix is ill-condition.

It is worth noting that linear transformations of the data, for example those produced by compensation, have no effect on the MM approach. More specifically, the configuration of clusters achieved on compensated and uncompensated data differ by the same linear transformation as the data. In addition, the standardized distance of each event to each cluster is the same regardless of the transformation. However, the best choices of initial conditions for the search are obviously different.

The EM algorithm used here requires that the number of populations be defined. However there are published methods for deciding on optimum criterion in an unsupervised way using a variety of methods (for example see (19–21)). In cases where a completely unknown sample is analyzed an unsupervised method may be best, whereas in cases where standard assays are available to subdivide populations of cells is already known it may not be necessary.

The question of sensitivity to starting conditions is a large one. The EM algorithm used here is guaranteed to converge to a locally optimal solution, but not necessarily a globally optimum solution (11). In fact, of the 390 files processed using the automated method we found 10 globally suboptimal final solutions. Readjusting the initial configuration resulted in solutions that produced a clustering pattern more consistent with other samples and prior expectations. Ideally, starting conditions are determined during the development of a particular assay that is robust to between sample variability. This could be done, for example, using the distributions from control samples. However, it is not guaranteed that any initial conditions that will work for all future samples, and therefore, there will be some cases where the optimum configuration of components is not found. In this vein, there are two factors worth considering. One is that, if outliers can be identified in an automated way, then a manual process to find a globally optimal solution could be applied to the few that required it. This would still represent a large time savings compared to the manual the method. Second, there are methods, such as simulated annealing (22) or genetic algorithms (23) or simply trying several starting points (19) that could be incorporated to help ensure that globally optimal solutions were found.

The idea of initial conditions can be useful as a quality control and research tool. Initial conditions can be used as the expected locations of subsets and then tests can be applied to determine the relationship between the initial and final configurations. Final configurations that apparently fit the data but that are different in some way from an expected result can be automatically flagged. The tool developed here identifies and flags three conditions: additive shifts, subset drift, and class swapping. Additive shifts occur when one or more dimension in the final configuration is related to the initial condition through an additive factor. This can happen, for example, if more label is added in a reaction or if different gain settings on the detector are used. Subset drift is when one or more classes have changed positions relative to each other. Class swapping is defined as a swapping of two or more clusters so that the final set of clusters are near different initial population. This can happen when the optimization path for one

cluster takes it toward a different initial cluster and the path for that cluster takes it toward another. The final solution may still be optimal, but the labels may be swapped. The latter did not occur with the analysis here and may be quite rare with reasonable starting conditions. Detection of unexpected final configurations is automated so a researcher can follow up on only the unexpected and potentially interesting results.

The MM approach uses a statistical model where each event is assigned a likelihood that it is from any subset. If desired, this information could be used to design a more efficient gating procedure. For example, the location and size of sorting gates could be calculated to reduce contamination to a predefined level. The gates would be automatically adaptive for new samples and would allow a level of control on gating not currently available.

To implement the assumption that some subsets of cells have some characteristics in common with other subsets, we introduce the idea of a hierarchical constraints to the MM. These constraints apply to some subsets and only in some dimensions. Physically it makes sense to constrain both the means and the covariances. However, there is no guarantee that the constrained covariance matrix will be semipositive definite and so may not be a valid covariance matrix. This can happen, for example, when the subsets being constrained are not sufficiently similar. For this reason, we allow only the means to be constrained. We call these constraints hierarchically constrained class structure to differentiate it from a previous method called hierarchical constrained mixture models (24), which is a fundamentally different kind of constraint applied to all dimensions in the model. The use of hierarchically constrained class structure can produce more desirable results because it restricts solutions to those that are assumed by the researcher to be plausible.

The MM approach offers improved performance at separating subsets which should ease demands on assay development and make new assays possible. Dramatic savings in time and effort have been realized through automated data analysis while achieving results as good as an expert on these data.

Acknowledgments

The authors thank James Chung, Michael Vincent (Executive Directors, Amgen Medical Sciences), Daniel Wallace and Michael Weisman (Cedars-Sinai Medical Center), who conducted the clinical portion of the study and provided whole blood samples to the Amgen Clinical Immunology Cytometry Laboratory for flow cytometry assessments. The authors also thank Brian Kotzin (Vice President, Amgen Medical Sciences), Steve Swanson (Executive Director, Amgen Medical Sciences), and Michael Bass for their support of this project. Finally, the authors thank Igor Fomenko for useful discussions on mixture modeling. This study was funded by and conducted at Amgen Inc., a publicly traded biotechnology company that discovers, develops, and delivers innovative human therapeutics.

Literature Cited

1. Rieseberg M, Kasper C, Reardon KF, Scheper T. Flow cytometry in biotechnology. Appl Microbiol Biotechnol 2001;56:350–360.
2. Nicholson JK, Hubbbar M, Jones BM. Use of CD45 fluorescence and side-scatter characteristics for gating lymphocytes when using the whole blood lysis procedure and flow cytometry. Cytometry 1996;26:16–21.
3. Schenker EL, Hultin LE, Bauer KD, Ferbas J, Margolick JB, Giorgi JV. Evaluation of dual-color flow cytometry immunophenotyping panel in multicenter quality assurance program. Cytometry 1993;14:307–317.
4. Rosa SCD, Herzenberg LA, Herzenberg LA, Roederer M. 11-color, 13-parameter flow cytometry: Identification of human naive T cells by phenotype, function, and T-cell receptor diversity. Nat Med 2001;7:245–248.
5. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. New York: Springer Verlag; 2001.
6. Ferbas J, Wallace DJ, Weisman MH, Belouski SS, Kesslak JP, Vincent M, Zack D, Hendricks L, Chung J. A systematic analysis of circulating B cell populations in healthy and SLE subjects. Ann Rheum Dis 2006;65:584.
7. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. Arthritis Rheum 1992;35:630–640.
8. McCoy JP. Preparation of cells from blood. In: Darzynkiewicz Z, Robinson JP, Crissman HA, editors. Methods in Cell Biology, 3rd ed. San Diego: Academic Press; 2001. p 214.
9. Hultin LE, Matud JL, Giorgi JV. Quantitation of CD38 activation antigen expression on $CD8^+$ T cells in HIV-1 infection using CD4 expression on $CD4^+$ T lymphocytes as a biological calibrator. Cytometry 1998;33:123–132.
10. Seamer LC, Bagwell CB, Barden L, Redelman D, Salzman GC, Wood JC, Murphy RF. Proposed new data file standard for flow cytometry. Version FCS 3.0. Cytometry 1997;28:118–122.
11. Dempster NMLAP, Rubin DB. Maximum likelihood from incomplete data via the EM-algorithm. J R Stat Soc B 1997;39:1–38.
12. Vlassis N, Likas A. A greedy EM algorithm for Gaussian mixture learning. Neural Process Lett 2002;15:77–87.
13. Loken MR, Brosnan JM, Bach BA, Ault KA. Establishing optimal lymphocyte gates for immunophenotyping by flow cytometry. Cytometry 1990;11:453–459.
14. Costa ES, Arroyo ME, Pedreira CE, Garcia-Marcos MA, Tabernero MD, Almeida J, Orfao A. A new automated flow cytometry data analysis approach for the diagnostic screening of neoplastic B-cell disorders in peripheral blood samples with absolute lymphocytosis. Leukemia 2006;20:1221–1230.
15. Chaudhuri P, Marron J. SiZer for exploration of structures of curves. J Acoust Soc Am 1999;94:807–823.
16. Jacobi AM, Odendahl M, Reiter K, Bruns A, Burmester GR, Radbruch A, et al. Correlation between circulating $CD27^{high}$ plasma cells and disease activity in patients with systemic lupus erythematosus. Arthritis Rheum 2003;48:1332–1342.
17. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc B 1964;26:211–252.
18. Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Comput 1996;8:129–151.
19. Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput Stat Data Anal 2003;41:561–575.
20. Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Machine Intell 2002:24;351–396.
21. Verbeek JJ, Vlassis N, Kröse BJA. Efficient greedy learning of Gaussian mixture models. Neural Comput 2003;15:469–485.
22. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220:671–680.
23. Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Boston, MA: Addison-Wesley Professional; 1989.
24. Titsias MK, Likas A. Mixture of experts classification using a hierarchical mixture model. Neural Comput 2002;14:2221–2244.