Information-Geometric Dimensionality Reduction

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

We consider the problem of dimensionality reduction and manifold learning when the domain of interest is a set of probability distributions instead of a set of Euclidean data vectors. In this problem, one seeks to discover a low dimensional representation, called an embedding, that preserves certain properties such as distance between measured distributions or separation between classes of distributions. Such representations are useful for data visualization and clustering. While a standard Euclidean dimension reduction method like PCA, ISOMAP, or Laplacian Eigenmaps can easily be applied to distributional data – e.g. by quantization and vectorization of the distributions – this may not provide the best low-dimensional embedding. This is because the most natural measure of dissimilarity between probability distributions is the information divergence and not the standard Euclidean distance. If the information divergence is adopted then the space of probability distributions becomes a non-Euclidean space called an information geometry. This article presents methods that are specifically designed for the low-dimensional embedding of information-geometric data, and we illustrate these methods for visualization in flow cytometry and demography analysis.

Index Terms

Information geometry, dimensionality reduction, statistical manifold, classification

I. Introduction

High dimensional data visualization and interpretation have become increasingly important for data mining, information retrieval, and information discrimination applications arising in areas such as search engines, security, and biomedicine. The explosion in sensing and storage capabilities has generated a vast amount of high dimensional data and led to the development of many algorithms for feature extraction and visualization, known variously as dimensionality reduction, manifold learning, and factor analysis.

Dimensionality reduction strategies fall in two categories: supervised task-driven approaches and unsupervised geometry-driven approaches. Supervised task-driven approaches reduce data dimension according to optimize a performance criterion that depends on both the reduced data and ground truth, e.g., class labels. Examples include linear discriminant analysis (LDA) [1], supervised principal components [2], and multi-instance dimensionality reduction [3]. Unsupervised geometry-driven approaches perform dimension reduction without ground truth and try to preserve geometric properties such as distances or angles between data points. Examples include principal components analysis (PCA) and multidimensional scaling (MDS) [4], and ISOMAP [5]. Most of these approaches use Euclidean distances between sample points to drive the dimensionality reduction algorithm.

Recently it has been recognized that these Euclidean algorithms can be generalized to non-Euclidean spaces by replacing the Euclidean distance metric with a more general dissimilarity measure. In particular, when the data samples are probability distributions, use of an information divergence such as Kullback-Leibler (KL) instead of Euclidean distance leads to a class of information geometric algorithms for dimensionality reduction [6], [7]. In this article we motivate and explain the application of information-geometric dimensionality reduction for two real-world applications.

Information-geometric dimensionality reduction (IGDR) operates on a statistical manifold of probability distributions instead of the geometric manifold of Euclidean data points. When such distributional information can be extracted from the data, IGDR results in significant improvements in information retrieval, visualization, and classification performance [6]–[10]. This improvement can be understood from the point of view of information-theoretic bounds: information divergence is generally more relevant to statistical discrimination performance than Euclidean distance.

For example, for binary classification the minimum probability of error converges to zero at an exponential rate with rate constant equal to the the Kullback-Leibler information divergence between the distributions of the data over each class [11]. The KL divergence not a function of the Euclidean distances between data points unless these distributions are spherical Gaussian. Therefore, as it preserves information divergence, in many cases IGDR can produce more informative dimension reductions than classical Euclidean approaches.

Implementation of information-geometric versions of PCA, ISOMAP and others is often not as straightforward as the Euclidean counterparts, which are frequently convex and solvable as generalized eigenvalue problems. Nonetheless, as shown in this paper, the added complexity of implementation can be well

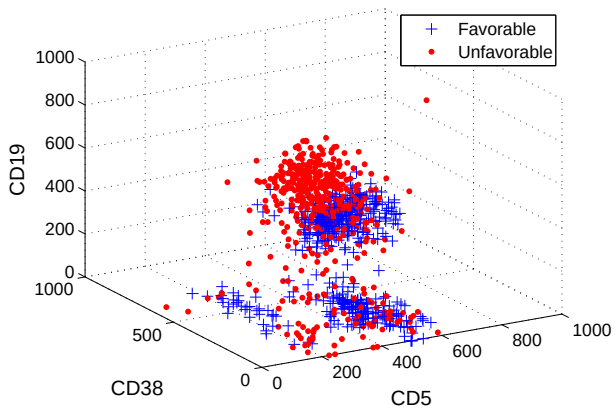


Fig. 1. In clinical flow cytometry, diagnoses and prognoses are made through the analysis of high-dimensional point clouds the measurement space of selected bio-markers.

worth the effort. We illustrate the power of information geometric dimensionality reduction by presenting generalizations of ISOMAP, PCA and LDA. These implementations are called Fisher Information Nonparametric Embedding (FINE) [6], Information Preserving Components Analysis (IPCA) [9], and Information Maximizing Components Analysis (IMCA) [12], respectively. Each of these algorithms solves a well-posed optimization problem over the information-geometric embedding of each sample point's distribution.

Probability distributions and information divergence can arise as useful targets for dimensionality reduction in several ways. In image retrieval applications, the most discriminating properties of an image may be invariants such as the relative frequency of occurrence of different textures, colors, or edge features. The histogram of these relative frequencies is a probability distribution that is specific to the particular image; up to scale, translation, rotation or other unimportant spatial transformations. Dimensionality reduction on these probabilities can accelerate retrieval speed without negatively affecting precision or recall rates. Furthermore, visualization of the database, e.g. as manifested by clusters of similar images, can be useful for understanding database complexity or for comparing different databases.

In other applications, each object in the database is itself stored as a cloud of high dimensional points and the shape of this point cloud is what naturally differentiates the objects. For example, in the flow cytometry application, discussed in Section V of this paper, the objects are different patients, the data points are vector attributes of a population of the patient's blood cells, and it is the shape of the point cloud that is of interest to the pathologist. This is demonstrated in Fig. 1, where we compare the point clouds, with respect to 3 bio-markers, of two patients with favorable and unfavorable prognoses. Another example, discussed in Section VI, is spatio-demographic analysis of crime data where the analyst is interested in

comparing patterns of crime in different cities based on distributions of community and law enforcement characteristics.

All the algorithms presented here are available for download as MATLAB code on our reproducible research website [13].

II. Distance on Statistical Manifolds

Information geometry is a field that has emerged from the study of geometrical constructs on manifolds of probability distributions. These investigations analyze probability distributions as geometrical structures in a Riemannian space. Using tools and methods deriving from differential geometry, information geometry is applicable to information theory, probability theory, and statistics[1].

As most dimensionality reduction techniques are designed to either preserve pairwise sample distances (unsupervised) or maximize between-class distances (supervised), it is first necessary to understand the principles of distance in information geometry. Similar to points on a Riemannian manifold in Euclidean space, PDFs which share a parametrization lie on a *statistical* manifold. A statistical manifold may be viewed as a set $\mathcal{M}$ whose elements are probability distributions. The coordinate system of this manifold is equivalent to the parametrization of the PDFs. For example, a d-variate Gaussian distribution is entirely defined by its mean vector μ and covariance matrix Σ, leading to a $d + d(d+1)/2$-dimensional statistical manifold which is of a higher dimension than the dimension d of a sample realization $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$ from this distribution.

For a parametric family of probability distributions on a statistical manifold, it is possible to define a Riemannian metric using the Fisher information metric, which measures the amount of information a random variable contains in reference to an unknown parameter θ. This metric may then be used to compute the Fisher information distance $D_F(p_1, p_2)$ between two distributions $p(x;\theta_1), p(x;\theta_2) \in \mathcal{M}$. This distance is the length of the shortest path – the geodesic – on $\mathcal{M}$ connecting coordinates θ_1 and θ_2.

While the Fisher information distance cannot be exactly computed without a priori knowledge about the parametrization θ of the manifold, the distance between PDFs p_1 and p_2 may be approximated with a variety of pseudo-metrics such as the Kullback-Leibler (KL) divergence,

$$KL(p_1 \,\|\, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)}\, dx. \tag{1}$$

[1]For a more thorough introduction to information geometry, we suggest [14], [15]

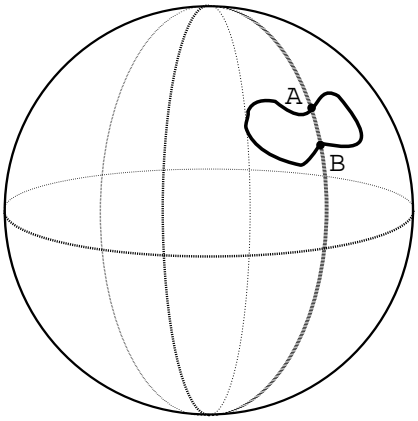


Fig. 2. Given a 1-dimensional submanifold (the curvy dark line) of interest lying on a 2-dimensional sphere manifold, the Fisher information distance is the shortest path connecting the points A and B along the 1-D submanifold, rather than the length of a portion of the great circle connecting the points on the sphere.

The KL-divergence is very important in information theory, and is commonly referred to as the relative entropy of one PDF to another. As a pair of densities approach each other, the Kullback-Leibler divergence is a good approximation to the Fisher information distance between them[2] [14]:

$$\sqrt{2KL(p_1\|p_2)} \to D_F(p_1, p_2)$$

as $p_1 \to p_2$. This allows for a data-driven approximation of the Fisher information distance, through the use of the empirically determined PDFs in the absence of information about the Fisher information metric. While the KL-divergence is not a symmetric measure, we can add symmetry by defining, $D_{KL}(p_1, p_2) = KL(p_1 \,\|\, p_2) + KL(p_2 \,\|\, p_1)$, which maintains similar convergence properties. We note that there are several other metrics which approximate the Fisher information distance – such as the Hellinger and cosine distances – although for brevity we utilize the KL-divergence throughout this paper. For additional measures of probabilistic distance and details on their computation for empirical data, we refer the reader to [16], [17].

As the two densities p_1 and p_2 in (1) become more dissimilar, the KL-divergence approximation of the Fisher information distance becomes weak. Additionally, when PDFs are constrained to form a submanifold of interest, the "straight shot" distance is no longer an accurate description of the manifold distance. This is illustrated in Fig. 2 in which we represent a 1-dimensional submanifold which occupies a subspace of the 2-dimensional hyper-sphere. The Fisher information distance is equal to the shortest path along the submanifold (curvy line), and is not equal to the distance on the full manifold, i.e. the portion of a

[2]More precisely, $2KL(p_1\|p_2) = D_F^2(p_1, p_2)(1 + O(\|p_1 - p_2\|)$ where $\|p_1 - p_2\|$ denotes the L_2 norm of the difference between the densities.

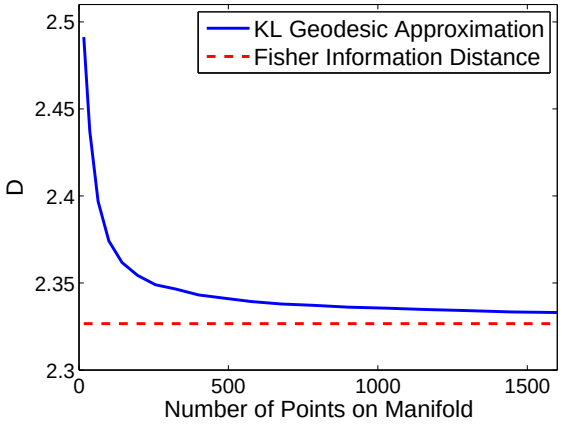


Fig. 3. Convergence of the graph approximation of the Fisher information distance using the Kullback-Leibler divergence. As the manifold is more densely sampled, the KL divergence approaches the Fisher information distance.

great circle on a hyper-sphere connecting the two points. Hence, there are situations in which standard approximations of the information distance do not converge to the true distance, and it is necessary to approximate the geodesic along the manifold.

Using a connected graph, we may define the path between p_1 and p_2 as a series of connected segments. The geodesic distance may then be approximated as the sum of the lengths of those segments. Specifically, given the collection of N PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ and using the KL-divergence as approximation of the Fisher information distance, we can now define an approximation function G for all pairs of PDFs:

$$G(p_1, p_2; \mathcal{P}) = \min_{M, \mathcal{P}} \sum_{i=1}^{M-1} D_{KL}(p_{(i)}, p_{(i+1)}), \quad p_{(i)} \to p_{(i+1)} \, \forall \, i. \tag{2}$$

Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well sampled manifold, and as such $G(p_1, p_2; \mathcal{P}) \to D_F(p_1, p_2)$ as $N \to \infty$. Empirically, (2) may be solved with Dijkstra's shortest path algorithm. This is similar to the manner in which ISOMAP [5] approximates distances on Euclidean manifolds. Figure 3 illustrates this approximation by comparing the KL graph approximation to the actual Fisher information distance for the univariate Gaussian case. As the manifold is more densely sampled (uniformly sampling over the range of mean and variance parameters for this simulation), the approximation converges to the true Fisher information distance.

III. Dimensionality Reduction in the Density Space

Consider the collection of PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ lying on some statistical manifold $\mathcal{M}$. By performing dimensionality reduction in the space of probability densities, one wishes to reconstruct $\mathcal{M}$ using only the information available in $\mathcal{P}$. Specifically, the aim is to find an embedding $A : p(x) \to y$, where $y \in \mathbb{R}^m$.

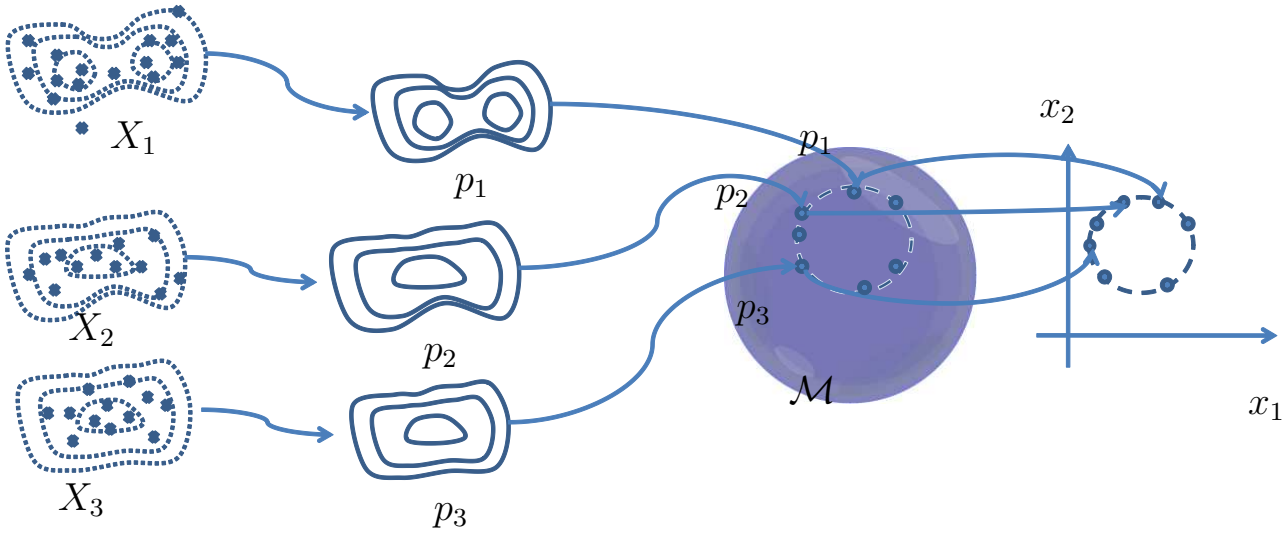


Fig. 4. FINE: first, a probability density function (PDF) p_i is estimated for each dataset X_i. Then, an information-geometric metric is used to learn the geometry of the manifold of PDFs from pairwise distance measurements. Finally, a Euclidean embedding from the manifold $\mathcal{M}_x$ to $\mathbb{R}^d$ is obtained associating each original dataset X_i with its embedded point in Euclidean space x_i.

This is a similar setting to traditional manifold learning algorithms which aim to reconstruct Riemannian manifolds based on a finite sampling, extended to the properties of statistical manifolds.

By performing dimensionality reduction on a family of PDFs, we are better able to both visualize and classify the data. In order to obtain a lower dimensional embedding, we calculate the pairwise KL-divergences within $\mathcal{P}$. In problems of practical interest, however, the parameterization of the probability densities is usually unknown. We instead are given a family of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, in which we may assume that each data set $\boldsymbol{X}_i$ is a realization of some underlying probability distribution to which we do not have knowledge of the parameters. As such, we rely on nonparametric techniques to estimate both the probability density and the KL-divergence. For the purposes of this paper, we implement kernel density estimation methods, although other estimation methods are also applicable.

In previous work [6] we developed an algorithm for dimensionality reduction in the density space, which we called Fisher Information Nonparametric Embedding (FINE). By assuming each data set is a realization of an underlying PDF, and each of those distributions lie on a manifold with some natural parametrization, then this embedding can be viewed as an embedding of the actual manifold into Euclidean space. We illustrate the FINE algorithm in Fig. 4.

Through information geometry, FINE enables the joint embedding of multiple data sets $\boldsymbol{X}_i$ into a single low-dimensional Euclidean space. By viewing each $\boldsymbol{X}_i \in \mathcal{X}$ as a realization of $p_i \in \mathcal{P}$, we reduce the numerous samples in $\boldsymbol{X}_i$ to a single point. The dimensionality of the statistical manifold may be significantly less than that of the Euclidean realizations. MDS methods reduce the dimensionality of p_i

from the Euclidean dimension to the dimension of the statistical manifold on which it lies.

A. Adding Application-specific Constraints

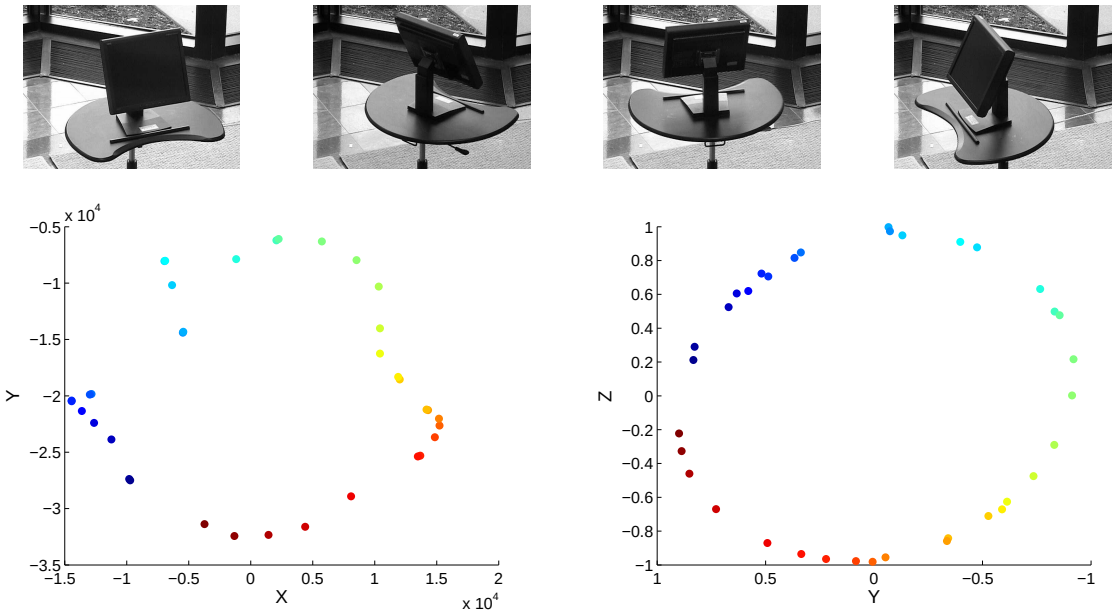
FINE was developed to be applied to the general case of dimensionality reduction in the space of PDFs, making no assumptions on the data distributions or the geometry of the underlying statistical manifolds. However, there are several applications where known intrinsic properties which may be exploited when performing information geometric dimensionality reduction. By incorporating these properties into algorithm constraints, one may be able to obtain improved performance.

Lee *et al*. [18] have demonstrated the use of IGDR for image segmentation, using multinomial distributions as points which lie on an n-simplex (or projected onto an $(n+1)$-dimensional sphere). By framing their problem as such, they are able to exploit the properties of such a manifold – using the cosine distance as an exact computation of the Fisher information distance, and using linear methods (PCA) of dimensionality reduction. They have shown very promising results for the problem of image segmentation.

If there exists *a priori* knowledge that the geometry of the underlying manifold is that of a (hyper)sphere, adding such a constraint results in an improved embedding. In [8], we presented a special case of FINE which we called Spherical Laplacian Information Maps (SLIM), which restricted the final embedding to constrain all points to lie on the surface of a sphere. SLIM is useful when the user wants to preserve the spherical geometry of the ambient space, as arises for example when dimensionality reduction is used to extract object pose trajectories from video. This is illustrated in Fig. 5, where we embed the rotation of an object captured by a stationary camera with SLIM and PCA. Each of 36 images was featured as a multinomial distribution over the pixel space prior to embedding. While PCA discerns the order of the change in angle, it does not properly identify the shape of the trajectory (i.e. circular) as SLIM does.

IV. Dimensionality Reduction in the Sample Space

For many learning methods, it is often desirable to reduce the dimensionality of $\boldsymbol{X}$, finding a transformation $A : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ where $\boldsymbol{Y} = [y_1, \ldots, y_n]$ and each $y_i \in \mathbb{R}^m$, $m < d$. Typically, each set would be reduced in an individual manner; if there is deemed a relationship between the sets, it has generally been approached as a classification problem in which each signal $\boldsymbol{X}_i$ is considered a set of points belonging to class i. An example of this situation would be supervised dimensionality reduction with Fisher's linear discriminant analysis (LDA) [1].



(a) PCA (b) SLIM

Fig. 5. The embedding of an object captured at various rotation points with SLIM and PCA. SLIM preserves the spherical nature of the intrinsic manifold.

Viewing this problem from an information-geometric perspective presents a different vantage point; rather than considering each $\boldsymbol{X}_i$ to be a collection of points in a specific class, let us generalize the relationship between sets $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. Specifically, consider the case for which each $\boldsymbol{X}_i$ is a realization of some unknown generating function p_i, in which p_i and p_j may or may not be equivalent. This agrees with the standard classification problem, in which each p_i represents a class PDF, but it also allows for different relationships between PDFs. Specifically, rather than having a number classes equal to the number of data sets N, there may be significantly fewer classes $M \ll N$, in which M is unknown and no labels are available. In this generalized scenario, dimensionality reduction dimensionality reduction is desirable for the purposes of classification, feature extraction, and/or visualization.

Let us illustrate with a simple example. Every 10 years, a US census is performed generating a collection of data about each of its residents such as height, weight, income, ethnicity, education level, etc.. Let us now partition the data such that each county within the same state is represented by its own set $\boldsymbol{X}$. Standard methods of feature extraction will find the features which best describe each county on an individual level. We are interested in determining the most important features when comparing all counties at the same time. While median income may not be a distinguishing characteristic within a single county, and may not be recognized as such when solely extracting features from that individual county, it would be quite informative when comparing all counties across the state.

The construct of comparison across data sets can be directly abstracted to the biomedical fields, where it is necessary to compare patients who have been analyzed with the same set of features, and identify which of those features best distinguishes the patient corpus. We have presented a method of information geometric dimensionality reduction – which we refer to as *Information Preserving Component Analysis (IPCA)* – to solve this problem for flow cytometry data [9]. IPCA aims to find the optimal transformation of PDFs $A : p(x) \to p(y)$. By preserving the KL-divergence the estimated PDFs generating the data 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.

With some abuse of notation, we will further refer to $D_{KL}(p_i, p_j)$ as $D_{KL}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, recalling that the KL-divergence is calculated with respect to PDFs, not realizations. We define the IPCA projection matrix $A \in \mathbb{R}^{m \times d}$, in which A reduces the dimension of $\boldsymbol{X}$ from d to m ($m \leq d$), such that

$$D_{KL}(A\boldsymbol{X}_i, A\boldsymbol{X}_j) = D_{KL}(\boldsymbol{X}_i, \boldsymbol{X}_j), \; \forall \, i, j. \tag{3}$$

This can be formulated as an optimization problem:

$$A = \arg \min_{A : AA^T = I} J(A), \tag{4}$$

where I is the identity matrix and $J(A)$ is some cost function designed to implement (3). Note that we include the optimization constraint $AA^T = I$ to ensure our projection is orthonormal, which keeps the data from scaling or skewing as that would undesirably distort the data. Let $D(\mathcal{X})$ be 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. $D_{ij}(\mathcal{X}; A) = D_F(A\boldsymbol{X}_i, A\boldsymbol{X}_j)$. This formulation results in the following cost function:

$$J(A) = \sum_i \sum_j W_{ij} \left(D_{ij}(\mathcal{X}) - D_{ij}(\mathcal{X}; A) \right)^2, \tag{5}$$

where W_{ij} is some weighting factor.

The weights W_{ij} can be selected based on $D_{ij}(\mathcal{X})$ to de-emphasize the influence of certain pairs (i, j) on the embedding. For example, nearest neighbor (NN) weights of $W_{ij} = 1$ for some k-NN and $W_{ij} = 0$ will eliminate far-flung interactions for which the KL-divergence is a poor approximation to the Fisher metric. The use of heat kernel weights, similar to Laplacian Eigenmaps [19], will have a more gradual effect. These functions will ensure that more weight is given to preserve the pairwise distances of "close" PDFs. While the choice of cost weighting function is dependent on the problem, the overall projection

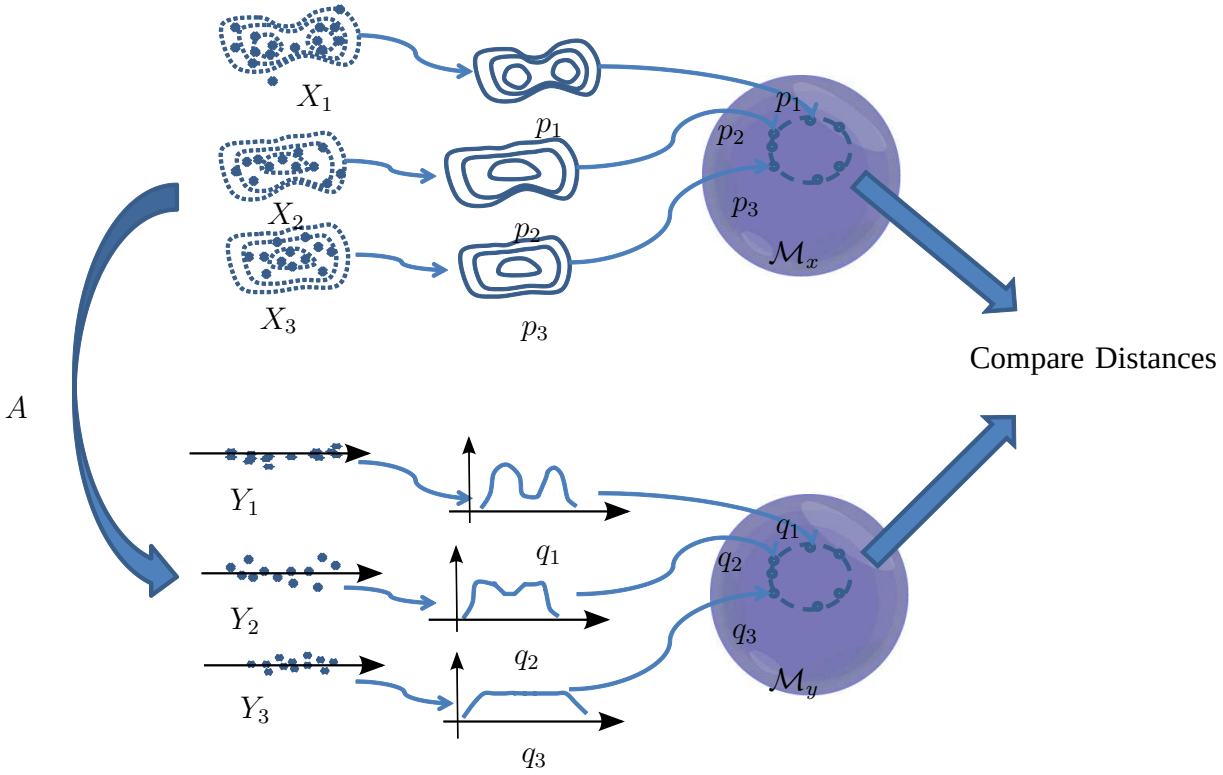


Fig. 6. IPCA/IMCA: first, a probability density function (PDF) p_i is estimated for each dataset X_i. Simultaneously, a probability density function (PDF) q_i is estimated for each dataset $Y_i = AX_i$. Then, an information-geometric metric is used to learn the geometry of the manifold $\mathcal{M}_x$ of PDFs p_is and manifold $\mathcal{M}_y$ from PDFs q_is from pairwise distance measurements. Finally, an objective is calculated to compare the geometry of the two manifolds $\mathcal{M}_x$ and $\mathcal{M}_y$. For IPCA, we consider the minimization of the sum of squared differences between each pairwise distance on $\mathcal{M}_x$ and its equivalent in $\mathcal{M}_y$. For IMCA, we consider the maximization of the sum of distances in $\mathcal{M}_y$.

method ensures that the similarity between data sets is maximally preserved in the desired low-dimensional space, allowing for comparative learning between sets.

We illustrate the IPCA and IMCA (see Section IV-A) in Fig. 6. While we omit the details in this paper (see [9], [17]), the cost function (5) may be minimized with various convex optimization techniques; we utilize gradient descent with random initializations for A. There are computational issues with gradient methods, namely local extrema. We find the global minimum by computing IPCA over several random initializations and taking the resultant A which minimizes the cost function. In most applications we have tested, this method has been very effective, and we have found most random initializations of A converge to the same minimum.

Recall that the information distance is entirely defined by those areas of input space in which PDFs differ. As the IPCA preserves the information distance between probability distributions, A is going to be highly weighted towards the variables which contribute most to that distance. Hence, the loading vectors of A give a ranking of the discriminative value of each variable in the full-dimensional feature space. This form of variable selection is useful in exploratory data analysis.

A. Supervised Learning

As mentioned previously, when developing ICPA we generalized the relationship between PDFs such that they may or may not represent unique classes in a classification task. We presented IPCA in the scenario for which sample classification is not the desired task, but we now extend the methods to supervised dimensionality reduction.

The Chernoff performance bound on classification error is used to bound the probability of error based on the probabilistic distance between classes. The Chernoff distance is a single-parameter class of probabilistic distances, and as the distance increases, the probability of misclassification decreases. A special member of the class of Chernoff distances, know as the Bhattacharya distance between PDFs, converges to the Fisher information distance, similarly to the KL-divergence. It is natural, therefore, to find a form of dimensionality reduction which will maximize the information distance between class PDFs, as that will enable control of the error probability.

This information geometric approach fits into the IPCA framework. Consider the following theorem:

Theorem 1: Let RVs $X, X' \in \mathbb{R}^d$ have PDFs f_X and $f_{X'}$, respectively. Using the $m \times d$ matrix A satisfying $AA^T = I_m$, construct RVs $Y, Y' \in \mathbb{R}^m$ such that $Y = AX$ and $Y' = AX'$. The following relation holds:

$$D_{KL}(f_X, f_{X'}) \geq D_{KL}(f_Y, f_{Y'}),$$

where f_Y and $f_{Y'}$ are the PDFs of Y, Y', respectively.

The proof of this theorem may be found in [17] and states that the KL-divergence cannot be increased through an orthonormal transform of the input space. This is intuitive, as an orthonormal transform is simply a rotation, which cannot increase distance. As such, maximizing the information distance between PDFs in a low-dimensional space is directly related to preserving said distance, albeit with a different formulation.

The first difference is in the setup of the data. We now specify that $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ where $\boldsymbol{X}_i$ consists of all points $x \in \mathbb{R}^d$ in class C_i; estimating the PDF of $\boldsymbol{X}_i$ as $p_i(x)$. Our objective function for the supervised scenario undergoes a slight modification to become:

$$A = \arg \max_{A:AA^T=I} \sum_i \sum_j W_{ij} D_{ij}(\mathcal{X}; A)^2. \tag{6}$$

We refer to this modified algorithm as *Information Maximizing Component Analysis* (IMCA) [12] []. By

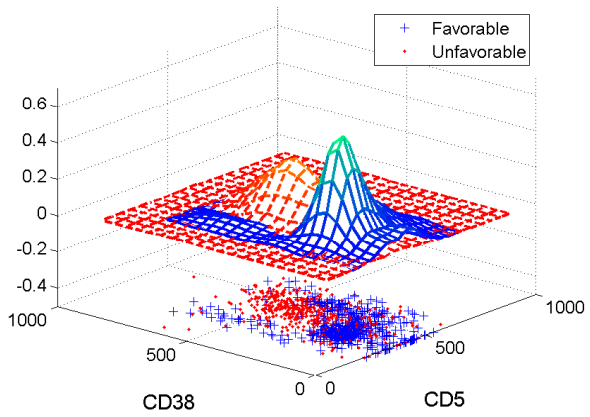


Fig. 7. The point cloud method of analyzing flow cytometry data is parallel to the analysis of the marginal densities of the data distributions.

maximizing the information distance between class PDFs, we not only ensure an optimal performance bound on classification error, but we also preserve the natural information geometry between classes. This fact is critical when class PDFs are not linearly separable (e.g. such is the assumption of standard LDA). Note that the optimization of the IMCA cost function may be done in a similar fashion to that of IPCA. In fact, for the 2-class problem, IPCA and IMCA are identical. For our purposes we use gradient *ascent*, as the objective is now a maximization, and the calculation/code is quite similar. Note that we may still use the IMCA projection matrix for variable selection, with the knowledge that the variables with the highest weights are those which contain the most discriminative value, which is critical for classification tasks.

It is worth explicitly pointing out that IMCA is similar to LDA. In fact, if the classes are Gaussian, IMCA would result in an orthogonal version of LDA. Recall that LDA assumes Gaussian classes and maximizes the between-class covariance while minimizing the within-class spread. This would maximize the information distance between the classes. Hence, IMCA can be viewed as a generalized and orthogonal version of LDA, which does not make assumptions on the class distribution.

V. Flow Cytometry

In clinical flow cytometry, pathologists gather readings of fluorescent markers and light scatter off of individual blood cells from a patient sample, leading to a characteristic multi-dimensional distribution that, depending on the panel of markers selected, may be distinct for a specific disease entity. Clinical pathologists generally interpret results in the form of two-dimensional scatter plots in which the axes each represent one of the many cell characteristics analyzed; the multi-dimensional nature of flow cytometry

Marker	Loading
Forward Light Scatter	0.1843
Side Light Scatter	0.1044
CD5	0.6270
CD38	0.8420
CD45	0.7228
CD19	0.5750

TABLE I
CLL ANALYSIS MARKERS THEIR CORRESPONDING IPCA LOADING WEIGHTS.

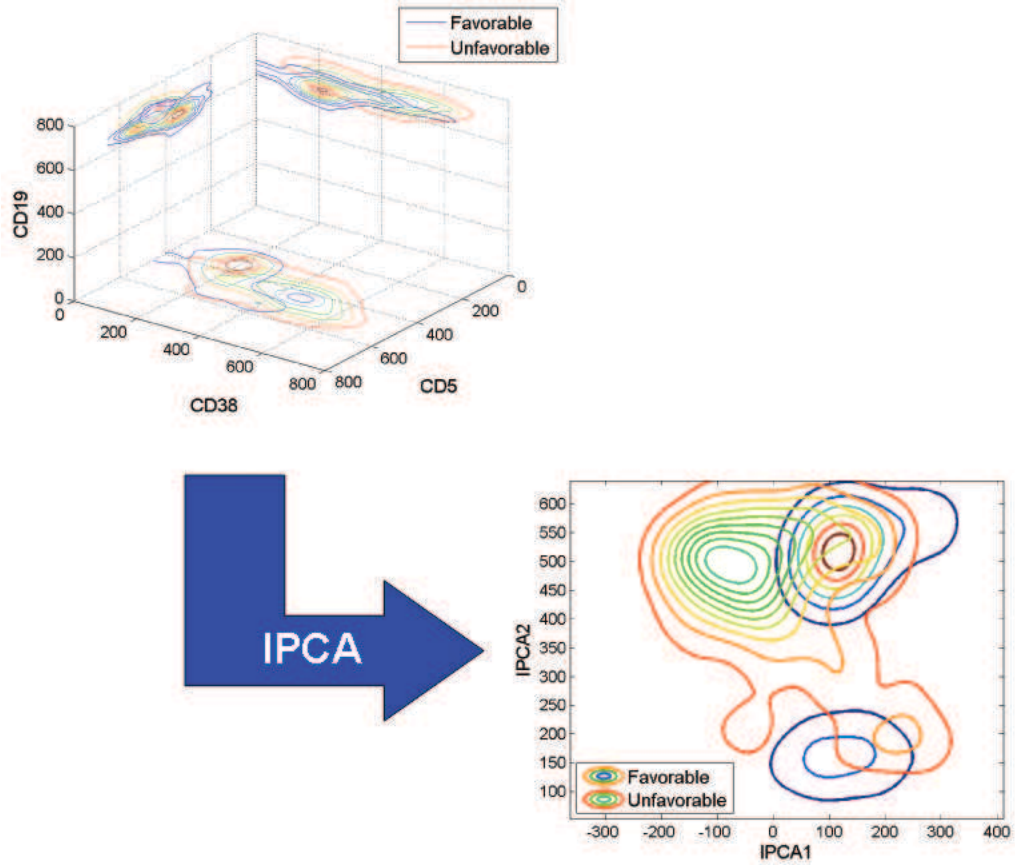
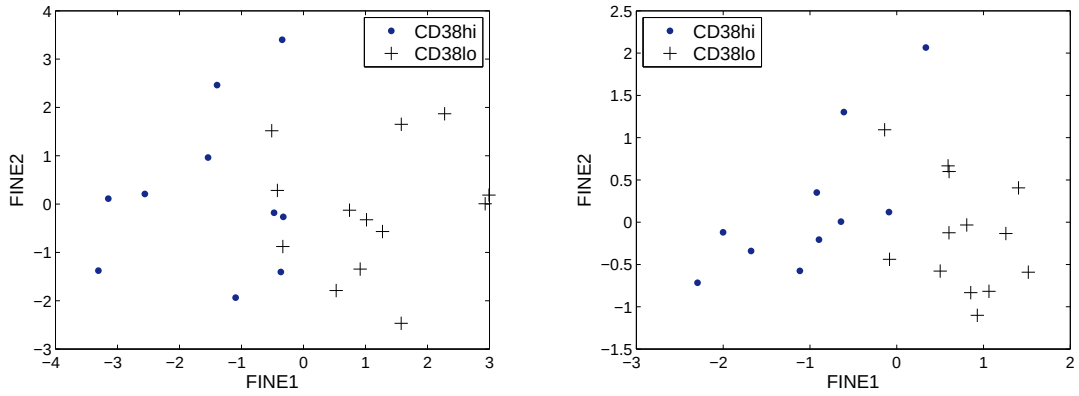


Fig. 8. Contour plots (i.e. PDFs) for 3 of the 6 analysis dimensions for CLL prognosis. The data for these patients is then transformed by IPCA, yielding a simple and easily discernable 2-dimensional analysis space. The patients chosen are the most similar favorable and unfavorable prognosis CLL patients.

is routinely underutilized in practice. Given the manner in which analysis is performed on point clouds, pathologists are actually performing a visual density analysis, as illustrated in Fig.7. Here we demonstrate the similar marginal densities (with respect to 2 bio-markers) of patients with differing prognosis. This enables the utilization of IGDR methods to provide a single analysis space for pathologists.

We present a study of chronic lymphocytic leukemia (CLL) patients, using IPCA to find a low-dimensional space which preserves the differentiation between patients with good and poor prognoses (i.e. favorable and unfavorable immunophenotypes). Using a collection of 23 patients diagnosed with



(a) Full Dimensional (b) IPCA Projection

Fig. 9. Comparison of CLL patient embeddings, obtained with FINE, using (a) the full dimensional and (b) the IPCA projection matrix. The patients with a poor immunophenotype (CD38hi) are generally well clustered against those with a favorable immunophenotype (CD38lo) in both embedddings.

CLL[3], 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 I. We use IPCA to determine the optimal information-preserving projection space, and illustrate this projection in Fig. 8. This image shows the 3-dimensional measurement space of markers CD5, CD38, and CD19; comparing two very similar patients with differing prognosis. It should be clear that IPCA provides a projection space for which discerning prognosis is simplified.

In Table I we also display the loading weights of each of the markers in the IPCA projection matrix. This is done by taking the vector norm of each column in the 2×6 IPCA matrix. Note that CD38 has the largest loading value; literature [20] has shown that patients whose leukemic cells are strong expressers of CD38 have significantly worse survival outcome. We also identify the possibility that CD45 and CD19 expression are also areas which may help prognostic ability, which is an area for further investigation.

Using FINE to embed the data (Fig. 9) for comparative visualization, we see that the different prognosis groups are very similar, although decent clusters are formed when labels are applied. These clusters are not well separated, however, which further illustrates the difficulties in forming an appropriate prognosis. There are also issues of sample size, as a larger database of patients may lead to a more clear separation of clusters. Nonetheless, IPCA and FINE were able to appropriately identify the important markers for assigning prognosis, and group patients accordingly with respect to immunophenotype. For additional details on this and other studies of FINE and IPCA with flow cytometry, we refer the reader to [9], [21]

[3]Courtesy of the Department of Pathology at the University of Michigan

VI. Crime in the 90s

We next illustrate IDGR to the analysis of crime indicators from 1990 U.S. census data. This data will be used to illustrate how information geometry can be used to discover which community and law enforcement features may be indicative of the level of crime seen in said community. We obtained the data from the UCI Machine Learning Repository [22], which is described in an abbreviated fashion as follows:

> The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. Attributes were picked if there was any plausible connection to crime ($N = 122$), plus the attribute to be predicted (Per Capita Violent Crimes). The variables included in the data set involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.
>
> All numeric data was normalized into the decimal range [0.00-1.00] using an unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small). E.g. An attribute described as 'mean people per household' is actually the normalized (0-1) version of that value.
>
> Since this data set was developed to identify potential crime indicators, the natural partitioning comes by grouping communities by the Per Capita Violent Crimes (PCVC) indicator variable. We note that while the data set contains 122 features, 22 of those features were only available for a small minority of communities, so we removed them from the set. This left us with a data set consisting of 1993 communities measured by 100 features. We omit the full feature list, which can be found in [22], however we will make explicit note of some selected features shortly.

A. A Distinct Difference

Although it is intuitive to think that communities with high rates of violent crime contain inherent differences than those on the opposite end of the spectrum, it is worth noting that none of the measured features are directly related to crime. Hence, it is worthwhile to first confirm our

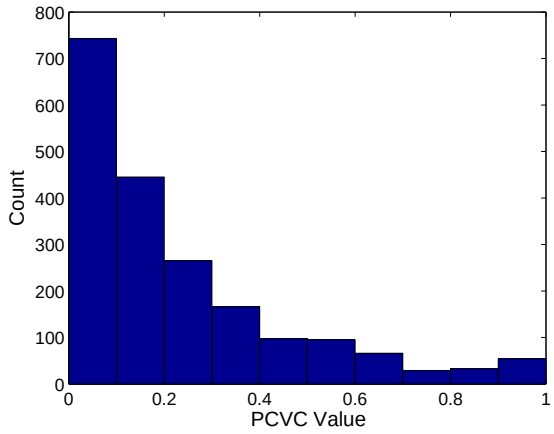


Fig. 10. Histogram of the per capita violent crime statistic for the measured communities.

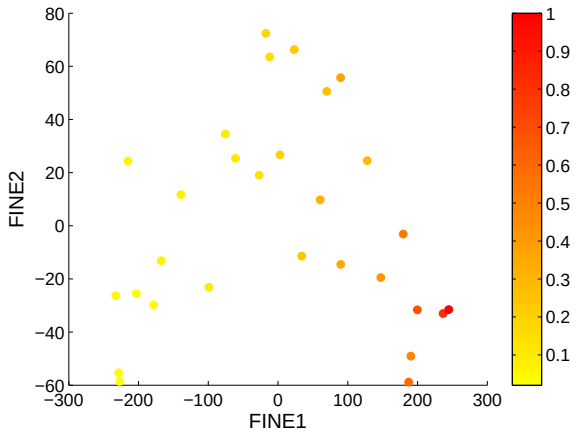


Fig. 11. Embedding the crime-based community groupings with FINE. The color of each sample corresponds to the maximum per capita violent crime rate within the group.

initial intuition. Additionally, if these features are truly indicative of violent crime, it is reasonable to expect a smooth gradient of change in the features from one end of the spectrum to the other. For example, if a low median family income "indicates" the potential for a high amount of crime, and vice versa, then it should be expected that a mid-range median family income should correspond to mid-range crime rates.

We set up this study by grouping communities with respect to their PCVC values. Recall that the range of PCVC is $[0.00, 1.00]$, with a distribution illustrated in the histogram of Fig. 10. As this distribution is highly non-uniform, we use non-uniform bin ranges to group the communities, intended to keep each bin with roughly the same number of samples. This leaves us with a set of 29 crime-based groupings $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{29}\}$, each consisting of between 50 and 122 sample points.

Using kernel density estimation to approximate group PDFs, we embed each crime grouping into a 2-dimensional space with FINE. The embedding results can be seen in Fig. 11, where each 2-dimensional sample point represents a collection of communities whose maximum PCVC value is identified by plot color. It is clear that our intuition was correct; there exists a smooth and continuous gradient of increasing crime rate. This leads to the natural conclusion that the collection of measured features (or some subset thereof) does indeed contain predictive indicators of violent crime rates.

B. Predicting Crime and Discriminating Features

Given the confirmation that the chosen features do indeed contain predictive value, we now test the classification capabilities when using IGDR as a pre-processing step. Specifically, we look to find the optimal subspace for classifying a community as having either low or high rates of violent crime. This sets up as a 2 class problem, and we determine the low-crime class as those communities having a PCVC value of 0.03 or less, and the high-crime class contains communities with a PCVC value greater than 0.53. These thresholds were chosen such that each class contained roughly the same number of samples (226 and 239 respectively).

Given that this is a classification problem, we use IMCA to determine our optimal *orthonormal* projection matrix. Note that we stress the orthonormality constraint here, as using Fisher's LDA, which does not result in orthogonality, may seem appropriate for this task. If classification was the only desirable task, then LDA would be sufficient. However, we also intend to analyze the projection matrix for variable selection, for which orthogonality becomes a necessity. The LDA projection is useful, however, as it gives us a means for initialization; we make the LDA matrix orthogonal with the classical Gram-Schmidt orthogonalization algorithm, and initialize our IMCA gradient methods with the resultant matrix.

We choose to perform our analysis in an $m = 3$-dimensional projection space for 2 reasons – the ability to visualize the data, and the 3-dimensional space optimized our objective, obtaining the maximum separation between classes for $m \in [2, 7]$. After obtaining the 3×100 IMCA matrix A, we project the data from each class into the same space, and perform our classification task. The projected data is shown in Fig. 12. It is interesting that the low-crime communities show much more variation than the high-crime communities, which exhibit form a tight cluster even though the range of PCVC value was much larger.

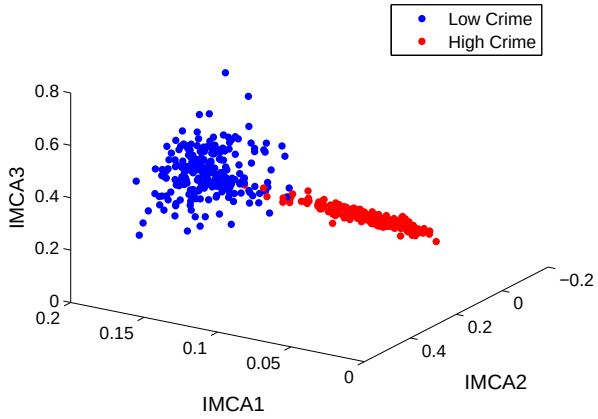


Fig. 12. The IMCA projection of communities based on the classes defined by low and high PCVC values.

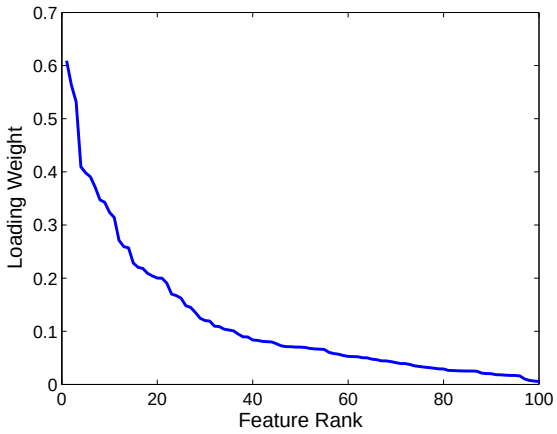


Fig. 13. The rankings of the 100 variables in IMCA prjection matrix.

To test classification performance, we use a simple linear classifier and perform leave-one-out cross-validation over all samples in the set. The results yield a 1.29% classification error – 1 low-crime and 5 high-crime communities were misclassified. For comparison sake, we note that principal components analysis (PCA), an orthonormal unsupervised method, results in a 3.44% error rate, and LDA yielded a 1.51% error rate. Recall, that LDA does not have the orthogonal constraint, yet IMCA still results in (slightly) better classification performance. In all cases, the projection data was projected to 3 dimensions.

We now use the IMCA matrix A to identify the most discriminating features. To do such, we calculate the L_2-norm of the vector of weights for each of the 100 features (columns) of the 3×100 projection matrix A. After sorting in descending order, we plot these ranks in Fig. 13. This shows that there are several features which offer some discriminative value, and many more that offer very little. In Table II, we report the 5 most and 5 least discriminating features. We

Top 5 Variables
Population For Community
Rental Housing - Median Rent
Number Of People Living In Areas Classified As Urban
Percentage Of Population Who Are Divorced
Median Household Income
Bottom 5 Variables
Per Capita Income For People With Hispanic Heritage
Percent Of Officers Assigned To Drug Units
Per Capita Income For People With Asian Heritage
Land Area In Square Miles
Median Year Housing Units Built

TABLE II
THE 5 MOST AND LEAST DISCRIMINATING FEATURES FOR PREDICTING HIGH OR LOW RATES OF VIOLENT CRIME.

preface these results by recalling that this data was from a 1990 census and 1995 crime reporting. Obviously much has changed since this data was reported, but the results do appear logical.

VII. CONCLUSION

In this article, we have presented IGDR; an information-geometric framework for dimensionality reduction. As contrasted to standard Euclidean approaches to manifold learning, which aim to reconstruct a Riemannian sub-manifold of Euclidean space, our objective is to learn statistical manifolds. We have shown that when the data produces realizations of probability density functions lying on a statistical manifold, we can perform information-driven dimensionality reduction in both the density space and the sample space. These techniques were illustrated on the problem of flow cytometry analysis, showing the ability to find a subspace in which a pathologist can better diagnose chronic lymphocytic leukemia patients. We were also able to compare patients one to another in a single low-dimensional embedding space. We also applied IGDR to a crime and community data set, identifying community indicators of violent crime and accurately clustering and classifying communities with high or low crime rates. The power of using information-geometry for dimensionality reduction has just begun to be explored and we hope this article will lead to further extensions and applications.

A. Acknowledgement

This research was supported in part by ARO grant W911NF-09-1-0310, NSF grant CCF 0830490 and an AFRL ATR Center grant through SIG Inc. The authors also thank Christine

Kim of the University of Michigan who collected the data used for Fig. 5 while a summer intern at AFRL.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[3] Y.Y. Sun, M.K. Ng, and Z.H. Zhou, "Multi-Instance Dimensionality Reduction," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, pp. 587–592.

[4] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.

[5] 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.

[6] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Fine: Fisher information non-parametric embedding," *IEEE Transactions on Pattern Analyis and Machine Intelligence*, vol. 31, no. 11, pp. 2093–2098, Nov. 2009.

[7] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," vol. 11, pp. 451–490, Feb. 2010.

[8] K. M. Carter, R. Raich, and A. O. Hero, "Spherical laplacian information maps (slim) for dimensionality reduction," in *Proc. IEEE Inter. Conf. on Statistical Signal Processing*, August 2009.

[9] 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: Special Issue on Digital Image Processing Techniques for Oncology*, vol. 3, no. 1, pp. 148–158, Feb. 2009.

[10] J. Peltonen, "Visualization by linear projections as information retrieval," in *Proc. of the 7th International Workshop on Advances in Self-Organizing Maps*, 2009, pp. 237–245.

[11] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer-Verlag, NY, 1998.

[12] K. M. Carter, R. Raich, and A. O. Hero III, "An information geometric approach to supervised dimensionality reduction," in *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, pp. 1829–1832.

[13] *Information Geometric Dimensionality Reduction Toolbox*, http://tbayes.eecs.umich.edu/kmcarter/igdr/index.html.

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

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

[16] S.K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, June 2006.

[17] K. M. Carter, *Dimensionality Reduction on Statistical Manifolds*, Ph.D. thesis, University of Michigan, Jan. 2009.

[18] S-M. Lee, A. L. Abbott, and P. A. Araman, "Dimensionality reduction and clustering on statistical manifolds," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2007, pp. 1–7.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems, Volume 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.

[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] 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, pp. 1–7, Jan. 2009.

[22] *UCI Machine Learning Repository: Communities and Crime Data Set*, available at http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.