AN INFORMATION GEOMETRIC APPROACH TO SUPERVISED DIMENSIONALITY REDUCTION

Kevin M. Carter[1]*, *Raviv Raich*[2], *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
{kmcarter,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

Due to the *curse of dimensionality*, high-dimensional data is often pre-processed with some form of dimensionality reduction for the classification task. Many common methods of supervised dimensionality reduction have focused on separating and collapsing the data near the class centroids. These methods often make assumptions on the distributions of the data classes – namely Gaussianity – which can lead to ad-hoc and sub-optimal implementation. In this paper we present a method of supervised dimensionality reduction which takes an information-geometric approach by maximizing the between class information distances. This is shown to have direct relation to the Chernoff and Bhattacharya performance bounds for classification error. We illustrate our methods on real data and compare to several existing methods.

Index Terms— Information geometry, statistical manifold, dimensionality reduction, classification

1. INTRODUCTION

As the efficiency of data retrieval increases, and the costs of storage decreases, many applications have been developed which generate massive amounts of data. This raises the issue of analysis, as high-dimensional data sets suffer from the *curse of dimensionality*. There has been much work presented in which high dimensional data is projected into a low dimensional space to aid in various learning tasks such as classification. While several supervised algorithms [1, 2] have been presented which operate in the non-linear framework, many of the commonly presented methods of supervised dimensionality reduction focus on linear projections, which do not require a re-processing of the low-dimensional space when new data becomes available. An extensively utilized method of linear supervised dimension reduction, Fisher's linear discriminant analysis (LDA) [3], has been the inspiration for many derivatives [4, 5]. These methods have been created specifically for the classification task, and offer projections which are ideal for separating Gaussian data classes. When classes do not follow a normal distribution, these methods often suffer performance losses, although they have proven to be robust for many applications.

In this paper we use an information-geometric form of dimensionality reduction deemed *information preserving component analysis* (IPCA) towards the problem of pre-processing for the classification task. We have recently utilized IPCA in an unsupervised manner [6] for the tasks of visualization and feature extraction on multiple related high dimensional, large sample size data sets (e.g. samples of patient blood cells). We characterized each data set as some generative model and found the linear projection which preserved the high dimensional Fisher information distances (between all pairs of data sets) in the low dimensional space.

We now adapt IPCA to work in the supervised realm, in which each class is characterized by a probability density function (PDF), and we find the low-dimensional space which maximizes the information distances between class PDFs. Unlike traditional LDA methods, IPCA makes no assumption on the class distributions, only that each PDF lies on some *statistical manifold* for which the Fisher information distance is an appropriate metric. Previous methods [7] have used information theory for dimensionality reduction, but focus mainly on measuring the entropy or mutual information of the samples to their class labels. IPCA uses the information geometry of the statistical manifold to find the optimal low-dimensional subspace which maximizes the information distance between classes. This is directly related to the Bhattacharya and Chernoff performance bounds, resulting in superior classification.

This paper proceeds as follows: In Section 2, we formulate the problem we will attempt to solve. We present our methods for finding the IPCA projection in Section 3. Simulation results for real data are shown in Section 4, followed by a discussion and areas for future work in Section 5.

2. PROBLEM FORMULATION

The Chernoff performance bound [8] is related to the Chernoff distance between two probability density functions (PDFs)

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

$f(x)$ and $g(x)$,

$$D_{CH}(f,g) = -\log \int f(x)^{\alpha} g(x)^{1-\alpha}\, dx,$$

where $0 \leq \alpha \leq 1$. Let $f(x)$ and $g(x)$ be the PDFs of two distinct data classes $\boldsymbol{X}_f$ and $\boldsymbol{X}_g$ respectively. As $D_{CH}(f,g)$ increases, the lower bound on the probability of classification error between points in $\boldsymbol{X}_f$ and $\boldsymbol{X}_g$ decreases. A special case of the Chernoff distance is when $\alpha = \frac{1}{2}$, and is known as the Bhattacharya distance,

$$D_B(f,g) = -\log \int \sqrt{f(x)g(x)}\, dx,$$

which has been used to bound the classification error for dimensionality reduction [9]. Hence, an ideal form of dimensionality reduction would ensure that the Bhattacharya distance between all classes is maximized, which would allow for control of error probability.

Specifically, given a data set $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$, where $\boldsymbol{X}_i$ consists of data points $x \in \mathbb{R}^d$ belonging to class i we can define a similarity between data classes $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ with the Bhattacharya as $D_B(p_i(x), p_j(x))$, where p_i and p_j are the estimated PDFs of classes i and j respectively. Can we find a mapping

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

in which the elements $y \in \boldsymbol{Y}$ exist in $\mathbb{R}^m$, $m < d$ which maximizes $D_B(p_i(y), p_j(y)),\ \forall\, i,j$? As to minimally alter the natural geometry of the data, we focus solely on linear and orthonormal projection matrices (i.e. rotations).

3. METHODS

It should be noted that the Bhattacharya distance is a monotonic transformation of the Hellinger distance,

$$D_H(f,g) = \sqrt{\int \left(\sqrt{f} - \sqrt{g}\right)^2 dx},$$

such that

$$D_B(f,g) = -\log\left(1 - \frac{1}{2} D_H^2(f,g)\right).$$

This transformation is important as it allows us to modify our original desire of maximizing the Bhattacharya distance between class PDFs to that of maximizing the Hellinger distance between classes. The Hellinger distance has been shown to converge to the Fisher information distance, which is the natural metric on a *statistical* manifold (i.e. a manifold of PDFs) [10]. While some may argue to simply maximize the Bhattacharya distance between class PDFs, as was parametrically done in [11], by framing the problem in the manner in which we have, we now offer an information geometric approach with no increase in complexity.

This information geometric approach fits into a framework which we have recently presented deemed *information preserving component analysis* (IPCA) [6], which is an unsupervised method of dimensionality reduction which preserves the high-dimensional information distances between data sets in a low-dimensional space. Specifically, in [6] we found the projection matrix A which optimized a cost function similar to

$$A = \arg \min_{A:AA^T=I} \|D(\mathcal{X}) - D(\mathcal{X}; A)\|_F^2, \qquad (1)$$

where $\|\cdot\|_F$ is the standard Frobenius norm, I is the identity matrix, $D(\mathcal{X})$ is a distance matrix such that $D_{ij}(\mathcal{X}) = D_H(\boldsymbol{X}_i, \boldsymbol{X}_j)$, and $D(\mathcal{X}; A)$ is a similar matrix where the elements are perturbed by the projection matrix A, $D(\mathcal{X}; A) = D_H(A\boldsymbol{X}_i, A\boldsymbol{X}_j)$. With an abuse of notation, we refer to $D_H(p_i, p_j)$ as $D_H(\boldsymbol{X}_i, \boldsymbol{X}_j)$, with the knowledge that the distance is calculated with respect to PDFs, not realizations.

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:*

$$D_H(f_X, f_{X'}) \geq D_H(f_Y, f_{Y'}), \qquad (2)$$

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

Due to space restrictions, we omit the proof of this theorem which may be found in [12]. The nature of this proof is two-fold – first, we show that the Hellinger distance is constant over an arbitrary dimension preserving orthonormal transformation. Next, we show that the same truncation of two random vectors does not increase distance. This implies that maximizing the Hellinger distance in the lower dimensional space is directly related to minimizing the difference (i.e. preserving) between the high and low dimensional distances; they are indeed equivalent statements in the 2 class case. Hence, our objective of finding the projection which maximizes the distance between PDFs is parallel to the objective of preserving the distances between PDFs, albeit with a different formulation. With this knowledge, we will still refer to our supervised framework as IPCA.

Let us now define the IPCA projection as one that maximizes the information distance between data sets. Specifically, let $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ where $\boldsymbol{X}_i$ consists of all points $x \in \mathbb{R}^d$ in class i; estimating the PDF of $\boldsymbol{X}_i$ as p_i. Formatting as an optimization problem, we would like to solve:

$$A = \arg \max_{A:AA^T=I} \|D(\boldsymbol{X}; A)\|_F^2. \qquad (3)$$

By 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.

3.1. Optimization

Gradient ascent (or the method of *steepest* ascent) allows for the solution of convex optimization problems by traversing a surface or curve in the direction of greatest change, iterating until the maximum is reached. Specifically, let $J(A) = \|D(\boldsymbol{X}; A)\|_F^2$ be our objective function, measuring the total information distance between all class PDFs in our projection space. The direction of the gradient is solved by taking the partial derivative of J w.r.t. a projection matrix A,

$$\frac{\partial}{\partial A} J(A) = \sum_i \sum_j 2D_{ij}(\mathcal{X}; A) \frac{\partial}{\partial A} D_{ij}(\mathcal{X}; A).$$

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

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

is the direction of the gradient, constrained to force A to remain orthonormal, and μ is a small number regulating the step size. Variables Q_0 and Q_1 are defined as

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

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

The full derivation of this constraint, as well as specific implementation details for estimating the information distances, can be found in [12]. This process is iterated until the objective $J(A)$ converges.

3.2. Algorithm

The full method for IPCA, specialized towards the classification task, is described in Algorithm 1. We note that 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 maximum rather than an absolute maximum, experimental results on our available data has shown that the algorithm converges to the same point given several random initializations. If *a priori* knowledge of the global maximum is available, one would initialize A in its vicinity.

Algorithm 1 Information Preserving Component Analysis

Input: Collection of data classes $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ in $\mathbb{R}^d$; projection dimension m; search step size μ; threshold ϵ

1: Initialize $A_1 \in \mathbb{R}^{m \times d}$ as a random orthonormal projection matrix
2: Calculate $D(\mathcal{X}; A_1)$, the information distance matrix in the projected space
3: **while** $|J_i - J_{i-1}| > \epsilon$ **do**
4: Calculate $\frac{\partial}{\partial A_i}\tilde{J}$, the direction of the gradient, constrained to $AA^T = I$
5: $A_{i+1} = A_i + \mu \frac{\partial}{\partial A_i}\tilde{J}$
6: Calculate $D(\mathcal{X}; A_{i+1})$
7: $J = \|D(\mathcal{X}; A_{i+1})\|_F^2$
8: $i = i + 1$
9: **end while**

Output: Projection matrix $A \in \mathbb{R}^{m \times d}$, which maximizes the information distances between class PDFs.

4. SIMULATION

We now study the performance of IPCA for supervised dimensionality reduction, utilizing the well studied Landsat satellite imagery database [13]. This data set consists of satellite images of 6 differing soil types. Each sample point is a 36-dimensional vector corresponding to the 9 intensity values of a 3×3 pixel region (with overlapping regions) in 4 different spectral bands. The data set has been pre-defined as containing 4435 training samples and 2000 test samples.

We compare IPCA performance to other methods of linear, supervised dimensionality reduction: linear discriminant analysis (LDA) [3] and quadratic discriminant analysis with slice average variance estimation (QDA-SAVE) [14]. We implement several different classification methods – linear, radial, and quadratic kernel support vector machines (SVMs) [15], and a k-nearest neighbor (k-NN) classifier – as different methods of dimensionality reduction may be optimized specifically for certain classification methods (e.g. LDA and linear classification). In Table 1, we illustrate the "best case" classification performance for all simulations, in which the lowest error rate is reported over all projection dimensions with values in the range $m \in \{3 - 25\}$, emphasizing the best performance for each classifier. We see that IPCA outperforms LDA and QDA-SAVE for all classifiers except the quadratic kernel SVM, for which QDA-SAVE narrowly shows better performance.

	Linear	Radial	Quadratic	k-NN
IPCA	**13.60 %**	**9.85 %**	10.05 %	**9.70 %**
LDA	13.70 %	11.35 %	11.25 %	12.60 %
QDA-SAVE	13.65 %	10.15 %	**9.90 %**	10.15 %

Table 1. Classification error probability

We further investigate the performance with the quadratic kernel SVM by plotting the classification error as a function of dimension in Fig. 1. It is clear that QDA-SAVE has significant difficulties in the low dimensional regime, which may be an issue if significant dimensionality reduction is required

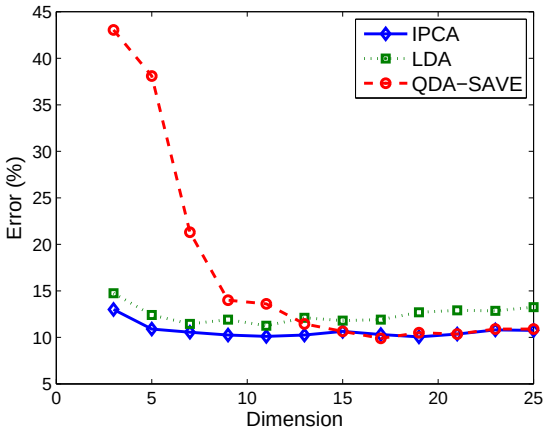

Fig. 1. Classification error probability as a function of dimension when using a quadratic kernel SVM. IPCA shows far superior performance in low dimensions, while still performing comparable to QDA-SAVE in the 'best case'.

(e.g. compression). In contrast, IPCA shows far superior performance in low dimensions, while still maintaining strong competitiveness in high dimensions. While not illustrated, we noticed similar performance with the other classifiers.

5. CONCLUSIONS

In this paper we have shown the ability to find an information geometric projection for supervised dimensionality reduction using information preserving component analysis. By maximizing the information distance between class PDFs, we find a low-dimensional projection which alleviates the *curse of dimensionality* and improves classification performance. We have theoretically shown a direct relation to the Bhattacharya and Chernoff performance bounds, and experimentally demonstrated that space defined by IPCA gives superior classification performance to comparable methods of supervised dimensionality reduction, and is not biased towards any single classifier. In future work we plan to continue applying IPCA towards the classification task, and extend to semi-supervised learning problems.

6. REFERENCES

[1] R. Raich, J. A. Costa, and A. O. Hero, "On dimensionality reduction for classification and its applications," in *Proc. IEEE Intl. Conference on Acoustic Speech and Signal Processing*, May 2006.

[2] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood component analysis," in *Neural Information Processing Systems*, 2004, number 17, pp. 513–520.

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

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

[5] 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.

[6] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero III, "Dimensionality reduction of flow cytometric data through information preservation," in *IEEE Machine Learning for Signal Processing Workshop*, Oct. 2008.

[7] K. E. Hild II, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1385–1392, Sept. 2006.

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

[9] 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.

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

[11] M. Thangavelu and R. Raich, "Multiclass linear dimension reduction via a generalized chernoff bound," in *IEEE Machine Learning for Signal Processing Workshop*, Oct. 2008.

[12] 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.

[13] *UCI Machine Learning Repository: Statlog (Landsat Satellite) Data Set*, available at `http://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)`.

[14] I. Pardoe, X. Yin, and R. D. Cook, "Graphical tools for quadratic discriminant analysis," *Technometrics*, vol. 49, no. 2, May 2007.

[15] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.