DIMENSIONALITY REDUCTION OF FLOW CYTOMETRIC DATA THROUGH INFORMATION PRESERVATION

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

Like many biomedical applications, flow cytometry is a field in which dimensionality reduction is important for analysis and diagnosis. Through expression patterns of various fluorescent biomarkers, flow cytometry is often used to characterize the malignant cells in cancer patients, traced to the level of the individual cell. Typically, diagnosticians analyze cytometric data through a series of 2-dimensional histograms of the expression of various marker combinations, which does not exploit the high-dimensional nature of the data. In this paper we utilize a form of dimensionality reduction – which we refer to as Information Preserving Component Analysis (IPCA) – that preserves the information distance between multi-dimensional data sets. As such, we offer a method for clinicians to visualize patient data in a low-dimensional projection space defined by a linear combination of all available markers. We illustrate these results on actual patient data.

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

1. INTRODUCTION

Clinical flow cytometry typically involves data retrieved from cancerous blood samples which have been treated with different fluorescent markers. This offers a high-dimensional data set which contains simultaneous analysis of several measurements, such as marker expression and light scatter angle. Typically, diagnosis of flow cytometry data is performed by analyzing a series of 2-dimensional projections onto the axes of the data, which correspond to different biomarker combinations determined through years of clinical experience. These projections methods do not fully exploit the high-dimensional nature of the data, and are typically used due to the critical importance of visualization in diagnosis.

Given the desired for visualization, dimensionality reduction is important in flow cytometry analysis. The key principle of the low-dimensional space is that it must preserve the relationship between data sets such that patients with the same disease should exhibit similar expression patterns in the projected space. 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, which is goal of common dimension reduction methods such as [1, 2].

In this paper we present the utilization of a method of dimensionality reduction – which we refer to as *Information Preserving Component Analysis* (IPCA) [3] – that preserves the Fisher information distance between data sets. IPCA operates in the space of linear and unsupervised dimensionality reduction methods, such as Principal Components Analysis (PCA) and Independent Component Analysis (ICA) [4]. IPCA ensures that the low-dimensional representation maintains the similarities (i.e. information distance) 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 is largely a synopsis of the submitted full journal paper [3], which is publicly available on ArXiv, but the analysis of sample size effects on the IPCA projection (Sec. 4.1) is new.

This paper proceeds as follows: In Section 2 we give a background of flow cytometry as well as a formulation of the problem we will attempt to solve. We present our methods for finding the IPCA projection in Section 3. Simulation results for actual clinical data are illustrated in Section 4, followed by a discussion and areas for future work in Section 5.

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

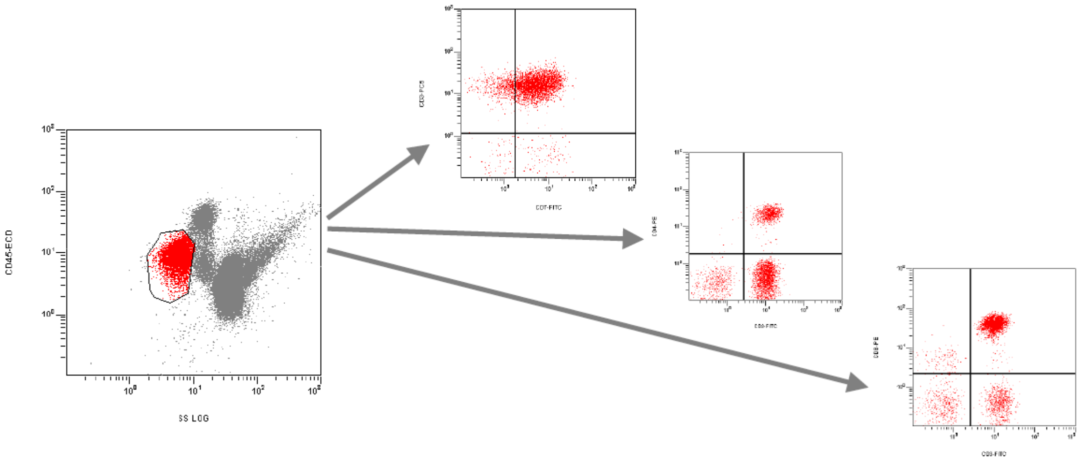


Fig. 1. Historically, clinical flow cytometric analysis relies on a series of 2-dimensional scatter plots in which cell populations are selected for further evaluation. This process does not take advantage of the multi-dimensional nature of the problem.

2. 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). At the basic level, flow cytometry is the process of analyzing a blood sample with a collection of different fluorescent markers, selected due to known expression patterns with certain disease types. In routine flow cytometric immunophenotyping, the expression patterns of each marker in a given sample can be traced to the level of the single cell.

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. For visualization, diagnosticians typically analyze a series of 2-dimensional histograms defined by any 2 of the 6 characteristics measured in a given tube (see Fig. 1). 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 set.

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.

2.1. 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, 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.

Specifically, given a collection of flow cytometer outputs $\mathcal{X} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_N\}$ 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

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

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?

3. METHODS

Flow cytometry data is often analyzed through the statistics of the patient data set. Essentially, each patient can be viewed as a realization of some overriding probability density function (PDF) lying on a statistical manifold [5]. The Kullback-Leibler (KL) divergence, defined as

$$KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx, \tag{1}$$

is a common, non-parametric, means of determining a similarity between PDFs $p(x)$ and $q(x)$. The KL-divergence is a very important metric in information theory, and is commonly referred to as the relative entropy of one PDF to another. This divergence stresses the differences in PDFs, yielding very large values when $p(x)$ and $q(x)$ are highly dissimilar, specifically at the tails of the distributions.

It should be noted that the KL-divergence is not a distance metric, as it does not satisfy the symmetry properties of a distance metric, $KL(p\|q) \neq KL(p\|q)$. To obtain this symmetry, we utilize the symmetric KL-divergence:

$$D_{KL}(p,q) = KL(p\|q) + KL(q\|p). \tag{2}$$

Using the KL-divergence, we are able to define a similarity measure between patient data sets. In order to calculate the KL-divergence, the generative PDF for each patient needs to be approximated, and there are a multitude of methods for this task, such as kernel density estimation and mixture models.

3.1. Objective Function

We define the *Information Preserving Component Analysis* (IPCA) projection as one that preserves the Fisher information distance (or some approximation thereof, such as the KL-divergence [6]) between data sets. As such, the divergence between data PDFs should be minimally altered when projecting from the full data dimension to the low-dimensional space. Specifically, let $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ where $\boldsymbol{X}_i \in \mathbb{R}^{d \times n_i}$ is the i^{th} data set, containing n_i elements of dimension d. 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:

$$A = \arg \min_{A:AA^T=I} \|D(\mathcal{X}) - D(\mathcal{X}, A)\|_F^2, \tag{3}$$

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 the projection matrix A, i.e. $D_{ij}(\mathcal{X}, A) = D_{KL}(A\boldsymbol{X}_i, A\boldsymbol{X}_j)$.

3.2. 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 a global minimum, so typically it is important to initialize x_0 near the estimated minimum.

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 $\mu$
Calculate $D(\mathcal{X})$, the Kullback-Leibler dissimilarity matrix
Initialize $A_1 \in \mathbb{R}^{m \times d}$ as a random orthonormal projection matrix
Calculate $D(\mathcal{X}, A_i)$, the Kullback-Leibler dissimilarity matrix in the projected space
for $i = 1$ to $\infty$ do
    Calculate $\frac{\partial}{\partial A_i}\tilde{J}$, the direction of the gradient, constrained to $AA^T = I$
    $A_{i+1} = A_i - \mu \frac{\partial}{\partial A_i}\tilde{J}$
    Calculate $D(\mathcal{X}, A_{i+1})$
    $J = \|D(\mathcal{X}) - D(\mathcal{X}, A_{i+1})\|_F^2$
    Repeat until convergence of $J$
end for
Output: Projection matrix $A \in \mathbb{R}^{m \times d}$, which preserves the information distances between sets in $\mathcal{X}$.
```

Using gradient descent, we are able to solve (3). Specifically, let $J = \|D(\mathcal{X}) - D(\mathcal{X}, A)\|_F^2$ 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 with respect to a projection matrix A,

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

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

$$A = A - \mu \frac{\partial}{\partial A}\tilde{J}(A), \tag{4}$$

where

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

is the direction of the gradient, constrained to force A to remain orthonormal (we omit the derivation of this constraint, which can be found in [3]). This process is iterated until the error J converges.

3.3. Algorithm

The full method for IPCA, specialized towards the current problem, is described in Algorithm 1. We note that A is ini-

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

Table 1. Data dimensions and corresponding markers for analysis of CLL and MCL.

tialized 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 have shown the flow cytometry problem is sufficiently convex, at least for our available data, yielding significantly similar convergence values.

4. SIMULATION

We now present simulation results for using IPCA to find a projection matrix for flow cytometric data analysis. By using IPCA and projecting the data down to 2 dimensions, pathologists can obtain this visualization through a linear combination of all available markers, weighted by importance in preserving information, rather than just 2 at a time. We offer this as a proof of concept for diagnosis and exploratory research. Patient data was obtained and diagnosed by the Department of Pathology at the University of Michigan.

4.1. Lymphoid Leukemia Study

In this study, we compare patients with two 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.

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 1), 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

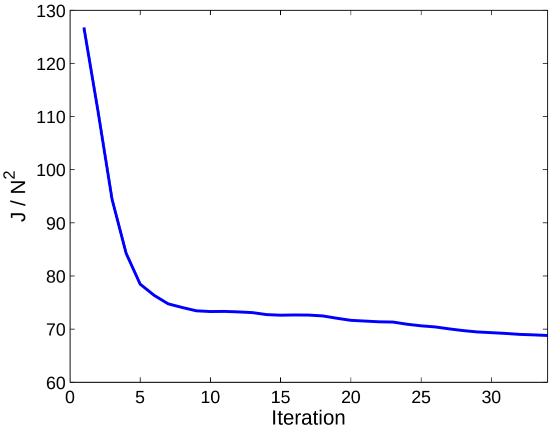


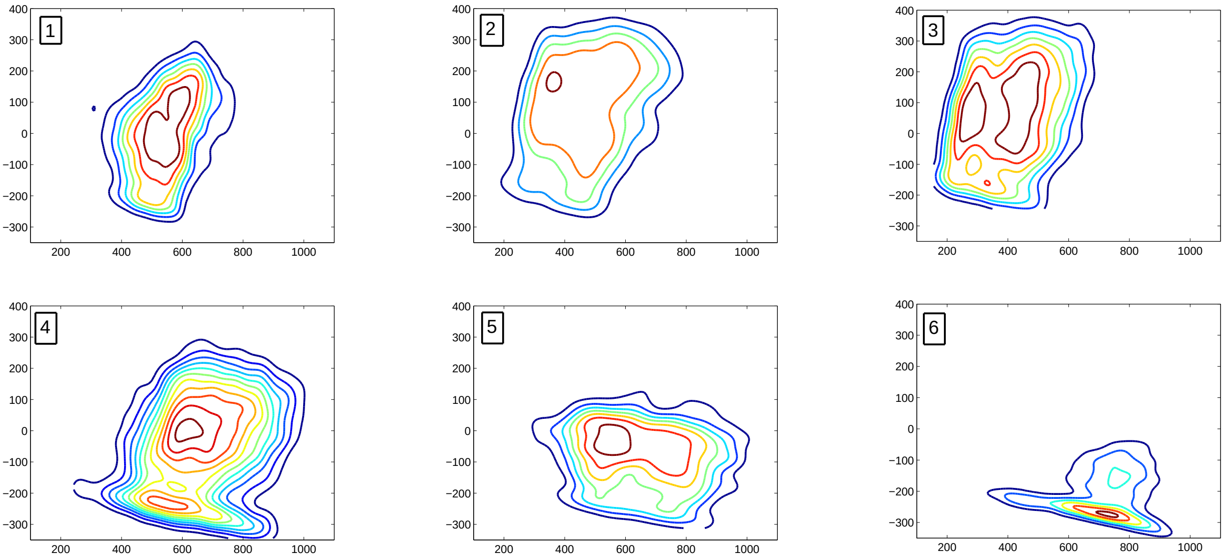
Fig. 2. Evaluating the objective as a function of time. As the iterations increase, the objective function eventually converges.

$$A = \begin{pmatrix} 0.064 & 0.036 & 0.906 & 0.208 & 0.355 & -0.084 \\ -0.019 & -0.197 & -0.145 & -0.956 & 0.165 & -0.011 \end{pmatrix}. \tag{5}$$

This projection was calculated by minimizing the objective function with respect to A, as illustrated in Fig. 2 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 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. 3. 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 PDFs, there is still a significant enough difference to distinguish them, especially given the similarities to other patient PDFs.

We now illustrate the embedding obtained with FINE [7] of the projected data (see Appendix A). The embedding results are shown in Fig. 4(b), in which the separation between

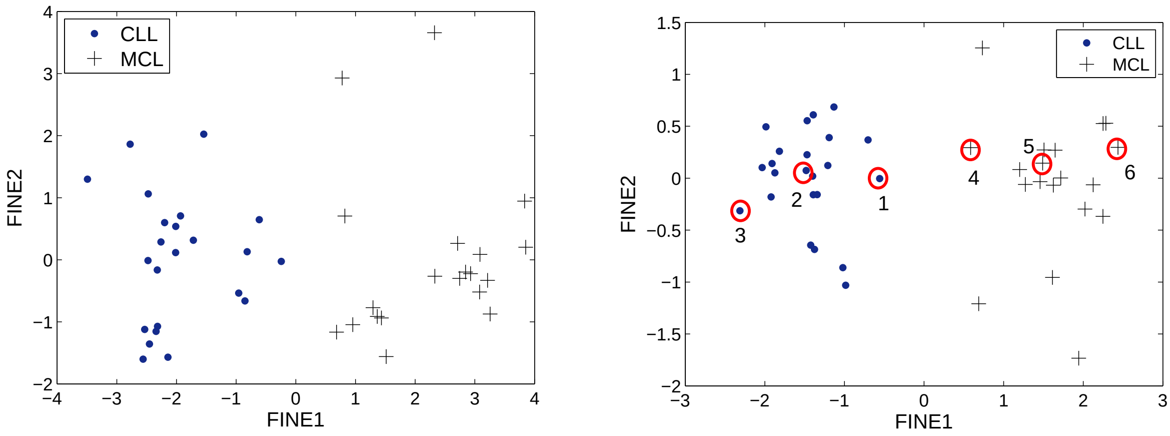


(a) Most Similar

(b) Centroids

(c) Least Similar

Fig. 3. 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 illustrated in Fig. 4.



(a) Full Dimension

(b) IPCA Projection

Fig. 4. 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. 3, numbered respectively.

classes is preserved when using the projected data as compared to using the full-dimensional data in Fig. 4(a). In both embeddings, each point represents the estimated PDF of an entire patient data set, and those which are circled correspond to the PDFs shown in Fig. 3. IPCA maintains the relationships between different sets, allowing for a consistent analysis.

Using the projection matrix (5) for variable selection, the loading vectors are highly concentrated towards the 3$^{\text{rd}}$ and 4$^{\text{th}}$ dimensions, which correspond to fluorescent markers FMC7 and CD23. This marker combination, which is well known in the clinical pathology community for differentiating CLL and MCL, was able to be independently validated 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.

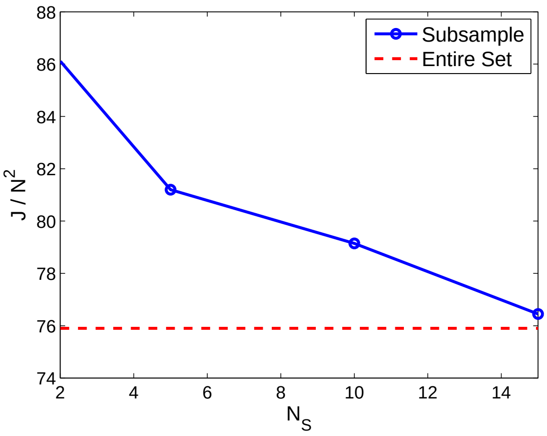


Fig. 5. IPCA performance using subset of collection the $\mathcal{X}_S \subset \mathcal{X}$, where N_S is the number of randomly selected patients from each disease class. Results shown over a 5-fold cross validation, with the IPCA projection determined by $\mathcal{X}$ shown as a lower bound with the dotted line.

One concern when implementing IPCA is the number of data sets necessary to find a proper projection. Specifically, given a subset of $\mathcal{X}_S \subset \mathcal{X}$, how close does IPCA approach the value of the objective function obtained when utilizing the entire collection $\mathcal{X}$? To determine this, we 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. We then calculate the value of the objective function on the entire set $\mathcal{X}$. The mean results over a 5-fold cross validation are illustrated in Fig. 5, where we signify the value of the objection function when using IPCA on the entire data set with the dotted line. Note that this value is not identical to that in Fig. 2 as the initial starting matrix was randomly generated (although held constant for all trials in this experiment). Given that the value of the objection function with the initial random projection matrix was $\frac{J}{N^2} = 180.5941$, the relative performance of IPCA with few available data sets is promising.

5. CONCLUSIONS

In this paper we have shown the ability to find an information-based projection for high-dimensional data analysis using Information Preserving Component Analysis (IPCA). By preserving the information distance between data sets (through the use of the KL-divergence), we find a low-dimensional projection which allows for visualization in cancer diagnosis. Analysis of the loading vectors of the IPCA projection matrix may be used as a form of variable selection, which enables the use of IPCA for exploratory research. In future work we plan to continue applying IPCA towards flow cytometric analysis, including looking for subgroups within disease classes. Additionally, we plan to look for other applications which would benefit from this form of dimensionality reduction.

Algorithm 2 Fisher Information Non-parametric Embedding

```
Input: Collection of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$; the desired embedding dimension $d$
1: for $i = 1$ to $N$ do
2:     Calculate $\hat{p}_i(\boldsymbol{x})$, the density estimate of $\boldsymbol{X}_i$
3: end for
4: Calculate $D(\mathcal{X})$
5: $\boldsymbol{Y} = \text{cMDS}(D(\mathcal{X}), d)$
Output: $d$-dimensional embedding of $\mathcal{X}$, into Euclidean space $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$
```

A. FINE ALGORITHM

We have previously [7] presented an algorithm for determining a low-dimensional Euclidean embedding of high dimensional data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$. Coined *Fisher Information Non-parametric Embedding* (FINE), we determine a mapping:

$$\psi : \boldsymbol{X}_i \rightarrow y_i, \; y_i \in \mathbb{R}^d$$

where y_i is the location of the PDF of $\boldsymbol{X}_i$ on the reconstructed manifold in Euclidean space. Details can be found in Algorithm 2, where line 5 refers to using classical multidimensional scaling to embed the dissimilarity matrix $D(\mathcal{X})$ into a Euclidean space with dimension d.

B. REFERENCES

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

[2] 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.

[3] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, “Information preserving component analysis: Data projections for flow cytometry analysis,” *IEEE Journal of Selected Topics in Signal Processing*, 2008, submitted.

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

[5] W. G. Finn, K. M. Carter, R. Raich, and A. O. Hero, “Analysis of flow cytometric immunophenotyping data by clustering on statistical manifolds,” *Cytometry Part B: Clinical Cytometry*, 2008, submitted.

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

[7] 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.