Batch Discovery of Recurring Rare Classes toward Identifying Anomalous Samples

Murat Dundar[*]
Computer Science
Department
IUPUI
723 W. Michigan St.
Indianapolis, IN 46202
dundar@cs.iupui.edu

Halid Ziya Yerebakan
Computer Science
Department
IUPUI
723 W. Michigan St.
Indianapolis, IN 46202
hzyereba@cs.iupui.edu

Bartek Rajwa
Bindley Bioscience Center
Purdue University
1203 W. State Street
W. Lafayette, IN 47907
rajwa@cyto.purdue.edu

ABSTRACT

We present a clustering algorithm for discovering rare yet significant recurring classes across a batch of samples in the presence of random effects. We model each sample data by an infinite mixture of Dirichlet-process Gaussian-mixture models (DPMs) with each DPM representing the noisy realization of its corresponding class distribution in a given sample. We introduce dependencies across multiple samples by placing a global Dirichlet process prior over individual DPMs. This hierarchical prior introduces a sharing mechanism across samples and allows for identifying local realizations of classes across samples. We use collapsed Gibbs sampler for inference to recover local DPMs and identify their class associations. We demonstrate the utility of the proposed algorithm, processing a flow cytometry data set containing two extremely rare cell populations, and report results that significantly outperform competing techniques.

The source code of the proposed algorithm is available on the web via the link: `http://cs.iupui.edu/~dundar/aspire.htm`.

Categories and Subject Descriptors

I.5.3 [**Pattern recognition**]: Clustering—*algorithms*

Keywords

hierarchical Dirichlet process, random effects, batch clustering, recurring classes, rare classes, anomaly detection

1. INTRODUCTION

Rare-class discovery is a difficult machine-learning problem that occurs in various practical settings, including visual surveillance and monitoring, quality control, astronomy, physics, and – last but certainly not least – life sciences. A solution to the detection of rare classes is essential for rapid identification of samples with anomalous patterns of data. In this context a normal sample can be considered to be a composition of data points each originating from a *predefined*, i.e., known, class. Unlike normal samples, anomalous samples contain data points originating from classes not known beforehand and thus are considered *undefined*. An anomalous sample may contain data points from both defined and undefined classes; however, the points belonging to undefined classes are usually far less frequent than those originating from predefined ones, hence the term *rare* classes. Predefined classes are recurring and form reproducible patterns (in terms of class membership proportions) across all normal samples, whereas rare classes do not necessarily recur, and when they do, they may form varying patterns of class proportions in each anomalous sample. Therefore, anomalous samples can be as different from each other as they are from normal samples, in terms of the specific subset of rare classes present and their membership proportions.

We assume that data for each sample are generated by local distributions of classes present in that sample. The total number of classes across all samples is not known. The number and the specific subset of classes locally realized in each sample are also not known. Ideally, local distributions of a given class across all samples should be identical, as they are snapshots of the same underlying model. However, random effects that arise from various sources affecting sample-to-sample heterogeneity cause local distributions of the same class to vary significantly from one sample to other. This makes automated matching of local distributions across samples an arduous task, which is further complicated when some classes are represented by only a small number of data points. As a result, identifying the subset of classes present in each sample and recovering true class distributions become impractical without modeling random effects. Thus, the main objective of this study reaches beyond clustering on a per-sample basis, but addresses the issue of grouping local clusters across multiple samples to identify subsets of classes present in each sample. This goal is achieved under the severe constraint imposing the mo1(n)1(d)- in which some

[*]Corresponding author.

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623695.

classes are rare, local class distributions vary from sample to sample owing to random effects, and classes may disappear altogether from some samples.

1.1 Motivation

Our research has been motivated mainly by a practical problem related to automated clinical diagnostics, involving flow cytometry (FC) data analysis.

FC is a single-cell screening, analysis, and sorting technology that plays a crucial role in research and clinical immunology, hematology, and oncology. The power of FC lies in its ability to quantify phenotypic characteristics of individual cells in a high-throughput manner. This unique capability allows FC to study complex inter-cellular networks, such as the immune system as it responds to various external perturbants, including pathogens, chemical compounds (drugs), or vaccination. The cellular phenotypes are defined in FC by combinations of morphological features (measured by elastic light scatter) and abundances of surface and intracellular markers. Each biological sample contains multiple, functionally distinct cell types, or "cell populations" in FC vernacular. These populations form multidimensional clusters in the space defined by measured biological features. Although the characteristics of cell populations present in normal samples are generally known, the number of populations and the proportions of cells present in them could be substantially different in anomalous (often diagnostically relevant) samples.

Given the rapid increase in FC data abundance and the unsatisfactory level of engagement from the machine-learning community, FC researchers have been organizing the annual FlowCAP (Flow Cytometry Critical Assessment of Population Identification Methods) competition in order to increase awareness and elicit help from data scientists. The problem that our study tackles is related to the rare-class classification challenge introduced in FlowCAP 2012 [2]. The data set used in this challenge was produced by multiple laboratories participating in the External Quality Assurance Program Oversight Laboratory (EQAPOL) project [1].

The data sets containing two biologically important rare-cell populations represent samples that were subject to several potential sources of variation, including natural biological variability, different stimulation levels, and data acquisition in different laboratories. The challenge provided several data sets representing three biological samples, at three levels of stimulation, collected in fifteen FC laboratories across the US. For the purpose of method verification the data points (individual cells analyzed by FC) belonging to two rare classes were manually labeled by experts. The remaining data points, considered "normal" (and hence not interesting from the perspective of rare-class discovery), were all labeled as a single predefined abundant class. A typical sample contained about three hundred thousand data points of which only less than one percent belonged to one of the two rare classes.

The FlowCAP challenge framed this problem in a standard supervised classification setting in which the contestants were provided with half the samples as training data and were required to build classifier models subsequently assessed by the organizers using the remaining test data.

Although we appreciate the complexity and the difficulty of the challenge, we believe it represented the best-case scenario and a relatively easy problem setting. Therefore, our problem formulation presented in this report differs significantly from the FlowCAP challenge description. We recognize that biologically the rare classes may emerge as a result of various external perturbants, some of which may be unknown *a priori*. Thus, defining rare classes in an exhaustive fashion may not be realistic. In other words, defining rare classes on the basis of a small available subset present in the training data inevitably leads to classifiers that are biased towards those particular types of rare classes. Such models may not generalize well when applied to future samples in which rare classes may originate owing to other biological mechanisms. Therefore, our problem formulation requires that rare class discovery be performed in the absence of labeled data points representing these classes in the training sets. Herein, we present a non-parametric Bayesian algorithm called ASPIRE (anomalous sample phenotype identification with random effects) that identifies biologically significant phenotypes across a batch of samples in the presence of random effects.

1.2 Proposed Approach

We model each sample data by a mixture of potentially infinitely many Dirichlet-process Gaussian-mixture models (DPMs) with each individual DPM modeling the local distribution of a single class. Under fairly weak assumptions and given enough components, finite mixtures of Gaussian distributions can model a given density arbitrarily closely [9]. The DPM itself is a mixture of potentially infinitely many Gaussian distributions with the actual number of mixture components determined directly from the data during inference. Thus, modeling local class distributions by DPMs offers the flexibility needed to accommodate class data that may arise in samples subjected to significant sources of variations.

As local distributions of a given class are noisy realizations of the true class distribution we introduce a sharing mechanism to create dependencies across DPMs associated with the same class. This is achieved by centering the base distributions of DPMs associated with the same class on a unique global parameter, which itself is distributed according to a higher level DPM. This global DPM not only associates local distributions of a given class with one another but also models the number and proportions of classes in each sample.

We use a collapsed Gibbs sampler to perform inference. Model learning, which is performed in a single unified process, involves three main tasks: recovering DPMs in each sample, finding class associations of DPMs, and identifying the total number of classes and their proportions in each sample.

ASPIRE is capable of identifying recurring classes (both normal and rare) in a completely unsupervised way across a batch of samples that are significantly perturbed by random effects and can characterize normal as well as anomalous states given only very weak assumptions regarding sample characteristics and origin.

1.3 Related Work

Existing lines of work that can be adapted to solve the described problem can be broadly grouped into three categories.

The first approach involves pooling data from all samples and applying a standard clustering algorithm to cluster pooled data. Such an approach will have limited success with most real biological data sets because in the presence of random effects, local distributions belonging to one class may significantly overlap with local distributions of another class. The degree of overlap will be more severe in the presence of rare classes. As a result, clusters recovered this way are unlikely to have any meaningful correspondence with the true class distributions.

The second approach involves identifying clusters on a per-sample basis and then matching local clusters across samples to recover actual class distributions. Although this technique may perform better than the first solution operating with pooled data, the cluster-matching part will remain a big challenge in the presence of random effects and rare classes. As a result, local distributions corresponding to larger classes may not be recovered as a whole and clusters corresponding to rare classes may be incorrectly matched with the distributions of other dominant classes, failing to indicate rare classes. FLAME (flow analysis with automated multivariate estimation) [11] is a well-known specialized FC algorithm that can be considered an example belonging to this category. FLAME fits a mixture model into each sample data with four possible choices of density functions (Gaussian, skewed-Gaussian, t-distribution, skewed-t-distribution) available for individual mixture components. Local modes are pooled and then clustered to obtain a global template of meta-clusters. Local clusters are then assigned to these meta-clusters using graph-matching techniques. FLAME is somewhat similar to ASPIRE in the narrow sense that both techniques model individual sample data by a mixture model. However, there are significant differences in model learning. FLAME divides model learning into three tasks: clustering data in individual samples, finding the optimal number of local clusters in each sample, and matching local clusters across samples to recover classes. These three tasks are performed by FLAME independently in a sequential manner. Unlike FLAME, the model learning by ASPIRE is performed as a single unified process. Thus, ASPIRE can take advantage of recurring patterns of similarities across samples. For example, groups of isolated data points forming rare classes that would be ignored as outliers by clustering followed by cluster matching can be successfully identified as a rare class when these two tasks are performed simultaneously.

The third approach involves performing sample clustering jointly with cluster matching. The proposed ASPIRE model, the hierarchical Dirichlet-process Gaussian-mixture model (HDPM) [5], and HDPM with random effects (HDPM-RE) [8] all belong to this category. Thanks to their nonparametric nature, the number of local clusters and classes can arbitrarily grow in all three models to better accommodate data as needed. Both HDPM and HDPM-RE model individual sample data by a single DPM. HDPM uses the standard hierarchical Dirichlet process prior [13], assuming exact sharing of class parameters across all samples and ignoring the presence of random effects. In the presence of random effects this assumption leads to the creation of several extraneous classes. HDPM tackles this problem by post-processing the results to combine local clusters sharing a common mode. However, such a post-processing technique may have limited success, as local clusters of a given class may not necessarily share the same mode. Unlike HDPM, HDPM-RE assumes that local clusters are noisy realizations of true class distributions and probabilistically models the deviations of the local cluster means from the mean of the corresponding class distribution.

One key limitation of HDPM-RE is the assumption that local class distributions can be effectively captured using a single Gaussian distribution. This assumption is often violated in many real-world settings because different sources of variation introduced at different stages of the data collection and processing pipeline create class data that may not be closely approximated by a single Gaussian distribution. In the case of HDPM-RE, additional local clusters of a given class are treated as if they belong to another class, thereby splitting a single class into multiple subclasses. Unlike HDPM-RE, which uses a single Gaussian distribution for each local distribution of a class, ASPIRE uses a single DPM for each local distribution, allowing for an arbitrarily large number of Gaussian distributions for modeling of local class data. Individual DPMs across samples are linked through class-specific global parameters, which are in turn distributed according to a higher-level DPM model. In addition to modeling random effects, ASPIRE offers a more flexible data model that can recover class distributions with arbitrary shapes, avoiding the creation of artificial classes.

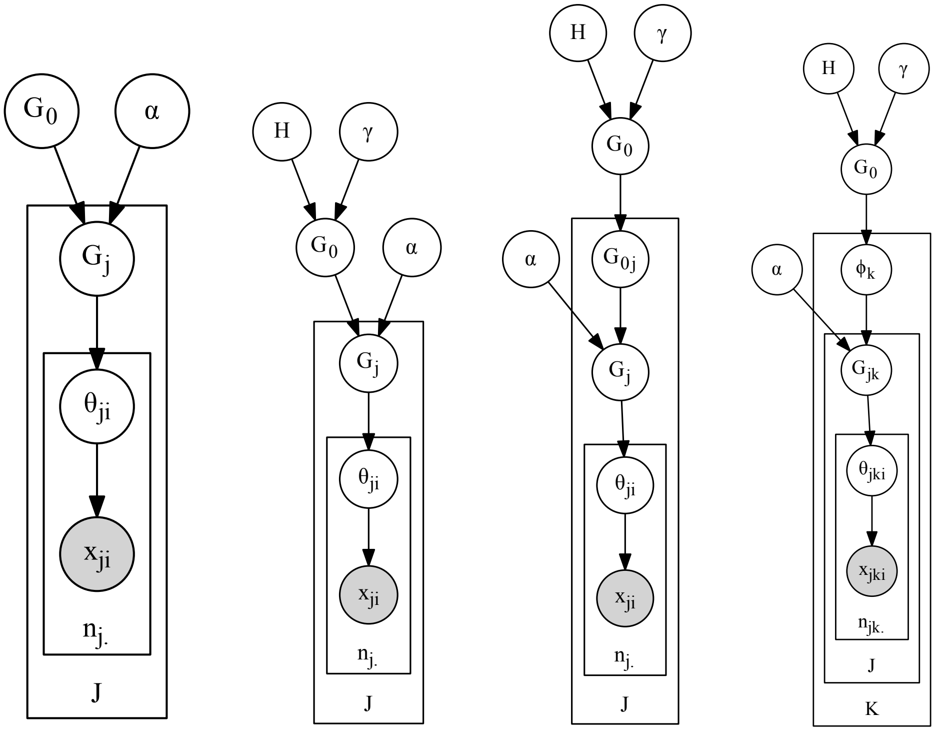
The rest of this report is organized as follows. In Section 2 we compare data models for DPM, HDPM, HDPM-RE, and ASPIRE. In Section 3 we discuss model inference for ASPIRE. In Section 4 we demonstrate the performance of ASPIRE with two experiments and compare results with three other competing techniques. In Section 5 we conclude by summarizing our contributions and offering future research directions.

2. ASPIRE GENERATIVE MODEL

We describe the technical details of our data model in four incremental stages. In the first stage we assume that each sample is modeled by a single DPM and that DPMs across multiple samples are independent. In the second stage we introduce dependencies across DPMs and impose exact sharing of mixture components corresponding to classes across samples. This is equivalent to the HDPM model. In the third stage we tackle random effects by relaxing the exact sharing of mixture components to allow local clusters to inherit noisy realizations of classes in individual samples. This is equivalent to the HDPM-RE model. In the fourth stage we describe the proposed data model for ASPIRE, which models each sample by a potentially infinite mixture of DPMs.

2.1 Independent Modeling of Samples by DPM

We denote point i in sample j by $\boldsymbol{x}_{ji} \in \Re^d$, where $i =$ $\{1, \ldots, n_j\}$ and $j \in \{1, \ldots, J\}$, n_j is the number of points b/T1_1 9.96 Tf 8.9 G(6)(T)f 3407 /(0)-1J(t)6727 2816 Td/[(j)T44 T.0221 0.9
2a1kto j67 i initializationsgb8(w)15ndgin=d1(e)1(l.)-419(l)1(n)-233(th)17 0 Td (j)Tj /T1_0 r



(a) DPM (b) HDPM (c) HDPM-RE (d) ASPIRE

Figure 1: Plate diagrams for DPM, HDPM, HDPM-RE, and ASPIRE.

Using the stick-breaking construction according to [7], we can express G_j as

$$G_j \;=\; \mathsf{P}^{\infty}_{t=1}\ _{jt}\ \psi_{jt} \qquad (3)$$

where

$$\begin{array}{lcl} _{jt} & = & ^{\theta}_{jt}\ \mathsf{Q}^{t-1}_{l=1}(1-\ ^{\theta}_{jl}) \\ ^{\theta}_{jt} & \sim & Beta(1;\ \ ) \\ _{jt} & \sim & G_0 \end{array}$$

The points $_{jt}$ are called the *atoms* of G_j. Note that unlike continuous distributions, the probability of sampling the same $_{jt}$ twice from G_j is not zero and is proportional to $_{jt}$. Thus, G_j is considered a discrete distribution and offers a clustering property, as the same $_{jt}$ can be sampled for different $_{ji}$. In this model is the parameter that controls the prior probability of assigning a point to a new cluster and thus plays a critical role in the number of clusters generated.

For the base distribution G_0, from which $_{jt}$ are drawn, we define a bivariate prior:

$$p\,(\boldsymbol{\mu};\Sigma) = N \quad \boldsymbol{\mu}|\boldsymbol{\mu}_0;\frac{\Sigma}{_0} \quad \times W^{-1}\,(\Sigma|\Sigma_0;m) \qquad (4)$$

where $\boldsymbol{\mu}_0$ is the prior mean and $_0$ is a scaling constant that controls the deviation of the cluster means from the prior mean. The smaller the $_0$, the larger the separation will be between the cluster means. The parameter Σ_0 is a positive definite matrix that encodes our prior belief about the expected Σ, i.e., $E(\Sigma) = \frac{\theta}{m-d-1}$. The parameter m is a scalar that is negatively correlated with the degrees of freedom. In other words the larger the m, the less Σ will deviate from $E(\Sigma)$, and vice versa. The plate model for independent modeling of samples using one DPM for each sample is available in Figure 1a.

2.2 Introducing dependencies across samples by HDPM

In the previous section we introduced a clustering property across points in an individual sample by placing a DP prior over G_j as in (2). Since G_j is a discrete distribution, this prior enables sharing of the same cluster parameter by different points. When dealing with multiple samples, in addition to sharing of clusters by points formed within individual samples, a higher level of sharing occurs. Each local cluster in an individual sample is associated with a class. Thus, as we cluster points in each sample we also need to group local clusters into appropriate classes so that we can identify class associations of local clusters. This grouping can be achieved by introducing dependencies across individual DPMs by placing a hierarchical DP prior over G_0 [13]. The HDPM for joint clustering and cluster matching across multiple samples becomes

$$\begin{array}{lcl} \boldsymbol{x}_{ji} & \sim & p\,(\cdot|\ _{ji}) \\ _{ji} & \sim & G_j \\ G_j & \sim & DP\,(G_0;\ \ ) \\ G_0 & \sim & DP\,(H;\ \ ) \end{array} \qquad (5)$$

where is the precision parameter for the higher-level DP prior and H is defined as in (4).

Using the stick-breaking construction we can express G_0 as

$$G_0 \quad = \quad \mathsf{P}_{k=1}^{\infty} \quad _k \;\; \phi_k \qquad (6)$$

where

$$\begin{array}{rcl} _k & = & {}^{\theta}_{k} \; \mathsf{Q}_{l=1}^{k-1}(1 - \; {}^{\theta}_{k}) \\ {}^{\theta}_{k} & \sim & Beta(1; \quad) \\ _k = \{\boldsymbol{\mu}_k; \Sigma_k\} & \sim & H \end{array}$$

With this update, instead of letting G_0 be distributed according to (4) as in the independent modeling of samples we let H be distributed according to (4) and let the atoms of G_0 be distributed according to H. The distinct set of parameters $\;_k$ corresponding to classes is sampled from H and local cluster parameters are sampled from G_j. Since G_j is a discrete distribution with its atoms sampled from G_0, and G_0 is a discrete distribution with its atoms sampled from H, each local cluster in turn inherits one of the $\;_k$, i.e., $\;_{jt} \in \{ \;_k\}_{k=1}^{K}$ and $\;_{ji} \in \{ \;_{jt}\}_{t=1}^{m_{j\cdot}}$, where K is the number of classes and $m_{j\cdot}$ is the number of local clusters in sample j.

Therefore, this model not only groups data points within each sample into clusters, but also groups local clusters across samples into classes. In other words, clustering and cluster matching are simultaneously addressed and depend on one another. The plate model for HDPM is available in Figure 1b.

2.3 Modeling random effects by HDPM-RE

In the standard HDPM the same parameters are inherited by all local realizations of a class. However, owing to the potential random effects this surmise may be unrealistic. Therefore, to account for random effects the HDPM-RE model [8] would be more suitable for the discovery of recurring classes. HDPM-RE presumes that sample data are generated by noisy versions of parameters defining classes. This change can be incorporated into the data model by updating the model in (5) as follows:

$$\begin{array}{rcl} \boldsymbol{x}_{ji} & \sim & p\,(\cdot| \;_{ji}) \\ _{ji} & \sim & G_j \\ G_j & \sim & DP\,(G_{0j}; \quad) \\ G_0 & \sim & DP\,(H; \quad) \end{array} \qquad (7)$$

where G_{0j} is a discrete distribution whose atoms are noisy versions of the corresponding atoms in G_0. With this change in the model each individual sample now inherits different noisy realizations of global parameters. The plate model for HDPM-RE is available in Figure 1c.

2.4 Modeling individual sample data with multiple DPMs

Both HDPM and HDPM-RE assumes that local distributions of classes can be closely approximated by a single Gaussian distribution. This assumption is often quite restrictive for many practical settings, as local class data, which are produced subject to random effects, may emerge in the form of skewed as well as multi-mode distributions. As a result, fitting a single Gaussian distribution for local class distributions creates artificial classes that may not be easily distinguished from other significant classes.

ASPIRE uses a potentially infinite mixture of DPMs to model each sample data where individual DPMs are linked together through a hierarchical DP prior. This hierarchical prior not only identifies local DPMs associated with the same class through sharing of a global parameter but also models the specific subset of classes present and their proportions in each sample.

We update our indexing notation from previous sections to introduce an additional subscript k to account for multiple DPMs in each sample. We denote point i of class k in sample j by $\boldsymbol{x}_{jki} \in \Re^d$, where $i = \{1; : : : ; n_{jk\cdot}\}$, $k = \{1; : : : ; K\}$, and $j = \{1; : : : ; J\}$, $n_{jk\cdot}$ is the number of points from class k in sample j, K is the total number of classes, and J is the total number of samples. The proposed ASPIRE data model is as follows.

$$\begin{array}{rcl} \boldsymbol{x}_{jki} & \sim & p\,(\cdot| \;_{jki}) \\ _{jki} & \sim & G_{jk} \\ G_{jk} & \sim & DP\,(F_{\phi_k}; \quad) \\ _k & \sim & G_0 \\ G_0 & \sim & DP\,(H; \quad) \end{array} \qquad (8)$$

where $\;_k$ are global parameters each of which is associated with a different class. Individual DPMs associated with the same class inherit the same $\;_k$ across samples. The notation F_{ϕ_k} indicates a distribution F centered at $\;_k$ and defines class-specific base distributions of individual DPMs. Although F_{ϕ_k} is same for all DPMs associated with the same class, local clusters between samples are generated i.i.d. given $\;_k$ of corresponding DPMs. Thus, each local realization of a given class is modeled by a different DPM, allowing for the modeling of sample-to-sample variations in samow1(o)l(a)1(1o)1()1(1cF(en)-37gug)1(en)1(1c1n)1(g)-.461 T708 1.[(rea

$$\bar{} = \frac{\left(\sum_{jkt:c_{jkt}=k} \frac{n_{jkt}\kappa_1}{(n_{jkt}+\kappa_1)} + {}_0\right) {}_1}{\sum_{jkt:c_{jkt}=k} \frac{n_{jt}\kappa_1}{(n_{jkt}+\kappa_1)} + {}_0 + {}_1} \quad (22)$$

Once the distributions in (17)-(20) are substituted into (14) a closed-form expression for $p(\boldsymbol{\mu}_{jkt}; \Sigma_k | D_{.c_{jkt}.}; D_{jkt})$ can be obtained. When we substitute this solution into (13) we obtain $p(\boldsymbol{x}_{jki} | D_{.c_{jkt}.}; D_{jkt})$ in the form of a multivariate Student-t distribution with three parameters.

$$p(\boldsymbol{x}_{jki} | D_{.c_{jkt}.}; D_{jkt}) = stu - t(\hat{\boldsymbol{\mu}}; \hat{\Sigma}; v) \quad (23)$$

The location vector ($\hat{}$), the scale matrix ($\hat{\Sigma}$), and the degrees of freedom (v) are given below. Location vector:

$$\hat{\boldsymbol{\mu}} = \frac{n_{jkt}\bar{\boldsymbol{x}}_{jkt} + \bar{}\bar{\boldsymbol{\mu}}}{n_{jkt} + \bar{}} \quad (24)$$

Scale matrix:

$$\hat{\Sigma} = \frac{\Sigma_0 + A_k + A_{jkt} + \frac{n_{jkt}\kappa}{n_{jkt}+\kappa}(\bar{\boldsymbol{x}}_{jkt} - \bar{\boldsymbol{\mu}})(\bar{\boldsymbol{x}}_{jkt} - \bar{\boldsymbol{\mu}})^T}{\frac{(\kappa + n_{jkt})\, v}{(\kappa + n_{jkt} + 1)}} \quad (25)$$

Degrees of freedom:

$$v = m + \sum_{jkt:c_{jkt}=k} (n_{jkt} - 1) + n_{jkt} - d + 1 \quad (26)$$

The predictive distribution of a class can be readily obtained from $p(\boldsymbol{x}_{jki} | D_{.c_{jkt}}; D_{jkt})$ by setting D_{jkt} an empty set. This is equivalent to dropping terms related to local clusters in equations (24), (25), and (26). Finally, the predictive distribution of an empty cluster can be obtained from $p(\boldsymbol{x}_{jki} | D_{.k.})$ by setting $D_{.k.}$ an empty set. This is equivalent to dropping terms in $p(\boldsymbol{x}_{jki} | D_{.k.})$ related to classes.

4. RESULTS AND DISCUSSIONS

We report results of experiments performed with two different data sets. The first experiment demonstrated the functionality of the algorithms tested using simulated data, while the second experiment utilized real FC data.

Aside from the proposed ASPIRE algorithm, three other techniques were considered: DPM, HDPM, and HDPM-RE. In Section 1.3 we described three different approaches to the clustering problem set forth in this study. The first method uses standard clustering algorithms applied to pooled data, the second approach performs clustering and cluster matching in a sequential way, and the third performs clustering jointly with cluster matching. Among the three benchmark techniques DPM belongs to the first category; HDPM and HDPM-RE along with ASPIRE belong to the third category. We chose the well-known FC algorithm FLAME to represent the second category. Unfortunately the implementation of FLAME available through GenePattern [12] produced errors during processing of many of the samples in the two data sets, so we were forced to exclude FLAME from this analysis. For HDPM we used the software provided by the authors in [5]. For the other three algorithms we used our own implementations. Each algorithm is run for a thousand sweeps, and the state with the best likelihood is recorded for subsequent analysis.

The F_1 score is used as the performance measure for comparing performances of these four techniques. As one-to-many matchings are expected between true and recovered classes, the F_1 score for each class is computed as the maximum of the F_1 scores for all recovered classes, similar to [3].

Table 1: F1 scores achieved and the number of classes recovered by each of the four techniques on the artificial data set.

	Class F1 Scores			
Method	1 (98.7%)	2 (0.3%)	3 (1%)	# Classes
DPM	1.00	0.75	0.56	5
HDPM	0.84	0.74	0.66	11
HDPM-RE	0.68	0.94	0.85	7
ASPIRE	1.00	1.00	0.90	3

Table 2: Number of points available from three classes in the FC data set before and after subsampling. Numbers in parentheses indicate percentage of the total number of points in the corresponding set.

	# points		
Method	Normal	Rare 1	Rare 2
Original	56.2M (99.94%)	10.2K (0.02%)	24.3K (0.04%)
Subsampled	1.9M (98.23%)	9.5K (0.50%)	24.1K (1.27%)

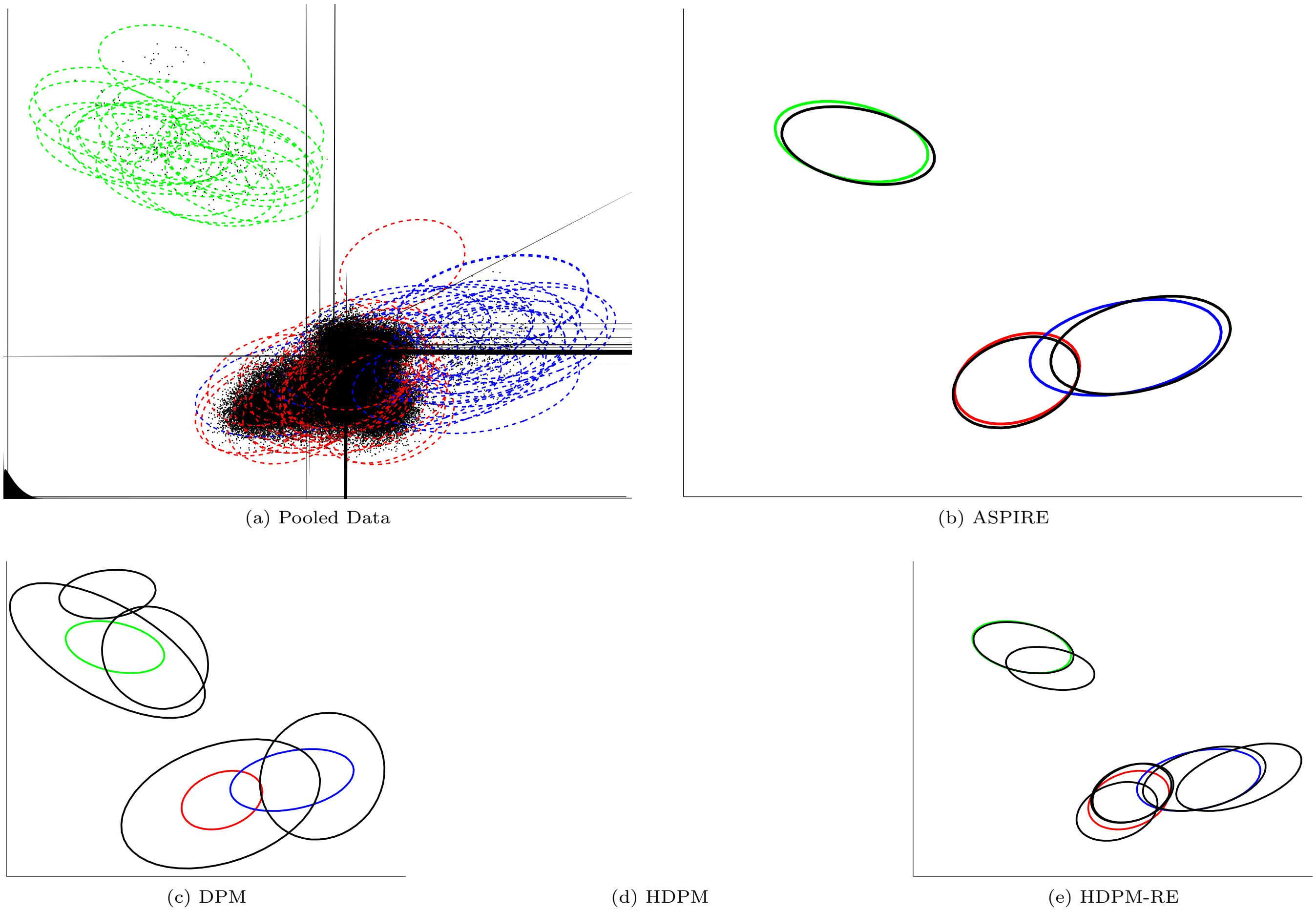
4.1 Experiment 1: Artificial Data Set

We generated twenty samples, each with five thousand data points in a two-dimensional feature space, using the model in (8) and the following values of the model parameters: ${}_0 = 0{:}01$, ${}_1 = 0{:}2$, $m = 20$, ${}_0 = [0\ 0]^T$, $\Sigma_0 = I$, $= 0{:}2$, $= 0{:}2$, where I denotes the identity matrix. After all data points were sampled, three classes were produced by this model with overall class proportions of 0.987, 0.003, and 0.01, which indicates that two of the three recurring classes can be considered rare. For the pooled data, distributions of local clusters and the true values of the global parameters, i.e., ${}_k$, are shown in Fig. 2a by dashed and solid contours, respectively. The ellipses correspond to data distributions that are at most four standard deviations from the mean. Individual data points are shown by black dots.

We ran all four techniques (ASPIRE, DPM, HDPM, and HDPM-RE) on this data set and plotted contours representing recovered classes in Figures 2b, 2c, 2d, and 2e, respectively. F1 scores obtained for each class and numbers of classes recovered by all four techniques are included in Table 1. Results suggest that ASPIRE not only correctly predicts the true number of classes but also estimates global parameters with almost no bias, which in turn produces almost perfect F1 scores for each class. DPM produces a reasonable number of classes but estimates global parameters with a large bias. HDPM fails to consistently match local clusters across samples and substantially overpredicts the actual number of classes. HDPM-RE performs better compared to DPM and HDPM but generates several artificial classes, a direct result of modeling local class data by a single Gaussian distribution.

4.2 Experiment 2: Flow Cytometry Data Set with Two Rare Classes

We evaluated the performance of ASPIRE in discovering rare classes with a FC data set used in the FlowCAP 2012 competition [2]. The data set contained FC measurements of multiple aliquots of three biological samples exposed to three



(a) Pooled Data

(b) ASPIRE

(c) DPM

(d) HDPM

(e) HDPM-RE

Figure 2: An illustrative example showing the performance of DPM, HDPM, HDPM-RE, and ASPIRE algorithms in estimating global parameters corresponding to classes. Solid color contours plotted using true values of global parameters represent true classes. Dashed color contours indicate true distributions of local clusters with the color identifying the class origin. Solid black contours plotted using estimated values of the global parameters represent recovered classes. Black dots denote data points.

different stimulation levels. The samples were examined independently by fifteen FC laboratories. In this context the term “sample” denotes a tube containing white blood cells. Each cell is separately measured by a flow cytometer. The measurement provides the small-angle and large-angle light-scatter characteristics as well as four fluorescence intensity values. Thus, each cell is characterized by a six-dimensional feature vector. The goal is to recognize the cells belonging to two rare-cell populations, manually labeled by experts, without access to information about characteristics of these populations in the training data set. Cells not belonging to one of the two rare populations are considered “normal” and were all labeled as a single predefined abundant class. Thus, including the normal class there are three classes in this data set.

The original data set contained data points for about 60 million cells across 202 samples. To obtain a more manageable data-set size while preserving cells from rare classes we used a density-based subsampling technique and reduced the data size to 1.9 million points. The number of points available from each of the two rare classes as well as the normal class before and after subsampling and their percentages are shown in Table 2.

As in the previous experiment, we compare ASPIRE against DPM, HDPM, and HDPM-RE. The DPM model has five free parameters ($;\Sigma_0; m;\ _0;\ _0$), the HDPM model has one more parameter ( ) than DPM, and HDPM-RE and ASPIRE have one more parameter ($_1$) than HDPM. These parameters are selected using the following strategy.

Each feature is normalized to have zero mean and unit variance. As the sample batch may contain anomalous samples, prior information about the potential number of local clusters and global classes may not exist for most real-world FC data. Thus, for   and   we use vague priors by fixing their value to one. We set m to the minimum feasible value, which is $d+2$, to achieve maximum degrees of freedom. By doing this we let the actual covariance matrices deviate significantly from the expected covariance matrix, which is $E(\Sigma) = \frac{_0}{m-d-1}$. The prior mean $_0$ is set to the mean of the entire data. The scale matrix Σ_0 is set to $I{=}s$, where I is the identity matrix. This leaves the scaling constant s of Σ_0, $_0$, and $_1$ as the three free parameters that require

Table 3: F$_1$ scores achieved and the number of classes recovered by each of the four techniques on the entire FC data set. Results for ASPIRE are averages over ten repetitions. Numbers in parenthesis indicate standard deviations.

	Class F$_1$ Scores			
Method	Normal	Rare 1	Rare 2	# Classes
DPM	0.22	0.20	0.39	175
HDPM	0.23	0.01	0.02	75
HDPM-RE	0.22	0.46	0.63	91
ASPIRE	0.62	0.59	0.77	38.7
	(0.02)	(0.03)	(0.01)	(3.20)

tuning. The parameter $_1$ models the deviation of cluster means from their corresponding class mean in the generative model. Thus, increasing $_1$ while $_0$ and s are fixed potentially increases the number of classes generated. The parameter $_0$ models the deviation of cluster means from the prior mean in the generative model. Thus, increasing $_0$ while $_1$ and s are fixed potentially increases the number of clusters generated. The parameter s models the expected size of clusters. Increasing s potentially increases the number of clusters generated. These three parameters were coarsely tuned using a generic 5-parameter peripheral-blood immunophenotyping data set previously collected and analyzed in our lab as part of an earlier study without retuning them for the FC data used in this experiment. The following values were used: $_0 = 0{:}05$, $_1 = 0{:}1$, $s = 10$.

F$_1$ scores computed for all three classes are shown in Table 3. Results for ASPIRE are averages of ten repetitions. As the run time for ten repetitions of the other algorithms would take on the order of weeks, we included results of a single run for these algorithms. Results in Table 3 favor methods modeling random effects (HDPM-RE and ASPIRE) over those that do not (DPM and HDPM) in terms of higher F$_1$ scores achieved for both rare classes. Between techniques that model random effects ASPIRE significantly outperforms HDPM-RE in terms of producing a more realistic number of classes and higher F$_1$ scores for all classes. ASPIRE models local realizations of classes by an infinite mixture of Gaussians, which allows for associating multiple clusters with individual classes during inference. The other three techniques use a single Gaussian distribution to model local realization of classes. If a local distribution of a class cannot be effectively modeled by a single Gaussian distribution, these techniques tend to produce multiple local clusters all of which are assigned to a distinct class. As a result ASPIRE tends to generate a fewer number of classes and achieves higher F$_1$ scores compared to the other three techniques.

We also compared ASPIRE with a supervised classifier to find out how F$_1$ scores would improve if a subset of the labeled data were to be used during training. We used all samples belonging to one of the biological samples for training and sequestered all samples for the other two biological samples for testing. The support vector machine toolbox in [6] was used to train and test a supervised classifier on this data. Parameters of this classifier are extensively tuned to optimize test performance. These results along with the results obtained by ASPIRE on the test data are shown in Table 4. Results suggest that ASPIRE can predict rare classes with F$_1$ scores comparable to those of a supervised classifier without using any labeled data. The F$_1$ score achieved by ASPIRE for the normal class is worse than that of the supervised classifier, mainly because the normal class is a combination of multiple uninteresting subclasses for which ASPIRE produces multiple classes to more effectively model the underlying class distribution. However, we do not believe this is a major limitation, as in most practical settings labeled data are present for normal classes as these are classes that are known and predefined. On the other hand, for rare classes, labeled data may not exist because rare classes are usually not known a priori and cannot be predefined. Under such circumstances training a supervised classifier that requires labeled data for all classes may not be very realistic. On the other hand, ASPIRE can cluster data in a fully unsupervised manner and with the help of a limited amount of labeled data from normal classes results can be post-processed to distinguish unknown classes from known ones.

Table 4: F$_1$ scores achieved by ASPIRE and SVM on the test portion of the FC data set. Results are averages over ten repetitions. Numbers in parenthesis indicate standard deviations.

	Class F$_1$ Scores		
Method	Normal	Rare 1	Rare 2
ASPIRE	0.62	0.54	0.75
	(0.02)	(0.03)	(0.01)
Supervised	1.00	0.66	0.83
	(0.00)	(0.01)	(0.01)

For ASPIRE, one sweep of the Gibbs sampler involves two main iterative loops. In the first loop, cluster indicator variables are sampled for all data points across all samples. In the second loop, class indicator variables are sampled for all local clusters across all samples. As the first loop iterates over all points across all samples it is usually more computationally expensive than the second loop. Fortunately, during the sampling of the cluster indicator variables class parameters are fixed. This allows us to sample cluster indicator variables independently for each sample during a single sweep and leads to improvement in processing time on multi-processor machines. The actual run time for ASPIRE to process the FC data set containing 1.9 million points is about five and eleven hours with and without parallelization, respectively, on an eight-core workstation. The reduction in the overall computational time is not proportional to the number of processors, as the computational gain by parallelizing the first loop will be limited after a certain point by the computational time of the second loop.

5. CONCLUSIONS

We introduced ASPIRE as a new method for discovering recurring yet significant rare classes in the presence of random effects and showed experimental results that clearly favor ASPIRE over other benchmark techniques. We believe that ability to recover rare classes in FC data sets obtained in fifteen different laboratories convincingly demonstrates that automated identification of anomalous samples in research or diagnostic settings is indeed feasible.

Labeled information about normal, i.e., known classes, can be directly incorporated into the learning process by adopting a restricted Gibbs sampler scheme similar to the one introduced in [4]. Our research was mainly driven by a rare-class discovery problem in a clinical setting. However, ASPIRE is a general clustering technique that can be used in other disciplines to discover classes with recurring nature irrespective of whether they are rare or normal. ASPIRE can also be utilized for problems involving the detection of group anomalies [10, 14].

ASPIRE is implemented in C++. The source code is available on the web via the link `http://cs.iupui.edu/~dundar/aspire.htm`.

6. ACKNOWLEDGMENTS

This research was sponsored by the National Science Foundation (NSF) under Grant Number IIS-1252648 (CAREER), by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under Grant Number 5R21EB015707, and by the PhRMA Foundation (2012 Research Starter Grant in Informatics). The content is solely the responsibility of the authors and does not represent the official views of NSF, NIBIB or PhRMA.

7. REFERENCES

[1] External quality assurance program oversight laboratory (EQAPOL). `http://eqapol.dhvi.duke.edu/`.

[2] FlowCAP - flow cytometry: Critical assessment of population identification methods. `http://flowcap.flowsite.org/`.

[3] N. Aghaeepour, G. Finak, FlowCAP Consortium, DREAM Consortium, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, and R. H. Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238, mar 2013.

[4] F. Akova, Y. Qi, B. Rajwa, and M. Dundar. Self-adjusting models for semi-supervised learning in partially-observed settings. In *IEEE International Conference on Data Mining (ICDM'12)*, 2012. under review.

[5] A. J. Cron, C. Gouttefangeas, J. Frelinger, L. Lin, S. K. Singh, C. M. Britten, M. J. P. Welters, S. H. van de Burg, M. West, and C. Chan. Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS Computational Biology*, 9:e1003130, 2013.

[6] N. Djuric, L. Lan, S. Vucetic, and Z. Wang. Budgetedsvm: A toolbox for scalable svm approximations. *Journal of Machine Learning Research*, 14:3813–3817, 2013.

[7] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):pp. 161–173, 2001.

[8] S. Kim and P. Smyth. Hierarchical Dirichlet processes with random effects. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 697–704, Cambridge, MA, 2007. MIT Press.

[9] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, 2001.

[10] K. Muandet and B. Schölkopf. One-class support measure machines for group anomaly detection. *CoRR*, abs/1303.0309, 2013.

[11] S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. De Jager, and J. P. Mesirov. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A*, 106(21):8519–24, 2009.

[12] M. Reich, T. Liefeld, J. Gould, J. Lerner, T. P., and M. J.P. Genepattern 2.0. *Nature Genetics*, 38(5):500–1, 2006.

[13] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[14] L. Xiong, B. Poczos, and J. Schneider. Group anomaly detection using flexible genre models. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1071–1079. 2011.