Information Preserving Component Analysis: Data Projections for Flow Cytometry Analysis

Kevin M. Carter[1], Raviv Raich[2], William G. Finn[3], and Alfred O. Hero III[1]

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109

[2] School of EECS, Oregon State University, Corvallis, OR 97331

[3] Department of Pathology, University of Michigan, Ann Arbor, MI 48109

{kmcarter,wgfinn,hero}@umich.edu, raich@eecs.oregonstate.edu

Abstract

Flow cytometry is often used to characterize the malignant cells in leukemia and lymphoma patients, traced to the level of the individual cell. Typically, flow cytometric data analysis is performed through a series of 2-dimensional projections onto the axes of the data set. Through the years, clinicians have determined combinations of different fluorescent markers which generate relatively known expression patterns for specific subtypes of leukemia and lymphoma – cancers of the hematopoietic system. By only viewing a series of 2-dimensional projections, the high-dimensional nature of the data is rarely exploited. In this paper we present a means of determining a low-dimensional projection which maintains the high-dimensional relationships (i.e. information distance) between differing oncological data sets. By using machine learning techniques, we allow clinicians to visualize data in a low dimension defined by a linear combination of all of the available markers, rather than just 2 at a time. This provides an aid in diagnosing similar forms of cancer, as well as a means for variable selection in exploratory flow cytometric research. We refer to our method as Information Preserving Component Analysis (IPCA).

Index Terms

Flow cytometry, statistical manifold, information geometry, multivariate data analysis, dimensionality reduction

Acknowledgement: This work is partially funded by the National Science Foundation, grant No. CCR-0325571.

I. INTRODUCTION

Clinical flow cytometric data analysis usually involves the interpretation of data culled from sets (i.e. cancerous blood samples) which contain the simultaneous analysis of several measurements. This high-dimensional data set allows for the expression of different fluorescent markers, traced to the level of the single blood cell. Typically, diagnosis is determined by analyzing individual 2-dimensional scatter plots of the data, in which each point represents a unique blood cell and the axes signify the expression of different biomarkers. By viewing a series of these histograms, a clinician is able to determine a diagnosis for the patient through clinical experience of the manner in which certain leukemias and lymphomas express certain markers.

Given that the standard method of cytometric analysis involves projections onto the axes of the data (i.e. visualizing the scatter plot of a data set with respect to 2 specified markers), the multi-dimensional nature of the data is not fully exploited. As such, typical flow cytometric analysis is comparable to hierarchical clustering methods, in which data is segmented on an axis-by-axis basis. Marker combinations have been determined through years of clinical experience, leading to relative confidence in analysis given certain axes projections. These projection methods, however, contain the underlying assumption that marker combinations are independent of each other, and do not utilize the dependencies which may exist within the data. Ideally, clinicians would like to analyze the full-dimensional data, but this cannot be visualized outside of 3-dimensions.

There have been previous attempts at using machine learning to aid in flow cytometry diagnosis. Some have focused on clustering in the high-dimensional space [1], [2], while others have utilized information geometry to identify differences in sample subsets and between data sets [3], [4]. These methods have not satisfied the problem because they do not significantly approach the aspect of visualization for 'human in the loop' diagnosis, and the ones that do [5], [6] only apply dimensionality reduction to a single set at a time. The most relevant work, compared to what we are about to present, is that which we have recently presented [7] where we utilized information geometry to simultaneously embed each patient data set into the same low-dimensional space, representing each patient as a single vector. The current task differs in that we do not wish to reduce each set to a single point for comparative analysis, but to use dimensionality reduction as a means to individually study the distributions of each patient. As such, we aim to reduce the dimension of each patient data set while maintaining the number of

data points (i.e. cells).

With input from the Department of Pathology at the University of Michigan, we have determined that the ideal form of dimensionality reduction for flow cytometric visualization would contain several properties. The data needs to be preserved without scaling or skewing, as this is most similar to the current methods in practice (i.e. axes projections). Hence, the ideal projection should be orthonormal. Secondly, the methods should be unsupervised, relying solely on the geometry of the data. This requirement is straight forward as the dimensionality reduction would be an aid for diagnosis, so no labels would be available. As such, common supervised methods geared towards dimensionality reduction for classification tasks (e.g. LDA methods [8], [9]) are not applicable towards this problem.

Clinicians would also like to work in a low-dimensional space similar to what they have grown accustomed to through years of experience. Once determined, the subspace should be consistent, and should not change when processing new data. Therefore non-linear methods of dimensionality reduction such as [10], [11] are not ideal for this task. Adding new data to non-linear methods forces a re-computation of the subspace, which may be noticeably different than previous spaces (e.g. scaled or rotated differently). This has been approached with out-of-sample extension methods [12], but it is still a relatively open problem. Finally, the projection space needs to preserve the relationship between data sets; patients in the same disease class should show similar expressions in the low-dimensional space, while differing disease classes should be distinct from one another. This requirement leads directly to a projection method which maintains the similarity between multiple data sets, rather than preserving similarities between the elements of a single set.

Given the desired properties, one might immediately consider principal component analysis (PCA) [13] for unsupervised, linear dimensionality reduction. However, PCA has well known issues with data sets in which the interesting directions do not have the largest variance. This is illustrated in Fig. 1, where we illustrate two different patient data sets with 2 distinct diseases (see Section IV-C), and attempt to use PCA to find a 1-dimensional projection. This projection would have the adverse effect of making distinguishable (albeit potentially difficult) patients practically indistinguishable in the lower dimensional space. While other methods such as projection pursuit (PP) [14] and independent component analysis (ICA) [15] may not suffer from the same setbacks, they too have their own drawbacks in relation to the cytometry problem. Namely, while this is an

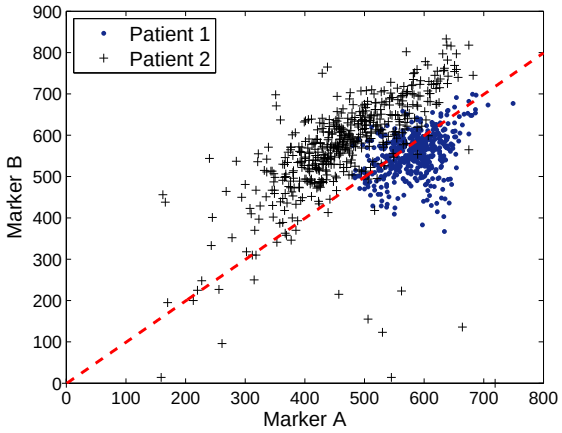


Fig. 1. Using PCA for dimensionality reduction for the comparison of two patients with differing diseases. The PCA projection (dashed line) does not discern the ideal direction due to the variance of the patient 2. Each point represents a unique blood cell analyzed with 2 different markers.

unsupervised problem w.r.t. disease classes, these methods do not operate with any measure of similarity between patient data sets. While we only illustrate two patients, this problem becomes significantly more complex as additional patient data is included.

In this paper we present a method of dimensionality reduction – which we refer to as *Information Preserving Component Analysis (IPCA)* – that preserves the Fisher information distance between data sets. We have shown in previous work [16], [17] that the Fisher information distance is the appropriate means for determining the similarity between non-Euclidean data. This is the case for flow cytometry data, as certain channels may represent light scatter angles, while other channels correspond to the expression of a specific fluorescent marker. Hence, there is no straight-forward Euclidean representation of the data.

By preserving the Fisher information distance between sets, IPCA ensures that the low-dimensional representation maintains the similarities between data sets which are contained in the full-dimensional data, minimizing the loss of information. This low-dimensional representation is a linear combination of the various markers, enabling clinicians to visualize all of the data simultaneously, rather than the current process of axes projections, which only relays information in relation to two markers at a time. Additionally, analysis of the loading vectors within the IPCA projection matrix offers a form of variable selection, which relays information describing which marker combinations yield the most information. This has the significant benefit of allowing for

exploratory data analysis.

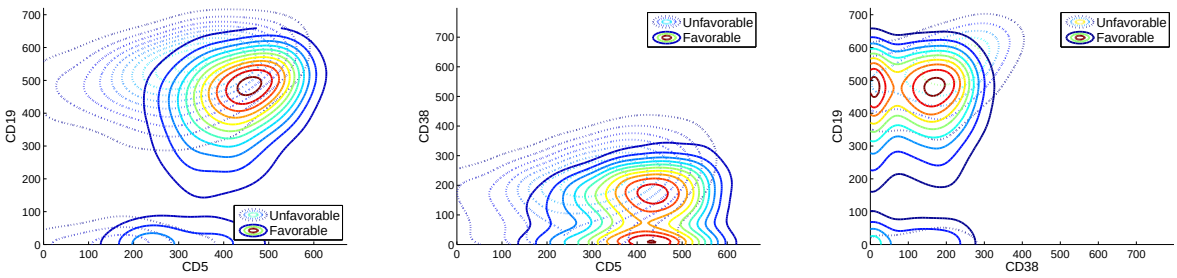
This paper proceeds as follows: Section II gives a background of flow cytometry and the typical clinical analysis process, as well as a formulation of the problem we will attempt to solve. We present our methods for finding the IPCA projection in Section III. Simulation results for clinical cytometric data are illustrated in Section IV, followed by a discussion and areas for future work in Section V.

II. Background

Clinical flow cytometry is widely used in the diagnosis and management of malignant disorders of the blood, bone marrow, and lymph nodes (leukemia and lymphoma). In its basic form, flow cytometry involves the transmission of a stream of cells through a laser light source, with characteristics of each cell determined by the nature of the light scattered by the cell through disruption of the laser light. Application to leukemia and lymphoma diagnosis is usually in the form of flow cytometric immunophenotyping, whereby cells are labeled with antibodies to specific cellular antigens, and the presence of these antigens detected by light emitted from fluorescent molecules (of different "colors") conjugated to the target antibody.

Clinical grade flow cytometers typically assess the size and shape of cells through the detection of light scattered at two predetermined angles (forward angle light scatter, and side angle or orthogonal light scatter), and are also capable of simultaneously detecting the expression patterns of numerous cellular antigens in a single prepared cell suspension ("tube"). The analysis of multiple tubes then allows for any number of antigen expression patterns to be assessed. Although 8-color flow cytometry is possible with the latest generation of clinical grade analyzers, most clinical flow cytometry laboratories utilize 3 or 4 color approaches.

In routine flow cytometric immunophenotyping, the expression patterns of each marker in a given tube can be traced to the level of the single cell, giving flow cytometry a uniquely spatial characteristic when compared to other immunophenotyping or proteomic analysis methods. When measurements of forward and side angle light scatter characteristics are included, each cell analyzed via 4-color flow cytometry can be thought of as occupying a unique point in 6-dimensional space, with the dimensions of each point defined by the magnitude of expression of each antigen or light scatter characteristic. Since all 6 dimensions cannot be projected simultaneously onto a single histogram, diagnosticians typically analyze a series of 2-dimensional histograms



(a) CD5 vs. CD19 (b) CD5 vs. CD38 (c) CD38 vs. CD19

Fig. 2. Typically, flow cytometric analysis is performed using multiple 2-dimensional projections onto the various marker combinations. This can lead to ambiguity and does not fully exploit the high-dimensional nature of the data. We illustrate this difficulty in distinguishing a patient with an unfavorable immunophenotype to that of a favorable patient, using their marginal PDFs over 3 of a possible 15 marker combinations from the 6-marker assay.

– marginal probability density functions (PDFs) – defined by any 2 of the 6 characteristics measured in a given tube (see Fig. 2). Often one or more measured characteristics are used to restrict immunophenotypic analysis to a specific subset of cells in a process commonly known as *gating*, which allows for limited exploitation of the dimensionality of the flow cytometry data.

The use of each single measured characteristic as an axis on a 2-dimensional histogram is a convenient method for visualizing results and observing relationships between cell surface markers, but is equivalent to viewing a geometric shape head-on, and therefore does not necessarily take full advantage of the multidimensional nature of flow cytometry. Just as it is possible to rotate an object in space to more effectively observe that object's characteristics, so too is it possible to "rotate" the 2-dimensional projection of a 6-dimensional flow cytometry analysis to optimally view the relationships among the 6 measured characteristics.

A. Problem Formulation

Given the critical importance of visualization in the task of flow cytometric diagnosis, we wish to find the low-dimensional projection which best preserves the relationships between patient data sets. Rather than viewing a series of axes projections determined by clinical experience as in Fig. 2 (where we illustrate only 3 of the 15 possible axes projections of the 6-dimensional data set), a projection which is a linear combination of several biomarkers will allow a clinician to visualize all of the data in a single low-dimensional space, with minimal loss of information.

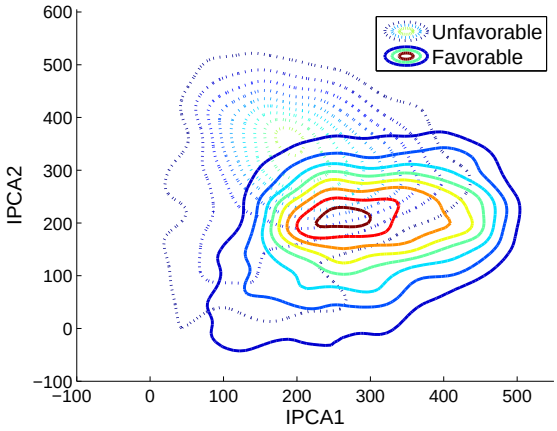


Fig. 3. Projecting the same data in Fig. 2 down to 2-dimensions using a linear combination of all available markers. It is a much easier task to discriminate these joint PDFs.

An example is shown in Fig. 3, where it is easy to differentiate the patient with an unfavorable immunophenotype from that of a favorable patient[1]. This ability becomes even more substantial with ever advancing technology, leading to flow cytometers that have 9 or more available parameters (yielding 36+ two-dimensional plots for analysis).

Specifically, given a collection of flow cytometer outputs $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ in which each element of $\boldsymbol{X}_i$ exists in $\mathbb{R}^d$, we can define similarity between data sets $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ (e.g. patients i and j) with some metric as $D(\boldsymbol{X}_i, \boldsymbol{X}_j)$. Can we find a mapping

$$A : \boldsymbol{X} \rightarrow \boldsymbol{Y}$$

in which the elements of $\boldsymbol{Y}$ exist in $\mathbb{R}^m$, $m < d$ ($m = 2$ or 3 for visualization) such that

$$D(\boldsymbol{X}_i, \boldsymbol{X}_j) = D(\boldsymbol{Y}_i, \boldsymbol{Y}_j), \ \forall\, i, j?$$

Can we define this mapping as a linear projection $A \in \mathbb{R}^{m \times d}$? Can we ensure that the projection minimally alters the data itself (i.e. ensure A is orthonormal)? Additionally, by analyzing the loadings in A, can we determine which biomarkers are best at differentiating between disease classes?

[1]The data presented here is from patients with chronic lymphocytic leukemia, and is further explained in Section IV-B

III. METHODS

In our previous work on Fisher Information Nonparametric Embedding (FINE) [16], [17], we have shown that we can derive an information-based embedding for the purposes of flow cytometric analysis [7]. By viewing each patient as a probability density function on a statistical manifold, we were able to embed that manifold into a low-dimensional Euclidean space, in which each patient is represented by a single point. This visualization allows for a diagnostician to view each patient in relation to other selected patients in a space where disease classes are well distinguished. The similarity between patients was determined by using an approximation of the Fisher information distance between PDFs parameterized by $\theta = [\theta^1, \ldots, \theta^n]$. The Fisher information distance between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ is:

$$D_F(\theta_1,\theta_2) = \min_{\substack{\theta(\cdot):\\ \theta(0)=\theta_1\\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{dt}\right)^T [\mathcal{I}(\theta)] \left(\frac{d\theta}{dt}\right)}\, dt, \tag{1}$$

where $\theta = \theta(t)$ is the parameter path along the manifold [18], [19] and $[\mathcal{I}(\theta)]$ is the Fisher information matrix whose elements are

$$[\mathcal{I}(\theta)]_{ij} = \int f(X;\theta) \frac{\partial \log f(X;\theta)}{\partial \theta^i} \frac{\partial \log f(X;\theta)}{\partial \theta^j}\, dX. \tag{2}$$

The Fisher information distance is the best way to characterize similarity between PDFs as it is an exact measure of the geodesic (i.e. shortest path) between points along the manifold. While the Fisher information distance cannot be exactly computed without knowing the parameterization of the manifold, it may be approximated with metrics such as the Kullback-Leibler (KL) divergence, Hellinger distance, and Rényi-alpha entropy [18]. For our work, we focus on the KL divergence, which is defined as

$$KL(p_1\|p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx, \tag{3}$$

where p_1 and p_2 are PDFs of possibly unknown parameterization. It should be noted that the KL divergence is not a distance metric, as it is not symmetric, $KL(p_1\|p_2) \neq KL(p_1\|p_2)$. To obtain this symmetry, we will define the KL divergence as:

$$D_{KL}(p_1,p_2) = KL(p_1\|p_2) + KL(p_2\|p_1). \tag{4}$$

The KL divergence approximates the Fisher information distance [18],

$$\sqrt{D_{KL}(p_1,p_2)} \to D_F(p_1,p_2), \tag{5}$$

as $p_1 \to p_2$. Hence, we are able to use the KL divergence as a means for calculating similarity between patient data sets. For our purposes, we choose to nonparametrically estimate patient PDFs and KL divergences through kernel density estimation [17], although other methods are available (e.g. mixture models, k-nearest neighbor methods).

A. Objective Function

In FINE, we found an embedding which mapped information distances between PDFs as Euclidean distances in a low-dimensional space. This allowed us to embed an entire PDF, and therefore all of the cells which were realizations of that PDF, into a single low-dimensional vector. This provided for the direct comparison of patients in the same normalized space. In our current task, we are not interested in embedding a group of patients into the same space, but rather projecting each patient individually in its own space. However, it is important that we maintain differences between patients, as we have found that is a great way to differentiate disease classes.

We define our *Information Preserving Component Analysis (IPCA)* projection as one that preserves the Fisher information distance between data sets. Specifically, let $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ where $\boldsymbol{X}_i$ corresponds to the flow cytometer output of the i^{th} patient containing n_i blood cells measured with d different markers; estimating the PDF of $\boldsymbol{X}_i$ as p_i. With an abuse of notation, we refer to $D_{KL}(p_i, p_j)$ as $D_{KL}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ with the knowledge that the divergence is calculated with respect to PDFs, not realizations. We wish to find a single projection matrix A such that

$$D_{KL}(A\boldsymbol{X}_i, A\boldsymbol{X}_j) = D_{KL}(\boldsymbol{X}_i, \boldsymbol{X}_j),\ \forall\, i, j.$$

Formatting as an optimization problem, we would like to solve:

$$A = \arg \min_{A:AA^T=I} \|D(\mathcal{X}) - D(\mathcal{X}; A)\|_F^2, \tag{6}$$

where I is the identity matrix, $D(\mathcal{X})$ is a dissimilarity matrix such that $D_{ij}(\mathcal{X}) = D_{KL}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, and $D(\mathcal{X}; A)$ is a similar matrix where the elements are perturbed by A, i.e. $D_{ij}(\mathcal{X}; A) = D_{KL}(A\boldsymbol{X}_i, A\boldsymbol{X}_j)$.

Since pathologists view projections in order diagnose based on similar marker expression patterns, maintaining similarities within disease class (and differences between class) is of the utmost importance. These measures are expressed quantitatively through information. By finding

the projection solving the objective function (6), we ensure that the amount of information between patients which is lost due to the projection is minimized.

B. Gradient Descent

Gradient descent (or the method of *steepest* descent) allows for the solution of convex optimization problems by traversing a surface or curve in the direction of greatest change, iterating until the minimum is reached. Specifically, let $J(x)$ be a real-valued objective function which is differentiable about some point x_i. The direction in which $J(x)$ decreases the fastest, from the point x_i, is that of the negative gradient of J at x_i, $-\frac{\partial}{\partial x}J(x_i)$. By calculating the location of the next iteration point as

$$x_{i+1} = x_i - \mu\frac{\partial}{\partial x}J(x_i),$$

where μ is a small number regulating the step size, we ensure that $J(x_i) \geq J(x_{i+1})$. Continued iterations will result in $J(x)$ converging to a local minimum. Gradient descent does not guarantee that the process will converge to an absolute minimum, so typically it is important to initialize x_0 near the estimated minimum.

Let $J(A) = \|D(\mathcal{X}) - D(\mathcal{X}; A)\|_F^2$ be our objective function, measuring the error between our projected subspace and our full-dimensional space. The direction of the gradient is solved by taking the partial derivative of J w.r.t. a projection matrix A,

$$\frac{\partial}{\partial A}J(A) = \sum_i \sum_j \frac{\partial}{\partial A}\left[D_{ij}(\mathcal{X}; A)^2 - 2D_{ij}(\mathcal{X})D_{ij}(\mathcal{X}; A)\right].$$

Given the direction of the gradient, the projection matrix can be updated as

$$A = A - \mu\frac{\partial}{\partial A}\tilde{J}(A), \tag{7}$$

where

$$\frac{\partial}{\partial A}\tilde{J}(A) = \frac{\partial}{\partial A}J(A) + Q_0 A + \mu Q_1 A$$

is the direction of the gradient, constrained to force A to remain orthonormal. Variables Q_0 and Q_1 are defined as

$$Q_0 = -\frac{1}{2}\left(\left(\frac{\partial}{\partial A}J(A)\right)A^T + A\left(\frac{\partial}{\partial A}J(A)\right)^T\right)$$

$$Q_1 = \frac{1}{2}\left(\frac{\partial}{\partial A}J(A) + Q_0 A\right)\left(\frac{\partial}{\partial A}J(A) + Q_0 A\right)^T.$$

Algorithm 1 Information Preserving Component Analysis

Input: Collection of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, $\boldsymbol{X}_i \in \mathbb{R}^{d \times n_i}$; the desired projection dimension m; search step size μ

1: Calculate $D(\mathcal{X})$, the Kullback-Leibler dissimilarity matrix
2: Initialize $A_1 \in \mathbb{R}^{m \times d}$ as an orthonormal projection matrix
3: Calculate $D(\mathcal{X}; A_1)$, the Kullback-Leibler dissimilarity matrix in the projected space
4: **for** $i = 1$ to ∞ **do**
5: Calculate $\frac{\partial}{\partial A_i}\tilde{J}(A_i)$, the direction of the gradient, constrained to $A_i A_i^T = I$
6: $A_{i+1} = A_i - \mu \frac{\partial}{\partial A_i}\tilde{J}(A_i)$
7: Calculate $D(\mathcal{X}; A_{i+1})$
8: $J(A_{i+1}) = \|D(\mathcal{X}) - D(\mathcal{X}; A_{i+1})\|_F^2$
9: Repeat until convergence of J
10: **end for**

Output: Projection matrix $A \in \mathbb{R}^{m \times d}$, which preserves the information distances between sets in $\mathcal{X}$.

The full derivation of this constraint can be found in Appendix A. This process of gradient descent is iterated until the error $J(A)$ converges.

C. Algorithm

The full method for IPCA is described in Algorithm 1. We note that typically A is initialized as a random orthonormal projection matrix due to the desire to not bias the estimation. While this may result in finding a local minimum rather than an absolute minimum, experimental results on our available flow cytometry data have shown that the algorithm converges near the same result given several random initializations. If *a priori* knowledge of the global minimum was available, one would initialize A in its vicinity. At this point we stress that we utilize gradient descent due to its ease of implementation. There are more efficient methods of optimization, but that is out of the scope of the current contribution and is an area for future work.

D. Variable Selection

One immediate benefit of IPCA is that we may use the loading vectors of A towards the problem of variable selection. IPCA finds the linear combination of channels which best preserves the information divergence between patient data sets (i.e. realizations of PDFs). Given the definition of the Kullback-Leibler divergence (3), the dimensions which contribute most to the information are those in which data sets differ most in probability distribution. For example, if two multivariate PDFs p and q are independent and identically distributed in a certain dimension, that dimension will offer zero contribution to the KL divergence between p and q. When finding a projection which preserves the information divergence between p and q, A is going to be highly weighted towards the variables which contribute most to that distance. Hence, the loading vectors of A essentially give a ranking of the discriminative value of each variable. This form of variable selection is useful in exploratory data analysis.

IV. Simulations

We now present simulation results for using IPCA to find a projection matrix for flow cytometric data analysis. We demonstrate three distinct studies involving differing disease classes to show that our methods are not just beneficial to a single example. We offer a proof of concept that shall allow pathologists to utilize our methods on many different studies and for exploratory data analysis. In all cases, patient data was obtained and diagnosed by the Department of Pathology at the University of Michigan.

A. Lymphoid Leukemia Study

For our first study, we will compare patients with two distinct but immunophenotypically similar forms of lymphoid leukemia – mantle cell lymphoma (MCL) and chronic lymphocytic leukemia (CLL). These diseases display similar characteristics with respect to many expressed surface antigens, but are generally distinct in their patterns of expression of two common B lymphocyte antigens: CD23 and FMC7. Typically, CLL is positive for expression of CD23 and negative for expression of FMC7, while MCL is positive for expression of FMC7 and negative for expression of CD23. These distinctions should lead to a difference in densities between patients in each disease class.

Dimension	Marker
1	Forward Light Scatter
2	Side Light Scatter
3	FMC7
4	CD23
5	CD45
6	Empty

TABLE I

DATA DIMENSIONS AND CORRESPONDING MARKERS FOR ANALYSIS OF CLL AND MCL.

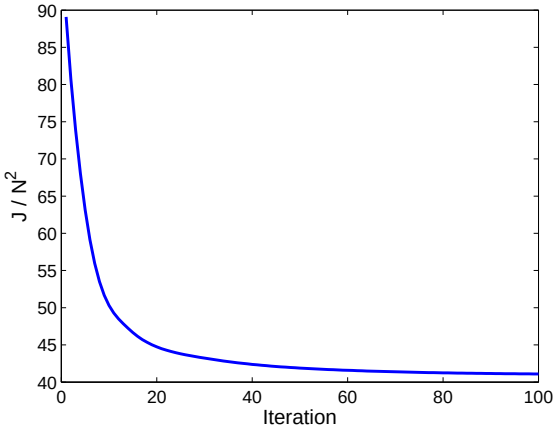
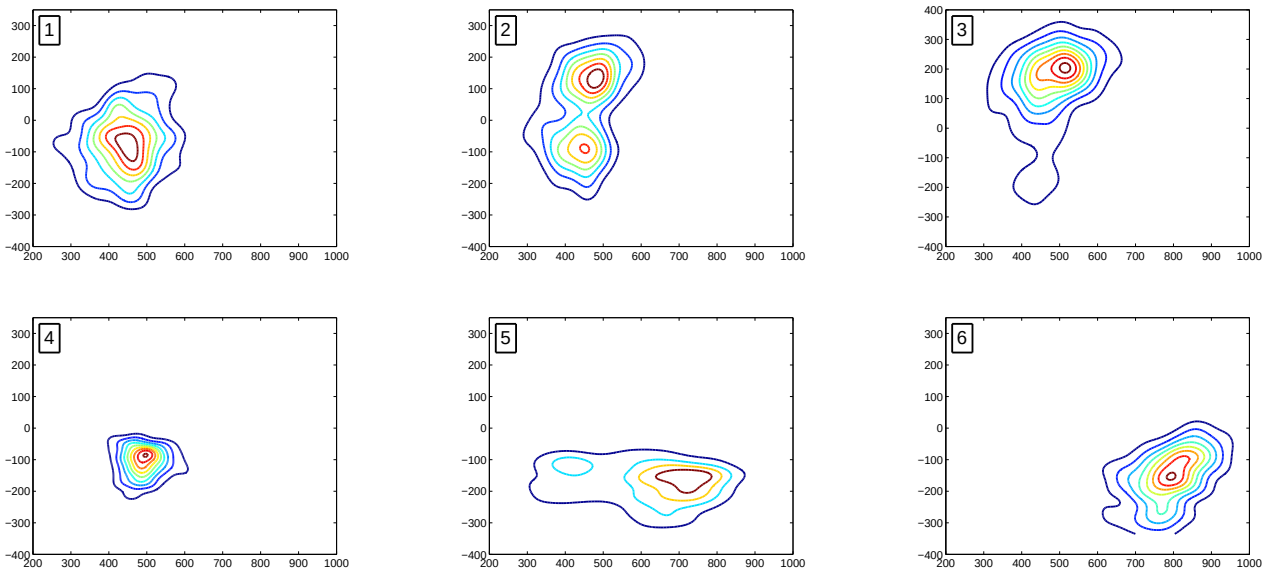


Fig. 4. CLL and MCL Study: Evaluating the objective as a function of time. As the iterations increase, the objective function eventually converges.

The data set $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{43}\}$ consists of 43 patients, 23 of which have been diagnosed with CLL and 20 diagnosed with MCL. Each $\boldsymbol{X}_i$ is a 6-dimensional matrix, with each dimension corresponding to a different marker (see Table I), and each element representing a unique blood cell, totaling $n_i \sim 5000$ total cells per patient. We calculate $D(\mathcal{X})$, the matrix of Kullback-Leibler similarities, and desire to find the projection matrix A that will preserve those similarities when all data sets are projected to dimension $d = 2$.

Using the methods described in this paper, we found the IPCA projection as

$$A = \begin{pmatrix} -0.1177 & 0.0693 & 0.8979 & 0.2513 & 0.3346 & -0.0032 \\ 0.0077 & -0.2678 & -0.1541 & 0.9243 & -0.2224 & 0.0270 \end{pmatrix}. \tag{8}$$

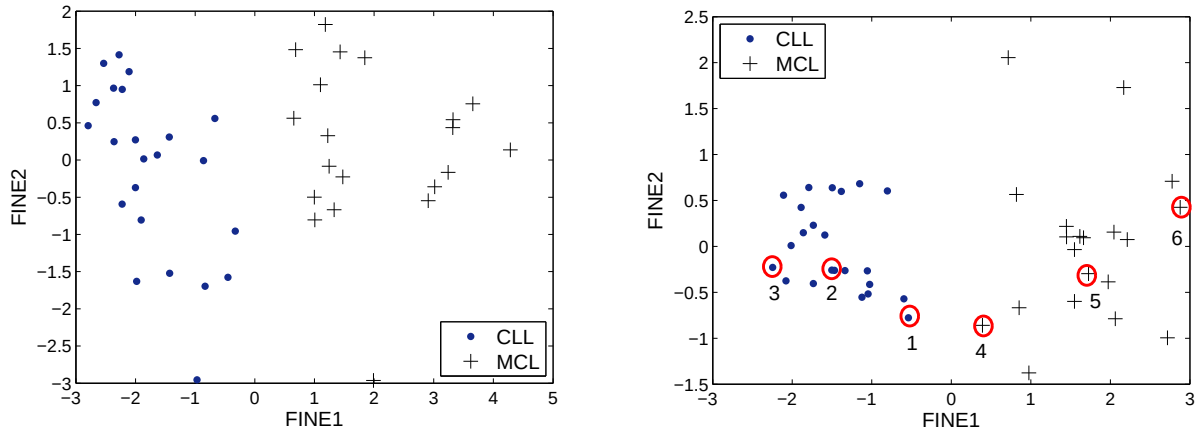


(a) Most Similar (b) Centroids (c) Least Similar

Fig. 5. CLL and MCL Study: Contour plots (i.e. PDFs) of the IPCA projected data. The top row corresponds to the PDFs the CLL patients, while the bottom row represents PDFs of MCL patients. The selected patients are those most similar between disease classes, the centroids of disease classes, and those least similar between disease classes, as highlighted in Fig. 6(b).

This projection was calculated by minimizing the objective function with respect to A, as illustrated in Fig. 4 in which the squared error (per element pair) is plotted as a function of time. As the iteration i increases, J converges and A_i is determined to be the IPCA projection matrix. We note that while dimension 6 corresponds to no marker (it is a channel of just noise), we do not remove the channel from the data sets, as the projection determines this automatically (i.e. loading values approach 0). Additionally, due to computational complexity issues, each data set was randomly subsampled such that $n_i = 500$. While we would not necessarily suggest this decimation in practice, we have found it to have a minimal effect during experimentation.

Given the IPCA projection, we illustrate the 2-dimensional PDFs of several different patients in the projected space in Fig. 5. We selected patients based on the KL divergence values between patients of different disease class. Specifically, we selected the CLL and MCL patients with a small divergence (i.e. most similar PDFs), patients with a large divergence (i.e. least similar PDFs), and patients which represented the centroid of each disease class. These low-dimensional PDFs, which are what would be utilized by a diagnostician, are visibly different between disease classes. While the most similar CLL and MCL patients do share much similarity in their IPCA



(a) Full Dimension (b) IPCA Projection

Fig. 6. CLL and MCL Study: Comparison of embeddings, obtained with FINE, using the full dimensional data and the data projected with IPCA. IPCA preserves the separation between disease classes. The circled points correspond to the density plots in Fig. 5, numbered respectively.

PDFs, there is still a significant enough difference to distinguish them, especially given the similarities to other patient PDFs.

We now illustrate the embedding of the projected data obtained with FINE, which performs classical multidimensional scaling on the matrix of dissimilarities formed by the KL divergence (see [7], [17] for additional details). The embedding results are shown in Fig. 6(b), in which the separation between classes is preserved when using the projected data as compared to using the full-dimensional data in Fig. 6(a). Each point represents an entire patient data set, and those which are circled correspond to the PDFs shown in Fig. 5. By finding the projection which minimizes the difference in KL divergence between the full and projected data, we maintain the relationships between different sets, allowing for a consistent analysis.

Using the projection matrix (8) for variable selection, the loading vectors are highly concentrated towards the 3^{rd} and 4^{th} dimensions, which correspond to fluorescent markers FMC7 and CD23. We acknowledge that this marker combination is well known and currently utilized in the clinical pathology community for differentiating CLL and MCL[2]. We stress, however, that what had previously been determined through years of clinical experience was able to be

[2]CD45 and light scatter characteristics are often used as gating parameters for selection of lymphocytes among other cell types prior to analysis, but CD23 and FMC7 are the main analytical biomarkers in this 3-color assay.

Dimension	Marker
1	Forward Light Scatter
2	Side Light Scatter
3	CD5
4	CD38
5	CD45
6	CD19

TABLE II

DATA DIMENSIONS AND CORRESPONDING MARKERS FOR ANALYSIS OF CLL.

independently validated quickly using IPCA. This is important as it could enable pathologists to experiment with new combinations of fluorescent markers and see which may have strong effects on the discernment of similar leukemias and lymphomas.

B. Chronic Lymphocytic Leukemia Study

Continuing our study of patients with chronic lymphocytic leukemia, we wish to determine subclasses within the CLL disease class. Specifically, we now use IPCA to find a low-dimensional space which preserves the differentiation between patients with good and poor prognoses (i.e. favorable and unfavorable immunophenotypes). Literature [20] has shown that patients whose leukemic cells are strong expressors of CD38 have significantly worse survival outcome. Genotypic studies have shown that the absence of somatic mutation within immunoglobulin genes of CLL cells (a so-called "pre-follicular" genotype) is a potent predictor of worse outcome. High levels of CD38 expression are an effective surrogate marker for the absence of somatic immunoglobulin gene mutation, and also have been shown to be an independent predictor of outcome in some studies. Since patients can generally be stratified by CD38 expression levels, and CD38 has been shown to emerge as a defining variable of CLL subsets in hierarchical immunophenotypic clustering [21], we would expect IPCA to localize the CD38 variable as one of importance when analyzing CLL data.

Using the same patients (those diagnosed with CLL) as in the above simulation, we define $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{23}\}$, where each $\boldsymbol{X}_i$ was analyzed with by the series of markers in Table II.

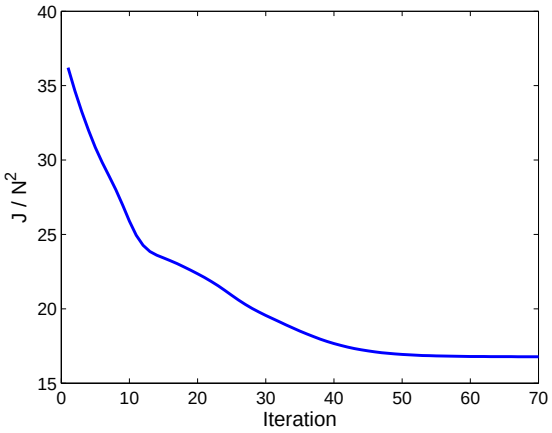


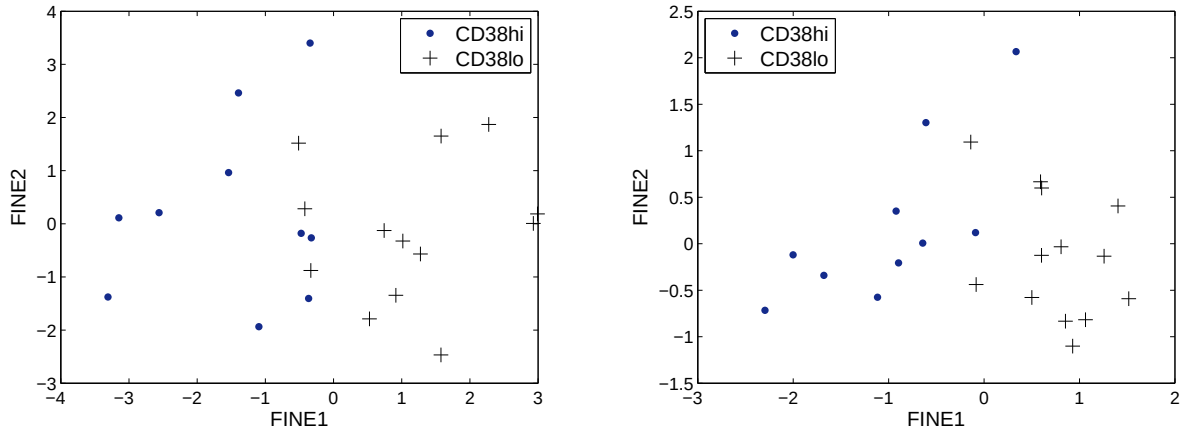
Fig. 7. CLL Prognosis Study: The value of the objective function vs. time.

Minimizing the objective function (see Fig. 7), we calculate the IPCA projection matrix as

$$A = \begin{pmatrix} -0.0700 & 0.0950 & 0.5006 & -0.8361 & 0.1834 & -0.0519 \\ -0.1705 & -0.0434 & -0.3775 & -0.0988 & 0.6992 & 0.5727 \end{pmatrix}.$$

This projection matrix has very high loadings in variables 4, 5, and 6, which correspond to markers CD38, CD45, and CD19 respectively. This identifies the isolation of B cells by CD19 expression (a B lymphocyte restricted antigen always expressed on CLL cells) and assessment of CD38 on these B cells. As expected, we identify CD38 as a marker of importance in differentiating patient groups. We also identify the possibility that CD45 and CD19 expression are areas which may help prognostic ability. This is an area for further interrogation.

Using FINE to embed the data (Fig. 8) for comparative visualization, we see that the IPCA projection preserves the grouping of patients with unfavorable immunophenotype (CD38hi) and favorable immunophenotype (CD38lo). CD38hi versus CD38lo for each patient was determined using cutoff values endorsed in the literature [20]. Although complete follow-up data for this retrospective cohort were not available, the findings were indirectly further validated by the fact that, of the patients with follow-up information available, zero of six CD38lo patients died, while four of nine CD38hi patients died within a median follow-up interval of 25 months (range 1 to 102 months). Hence, we find that IPCA can help identify sub-classes and may be useful for possible help towards prognosis.



(a) Full Dimensional

(b) IPCA Projection

Fig. 8. CLL Prognosis Study: Comparison of embeddings, obtained with FINE, using the IPCA projection matrix A and the full dimensional data. The patients with a poor prognosis (CD38hi) are generally well clustered against those with a favorable prognosis (CD38lo) in both embedddings.

C. Acute Lymphoblastic Leukemia vs. Hematogone Hyperplasia Study

We now demonstrate a study involving the diseases acute lymphoblastic leukemia (ALL) and a benign condition known as hematogone hyperplasia (HP). ALL is marked by the neoplastic proliferation of abnormal lymphocyte precursors (lymphoblasts). Our study specifically focused upon ALL consisting of B cell precursor lymphobalsts (B-precursor ALL), the most common form of this disease, since the normal counterpart to B-precursor lymphoblasts, termed hematogones, are detectable in the bone marrow of most healthy individuals, and hematogones can proliferate in benign reversible fashion in numerous clinical states [22]. The distinction between hematogones and leukemic B-precursor lymphoblasts is highly relevant in clinical practice since these cell types exhibit substantial immunophenotypic overlap, many transient conditions associated with hematogone hyperplasia can present with clinical suspicion for leukemia, and patients with ALL can develop HP during recovery from chemotherapy for their leukemia.

For this study, let us define the data set $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{54}\}$, which consists of 54 patients, 31 of which have been diagnosed with ALL and 23 diagnosed with HP. Patient samples were analyzed with a series of markers (see Table III) designed for the isolation of hematogones and aberrant lymphoblast populations, based on known differential patterns of these markers in these cell types. Specific details of how the data was retrieved can be found in [7].

Dimension	Marker
1	Forward Light Scatter
2	Side Light Scatter
3	CD38
4	CD19
5	CD45
6	CD10

TABLE III

DATA DIMENSIONS AND CORRESPONDING MARKERS FOR ANALYSIS OF ALL AND HP.

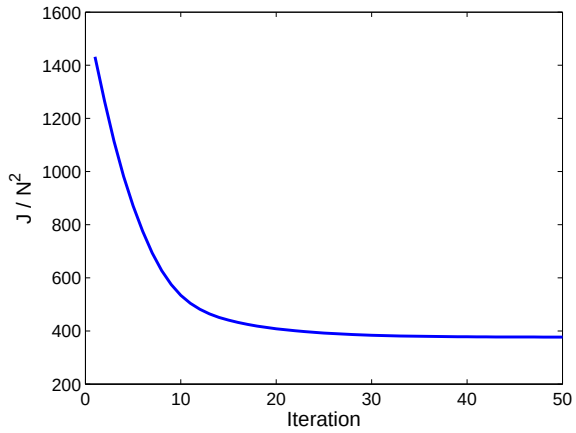


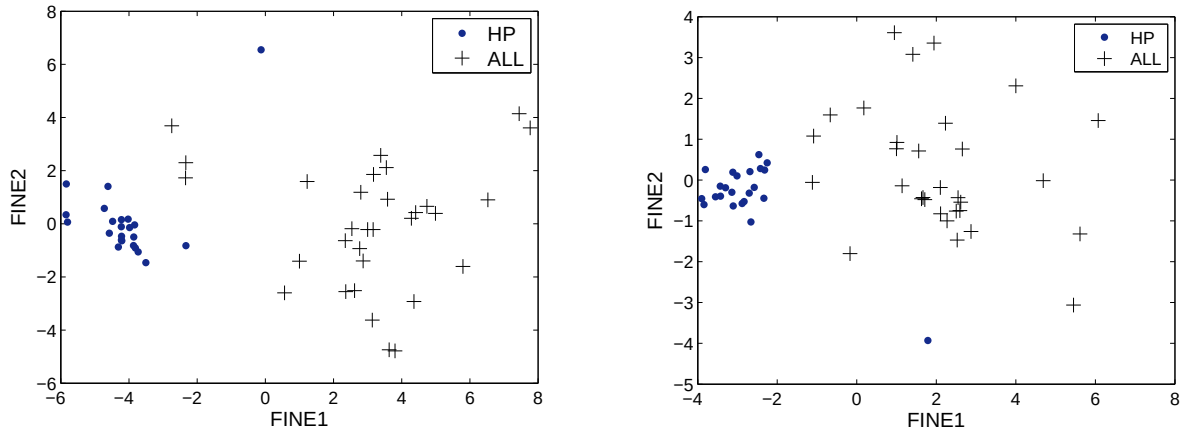
Fig. 9. The value of the objective function (vs. time) for the analysis of ALL and HP diagnosis.

By minimizing the objective function (Fig. 9), we find the IPCA projection as

$$A = \begin{pmatrix} -0.1805 & -0.1448 & 0.8691 & 0.0848 & 0.4084 & 0.1310 \\ -0.0336 & 0.1143 & -0.0291 & 0.2506 & -0.2608 & 0.9242 \end{pmatrix}.$$

Using FINE, we compare the embedding of the full-dimensional data to that of the projected data in Fig. 10. The embeddings are very similar, which illustrates once again that IPCA preserves the similarities between different sets. This allows for a low-dimensional analysis in the projected space with the security of knowing the relationships between patients have been minimally effected.

We also observe that the IPCA projection matrix has strong loadings corresponding to markers CD38 and CD10. In clinical practice, it is often noted that hematogones have a very uniform and



(a) Full Dimensional (b) IPCA Projection

Fig. 10. ALL and HP Study: Comparison of embeddings, obtained with FINE, using the full dimensional data and the IPCA projection matrix A. The embedding is very similar when using the projected data, which preserves the similarities between patients.

strong CD38 expression pattern, while lymphoblasts can have quite a range of CD38 expression [22]. This analysis seems to provide independent validation for that observation. Furthermore, this analysis identifies CD10 as a principal distinguishing marker among the others analyzed in this 4-color assay. This finding is not intuitive, since in day-to-day practice CD10 is not obviously of greater distinguishing value than marker such as CD45 or side angle light scatter. These markers, like CD10, are used for their different expression patterns in lymphoblasts versus hematogones, but that may show considerable overlap in expression intensity between these two cell types. Our identification of CD10 as a marker of importance identifies an area for further clinical investigation.

D. Performance Comparison

We now compare IPCA to the PCA and ICA projection matrices for the preceding studies. Given that an ultimate task is visualization for diagnosis and validation, it is important that the disease classes are easily distinguished. For our comparison, we utilize the Bhattacharya distance to measure how distinguishable the "worst case" scenarios are in the projected space – essentially we desire the most similar patients in differing disease classes (i.e. "worst case") to have as little similarity as possible. The Bhattacharya distance has been used to bound classification error in

Study	DR Method		
	IPCA	PCA	ICA
Lymphoid	**0.1573**	0.0821	0.0220
CLL	**0.0550**	0.0409	0.0326
ALL/HP	**0.0624**	0.0532	0.0335

TABLE IV

'WORST CASE' PERFORMANCE COMPARISON OF DIMENSION REDUCTION (DR) METHODS FOR FLOW CYTOMETRY STUDIES. RESULTS REPORTED FOR EACH CASE STUDY ARE OF THE LOWEST VALUES OF THE BHATTACHARYA DISTANCE BETWEEN PATIENT PAIRS WITH DIFFERING DISEASES IN THE PROJECTED SPACE. IPCA OUTPERFORMS BOTH PCA AND ICA IN ALL CASES.

dimension reduction problems [23], and is directly related to the Chernoff performance bound [13]. Results are illustrated in Table IV, where the best performance is emphasized (larger numbers are more desirable). It is clear that IPCA consistently outperforms both other methods of dimension reduction; concluding that the projection subspace defined by IPCA is best at distinguishing between disease types. Although we do not present them here, we have observed similar results with several other measures of probabilistic distance and cluster similarity. Note that ICA was performed using the FastICA algorithm [15], and the data was pre-processed by whitening and PCA in accordance with [24].

E. Subsampling Performance

One concern when implementing IPCA is the number of data sets necessary to find a proper projection. Specifically, given a subset of patients $\mathcal{X}_S \subset \mathcal{X}$, how close does IPCA approach the value of the objective function obtained when utilizing the entire patient collection $\mathcal{X}$? To determine this, we return to our lymphoid leukemia study and subsample from $\mathcal{X}$, with N_S patients randomly selected from each disease class ($N_S \in [2, 5, 10, 15]$), and use IPCA to determine the projection matrix A. We then calculate the value of the objective function on the entire set $\mathcal{X}$ projected by A. The mean results over a 10-fold cross validation are illustrated in Fig. 11, where we signify the value of the objection function when using IPCA on the entire data set with the dashed line. Given that the value of the objection function with the initial random projection matrix was $\frac{J}{N^2} = 89.0802$, the relative performance of IPCA with few available data

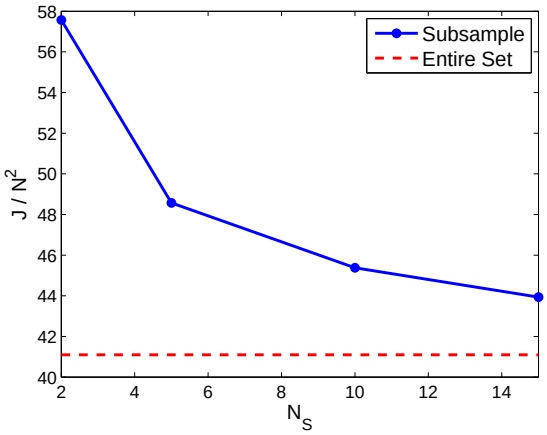


Fig. 11. IPCA performance using subset of patients $\mathcal{X}_S \subset \mathcal{X}$ from the lymphoid leukemia collection, where N_S is the number of randomly selected patients from each disease class. Results shown over a 10-fold cross validation, with the IPCA projection determined by $\mathcal{X}$ shown as a lower bound with the dotted line.

sets is promising.

V. Conclusions

In this paper we have shown the ability to find an information-based projection for flow cytometric data analysis using Information Preserving Component Analysis (IPCA). By preserving the Fisher information distance between oncological data sets (i.e. patients), we find a low-dimensional projection that allows for visualization in which the data is discernable between cancerous disease classes. As such, we use machine-learning to provide a projection space that is usable for verification of cancer diagnosis. Additionally, analysis of the loading vectors in the projection matrix allows for a means of variable selection. We have shown independent verification for determining optimal marker combinations in distinguishing immunophenotypically similar cancers, as well as validating variables which help to identify prognostic groups. Verifying these known results through independent methods provides a solid *proof-of-concept* for the ability to utilize IPCA for exploratory research of different marker assays.

In future work we plan to study the effects of preserving only the local distances between data sets. As we have stated, the KL divergence becomes a weak approximation as the densities separate on the statistical manifold. As such, performance may improve by putting more emphasis on preserving the close distances. However, this may have the adverse effect of diminishing the

ability to distinguish between disease classes if they are well separated, as those far distances may not be well preserved. Additionally, we would like to utilize different methods for optimizing the cost function. While we currently utilize gradient descent for ease of implementation, it is relatively slow and there are more efficient methods to use (e.g. fixed point iteration). The optimization method is not the focus of our work, but faster methods may be required for practical usage. Finally, we would like to apply our methods towards exploratory research and determine other applications of interest.

APPENDIX

A. *Orthonormality Constraint on Gradient Descent*

We derive the orthonormality constraint for our gradient descent optimization in the following manner; solving

$$A = \arg \min_{A:AA^T=I} J(A),$$

where I is the identity matrix. Using Lagrangian multiplier M, this is equivalent to solving

$$A = \arg\min_{A} \tilde{J}(A),$$

where $\tilde{J}(A) = J(A) + \mathrm{tr}(A^T M A)$. We can iterate the projection matrix A, using gradient descent, as:

$$A \leftarrow A - \mu \frac{\partial}{\partial A} \tilde{J}(A), \tag{9}$$

where $\frac{\partial}{\partial A}\tilde{J}(A) = \frac{\partial}{\partial A}J(A) + (M + M^T)A$ is the gradient of the cost function w.r.t. matrix A. To ease notation, let $\Delta \triangleq \frac{\partial}{\partial A}J(A)$ and $\tilde{\Delta} \triangleq \frac{\partial}{\partial A}\tilde{J}(A)$. Continuing with the constraint $AA^T = I$, we right-multiply (9) by A^T and obtain

$$0 = -\mu A\tilde{\Delta}^T - \mu\tilde{\Delta}A^T + \mu^2\tilde{\Delta}\tilde{\Delta}^T,$$

$$\mu\tilde{\Delta}\tilde{\Delta}^T = \tilde{\Delta}A^T + A\tilde{\Delta}^T, \tag{10}$$

$$\mu(\Delta + (M + M^T)A)(\Delta + (M + M^T)A)^T = (\Delta A(M + M^T)A)A^T + A(\Delta A^T(M + M^T)A).$$

Let $Q = M + M^T$, hence $\tilde{\Delta} = \Delta + QA$. Substituting this into (10) we obtain:

$$\mu(\Delta\Delta^T + QA\Delta^T + \Delta A^T Q + QQ^T) = \Delta A^T + A\Delta^T + 2Q. \tag{11}$$

Next we substitute the Taylor series expansion of Q around $\mu = 0$ back into (11): $Q = \sum_{j=0}^{\infty} \mu^j Q_j$. The dependence of Q on μ is somewhat artificial, but helps to establish a relationship between Q and the gradients. By equating corresponding powers of μ (i.e. setting $\frac{\partial^j}{\partial \mu^j}|_{\mu=0}(\cdot) = 0$), we identify:

$$Q_0 = -\frac{1}{2}(\Delta A^T + A\Delta^T),$$

$$Q_1 = \frac{1}{2}(\Delta + Q_0 A)(\Delta + Q_0 A)^T.$$

Replacing the expansion of Q in $\tilde{\Delta} = \Delta + QA$:

$$\tilde{\Delta} = \Delta - \frac{1}{2}(\Delta A^T + A\Delta^T)A + \mu\, Q_1 A + \mu^2\, Q_2 A + \ldots.$$

Finally, we would like to assure a sufficiently small step size to control the error in forcing the constraint due to a finite Taylor series approximation of Q. Using the L_2 norm of $\tilde{\Delta}$ allows us to calculate an upper bound on the Taylor series expansion:

$$\|\tilde{\Delta}\| \;\leq\; \|\Delta - \frac{1}{2}(\Delta A^T + A\Delta^T)A\| + \mu\, \|Q_1 A\| + \mu^2\, \|Q_2 A\| + \ldots.$$

We condition the norm of the first order term in the Taylor series approximation to be significantly smaller than the norm of the zeroth order term. If $\mu \ll \|\Delta - \frac{1}{2}(\Delta A^T + A\Delta^T)A\|/\|Q_1 A\|$ then:

$$\frac{\partial}{\partial A}\tilde{J}(A) = \frac{\partial}{\partial A}J(A) + Q_0 A + \mu Q_1 A, \tag{12}$$

where

$$Q_0 = -\frac{1}{2}\left(\left(\frac{\partial}{\partial A}J(A)\right)A^T + A\left(\frac{\partial}{\partial A}J(A)\right)^T\right)$$

$$Q_1 = \frac{1}{2}\left(\frac{\partial}{\partial A}J(A) + Q_0 A\right)\left(\frac{\partial}{\partial A}J(A) + Q_0 A\right)^T,$$

is a good approximation of the gradient constrained to $AA^T = I$. We omit the higher order terms as we experimentally find that they are unnecessary, especially as even $\mu^3 \to 0$. We note that while there are other methods for forcing the gradient to obey orthogonality [25], [26], we find our method is straightforward and sufficient for our purposes.

REFERENCES

[1] Q. T. Zeng, J. P. Pratt, J. Pak, D. Ravnic, H. Huss, and S. J. Mentzer, "Feature-guided clustering of multi-dimensional flow cytometry datasets," *Journal of Biomedical Informatics*, vol. 40, pp. 325–331, 2007.

[2] E. Zamir, B. Geiger, N. Cohen, Z. Kam, and B. Katz, "Resolving and classifying haematopoietic bone-marrow cell populations by multi-dimensional analysis of flow-cytometry data," *British Journal of Haematology*, vol. 129, pp. 420–431, 2005.

[3] M. Roederer and R. Hardy, "Frequency difference gating: A multivariate method for identifying subsets that differ between samples," *Cytometry*, vol. 45, no. 1, pp. 56–64, 2001.

[4] M. Roederer and R. Hardy, "Probability binning comparison: A metric for quantitating multivariate distribution differences," *Cytometry*, vol. 45, no. 1, pp. 47–55, 2001.

[5] R. C. Mann, D. M. Popp, and R. E. Hand Jr., "The use of projections for dimensionality reduction of flow cytometric data," *Cytometry*, vol. 5, no. 3, pp. 304–307, 1984.

[6] R. C. Mann, "On multiparameter data analysis in flow cytometry," *Cytometry*, vol. 8, no. 2, pp. 184–189, 1987.

[7] W. G. Finn, K. M. Carter, R. Raich, and A. O. Hero, "Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: Treating flow cytometry data as high-dimensional objects," *Cytometry Part B: Clinical Cytometry*, vol. 76B, no. 1, Jan. 2009.

[8] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, March 1989.

[9] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Neural Networks for Signal Processing Workshop*, 1999.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[11] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 1, pp. 2323–2326, 2000.

[12] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990, 2nd edition.

[14] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, vol. c-23, no. 9, pp. 881–890, September 1974.

[15] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, NY, USA, 2001.

[16] K. M. Carter, R. Raich, and A. O. Hero, "Fine: Information embedding for document classification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, April 2008, pp. 1861–1864.

[17] K. M. Carter, R. Raich, and A. O. Hero, "An information geometric framework for dimensionality reduction," Tech. Rep., University of Michigan, 2008, arXiv:0809.4866.

[18] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics. John Wiley and Sons, NY, USA, 1997.

[19] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191, American Mathematical Society and Oxford University Press, 2000, Translations of mathematical monographs.

[20] R. N. Damle, T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S. L Allen, and et. al., "Ig v gene mutation status and cd38 expression as novel prognostic indicators in chronic lymphocytic leukemia," *Blood*, vol. 95, no. 7, pp. 1840–1847, 1999.

[21] L. K. Habib and W. G. Finn, "Unsupervised immunophenotypic profiling of chronic lymphocytic leukemia," *Cytometry Part B: Clinical Cytometry*, vol. 70B, no. 3, pp. 124–135, 2006.

[22] R. W. McKenna, L. T. Washington, D. B. Aquino, L. J. Picker, and S. H. Kroft, "Immunophenotypic analysis of hematogones (b-lymphocyte precursors) in 662 consecutive bone marrow specimens by 4-color flow cytometry," *Blood*, vol. 98, no. 8, pp. 2498–2507, 2001.

[23] P. F. Hsieh, D. S. Wang, and C. W. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *IEEE Transactions on Pattern Analalysis and Machine Intelligence*, vol. 28, no. 2, pp. 223–235, Feb. 2006.

[24] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, May 2000.

[25] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, April 1999.

[26] S. C. Douglas, "On the design of gradient algorithms employing orthogonal matrix constraints," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2007, pp. 1401–1404.