BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 23 2014, pages 3372–3378
doi:10.1093/bioinformatics/btu575

Data and text mining

Advance Access publication August 27, 2014

Cell population identification using fluorescence-minus-one controls with a one-class classifying algorithm

Kristen Feher*, Jenny Kirsch, Andreas Radbruch, Hyun-Dong Chang and Toralf Kaiser*
Deutsches Rheuma-Forschungszentrum, Berlin 10117, Germany

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: The tried and true approach of flow cytometry data analysis is to manually gate on each biomarker separately, which is feasible for a small number of biomarkers, e.g. less than five. However, this rapidly becomes confusing as the number of biomarker increases. Furthermore, multivariate structure is not taken into account. Recently, automated gating algorithms have been implemented, all of which rely on unsupervised learning methodology. However, all unsupervised learning outputs suffer the same difficulties in validation in the absence of external knowledge, regardless of application domain.

Results: We present a new semi-automated algorithm for population discovery that is based on comparison to fluorescence-minus-one controls, thus transferring the problem into that of one-class classification, as opposed to being an unsupervised learning problem. The novel one-class classification algorithm is based on common principal components and can accommodate complex mixtures of multivariate densities. Computational time is short, and the simple nature of the calculations means the algorithm can easily be adapted to process large numbers of cells (10^6). Furthermore, we are able to find rare cell populations as well as populations with low biomarker concentration, both of which are inherently hard to do in an unsupervised learning context without prior knowledge of the samples' composition.

Availability and implementation: R scripts are available via https://fccf.mpiib-berlin.mpg.de/daten/drfz/bioinformatics/with{user name,password} = {bioinformatics,Sar = Gac4}.

Contact: kristen.feher@drfz.de or kaiser@drfz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 12, 2014; revised on July 14, 2014; accepted on August 19, 2014

1 INTRODUCTION

The mainstream method of flow cytometry data analysis is to 'manual hierarchical gate'. In other words, given N observations (cells) of a p-dimensional random vector, where each $i \in \{1, \ldots, p\}$ represents a biomarker coupled to a fluorescent marker, and f_i is the marginal density of each biomarker, the cells are sequentially split into ≥ 2 groups corresponding to presence, absence, and possibly discretized biomarker concentration levels following visual inspection of each f_i. Biomarkers are most typically antibodies; however, they can also be fluorescent proteins, among others.

Gating order is based on prior biological knowledge: e.g. only a particular cell type can be positive for certain biomarkers and not others; therefore, the general strategy is to gate on highly specific biomarkers early to separate broad categories of cells. Following this, separate gating strategies may be applied to find the nuances within the broad categories. This has been a successful approach over the past three decades, where it has typically been applied to experiments comprising less than five biomarkers. With increasing numbers of biomarkers and/or novel biomarkers, manual gating becomes difficult, as well as neglecting the multivariate structure (Chattopadhyay *et al.*, 2008; Baumgarth, 2000).

Compounding these difficulties is the impact of data pre-processing on gating, namely 'compensation'. Each biomarker is coupled to a fluorescent marker that is excited by the appropriate laser, and in turn, it emits light over a typical range of 400–800 nm. Each biomarker is measured with an optical bandpass filter (BP) tuned to its peak, e.g. the intensity of light between 600 and 630 nm is measured. However, the spectra of different fluorescent markers overlap; thus, measurement of total light captured in a BP has contributions from up to p dyes. This is known as 'spectral mixing', and its theoretical form is simply a linear transform from 'biomarker space' to 'spectral space'. Compensation is simply the reverse linear transformation back to the biomarker space, where the features have a direct biological meaning as they correspond directly to the biomarkers.

However, compensation is based on the assumption of Gaussian distributed error, and thus distorts the data owing to the underlying Poisson distribution of the emitted photons (Roederer, 2001; Novo *et al.*, 2013). Additionally, it cannot account for autofluorescence (AF) (Aubin, 1979), which occurs when a cell falsely appears to have a biomarker concentration above background, which is simply because of the cell's properties. Prior knowledge is required to account for AF, and if not accounted for, can lead to false conclusions in manual gating. Additionally, in the case of antibodies, the measurement can be confounded by unspecific binding (UB), for which there is no correction.

In fact, these are such pervasive problems that a specific type of biological control has been designed for their detection: the fluorescence-minus-one (FMO) control (Roederer, 2002). For each biomarker $i \in \{1, \ldots, p\}$, a new sample is measured with all biomarkers minus the i^{th} biomarker. The cells that are positive for the i^{th} biomarker shift with magnitude proportional to the original biomarker concentration, and no shift occurs

*To whom correspondence should be addressed.

when the cells are negative. The effects of AF and UB are minimized, as they should be unchanged between the fully stained sample (full staining) and the FMO. Superposing the full staining with each FMO is the multivariate spectral analogue of univariate gating on the biomarker marginal densities, while simultaneously taking AF and UB into account. The use of FMOs is the gold standard in experimental design, however, their use has been restricted to visual inspection thus far, to confirm that the results of manual hierarchical gating is unlikely to contain artifacts.

Here we use FMOs as training sets in a one-class classifier to systematically identify cell populations in a semi-automated fashion. By definition of the spectral mixing equations, the FMOs are training sets of negative examples for each BM. The full staining (test set) is compared with the FMOs to determine positive cells for each BM, as given by those that lie outside the FMOs (Tax, 2001). The output of all the one-class classifiers is combined to yield the final populations. Furthermore, we can use the physical constraints to improve sensitivity of positive cell detection, namely, that (i) cells are not independent events but rather they are correlated; (ii) the composition of cells is preserved between the full staining and FMO; and (iii) the cells that are negative for a given BM have the same covariance in both the full staining and FMO. Additionally, we apply the algorithm iteratively to remove populations that mask others.

This gives the following statistical and technical advantages. Populations are found in direct relation to a biological control, thus avoiding unsupervised learning. As larger populations are successively removed, it becomes easier to find smaller populations even without prior knowledge. Furthermore, no prior knowledge is required as to the form of AF and UB. Given that population identification occurs directly in spectral space, spectral overlap no longer needs to be minimized to such a great extent, giving more flexibility in fluorescent marker selection, and experiments that are currently difficult are possible on standard instruments, without requiring the purchase of specialized equipment.

All flow cytometry applications suffer from contamination arising from various sources: 'junk' (e.g. debris, dead cells that were not gated out, doublets, bubbles in the sheath fluid and unbound fluorescent molecules), altered fluorescence (e.g. unspecific staining, bleaching and FRET) or instrumentation (e.g. electronic noise, laser fluctuations, optical properties and variation in flow speed), among others. It can be hard to distinguish rare populations from contamination, and in traditional manual hierarchical gating, an operator would probably ignore small percentages, with the exact threshold depending on the total number of cells collected, the experience of the operator and the biological question at hand. In fact, if rare cells are indeed the focus of the biological question, then different experimental approaches would almost certainly have been taken in the first place, such as magnet pre-enrichment of the rare cell population (Bacher, 2013). For these reasons, the full algorithm is quite deliberately only semi-automated in nature, and this is based on the fact that successive iterations find increasingly smaller groups of positive cells. The issues listed above are too numerous and complex to be solved by an algorithm. The approach we therefore take is to let the operator decide the minimum population size, and return a list of putative populations for further human interpretation.

The article is organized as follows. The algorithm is described in Section 2. In Section 3, the algorithm is tested on simulated data to demonstrate its increased sensitivity over alternatives. Selected clustering methods are also applied to the simulated data, to demonstrate that it is difficult to find the correct clustering in the absence of prior or external knowledge. The clustering methods discussed were evaluated in FlowCAP (Aghaeeour, 2013) and are available via Bioconductor. Finally, the algorithm is applied to a four-marker staining, and the output is confirmed with manual hierarchical gating. Selected clustering methods are also applied but they fail to find the small populations.

2 METHODS

2.1 Sample preparation and data acquisition

Blood collection and staining Human blood was obtained from a healthy donor and incubated with erythrocyte lysis buffer (Qiagen, Hilden, Germany) according to the manufacturer instructions to remove the red blood cells. Cells were centrifuged (10 min, 500 g) and washed twice with PBS/BSA/EDTA. Absolute cell numbers were counted by the CASY®cell counter (Schärfe System, Reutlingen, Germany). Total leukocytes were stained on ice for 15 min in the dark with optimal concentration (as determined by titration, data not shown) of fluorescent-conjugated anti-human antibodies: CD14 PE-Alexa Fluor 700 (BP 720/30) (Invitrogen, Darmstadt, Germany) and CD3 PerCP (BP 670/14), CD4 FITC (BP 520/30) and CD8 PE-Cy7 (BP 780/60) (all obtained from Biolegend, San Diego, USA). After staining, the cells were centrifuged (10 min, 400 g) and resuspended in PBS/BSA/EDTA. Additionally, FMO controls were prepared for each cell marker. For dead cell exclusion, 1 μl DAPI (2 μg/ml) was added to each sample.

Flow Cytometry Data were collected on a LSRII (BD, San Jose, USA) equipped with four fixed-aligned 355, 405, 488 and 633 nm lasers. The BD Cytometer setup and tracking application (CST) was used to determine the optimal baseline PMT voltage for each fluorescent channel. In all, 50 000 events were acquired for each sample (gated on lymphocytes/monocytes and live cells). Data were collected by using the BD FACSDiva software (version 6.1.3) and saved as FCS 3.0 data files. The data analysis including sequential gating was performed by using FlowJo 9.8 (Tree Star, USA).

2.2 Hardware and software

All calculations (apart from those in FlowJo) were performed on a MacBook Pro, OS X 10.8.5, 3 GHz Intel Core i7 Processor, and 8 GB 1600 MHz DDR3 Memory.

All calculations were performed with R (R Core Team, 2013) version 3.0.2 (September 25, 2013), Platform: x86_64-apple-darwin10.8.0 (64-bit).

Data were pre-processed in R core team using Bioconductor package `flowCore` (Hahne, 2009). All data were logicle transformed with $w = 1$, and live-cell/monocyte/lymphocyte gated on forward scatter, side scatter and DAPI parameters.

2.3 Formal definition

There is a set of p biomarkers $\{\alpha_1, \ldots, \alpha_i, \ldots, \alpha_p\}$ that forms the basis vectors for the 'biomarker space' ($\mathbb{R}^p$). Each cell C'_n, $1 \leq n \leq N$ is a location in $\mathbb{R}^p$ defined by a linear combination of α_i: $C'_n = C'_{1n}\alpha_1 + \ldots + C'_{pn}\alpha_p$.

There is a set of p BPs $\{\beta_1, \ldots, \beta_i, \ldots, \beta_p\}$ that forms the basis vectors for 'spectral space', also $\mathbb{R}^p$. Each biomarker can be expressed as a linear combination of BPs in spectral space, corresponding to the proportion of its spectrums light falling in each BP (i.e. spectral mixing): $\alpha_i = B_{i1}\beta_1 + B_{12}\beta_2 + \ldots + B_{ip}\beta_p$

Each cell $C_n = (C_{1n}, \cdots, C_{pn})$ is a location in $\mathbb{R}^p$ (spectral space) and is thus defined as follows:

$$\begin{pmatrix} C_{1n} \\ \vdots \\ C_{pn} \end{pmatrix} = \begin{pmatrix} B_{11} & \cdots & B_{p1} \\ \vdots & \ddots & \vdots \\ B_{1p} & \cdots & B_{pp} \end{pmatrix} \begin{pmatrix} C'_{1n} \\ \vdots \\ C'_{pn} \end{pmatrix}$$

The α_is are not orthogonal in spectral space, and their exact arrangement corresponds to spectral overlap. Using the full complement of antibodies is known as a 'full staining'.

The location of C_n in spectral space in the i^{th} FMO is defined as follows:

$$\begin{pmatrix} C_{1n} \\ \vdots \\ C_{pn} \end{pmatrix}_{FMO_i} = \left(\begin{pmatrix} B_{11} & \cdots & B_{p1} \\ \vdots & \ddots & \vdots \\ B_{1p} & \cdots & B_{pp} \end{pmatrix} - \begin{pmatrix} \mathbf{0} & B_{i1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & B_{ip} & \mathbf{0} \end{pmatrix} \right) \begin{pmatrix} C'_{1n} \\ \vdots \\ C'_{pn} \end{pmatrix},$$

i.e. the cell shifts in a direction defined by $-\alpha_i$ and proportional to C_{in}.

All C'_{in} and all B_{ii} must be ≥ 0; thus $C_{in} \geq 0$. A shift of $-\alpha_i$ is towards the origin, and thus, less total light is collected for positive cells in the FMO compared with the full staining. Thus, positive cells are 'outside' the FMO, and further from the origin. This fact forms the basis for the application of common principal components (CPC). This theoretical description of spectral mixing does not take AF or UB into account.

2.4 Empirical histogram

Allocating full staining cells as positive/negative is done by comparison of the distribution of their Mahalanobis distances from the FMO to that of the cells within the FMO. We can refine simple thresholding by incorporating the physical restraints: cells are correlated, cell composition is preserved and negative cells have the identical covariance in the full staining and FMO. Thus, when there are many full staining cells occurring in the tail of the FMO distance distribution, it is not possible that they are all negative. It is desired to detect cells that fall outside the FMO with respect to both distance and frequency.

Given a set of observations $\{x_n : 1 \leq n \leq N\}$ of random variable $\mathcal{X}$, the empirical cumulative density function is defined as follows:

$$\hat{F}(y) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{x_n \leq y\},$$

where $\mathbf{1}$ is the indicator function. The concentration of observations with respect to y is encoded in the gradient of F, the inverse of which is given by $\Delta x_n = x'_n - x'_{n-1}$, where $\{x'_n\}$ is $\{x_n\}$ after permutation such that $x'_n \geq x_{n-1}{}'$ and $x'_n \leq x_{n+1}{}' \forall n \in 2 \ldots N-1$.

The empirical histogram (EH) H is given by $\{(x_n, \Delta x_n) : n = 1, \ldots, N\}$. Low values of Δx_n correspond to a region of high probability of $\mathcal{X}$ occurring, and high values correspond to low probability, either the tails of a distribution or regions between modes of a multi-modal distribution.

Define data matrices X_a and X_b, of respectively N_a and N_b rows (observations) and p columns, and both with some fraction of observations that share a common covariance. Then, H can be used determine those datapoints unlikely to arise from the common covariance. This method is not proposed as an alternative in more standard contexts that rely on Mahalanobis distance (e.g. outlier detection), as it directly depends on a common fraction for its success.

The criteria for rejecting a datapoint in X_b for not arising from the density of X_a are as follows. For $A = a, b$, define

$$M_A = \{M_{n_A}(X_{n_A}, \mu_a, \Sigma_a) : n_A = 1, \ldots, N_A\} \quad (1)$$

as the set of Mahalanobis distances of each observation in either X_a or X_b, with respect to the p-dimensional multivariate mean μ_a and covariance Σ_a of X_a. Thus,

$$H_A = \{(M_{n_A}, \Delta M_{n_A}) : n_A = 1, \ldots, N_A\} \quad (2)$$

is the EH corresponding to X_A with respect to X_a. Let $M_A(t) = \{M_{n_A} : \Delta M_{n_A} = t\}$ and $m_{\max}(t) = Q_q(M_a(t))$ where $Q_q(W)$ is the q^{th} quantile of W. The observations X_{b+} that do not arise from the density of X_a are thus defined by the following:

$$X_{b+} = \{x_{n_b} : M_b(t) > m_{\max}(t), \forall t \in [\min(\Delta M_{n_b}), \max(\Delta M_{n_b})]\} \quad (3)$$

In practice, this is done by partitioning based on quantiles of ΔM_{n_a}, i.e. for $q \in (q_1, q_2, \ldots, q_r)$ and redefining M_A as follows:

$$M_A(q_j) = \{M_{n_A} : \Delta M_{n_A} > Q_{q_j}(\Delta M_{n_a}) \text{ and } \Delta M_{n_A} \leq Q_{q_{j+1}}(\Delta M_{n_a})\}$$

for $j = 1, \ldots, r-1$, as well as end sets:

$$M_{A\text{low}} = \{M_{n_A} : \Delta M_{n_A} \leq Q_{q_1}(\Delta M_{n_a})\}$$

$$M_{A\text{high}} = \{M_{n_A} : \Delta M_{n_A} > Q_{q_r}(\Delta M_{n_a})\}.$$

This is graphically depicted in the Appendix.

2.5 Separation-improving dimension reduction via CPC

The aim of CPC (Flury, 1983) is to find a linear combination of features $\mathbf{z}$ that maximize the ratio of variances of populations a and b with respective population covariances Σ_a and Σ_b, i.e. to maximize the function:

$$h(\mathbf{z}) = \frac{\mathbf{z}^T \Sigma_b \mathbf{z}}{\mathbf{z}^T \Sigma_a \mathbf{z}}, \; \mathbf{z} \in \mathbb{R}^p.$$

The solution of $h(\mathbf{z})$ corresponds to the eigensolution of $\Sigma_a^{-1}\Sigma_b$. An alternative interpretation of CPC is to find linear combinations along which Mahalanobis distance for X_a and X_b is as different as possible, and Σ_a and Σ_b are the respective sample covariances. Define $v_i(X_a, X_b)$ as the i^{th} eigenvalue of $\Sigma_a^{-1}\Sigma_b$, and $V_i(X_a, X_b)$ as its i^{th} eigenvector.

In the case when portions of X_a and X_b have common covariance, and covariance difference is caused by shifting populations, CPC finds the projection that contains maximum separation between the shifted population and common portions. By the spectral mixing equations, all separation should occur along one direction in the absence of noise. Here we take the top two components to account for variation.

2.6 Expansion to polynomial space

Even after CPC-based dimension reduction, the separating boundary can still take on a complex shape. Typically, most separation has been found to occur in the first two eigenvectors of $\Sigma_a^{-1}\Sigma_b$. Thus, it is feasible to perform a polynomial basis expansion directly (e.g. $\mathbb{R}^2 \mapsto \mathbb{R}^9$ for cubic expansion). This has the advantage of avoiding tuning additional parameters, as is necessary in the proper use of support vector machines (SVM) (Chapelle, 2002).

Here, we apply the following expansion:

$$(x, y) \mapsto (x, y, x^3, y^3, x^2y, xy^2, x^2, y^2, xy) \quad (4)$$

A polynomial expansion of order >3 often leads to singular covariance because of the resulting features spanning over orders of magnitudes. Rescaling in such cases leads to complex-valued solutions of CPC.

2.7 One-class classifying algorithm

The algorithm to discover each successive population is as follows. Define a reference dataset X_{ref} and the comparison dataset X_{comp} (assuming a common covariance is shared among some fraction of the datapoints). The following separating algorithms find the positive datapoints of X_{comp} and are the cornerstones of the full algorithm:

Separating algorithm The operator CPC_proj($X_{\mathrm{ref}}, X_{\mathrm{comp}}$) is defined by the following steps:

1. Define projection $\mathrm{Proj} = \left[V_1(X_{\mathrm{ref}}, X_{\mathrm{comp}})V_2(X_{\mathrm{ref}}, X_{\mathrm{comp}})\right]$, where columns are the CPC eigenvectors.
2. Transform data: $X'_{\mathrm{ref}} = (X_{\mathrm{ref}})(\mathrm{Proj})$ and $X'_{\mathrm{comp}} = (X_{\mathrm{comp}})(\mathrm{Proj})$
3. Define Mahalanobis distance EHs using $\log(M(X'_{\mathrm{ref}}, X'_{\mathrm{ref}}))$ and $\log(M(X'_{\mathrm{comp}}, X'_{\mathrm{ref}}))$ (Eq. 1 and 2).
4. The positive datapoints $X_{\mathrm{comp}+}$ of X_{comp} are given by Eq. 3.
5. Return $X_{\mathrm{comp}+}$.

Polynomial expanded separating algorithm The operator CPC_proj_poly($X_{\mathrm{ref}}, X_{\mathrm{comp}}$) is defined by the following steps:

1. Steps 1 and 2 as for separating algorithm above.
2. Polynomial expand each row of X'_{ref} and X'_{comp} according to Equation 4 to yield X''_{ref} and X''_{comp}
3. Define projection $\mathrm{Proj}' = \left[V_1(X''_{\mathrm{comp}}, X''_{\mathrm{ref}})V_p(X''_{\mathrm{comp}}, X''_{\mathrm{ref}})\right]$, where columns are the CPC eigenvectors.
4. Transform data: $X'''_{\mathrm{ref}} = (X''_{\mathrm{ref}})(\mathrm{Proj}')$ and $X''_{\mathrm{comp}} = (X''_{\mathrm{comp}})(\mathrm{Proj}')$
5. Define Mahalanobis distance EHs using $\log(M(X'''_{\mathrm{ref}}, X'''_{\mathrm{ref}}))$ and $\log(M(X'''_{\mathrm{comp}}, X'''_{\mathrm{ref}}))$ (Equations 1 and 2).
6. The positive datapoints $X_{\mathrm{comp}+}$ of X_{comp} are given by Equation 3.
7. Return $X_{\mathrm{comp}+}$.

Main algorithm Denote the full staining dataset as X_2 and the i^{th} FMO as $X_1(i)$. There are N_{tot} cells in X_2.

- Start: set $X'_2 \leftarrow X_2$.
- While $N_{\mathbb{P}} > T \in (0, 1)$:
  1. For each i:
     - (a) Filter positive cells out of FMO: CPC_proj($X_1(i), X'_2$)
     - (b) $X_{1-}(i) \leftarrow X_1(i) \setminus X_{1+}(i)$, where $S_1 \setminus S_2$ is the relative complement of S_2 in S_1, i.e. the set of elements in S_1 but not in S_2.
     - (c) Determine positive cells in full staining: CPC_proj_poly($X'_2, X_{1-}(i)$)
     - (d) For each cell n_2 in X'_2, define a binary variable V indicating if it is positive or negative for BM i: $V_{n_2}(i) = 0$ if $x_{n_2} \in X'_{2+}(i)$ and $V_{n_2}(i) = 1$ if $x_{n_2} \notin X'_{2+}(i)$.
     - (e) return V(i)
  2. Combine all full staining/FMO comparisons. Define the set of all possible staining patterns as follows:
     $$S = \{(S_1, \ldots, S_p) : S_1 \in \{0, 1\}, \ldots, S_p \in \{0, 1\}\}$$
  3. Define the population corresponding to $s \in S$ as follows:
     $$\mathbb{P}' = \{x_{n_2} : V_{n_2}(1) = s_1, \ldots, V_{n_2}(p) = s_p\}$$
  4. Define the next population as follows:
     $$\mathbb{P} = \max_{s \in S} |\mathbb{P}'|$$
     where $|\cdot|$ corresponds to the cardinality of a set.
  5. Remove $\mathbb{P}$ and update X'_2:
     $$X'_2 \leftarrow X_2{}' \backslash \mathbb{P},$$
  6. Return $N_{\mathbb{P}} = |\mathbb{P}|/N_{\mathrm{tot}}, \mathbb{P}, X'_2$.

In short, the main algorithm is repeatedly applied while the percentage of total cells $N_{\mathbb{P}}$ found at each iteration is $> T$.

2.8 Simulated data

Spectral mixing equations in Section 2.3 are used to generate artificial full stainings and FMOs. First, a binary full staining matrix with populations as rows and BMs as columns is generated using the binomial distribution. If a population is negative, then its average intensity is uniformly distributed on [0, 1]. If the cell is positive, its average intensity is exponentially distributed on $(1, +\infty)$. After population average intensities are established, cells are generated from a multivariate normal distribution with variance of each BM set to $\exp(-\mu)/5$ where μ is the BM's average intensity, and zero covariance. The FMO corresponds to one column of the binary staining matrix being set to zero. Finally, the full staining and the FMO are multiplied by the spectral mixing matrix. Full details can be found in the Appendix.

3 RESULTS

3.1 Population masking in CPC

When a mixture of populations exists, it is possible that a small population is masked by a larger one in CPC. It is impossible to enumerate all conditions under which this can occur; however, we give one specific example to demonstrate the possibility, with all details in the Appendix. In real flow cytometry data, imbalanced population sizes are to be expected, and it is likely that some will be masked. In the proposed algorithm, this situation is thus accounted for by successively removing large populations and reapplying CPC.

3.2 Simulations

We demonstrate how the proposed algorithm works by applying to simulated data. Data are simulated by using a binary population staining matrix as a seed, whose elements are binomially distributed. The generated data are spectrally mixed, as given by Section 2.3. The binary staining matrix is also used to generate a seed for a corresponding FMO. Full simulations details can be found in Section 2.8 and Appendix. Randomly generated full staining–FMO pairs are used to demonstrate that the novel combination of the algorithm's components improve the detection of positive cells compared with Mahalanobis distance thresholding, while remaining conservative.

As the foundation of the algorithm is a one-class classifier, the simulations are judged by true-positive (TP) and false-positive (FP) rates, where 'positive' simultaneously means 'outside the FMO' and 'positively stained for BM'. Formally, let each cell have a true class that is either positive (P) or negative (N): $C_i \in \{P, N\}$ and an estimated class $\hat{C}_i \in \{P', N'\}$ for $i = 1, \ldots, n$. The TP rate is $\mathrm{TP} = |\{\{C_i, \hat{C}_i\} : C_i = P, \hat{C}_i = P'\}|/|\{C_i : C_i = P\}|$, where $|\cdot|$ is the cardinality of a set. Similarly, the FP rate is $\mathrm{FP} = |\{\{C_i, \hat{C}_i\} : C_i = N, \hat{C}_i = P'\}|/|\{C_i : C_i = N\}|$. We present the results in ROC space and as such they are insensitive to an imbalanced number of positive and negative cells (Fawcett, 2006).

Four methods are tested: Mahalanobis distance thresholding, EH, CPC + EH, CPC + polynomial expansion + EH, demonstrating that the last method is the most advantageous. All graphical output is included in the Appendix. Here we supply a summary of the key results; a more detailed discussion is also included in the Appendix. Over all the methods, when TP rates decrease the following increase: p, number of positive populations and population variance. In other words, it is harder to detect positive cells as complexity increases, and can also be made easy (in silico) by ensuring population variance is small. Using Mahalanobis distance is conservative (few FP rates), but given its assumption of a single multivariate normal population, TP rates are low in the presence of a mixture of populations. TP rates are greatly improved with CPC + polynomial + EH while maintaining low FP rates. The TP rates are only slightly improved compared with CPC + EH; however, polynomial expansion is crucial to obtain FP rates comparable with Mahalanobis distance thresholding.

Concluding, this algorithm can successfully deal with complex mixtures of multivariate populations with non-constant variance in single full staining/FMO comparisons. Given that the cells are correlated, combining the output from multiple comparisons strengthens this (e.g. in the case when a population is positive for two BMs but only weakly so for one). Iteratively removing larger populations makes smaller and possibly masked populations apparent, and thus, the entire algorithm can deal with complex mixtures of populations in a computationally tractable manner.

3.3 Simulations with SVM-based one-class classifier

We now characterize the performance of SVM-based one-class classifiers (Schölkopf, 2000), with full details in the Appendix. A polynomial kernel of degree 3 is chosen to be comparable with our algorithm. The first thing to note is that in this context computational time is quadratic in the number of cells (via simulations, not shown), and $N_C = 10\,000$ was chosen. The dominant parameter is $\nu \in (0, 1)$, which controls FP rate. To summarize, as ν increases from 0 to 1, so do both the TP and FP rates, and there is no parameter setting that matches our algorithm. We do not pursue other kernels as computational time is likely to be prohibitive for real data (10^5 is not a large number of cells in an experiment).

3.4 Simulations with clustering methods

All unsupervised learning output is difficult to evaluate in the absence of prior knowledge (Hastie, 2009), and a specific concern is the choice of number of clusters K. At least two types of clustering have been applied to flow cytometry thus far: partitioning [e.g. flowMeans (Aghaeepour, 2011)] and model-based clustering [e.g. flowClust (Lo, 2009)], among others.

flowMeans flowMeans, a K-means variant with automatic choice of K, is applied to simulated data. It is evaluated via the number of clusters and cluster agreement (adjusted Rand index R). It is applied to simulated data, with details included in the Appendix. The simulated data contain 10 clusters with sizes ranging from ∼0.5 to ∼30% of the total number of cells, and additionally with the three smallest clusters removed. The presence of the small clusters gives rise to a greater range of K, as well as decreasing R and thus it appears that the local density influences clustering outcome, and that a more uniform density improves the clustering output. To test this, a new simulation was performed where all clusters were of equal size. However, K appears to be generally underestimated, and the cluster agreement is poor. The clustering could potentially be retrieved with a different choice of K, but this is difficult to know in advance. In summary, it appears that the data's characteristics (local density and population sizes) influence the outcome of a clustering algorithm, and these characteristics are not known in advance.

Model-based clustering Owing to the lengthy computational time, extensive simulations of density-based clustering are not performed. However, model-based clustering based on finite normal mixture modeling [R package `mclust` (Fraley, 2012)] was applied a small number of times and visually inspected. The clustering is almost perfectly recovered when K is known in advance. However, even in this ideal case, it is difficult to choose K without prior or additional knowledge. In practice, flow cytometry data are not multivariate normal and variants of mixture modeling have been developed [flowClust, FLAME (Pyne, 2009)]. It is not informative to apply these methods here, as the simulated data are multivariate normal.

3.5 Application to data

We apply our algorithm to a simple four-marker staining (Materials and Methods), with the terminating condition T being set to 0.001 (see 'Main Algorithm' in Section 2). Some very small 'populations' were visually judged to be artefacts, as their scatter is too high. First, the full staining is superposed with each FMO in Figure 1. Example CPC projections are included in the Appendix. The populations are summarized in Table 1 and

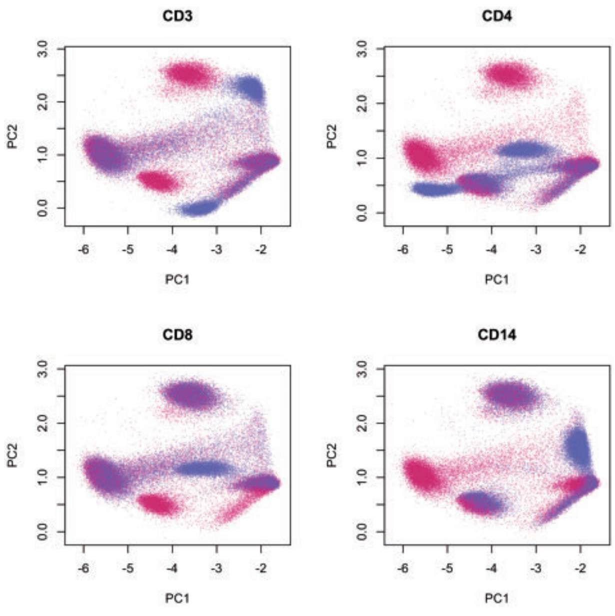

Fig. 1. The full staining is superposed with each FMO; shown here in a projection defined by PCA is the full staining. The full staining cells are red; the FMOs cells are blue

Table 1. Populations found by comparison to FMOs

Population	Proportion
CD4-CD14	0.335401270
CD3-CD4	0.179984676
unstained	0.1723967
CD3-CD8	0.154304357
CD14	0.067475716
CD8	0.059615907
CD4	0.015126424
CD3-CD4-CD8	0.010331447
CD3	0.005363454

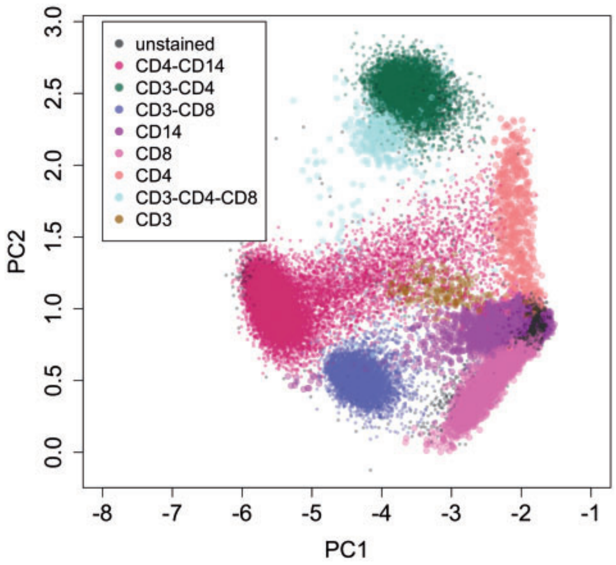

Fig. 2. The final populations. The cells of the small populations are plotted with larger circles for visibility

displayed in Figure 2, and are confirmed with hierarchical manual gating (FlowJo output in Appendix). There is a variety of population shapes. The CD3-CD4 and CD3-CD8 populations are compact and regular. In contrast, the CD4-CD14 population comprises of a dense core and a less dense tail. The unstained population is not 'stand-alone', but rather the single positive populations start directly from its boundaries (i.e. at the detection limit). The CD3-CD4-CD8 population is only weakly positive for CD8, and thus it is directly adjacent to the CD3-CD4 population. Despite the heterogeneity of the various population densities, and the lack of a clear separation between certain populations, they are all detectable using FMOs. Finally, these results were found in <40 s, but conceivably could be less when run in parallel. The algorithm is additionally applied to an unusual staining with a high degree of spectral overlap, which is difficult to compensate, with results included in the Appendix.

3.6 Comparison of algorithm to unsupervised methods

We compare the output of our algorithm to flowMeans, flowClust and SPADE. All output is in the Appendix.

flowMeans flowMeans is applied to both uncompensated and compensated data, with K being chosen as 6 and 4 respectively. In both cases, the major features of the data are retrieved [CD3-CD4, CD3-CD8, CD4-CD14 (dense), unstained], with the single positive populations tending to be subsumed into the unstained population. In the uncompensated data, CD4-CD14 is split into high density and low density portions, and the single CD3 population is also retrieved. Computational time is <10 s for each.

flowMeans was further applied on each cluster of the uncompensated data (excepting the single 3 cluster), to see if it could iteratively find the small populations. However, for each, K was found to be 1 and no further splits were possible.

flowClust flowClust is applied to both uncompensated and compensated data, calculated with $K = 1, \dots, 11$. In both cases, the Bayesian Information Criterion stops sharply increasing after K = 4. The exact clustering depends on the outlier identification quantile, which is effectively an additional parameter to tune. Both find similar results [CD3-CD4, CD3-CD8, CD4-CD14 (dense), unstained]. The CD4-CD14 population has its tail that approaches the unstained cells being designated as outliers when the quantile is set to 95%. Increasing K fails to find the small populations, but starts artificially splitting populations. Computational time is <15 min for each, when parallel computation is used.

SPADE SPADE (Qiu, 2011) is applied to compensated data alone, as it is ill-defined for uncompensated data. Downsampling target percentiles are 0.05 and 0.005. On visual inspection of the output trees, both percentiles appear to find the major features of the data (CD3-CD4, CD3-CD8, CD4-CD14, unstained). However, it is not clear how to choose the optimal down-sampling percentile, nor is there a systematic way to synthesize output across different trees. Its main purpose seems to be an exploration tool.

Summary The three methods find the four big populations of the data. However, they all require parameter tuning, which is difficult in the absence of prior knowledge. Additionally, it appears to be difficult to find small populations in the absence of prior knowledge.

4 DISCUSSION

We have demonstrated how FMOs can be compared with full stainings with a one-class classification approach to systematically discover the sample's composition of cell populations. The proposed algorithm requires no tuning of parameters, and each full staining-FMO comparison (CPC_proj_poly) is linear in the number of cells, and easily parallelizable. The cell populations found were heterogenous in terms of cell number, ranging from 0.05 to 30% of the total. When attempting to discover cell populations via unsupervised methods, the four largest populations masked the smaller ones, and it was not possible to find them using the investigated methods without further modification. It could be argued that this problem is easily avoided by first down-sampling (as is done in SPADE) to approximate a uniform density; however, the degree of down-sampling requires parameter tuning to ensure the smallest populations are not missed, and it is unclear how to do so without prior knowledge of the sample's composition. These difficulties are the manifestation of the general problem of unsupervised learning output validation, which is not possible in the absence of external information. When widely used biomarkers are applied, this problem is not so severe, as output can be compared against accumulated knowledge, but as novel biomarkers are increasingly used, the appropriate *post hoc* controls must be made, e.g. cell sorting and further tests. Alternatively, FMOs can be made at the outset. Additionally, it was noted in this data that many populations that we found (CD4-CD14, all single positive populations) are in fact on a continuum with long tails stretching down towards the unstained population. In other words, there is no distinct cut between stained and unstained, and if this is in fact a common feature of flow cytometry data, it would suggest that it is inherently ill-suited to being clustered.

Unspecific contamination is an issue common to all flow cytometry applications, and we do not attempt to solve this here. As such, the algorithm is presented as being semi-automated. The major problem is to decide whether a 'rare' population is a true population or an artefact caused by contamination. As discussed in Section 1, this cannot be decided from the data alone, but rather further experiments would be necessarily. More commonly, if rare populations are specifically of interest, then alternative experimental designs would be undertaken from the beginning. These issues are widely known and lack a 'cut-and-dried' solution (Roederer, 2008). Thus, our strategy has been to leave the minimum population size decision to the operator, and return a list of putative populations. A future improvement would be to additionally return a measure of scatter for each population, to improve the interpretation process.

This algorithm not only represents a statistical advance in the analysis of flow cytometry experiments, but it also confers significant technical advantages. A current limitation of flow cytometry is the choice of fluorescent markers: for optimal compensation, they must be chosen so as to have minimal spectral overlap. Because all calculations are done in spectral space, it is no longer crucial that spectral overlap is minimized to such large extent. This increases the range of experiments that are technically possible on standard instrumentation already in daily use.

5 CONCLUSION

The introduction of FMOs was a major step forward in the quality control of flow cytometry experiments, and they remain a gold standard but their use has largely been confined to visual inspection. We have presented a novel one-class classifying algorithm that harnesses the properties of FMOs to systematically identify cell populations. By doing so, the validation problem of unsupervised learning is circumvented. As all calculations are done in the spectral space, new technical possibilities are opened, such as a great deal more flexibility in the choice of fluorescent markers, as compensation is not required. Finally, this method is immediately available to any flow cytometry user, as it can be done on completely standard instrumentation without requiring the purchase of specialized equipment.

Conflict of interest: none declared.

REFERENCES

Aghaeepour,N. *et al.* (2011) Rapid cell population identification in flow cytometry data. *Cytometry A*, **79**, 6–13.

Aghaeeour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Aubin,J.E. (1979) Autofluorescence of viable cultured mammalian cells. *J. Histochem. Cytochem.*, **27**, 36–43.

Bacher,P. *et al.* (2013) Antigen-reactive T cell enrichment for direct, high-resolution analysis of the human naive and memory Th cell repertoire. *J. Immunol.*, **190**, 3967–3976.

Baumgarth,N. and Roederer,M. (2000) A practical approach to multicolor flow cytometry for immunophenotyping. *J. Immunol. Meth.*, **243**, 77–97.

Chapelle,O. *et al.* (2002) Choosing multiple parameters for support vector machines. *Mach. Learn.*, **46**, 131–159.

Chattopadhyay,P.K. *et al.* (2008) A chromatic explosion: the development and future of multiparamter flow cytometry. *Immunology*, **125**, 441–449.

Fawcett,T. (2006) An introduction to ROC analysis. *Patt. Recog. Lett*, **27**, 861–874.

Flury,B. (1983) Some relations between the comparison of covariance matrices and principal component analysis. *Comp. Stat. Data Anal.*, **1**, 97–109.

Fraley,C. *et al.* (2012) mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597*. Department of Statistics, University of Washington.

Hahne,F. *et al.* (2009) flowCore: a Bioconductor package for high-throughput flow cytometry. *BMC Bioinformatics*, **10**, 106.

Hastie,T. *et al.* (2009) Unsupervised learning. *Elements of Statistical Learning*. 2nd edn. Springer Verlag.

Lo,K. *et al.* (2009) flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, **10**, 145.

Novo,D. *et al.* (2013) Generalized unmixing model for multispectral flow cytometry utilizing nonsquare compensation matrices. *Cytometry A*, **83A**, 508–520.

Pyne,S. *et al.* (2009) Automated high-dimensional flow cytometric data analysis. *Proc. Natl Acad. Sci. USA*, **106**, 8519–8524.

Qiu,P. *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotech.*, **29**, 886–891.

R Core Team. (2013) R: a language and environment for statistical computing. *R Foundation for Statistical Computing*.

Roederer,M. (2001) Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry*, **45**, 194–205.

Roederer,M. (2002) Compensation in flow cytometry. *Curr. Prot. Cytometry*, **22**1.14.1–1.14.20.

Roederer,M. (2008) How many events is enough? Are you positive? *Cytometry A*, **73**, 384–385.

Schölkopf,B. *et al.* (2000) Estimating the support of a high-dimensional distribution. *Technical Report MSR-TR-99-87*. Microsoft Research.

Tax,D.M.J. (2001) One-class classification (concept-learning in the absence of counter-examples). PhD Thesis. Technische Universiteit Delft.