Neurocomputing 73 (2010) 1840–1852

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Principal component analysis based on non-parametric maximum entropy

Ran He [a,*], Baogang Hu [b], XiaoTong Yuan [c], Wei-Shi Zheng [d]

[a] School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, People's Republic of China
[b] National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[c] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[d] Department of Computer Science, Queen Mary University of London, London, UK

ARTICLE INFO

Available online 12 March 2010

Keywords:
PCA
Entropy
Subspace learning
Information theoretic learning

ABSTRACT

In this paper, we propose an improved principal component analysis based on maximum entropy (MaxEnt) preservation, called MaxEnt-PCA, which is derived from a Parzen window estimation of Renyi's quadratic entropy. Instead of minimizing the reconstruction error either based on L_2-norm or L_1-norm, the MaxEnt-PCA attempts to preserve as much as possible the uncertainty information of the data measured by entropy. The optimal solution of MaxEnt-PCA consists of the eigenvectors of a Laplacian probability matrix corresponding to the MaxEnt distribution. MaxEnt-PCA (1) is rotation invariant, (2) is free from any distribution assumption, and (3) is robust to outliers. Extensive experiments on real-world datasets demonstrate the effectiveness of the proposed linear method as compared to other related robust PCA methods.

1. Introduction

Subspace learning has been a fundamental problem in the study of machine learning and computer vision. It is a common preprocessing step to find a low-dimensional data representation from the raw input samples which may be strongly relevant and redundant [1]. It plays an important role in many learning tasks due to the curse of dimensionality [2].

From different standpoints, there are two major categories of subspace learning: supervised and unsupervised. In supervised subspace learning, class labels are known and incorporated to learn a discriminant low-dimensional representation [3]. The linear discriminant analysis (LDA) [4] is the most representative of these discriminators. In unsupervised subspace learning, the labels of data are unknown. The low-dimensional representation is formulated by minimizing the reconstruction error or by preserving the local relationships on a local patch [5–7]. The best-known representative of unsupervised methods is the principal component analysis (PCA) [8–10]. In this paper,we focus on the unsupervised learning.

In subspace learning, PCA is a linear data transformation technique which is often used as a data pre-processing step of other subspace learning methods. Many commonly used discriminators like LDA, Locality Preserving Projections (LPP) [5,11], and marginal Fisher analysis (MFA) [12] are typically performed on the principal component space produced by PCA. However, PCA also has some limitations. First, PCA is sensitive to outliers, which means that outliers may significantly change the principal subspaces [13–15]. Second, PCA is intrinsically based on the Gaussian distribution (i.e., with a single maximum) and thus is unable to derive a good approximation from multimodal distributions [16].

In order to alleviate above the two problems of PCA, many investigations have been reported. One important approach is to use more robust metric rather Euclidean distance to measure the reconstruction error in PCA. Typically, the L_1-norm has widely discussed and used in PCA for years. In [17,18], L_1-norm PCA was formulated by applying a maximum likelihood estimation to the original data. And a heuristic estimation method and convex programming methods were developed to detect outliers in [17,18], respectively. A major drawback of these work is that they are not rotationally invariant. To overcome this drawback, two rotationally invariant PCA methods have been recently developed [19,20] by relaxing the objective function of L_1-norm. R1-PCA in [19] weights each sample using Huber's M-estimator and removes the outliers by an iteration algorithm. PCA-L_1 in [20] provided a greedy algorithm to solve a simplified L_1-norm objective function. TPCA-L_1 in [21] further generalized PCA-L_1 to robust tensor analysis. However, those methods assume that the data mean is fixed, for example zero mean. When outliers occur, the data mean will be biased. Other important variants of PCA include robust PCA [22,23], Locally PCA [24–26], manifold learning based PCA [27,28], and generalized PCA [29].

Another important approach to develop new PCA is based on information theoretic learning (ITL) [30,31]. It has been shown that PCA can be formulated as a maximum entropy (MaxEnt)

* Corresponding author.
E-mail address: rhe1979@gmail.com (R. He).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.12.032

problem under Gaussian distribution assumption. An earlier work of unsupervised MaxEnt was EMMA based component analysis [32], where entropy and density are estimated iteratively on two data sets. But when dimensionality of data is high, there will be a large number of parameters in EMMA have to be estimated. In [33], the connection between Renyi's entropy and robust function is discussed. Recently, the close connection between the kernel methods and ITL has been discussed [34–36]. In [37], the kernel PCA has been proven to be equal to a MaxEnt problem. In [38], a new kernel-based nonlinear subspace technique is proposed based on MaxEnt preservation.

In this paper, the maximum entropy (MaxEnt) criterion, which provides a natural way to process information with constraints, is introduced to produce a robust linear subspace. A linear subspace technique, called MaxEnt-PCA, is proposed based on the Renyi's quadratic entropy estimated via a non-parametric Parzen window. A gradient based fixed-point algorithm is proposed to solve the MaxEnt problem. From the entropy point of view, MaxEnt-PCA is a natural extension of PCA and has several appealing advantages: (1) it has a solid theoretical foundation based on the concept of MaxEnt; (2) it is rotation invariant and its solution consists of eigenvectors of a Laplacian probability matrix corresponding to the MaxEnt distribution; (3) it makes use of high order statistics to estimate the energy matrix and is robust to outliers; and (4) it is free from any distribution assumption and thus it can effectively capture the underlying distribution of multimodal data statistics.

The rest of this work is organized as follows. We start our work with a brief review of PCA and its extensions in Section 2. In Section 3, a fixed-point algorithm is proposed to solve the MaxEnt-PCA, followed by the theoretical analysis. In Section 4, we evaluate our method in real-world datasets. Finally, we conclude the paper in Section 5.

2. PCA and related work

Consider a data set of samples $X=[x_1,\ldots,x_n]$ where x_i is a variable with dimensionality d, $U=[u_1,\ldots,u_m]\in R^{d\times m}$ is a projection matrix whose columns constitute the bases of an m-dimensional subspace, and $V=[v_1,\ldots,v_n]\in R^{m\times n}$ is the projection coordinates under the projection matrix U.

PCA can be defined as an orthogonal projection of the samples onto a lower dimensional subspace such that the variance of the projected data is maximized [8]. Equivalently, it can also be defined as an orthogonal projection that minimizes the average reconstruction error, which is the mean squared distance between the samples and their projections [39].

From reconstruction error point of view, PCA can be formulated as the following optimization problem:

$$\min_{U,V}\sum_{i=1}^{n}\|x_i-(\mu+Uv_i)\|^2 \tag{1}$$

where μ is the center of X. The optimization problem in (1) can also be written below

$$\min_{U,V}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{ij}-\left(\mu_{ij}+\sum_{p=1}^{m}v_{ip}u_{pj}\right)\right)^2 \tag{2}$$

By projection theorem [40], for a fixed U, V that minimizes (1) is uniquely determined by $V=U^TX$. Because (1) is based on L_2-norm (Euclidean distance), the PCA is often denoted as L_2-PCA. In order to develop a fast and robust subspace algorithm, the expectation maximization (EM) algorithm [41,42] and fixed-point algorithm are developed to solve (2).

The global minimum of (1) is provided by singular value decomposition (SVD) [43], whose optimal solution is also the solution of the following alternative formulation of PCA:

$$\max_{U^TU=I}\mathrm{Tr}(U^T\Sigma U) \tag{3}$$

where $\Sigma=\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$ is the covariance matrix , $\mathrm{Tr}(\cdot)$ is the matrix trace operation, and T denotes the transpose. The (3) searches for a projection matrix where the variances of U^TX are maximized. Based on (3), PCA can also be further unified in the patch alignment framework [44].

In graph embedding (GE) framework [12], PCA can also be formulated as the following optimization problem:

$$\max_{U^TU=I}\mathrm{Tr}\left(U^TX(I-W)X^TU\right) \tag{4}$$

where I is the identity matrix, W is a $n\times n$ matrix whose elements are all equal to $1/n$. The solutions of PCA can be obtained by solving the following eigenvalue decomposition problem:

$$X(I-W)X^Tu=\lambda u \tag{5}$$

In graph embedding, the matrix $L=I-W$ is often denoted as Laplacian matrix.

Since L_2-norm based PCA is sensitive to outliers [20], L_1-norm was used to substitute L_2-norm. From a statistical point of view, those methods based on L_1-norm are more robust to outliers than L_2-norm based ones [45,18,20]. In this case, the problem of PCA becomes finding the U that minimizes the following reconstruction error function:

$$\min_{U,V}\sum_{i=1}^{n}\sum_{j=1}^{d}\left\|x_{ij}-\left(\mu_{ij}+\sum_{p=1}^{m}v_{ip}u_{pj}\right)\right\|_{L_1} \tag{6}$$

Since the PCA based on L_1-norm is not invariant to rotations of coordinates, two rotationally invariant PCA methods are proposed [19,20].

Instead of using L_1-norm, weighted PCA is also developed. The temporal weighted PCA [46] produces a robust subspace by putting different weights on each x_i in X. The temporal weighted PCA tries to minimize the following weighted squared reconstruction error:

$$\min_{U,V}\sum_{i=1}^{n}w_i\|x_i-(\mu+Uv_i)\|^2 \tag{7}$$

where w_i is a given weight on x_i. The spatial weighted PCA [46] produces a robust subspace by putting different weights on each entry of x_i. It tries to solve the following optimization problem:

$$\min_{U,V}\sum_{i=1}^{n}\sum_{j=1}^{d}w_{ij}\left\|x_{ij}-\left(\mu_{ij}+\sum_{p=1}^{m}v_{ip}u_{pj}\right)\right\|^2 \tag{8}$$

where w_{ij} is the weight of the j th entry in the x_i. The EM algorithm [46] and Iteratively Reweighted Least Squares (IRLS) [22] were used to solve (8).

Table 1 summarizes PCA and its several major variations of PCA. The second column shows the objective of each method. The "R" indicates minimizing reconstruction error and the "S" indicates maximizing scatter matrix. The "A" means to solve an approximation problem. According to (15), MaxEnt-PCA aims to maximize a robust scatter matrix. The third column shows the distribution assumptions of some methods, while it is difficult to tell the distribution assumption of most PCA's extensions. Different distribution assumption will lead to a different subspace (see Fig. 4 for detail). In the fourth column, the "Y" indicates that the algorithm makes use of a given data mean or assume data is zero mean and the "N" indicates that the algorithm can calculate data mean by itself or they do not need to calculate data center. In the fifth column, the "Y" indicates that the algorithm is rotation invariant and the "N" indicates that the algorithm is not. In the sixth column, the "Y" indicates that the algorithm needs additional

Table 1
PCA and major variations of PCA (see text for details).

	Objective(s)	Distribution estimation	Zero mean	Rotational invariant	Additional weight	Robust oriented	Linear subspace	Entropy
PCA [47]	R & S	Gaussian	Y	Y	N	N	Y	MaxEnt
KPCA [48]	S	–	Y	Y	N	–	N	MaxEnt
kernel MaxEnt [38]	A	–	N	Y	N	–	N	MaxEnt
Local PCA [24]	S	–	N	Y	N	N	Y	–
TWPCA [46]	R & S	Gaussian	Y	Y	Y	Y	Y	–
R1-PCA [19]	R	Gaussian	Y	Y	N	Y	Y	–
WPCA [46]	R	–	N	N	Y	Y	Y	–
RPCA [22]	R	–	N	N	N	Y	Y	–
L_1 PCA [18]	R	–	N	N	N	Y	Y	–
PCA L_1 [20]	S	–	Y	Y	N	Y	Y	–
EMMA [32]	A	A	N	Y	N	Y	Y	MaxEnt
MaxEnt-PCA	S(Renyi)	Parzen	N	Y	N	Y	Y	MaxEnt

"R" represents minimizing reconstruction error; "S" represents maximizing scatter matrix; "A" represents "approximation"; "-" represents "unknown" or "unavailable"; "Y" represents "yes"; and "N" represents "no".

weights as input parameters and the "N" indicates that the algorithm does not. In the seventh column, the "Y" indicates that the algorithm is robust to outliers and the "N" indicates that the algorithms are not. In the eighth column, the "Y" means a linear subspace and the "N" means a nonlinear subspace. The last column shows whether the algorithm can be formulated as a problem of entropy maximization.

Although shown some more robust metric such as L_1-norm rather than the L_2-norm has been used to improve PCA, they still cannot explore high order statistical information about the difference between input variables. In this paper, we then develop a robust high-order PCA algorithm in terms of MaxEnt.

3. MaxEnt-PCA

3.1. Objective function

The aim of MaxEnt-PCA is to learn a new data distribution in a subspace such that entropy is maximized. Here we consider the Renyi's quadratic entropy of a random variable X with probability density function (P.D.F.) $f_X(x)$ defined by

$$H(X) = -\log \int f_X^2(x)\,dx \tag{9}$$

If $f_X(x)$ is a Gaussian distribution, the estimate of Renyi's quadratic entropy is obtained by [31,49]:

$$H(X) = \frac{1}{2}\log(|\Sigma|) + \frac{d}{2}\log 2\pi + \frac{d}{2} \tag{10}$$

where $|\cdot|$ is the absolute value of determinant [49, p. 254]. If Parzen window method is used to estimate the P.D.F., $f_X(x)$ can be obtained by

$$\widehat{f}_{X;\sigma}(x) = \frac{1}{n}\sum_{i=1}^{n} G(x-x_i,\sigma^2) \tag{11}$$

where $G(x-x_i,\sigma^2)$ is the Gaussian kernel with bandwidth $\Sigma = \sigma^2 I$

$$\begin{aligned} G(x-x_i,\sigma^2) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-x_i)^T\Sigma^{-1}(x-x_i)\right) \\ &= \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \end{aligned} \tag{12}$$

By substituting $f_X(x)$ in (9) with (11), the estimate of entropy by Parzen window method can be formulated as [50]:

$$H(X) = -\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_j-x_i,\sigma^2)\right) \tag{13}$$

In unsupervised subspace learning, one considers the following constraint MaxEnt problem:

$$\max_U H(U^TX) \quad \text{s.t. } U^TU = I \tag{14}$$

Note that the orthonormal constraint is necessary and important for extracting non-reductant features. When the formula of $f_X(x)$ is given, the MaxEnt distribution about $f_{U^TX}(x)$ in (14) is only relative to the subspace U. When Parzen window density estimation of entropy in (13) is adopted, the optimization problem in (14) becomes

$$\max_U \quad \left(-\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(U^Tx_j-U^Tx_i,\sigma^2)\right)\right)$$

$$\text{s.t.} \quad U^TU = I \tag{15}$$

We denote the above method as MaxEnt-PCA. From the entropy point of view, MaxEnt-PCA is a natural extension of PCA from Gaussian distribution assumption to Parzen window density estimation. Obviously the superiority of MaxEnt-PCA lies in the non-parametric density estimation from training data set, which can be more flexible and robust.

Furthermore, it is obvious that (15) is a robust *M-estimator* [51] formulation of scatter matrix in (4) with robust function $r(x) = \exp(-x^2)$ [51]. The $r(x)$ belongs to the so called redescending M-estimators, which have some special robustness properties in theory [52]. Therefore MaxEnt-PCA can be viewed as a robust extension of the classical PCA.

3.2. Algorithm of MaxEnt-PCA

Proposition 1. *The optimal solution of MaxEnt-PCA in* (15) *is given by the eigenvectors of the following generalized eigen-decomposition problem*:

$$XL(U)X^TU = 2U\Lambda \tag{16}$$

where

$$L(U) = D(U) - W(U) \tag{17}$$

$$W_{ij}(U) = \frac{G(U^Tx_i-U^Tx_j,\sigma^2)}{\sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n} G(U^Tx_i-U^Tx_j,\sigma^2)} \tag{18}$$

$$D_{ii}(U) = \sum_{j=1}^{n} W_{ij}(U) \tag{19}$$

This can be figured out by applying the Lagrangian factor on (15), where entries of Λ are the Lagrangian coefficients associated

to the orthonormal constraint on U as follows:

$$J_H \triangleq -\log\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(U^T x_i - U^T x_j, \sigma^2) - \mathrm{Tr}(\Lambda(U^T U - I))$$

where $\mathrm{Tr}(\cdot)$ is the matrix trace operation. The KKT condition for optimal solution specifies that the derivative of J_H with respect to U must be zero:

$$\frac{\partial J_H}{\partial U} = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(U)(x_i - x_j)\left(x_i^T - x_j^T\right)U - 2U\Lambda = 0$$

Then we have

$$XL(U)X^T U = 2U\Lambda \tag{20}$$

Intuitively, an optimal U is the eigenvectors of the symmetric matrix $XL(U)X^T$ and the Lagrangian multipliers Λ then becomes a diagonal matrix: $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$.

In Proposition 1, $W_{ij}(U)$ is an approximation of probability distribution on x_i under the j th Parzen estimate, and the D_{ii} (U) is an approximation of probability value on x_i under the Parzen estimate. We follow the notation of graph embedding, and denote $W(U)$ and $L(U)$ as Parzen probability matrix and Laplacian probability matrix, respectively. Compared to (5), MaxEnt-PCA is actually a special weighted PCA. However, PCA is based on Gaussian assumption,while MaxEnt-PCA is derived from Parzen estimation.

Since $L(U)$ in (16) is also a function of U, the eigenvalue decomposition problem in (16) has no closed-form solution. Fortunately, we can solve this MaxEnt problem by gradient-based fixed-point algorithm [53,19,54,55] which is often used in subspace learning. As a result, we use the following steps to update the projection matrix U.

$$U = (I + \beta XL(U)X^T)U \tag{21}$$

$$U = svd(U) \tag{22}$$

where β is a step length to ensure an increment of the objective function, and $svd(U)$ returns an orthonormal base by the Singular Value Decomposition (SVD) on matrix U. In (21), the U is updated by the gradient direction. In (22), an orthonormal solution of U is obtained. The convergence of the fixed-point algorithm is actually guaranteed by [19,43].

The fixed-point algorithm of MaxEnt-PCA is outlined in Algorithm 1. The step length β can be decided by the line search method [56]. Note that the estimate of $f_X(x)$ is performed on the reduced dimension instead of original input feature space. The bandwidth σ is an important parameter in MaxEnt-PCA, which is used in Parzen estimate of $f_X(x)$. Considering the theoretical analysis of non-parametric entropy estimators [38,32], we present a tunable way to set the bandwidth as a factor of average distance between projected samples:

$$\sigma^2 = \frac{1}{sn^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|U^T x_i - U^T x_j\|^2 \tag{23}$$

where s is a scale factor. The bandwidth σ is also a function of subspace U. In each update, the bandwidth σ is also updated on the projection dataset $U^T X$.

Algorithm 1. MaxEnt-PCA

Input: data matrix X, random orthonormal matrix U and a small positive value ε
Output: orthonormal matrix U

```
repeat
    Initialize converged=FALSE.
    Calculate σ according to (23), and L(U) according to (17)
    Select a suitable β, and update U according to (21) and (22)
    if the entropy delta is smaller than ε then
        converged=TRUE
    end if
until converged==TRUE
```

During each update in MaxEnt-PCA, the probability distribution is estimated by Parzen method. Since an outlier is far away from the data cluster, its contribution to estimation of the probability density function will be smaller so that it always receives a low value in the Parzen probability matrix W. Therefore, the outliers will have weaker influence on the estimation of the MaxEnt probability distribution as entropy increases. Hence, MaxEnt-PCA is robust against outliers.

Fig. 1 illustrates examples of principal direction produced by PCA and MaxEnt-PCA. In Figs. 1(a) and (b), we see the instability of PCA and the robustness of MaxEnt-PCA to outliers. When the data is drawn from Gaussian, the principal directions of PCA and MaxEnt-PCA overlap each other; when an outlier occurs, MaxEnt-PCA can still produce a robust principal direction. Fig. 1(c) further show an example of a bimodal Gaussian distribution where a small set of outliers exist. Group 1 is normally distributed with mean (0,0) and covariance matrix=diag(5,1), whereas the second group has mean (0,10) and covariance matrix=diag(1,3). The number of points in group 2 is 10% of that in group 1. The second group is used to simulate the outliers. The MaxEnt-PCA still find a robust principal direction.

3.3. Convergence analysis

In this subsection, we further discuss the convergence of MaxEnt-PCA in terms of (21) and (22). We begin the analysis with two theoretical properties of differential entropy.

Proposition 2. *The differential entropy is invariant to orthonormal linear transformations, i.e.,*

$$H(U^T X) = H(X) \tag{24}$$

where $U \in R^{d\times d}$ *and* $U^T U = I$

Proof. According to the properties of differential entropy [49], we have

$$H(U^T X) = H(X) + \log(|U^T|) \tag{25}$$

Because U is an orthonormal matrix, we can obtain $\log(|U^T|) = 0$, hence

$$H(U^T X) = H(X) + \log(|U^T|) = H(X) \qquad \square \tag{26}$$

We can easily prove that the objective in (15) is also rotationally invariant (i.e., $G(U^T x_i - U^T x_j, \sigma^2) = G(x_i - x_j, \sigma^2), U^T U = I$). Rotational invariance is a fundamental property of Euclidean space with L_2-norm and has been emphasized by [57]. For any orthonormal rotation, data transformation U is invariant under L_2-norm, i.e., $\|U^T x\|_2 = \|x\|_2$. Proposition 2 illustrates that the maximum entropy objective is also invariant to rotation. It is independent of the selection of a coordinate system for subspace learning.

Proposition 3. *The differential entropy is bounded under orthonormal linear transformations, i.e.*

$$0 \le H(U_F^T X) \le H(X) \tag{27}$$

where $U_F^T : R^d \to R^m$, $m < d$ *and* $U_F^T U_F = I$.

Proof. Let $U_B \in R^{d\times(d-m)}$ be a matrix whose columns constitute the complement subspace of U_F, and define a matrix U as

$$U^T X = \begin{bmatrix} U_F^T X & U_B^T X \end{bmatrix}, \quad U = [U_F \quad U_B] \tag{28}$$

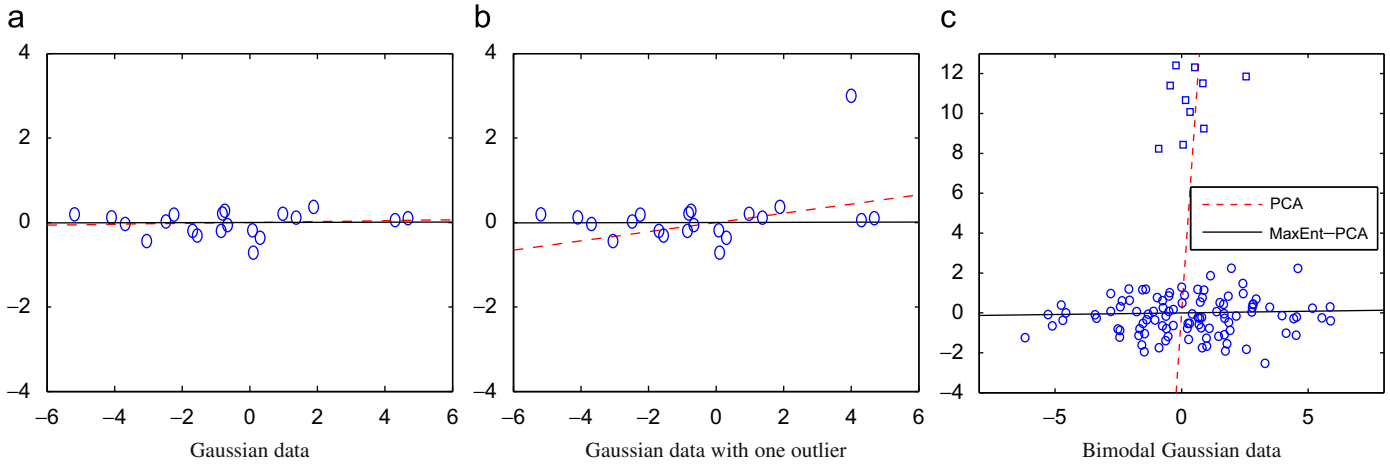


Fig. 1. Principal directions produced by PCA and MaxEnt-PCA. (a) Points are drawn from a Gaussian distribution. (b) Points are drawn from a Gaussian distribution with single outlier. (c) The data contains two groups. Group 1 is normally distributed with mean (0,0) and covariance matrix=diag(5,1), whereas the second group has mean (0,10) and covariance matrix=diag(1,3). The number of points in group 2 is 10% of that in group 1.

Since U is a $d \times d$ orthonormal matrix, it follows from Proposition 2 that $H(X)=H(U^T X)$. Consider that

$$H(X) \geq 0 \tag{29}$$

and the chain rule for entropies ([58, p.22]), we have

$$H(X) = H\left(U_F^T X U_B^T X\right) = H(U_F^T X) + H(U_B^T X | U_F^T X) \geq H(U_F^T X)$$

hence

$$0 \leq H(U_F^T X) \leq H(X) \qquad \square$$

Proposition 3 states that the differential entropy of orthonormal subspace of the original feature space is bounded. We can easily prove that the Renyi entropy in (15) is also bounded ($\|U_F^T x_i - U_F^T x_j\|^2 \leq \|U^T x_i - U^T x_j\|^2$). Since the entropy is a concave function of $f_X(x)$, there is at least one local maximum of (14).

In Algorithm 1, the objective is upper bounded and a new U is produced along gradient ascend direction in each update. Hence Algorithm 1 will increase the value of entropy until it converges. Fig. 2 demonstrates the convergence curves and eigenvalues of MaxEnt-PCA on two UCI datasets. Fig. 2(a) illustrates that MaxEnt-PCA increases the entropy step by step, and Fig. 2(b) shows the diagonal elements of Lagrangian multipliers at convergence. Figs. 2(c) and (d) list the top-left matrix of Λ on an Australian dataset at first iteration and at convergence, respectively (the rest has similar format). We can learn that the Lagrangian multiplier Λ becomes a diagonal matrix at convergence.

The computation of MaxEnt-PCA mainly involves two steps: calculation of a gradient and singular value decomposition. The cost of calculating matrix $XL(U)X^TU$ is $o(2n \times d \times m + n^2 \times m)$ and SVD requires $o(d \times m^2)$. Thus the cost of MaxEnt-PCA for each update requires $o(2n \times d \times m + n^2 \times m + d \times m^2)$. For occlusion problem in Section 4.4, PCA-L1, R1-PCA, and MaxEnt-PCA take 46 s, 59 s and 67 s, respectively. When the number of samples n is large, the complexity of MaxEnt-PCA will be relatively high. Fortunately, there have been several methods in ITL to address this issue. The stochastic gradient algorithm [50] can be used to draw part of the data to estimate the gradient without sacrificing the accuracy. Furthermore, we notice that $L(U)$ is a dense matrix derived from Parzen probability estimate. It is reasonable to assume that the probability at point x can be estimated from its several nearest Gaussian kernels. Then the $L(U)$ can be treated as a sparse matrix to reduce complexity.

4. Experiments

In this section, we applied the proposed MaxEnt-PCA algorithm to several real-world pattern recognition problems and compared it with PCA, spherical PCA [14], R1-PCA [19], PCA-L_1 [20], RoPCA [23],[1] RPCA [22], and local coordinates alignment (LCA) [7]. In all of the experiments, the Cauchy robust function was used for R1-PCA, and the convergence condition for R1-PCA, PCA-L_1 and MaxEnt PCA were set if the difference between the norms of projection matrix U in successive iterations was less than 10^{-5} or the maximum number of iterations (e.g. 50) was reached [19,20]. The Gaussian kernel is used in LCA. The size of nearest neighbors k and kernel parameter σ in LCA are set to 4 and 2, respectively. The scale factor s in (28) of MaxEnt-PCA is set to 2 and the β in (21) is always fixed to 1. All of the experiments were implemented by MATLAB on a P4 2.40 GHz Windows XP machines with 2 GB memory.

4.1. UCI balance scale data set

In this subsection, UCI Balance Scale data set was selected to demonstrate the iterative procedure of MaxEnt-PCA as well as to discuss the relationships and differences between entropy based PCA and the typical L_1-norm based PCA methods.

The Balance Scale data set [59] is a benchmark data set and is frequently used to verify the effectiveness of subspace learning algorithms. It consists of 625 examples in three categories. The numbers of instances in all three categories are 288, 288 and 49, respectively. Each instance has four raw attributes, i.e., Left-Weight, Left-Distance, Right-Weight, and Right-Distance. Table 2 lists a brief summary of this data set. Each dimension of raw data is normalized (with zero mean and standard variance). The data were projected into a two-dimensional subspace for visualization.

Fig. 3(d) plots the scatter of data with the first three dimensions. We see that the data were arranged in 25 clusters. It seems that the red circle instances occupy the top-left corner and the blue cross instances occupy the bottom-right corner. Fig. 3(a) shows a 2D scatter plot under an orthonormal projection matrix. Since this projection matrix was randomly initialized, the data in Fig. 3(a) are out-of-order. Figs. 3(b) and (c) further depict the visualization results of MaxEnt-PCA after the 10th and 30th iteration, respectively. After 10 iterations, the data distribution

[1] http://users.jyu.fi/~samiayr/DM/demot/LIBRA/

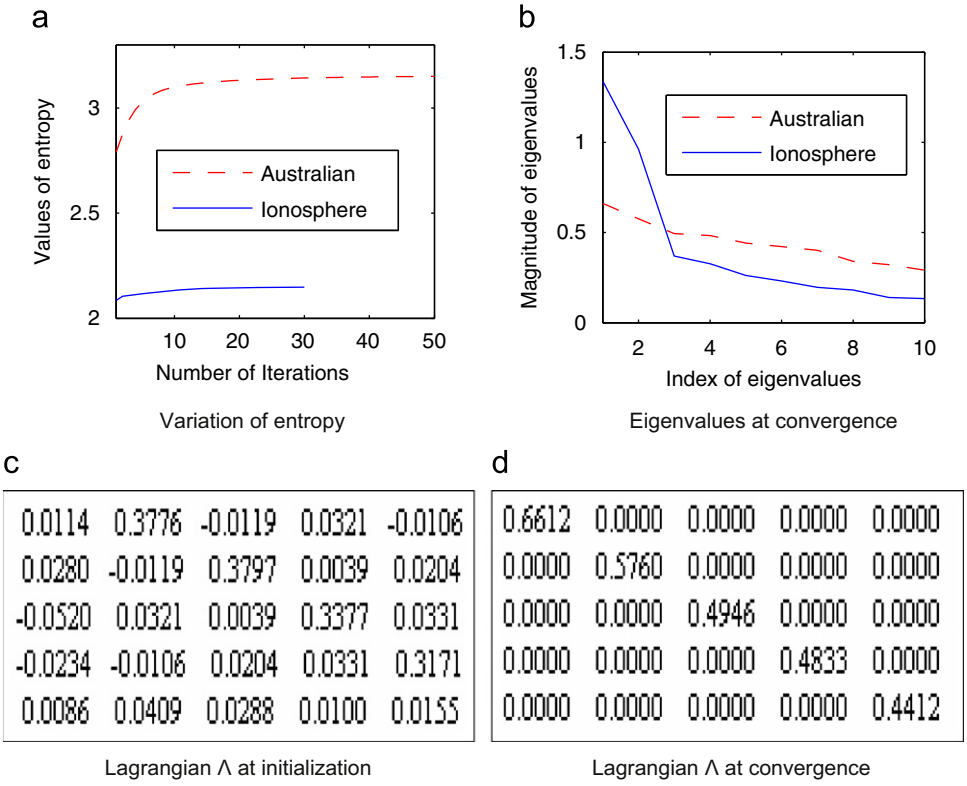


Fig. 2. Convergence of MaxEnt-PCA on UCI datasets where β is fixed at 1 and dimension of subspace is 10. (a) Variation of entropy on Australian and Ionosphere dataset. (b) The eigenvalues at convergence. (c) Top-left matrix of Λ at first iteration on Australian datatset. (d) Top-left matrix of Λ at convergence on Australian datatset.

Table 2
Description of data sets used in the experiments.

Data set	Australian	Balance	Ionosphere	Isolet	Pima	Yeast	FRGC
No. of dimension	14	4	33	617	8	8	1024
No. of classes	2	3	2	26	2	9	186
No. of samples	690	625	351	1559	768	1479	3720

becomed similar to the real data distribution. After 30 iterations, the data clusters in 25 groups and the margins between different groups were maximized. The corresponding variation of entropy is drawn in Fig. 3(e). The entropy increased step by step until the algorithm converged. It is interesting to observe that the entropy of structural data in Fig. 3(c) is larger than that of out-of-order data in Fig. 3(a).

The update process in Fig. 3 not only illustrates the principle of MaxEnt-PCA but also yields an interesting structure of the Balance dataset. If we make use of dash lines to link the black square instances, we see that the dash lines are all nearly on the boundaries between blue cross category and red circle category. It seemed that in the raw four-dimensional space the data were distributed in many clusters and in each cluster one category could separate the other two categories. This example also illustrates that MaxEnt-PCA can be used to find intrinsic structure of high-dimensional data.

Different from PCA, which has a unique global solution, MaxEnt-PCA and PCA-L_1 may in practice always learn a local maximum. Different initial projection matrices could lead to different local maximums. To alleviate the randomness, we have selected the first two eigenvectors of PCA as the initial projection the matrix to fairly compare different methods. Fig. 4 shows the 2D scatterplots of PCA, MaxEnt-PCA, and R1-PCA. We should note that the 2D scatterplot of MaxEnt-PCA are different in Fig. 3(c). This is due to random initialization of projection matrix.

Different distribution assumption will lead to an entirely different subspace. We see that the scatterplots of MaxEnt-PCA and PCA-L_1 are significantly different from those of PCA and R1-PCA. It is known that PCA is based on Gaussian assumption and MaxEnt-PCA is free from this assumption. Since the Balance Scale data set is obviously drawn from non-Gaussian distribution, PCA failed to keep the structure of data in the low dimensional subspace. Note that of R1-PCA is to produce a robust subspace to minimize the reconstruction error, and PCA-L_1 is to produce a robust subspace to maximize the variance, but they still did not keep the structure of data as well as MaxEnt-PCA because MaxEnt-PCA preserves the high order data distribution information.

4.2. Numerical results on dimension reduction

In this subsection, we quantitatively compared MaxEnt-PCA with PCA, R1-PCA [19], and PCA-L1 [20] so as to see the how entropy is different from the other typical metrics including L_2-norm, R_1-norm, and L_1-norm for PCA. All methods are applied to six data sets in the UCI machine learning repositories.[2]

Table 2 give a brief of these data sets, which have been used in many subspace learning studies [50,60]. For each data set, we performed 10-fold cross validation (CV) 10 times and computed the average correct classification rate. Each dimension of raw data was normalized using by the mean and variance. The 1-nearest-neighbor [20] algorithm was used, which is popularly used in subspace learning.

[2] UCI machine learning repositories [59] are well-known datasets to evaluate the performance of an algorithm for dimension reduction [50,60].

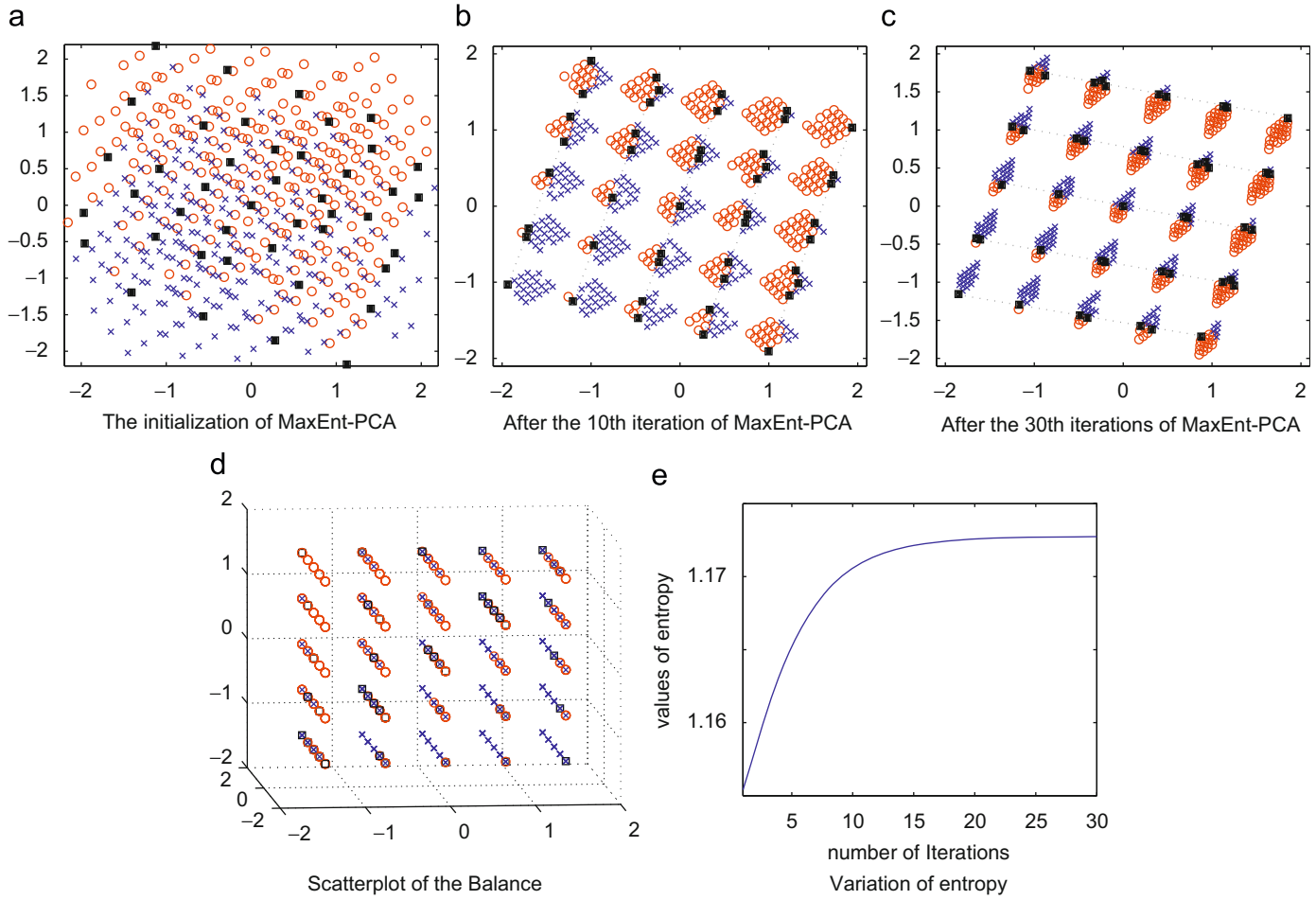


Fig. 3. Visualization results of MaxEnt-PCA on Balance Scale data set (see details in text).

Fig. 5 shows the average correct classification rates on each data set with various numbers of extracted features. The number of extracted features m varied from two to the dimension of original input space d. The results for all algorithms were the same when the number of extracted features were equal to d. For data sets with high dimension such as the "Ionosphere" and "Isolet", the dimension of subspace in Fig. 5 was truncated to 30 for a clear view.

Compared with other methods as shown in Fig. 5, MaxEnt-PCA can achieve better results except in Fig. 5(c). When the dimension of subspace was low, the curves of MaxEnt-PCA were higher than the others. As MaxEnt-PCA preserves the entropy of data, it can learn a subspace that model the variance of data more accurately. If the data is drawn from a noise-free Gaussian distribution, all methods yield similar subspaces.

4.3. Simulation results

Simulation studies [15,61–63] are often used to evaluate the robustness of different PCA methods to outliers.[3] In this subsection, we follow the lines of simulation study in [61] The true principal components V are taken as n samples from the m-variate standard normal distribution

$$v_i \sim N_m(0_m, I_m) \tag{30}$$

with I_m the $m \times m$ identity matrix and 0_m the null vector with m components. Then the base U_B(projection matrix) are defined as an orthogonal $d \times m$ matrix of uniformly distributed pseudorandom numbers. Given the principal components and the base, we can reconstruct the data matrix $X = U_{BV}$. For simulating outliers, we added the $100\alpha\%$ of observation of X with data from another distribution. The contaminated data $\hat{X}$ are constructed as $\hat{X} = [X_{(1)} X_{(2)}]$, where $X_{(1)}$ contains the first $100(1-\alpha)\%$ of the observations of X and $X_{(2)} = [x_{(2)1}, x_{(2)2}, \ldots, x_{(2)\alpha n}]$ is a $d \times \alpha n$ matrix taken from

$$x_{(2)i} \sim N_d(15_d, 8 \times I_d) \tag{31}$$

with 15_d a vector containing d elements equal to 15. Thus we can make use of the multivariate normal distribution defined in (31) for generating outlying values. Then standard normally distributed random noise (divided by 100) was added to our contaminated data matrix. This simulation process was repeated for levels of outlier fractions of 0%, 5%, 10%, 20%, and 30%, respectively. The simulations were repeated 200 times; the means of these 200 runs are reported.

In a simulation study, a measure of performance is necessary and important. However, it is not clear which measure [15,61,63] is the best one to evaluate a robust PCA [61]. Here, we make use of two measurements suggested by [61] based on the ratio of eigenvalues. The first measurement is defined as

$$\frac{\sum_{i=1}^{m} \hat{\lambda}_i}{\sum_{j=1}^{d} \hat{\lambda}_j} \tag{32}$$

[3] In pattern recognition, outliers are defined as points that deviate significantly from the rest of the data [19].

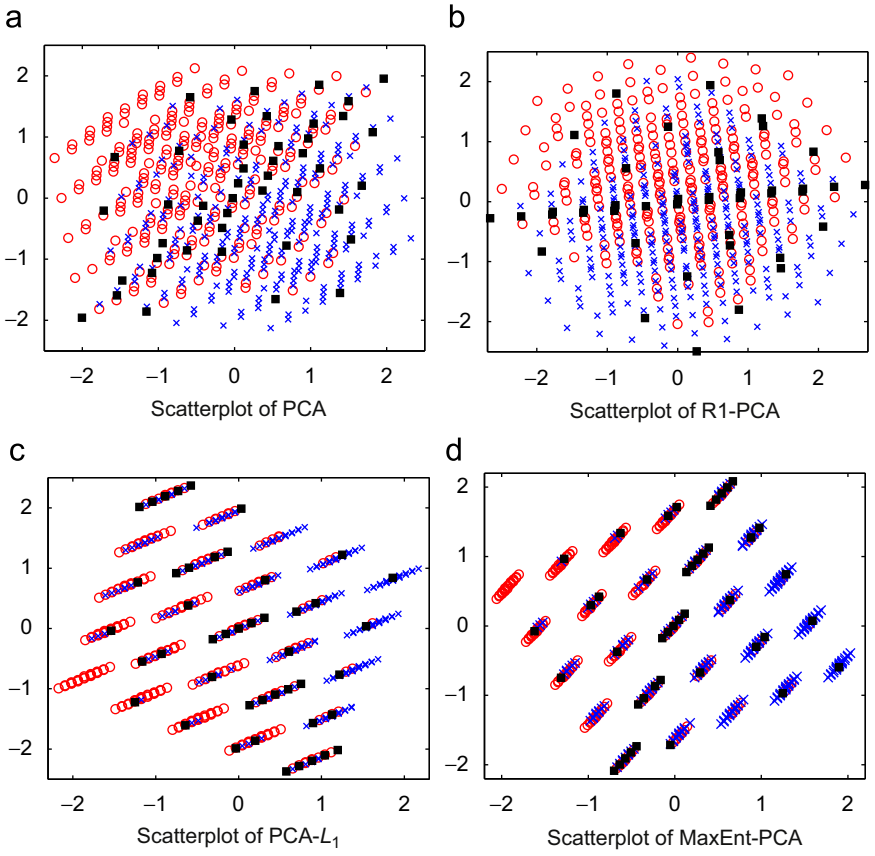


Fig. 4. Visualization results of different PCA algorithms on Balance data set where the initial projection matrix is generated by PCA.

where $\hat{\lambda}_i$ are the estimated eigenvalues. And the second measurement is defined as

$$\hat{\lambda}_1/\hat{\lambda}_2 \tag{33}$$

The second measure is an ancillary measure for the first measure. A PCA method is more robust if the two measurements are close to 1 [61].

We compare our algorithm with related methods.[4] Note that both R1-PCA and MaxEnt-PCA are rotation invariant, i.e., all orthonormal matrices are their solutions if the projection matrix $U \in R^{d\times d}$. Hence we should optimize the objectives in a lower dimensional subspace (i.e. $m < d$). Here, we set $U \in R^{10\times 5}$ for R1-PCA and MaxEnt-PCA. However, we can only obtain 5 eigenvalues rather than d eigenvalues. A simple way to get all d eigenvalues is to eigen-decompose the matrix $XL(U)X^T$ and $XW(U)X^T$($W(U)$ is the diagonal weight matrix in R1-PCA [19]), respectively, when two algorithms converge.

In RoPCA, there is a default parameter to control outliers. The value of this default parameter equals 0.25. To fairly evaluate different methods, we also introduce this parameter to R1-PCA and MaxEnt-PCA and set it to 0.25. A simple implementation is to remove $0.25n$ samples from $\hat{X}$ that have the smallest weights when two methods calculate the final d eigenvalues. For R1-PCA, we remove the samples according to weight matrix $W(U)$; and for MaxEnt-PCA, we remove the samples according to $D(U)$ that is the estimation of probability of x_i under the Parzen estimate. We denote the two methods as R1-PCA(0.75) and MaxEnt-PCA(0.75), respectively.

[4] Since there is no eigenvalues in RPCA [22] and PCA-L1 [20], they are not compared in the simulation experiment.

Simulation results were reported in Table 3. As expected, the classical PCA yields the best results if the data are not contaminated. The MaxEnt-PCA(0.75) outperform other methods when $\alpha \le 0.2$. Since the outliers are far away from the data cluster, the outliers will receive smaller values in the probability matrix $L(U)$ corresponding to the MaxEnt distribution. As a result, outliers were always removed in MaxEnt-PCA(0.75) and less affected the objective function. The MaxEnt-PCA can correctly detect outliers and hence yields stable results.

For 30% of outliers, R1-PCA(0.75) and LCA yield the best results under the first measure and spherical PCA yields the best result under the second measure. When $\alpha = 0.3$, the data $\hat{X}$ actually contains two normal distributions. There are $0.7n$ samples in $X_{(1)}$ and $0.3n$ samples in $X_{(2)}$. MaxEnt-PCA treats the data as a bimodal Gaussian distribution and tries to yield a MaxEnt distribution to fit the data. As a result, MaxEnt-PCA and MaxEnt-PCA(0.75) yield the lowest results under the first measure. But we consider this phenomenon as a coincidence with that data set.

4.4. Numerical results on FRGC data set

PCA is an important preprocessing step to reduce the dimension in face recognition (such as Eigenface [64] and Fisherface [4]) and discriminant methods (such as LDA and MFA). In this subsection, we evaluated different PCA methods on a challenging benchmark face recognition databases (FRGC version 2 face database [65]). There are 8014 images of 466 subjects in the query set for FRGC experiment 4. These uncontrolled still images contain the variations of illumination, expression, time, and blurring. The first 20 facial images were selected if the number of facial images is over 20. Finally we got 3720 facial

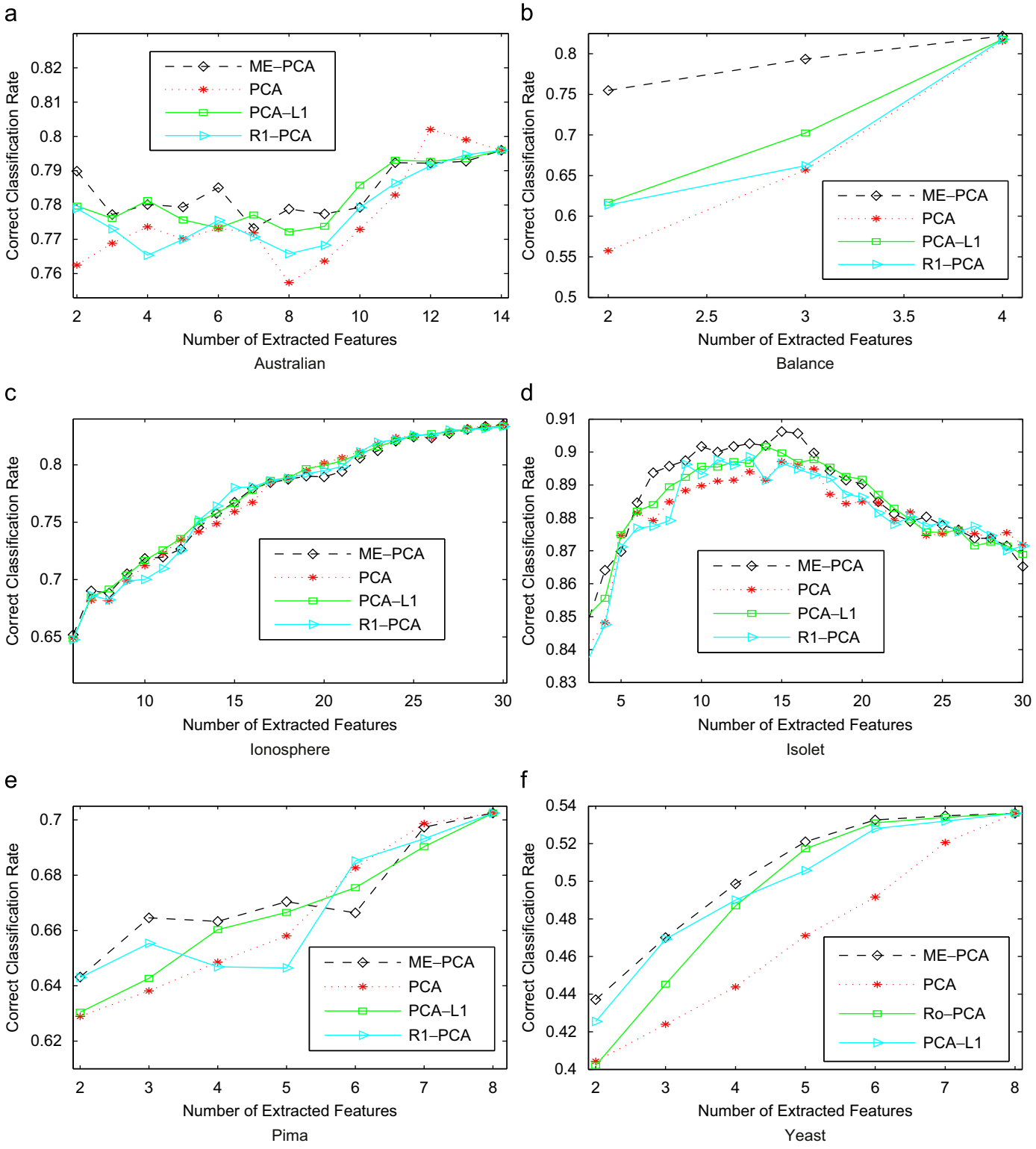


Fig. 5. Correct classification rates for UCI data sets. For ease of representation, we denote MaxEnt-PCA as ME-PCA.

images of 186 subjects. Each facial image was in 256 grey scales per pixel and cropped into size of 32 $\times$32 pixels by fixing the positions of two eyes. Table 2 summarizes the information of facial images. To simulate the outliers, we randomly blocked part of a facial image or randomly corrupted image pixels. Fig. 6 shows the original images in FRGC, occluded, and corrupted images.

The image set was randomly partitioned into a gallery and a probe set with different numbers. For ease of representation, *Gp* means *p* images per person are randomly selected for training and *Pq* means the remaining *q* images per person is used for testing. We performed on all extracted feature dimensions less than 200 and report the best results. To eliminate statistical deviations, all experiments were averaged over 10 random splits and the mean as well as the standard

Table 3
Simulation results for different levels of outliers ($n=100$, $p=10$, $h=2$).

Outliers	0%		5% ($\alpha=0.05$)		10% ($\alpha=0.1$)		20% ($\alpha=0.2$)		30% ($\alpha=0.3$)	
Method/measure	(32)	(33)	(32)	(33)	(32)	(33)	(32)	(33)	(32)	(33)
PCA	**1.00**	**1.0**	0.97	2.0	0.96	2.9	0.96	4.6	0.96	5.6
LCA	1.00	1.1	0.98	2.3	0.98	3.5	0.98	6.2	0.98	9.2
spherical PCA	1.00	1.1	0.99	1.1	0.99	1.2	0.98	1.3	0.97	1.8
Ro-PCA	1.00	1.2	1.00	1.4	1.00	1.3	**1.00**	**1.2**	0.97	2.4
R1-PCA	1.00	1.3	1.00	1.3	0.98	1.4	0.98	2.3	0.97	4.1
R1-PCA(0.75)	1.00	2.0	1.00	1.2	0.99	1.2	0.99	1.5	0.98	3.0
MaxEnt-PCA	1.00	1.1	**1.00**	**1.1**	0.99	1.1	0.96	1.1	0.88	1.6
MaxEnt-PCA(0.75)	1.00	1.2	1.00	1.2	**1.00**	**1.2**	**1.00**	**1.2**	0.96	2.8

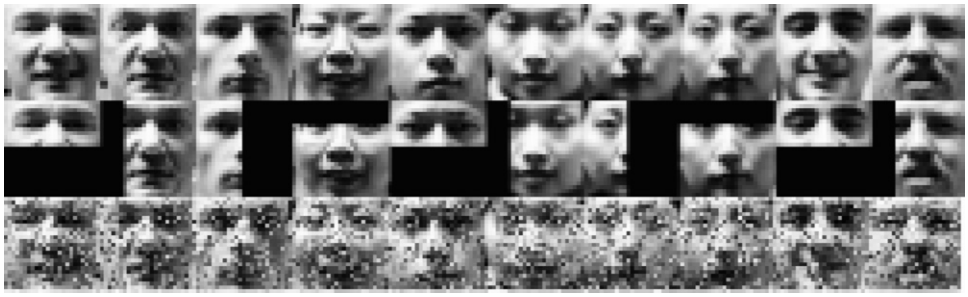


Fig. 6. Top row: original images in FRGC; middle row: occluded images; and bottom row: corrupted images.

Table 4
Comparison of PCA algorithms on FRGC database: average correct classification rate $\pm$ standard deviation.

FRGC	Original image		Occluded image		Corrupted image	
	G6/P14	G8/P12	G6/P14	G8/P12	G6/P14	G8/P12
PCA	54.1 $\pm$1.5	56.5 $\pm$0.9	49.0 $\pm$1.1	52.7 $\pm$1.2	52.1 $\pm$1.0	54.4 $\pm$0.9
Spherical PCA	54.2 $\pm$1.5	56.6 $\pm$0.9	52.4 $\pm$1.3	55.1 $\pm$1.0	52.3 $\pm$1.4	54.9 $\pm$0.8
RoPCA	48.5 $\pm$1.4	50.6 $\pm$0.9	46.3 $\pm$1.3	48.5 $\pm$1.1	47.1 $\pm$1.5	49.3 $\pm$0.9
RPCA	54.0 $\pm$1.5	56.3 $\pm$1.0	52.1 $\pm$1.2	54.8 $\pm$1.0	52.3 $\pm$1.5	54.5 $\pm$0.9
R1-PCA	54.2 $\pm$1.5	56.6 $\pm$0.9	49.1 $\pm$1.1	52.7 $\pm$1.2	52.5 $\pm$1.4	55.3 $\pm$0.9
PCA-L_1	54.1 $\pm$1.5	56.5 $\pm$0.9	49.0 $\pm$1.1	52.6 $\pm$1.2	52.9 $\pm$1.1	54.9 $\pm$0.9
LCA	56.3 $\pm$1.9	58.4 $\pm$1.4	52.4 $\pm$1.2	55.1 $\pm$1.1	53.8 $\pm$1.2	56.5 $\pm$1.3
MaxEnt-PCA	**58.5** $\pm$1.3	**61.0** $\pm$0.7	**55.6** $\pm$1.2	**58.6** $\pm$1.2	**55.2** $\pm$1.2	**58.3** $\pm$1.1

Best of all results are highlighted in bold.

Table 5
Simulation results for different levels of outliers ($n=100$, $p=10$, $h=2$).

Outliers	0%		5% ($\alpha=0.05$)		10% ($\alpha=0.1$)		20% ($\alpha=0.2$)		30% ($\alpha=0.3$)	
Method/Measure	(32)	(33)	(32)	(33)	(32)	(33)	(32)	(33)	(32)	(33)
Eq. (23)	1.00	1.1	1.00	1.1	0.99	1.1	0.96	1.1	0.88	1.6
Eq. (36)	1.00	1.1	1.00	1.1	0.98	1.1	0.96	1.7	0.92	2.6

deviation are reported. The nearest center classifier [66] was used for final classification. Considering that MaxEnt-PCA is based on a Parzen density estimation, we can calculate the probability of MaxEnt distribution in the subspace at x_i as

$$p(x_i) = D_{ii}(U) \Big/ \sum_{j=1}^{n} D_{jj}(U) \tag{34}$$

Then we can calculate the center x_c of a class C for MaxEnt-PCA as

$$\overline{x}_c = \frac{1}{\sum_{x_k \in C} p(x_k)} \sum_{x_k \in C} p(x_k) x_k \tag{35}$$

In the first experiment, there is no contamination in the gallery set. Table 4 tabulates the average correct classification rates of eight methods. Since LCA and MaxEnt-PCA can better model the data distribution, they outperformed six other methods. RoPCA obtained the lowest classification rates because it can only optimize at most 50 components. Still, MaxEnt-PCA achieved the highest accuracy, which show MaxEnt-PCA is also effective for dimensionality reduction on high dimensional data (Table 5).

In the second experiment, 20% of facial images in the gallery data set were randomly selected and occluded by a rectangular as shown in the second row of Fig. 6. There are obvious drops of performances for all methods. However, MaxEnt-PCA still achieved the highest

correct classification rate in two galleries. The experimental results show that MaxEnt-PCA is less affected by outliers.

In the third experiment, 20% of facial images in the gallery set were randomly corrupted by replacing a percentage of randomly selected pixels with random pixel value which followed a uniform distribution over [0,255]. MaxEnt-PCA achieves the highest correct classification rate in two galleries. The experimental results further validate that MaxEnt-PCA is also robust to corruption.

4.5. Parameter selection

The bandwidth σ is an important parameter which controls all robust properties of entropy [33]. This adjustable parameter provides an effective mechanism to eliminate the detrimental effect of outliers and noise, and makes entropy intrinsically different from the use of a threshold in conventional robust techniques [52]. The performance sensitivity to bandwidth is much smaller than the selection of thresholds [52]. In this work, we simply set the Gaussian kernel size as a single function of average distance in (23).

The selection of bandwidth is a hot issue in ITL based methods [67,68,60]. We can adopt the technique of simultaneous regression-scale estimation [69,67], Silverman's rule [70,52], Huber's rule [71,72] to select a robust bandwidth. In this paper, we follow the lines of Correntropy [52] and provide a tunable way ((23)) to select the bandwidth. A robust way suggested by [52] is the Silverman's rule [70]:

$$\sigma = 1.06 \times \min\{\sigma_E, R/1.34\} \times (n \times n)^{-1/5} \tag{36}$$

where σ_E is the standard deviation of the distance (i.e., $\|U^T x_i - U^T x_j\|$) and R is the interquartile range. To investigate the robustness of bandwidth selection, we compare the results of two bandwidth selection methods.

In the first experiment, we evaluate the performance of MaxEnt-PCA as the function of the s in bandwidth. The experimental setting is the same as that of the occlusion experiment in Section 4.4. And the p in Gp is 6. Fig. 7 shows the experimental results. We see that the classification rate estimated by (23) is even higher than that estimated by Silverman's rule when s is between 2.5 and 5. This phenomenon is coincident with the results in [52]. Although there are differences in classification rate, there is still a large range for selection of s to achieve a better result.

In the second experiment, simulation study is used to verify the robustness with and without contamination. The experimental setting is the same as that in Section 4.3. It is interesting to observe that MaxEnt-PCA under (23) outperforms MaxEnt-PCA under (36) when $\alpha \leq 0.2$. It seems that the bandwidth selection of (23) is more efficient than that of (36) in this simulation study. Note that the bandwidth of (23) is computed on the subspace rather than the high dimensional space. An outlier may be significantly far away from the data cluster, but it may be not in the subspace. Moveover, the bandwidth also controls the probability density estimation in MaxEnt. If the bandwidth can accurately reflect the MaxEnt distribution, MaxEnt-PCA can achieve better results. When $\alpha = 0.3$, the data $\hat{X}$ are from a bimodal Gaussian distribution. For the two measures, it is difficult to evaluate which bandwidth selection is better.

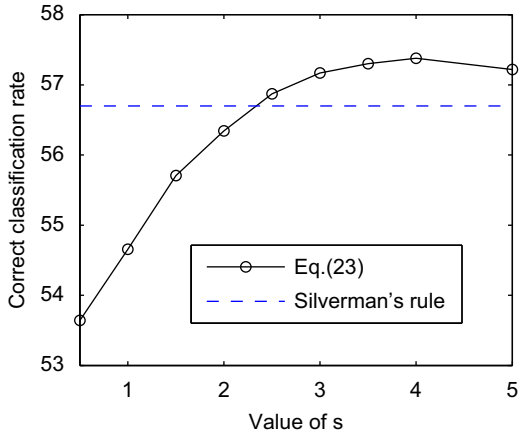


Fig. 7. Correct classification rate under different values of s in bandwidth σ (see text for details).

Table 6
Comparison of bandwidth selection methods on FRGC database: average correct classification rate ± standard deviation.

FRGC	Original image		Corrupted image	
	G6/P14	G8/P12	G6/P14	G8/P12
Eq. (23)	58.5 ± 1.3	61.0 ± 0.7	55.2 ± 1.2	58.3 ± 1.1
Eq. (36)	58.1 ± 1.3	60.6 ± 0.8	55.8 ± 1.0	58.8 ± 1.1

In the third experiment, we make use of a real-world data to investigate the bandwidth selection with and without contamination. The experimental setting is the same as that in Section 4.4. Table 6 shows the average correct classification rates for two bandwidth selection methods. The classification rate of (23) drops larger than that of (36). In this corruption case, the bandwidth selection by Silverman's rule seems more robust than that by (23). However, as discussed in Fig. 7, we can tune the parameter s to obtain a better result.

In this work, we study a simple method to estimate the bandwidth and choose a conservative way to set s to 2. Experimental results validate that MaxEnt-PCA can outperform other methods under this choice. A tunable bandwidth selection may be flexible for real-world unsupervised learning problems.

5. Conclusion

Based on the concept of maximum entropy (minimizing the information loss), a MaxEnt-PCA algorithm is proposed along with a theoretical analysis. The MaxEnt-PCA is rotation invariant and its optimal solution consists of the eigenvectors of a Laplacian probability matrix corresponding to a MaxEnt distribution. It has a clear theoretical foundation and provides a natural means to solve two major problems of traditional PCA (The sensitivity to noise and the Gaussian assumption). Experiments on real-world dimension reduction problems verify the superiority of MaxEnt-PCA.

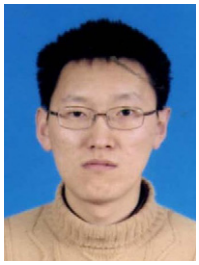
Acknowledgement

This project was supported by the NSFC (60675016, 60633030), the 973 Program (2006CB303104) and NSF-Guangdong (U0835005).

References

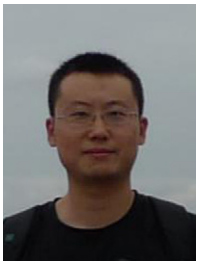
[1] I. Guyon, A. Elissee, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[2] J. Fan, R. Li, Statistical challenges with high dimensionality: feature selection in knowledge discovery, in: the Madrid International Congress of Mathematician, 2006, pp. 595–622.
[3] D. Tao, X. Li, X. Wu, S. Maybank, Tensor rank one discriminant analysis—a convergent method for discriminative multilinear subspace selection, Neurocomputing 71 (2008) 1866–1882.

[4] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1) (1997) 711–720.
[5] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using Laplacian faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.
[6] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: Advances in Neural Information Processing Systems, vol. 15, MIT Press, 2002, pp. 833–840.
[7] T. Zhang, X. Li, D.T.J. Yang, Local coordinates alignment (LCA): a novel manifold learning approach, IJPRAI 22 (4) (2008) 667–690.
[8] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology 24 (1933) 417–441.
[9] J. Li, X. Li, D. Tao, KPCA for semantic object extraction in images, Pattern Recognition 41 (2008) 3244–3250.
[10] Y. Pang, D. Tao, Y. Yuan, X. L1, Binary two-dimensional PCA, IEEE Transactions on Systems, Man, and Cybernetics, Part B 38 (2008) 1176–1180.
[11] S. Chen, H. Zhao, M. Kong, B. Luo, 2d-lpp: A two-dimensional extension of locality preserving projections, Neurocomputing 70 (2007) 912–921.
[12] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.
[13] L. Xu, A. Yuille, Robust principal component analysis by self-organizing rules based on statistical physics approach, IEEE Transactions on Neural Networks 6 (1995) 131–143.
[14] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, Robust principal component analysis for functional data, Test 8 (1) (1999) 1–73.
[15] R.A. Maronna, Principal components and orthogonal regression based on robust scales, Technometrics 47 (2005) 264–273.
[16] Y. Pang, X. Li, Y. Yuan, D. Tao, J. Pan, Fast haar transform based feature extraction for face representation and recognition, IEEE Transactions on Information Forensics and Security 4 (2009) 441–450.
[17] A. Baccini, P. Besse, A. Falguerolles, A l1-norm PCA and a heuristic approach, Ordinal and Symbolic Data Analysis 1 (1996) 359–368.
[18] Q. Ke, T. Kanade, Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: Computer Vision and Pattern Recognition, 2005.
[19] C. Ding, D. Zhou, X. He, H. Zha, R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization, in: International Conference on Machine Learning, 2006.
[20] N. Kwak, Principal component analysis based on l1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (9) (2008) 1672–1677.
[21] Y.W. Pang, X.L. Li, Y. Yuan, L1-norm based tensor analysis, IEEE Transactions on Circuits and Systems for Video Technology 20 (2) (2010) 172–178.
[22] F. Torre, M. Black, A framework for robust subspace learning, International Journal of Computer Vision 54 (1–3) (2003) 117–142.
[23] M. Hubert, P. Rousseeuwy, K. Branden, Robpca a new approach to robust principal component analysis, Technometrics 47 (2005) 64–79.
[24] J. Yang, D. Zhang, J. Yang, Locally principal component learning for face representation and recognition, Neurocomputing 69 (2006) 1697–1701.
[25] R. Moller, H. Hoffmannb, An extension of neural gas to local pca, Neurocomputing 62 (2004) 305–326.
[26] D. Huang, Z. Yia, X. Pu, A new local pca-som algorithm, Neurocomputing 71 (2004) 16–18.
[27] J. Park, Z. Zhang, H. Zha, R. Kasturi, Local smoothing for manifold learning, in: Computer Vision and Pattern Recognition, 2004.
[28] D. Zhao, Z. Lin, X. Tang, Laplacian pca and its applications, in: International Conference on Computer Vision, 2007.
[29] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis, in: Computer Vision and Pattern Recognition, vol. 1, 2003, pp. 621–628.
[30] J. Principe, D. Xu, J.W.F. Iii, Information-theoretic learning 〈http://www.cnel.ufl.edu/bib/./pdf_papers/chapter7.pdf〉.
[31] D. Xu, Energy, entropy and information potential for neural computation, Ph.D. Thesis, University of Florida, 1999.
[32] P. Viola, N. Schraudolph, T. Sejnowski, Empirical entropy manipulation for real-world problems, in: Neural Information Processing Systems (NIPS), 1995, pp. 851–857.
[33] X. Yuan, B. Hu, Robust feature extraction via information theoretic learning, in: International Conference on Machine Learning (ICML 2009), Montreal, Canada, 2009.
[34] M. Girolami, Orthogonal series density estimation and the kernel eigenvalue problem, Neural Computation 14 (3) (2002) 669–688.
[35] R. Jenssen, D. Erdogmus, J. Principe, T. Eltoft, Some equivalences between kernel methods and information theoretic methods, Journal of VLSI Signal Processing 45 (2006) 49–65.
[36] R. Jenssen, D. Erdogmus, J. Principe, T. Eltoft, Information theoretic angle-based spectral clustering: a theoretical analysis and an algorithm, in: International joint conference on neural networks, 2006, pp. 4904–4911.
[37] A.R. Paiva, J. Wu, J. Principe, kernel principal component are maximum entropy projection, in: ICA, 2006, pp. 846–853.
[38] R. Jenssen, T. Eltoft, M. Girolami, D. Erdogmus, Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm, in: Neural Information Processing Systems (NIPS), 2006.
[39] K. Pearson, On lines and planes of closest fit to systems of points in space, The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series 2 (1901) 559–572.
[40] D. Luenberger, Optimization by Vector Space Methods, Wiley, 1969.
[41] S. Roweis, Em algorithms for PCA and SPCA, in: Neural Information Processing Systems (NIPS), 1997, pp. 626–632.
[42] J. Ahna, J. Oha, S. Choib, Learning principal directions: integrated-squared-error minimization, Neurocomputing 70 (2007) 1372–1381.
[43] G. Golub, C.V. Loan, Matrix Computations, third ed., Johns Hopkins, Baltimore, 1996.
[44] T.H. Zhang, D.C. Tao, X.L. Li, J. Yang, Patch alignment for dimensionality reduction, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1299–1313.
[45] Q. Ke, T. Kanade, Robust subspace computation using l1 norm, Techical Report 〈http://citeseer.ist.psu.edu/ke03robust.html〉.
[46] S. Danijel, L. Ales, B. Horst, Weighted and robust learning of subspace representations, Pattern Recognition 40 (2007) 1556–1569.
[47] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[48] B. Scholkopf, A.J. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319.
[49] T. Cover, J.A. Thomas, Elements of Information Theory, second ed., John Wiley, New Jersey, 2005.
[50] K. Torkkola, Feature extraction by nonparametric mutual information maximization, Journal of Machine Learning Research 3 (2003) 1415–1438.
[51] P. Huber, Robust Statistics, Wiley, 1981.
[52] W. Liu, P.P. Pokharel, J.C. Principe, Correntropy: properties and applications in non-Gaussian signal processing, IEEE Transactions on Signal Processing 55 (11) (2007) 5286–5298.
[53] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks 10 (1999) 626–634.
[54] A. Sharma, K.K. Paliwal, Fast principal component analysis using fixed-point algorithm, Pattern Recognition Letters 28 (2007) 1151–1155.
[55] Y. Pang, Y. Yuan, X. Li, Iterative subspace analysis based on feature line distance, IEEE Transactions on Image Processing 18 (2009) 903–907.
[56] J. Nocedal, S.J. Wright, Numberical Optimization, Springer, 2000.
[57] A. Ng, Feature selection, l1 vs. l2 regularization, and rotational invariance, in: International Conference on Machine Learning, 2004.
[58] T. Cover, J.A. Thomas, Elements of Information Theory, second ed., John Wiley, New York, 2005.
[59] D. Newman, S. Hettich, C. Blake, C. Merz, Uci repository of machine learning databases 〈http://www.ics.uci.edu/mlearn/MLRepository.html〉.
[60] K.H. II, D. Erdogmus, K. Torkkola, C. Principe, Feature extraction using information-theoretic learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1385–1392.
[61] S. Serneels, T. Verdonck, Principal component analysis for data containing outliers and missing elements, Computational Statistics and Data Analysis 52 (2008) 1712–1727.
[62] A. van der Linde, Variational Bayesian functional PCA, Computational Statistics and Data Analysis 53 (2008) 517–533.
[63] M. Hubert, P. Rousseeuw, T. Verdonck, Robust pca for skewed data and its outlier map, Computational Statistics and Data Analysis 53 (2009) 2264–2274.
[64] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[65] P. Philips, P. Flynn, T. Sruggs, K. Bowyer, Overview of the face recognition grand challenge, in: Computer Vision and Pattern Recognition, 2005.
[66] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley-Interscience, 2000.
[67] S.H. Yang, H.Y. Zha, S. Zhou, B.-G. Hu, Variational graph embedding for globally and locally consistent feature extraction, in: Europe Conference on Machine Learning (ECML), 2009, pp. 538–553.
[68] R. He, B. Hu, X. Yuan, Robust discriminant analysis based on nonparametric maximum entropy, in: Asian Conference on Machine Learning (ACML), NanJing, China, 2009.
[69] I. Mizera, C. Muller, Breakdown points of cauchy regression-scale estimators, Statistics and Probability Letters 57 (2002) 79–89.
[70] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, UK, 1986.
[71] P.J. Rousseeuw, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[72] M. Rogers, J. Graham, Robust active shape model search, in: European Conference on Computer Vision, Springer, 2002, pp. 517–530.



Ran He received the B.S. degree in Computer Science from the Dalian University of Technology of China, and the Ph.D. degree in Pattern Recognition and Intelligent System from Institute of Automation, Chinese Academy of Sciences, in 2009. He is currently an Assistant Professor in Dalian University of Technology. His research interests include information theoretic learning and computer vision.

Bao-Gang Hu received his M.Sc. degree from the University of Science and Technology, Beijing, China in 1983, and his Ph.D. degree from McMaster University, Canada in 1993, all in Mechanical Engineering. From 1994 to 1997, Dr. Hu was a Research Engineer and Senior Research Engineer at C-CORE, Memorial University of Newfoundland, Canada. Currently, he is a Professor with NLPR (National Laboratory of Pattern Recognition), Institute of Automation, Chinese Academy of Science, Beijing, China. From 2000 to 2005, he was the Chinese Director of LIAMA (the Chinese-French Joint Laboratory for Computer Science, Control and Applied Mathematics). His main research interests include intelligent systems, pattern recognition, plant growth modeling. He is a Senior Member of IEEE.



XiaoTong Yuan received the B.S. degree in Computer Science from the Nanjing University of Posts and Telecommunications, China, in 2002, and the Ph.D. degree in Pattern Recognition and Intelligent System from Institute of Automation, Chinese Academy of Sciences. He is currently a Postdoctoral Research Fellow in National University of Singapore. His research interests include machine learning, data mining and computer vision.

Wei-Shi Zheng is a Postdoctoral Researcher at Department of Computer Science, Queen Mary University of London, UK. He is now working on the European SAMURAI Research Project with Prof. Gong Shaogang and Dr. Xiang Tao. Prior to that, he received his Ph.D. degree in Applied Mathematics at Sun Yat-Sen University, China, 2008. He has been a visiting student working with Prof. Li Stan Z. at Institute of Automation, Chinese Academy of Sciences, and an exchanged research student working with Prof. Yuen Pong C. at Hong Kong Baptist University. He was awarded the HP Chinese Excellent Student Scholarship 2008. Dr. Zheng is a member of IEEE. He has served as a regular reviewer for IEEE TPAMI, IEEE TNN, IEEE TCSVT, Pattern Recognition and etc in past two years. His current research interests are in object association and categorization. He is also interested in discriminant/sparse feature extraction and dimension reduction, kernel methods in machine learning, and face image analysis.