Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

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

Generalized mean for robust principal component analysis

Jiyong Oh, Nojun Kwak *

Graduate School of Convergence Science and Technology, AICT, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

ARTICLE INFO

Article history:
Received 11 August 2014
Received in revised form
9 December 2015
Accepted 4 January 2016

Keywords:
Generalized mean
Principal component analysis
Robust PCA
Dimensionality reduction

ABSTRACT

In this paper, we propose a robust principal component analysis (PCA) to overcome the problem that PCA is prone to outliers included in the training set. Different from the other alternatives which commonly replace L_2-norm by other distance measures, the proposed method alleviates the negative effect of outliers using the characteristic of the generalized mean keeping the use of the Euclidean distance. The optimization problem based on the generalized mean is solved by a novel method. We also present a generalized sample mean, which is a generalization of the sample mean, to estimate a robust mean in the presence of outliers. The proposed method shows better or equivalent performance than the conventional PCAs in various problems such as face reconstruction, clustering, and object categorization.

1. Introduction

Dimensionality reduction [1] is a classical problem in pattern recognition and machine learning societies, and numerous methods have been proposed to reduce the data dimensionality. Principal component analysis (PCA) [2] is one of the most popular unsupervised dimensionality reduction methods which tries to find a subspace where the average reconstruction error of the training data is minimized. It is useful in representation of input data in a low dimensional space and it has been successfully applied to face recognition [3,4], visual tracking [5], clustering [6,7], and so on.

When automatically collecting a large data set, outliers may be contained in the collected data since it is very difficult to examine whether each sample of data is outlier or not [8]. It is well known that, in this case, the conventional PCA is sensitive to outliers because it minimizes the reconstruction errors of training data in terms of the mean squared error and a few outliers with large errors dominate the objective function. This problem has been addressed in many studies [8–16]. Among them, some studies utilized L_1-norm instead of L_2-norm in the formulation of optimization problem to improve the robustness of PCA against outliers [9–11]. In [9], the cost function for optimization was constructed based on L_1-norm and a convex programming was employed to solve the problem. R_1-PCA [10] was presented to obtain a solution with the rotational invariance, which is a fundamental desirable property for learning algorithms [17]. In [11], PCA-L_1 was proposed, which maximizes an L_1 dispersion in the reduced space and an extension of PCA-L_1 using L_p-norm with arbitrary p was also proposed in [14]. Other method utilizing L_p-norm was also presented in [15]. On the other hand, some of robust PCAs were recently developed using information theoretic measures [12,13]. He et al. [12] proposed MaxEnt-PCA which finds a subspace where Renyi's quadratic entropy [18] is maximized. Renyi's entropy was estimated by a non-parametric Parzen window technique. In [13], HQ-PCA was developed based on the maximum correntropy criterion [19].

In this paper, we propose a new robust PCA method based on the power mean or the generalized mean [20], which can become the arithmetic, geometric, and harmonic means depending on the value of its parameter. The proposed method, *PCA-GM*, is a generalization of the conventional PCA by replacing the arithmetic mean with the generalized mean. The proposed method can effectively prevent outliers from dominating objective function by controlling the parameter in the generalized mean. Moreover, it is rotational invariant because it still uses the Euclidean distance as the distance measure between data samples. In doing so, we also propose a generalized sample mean, which is an enhancement of the conventional algebraic sample mean against outliers to address the problem that the sample mean is easily affected by outliers. It is used in the proposed PCA-GM instead of the arithmetic mean. The optimization problems based on the generalized mean are efficiently solved using a mathematical property of the generalized mean. Recently, Candés et al. proposed a robust PCA [21], which is sometimes referred to as RPCA in the literature, where data matrix is tried to be represented as a sum of a low rank matrix, which corresponds to reconstructions of data, and a sparse matrix, which corresponds to reconstruction errors different from the methods mentioned above. It can model pixel-wise noise effectively using the sparse matrix, thus it has been known that RPCA is useful in the applications such as background modeling from surveillance video and removing shadows and specularities from face images [21] by

* Corresponding author. Tel.: +82 31 888 9166; fax: +82 31 888 9148.
E-mail addresses: yong97@snu.ac.kr (J. Oh), nojunk@snu.ac.kr (N. Kwak).

http://dx.doi.org/10.1016/j.patcog.2016.01.002
0031-3203/

using each element in the reconstruction error vector (the column of the sparse matrix). On the other hand, in this paper, we will utilize distance metric in removing the effect of outliers like the previously mentioned methods, and an entire sample is considered as an outlier if it has a large norm of the reconstruction error vector.

The remainder of this paper is organized as follows. Section 2 briefly introduces PCA and the state-of-the-art robust PCAs. The proposed method is described in Section 3. It is demonstrated in Section 4 that the proposed method gives better performances in face reconstruction and clustering problems than other variants of PCA. Finally, Section 5 concludes this paper.

2. PCA and robust PCAs

Let us consider a training set of N n-dimensional samples $\{lbracesupsub\mathbf{x}_i\}_{i=1}^N$. Assuming that the samples have zero-mean, PCA is to find an orthonormal projection matrix $\mathbf{W} \in \mathbb{R}^{n\times m}$ $(m \ll n)$ by which the projected samples $\left\{\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i\right\}_{i=1}^N$ have the maximum variance in the reduce space. It is formulated as follows:

$$\mathbf{W}_{PCA} = \arg\max_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^T\mathbf{S}\mathbf{W}),$$

where $\mathbf{S} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$ is a sample covariance matrix and $\mathrm{tr}(\mathbf{A})$ is the trace of a square matrix $\mathbf{A}$. The projection matrix $\mathbf{W}_{PCA}$ can be also found from the viewpoint of projection errors, i.e., it minimizes the average of the squared projection errors or reconstruction errors. Mathematically, it is represented as the optimization problem minimizing the following cost function:

$$J_{L_2}(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^N \|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\|_2^2,$$

where $\|\mathbf{x}\|_2$ is the L_2-norm of a vector $\mathbf{x}$. The two optimization problems are equivalent and easily solved by obtaining the m eigenvectors associated with the m largest eigenvalues of $\mathbf{S}$. Although PCA is simple and powerful, it is prone to outliers [8,9] because $J_{L_2}(\mathbf{W})$ is based on the mean squared reconstruction error. To learn a subspace robust to outliers, Ke and Kanade [9] proposed to minimize an L_1-norm based objective function as follows:

$$J_{L_1}(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^N \|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\|_1,$$

where $\|\mathbf{x}\|_1$ is the L_1-norm of a vector $\mathbf{x}$. They also present an iterative method to obtain the solution for minimizing $J_{L_1}(\mathbf{W})$.

Although L_1-PCA minimizing $J_{L_1}(\mathbf{W})$ can relieve the negative effect of outliers, it is not invariant to rotations. In [10], Ding et al. proposed R_1-PCA, which is rotational invariant, at the same time is robust to outliers. It is to minimize the following objective function:

$$J_{R_1}(\mathbf{W}) = \sum_{i=1}^N \rho\left(\sqrt{\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\mathbf{W}\mathbf{W}^T\mathbf{x}_i}\right),$$

where $\rho(\cdot)$ is a generic loss function and the Cauchy function or Huber's M-estimator [22] was used for $\rho(\cdot)$ in [10]. Huber's M-estimator $\rho_H(s)$ is defined as

$$\rho_H(s) = \begin{cases} s^2 & \text{if } |s| \le c, \\ 2c|s| - c^2 & \text{otherwise} \end{cases} \quad (1)$$

where c is the cutoff parameter that controls the regularization effect of weights in a weighted covariance matrix. Note that $\rho_H(s)$ becomes a quadratic or a linear function of $|s|$ depending on the value of s. The solution for minimizing $J_{R_1}(\mathbf{W})$ was obtained by performing a subspace iteration algorithm [23].

On the other hand, PCA-L_1 was developed in [11] motivated by the duality between maximizing variance and minimizing reconstruction error. It maximizes an L_1 dispersion among the projected samples, $\sum_{i=1}^N \|\mathbf{W}^T\mathbf{x}_i\|_1$. A novel and efficient method for maximizing the L_1 dispersion was also presented in [11]. The method allows PCA-L_1 to be performed by much less computational effort than R_1-PCA.

HQ-PCA is formulated based on the maximum correntropy criterion in terms of information theoretic learning. Without the zero-mean assumption, which is necessary in other variants of PCA, HQ-PCA maximizes the correntropy estimated between a set of training samples $\{\mathbf{x}_i\}_{i=1}^N$ and the set of their reconstructed samples $\{\mathbf{W}\mathbf{y}_i + \mathbf{m}\}_{i=1}^N$, where $\mathbf{m}$ is a data mean. Mathematically, HQ-PCA tries to maximize the following objective function:

$$\arg\max_{\mathbf{W},\mathbf{m}} \sum_{i=1}^N g\left(\sqrt{\overline{\mathbf{x}}_i^T\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_i^T\mathbf{W}\mathbf{W}^T\overline{\mathbf{x}}_i}\right), \quad (2)$$

where $g(x) = \exp(-x^2/2\sigma^2)$ is the Gaussian kernel and $\overline{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}$. Note that HQ-PCA finds a data mean as well as a projection matrix. Using the Welsch M-estimator $\rho_W(x) = 1 - g(x)$, HQ-PCA is regarded as a robust M-estimator formulation because it is equivalent to finding $\mathbf{W}_H$ and $\mathbf{m}_H$ that minimize the following objective function:

$$J_{HQ}(\mathbf{W}, \mathbf{m}) = \sum_{i=1}^N \rho_W\left(\sqrt{\overline{\mathbf{x}}_i^T\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_i^T\mathbf{W}\mathbf{W}^T\overline{\mathbf{x}}_i}\right). \quad (3)$$

In [13], the optimization problem in (2) was effectively solved in the half-quadratic optimization framework, which is often used to address nonlinear optimization problems in information theoretic learning.

3. Robust principal component analysis based on generalized mean

3.1. Generalized mean for positive numbers

For a $p \ne 0$, the generalized mean or power mean $\mathcal{M}_p$ of $\{a_i > 0,\ i = 1, \dots, N\}$ [20] is defined as

$$\mathcal{M}_p\{a_1, \dots, a_N\} = \left(\frac{1}{N}\sum_{i=1}^N a_i^p\right)^{1/p}.$$

The arithmetic mean, the geometric mean, and the harmonic mean are special cases of the generalized mean when $p = 1, p \to 0$, and $p = -1$, respectively. Furthermore, the maximum and the minimum values of the numbers can also be obtained from the generalized mean by making $p \to \infty$ and $p \to -\infty$, respectively. Note that as p decreases (increases), the generalized mean is more affected by the smaller (larger) numbers than the larger (smaller) ones, i.e., controlling p makes it possible to adjust the contribution of each number to the generalized mean. This characteristic is useful in the situation where data samples should be differently handled according to their importance, for example, when outliers are contained in the training set.

In [24], it was shown that the generalized mean of a set of positive numbers can be expressed by a nonnegative linear combination of the elements in the set and, in this paper, it is further simplified as follows:

$$\sum_{i=1}^K a_i^p = b_1 a_1 + \dots + b_K a_K$$
$$b_i = a_i^{p-1}, \quad i = 1, \dots, K. \quad (4)$$

Note that each weight b_i has the same value of 1 if $p = 1$, where the generalized mean becomes the arithmetic mean. It is also noted that, if p is less than one, the weight b_i increases as a_i decreases. This means that, when $p < 1$, the generalized mean is more influenced by the small numbers in $\{a_i\}_{i=1}^K$, and the extent of the influence increases as p decreases. This equation plays an