Pattern Recognition 61 (2017) 524–536

Contents lists available at ScienceDirect

Pattern Recognition

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

Joint sparse principal component analysis

Shuangyan Yi [a], Zhihui Lai [b], Zhenyu He [a,*], Yiu-ming Cheung [c,d], Yang Liu [c,d]

[a] School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, China
[b] The College of Computer Science and Software Engineering, Shenzhen University, China
[c] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[d] The Institute of Research and Continuing Education, Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 16 January 2016
Received in revised form
21 August 2016
Accepted 22 August 2016
Available online 24 August 2016

Keywords:
Dimensionality reduction
Joint sparse
$\ell_{2,1}$-norm

ABSTRACT

Principal component analysis (PCA) is widely used in dimensionality reduction. A lot of variants of PCA have been proposed to improve the robustness of the algorithm. However, the existing methods either cannot select the useful features consistently or is still sensitive to outliers, which will depress their performance of classification accuracy. In this paper, a novel approach called joint sparse principal component analysis (JSPCA) is proposed to jointly select useful features and enhance robustness to outliers. In detail, JSPCA relaxes the orthogonal constraint of transformation matrix to make it have more freedom to jointly select useful features for low-dimensional representation. JSPCA imposes joint sparse constraints on its objective function, i.e., $\ell_{2,1}$-norm is imposed on both the loss term and the regularization term, to improve the algorithmic robustness. A simple yet effective optimization solution is presented and the theoretical analyses of JSPCA are provided. The experimental results on eight data sets demonstrate that the proposed approach is feasible and effective.

1. Introduction

Dimensionality reduction is an important issue in data classification. It aims to learn a transformation matrix to project the high-dimensional data into a low-dimensional subspace so that the data can be effectively classified in the low-dimensional subspace. There are many methods for dimensionality reduction [1–5] and the classical methods are principal component analysis (PCA) [6–9] and linear discriminant analysis (LDA) [10–12]. PCA is an unsupervised method, which projects data information into an orthogonal linear space. LDA is a supervised method, which extracts discriminative data information by maximizing the interclass scatter matrix and at the same time minimizing the intraclass scatter matrix [13,14].

It is well known that PCA is an unsupervised method and the unsupervised methods are important in the practical applications [15] since labeled data are expensive to obtain [16]. However, the original PCA is sensitive to outliers since its covariance matrix is derived from ℓ_2-norm and ℓ_2-norm is sensitive to outliers [7,17,18]. Thus, PCA fails to deal with the outliers that often appear in data sets in real-world applications. In terms of this problem, many variants of PCA [18,19,16] have been proposed to reduce the effect of outliers. One of the main strategies is to impose ℓ_1-norm on loss term [20,18,21,22,19]. In detail, PCA based on ℓ_1-norm maximization [18] uses a greedy strategy to solve the optimization problem and easy to get stuck in a local solution. Robust principal component analysis with non-greedy ℓ_1-norm maximization (RPCA) [19] is proposed to obtain a much better solution than that in [18]. Recently, $\ell_{2,1}$-norm has caused wide research interests [16,23,24]. Rotational invariant ℓ_1-norm PCA [23] imposes $\ell_{2,1}$-norm on loss term [16]. Optimal mean robust principal component analysis (OMRPCA) [16] based $\ell_{2,1}$-norm is proposed to learn the optimal transformation matrix and optimal mean simultaneously, which imposes $\ell_{2,1}$-norm on loss term.

Although the variants of PCA method mentioned above are able to reduce the effect of outliers to some extent, one major disadvantage of them is that each new feature in low-dimensional subspace is the linear combination of all the original features in high-dimensional space. Therefore, it is usually not good for classification due to the redundant features. Besides, it is often difficult to interpret the new features. Actually, the interpretation of the new features is very important especially when they have physical meanings in many applications such as gene representation and face recognition. To facilitate interpretation, sparse principal component analysis (SPCA) [25] is proposed. However, SPCA has no ability to jointly select the useful features because the ℓ_1-norm is imposed on each transformation vector and ℓ_1-norm cannot select the consistent features. Moreover, SPCA still suffers from the effect of outliers because the ℓ_2-norm is imposed on loss

* Corresponding author.
E-mail addresses: lai_zhi_hui@163.com (Z. Lai), zyhe@hitsz.edu.cn (Z. He), ymc@comp.hkbu.edu.hk (Y.-m. Cheung), csygliu@comp.hkbu.edu.hk (Y. Liu).

http://dx.doi.org/10.1016/j.patcog.2016.08.025
0031-3203/

term.

In this paper, we propose joint sparse principal component analysis (JSPCA), which integrates feature selection into subspace learning to exclude the redundant features. Specifically, JSPCA imposes joint $\ell_{2,1}$-norms on both loss term and regularization term. In this way, our method can discard the useless features on one hand and reduce the effect of outliers on the other hand. The main contributions are described as follows:

(1) JSPCA relaxes the orthogonal constraint of transformation matrix and introduces another transformation matrix to together recover the original data from the subspace spanned by the selected features, which makes JSPCA have more freedom to jointly select useful features for low-dimensional representation.

(2) Unlike PCA and its existing extensions, JSPCA uses joint sparse constraints on the objective function, i.e., $\ell_{2,1}$-norm is imposed on the loss term and the transformation matrix, to do feature selection and learn the optimal transformation matrix simultaneously.

(3) A simple yet effective optimal solution of JSPCA is provided. Furthermore, a series of theoretical analyses including convergence analysis, essence of JSPCA, and computational complexity are provided to validate the feasibility and effectiveness of JSPCA.

The remainder of this paper is organized as follows. In Section 2, we review some existing dimensionality reduction methods. In Section 3, we present the JSPCA model with an effective solution. In Section 4, we give the analyses of JSPCA in theory. In Section 5, we perform experiments and provide the observations. Finally, conclusion is drawn in Section 6.

2. Related work

In this section, we first give the basic notations and then review several variants of PCA. Suppose the given data matrix is $X = [x_1, \ldots, x_n] \in R^{m\times n}$, where m denotes the original image space dimensionality and n denotes the number of training samples. Without loss of generality, $\{x_j\}_{j=1}^{n}$ is assumed to have zero mean. The problem of linear dimensionality reduction is to project the data from the high-dimensional original space into a low-dimensional subspace. That is, we need to find a transformation matrix $A = [a_1, a_2, \ldots, a_d] \in R^{m\times d}$ with $d \lll m$, where each transformation vector a_k is with m loadings ($k = 1, 2, \ldots, d$). Then, the transformed data denoted by Y can be shown as follows:

$$Y = A^T X \in R^{d\times n}. \tag{1}$$

Notations: For the matrix A, we denote the (i,j)-th element by a_{ij}, the i-th row by A^i. In this paper, we denote $\|A\|_{2,1} = \sum_{i=1}^{m}\left\|A^i\right\|_2$, where $\left\|A^i\right\|_2$ means the ℓ_2-norm of vector A^i and $\left\|A^i\right\|_2 = \sqrt{\left\|A^i\right\|^T \left\|A^i\right\|}$.

The traditional PCA [6] based on ℓ_2-norm aims to project the high-dimensional data onto the low-dimensional linear subspace spanned by the leading eigenvectors of the data covariance matrix. RPCA [19] based on ℓ_1-norm aims to be robust to outliers by imposing ℓ_1-norm on the projected data. OMRPCA [16] based on $\ell_{2,1}$-norm aims to remove optimal mean automatically and enhance the robustness to outliers by imposing $\ell_{2,1}$-norm on the loss term.

All of the above methods focus on operating different norms such as ℓ_2-norm, ℓ_1-norm, and $\ell_{2,1}$-norm on the loss term. Although the above methods can get a prominent performance in many cases, one common disadvantage of the above methods is that each new feature is the linear combination of all the original features. To this end, the regularization term imposed by different norms is proposed to solve this problem. For example, based on PCA, SPCA [25] is proposed to learn a sparse projection matrix, where each new feature is the linear combination of some original features. Based on spectral regression [26], sparse subspace learning (SSL) [27] is proposed for learning a sparse projection matrix, which first regress the low-dimensional projection data and then solve the projection matrix. However, both SPCA and SSL still cannot exclude the redundant features. Furthermore, based on graph embedding [28], joint feature selection and subspace learning (JFSSL) [2] is proposed to integrate the ability of feature selection into subspace learning. Although JFSSL has the ability of feature selection, it is sensitive to outliers.

3. Joint sparse principal component analysis

In this section, we first present the motivation of this work. Then, we give the objective function of the proposed method. Finally, an iterative optimal solution is given for the proposed objective function.

3.1. Motivation of JSPCA

As the previous statement, SPCA attempts to interpret the selection of features. Intuitively, we use the right subfigure in Fig. 1

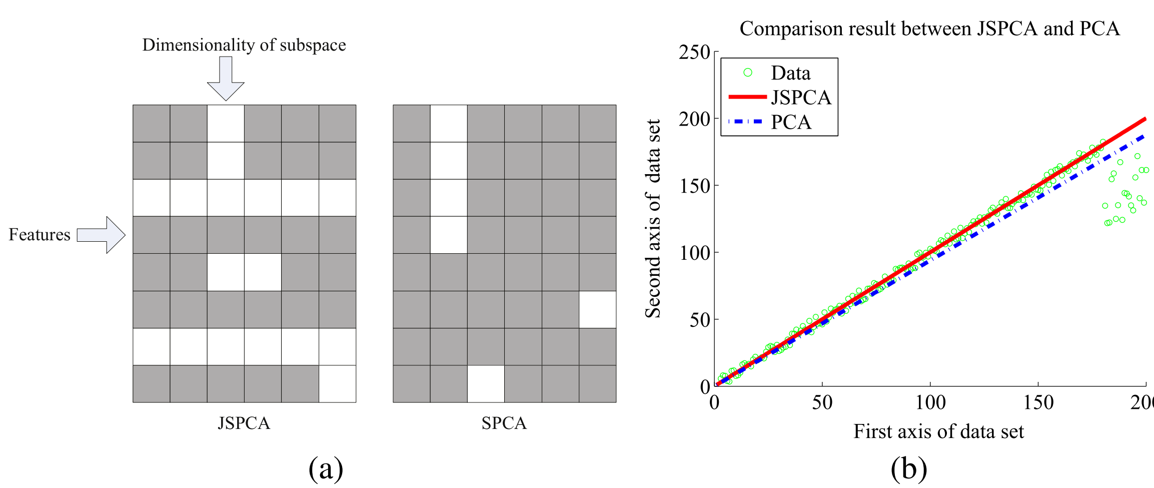

Fig. 1. Motivations of JSPCA. (a) Illustration of two transformation matrices got by JSPCA and SPCA, in which the white block means the zero loading and the gray block means the non-zero loading. JSPCA can tell us that the third and the seventh features are the useless features while SPCA cannot. (b) On a data set with some outliers, JSPCA shows more robustness to outliers than PCA.