Shrinkage Fisher Information Embedding of High Dimensional Feature Distributions

Xu Chen
Department of EECS
University of Michigan
Ann Arbor, MI, 48109
Email: xhen@umich.edu

Yilun Chen
Department of EECS
University of Michigan
Ann Arbor, MI, 48109
Email: yilun@umich.edu

Alfred Hero
Department of EECS
University of Michigan
Ann Arbor, MI, 48109
Email: hero@umich.edu

Abstract— **In this paper, we introduce a dimensionality reduction method that can be applied to clustering of high dimensional empirical distributions. The proposed approach is based on stabilized information geometrical representation of the feature distributions. The problem of dimensionality reduction on spaces of distribution functions arises in many applications including hyperspectral imaging, document clustering, and classifying flow cytometry data. Our method is a shrinkage regularized version of Fisher information distance, that we call shrinkage FINE (sFINE), which is implemented by Steinian shrinkage estimation of the matrix of Kullback Liebler distances between feature distributions. The proposed method involves computing similarities using shrinkage regularized Fisher information distance between probability density functions (PDFs) of the data features, then applying Laplacian eigenmaps on a derived similarity matrix to accomplish the embedding and perform clustering. The shrinkage regularization controls the trade-off between bias and variance and is especially well-suited for clustering empirical probability distributions of high-dimensional data sets. We also show significant gains in clustering performance on both of the UCI dataset and a spam data set. Finally we demonstrate the superiority of embedding and clustering distributional data using sFINE as compared to other state-of-the-art methods such as non-parametric information clustering, support vector machine (SVM) and sparse K-means.**

I. Introduction

The use of information theoretical measures for dimensionality reduction and machine learning has been studied over several decades in diverse applications including gene expression data clustering and document classification [1] [3] [6] [5]. The authors of [3] proposed parametric clustering algorithms based on Bregman divergences. The work in [6] uses non-parametric estimation of the average cluster entropies to search for a clustering that maximizes the estimated mutual information between data points and clusters. The Fisher information distance is the unique intrinsic distance metric to compute the distance between two probability density functions in a space of parameterized probability densities known as a statistical manifold in information geometry [2]. Our previous work [5] used a non-regularized estimate of Fisher information distance to perform non-linear embedding for visualization and document classification. However, for high dimensional data and a small number of samples, these estimators suffer from severe overfitting errors and therefore lead to inaccurate estimates. In this work, we introduce a new shrinkage regularized Fisher information distance as a similarity measure for dimensionality reduction on statistical manifolds. The corresponding embedding is called: shrinkage Fisher information distance (sFINE) and leads to more stable embeddings than the previous FINE framework [5] when the dimension is high.

The shrinkage Fisher information is a shrinkage regularized version of Fisher information. The traditional Fisher information relies on accurate estimation of the probability density functions (PDFs) of the features. The PDFs of these features are usually estimated using maximum likelihood for the multinomial distribution, leading to an empirical histogram estimator. However, for high dimensional data and small number of samples, the maximum likelihood estimator does not minimize mean square error (MSE) and therefore leads to inaccurate estimates. The estimates of the PDFs can be significantly improved by incorporating regularization. Unlike previous approaches to regularization of PDFs [13] [12], our estimation approach optimizes the MSE over the shrinkage estimator family and this translates into improved performance. In [13], Witten and Tibshirani proposed sparse K-means clustering by maximizing a weighted between-cluster sum of squares subject to L1 type constraints on the weights. In [12], Williams proposed to use a Laplace prior for sparse representation and regularization. Here we propose to use the James-Stein shrinkage estimator of Fisher information [8]. Compared to a Bayesian Laplace prior [12], our shrinkage estimator has the advantage of not requiring subjective assumptions on prior distributions of features.

The Fisher information can be approximated by Kullback Leibler (KL) divergence when the specific parameterization of the manifold is unknown [9]. In this paper, we estimate the shrinkage KL-divergence to approximate shrinkage Fisher information distance. We derive the bias and variance for shrinkage KL divergence estimator. In sFINE, the PDFs of features are embedded into a lower dimensional Euclidean space by applying Laplacian eigenmaps [4] on the distance matrices. Once the embedding is accomplished, we apply the K-means algorithm in the lower dimensional Euclidean space to cluster the data. The proposed embedding and clustering methods are evaluated on synthetic data, on a UCI dataset and on time-varying email spam data. The spam data has multiple attributes, span mail server and temporal usage. The proposed shrinkage Fisher information has the advantage that it can embed not only marginal PDFs for a single attribute but also joint PDFs for multiple attributes. The joint embedding incorporates more information than the marginal embedding and is therefore more discriminative. We demonstrate the effectiveness of the proposed methods for clustering spam data compared to a previous implementation using correlation-based spectral clustering [14]. When compared on standard UCI-dataset clustering, the sFINE clustering method outperforms the performance of state-of-the-art methods including: non-parametric information clustering [6]; sparse K-means [13]; and unsupervised support vector machine (SVM).

II. Problem Formulation

Let M be a family of probability density functions (PDF) parameterized by $\theta = [\theta^1, \ldots, \theta^n]$:

$$M = \{p(x|\theta) | \theta \in \Theta \subseteq R^n\} \ . \qquad (1)$$

M is a statistical manifold when the Fisher information is used as a Riemannian metric for measuring distances between the distributions in M. We define the Fisher information matrix $[I(\theta)]$ with elements

$$I_{ij} = E[\frac{\partial}{\partial \theta^i} log f(X;\theta) \frac{\partial}{\partial \theta^j} log f(X;\theta)] \ . \qquad (2)$$

The Fisher information distance can be accurately approximated by the KL divergence when the specific parameterization of the manifold is unknown. The Kullback Leibler divergence is an important metric in information theory and is commonly referred to as the relative entropy of one PDF to another, which is defined as

$$KL(p//q) = \int p(x) log \frac{p(x)}{q(x)} \ . \qquad (3)$$

It should be noted that the KL-divergence is not a distance metric, as it does not satisfy the symmetry $KL(p//q) \neq KL(q//p)$. To enforce symmetry, we will define the KL-divergence as:

$$D_{KL}(p,q) = KL(p//q) + KL(q//p) \ . \qquad (4)$$

As described in [9], we can relate the symmetric KL-divergence and approximate the Fisher information distance as

$$\sqrt{D_{KL}(p,q)} \rightarrow D_F(p,q)$$

as $p \rightarrow q$.

A. Shrinkage KL divergence

Since Fisher information distance can be approximated by KL-divergence, consider a m_x dimensional feature variable $X \in R^{m_x}$ and an associated codebook $e = \{C_i, x_i\}_{i=1}^{p}$, where C_i are quantization cells and x_i are quantization levels. Let $Z = [z_1, \ldots, z_p]$ be a vector containing the number of times that a set of instances $\{X_i\}_{i=1}^{n}$ fall into each cell, where $Z_k = \sum_{i=1}^{n} I(X_i \in C_k)$ and $I(A)$ is the indicator function of event A. Then if $\{X_i\}_{i=1}^{n}$ are i.i.d., Z is multinomial distributed.

$$Prob(z_1, \ldots, z_p; \theta_1, \ldots, \theta_p) = \frac{n!}{\prod_{k=1}^{p} z_k!} \prod_{k=1}^{p} \theta_k^{z_k},$$
$$\sum_{k=1}^{p} z_k = n, \sum_{n=1}^{p} \theta_k = 1 \ . \qquad (5)$$

The James-Stein shrinkage estimator of $\theta = [\theta_1, \ldots, \theta_p]^T$ is a modified version of the maximum likelihood (ML) estimator that reduces the MSE of the estimator. It is based on shrinking the maximum likelihood estimator towards a target [8],

$$\hat{\theta}_k^{\lambda} = \lambda t_k + (1-\lambda)\hat{\theta}_k^{ML} \ , \qquad (6)$$

where $\{t_k\}_{t=1}^{p}$ is the target distribution, here chosen as uniform distribution $t_k = \frac{1}{p}$, $\hat{\theta}_k^{ML} = \frac{z_k}{n}$. The resultant shrinkage estimator has reduced variance but increased bias. However we can guarantee a decrease in the mean squared error (MSE) by proper choice of the shrinkage parameter λ, as explained below. The James-Stein plug-in entropy estimator is then defined as [8]:

$$\hat{H}_{\theta}(X) = -\sum_{k=1}^{p} \hat{\theta}_k^{\lambda} \log(\hat{\theta}_k^{\lambda}) \ , \qquad (7)$$

The KL-divergence estimator can be represented as

$$KL_{\theta}(X \parallel Y) = \sum_{k=1}^{p} E[\theta_x(k) \log(\frac{\theta_x(k)}{\theta_y(l)})] \ , \qquad (8)$$

where $\theta_x(k) = P(x=k) = \sum_{l=1}^{p} \theta_{k,l}, \theta_y(l) = P(y=l) = \sum_{k=1}^{p} \theta_{k,l}$. Define the shrinkage estimator of KL-divergence:

$$KL^{\lambda}(X \parallel Y) = \sum_{k=1}^{p} E[\theta_x^{\lambda}(k) \log(\frac{\theta_x^{\lambda}(k)}{\theta_y^{\lambda}(l)})] \ , \qquad (9)$$

and the symmetrized KL estimate

$$\hat{D}_{KL}^{\lambda}(X,Y) = KL^{\lambda}(X \parallel Y) + KL^{\lambda}(Y \parallel X) \ . \qquad (10)$$

Our goal is to find the value of λ in the KL-divergence that has minimum MSE:

$$\lambda_{D_{KL}}^{\circ} = \arg\min_{\lambda} E\{(\hat{D}_{KL}(X,Y)^{\lambda} - D_{KL}(X,Y))^2\} \ , \qquad (11)$$

B. Shrinkage estimator of KL divergence

The MSE of shrinkage KL-divergence can be decomposed into the sum of the square of the bias and the variance. The theoretical expressions of bias and variance given Propositions 1 and 2 below, will be utilized to calculate the optimal shrinkage parameter by minimizing MSE over λ, where we initialize the random $0 < \lambda < 1$ and iteratively estimate the minimum MSE and the optimal shrinkage parameter.

Proposition 1: The bias of KL-divergence with James-Stein shrinkage estimator is given by

$$Bias(KL_{\theta}^{\lambda}) = C_{b1} + C_{b2}\frac{1}{n} + O\left[\frac{1}{n^2}\right] \ , \qquad (12)$$

where

$$C_{b1} = \sum_{k=1}^{p} \left[\theta_x(k) \log \frac{\theta_x(k)}{\theta_y(l)} \right] - \sum_{k=1}^{p} \left[\frac{\lambda}{p} + (1-\lambda)\theta_x(k) \right]$$
$$\log \frac{\frac{\lambda}{p} + (1-\lambda)\theta_x(k)}{\frac{\lambda}{p} + (1-\lambda)\theta_y(l)} \ . \qquad (13)$$

$$C_{b2} = \sum_{k=1}^{p} \frac{1}{2} \frac{\theta_x(k)}{1-\theta_y(l)} \frac{1}{[1-\lambda(1-\frac{1}{p\theta_y(l)})]^2} +$$
$$\sum_{k=1}^{p} \lambda(\frac{1}{p} - \theta_x(k))(1-\theta_y(l)) \frac{1}{[1-\lambda(1-\frac{1}{p\theta_y(l)})]}$$
$$-\sum_{k=1}^{p} \frac{1}{2\log 2} \frac{1}{1-\lambda(1-\frac{1}{p\theta_x(k)})} (\theta_x(k) - 1) \ . \qquad (14)$$

Proposition 2: The shrinkage KL-divergence estimator is asymptotically Gaussian, where the mean $\mu(KL_\theta^\lambda) = \sum_{k=1}^{p}[\frac{\lambda}{p} + (1-\lambda)\theta_x(k)] \log \frac{\frac{\lambda}{p}+(1-\lambda)\theta_x(k)}{\frac{\lambda}{p}+(1-\lambda)\theta_y(l)}$, the variance $Var(KL_\theta^\lambda) = \frac{1}{n}T_1\Sigma T_1'$, where $T_1 = [\frac{\partial KL_\theta^\lambda}{\partial\theta_x(k)}, \frac{\partial KL_\theta^\lambda}{\partial\theta_y(l)}]$ is a 1 by $2p$ vector,

$$\frac{\partial KL_\theta^\lambda}{\partial\theta_x(k)} = (1-\lambda)[\log(\frac{\frac{\lambda}{p} + (1-\lambda)\theta_x(k)}{\frac{\lambda}{p} + (1-\lambda)\theta_y(l)})$$
$$+\frac{\lambda}{p} + (1-\lambda)\theta_y(l)]$$
$$\frac{\partial KL_\theta^\lambda}{\partial\theta_y(l)} = \frac{\lambda}{p} + (1-\lambda)\theta_y(l) \ .$$

Σ is the $2p$ by $2p$ covariance matrix given by:

$$\Sigma_{kk} = n\theta_x(k)(1-\theta_x(k)), \forall k = 1, \dots, p$$
$$\Sigma_{kk} = n\theta_y(k)(1-\theta_y(k)), \forall k = p+1, \dots, 2p$$
$$\Sigma_{ij} = -n\theta_x(k)\theta_x(l), \forall k \in [1,p], l \in [1,p], k \neq l$$

where other non-diagonal elements in the covariance matrix Σ can be computed similarly as $-n\theta_y(k)\theta_y(l)$, $\forall k \in [p+1, 2p], l \in [p+1, 2p], k \neq l$ and $-n\theta_x(k)\theta_y(l)$, $\forall k \in [1,p], l \in [p+1, 2p]$.

Proposition 3: Let Z be a standard normal random variable with the mean $\mu(KL_\theta^\lambda)$. Then,

$$\lim_{n\to\infty} Pr \left[\frac{KL_\theta^\lambda - \mu(KL_\theta^\lambda)}{\sqrt{\frac{1}{n}T_1\Sigma T_1'}} \leq \alpha \right] = Pr(Z \leq \alpha) \ . \qquad (15)$$

It has to be noted that here KL_θ^λ in Proposition 3 is a scalar, while it is easy to extend to the case that KL_θ^λ is a vector.

III. Shrinkage FINE Embedding Algorithms

The sFINE embedding algorithm is defined as follows:

1) Using the training dataset, we learn an optimal codebook using the Lloyd-max procedure [10]. Using the codebook, we estimate the feature distribution $P_i, i = 1, 2, \dots, N$ of each quantization level of data in the testing dataset. Given the PDFs $P = \{p_1, p_2, \dots, p_N\}$, we calculate the matrix of **shrinkage** Fisher information distances D using regularized KL-divergence $\hat{D}_F^\lambda(i,j) = \sqrt{KL^\lambda(X \parallel Y) + KL^\lambda(Y \parallel X)}$.
2) Apply Laplacian Eigenmaps [4] on the distance matrix for dimensionality reduction to embed the distance matrix into lower dimensional space.
    - Construct adjacency graph: Given dissimilarity matrix D_x between pairs of PDFs in the set X, define the graph G over all the data points by adding an edge between points i and j if X_i is one of the k-nearest neighbors of X_j.
    - Compute the weight matrix W, if points i and j are connected, assign $W_{ij} = \exp^{-\frac{\hat{D}_F^\lambda(i,j)^2}{t}}$, otherwise $W_{ij} = 0$, where here t is the time in the heat kernel.
    - Construct low-dimensional embedding: Solve the generalized eigenvalue problem $Lf = \lambda Df$, where D is the diagonal weight matrix in which $D_{ii} = \sum_j W_{ji}$, and $L = D - W$ is the Laplacian matrix. If $[f_1, f_2, \dots, f_d]$ is the collection of the eigenvectors associated with d-dimensional embedding is defined by $y_i = (v_{i1}, \dots, v_{id})^T$, $1 \leq i \leq n$.

IV. Experimental Results

In this section, we evaluate the performance of embedding and clustering using shrinkage Fisher information using two experiments. The first experiment involves a very high-dimensional data set on spam harvesters and servers obtained through Project Honey Pot. We investigate the determining factors of shrinkage FINE including optimal shrinkage parameter in Fig.1. The clustering and embedding is compared to the ordinary FINE methods, which are not regularized, in addition to a previous study on this data set [14] (Fig.3 and Table I). The statistical reliability and temporal invariance for identifying strong patterns of the proposed method is also demonstrated (Fig.2). The second experiment involves several standard data sets from the UCI Machine Learning Repository [7]. We compare the performance of sFINE relative to that of several competing methods (Table II).

A. Project Honey Pot

Project Honey Pot[1] is an ongoing project targeted at identifying spammers. It consists of a distributed network of decoy web pages with trap email addresses, which are collected by automated email address harvesters. Both the decoy web pages and the email addresses are monitored, providing us with information about the harvester and email server used for each spam email received at a trap address. There have been two notable studies of the Project Honey Pot data. [11] found that harvesting is typically done in a centralized manner. [14] clustered harvesters using spectral clustering with correlation as the similarity measure. It was discovered that clusters divided into two types: phishing clusters, which consisted almost entirely of harvesters associated with phishing emails

[1] Additional information on Project Honey Pot is available at the website `http://www.projecthoneypot.org`.

(phishers), and non-phishing clusters, which consisted almost entirely of harvesters associated with non-phishing emails (non-phishers). Using phisher and non-phisher as labels for the harvesters, [14] computed the Rand index between the clustering results and labels and found very good agreement.

1) Experiment set-up: We set up this experiment in the same manner as [14]. Two types of attributes are used: email servers and timestamps. For the email servers attribute, we consider the $M \times N$ coincidence matrix $H = [h_{ij}]_{i,j=1}^{M,N}$, where M is the number of harvesters and N is the number of servers. We choose $h_{ij} = p_{ij}/(d_j e_i) \in [0,1]$ where p_{ij} denotes the number of emails sent using harvester i and server j, d_j denotes the total number of emails sent (by all harvesters) through server j, and e_i denotes the total number of email addresses harvester i has acquired. We can interpret h_{ij} as harvester i's of usage of spam server j per address it has acquired. For the timestamps attribute, we examine the timestamps of all emails associated with a particular harvester and bin them into 1-day intervals, which results in a vector indicating how many emails a harvester is associated with in each interval. Doing this for all of the harvesters, we get another coincidence matrix H but with the columns representing time rather than servers. The entries of H are $h_{ij} = s_{ij}/e_i$ where s_{ij} denotes the number of emails associated with harvester i in the jth time interval, and e_i is defined as before.

Given the coincidence matrices, we calculate both the marginal PDFs of the attributes and the joint PDFs with histogram estimation. We use the data from October 2006, which was highlighted in [14] as a month of interest. During this month, there were $2,627$ harvesters and about 2.7×10^5 servers active. We perform embedding and clustering using sFINE on both the marginal and joint PDFs. We compare clustering performance in terms of the Rand index using phisher and non-phisher as labels to that of [14] and to that of the ordinary FINE without shrinkage regularization. In addition, we further explore the individual clusters and discover several interesting groups of harvesters that also appear close together in the embedding.

2) Performance Evaluation: We first validate the proposed approach by evaluation of the determining factors in shrinkage FINE embedding. In Fig.1, we demonstrate the choice of the optimal shrinkage parameter can be translated into better clustering performance using email server attribute for spam data. In Fig.1 (a), we show the sFINE embedding using two steps including Laplacian Eigenmaps and K-means and we obtain the classification error 4.7%. Fig.1 (b) illustrates the classification performance of unregularized FINE. Fig.1 (c) demonstrates the classification performance by first using the principal component analysis (PCA) to reduce the data into 2 dimensional space and then implement K-means clustering, which gives the classification error 36.5%. As shown in the bounding box in Fig.1 (b), the unregularized KL divergence estimators are very close to zero and therefore misclassify the data. While the optimal shrinkage estimator smooths these data and successfully discriminate them with optimal shrinkage terms shown in Fig.1 (a). The poor performance of the unregularized KL estimator is due to the fact that without optimal shrinkage, the number of elements in most of the bins is zero due to insufficient samples, which severely underestimates the entropy and cross entropy. The superiority of Fig.1 (a) over Fig.1 (c) can be mainly attributed to the nonlinear dimensionality reduction with Laplacian Eigenmaps in sFINE since there is no straightforward representation in the Euclidean distance between the email server attributes.

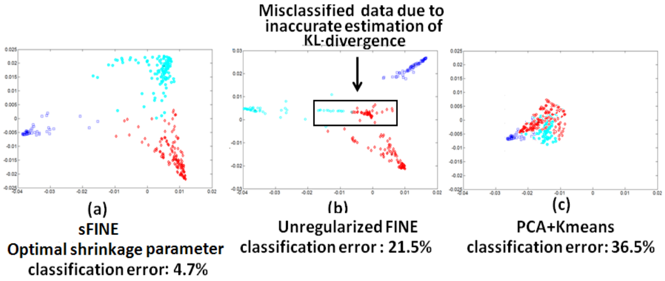


Fig. 1. The comparison of clustering performances with optimal shrinkage parameter in sFINE, unregularized FINE and PCA+Kmeans. (a) shows the shrinkage FINE embedding with optimal shrinkage parameter with classification error 4.7%, (b) shows unregularized FINE with classification error 21.5%, (c) shows the performance of first using PCA to reduce the data into 2 dimensional space and then implementing K-means clustering with classification error 36.5%.

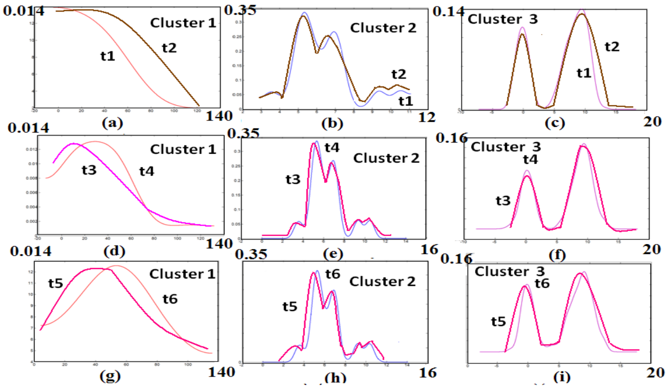
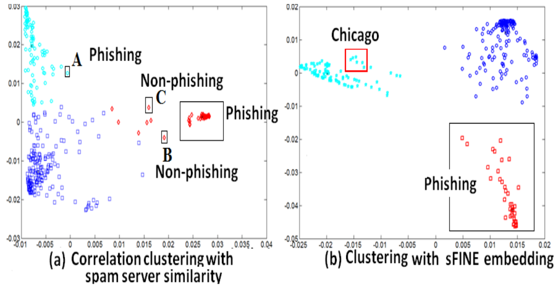


Fig. 2. The visual illustration of kernel density estimator for revealing the patterns in different clusters using sFINE for temporal behaviors with the window size $\tau_1 = 25$ days where the shift of time interval is 2 days.

To demonstrate the statistical reliability and time invariance of our shrinkage FINE embedding, we analyze the patterns contained in different clusters for spam data. By applying kernel density estimator over the window size $\tau_1 = 25$ days in each cluster in Fig.2, we find that that the PDF of cluster 1 is smooth with long tails; the PDF of cluster 2 is more concentrated; the PDF of cluster 3 is characterized by two sharp peaks. The similar performance is shown in each time interval in Fig.2. These results demonstrate that shrinkage FINE is a robust reliable statistical measure with time invariance.

Comparison: We compare the performance for sFINE with correlation-based clustering [14]. Fig.3(a) indicates that using correlation-based clustering [14] non-phishing harvesters B and C are incorrectly classified as phishing harvesters. While



(a) Correlation clustering with spam server similarity

(b) Clustering with sFINE embedding

Fig. 3. (a) shrinkage FINE embedding with both of two similarity by estimation of joint PDF for harvester, spam server and time in Oct 2006, (b) correlation-based clustering with spam server similarity

shown in Fig.3(b) FINE embedding is able to achieve perfect results by completely separating the phishing emails. The clustered harvesters with FINE embedding is labeled by phishing levels, where the harvesters corresponding to phishers are highlighted with the bounding box. It can be mainly attributed to the ability of our embedding to incorporate both spam server and temporal similarities to be more discriminative in clustering. The qualitative results demonstrated in Table I indicates the 8% on average improvement in accuracy for the proposed FINE embedding compared to correlation-based clustering [14]. Moreover, the clusterings with shrinkage methods further improves the accuracy by 3% on average as compared to unregularized methods.

TABLE I

COMPARISON OF VALIDATION INDICES USING CORRELATION, FINE AND SHRINKAGE FINE (SFINE) EMBEDDING USING SPAM SERVER CLUSTERING (SS), SPAM SERVER AND TEMPORAL CLUSTERING (ST) EVALUATED BY RAND INDEX (RI) AND ADJUST RAND INDEX (ARI) IN OCT 2006.

algorithm	RI-SS	ARI-SS	RI-ST	ARI-ST
correlation	0.92	0.87	-	-
FINE	0.94	0.89	0.95	0.93
sFINE	**0.96**	**0.92**	**0.97**	**0.94**

B. UCI Machine Learning Repository data

Next we compare the performance of shrinkage FINE with several state-of-the-art approaches. We select eight standard data sets from the UCI Machine Learning Repository [7] and compare the clustering performance of sFINE against three competing methods. The first is an information theoretic clustering algorithm [6] where non-parametric estimation of cluster entropies is utilized for clustering by maximizing the estimated mutual information between data points and clusters. Secondly, [13] proposed a sparse K-means algorithm by maximizing a weighted between-cluster sum of squares subject to L1 type constraints on the weights. L1 regularization can be interpreted as a form of regularization using hard thresholding instead of shrinkage. Finally, [15] presented unsupervised SVM training by formulating convex relaxations of the natural training criterion: find a labeling that would yield an optimal SVM classifier on the resulting training data.

1) Experiment set-up: We randomly selected 50% of a data set as testing dataset, the other 50% of the data as training dataset and cross-validation is carried on. In the training dataset, we learn the optimal quantization levels and ranges with Lloyd-Max quantization. In the testing dataset, we estimate frequency parameters of the multinomial distribution of the data using the quantization levels and ranges learned from the training dataset with maximum likelihood. We determine the clustering accuracy relying on shrinkage Fisher information embedding with the optimal shrinkage parameters we derived. The clustering performance was evaluated using the Rand index, which is a standard criterion for evaluating clustering accuracy. Since sparse K-means is sensitive to its intrinsic parameters, the sparse K-means clustering results are reported using the best tuning parameters.

2) Performance Evaluation: The results reported in Table II demonstrate that sFINE provides the best performance, with improvements in the Rand index by approximately 9% compared to unsupervised SVM, 7% compared to sparse K-means, and 4% compared to non-parametric information theoretic clustering. We compare the clustering performances by first applying principal component analysis (PCA) to reduce the data into 2 dimensional space before running the other algorithms. It can be mainly attributed to the fact that the sFINE embedding has provided a better similarity measure approximated by shrinkage regularized KL-divergence which leads to better clustering results. Compared to sFINE, non-parametric information theoretic clustering and unsupervised SVM do not take into account the high dimensionality and small number of samples and do not fully utilize the information of the distribution of the data. The sparse K-means algorithm can be viewed as hard thresholding instead of shrinkage and is sensitive to tuning parameters.

TABLE II

COMPARISON OF CLUSTERING ACCURACY (RAND INDEX SCORE) FOR OUR SHRINKAGE FINE (SFINE) TO OTHER ALGORITHMS INCLUDING WHERE WE APPLY PRINCIPAL COMPONENT ANALYSIS (PCA) TO REDUCE THE DATA INTO 2 DIMENSIONS PRIOR TO RUNNING OTHER ALGORITHMS.

Attributes	SVM	sK-means	NIC	standard K-means	**sFINE**
wine	87.1%	90.6%	91.9%	72.5%	**94.3%**
statlog	82.6%	88.7%	89.5%	77.3%	**92.7%**
segmentation	86.9%	88.3%	89.0%	68.9%	**92.1%**
vowel	83.3%	85.9%	87.2%	75.6%	**89.3%**
iris	85.5%	86.4%	90.3%	70.8%	**93.6%**
abalone	59.6%	60.8%	67.8%	55.4%	**68.2%**
balance	57.8%	61.8%	68.1%	56.1%	**70.2%**

ACKNOWLEDGEMENT

This research was partially supported by Army Research Office (ARO) grant number W911NF-09-1-0310. The authors also would like to thank Kevin Xu for his valuable suggestions for the paper.

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] S. Amari. Differential-geometrical methods in statistics. *Lecture notes in statistics*, 1985.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research (JMLR)*, 6, 2005.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[5] K. Carter, R. Raich, W. G.Finn, and A. O.Hero. Fisher information nonparametric embedding. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31:2093–2098, 1993.

[6] L. Faivishevsky and J. Goldberger. A nonparametric information theoretic clustering algorithm. *International Conference on Machine Learning (ICML)*, 2010.

[7] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[8] J. Hausser and K. Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research (JMLR)*, 10:1469–1484, 2009.

[9] R. Kass and P. Vos. Geometrical foundations of asymptotic inference. *Wiley Series in Probability and Statistics. John Wiley and Sons*, 1997.

[10] S. Lloyd. Least squares quantization in pcm. In *IEEE Transactions on Information Theory*, volume 28, pages 129–137, 1982.

[11] M. Prince, B. Dahl, L. Holloway, A. Keller, and E. Langheinrich. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from Project Honey Pot. In *Proc. 2nd Conference on Email and Anti-Spam*, 2005.

[12] P. Williams. Bayesian regularization and pruning using a laplace prior. In *Neural Computation*, volume 7, pages 117–143, 1995.

[13] D. Witten and R. Tibshrani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 15:713–726, 2010.

[14] K. S. Xu, M. Kliger, Y. Chen, P. Woolf, and A. O.Hero. Social networks of spammers through spectral clustering. *IEEE Intl. Conf. on Communications (ICC)*, 2009.

[15] L. Xu and D. Schuurmans. Unsupervised and semi-supervised multiclass support vector machines. In *National Conference on Artificial Intelligence (AAAI)*, 2005.