Pattern Recognition 60 (2016) 901–907

Contents lists available at ScienceDirect

Pattern Recognition

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

L1-norm-based principal component analysis with adaptive regularization

Gui-Fu Lu *, Jian Zou, Yong Wang, Zhongqun Wang

School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui 241000, China

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form
7 July 2016
Accepted 7 July 2016
Available online 8 July 2016

Keywords:
Principal component analysis
Dimensionality reduction
L1-norm
Trace lasso
L2-norm

ABSTRACT

Recently, some L1-norm-based principal component analysis algorithms with sparsity have been proposed for robust dimensionality reduction and processing multivariate data. The L1-norm regularization used in these methods encounters stability problems when there are various correlation structures among data. In order to overcome the drawback, in this paper, we propose a novel L1-norm-based principal component analysis with adaptive regularization (PCA-L1/AR) which can consider sparsity and correlation simultaneously. PCA-L1/AR is adaptive to the correlation structure of the training samples and can benefit both from L2-norm and L1-norm. An iterative procedure for solving PCA-L1/AR is also proposed. The experiment results on some data sets demonstrate the effectiveness of the proposed method.

1. Introduction

Dimensionality reduction [1] is of great importance in many applications, e.g., pattern recognition, text categorization and computer vision where the dimensionality of data is often very high. It can reduce the computational complexity and discover the intrinsic manifold structure of high-dimensional data. Principal component analysis (PCA) [1,2] is perhaps the most famous dimensionality reduction technique due to its simplicity and effectiveness. Generally, PCA finds a set of projection vectors such that the variance of the projected data points is maximized. By projecting the data onto the set of projection vectors, the data structure in the original input space can be discovered.

Every projection vector obtained by PCA is a nonzero linear combination of all the data, and then each variable in data point is regarded as equally important in dimensionality reduction. Hence, the extracted features obtained by PCA are difficult to interpret. The original variables in the high-dimensional data, however, have meaningful physical interpretation in many applications. In this case, the interpretation of the obtained projection vectors can be enhanced if the obtained projection vectors involve more zero entries.

As a consequence sparse PCA (SPCA) [3], which reformulates the conventional PCA as a regression-type optimization problem with the elastic net regularization, has been proposed and can gain good experiment results. Some different implementations of SPCA have also proposed [4,5]. SPCA has gained success in many applications for extracting interpretable principal components. By using structured regularization, Jenatton et al. [6] generalized SPCA to structured sparse PCA.

However, the objective functions of the above mentioned PCA and SPCA are both based on L2-norm, which makes these methods to be sensitive to noise and outliers since the square operation in L2-norm will exaggerate the effect of noise and outliers. It is generally believed that L1-norm is more robust to noise and outliers than L2-norm. Then, in recent years, some L1-norm-based principal component analysis methods have been developed in the literature [7–17]. Due to the use of the absolute value operator in L1-norm, however, it is much more difficult to obtain the optimal projection vectors of L1-norm-based PCA than those of L2-norm-based PCA.

Recently, Kwak [11] proposed the PCA-L1 method, which is also based on L1-norm and rotational invariant. A greedy iterative algorithm for solving PCA-L1 is also presented in [11]. Experiment results on data sets shows the effectiveness of PCA-L1. Nie et al. [13] proposed a non-greedy procedure to calculate the projection vectors of PCA-L1. Nie's method can obtain all the projection vectors of PCA-L1 simultaneously while the original PCA-L1 method obtains the projection vectors one by one. Li et al. [8] proposed the L1-norm-based 2DPCA algorithm (2DPCA-L1), which is a robust version of the 2DPCA method [18]. Further, Pang et al. [12] proposed the L1-norm-based tensor PCA method (TPCA-L1). Motivated by Nie's non-greedy PCA-L1, Wang et al. [16] and Cao

* Corresponding author.
E-mail address: luguifu_tougao@163.com (G.-F. Lu).

http://dx.doi.org/10.1016/j.patcog.2016.07.014

et al. [17], respectively, proposed the non-greedy versions of 2DPCA-L1 and TPCA-L1.

In order to improve the interpretation of the basis vectors of PCA-L1, Meng et al. [14] proposed a sparse PCA-L1 method called PCA-L1 with sparsity (PCA-L1S). Not only the objective function of PCA-L1S is based on L1-norm, but the basis vectors are also penalized by L1-norm. Similarly, Wang et al. [7] proposed 2DPCA-L1 with sparsity (2DPCA-L1S).

The L1-norm regularization can work optimally on high-dimensional low-correlation data [19–22]. However, there are various correlation structures among a lot of data. In this situation, the L1-norm regularization encounters instability problems. Recently, trace Lasso [19,23–25] has been proposed to remedy this instability problem. Trace lasso is adaptive and interpolates between L1-norm and L2-norm.

In this paper, we use trace Lasso to regularize the basis vectors of PCA-L1 and propose a novel L1-norm-based principal component analysis, called PCA-L1 with adaptive regularization (PCA-L1/AR). PCA-L1/AR, which can consider sparsity and correlation simultaneously, is adaptive to the correlation structure and can benefit both from L2-norm and L1-norm. We also present an iterative algorithm for solving PCA-L1/AR. The experiments on some publicly available data sets confirm the effectiveness of the proposed method.

The remainder of the paper is organized as follows. In Section 2, we review briefly the PCA and PCA-L1 techniques. In Section 3, we propose the PCA-L1/AR approach, including its objective function and algorithmic procedure. The experiment results are reported in Section 4. Finally, we conclude the paper in Section 5.

2. Outline of PCA, PCA-L1 and PCA-L1S

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in R^{d\times n}$ be a d-dimensional sample set with n elements. Without loss of generality, we assume that X has been centered. The classical PCA method (termed as PCA-L2) aims to maximize the variance of data points in the projected subspace. The optimal projection vector $\mathbf{w} \in R^d$ can be obtained by solving the following criterion function:

$$\max_{\mathbf{w}^T\mathbf{w}=1} \mathbf{w}^T S_t \mathbf{w} \tag{1}$$

where $S_t = \frac{1}{n}XX^T$ is the covariance matrix. The optimal subspace of PCA is spanned by the eigenvectors of S_t corresponding to the largest m eigenvalues. Eq. (1) can be reformulated as

$$\max_{\mathbf{w}^T\mathbf{w}=1} \frac{1}{n}\|\mathbf{w}^T X\|_2^2 \tag{2}$$

where $\|\bullet\|_2$ denotes the L2-norm of a vector.

Obviously the conventional PCA is based on L2-norm. In [11], Kwak proposed PCA-L1, where L2-norm in PCA-L2 is replaced with L1-norm. Then the robustness to noise and outliers of PCA-L2 is improved. PCA-L1 aims to maximize the following objective function

$$\max_{\mathbf{w}^T\mathbf{w}=1} \|\mathbf{w}^T X\|_1 \tag{3}$$

where $\|\bullet\|_1$ denotes the L1-norm of a vector. Kwak proposed a greedy iterative procedure to compute $\mathbf{w}$ since it is difficult to solve Eq. (3) directly.

In order to improve the interpretation of the basis vectors of PCA-L1, Meng et al. [14] proposed PCA-L1S, where L1-norm is not only used in the objective function, but also used to regularize the basis vector of PCA-L1. PCA-L1S aims to solve the following optimization problem

$$\max \|\mathbf{w}^T X\|_1, \quad \text{subject to } \mathbf{w}^T\mathbf{w} = 1, \ \|\mathbf{w}\|_1 < k \tag{4}$$

where k is a positive integer. An efficient iterative procedure to solve Eq. (4) is also presented in [14].

3. L1-norm-based principal component analysis with adaptive regularization (PCA-L1/AR)

3.1. Problem formulation

In this subsection, we will present our proposed L1-norm-based principal component analysis with adaptive regularization (PCA-L1/AR).

The L1-norm regularization will encounter stability problems if the data samples exhibit strong correlations [19]. In this paper, we impose the trace norm onto $\mathbf{w}$ inspired by [19]. Specifically, we integrate the trace norm into the objective function of PCA-L1 and the objective function of PCA-L1/AR is formulated as

$$\arg\max_{\mathbf{w}} \|X^T\mathbf{w}\|_1 - \lambda\|X^T Diag(\mathbf{w})\|_* \tag{5}$$

or

$$\arg\min_{\mathbf{w}} \lambda\|X^T Diag(\mathbf{w})\|_* - \|X^T\mathbf{w}\|_1 \tag{6}$$

where $\|\bullet\|_*$ denotes the trace norm of a matrix, i.e., the sum of its singular values, $Diag(\bullet)$denotes to convert a vector into a diagonal matrix. In Section 3.2, we will introduce how to solve the objective function of PCA-L1/AR, i.e., Eq. (6). The main difference between trace norm $\|X^T Diag(\mathbf{w})\|_*$and other norm, e.g. L1-norm and L2-norm, is that $\|X^T Diag(\mathbf{w})\|_*$ contains the data sample matrix X. $\|X^T Diag(\mathbf{w})\|_*$ is adaptive to the correlation structure and interpolates between L1-norm and L2-norm [19]. If $XX^T = I$, i.e., the data are uncorrelated, then we have

$$\begin{aligned}\|X^T Diag(\mathbf{w})\|_* &= tr\left[(X^T Diag(\mathbf{w}))^T(X^T Diag(\mathbf{w}))\right]^{0.5}\\ &= tr\left[Diag(\mathbf{w})XX^T Diag(\mathbf{w})\right]^{0.5} = \|\mathbf{w}\|_1\end{aligned} \tag{7}$$

Thus, the trace norm regularization, i.e., $\|X^T Diag(\mathbf{w})\|_*$, is equal to the L1-norm. If $X = \mathbf{1}\mathbf{x}^1$, i.e., the data are highly correlated, where $\mathbf{x}^1$ denotes the first row of X and $\mathbf{1} \in R^d$ is a column vector taking one at each entry, then we have

$$\begin{aligned}\|X^T Diag(\mathbf{w})\|_* &= \|(\mathbf{x}^1)^T\mathbf{w}\|_*\\ &= \|\mathbf{x}^1\|\|\mathbf{w}\|_2 = \|\mathbf{w}\|_2\end{aligned} \tag{8}$$

Thus in the case $\|X^T Diag(\mathbf{w})\|_*$ is equal to the L2-norm. For other cases, trace Lasso interpolates between the L1-norm and L2-norm depending on correlations [19], i.e.,

$$\|\mathbf{w}\|_2 \le \|X^T Diag(\mathbf{w})\|_* \le \|\mathbf{w}\|_1 \tag{9}$$

This means that trace Lasso can benefit both from L2-norm and L1-norm according to the correlations among data.

3.2. Optimization procedure for PCA-L1/AR

Motivated by the optimization method used in [26], we use the augmented Lagrange multipliers (ALM) method [27] to solve Eq. (6). In [27], the ALM method is introduced for solving the following constrained optimization problem:

$$\min f(X) \quad s.t. \quad h(X) = 0 \tag{10}$$

where $f: R^n \to R$ and $h: R^n \to R^m$. We can define the augmented Lagrangian function to solve Eq. (10):