Pattern Recognition Letters 34 (2013) 1037–1045

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Feature extraction based on Lp-norm generalized principal component analysis

Zhizheng Liang [a,*], Shixiong Xia [a], Yong Zhou [a], Lei Zhang [a], Youfu Li [b]

[a] Dept. of Computer Science, China University of Mining and Technology, China
[b] Dept. of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 11 August 2012
Available online 15 February 2013

Communicated by S. Sarkar

Keywords:
Generalized PCA
Lp-norm
Convex function
Face images
UCI data sets

ABSTRACT

In this paper, we propose Lp-norm generalized principal component analysis (PCA) by maximizing a class of convex objective functions. The successive linearization technique is used to solve the proposed optimization model. It is interesting to note that the closed-form solution of the subproblem in the algorithm can be achieved at each iteration. Meanwhile, we theoretically prove the convergence of the proposed method under proper conditions. It is observed that sparse or non-sparse projection vectors can be obtained due to the applications of the Lp norm. In addition, one deflation scheme is also utilized to obtain many projection vectors. Finally, a series of experiments on face images and UCI data sets are carried out to demonstrate the effectiveness of the proposed method.

1. Introduction

Principal component analysis (PCA) is one of the most widely used statistical techniques for dimensionality reduction. It is known that standard PCA constructs the optimal subspace approximation of data in terms of the quadratic error criterion. Due to its least squares formulation, PCA is highly sensitive to contaminated data. To alleviate the effect of contaminated data, many robust PCA methods (Black and Jepson, 1996; Torre and Black, 2001; Torre and Black, 2003; Aanas et al., 2002; Ke and Kanade, 2005; Ding et al., 2006; Liang and Li, 2010) have been proposed during past several years. In order to keep the rotational invariance property, Ding et al. (2006) devised a robust covariance matrix which can soften the effect of outliers. Black and Jepson (1996) replaced the quadratic error norm with a robust one in terms of an m-estimator technique. In (Kwak, 2008), a simple yet effective algorithm based on L1-norm optimization techniques is proposed to deal with contaminated data and the solution has a rotational invariance property. Based on this, Pang and Yuan (2010) applied the L1 norm to handle the problem of graph embedding. In addition, Pang et al. (2010) also generalized Kwak's method to deal with tensor data (Yang and Yang, 2002; Yang et al., 2004).

Some robust methods mentioned above try to overcome the drawbacks of standard PCA in dealing with contaminated data. However, the solution from these algorithms is generally not sparse. It is noted that sparsity is very important in most cases as it provides physical interpretations and improves the generalization performance in learning algorithms. As a result, some sparse PCA algorithms (Wright et al., 2009; Luss and Teboulle, 2011) have been proposed. Most sparse PCA algorithms involve solving a hard combinational problem. For example, Zou et al. (2006) proposed sparse principal component analysis (SPCA) based on the elastic net. Luss and Teboulle (2011) derived a convex relaxation for cardinality constraints based on a representation of the L1 norm. Later, Sriperumbudur et al. (2011) further proposed the sparse generalized eigenvalue problem. In fact, these sparse methods have been successfully used to extract sparse and interpretable components from the given raw data.

Note that the sparse PCA algorithms above generally impose L0 norm or L1 norm constraints on projection vectors. To the best of our knowledge, there is no study on imposing the general norm for PCA and its variants. To this end, we propose the Lp norm constraint for generalized principal component analysis (GPCA) in this paper. We refer to this new model as Lp norm GPCA. This new model not only suppresses the contaminated data by choosing robust functions, but also provides sparse or non-sparse solutions by applying the Lp norm. As a result, this offers a general scheme for PCA and its some variants. To solve the optimization problem, we use the successive linearization technique. We also theoretically show the convergence of the iterative algorithm. Experiments on face images and the datasets from the UCI machine learning repository are done to demonstrate the effectiveness of the proposed method.

2. Related work

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \Re^{d\times n}$ be a collection of n data points. Data points can be reduced to $\boldsymbol{y}_i = \boldsymbol{G}^T\boldsymbol{x}_i$ by a projection matrix

* Corresponding author. Tel./fax: +86 0516 83995918.
E-mail address: zzliang76@yahoo.com.cn (Z. Liang).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.01.030

$\boldsymbol{G} \in \Re^{d\times m}$. Without loss of generality, $\{\boldsymbol{x}_i\}_{i=1}^n$ are assumed to have zero mean. In fact, this is easily obtained by a translation of data.

2.1. Classical PCA

In PCA, the optimal $m(< d)$ dimensional linear subspace can be obtained by minimizing the following error function:

$$\min_{\boldsymbol{G},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{G}\boldsymbol{V}\|_F^2 := \sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_i^j - \sum_{k=1}^{m} g_k^j v_i^k\right)^2, \tag{1}$$

where $\boldsymbol{G} \in \Re^{d\times m}$ is a projection matrix whose jth element in the kth column is g_k^j, $\boldsymbol{V} \in \Re^{m\times n}$ is a coefficient matrix whose kth component in the ith column is v_i^k, x_i^j denotes the jth component of $\boldsymbol{x}_i$, and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

2.2. L1PCA

Kwak (2008) proposed to maximize the L1 dispersion using the L1-norm in the projection space instead of minimizing the L1 error function in the original d-dimensional input space. The optimization problem is

$$\begin{aligned} &\boldsymbol{w}^* = \arg\max \sum_{i=1}^{n} |\boldsymbol{w}^T\boldsymbol{x}_i|, \\ &\text{s.t. } \|\boldsymbol{w}\|_2 = 1. \end{aligned} \tag{2}$$

2.3. Sparse PCA

The sparse PCA problem (Luss and Teboulle, 2011) in the case of the covariance matrix is described as follows:

$$\begin{aligned} &\boldsymbol{w}^* = \arg\max(\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w}), \\ &\text{s.t. } \|\boldsymbol{w}\|_2 = 1,\ \|\boldsymbol{w}\|_0 \leqslant k \text{ or } \|\boldsymbol{w}\|_1 \leqslant k, \end{aligned} \tag{3}$$

where k is a parameter controlling the sparsity of projection vectors. Recently, Journee et al. (2011) also proposed two single-unit optimization formulations of the sparse PCA problem based on L1 or L0 norm penalties.

3. Our proposed method

3.1. Lp norm GPCA

Inspired by robust PCA or sparse PCA, we propose the following optimization problem to obtain the projection vector $\boldsymbol{w}$:

$$\begin{aligned} &\boldsymbol{w}^* = \arg\max_{w} \sum_{i=1}^{n} \phi(\boldsymbol{w}^T\boldsymbol{x}_i), \\ &\text{s.t. } \|\boldsymbol{w}\|_p^p \leqslant 1,\ p > 0, \end{aligned} \tag{4}$$

where $\|\boldsymbol{w}\|_p = (\sum_{i=1}^{d} |w_i|^p)^{1/p}$ and w_i is the ith component of vector $\boldsymbol{w}$. The function $\phi(s) : (-\infty, +\infty) \to [0, +\infty)$ in Eq. (4) should satisfy the following conditions: (1) $\phi(s)$ is a continuously convex function and its (sub)gradient at each point lies in a compact set; (2) $\phi(s)$ is an even function , i.e., $\phi(s) = \phi(-s)$. Note that $\boldsymbol{w}^T\boldsymbol{x}_i$ is an affine function and $\phi(\,)$ is a convex function, so the composition of them is also a convex function. As a result, the objective function is convex with respect to $\boldsymbol{w}$. It is known that the solution from the maximization of a convex function over a compact set is achieved at an extreme point. When the objective function is not strictly convex, one may obtain the global optimum at non-extreme points. However, one may always find the optimal solution of Eq. (4) from extreme points. If there are infinite extreme points for the constraint set, it is impractical to find the optimal solution by checking extreme points. In the following, we give two remarks for the parameter $\boldsymbol{p}$.

Remark 1. When $p = 1$ in Eq. (4), it is easy to obtain the extreme points of the constraint set, i.e., $w_i = \pm 1,\ w_1, \cdots w_{i-1},\ w_{i+1} \cdots w_d = 0,\ i = 1, \cdots, d.$ Specifically, there are $2d$ extreme points for this constraint set. One may find the optimal solution from these extreme points. Thus the projection vector only contains one nonzero component. This corresponds to selecting one feature if one projection vector is used. Note that the optimal solution may be obtained at non-extreme points if the objective function is not strictly convex.

Remark 2. When p in Eq. (4) takes the infinity, there are $2d$ extreme points for the constraint set, i.e., $w_i = \pm 1,\ i = 1, \cdots, d$. This produces the so-called binary GPCA if we search the optimal solution from extreme points. This may be beneficial in some cases where the dot product operation in obtaining projected data can be effectively computed by addition operations instead of multiplication operations, which is shown in (Pang et al., 2009). However, as the dimension (d) of samples grows, the number of extreme points exponentially increases. Moreover, it is NP-hard from the complexity of viewpoints.

In addition, one can observe from Eq. (4) that the function $\phi(\,)$ should be given in advance. In real applications, in order to suppress contaminated data, one may further restrict the range of convex functions. For example, $\phi(\,)$ increases more slowly than a quadratic function, i.e., $\lim_{s\to\infty} \frac{\varphi(s)}{s^2} = 0$. For clarity, we list some convex functions used in Eq. (4): (i) $\phi(s) = |s|^\theta,\ \theta \geqslant 1$; (ii) $\phi(s) = \max(|s| - \theta, 0)$; (iii) $\phi(s) = \max(|s| - \theta, 0)^2$; (iv) $\phi(s) = \sqrt{s^2 + \theta^2}$; (v) $\phi(s) = s^2$ if $|s| \leqslant \theta,\ \phi(s) = \theta(\theta + 2|s - \theta|)$ if $|s| > \theta$.

The main aim of adopting convex functions in our optimization problem is to guarantee that the objective function in Eq. (4) is convex. Thus there are some effective algorithms in global optimization that can be used to solve Eq. (4). In this paper, we adopt the successive linearization technique (SLT) to solve Eq. (4) as done in (Bradely and Mangasarian, 1998; Journee et al., 2011). First, we linearizate the objective function in Eq. (4) at $\boldsymbol{w}^{(k)}$ in the kth iteration.

$$\begin{aligned} &\boldsymbol{w}^{(k+1)} = \arg\max\{\sum_{i=1}^{n} \phi(\boldsymbol{x}_i^T\boldsymbol{w}^{(k)}) + (\boldsymbol{w}^T - \boldsymbol{w}^{(k)})\sum_{i=1}^{n}[\phi'(\boldsymbol{x}_i^T\boldsymbol{w}^{(k)})\boldsymbol{x}_i]\} \\ &\text{s.t. } \|\boldsymbol{w}\|_p^p \leqslant 1,\ p > 0. \end{aligned} \tag{5}$$

For clarity, we briefly describe how to solve Eq. (4) in terms of SLT in Algorithm 1.

Algorithm 1: the solution to Eq. (4)

Set $k = 0$, $\boldsymbol{w}^{(0)}$ is a column vector whose Lp norm is 1
While passing the stopping criterion
k: = k + 1
  if $\boldsymbol{w}^{(k)}$ is a non-differential point
    find some subgradient at $\boldsymbol{w}^{(k)}$ which is different from the previous (sub)gradient at $\boldsymbol{w}^{(k-1)}$ to compute $\boldsymbol{w}^{(k+1)}$ by Eq. (5);
  else
    solve Eq. (5) to obtain $\boldsymbol{w}^{(k+1)}$;
  end if
End while

It is observed that at each iteration in Algorithm 1, one needs to solve an optimization problem with the linear objective function and nonlinear constraints, i.e., Eq. (5). Since ϕ is a convex function and may not be differential, e.g., $\phi(s) = |s|$, we use ϕ' to denote any