Neurocomputing 207 (2016) 754–762

Contents lists available at ScienceDirect

Neurocomputing

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

Schatten p-norm based principal component analysis

Heyou Chang [a], Lei Luo [a], Jian Yang [a,*], Meng Yang [b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, PR China
[b] Shenzhen University, Shenzhen, PR China

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
2 April 2016
Accepted 3 May 2016
Communicated by Liu Guangcan
Available online 4 June 2016

Keywords:
Structured sparse PCA
Low rank
Schatten p-norm
ADMM

ABSTRACT

Structured sparse PCA (SSPCA) is a new emerging method regularized by structured sparsity-inducing norms. However, these regularization terms are not necessarily optimal because of the noisy and irrelevant features embedded in predefined patterns. This paper presents a method called Schatten p-norm based principal component analysis (S_pPCA) to learn interpretable and structured elements (or factors). In S_pPCA, a low-rank assumption is used to characterize structured elements in a two-dimensional matrix form. Compared to SSPCA, the low-rank assumption of S_pPCA is more intuitive and effective for describing object parts of an image. Moreover, S_pPCA can deal with some scenarios, where the dictionary element matrixes have complex structures. We also propose an efficient and simple optimization procedure to solve the problem. Extensive experiments on denoising of sparse structured signals and face recognition on different databases (e.g. AR, Extend Yale B and Multi-PIE) demonstrate the superior performance over some recently proposed methods.

1. Introduction

Principal component analysis (PCA) [1] is a classical technique widely used in feature extraction and unsupervised dimensionality reduction. It aims to find an orthogonal transformation to convert a set of correlated variables into uncorrelated ones. In the last decade, several alternatives to PCA have been proposed, notably independent component analysis (ICA) [2] and non-negative matrix factorization (NMF) [3].

Since each principal component in PCA is a linear combination of all the original variables, it is often difficult to interpret the results. For solving this problem, Zou et al. [4] proposed sparse PCA (SPCA) using the lasso (elastic net) to produce modified principal component. Christophe et al. [5] proposed robust SPCA to make the analysis resistant to outlying observations. However, these methods seem inappropriate in many applications because they only constrain the size or the sparsity of the principal factors without considering the important structures. For example, the pixels of an image, the common most data in compute vision, are naturally organized on a grid. The structural information in the image is very helpful. Meanwhile, the supports of factors explaining the variability of images can be expected to be localized or connected, such as eyes or mouth in face images. These structured relationships among variables can help us better interpret data and provide new insights into the underlying processes. Some works have been reported in the context of regression and classification [6,7], occlusion pattern learning [8], and back ground subtraction [9] by exploiting such structure. Particularly, Jenatton et al. [10] proposed structured sparse PCA (SSPCA) to yield a structured and sparse formulation of principal component analysis by adding sparse and some prior structural constraints in elements. Nonetheless, the convex structured sparsity constraints in [10] may not necessarily consist in real-world applications. In order to capture more flexible and general structure, Ren et al. [11] introduced binary matrices as auxiliary variables and proposed Markov Random Field (MRF) based SSPCA (MS^2PCA) for gene interaction. It is worth noting that the methods mentioned above worked off-line. Then Zoltan et al. [12] proposed online group-structured dictionary learning (OSDL) to fit large or slowly varying systems. Although these methods could achieve structured local factors, the structures of factors need to be pre-given in advance, which is likely to be not adequate and accurate in many practical applications because of the noisy and irrelevant features.

Recently, low-rank hypothesis has become an effective technique in depicting the structure in data. For instance, Wright et al. established a robust principal component analysis (RPCA) [13] method, which assumes that the error matrix is sparse and the clean data matrix is low rank. As an important extension of RPCA, the low-rank representation (LRR) [14] was presented to recover the subspace structures among data samples. Given a data matrix,

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.05.068
0925-2312/

LRR seeks the lowest-rank representation among all the candidates that can represent the data samples as linear combinations of the bases in a given dictionary. Unlike RPCA and LRR, Zhang et al. [15] focused on dealing with image data which were corrupted by continuous occlusion, and proposed double nuclear norm-based matrix decomposition (DNMD) to recover the low-rank data in image vector space and remove the low rank error in the image (matrix) space simultaneously. To make full use of the low-rank structural information of error image, Yang et al. [16] proposed a two-dimensional image matrix based matrix regression model, named nuclear norm based matrix regression (NMR), to carry out the image representation and classification. To handle face images with mixed noise, i.e., the structural noise plus the sparse noise, Luo et al. [17] proposed nuclear-L_1 norm joint matrix regression (NL_1R) model for face recognition with mixed noise, which are derived by using MAP (maximum a posteriori probability) estimation. All these methods use rank function to characterize the structural information in data. Additionally, to facilitate the computation, they replace the rank function with the nuclear norm. And the experimental results show that nuclear norm based models indeed obtain the effective low rank solutions in a variety of scenarios.

In this paper, we aim to learn the interpretable structural elements (e.g., mouth, eyes, or forehead in face images), which generally lead to a low-rank image in contrast to the full-rank original image. In [10–12], using the pre-defined structural constraints in factors is away from the actual situation and cannot depict complex structural information. Here we use nuclear norm to capture the structures in data. Good performance has been reported in [14–16] by using nuclear norm to depict low-rank structures. As we know, the nuclear norm of a matrix is equal to the L_1 norm of a vector formed by the singular values of the same matrix. Inspired by the experimental observations and theoretical guarantees showing superiority of L_p quasi-norm minimization to L_1 minimization in compressive sampling (CS) [18], some approaches in [19,20] replace the rank function with the Schatten-p quasi-norm and verify that Schatten-p quasi-norm minimization is superior to nuclear norm minimization. Specially, Nie et al. [21] proposed to jointly use Schatten p-norm and L_p-norm to approximate the rank minimization problem and enhance the robustness to outliers, and achieved promising performance on collaborative filtering and social network link prediction. This paper presents a Schatten p-norm based principal component analysis method to perform the image denoising and classification. Instead of pre-defining certain structures in the elements, the proposed model processes the elements in a two-dimensional matrix form and learns the structures effectively by using low-rank constraint. Compared with SSPCA, the pre-given structures are not required and more discriminative and complex structural variables can be captured. Alternating direction method of multipliers (ADMM) is utilized to solve the proposed model. We perform experiments on the denoising of sparse structured synthetic signals and face recognition on the AR [22], Extend Yale B [23] and MultiPIE [24] databases. The experimental results clearly demonstrate that the proposed method is more effective than state-of-the-art methods for image denoising and face recognition.

The rest of this paper is organized as follows. Section 2 presents the proposed S_pPCA model and the optimization of S_pPCA. Section 3 conducts the experiments on denoising and faces recognition, and Section 4 concludes the paper.

Notations: The extended Schatten p-norm $(0 < p < \infty)$ of a matrix $\mathbf{A} \in \mathcal{R}^{l \times m}$ is defined as $\| \mathbf{A} \|_{S_p} = (\sum_{i=1}^{\min\{l,m\}} \sigma_i^p)^{1/p} = (\mathrm{tr}(\mathbf{A}^T\mathbf{A})^{p/2})^{1/p}$, where $\mathrm{tr}(\cdot)$ means the trace operator. The Schatten p-norm of matrix $\mathbf{A} \in \mathcal{R}^{l \times m}$ to the power p is $\| \mathbf{A} \|_{S_p}^p = \sum_{i=1}^{\min\{l,m\}} \sigma_i^p = \mathrm{tr}(\mathbf{A}^T\mathbf{A})^{p/2}$. If $p=1$ and 2, the Schatten p-norm becomes nuclear norm (denoted by $\|\cdot\|_*$) and Frobenius norm (denoted by $\|\cdot\|_F$), respectively. The extended L_p-norm $(0 < p < \infty)$ of a vector $\mathbf{x} \in \mathcal{R}^{n \times 1}$ is denoted by $\| \mathbf{x} \|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$, where x_i is the ith element of the vector $\mathbf{x}$.

Remark 1. When $p < 1$, the extended Schatten p-norm is only a quasi-norm. But for convenience, we still call it Schatten p-norm.

2. The proposed model

In this section, we first propose Schatten p-norm based principal component analysis (S_pPCA, for short) model by using Schatten p-norm to constrain each element of the dictionary, and then apply ADMM to solve S_pPCA. Finally, we analyze the complexity and convergence of the proposed algorithm.

2.1. Schatten p-norm based principal component analysis model

Suppose that we are given a set of samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathcal{R}^{t \times n}$ $(t > n)$, the dictionary learning problem is to find a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K] \in \mathcal{R}^{t \times K}$ $(t \gg K)$ and coefficient $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n] \in \mathcal{R}^{K \times n}$, such that $\mathbf{X} \approx \mathbf{DZ}$. The matrix product $\mathbf{DZ}$ is called a decomposition of $\mathbf{X}$. To address this problem, dictionary learning has been widely investigated during the last decade [25–28]. It is natural, when learning the dictionary, to penalize or constrain some norms or quasi-norms on $\mathbf{D}$ and $\mathbf{Z}$, respectively, to encode prior information about the decomposition of $\mathbf{X}$. Dictionary learning (also called matrix decomposition [29]) is a general problem that contains, e.g., PCA, ICA, NMF, among many others. In particular, Jenatton et al. [10] casted SSPCA problem in the dictionary learning framework and considered some structural constraints to control the structure of the supports of dictionary elements. Zoltan et al. [12] extended SSPCA and proposed online group-structured dictionary learning, which not only was online, but also could represent general overlapping group structures and deal with missing information at a time.

Although the above approaches integrate the idea of structured sparsity into modeling, which use the mixed (L_1, L_2) norm to attempt to characterize the structure information of the overall dictionary $\mathbf{D}$, it actually destroys global structure of each element. As we know, for a face image which is spatially continuous, the global structure information is extremely significant for recognition task [15,16], thus, it should not be ignored. To emphasize this information, we do not stretch each element into a vector form, but use low rank constraint to characterize each element and preserve its spatial structure, which induces a low rank principal component analysis model:

$$\min_{\mathbf{D},\mathbf{Z}} \| \mathbf{X} - \mathbf{DZ} \|_F^2 + \lambda \sum_{i=1}^{K} \mathrm{rank}(\Psi(\mathbf{d}_i))$$
$$\text{s. t. } \| \mathbf{Z}^i \|_2 \le 1, \quad \forall\ i \in \{1, 2, \ldots, K\} \tag{1}$$

where $\lambda \ge 0$ is a regularization parameter, $\Psi(\mathbf{d})$ converts a vector $\mathbf{d} \in \mathcal{R}^t$ to a matrix in $\mathcal{R}^{l \times m}$ (suppose that the input image size is $l \times m$ and $t = l \times m$), $\mathbf{Z}^i$ is the ith row of $\mathbf{Z}$ to avoid trivial solutions.[1]

It is well-known that the rank minimization problem is difficult to solve since it is NP-hard. Recently, some scholars replace rank function using its convex envelope: nuclear norm [30]. However, the nuclear norm relaxation may deviate the solution away from the real solution of original rank minimization problem. As we know, when $p \to 0$, the Schatten

[1] For example, if using the nuclear norm or schatten p-norm to approximate rank function in Eq. (1), the objective can be decreased by respectively dividing and multiplying $\mathbf{d}_i$ and $\mathbf{Z}^i$ by a constant factor.