Robust Principal Component Analysis Based on Maximum Correntropy Criterion

Ran He, Bao-Gang Hu, *Senior Member, IEEE*, Wei-Shi Zheng, *Member, IEEE*, and Xiang-Wei Kong

***Abstract*—Principal component analysis (PCA) minimizes the mean square error (MSE) and is sensitive to outliers. In this paper, we present a new rotational-invariant PCA based on maximum correntropy criterion (MCC). A half-quadratic optimization algorithm is adopted to compute the correntropy objective. At each iteration, the complex optimization problem is reduced to a quadratic problem that can be efficiently solved by a standard optimization method. The proposed method exhibits the following benefits: 1) it is robust to outliers through the mechanism of MCC which can be more theoretically solid than a heuristic rule based on MSE; 2) it requires no assumption about the zero-mean of data for processing and can estimate data mean during optimization; and 3) its optimal solution consists of principal eigenvectors of a robust covariance matrix corresponding to the largest eigenvalues. In addition, kernel techniques are further introduced in the proposed method to deal with nonlinearly distributed data. Numerical results demonstrate that the proposed method can outperform robust rotational-invariant PCAs based on L_1 norm when outliers occur.**

***Index Terms*—Correntropy, half-quadratic optimization, principal component analysis (PCA), robust.**

I. Introduction

PRINCIPAL component analysis (PCA) [1] and [2] is a linear data transformation technique which plays an important role in image processing and machine learning. It has been widely used for the representation of high-dimensional data such as image data for appearance, shape, and visual tracking and is also popularly used as a preprocessing step to project high-dimensional data into a low-dimensional subspace. However, PCA also has limitations. Since large errors will dominate the mean square error (MSE), standard PCA is prone to the presence of outliers that are significantly far away from the rest of the data points [3]–[5].

Manuscript received December 23, 2009; revised May 25, 2010; accepted December 24, 2010. Date of publication January 06, 2011; date of current version May 18, 2011. This work was supported in part by the Research Foundation for the Doctoral Program of the Ministry of Education of China under Grant 20100041120009, the Natural Science of Foundation of China under Grant 61075051 and Grant 60971095, the NSFC-GuangDong under Grant U0835005, and the Sun Yat-sen 985 Project under Grant 35000–3181305. The associate editor coordinating the review of this manuscript and approving it for publication was B. Sankur.

R. He and B. G. Hu are with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China (e-mail: rhe@nlpr.ia.ac.cn; hubg@nlpr.ia.ac.cn).

W.-S. Zheng is with the School of Information Science and Technology, Sun Yat–sen University, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

X. W. Kong is with the School of Electronic and Information Engineering, The Dalian University of Technology, Dalian 116024, China (e-mail: kongxw@dlut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2103949

In order to alleviate the negative effect of outliers, various robust alternatives have been proposed. In [6] and [7], L_1-norm PCA was formulated by applying maximum-likelihood estimation to input data. Heuristic estimation method and convex programming methods were proposed to detect outliers in [6] and [7], respectively. In [8], hard redescending nonconvex M-estimators were used as objectives to learn a robust representation of color images. Despite the robustness of these three methods, they have a common limitation that they are not rotationally invariant, which is a fundamental property in the context of learning algorithms [9]. Hence, rotationally invariant PCA algorithms are developed [4], [10], [11]. R1-PCA [4] utilizes Cauchy robust function to calculate the weight of each data point and removes outliers by a subspace iteration algorithm. PCA-L_1 [10] adopts a greedy strategy to maximize a L_1-norm dispersion. Φ-PCA [11] formulates the objective function as a twice-differentiable and convex function that can be optimized by the Newton gradient algorithm. However, R1-PCA and PCA-L_1 assume that data are already centered, which is difficult to ensure in practice especially when outliers occur [12]. Outliers will make the data mean biased so that the robustness of algorithms decreases. Φ-PCA needs to calculate the Hessian matrix and can only optimize the data mean and principal components separately.

In this paper, we address the issue of the robustness of the rotational invariant PCA algorithm based on maximum correntropy criterion (MCC) [13] that is a useful measurement to handle nonzero mean and non-Gaussian noise with large outliers. Gaussian kernel function is selected as the objective in MCC, which also belongs to redescending M-estimators [14] and [15]. Since the correntropy objective can be optimized efficiently via Half-Quadratic (HQ) optimization framework in an iterative manner, we denote the new PCA method as HQ-PCA. The complex optimization problem can thereby be reduced to a quadratic optimization problem so that it can be efficiently solved by a standard optimization method. The HQ framework can also be easily extended to solve other correntropy problems or robust PCAs based on M-estimators. From the viewpoint of Information Theoretic Learning ($\mathbf{ITL}$) [16], HQ-PCA is a natural extension of PCA by replacing MSE criterion by MCC and has several appealing advantages, which are given here.

1) It is rotationally invariant and robust to outliers.
2) It can handle noncentered data and can naturally estimate data mean.
3) Optimal solutions of the proposed method are the principal eigenvectors of a robust covariance matrix corresponding to the largest eigenvalues.

In addition, a kernel method for performing a nonlinear form of HQ-PCA is developed to deal with nonlinearly distributed data [17], [18].

The remainder of this work is organized as follows. In Section II, we briefly review previous robust PCA methods and point out their main limitations. In Section III, we discuss a new objective for robust PCA and propose an algorithm based on half-quadratic optimization. In Section IV, we evaluate our method on face reconstruction, clustering and dimension reduction tasks. Finally, we summarize the paper in Section V.

II. PCA AND L_1 PCA

Consider a data set of samples $X = [x_1, \ldots, x_n]$, where x_i is a variable in Euclidean space with dimensionality d, $U = [u_1, \ldots, u_m] \in R^{d\times m}$ is a projection matrix whose columns constitute the bases of a m-dimensional subspace and $V = [v_1, \ldots, v_n] \in R^{m\times n}$ is the principal components that are projection coordinates under the projection matrix U. Based on MSE, PCA can be formulated as the following optimization problem:

$$\min_{U,V} \sum_{i=1}^{n} \|x_i - (\mu + Uv_i)\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} \left(x_{ij} - \left(\mu_j + \sum_{p=1}^{m} v_{ip}u_{pj} \right) \right)^2 \quad (1)$$

where $\|.\|$ is the L_2-norm and μ is the center of X. By projection theorem [19], for a fixed U, the V that minimizes (1) is uniquely determined by $V = U^T X$. Because (1) is based on L_2-norm (Euclidean distance), the PCA is often denoted as L_2-PCA.

The global minimum of (1) is provided by singular value decomposition (SVD) [20], whose optimal solution is also the solution of the following alternative formulation of PCA:

$$\max_{U^TU=I} Tr\left(U^T \Sigma U\right) \quad (2)$$

where $\Sigma = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$ is the covariance matrix, $Tr(.)$ denotes the matrix trace operation and T denotes the transpose. The (2) learns a projection matrix where the variances of $U^T X$ are maximized.

Since the L_2-norm based PCA is sensitive to outliers [10], L_1-norm was applied in (1) to substitute the L_2-norm. From the statistical viewpoint, those methods based on L_1-norm are more robust to outliers than L_2-norm based ones [7], [10] and [21]. In this case, the problem of PCA becomes finding the U that minimizes the following reconstruction error function:

$$\min_{U,V} \sum_{i=1}^{n}\sum_{j=1}^{d} \left\| x_{ij} - \left(\mu_j + \sum_{p=1}^{m} v_{ip}u_{pj} \right) \right\|_{L_1}. \quad (3)$$

Although methods based on (3) can improve robustness to outliers, those methods based on L_1-norm [6], [7] and [21] are not invariant to rotation of input feature space and the equidistance surface will be very skewed [4]. To overcome these problems, rotational invariant L_1-norm PCAs are developed by relaxing the objective function in (3) [4], [10]. Rotational invariant PCAs show their superiorities in clustering tasks [4].

However, those rotationally invariant PCAs are often based on heuristic strategy to remove outliers and lack theoretical foundations. Furthermore, they assume that data are already centered, which is difficult to ensure in practice especially when outliers occur [3], [5] and [12]. A single outlier may pull the data mean outside the range of data cluster [3]. In ITL [16], it has been shown that PCA can also be formulated as a maximum entropy problem and this work follows the line of ITL to develop a novel robust and rotational invariant PCA algorithm based on the maximum correntropy framework, which enables a close relationship between the proposed robust PCA and a deep investigation of ITL.

III. ROBUST PCA BASED ON MCC

Here, we propose a new robust PCA algorithm based on MCC and then develop a half-quadratic optimization algorithm to maximize the objective, along with convergence analysis.

A. Maximum Correntropy Criterion

Recently, the concept of correntropy [13] was proposed for ITL. It is derived from the generalized correlation function of random processes and is directly related to the information potential (IP) of Renyi's quadratic entropy [16] in which Parzen windowing method is used to estimate the data's probability distribution [22]. Based on the information potential, the correntropy is defined as a generalized similarity measure between two arbitrary random variables A and B

$$V_\sigma(A, B) = E[k_\sigma(A - B)] \quad (4)$$

where $k_\sigma(.)$ is the kernel function that satisfies Mercer's theory [23] and $E[.]$ denotes the mathematical expectation. It takes advantage of the kernel technique that nonlinearly maps the input space to a higher dimensional space. Different from conventional kernel methods, it works independently with pairwise samples. It has a clear theoretical foundation and is symmetric, positive, and bounded.

In practice, the joint probability density function is often unknown, and only a finite number of data $\{(a_i, b_i)\}_{i=1}^{n}$ are available, which lead to the following sample estimator of correntropy:

$$\hat{V}_{n,\sigma}(A, B) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(a_i - b_i). \quad (5)$$

When k_σ is the Gaussian kernel $g(x) \triangleq \exp(-x^2/2\sigma^2)$, we can rewrite (5) as

$$\hat{V}_{n,\sigma}(A, B) = \frac{1}{n}\sum_{i=1}^{n} g(a_i - b_i). \quad (6)$$

The maximum of correntropy of error in (5) is called the maximum correntropy criterion (MCC) [13]. Compared with the global measure—mean square error (MSE), MCC is local, which means that the value of correntropy is mainly decided by the kernel function along the line $A = B$ [13]. Correntropy has a close relationship with m-estimators [14]. If we define $\rho(x) \triangleq 1 - g(x)$, (6) is the robust formulation of Welsch m-estimator [13]. A main merit of correntropy is that the kernel size controls all of the properties of correntropy [13]. It establishes a close relationship between the m-estimation and methods of ITL and provides a practical way to choose an appropriate kernel size [13]. Moverover, the optimization of MCC-based criterion is easier than that of the methods based on L_1-norm.

B. PCA Based on MCC

Substituting $a_i = x_i$ and $b_i = \mu + Uv_i$ into (5), we obtain the following maximum correntropy problem:

$$\max_{\theta,V} J_1(\theta) = \sum_{i=1}^{n} g(x_i - \mu - Uv_i) \tag{7}$$

where $\theta \triangleq (\mu, U)$, and we have

$$\arg\max_{\theta,V} \sum_{i=1}^{n} g(x_i - \mu - Uv_i) = \arg\min_{\theta,V} \sum_{i=1}^{n} \rho(x_i - \mu - Uv_i). \tag{8}$$

Compared with (1), (8) replaces the L_2-norm with the Welsch m-estimator in the objective function. Hence, (8) is a robust *m-estimator* formulation of L_2-PCA. Provided that the U is orthonormal, i.e., $U^TU = I$, we can obtain

$$\sqrt{\|x - UU^Tx\|^2} = \sqrt{x^Tx - x^TUU^Tx}. \tag{9}$$

Substituting $v_i = U^T(x_i - \mu)$ (projection theorem [19]) into (7) and according to (9), we get the following optimization problem:

$$\max_{\theta} J_{\mathrm{HQ}}(\theta) = \sum_{i=1}^{n} g\left(\sqrt{x_i^{\mu T}x_i^{\mu} - x_i^{\mu T}UU^Tx_i^{\mu}}\right) \tag{10}$$

where U is an orthonormal matrix and $x_i^{\mu} = x_i - \mu$. We denote the new method to solve J_{HQ} based on half-quadratic (HQ) optimization as HQ-PCA. Since the J_{HQ} is based on the Huber's M-estimator of reconstruction error, the large reconstruction error in outliers makes detection easier.

C. Optimization Procedure via Half-Quadratic

In ITL, the half-quadratic technique [8], [24]–[26] is often used to solve nonlinear ITL optimization problem. In this section, we derive an algorithm to solve (10) based on the half-quadratic. Based on the theory of convex conjugated functions [24], we can easily derive the following proposition.

Proposition 1: There exists a convex conjugated function φ of $g(x)$ such that

$$g(x) = \max_{p'} \left(p' \frac{\|x\|^2}{\sigma^2} - \varphi(p')\right) \tag{11}$$

where $p' \in R$ is a scalar variable, and, for a fixed x, the maximum is reached at $p' = -g(x)$, [25], [27].

Substituting (11) into (10), we have the augmented objective function in an enlarged parameter space

$$\hat{J}_{\mathrm{HQ}}(\theta, p) = \sum_{i=1}^{n} \left(p_i \left(x_i^{\mu T}x_i^{\mu} - x_i^{\mu T}UU^Tx_i^{\mu}\right) - \varphi(p_i)\right) \tag{12}$$

where $p = [p_1, \ldots, p_n]^T$ is storing the auxiliary variables introduced in the Half-Quadratic optimization. According to Proposition 1, for the fixed θ, the equation $J_{\mathrm{HQ}}(\theta) = \max_p \hat{J}_{\mathrm{HQ}}(\theta, p)$ holds true. It follows that

$$\max_{\theta} J_{\mathrm{HQ}}(\theta) = \max_{\theta,p} \hat{J}_{\mathrm{HQ}}(\theta, p). \tag{13}$$

Then, we can conclude that maximizing $J_{\mathrm{HQ}}(\theta)$ is identical to maximizing the augmented function $\hat{J}_{\mathrm{HQ}}(\theta, p)$. Obviously, one local maximizer (θ, p) can be calculated in an alternating maximization way

$$p_i^{t+1} = -g\left(\sqrt{x_i^{\mu T}x_i^{\mu} - x_i^{\mu T}(U^t)(U^t)^Tx_i^{\mu}}\right) \tag{14}$$

$$\mu^{t+1} = \frac{1}{\left(\sum_{i=1}^{n} p_i^{t+1}\right)} \sum_{i=1}^{n} p_i^{t+1}x_i \tag{15}$$

$$U^{t+1} = \arg\max_{U} Tr\left(U^T X_c^{t+1} P^{t+1} \left(X_c^{t+1}\right)^T U\right)$$
$$s.t.\ U^TU = I \tag{16}$$

where t is the tth iteration, $x_i^{\mu} = x_i - \mu^t$, $X_c^{t+1} = [x_1 - \mu^{t+1}, \ldots, x_n - \mu^{t+1}]$, and matrix P^{t+1} is a diagonal matrix whose diagonal entity $P^{t+1}(i,i) = -p_i^{t+1}$. The optimization problem in (16) is the optimization problem of weighted PCA. The subspace U on the right-hand side in (16) is a variable that should be estimated in the alternating maximization. Its solution is given by the following eigenvalue problem:

$$X_c^{t+1} P^{t+1} \left(X_c^{t+1}\right)^T U = U\Lambda \tag{17}$$

where Λ is a diagonal matrix whose diagonal elements are the m largest eigenvalues. U consists of m eigenvectors of the weighted covariance matrix $X_c^{t+1} P^{t+1} \left(X_c^{t+1}\right)^T$ which corresponds to the m largest eigenvalues. If the number of data points n is smaller than the number of dimension d, the U can be calculated by the following eigenvalue problem [28]:

$$\sqrt{P^{t+1}} \left(X_c^{t+1}\right)^T X_c^{t+1} \sqrt{P^{t+1}} \hat{U} = \hat{U}\Lambda \tag{18}$$

$$U = \left(\sqrt{\sum_{i=1}^{n} -p_i^{t+1}}\right)^{-1} X_c^{t+1} \hat{U} \left(\sqrt{\Lambda}\right)^{-1}. \tag{19}$$

We first compute a subspace $\hat{U}$ according to (18). We then obtain the desired subspace U according to (19).

The algorithm of HQ-PCA is summarized in Algorithm 1. The HQ-PCA reduces the complex optimization problem to a weighted PCA problem and increases the objective step by step until it converges (Proposition 2). Compared with (2) and (16), the problem of (16) is actually a weighted PCA. Its solution U ($U \in R^{d \times m_r}$) consists of the principal eigenvectors corresponding to the m_r largest eigenvalues. Hence, HQ-PCA's optimal solution consists of the principal eigenvectors of $X_c^{t+1} P^{t+1} (X_c^{t+1})^T$, which is computed in the last iteration when it converges. Furthermore, it can naturally estimate data mean during optimization.

If we want to learn a m ($m_r < m \leq d$) dimensional subspace, we can directly learn the subspace by calculating the eigenvectors of matrix $X_c^{t+1} P^{t+1} (X_c^{t+1})^T$ (Step 10) where outliers have received small values in P^{t+1}. This property also suggests that if we want to learn a m dimensional subspace that is robust to outliers, we can first learn a m_r ($m_r < m$) dimensional subspace where outliers can be detected. This can significantly reduce the computation cost of HQ-PCA.

Proposition 2: The sequence $\left\{\hat{J}_{\mathrm{HQ}}^t\left(\mu^t, U^t, p^t\right), t = 1, 2, \ldots\right\}$ generated by HQ-PCA converges.

Proof: According to (15), (16) and Proposition 1, we have

$$\begin{aligned}\hat{J}_{\mathrm{HQ}}\left(\mu^t, U^t, p^t\right) &\le \hat{J}_{\mathrm{HQ}}\left(\mu^t, U^t, p^{t+1}\right)\\ &\le \hat{J}_{\mathrm{HQ}}\left(\mu^{t+1}, U^t, p^{t+1}\right)\\ &\le \hat{J}_{\mathrm{HQ}}\left(\mu^{t+1}, U^{t+1}, p^{t+1}\right).\end{aligned}$$

The cost function increases at each alternating maximization step. Therefore, the sequence $\left\{\hat{J}_{\mathrm{HQ}}\left(\mu^t, U^t, p^t\right),\ t = 1, 2, \ldots\right\}$ is non-decreasing. It can be verified that $J_{\mathrm{HQ}}(\theta)$ is bounded (property of correntropy [13]) and by (13) we get that $\hat{J}_{\mathrm{HQ}}^t\left(\mu^t, U^t, p^t\right)$ is also bounded. Consequently we can conclude that HQ-PCA will converge. ∎

In ITL, it has been pointed out that MCC is a local measurement whereas MSE is a global measurement [13]. By global, all the data points in the joint space will contribute equally to the value of the measurement and the locality of MCC means that the value is mainly determined by the kernel function along the $A = B$ [in (4)] line [13]. Since an outlier is far away from the data cluster, its contribution to estimating correntropy will be smaller so that it always receives a low value in the matrix P^{t+1}. Therefore, the outliers will have weaker influence on the estimation of θ as correntropy increases. As a result, HQ-PCA is robust against outliers. Algorithm 1 also provides a new means to solve the maximum correntropy problem and an alternative viewpoint to analyze the relationship between MCC and MSE. MCC achieves its locality by softening the data points far away from the $A = B$ step by step.

Algorithm 1 HQ-PCA

Input: data matrix X, a small positive value ε and an orthonormal matrix $U \in R^{d\times m_r}$ $(m_r < m \le d)$

Output: μ and $U \in R^{d\times m}$

1: **repeat**
2: Initialize $converged = \mathrm{FALSE}$.
3: Update p according to (14),
4: Update μ according to (15),
5: Update U according to (17) or (19).
6: **if** the difference of the correntropy in (10) is smaller than ε **then**
7: $converged = \mathrm{TRUE}$
8: **end if**
9: **until** $converged == \mathrm{TRUE}$
10: Calculate U that consists of m eigenvectors of $X_c^{t+1}P^{t+1}\left(X_c^{t+1}\right)^T$ corresponding to the m largest eigenvalues if it is necessary to learn a higher dimensional subspace.

From the robust M-estimator point of view, HQ-PCA can also be treated as a generalized one of R1-PCA, Φ-PCA[1] and M-Scale PCA[2] [5]. However, R1-PCA has to assume that the data are already centered, and Φ-PCA requires that the objective function must be a twice-differentiable and convex function and in comparison the proposed HQ-PCA is free from these requirement. Compared with the m-estimators used in other robust PCAs, there is no threshold in the Gaussian function of correntropy. The kernel size σ controls all properties of this robust estimator [13]. The Gaussian-like weighting function in the alternating maximum step of HQ-PCA attenuates the large error terms so that outliers would have a less impact on the adaptation.

Furthermore, when the dimension (d) is very high, HQ-PCA can avoid directly optimizing on the original dimension of data, which can reduce the computation cost. The computation cost of HQ-PCA mainly involves the eigen-decomposition in Step 5 of Algorithm 1. Assuming that $m_r \ll \min(d, n)$, when $d < n$, the computation cost of eigen-decomposition is $o(dm_r^2)$; when $n < d$, the computation cost of eigen-decomposition is $o(nm_r^2)$. It is clear that the computation cost of HQ-PCA mainly depends on m_r. When $n < d$ and dimension of input space (d) is large, the computation cost of HQ-PCA does not depend on the dimension d. For R1-PCA, the cost of the Gram–Schmidt method to maintain orthogonality of U requires $o(dm_r^2 - 1/3m_r^3)$ [29]. In image-based recognition, the dimension of a visual feature vector is usually high $(n < d)$ so that $o(nm_r^2) < o(dm_r^2 - 1/3m_r^3)$. Therefore, HQ-PCA can be smoothly applied to high dimensional data without adding much computation cost.

D. Kernel Extension

Like kernel PCA, the linear projections of HQ-PCA can be directly extended to the nonlinear case by using kernel trick under the half-quadratic optimization framework. The main idea of kernel trick is to map the input data X to another higher dimensional Hilbert space Z through a nonlinear mapping $\phi : X \mapsto Z$ and then perform the linear algorithm in this new feature space [30], [31]. This approach is well suited to algorithms that only need to compute the inner product of data pairs $k(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$ without knowing the nonlinear mapping explicitly. Assuming that the projection matrix $W = U\Phi$, where $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ and $\mathcal{K}$ is the kernel Gram matrix with entity $\mathcal{K}(i, j) = k(x_i, x_j)$, by adopting a similar strategy in [25], we have the following kernelization of model (7):

$$\sum_{i=1}^{n} g(\mathcal{K}_i - \mu - Uv_i) \tag{20}$$

where $\mathcal{K}_i$ indicates the ith column vector of the kernel Gram matrix $\mathcal{K}$. Accordingly, we can derive the so called HQ-KPCA algorithm for robust kernel-based PCA.

IV. Experiments

Here, we verify the robustness of our proposed rotationally invariant HQ-PCA algorithm and compare the performance with two state-of-the-art PCAs: R1-PCA [4] and PCA-L_1 [10]. The parameters of R1-PCA and PCA-L_1 follow the suggestion in [4] and [10], respectively. The convergence condition for R1-PCA, PCA-L_1 and HQ-PCA was set if the difference between norms of projection matrix U in successive iterations was less than 10^{-5} or the maximum number of iterations of 50

[1]Φ-PCA also minimizes the robust objective in (8). However, the $\rho(x)$ in Φ-PCA needs to be a twice-differentiable and convex function so that Φ-PCA can be optimized by Newton's gradient method.

[2]M-Scale PCA yields the eigenvectors corresponding to the $(d-m)$ smallest eigenvalues.

Fig. 1. Cropped facial images and their corresponding noisy images.

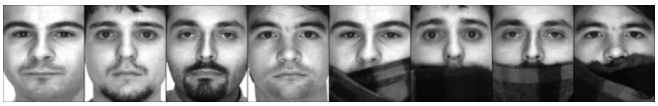


Fig. 2. Cropped facial images and their corresponding occluded images.

was reached [4], [10]. The data mean μ of both R1-PCA and PCA-L_1 is the same as that of PCA. The Huber's M-estimator is used for R1-PCA. As pointed out in [32], the redescending M-estimators are sensitive with respect to the scale parameter σ. This work follows the lines of correntropy [13] and estimates the bandwidth by Silverman's rule [33]

$$(\sigma^t)^2 = 1.06 \times \min\left\{\sigma_E, \frac{R}{1.34}\right\} \times (n)^{-1/5} \qquad (21)$$

where σ_E is the standard deviation of the distance $\left(||(x_i - \mu^t) - U^t(U^t)^T(x_i - \mu^t)||^2\right)$ and R is the interquartile range.

A. Data Sets

Two public face databases, the MNIST handwritten database[3] and the TDT2 Document Database[4] were selected for performance evaluation. Some basic information about four data sets is given here.

1) Yale Face Database: The Yale face database [34] consists of 165 grayscale images of 15 individuals. There are 11 images per subject, with variations of facial expressions or different configurations. Each facial image is in 256 gray scales per pixel and resized to 64×64 pixels aligned by the positions of the two eyes. Fig. 1 shows four facial images (first four images) in Yale database.

2) AR Database: The AR database [35] consists of over 4000 facial images of 126 subjects. For each subject, 26 facial images were taken in two separate sessions. These images are with different facial variations including various facial expressions, illumination variations and occlusion modes. This database is always used for the evaluation of robust face recognition algorithm. Fig. 2 shows four facial images and their corresponding occluded images in AR database.

3) TDT2 Document Database: The TDT2 corpus consists of 11 201 on-topic documents which are classified into 96 semantic categories. We use the top nine categories for our experimental evaluation. Each document is represented as a normalized term-frequency vector, with top 500 words selected according to mutual information. We randomly selected 270, 540, and 900 documents for training (each category has the same number of documents) and the rest were used for testing.

4) MNIST Handwritten Digits Database: The MNIST database has a training set A of 60 000 examples and a test set B of 10 000 examples. The digits were centered in a fixed-size (28×28) and normalized to 1. In our experiment, we use the digits {3, 8, 9} which represent difficult visual discrimination problem [25]. we took the {3, 8, 9} digits in the first 10 000 samples from set A as our training set and those in the first 10 000 from set B as our testing set. A subset with (100, 200, 300) samples per digit from set A was randomly selected for training. The number of samples in testing set was 2993.

[3]http://yann.lecun.com/exdb/mnist/

[4]http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

(a) (b)

Fig. 3. Selected digital images in MNIST Database (a) Images of "3," "8," and "9." (b) Outliers from remaining digits.

B. Face Reconstruction

In the face reconstruction experiment, the average reconstruction error is defined by the average distance between an original unoccluded image and the reconstructed image as follows:

$$e(m) = \frac{1}{n}\sum_{i=1}^{n}\left\|(x_i^{org} - \mu) - \sum_{j=1}^{m} u_j u_j^T (x_i - \mu)\right\| \qquad (22)$$

where x_i^{org} and x_i are the original unoccluded image and the corresponding occluded image in the training set respectively, m is the number of principal components, and μ is the data mean. Eigenvectors of PCA were used as the initial projection of both HQ-PCA and R1-PCA and the sample with the largest L_2-norm [10] was used for that of PCA-L_1. The m_r and m of HQ-PCA (Algorithm 1) were set to 30 and 70, respectively.

1) Artificial Outliers: The outliers were generated by randomly blocking parts of the facial images in the Yale database. In the first experiment, 30 images among the 165 images were randomly selected and occluded with noises consisting of random black and white dots. That is the numbers of outliers and inliers are 30 and 135, respectively. All noises were within a rectangle, size, and position of which were randomly generated. Fig. 1 shows the original unoccluded images and their corresponding noisy images.

Fig. 4(a) shows the average reconstruction errors of different robust PCAs. When the number of principal components is small (less than 10), the average reconstruction errors for different methods are almost the same. However, when the number of principal components is larger than 20, the difference among different methods becomes more apparent and HQ-PCA becomes to perform better than the other methods. In this experiment, PCA-L_1 performs similarly to PCA. This may be due to the fact that PCA-L_1 tries to maximize a L_1-norm dispersion instead of minimizing reconstruction error.

The principal components of PCA are often called eigenfaces. Fig. 5(a) shows the mean faces and eigenfaces of four different methods. The mean faces of four methods seem to be similar, but the eigenfaces are entirely different. Since there are noisy images in the training set, most of eigenfaces in Fig. 5(a) are contaminated by the noise (highlighted by a red rectangle). We can find that eigenfaces of HQ-PCA are less affected by the noise than the other methods. There is no contamination on the first three eigenfaces of HQ-PCA corresponding to the largest three eigenvalues. Fig. 5(b) further shows images in the training set

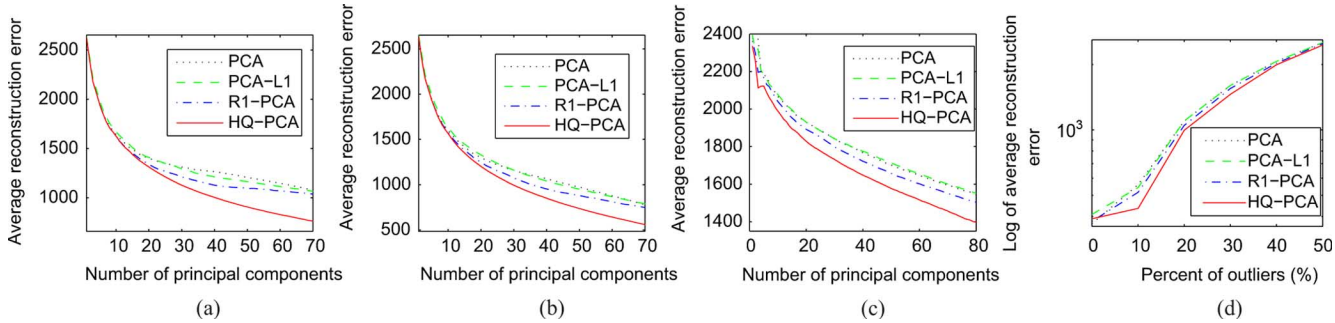


Fig. 4. Average reconstruction errors of different robust PCAs. (a) Data set with occluded images on the Yale database. (b) Data set with dummy images on the Yale database. (c) Data set with occluded images on the AR database. (d) Data set with varying level of outliers on the AR database.

Fig. 5. (a) Mean face and eigenfaces of four different methods. The first column shows the mean face and rest of columns show the eigenfaces. The first row shows mean face and eigenfaces of PCA, the second row PCA-L_1, the third row R1-PCA, and the last row HQ-PCA. (b) Reconstructed images of four different methods. The first column shows Images in training set. The second column shows reconstructed images by PCA. The third column shows PCA-L_1. The fourth column shows R1-PCA. The fifth column shows HQ-PCA.

and the reconstructed images using 70 projection vectors, respectively. Although the reconstructed images by HQ-PCA still have noisy dots, HQ-PCA performs better than other methods.

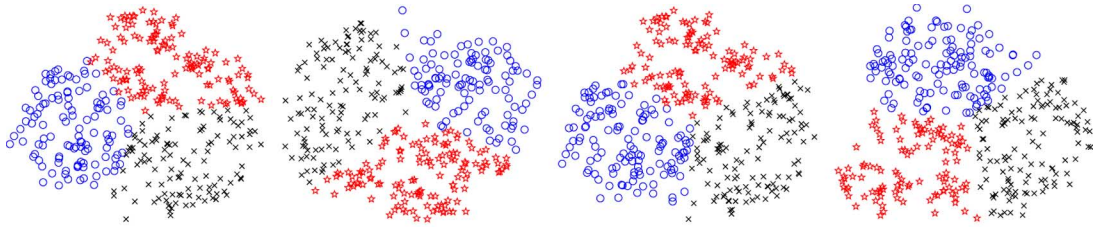
From Fig. 5(b), we can also observe that HQ-PCA seems to preserve some of the facial characteristics outside the rectangular outlier areas not as well as those of PCA. If a sample is an outlier, it will obtain a small weight in the alternate maximum process of HQ-PCA. In the ideal case, the p_i corresponding to the outlier would be zero. That means that outliers will be removed from the training set so that the eigenvectors computed by HQ-PCA contain no information of the outliers. Thus, HQ-PCA could not perfectly reconstruct the unterminated part.

In the second experiment, we added 30 dummy images that consisted of random black and white dots added to the original 165 Yale images, that is, the number of outliers and inliers are 30 and 165, respectively. When computing reconstruction error, 30 dummy images were excluded, and then x_i^{org} and x_i in (22) would be the same as those in this dummy case. Fig. 4(b) shows the average reconstruction errors of four methods with various numbers of principal components. It is clear that HQ-PCA performs the best over the other methods in reconstructing original images.

2) Malicious Occlusion: In this subsection, we made use of occluded faces as outliers to evaluate different PCA methods. The tests were performed on a subset of the AR database. Seventy subjects were selected from the AR database and two frontal face images per subject were used. Then we obtained 140 face images for inliers. First, 60 occluded face images by scarf corresponding to the first 60 subjects in the training set were selected as outliers. Fig. 4(c) shows the average reconstruction errors of different methods. As expected, HQ-PCA outperforms other PCA methods.

We next simulated various levels of outliers, from 0% to 50%. We selected two occluded face images per subject to form the outlier dataset. Then, 15, 35, 60, 93, and 140 face images from the outlier dataset were added to the set of inliers, corresponding to 10%, 20%, 30%, 40%, and 50% of outliers contained in the training set, respectively. The number of inliers is always set to 140. Fig. 4(d) shows average reconstruction errors of all four methods under varying levels of outliers. We see that the differences in average reconstruction errors between different methods becomes insignificant as the percent of outliers increases. Note that the outliers are significantly far away from the rest of the data points [3]–[5]. When the percent of outliers is larger than 40%, the occluded images by scarf become another group and could not be treated as outliers any more. Hence, the average reconstruction errors of different methods are close.

Fig. 4(d) also shows the reconstruction errors of different methods in the absence of any outliers (i.e., when the percentage is 0). We observe that PCA achieves the lowest reconstruction error, and the following are R1-PCA, HQ-PCA, and PCA-L_1.



(a) (b) (c) (d)

Fig. 6. Two-dimensional subspaces of four methods. The circle points belong to digit "3"; the pentagram points belong to digit "8"; the cross points belong to digit "9." (a) PCA. (b) PCA-L_1. (c) R1-PCA. (d) HQ-PCA.

TABLE I
CLUSTERING ACCURACY OF K-MEANS ON SUBSPACES

	'3'	'8'	'9'
PCA + K-means	0.76	0.57	0.65
PCA-L_1 + K-means	0.78	0.57	0.66
R1-PCA + K-means	0.80	0.65	0.71
HQ-PCA + K-means	**0.81**	**0.68**	**0.78**

This suggests that, when there are no outliers, PCA achieves the lowest reconstruction error and no robust PCA can outperform PCA. However, when there are 10% outliers, the reconstruction error of PCA increases rapidly, while the other PCA methods, especially the proposed HQ-PCA, are less affected.

C. Clustering

Theoretical analysis and experimental results [4], [36] show that PCA relates to K-means in a relaxed solution given by principal components and PCAs can be used as a preprocessing step to further improve the clustering accuracy of K-means. We performed experiments to show that HQ-PCA's subspace also is better than the other PCAs' subspaces for clustering when outliers exist. The K-means was initialized with the same starting vectors for all methods, and experiments were performed on a subset of the MNIST database. Fig. 3(a) shows the images of three digits in the training set. To simulate outliers, we randomly selected 60 samples from the remaining digits in the first 10 000 samples from set A. The numbers of outliers and inliers are 60 and 300, respectively. Fig. 3(b) shows the outliers from the remaining digits.

Fig. 6 shows the clustering results on the subspace spanned by the eigenvectors corresponding to the two largest eigen values. The circle points belong to digit "3"; The pentagram points belong to digit "8"; the cross points belong to digit "9." There are obviously three clusters in all of the PCAs' subspace. However, the data points overlap each other on the boundaries of three clustering in subspaces of PCA, PCA-L_1, and R1-PCA. It is more evident that the three clusterings are well separated in HQ-PCA's subspace.

To quantitatively evaluate the robustness of different methods, we compared the clustering accuracy of K-means algorithm on four subspaces. Clustering accuracy was computed by using the known class labels. Table I tabulates the cluster accuracy for each digit. The clustering accuracy of PCA and PCA-L_1 are very close, and both R1-PCA and HQ-PCA can significantly improve the clustering accuracy. HQ-PCA achieves the highest clustering accuracy on three digits. The results indicate that MaxEnt-PCA's subspace outperforms other methods' subspaces for clustering.

D. Dimension Reduction

Here, we evaluate the robustness of different PCA methods for dimension reduction. Since automatic selection of principal components is still an ongoing research problem in PCA, we search the optimal dimension of principal components of PCA on which PCA performs the best by following the approach in [37]. The learned dimension was then used for all PCA methods. Therefore, the dimensions of principal subspaces are 127, 246, and 181 for three different datasets in TDT2, respectively, and the dimensions of principal subspaces are 64, 52, and 66 for three different datasets in MNIST, respectively. Classification was then performed in the reduced space. Note that, for HQ-PCA, the value of m_r was the same as that of m.

1) Artificial Outliers: In computer vision, one assumes that outliers are significantly far away from the rest of the data points [4]. Here, we used nonnormalized samples as outliers, e.g., the outlier's norm is significantly larger than 1. We randomly selected 2% of samples in the training set as outliers and the remaining 98% of samples as inliers. To eliminate statistical deviations, all experimental results were reported over 20 random trials. Random orthonormal matrices were used for the initial projection of both HQ-PCA and R1-PCA, and the sample with the largest L_2-norm [10] was used for that of PCA-L_1. The *nearest center classifier* was finally used as the evaluation metric. In addition, the center of a class C for all methods was calculated as

$$\bar{x}_c = \frac{1}{\sum_{x_k \in C} w(x_k)} \sum_{x_k \in C} w(x_k) U^T x_k \tag{23}$$

$$w(x_i) = \exp\left(- \left\| \frac{(x_i - \mu) - UU^T(x_i - \mu)}{\sigma^2} \right\|^2 \right) \tag{24}$$

where σ is calculated by (21).

Tables II and III show the classification rates of different robust methods on the MNIST database and the TDT2 database, respectively. When outliers occur, PCA learns a bias subspace so that its correct rate significantly decreases and its deviation is large. Although PCA-L_1 can achieve higher classification rates than PCA, there are still large declines compared with HQ-PCA. It is clear that HQ-PCA achieves the highest correct rate and the deviation of HQ-PCA is small. HQ-PCA performs better than the two other L_1-norm PCAs when outliers occur.

Fig. 7(a) shows the classification rates of HQ-PCA under two conditions: updating data mean μ and fixing data mean μ. This experiment was performed on MNIST database and 2% of the samples were selected from 100×3 training samples as outliers. The outliers were generated by $x = ax_{\text{org}}$, where x_{org} is a normalized data and a is used to control the magnitude of

TABLE II
COMPARISON OF PCA ALGORITHMS ON MNIST DATABASE: AVERAGE CORRECT RATE ± STANDARD DEVIATION

Number	PCA	R1-PCA	PCA-L_1	HQ-PCA
100×3	37.5± 6.0	37.7± 6.0	40.8±9.2	**88.8**±0.2
200×3	42.8±11.3	42.9±11.4	74.6±4.0	**90.0**±0.1
300×3	47.3±13.6	47.4±13.6	79.0±1.5	**89.7**±0.1

TABLE III
COMPARISON OF PCA ALGORITHMS ON TDT2 DATABASE: AVERAGE CORRECT RATE ± STANDARD DEVIATION

Number	PCA	R1-PCA	PCA-L_1	HQ-PCA
30×9	63.6±8.2	64.0±8.4	69.8±7.5	**86.7**±0.4
60×9	71.7±7.5	70.9±7.8	80.5±5.4	**90.0**±0.1
100×9	81.5±4.3	81.2±4.4	89.0±0.6	**90.4**±0.1

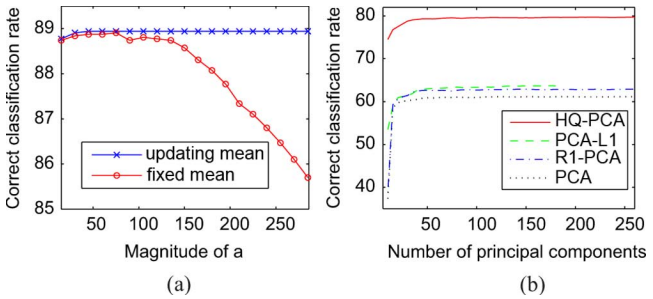


Fig. 7. (a) Correct classification rates of HQ-PCA under two conditions: updating data mean and fixing data mean. The axis presents the magnitude of outliers. (b) Correct classification rates under different numbers of principal components on the TDT2 dataset.

TABLE IV
COMPARISON OF PCA ALGORITHMS WITH REAL-WORLD OUTLIERS: AVERAGE CORRECT RATE ± STANDARD DEVIATION

Number	PCA	R1-PCA	PCA-L_1	HQ-PCA
100×3	72.2±0.9	72.4±2.4	71.8±1.0	**74.1**±0.3
200×3	71.3±0.6	72.9±0.6	71.4±1.4	**74.6**±0.6
300×3	71.9±0.4	71.9±0.5	70.9±1.2	**73.0**±0.5

outliers. We can find that, if we fixed the data mean (weights of data are equal), the accuracy of HQ-PCA will drop rapidly when a increases. However, if we update the data mean according to (15), HQ-PCA can achieve a stable correct rate. Fig. 7(b) shows correct classification rates under different numbers of principal components on the TDT2 dataset. We can observe that the variation of classification rates is small as the number of principal components increases.

2) Real Outliers: To simulate outliers, 60, 120, and 180 samples were randomly selected from the remaining digits in the first 10 000 samples from set A, corresponding to 300, 600, and 900 inliers, respectively. The *nearest center classifier* [25] was finally used as the evaluation metric. All experiments were averaged over 20 random trials.

Table IV tabulates the correct classification rates and standard deviations of different PCA methods. HQ-PCA still achieves the highest classification rates on all three training sets. Although PCA-L_1 can work well on the artificial outliers, PCA-L_1 performs no better than PCA on the real outliers.

Table V further shows the correct classification rates and standard deviations of two kernel PCA methods. The Laplacian kernel (i.e., $k(x_i, x_j) = \exp(-\sqrt{r}\|x_i - x_j\|_{L_1})$) was used in this experiment. The r was set as $1/d$ (d is the feature dimension). Since there are no kernel extensions of PCA-L_1 and R1-PCA, we only compare kernel PCA and HQ-KPCA. HQ-PCA can also improve the robustness in the kernel space.

TABLE V
COMPARISON OF KERNEL PCA ALGORITHMS WITH REAL-WORLD OUTLIERS: AVERAGE CORRECT RATE ± STANDARD DEVIATION

Number	KPCA	HQ-KPCA
100×3	63.5±1.3	68.5±1.2
200×3	66.7±0.6	72.6±0.7
300×3	72.1±0.5	73.4±0.5

E. Parameter Selection

There are two parameters (σ and m_r in Algorithm 1) affecting the performance of HQ-PCA. For demonstration, Fig. 8 shows the average reconstruction errors as functions of values of σ and m_r. The reconstruction errors were calculated with 70 principal components. The experimental setting in Fig. 8(a)–(c) is the same as that of the first experiment in Section IV-B1), and the experimental setting in Fig. 8(d) is the same as that of the first experiment in Section IV-B2).

The average reconstruction errors as a function of the values of m_r (Algorithm 1) are given in Fig. 8(a). We can find that different values of m_r will lead to different reconstruction errors. However, compared with the reconstruction error of PCA in Fig. 4(a) that is larger than 1000 with 70 principal components, the variation of reconstruction errors of HQ-PCA is very small. Hence, HQ-PCA is less sensitive to the choice of m_r, and can achieve smaller reconstruction errors over a large range of m_r. We can detect outliers under a small value of m_r and then obtain additional principal components by eigen-decomposition of $X_c^{t+1}P^{t+1}\left(X_c^{t+1}\right)^T$.

The kernel size σ is an important parameter which controls all robust properties of entropy [13]. A well-tuned kernel size value can effectively eliminate the effect of outliers and noise. We further investigated the kernel size σ as follows:

$$\left(\sigma^t\right)^2 = \frac{1}{sn}\sum_{i=1}^{n}\left\|\left(x_i - \mu^t\right) - U^t\left(U^t\right)^T\left(x_i - \mu^t\right)\right\|^2 \quad (25)$$

where s is a scale factor to control the value of the kernel size. Fig. 8(b) compares the average reconstruction errors as the scale factor s increases from 0.7 to 2. We see that the average reconstruction error increases rapidly when s is decreased from 1.1 to 0.9. However, compared with the improvements of other methods in Fig. 4(a) (more than 200), the variation of HQ-PCA is less than 20 and hence is relatively small. HQ-PCA can achieve a good robustness over a large range of s.

Fig. 8(c) shows the convergence curves of HQ-PCA. Each convergence curve shows a variation of the J_{HQ} under a fixed parameter. It is clear that HQ-PCA increases the objective function step-by-step and converges rapidly (less than ten iterations). The rapid convergence of HQ-PCA may be due to the fact that maximization (or minimization) using HQ optimization can speed up computation [38].

Fig. 8(d) shows the convergent curves of HQ-PCA as the size of the training set increases from 130 to 620. We selected 1, 2, 4, and 8 images per subject and 60 occluded images as outliers. We can observe that the value of the J_{HQ} increases rapidly over

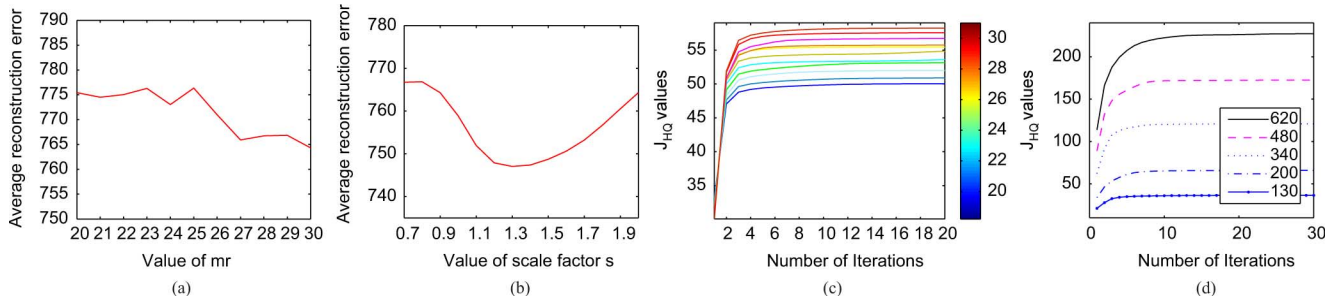


Fig. 8. (a) Reconstruction errors under different values of m_r in Algorithm 1. (b) Reconstruction errors under different values of s in kernel size σ estimated by (25). (c) Convergent curves of J_{HQ} under different values of m_r. A special color curve is corresponding to a value of m_r. (d) Convergent curves of J_{HQ} under different values of the training size. The number in the legend is the size of training set.

the first several iterations. Although HQ-PCA would take more iterations to converge as the size of training set increases, it still converges less than ten iterations.

V. Summary

This paper proposes a rotationally invariant PCA algorithm by replacing MSE criterion with MCC. The proposed objective function is robust to outliers and can be efficiently optimized by the half-quadratic optimization technique. At each iteration, the complex correntropy objective can thereby be reduced to a quadratic optimization problem. The proposed method is rotation-invariant and can correctly update the data mean. Its principal components are the principal eigenvectors of a robust covariance matrix corresponding to the largest eigenvalues. Experimental results illustrate that the proposed method can outperform the other robust PCAs which are based on L_1-norm.

Appendix
Information Potential

To better understand the relationship among correntropy, information potential, and Renyi's quadratic entropy, here we briefly review some derivations in information theory learning.

The Renyi's quadratic entropy of a random variable X with probability density function (pdf) $f_X(x)$ is defined as

$$H(X) = -\log \int f_X^2(x)dx. \tag{26}$$

If Parzen window method is used to estimate the pdf, $f_X(x)$ can be obtained by

$$\widehat{f}_{X;\sigma}(x) = \frac{1}{n}\sum_{i=1}^{n} G\left(x - x_i, \sigma^2\right) \tag{27}$$

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

$$G\left(x - x_i, \sigma^2\right) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right). \tag{28}$$

By substituting $f_X(x)$ in (26) with (27), the estimate of entropy by Parzen window method can be formulated as

$$H(X) = -\log\left(\mathrm{IP}(X)\right) \tag{29}$$

$$\mathrm{IP}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_j - x_i, \sigma^2) \tag{30}$$

where $\mathrm{IP}(X)$ stands for the information potential (IP).

According to IP in (30), correntropy is defined as a generalized similarity measure between two arbitrary random variables A and B. It is directly related to Renyi's quadratic entropy in which Parzen windowing method is used to estimate the data's probability distribution [13], [22].

Acknowledgment

The authors would like to greatly thank the associate editor and the reviewers for their valuable comments and advice.

References

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh and Dublin Philosop. Mag. J. Sci.*, vol. 2, pp. 559–572, 1901.

[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educational Psychol.*, vol. 24, pp. 417–441, 1933.

[3] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, and K. Cohen, "Robust principal component analysis for functional data," *Test*, vol. 8, no. 1, pp. 1–73, 1999.

[4] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. Int. Conf. Mach. Learning*, Jun. 2006, pp. 281–288.

[5] R. A. Maronna, "Principal components and orthogonal regression based on robust scales," *Technometrics*, vol. 47, pp. 264–273, 2005.

[6] A. Baccini, P. Besse, and A. Falguerolles, "A L_1-norm PCA and a heuristic approach," *Ordinal Symbol. Data Anal.*, pp. 359–368, 1996.

[7] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. Comput. Vis. Pattern Recogn. Conf.*, 2005, pp. 739–746.

[8] R. Dahyot, P. Charbonnier, and F. Heitz, "Robust visual recognition of colour images," in *Proc. Comput. Vis. Pattern Recogn.*, 2000, pp. 1685–1690.

[9] A. Ng, "Feature selection, L_1 versus L_2 regularization and rotational invariance," in *Proc. Int. Conf. Mach. Learning*, 2004, pp. 78–86.

[10] N. Kwak, "Principal component analysis based on L_1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1677, Sep. 2008.

[11] J. Iglesias, M. d. Bruijne, M. Loog, F. Lauze, and M. Nielsen, "A family of principal component analyses for dealing with outliers," *Med. Image Comput. Comput.-Assisted Intervention*, pp. 178–185, 2007.

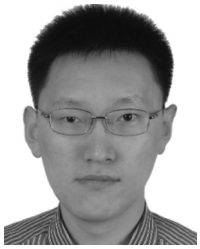
[12] R. Subbarao and P. Meer, "Subspace estimation using projection based m-estimators over Grassmann manifolds," in *Proc. ECCV*, 2006, pp. 301–312.

[13] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. , pp. 5286–5298, 2007.

[14] P. Huber, *Robust Statistics*. New York: Wiley, 1981.

[15] R. He, B.-G. Hu, X.-T. Yuan, and W.-S. Zheng, "Principal component analysis based on non-parametric maximum entropy," *Neurocomputing*, vol. 73, no. 10-12, pp. 1840–1852, 2010.

[16] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, vol. 1, Blind-Source Separation, pp. 265–319.
[17] X. D. Xie and K. M. Lam, "Gabor-based kernel PCA with doubly nonlinear mapping for face recognition with a single face image," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2481–2492, Sep. 2006.
[18] T. J. Chin and D. Suter, "Incremental kernel principal component analysis," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1662–1674, Jun. 2007.
[19] D. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[20] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ., 1996.
[21] Q. Ke and T. Kanade, Robust Subspace Computation Using L_1 Norm [Online]. Available: http://citeseer.ist.psu.edu/ke03robust.html
[22] I. Santamaria, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: Definition, properties and application to blind equalization," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2187–2197, Jun. 2006.
[23] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[24] R. Rockfellar, *Convex Analysis*. Princeton, NJ: Princeton Univ., 1970.
[25] X. T. Yuan and B. G. Hu, "Robust feature extraction via information theoretic learning," in *Proc. Int. Conf. Mach. Learning*, 2009, pp. 1–8.
[26] R. He, B.-G. Hu, W.-S. Zheng, and Y. Guo, "Two-stage sparse representation for robust recognition on large-scale database," in *Proc. AAAI*, 2010, pp. 1–6.
[27] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, 2010.
[28] C. Bishop, *Patten Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.
[29] G. Stewart, *Matrix Algorithms Volume I: Basic Decompositions*. Philadelphia, PA: SIAM, 1998.
[30] S. Xiang, F. Nie, C. Zhang, and C. Zhang, "Interactive natural image segmentation via spline regression," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1623–1632, Jul. 2009.
[31] S. Xiang, F. Nie, and C. Zhang, "Semi-supervised classification via local spline regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2039–2053, Nov. 2010.
[32] I. Mizera and C. Muller, "Breakdown points of Cauchy regression-scale estimators," *Stat. Probabil. Lett.*, vol. 57, pp. 79–89, 2002.
[33] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman and Hall, 1986.
[34] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[35] A. Martinez and R. Benavente, The AR Face Database, CVC 1998.
[36] C. Ding and X. He, "K-means clustering and principal component analysis," in *Proc. Int. Conf. Mach. Learning*, 2004, pp. 225–232.
[37] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.
[38] M. Nikolova and M. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J.*, vol. 27, no. 3, pp. 937–966, 2005.

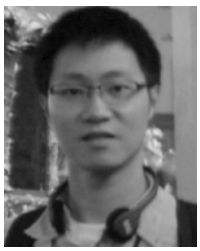


Ran He received the B.S. degree in computer science from the Dalian University of Technology of China, Dalian, China, in 2001, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China. His research interests include information theoretic learning and computer vision.

Bao-Gang Hu (M'94–SM'99) received the M.Sc. degree from the University of Science and Technology, Beijing, China, in 1983, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 1993, both in mechanical engineering.

From 1994 to 1997, he was a Research Engineer and Senior Research Engineer with C-CORE, Memorial University of Newfoundland, Canada. Currently, he is a Professor with the 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, and plant growth modeling.



Wei-Shi Zheng (M'08) received the Ph.D. degree in applied mathematics from Sun Yat-Sen University, Guangzhou, China, 2008.

After that, he has been a Postdoctoral Researcher on the European SAMURAI Research Project at Queen Mary University of London, London, U.K. He has joined Sun Yat-sen University, Guangzhou, China, under the one-hundred-people program of Sun Yat-sen University in 2011. His current research interests are in object association and categorization for visual surveillance. He is also interested in feature extraction, kernel methods in machine learning, transfer learning, and face image analysis.

Xiang-Wei Kong received the B.E. and M.Sc. degrees from Harbin Shipbuilding Engineering Institute, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree from Dalian University of Technology, Dalian, China, in 2003.

She is a Professor with the Department of Electronic and Information Engineering and vice-director of the Information Security Research Center, Dalian University of Technology, Dalian, China. She is also the vice-director of the Multimedia Security Session of the Chinese Institute of Electronics. Her research interests include multimedia security and forensics, digital image processing, and pattern recognition.