ARTICLE IN PRESS

Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Analytic separable dictionary learning based on oblique manifold

Fengzhen Zhang[a,b], Yigang Cen[a,b,*], Ruizhen Zhao[a,b], Hengyou Wang[a,b,c], Yi Cen[d], LiHong Cui[e], ShaoHai Hu[a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing, China
[b] Key Laboratory of Advanced Information Science and Network Technology of Beijing, Beijing, China
[c] School of Science, Beijing University of Civil Engineering and Architecture, Beijing, China
[d] School of Information Engineering, Minzu University of China, Beijing, China
[e] Department of Mathematics and Computer Science, Beijing University of Chemical Technology, Beijing, China

ARTICLE INFO

Keywords:
Dictionary learning
Oblique manifold
Sparse representation
2D-OMP

ABSTRACT

Sparse representation based on dictionary has gained increasing interest due to its extensive applications. Because of the disadvantages of computational complexity of traditional dictionary learning, we propose an algorithm of analytic separable dictionary learning. Considering the differences of sparse coefficient matrix and dictionary, we divide our algorithm into two phases: 2D sparse coding and dictionary optimization. Then an alternative iteration method is used between these two phases. The algorithm of 2D-OMP (2-dimensional Orthogonal Matching Pursuit) is used in the first phase because of its low complexity. In the second phase, we create a continuous function of the optimization problem, and solve it by the conjugate gradient method on oblique manifold. By employing the separable structure of the optimized dictionary, a competitive result is achieved in our experiments for image de-noising.

1. Introduction

Recently, sparse representation has received many attentions for its extensive applications to face recognition [1] and image processing, such as image de-noising [2], image super resolution reconstruction [3–5], etc. The model supposes that a signal can be represented as a sparse linear combination of a few columns of a dictionary, i.e., suppose that $\mathbf{D} \in \mathbb{R}^{m^2 \times n^2}$ is a dictionary with $m \ll n$, a signal $\mathbf{y} \in \mathbb{R}^{m^2}$ can be expressed as $\mathbf{y} \approx \mathbf{Dx}$, such that $\| \mathbf{y} - \mathbf{Dx} \|_p \le \epsilon$. Here the vector $\mathbf{x} \in \mathbb{R}^{n^2}$ is a sparse vector, i.e., $\| \mathbf{x} \|_0 = k \ll m^2$, which means that most of its entries are zeros or have small magnitudes [6–10].

The properties of dictionary $\mathbf{D}$ decide the sparsity of $\mathbf{y}$. The constraints on the dictionary include [6,7]:

(1) The columns of dictionary have unit Euclidean norm, i.e., $\| \mathbf{D}_i \|_2 = 1$ for $i = 1, 2, \ldots, n^2$.
(2) The dictionary has a full rank, i.e., $rank(\mathbf{D}) = m^2$.
(3) The dictionary does not have linearly dependent columns, i.e., $\mathbf{D}_i \neq \pm \mathbf{D}_j$ for $i \neq j$.

Therefore, it is crucial to find a dictionary so that the interested signal can be represented as accurate as possible with a coefficient vector $\mathbf{x}$ that is as sparse as possible [9,10]. There are two different types of dictionaries: analytic dictionaries and learned dictionaries [11]. In many cases, the analytic dictionaries can lead to simple and rapid algorithms for the problem of sparse representation, such as Wavelets [12], Curvelets, and Fourier transform matrix. But for some special signals, e.g., the natural facial images, the analytic dictionaries cannot lead to a perfect result, because the structure of signals is so complicated that a simple dictionary cannot capture the most salient features of them. Thus, algorithms of dictionary construction were proposed based on learning samples [10,13–15]. The typical dictionary learning problem is

$$\underset{\mathbf{D},\mathbf{X}}{argmin} \| \mathbf{X} \|_0 \quad s.\,t. \quad \| \mathbf{DX} - \mathbf{Y} \|_F^2 \le \epsilon \tag{1}$$

Here, $||*||_0$ denotes the ℓ_0 norm, which counts the nonzero entries of a matrix. $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N] \in \mathbb{R}^{m^2 \times N}$ is the matrix containing N learning samples, and the matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N] \in \mathbb{R}^{n^2 \times N}$ contains the corresponding sparse coefficients.

The most popular dictionary learning algorithm is the K-SVD, which is an alternative iteration method [10,15,16]. Recently, because of the classification inability of K-SVD, discriminative K-SVD was proposed [17]. In [11], a Fisher Discriminative K-SVD (FD-KSVD) was proposed, which not only employed the Fisher discrimination criterion to obtain discriminative coding coefficients, but also introduced a linear

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, Beijing, China.
E-mail address: 13112062@bjtu.edu.cn (F. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.09.099
Received 29 February 2016; Received in revised form 18 July 2016; Accepted 24 September 2016
0925-2312/

Please cite this article as: Cen, Y., Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.09.099

predictive classifier. In [6], they presented an analysis sparse model and projected the analysis dictionary to the oblique manifold due to the constraints of the dictionary. In this algorithm, the dictionary is updated as a whole part, which is different with the previous algorithms that updated the atoms of dictionary one by one. However, these algorithms are used for 1D vectors. When we deal with 2D signals, such as natural images, we need to convert images to 1D vectors. This kind of process not only breaks the potential correlations within the images, but is also restricted due to limitations in memory and computational ability [6,7,18]. Thus, separable learning schemes were proposed, which tackled this problem by enforcing additional structure on the learned dictionary, as follows:

$$\underset{\mathbf{A},\{\mathbf{X}_j\}_{j=1}^N,\mathbf{B}}{\text{argmin}} \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{X}_j\|_0 \ s.\ t. \quad \mathbf{A}\mathbf{X}_j\mathbf{B}^T=\mathbf{S}_j,\ \mathbf{X}_j\in\mathbb{R}^{n_\mathbf{A}\times n_\mathbf{B}},\ ddiag(\mathbf{A}^T\mathbf{A})=\mathbf{I}_{n_\mathbf{A}},$$
$$rank(\mathbf{A})=m_\mathbf{A},\ ddiag(\mathbf{B}^T\mathbf{B})=\mathbf{I}_{n_\mathbf{B}},\ \ rank(\mathbf{B})=m_\mathbf{B}. \tag{2}$$

Here, $ddiag(*)$ denotes a diagonal matrix consisting of the main diagonal elements of $*$. $\{\mathbf{S}_j\}_{j=1}^N$ are the learning samples and $\mathbf{S}_j\in\mathbb{R}^{m_\mathbf{A}\times m_\mathbf{B}},\ \forall\, j=1,\ldots,N$. The sizes of $\mathbf{A}\in\mathbb{R}^{m_\mathbf{A}\times n_\mathbf{A}}$ and $\mathbf{B}\in\mathbb{R}^{m_\mathbf{B}\times n_\mathbf{B}}$ may be different. In this paper, we let $m_\mathbf{A}=m_\mathbf{B}=m,\ n_\mathbf{A}=n_\mathbf{B}=n$ for simplicity.

In [16], 2D synthesis sparse model was proposed, which made full use of the local correlations within natural images. In this model, the dictionary $\mathbf{D}\in\mathbb{R}^{m^2\times n^2}$ has a separable structure, which can be given by the Kronecker product of two smaller dictionaries $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{B}\in\mathbb{R}^{m\times n}$, formulated as $\mathbf{D}=\mathbf{B}\otimes\mathbf{A}$. Then, it can be seen as a combination of twice K-SVD along the horizontal and vertical directions, respectively. In [7], based on the same assumption of the structure of $\mathbf{D}$, they used a continuous function to measure the sparsity and optimize the problem through the conjugate gradient algorithm. With the constraints on $\mathbf{A}$ and $\mathbf{B}$, they restricted these two dictionaries to be elements of the oblique manifold, i.e., $OB(m,n){:=}\{\mathbf{Q}\in\mathbb{R}^{m\times n}\mid rank(\mathbf{Q})=m,\ \ ddiag(\mathbf{Q}^T\mathbf{Q})=\mathbf{I}_n\}$. However, they ignored the influence of dictionary rank condition for the sparse representation. Also, a same step size was used for updating $\mathbf{A}$, $\mathbf{B}$, and $\{\mathbf{X}_j\}_{j=1}^N$.

In this paper, we propose an algorithm for analytic separable dictionary learning. Inspired by the K-SVD, we optimize the sparse coefficients and the dictionaries in two phases. In the first phase, the sparse coefficients $\{\mathbf{X}_j\}_{j=1}^N$ are updated by the 2D-OMP (2-dimensional Orthogonal Matching Pursuit). In the second phase, the dictionaries are optimized with the given $\{\mathbf{X}_j\}_{j=1}^N$. This optimization procedure alternated between these two phases until the stop criterion is satisfied. The main contributions are as follows:

(1) Different with [7] that updates $\{\mathbf{X}_j\}_{j=1}^N$, $\mathbf{A}$ and $\mathbf{B}$ in a same step size, they are updated respectively in our proposed algorithm. 2D-OMP is used to update $\{\mathbf{X}_j\}_{j=1}^N$, which not only ensures the sparsity of $\{\mathbf{X}_j\}_{j=1}^N$, but also speeds up the convergence of the objective function. Moreover, $\mathbf{A}$ and $\mathbf{B}$ are projected onto the oblique manifold and updated together, which uses the matrix multiplication instead of the SVD decomposition used in the traditional K-SVD. This will save the computational cost for the dictionary optimization, especially when the dictionary dimension is high.

(2) To emphasize the constraints of full rank and incoherence, two log-barrier functions are added to the objective function such that a continuous differentiable objective function can be obtained, which ensures that the conjugate gradient method can be applied to optimize the dictionary. Also, full rank constraint will improve the robustness of the optimized dictionary.

(3) Finally, an analytic separable dictionary learning algorithm is proposed, which not only reduces the computational cost of the dictionary optimization, but also optimizes the dictionary quickly even in the high-dimensional situation.

The remainder of this paper is organized as follows. In Section 2, we describe our proposed approach in details. The results are presented in Section 3. Section 4 concludes this paper.

2. Learning schemes

To solve the optimization problem (2), we divide it into two phases. The first phase calculates the sparse matrices $\{\mathbf{X}_j\}_{j=1}^N$ with the given dictionaries, which is called as the 2D sparse coding. In the second phase, we add the constraints of dictionaries through coefficients to the objective function. Due to the constraints of the dictionary and the continuous differentiable objective function, the dictionaries $\mathbf{A}$ and $\mathbf{B}$ are projected to the oblique manifold, and the geometric conjugate gradient method is employed to solve this function. By projecting the dictionaries to the oblique manifold, they can be updated as a whole part, which avoids updating the atoms of dictionary one by one such that the updating speed can be improved.

2.1. 2D Sparse coding

The aim of this phase is to get the sparse matrices $\{\mathbf{X}_j\}_{j=1}^N$ with the given dictionaries $\mathbf{A}$ and $\mathbf{B}$. We re-formulate this problem as

$$\underset{\mathbf{X}_j}{\arg\min}\ \|\mathbf{X}_j\|_0 \quad s.\ t.\ \|\mathbf{A}\mathbf{X}_j\mathbf{B}^T-\mathbf{S}_j\|_F^2\le\epsilon,\ \forall\, j=1,2,\ldots,N \tag{3}$$

[19] developed a 2D-OMP for 2D sparse signal recovery. With the advantages of low complexity and good performance, we use this algorithm to solve the problem (3). In 2D-OMP, each atom of dictionary is a matrix, which is the outer product of two columns of $\mathbf{A}$ and $\mathbf{B}$ respectively, i.e.,

$$\mathbf{D}_{p,q}=\mathbf{A}_p\times\mathbf{B}_q^T=\begin{pmatrix}\mathbf{A}_{1,p}\\ \mathbf{A}_{2,p}\\ \vdots\\ \mathbf{A}_{m,p}\end{pmatrix}\times\begin{pmatrix}\mathbf{B}_{1,q}\\ \mathbf{B}_{2,q}\\ \vdots\\ \mathbf{B}_{m,q}\end{pmatrix}^T=\begin{pmatrix}\mathbf{A}_{1,p}\mathbf{B}_{1,q} & \cdots & \mathbf{A}_{1,p}\mathbf{B}_{m,q}\\ \vdots & & \vdots\\ \mathbf{A}_{m,p}\mathbf{B}_{1,q} & \cdots & \mathbf{A}_{m,p}\mathbf{B}_{m,q}\end{pmatrix} \tag{4}$$

Here, $\mathbf{A}_p$ is the pth column of the matrix $\mathbf{A}$. $\mathbf{A}_{i,p}$, $1\le i\le m$ is an entry of the matrix $\mathbf{A}$, and so do $\mathbf{B}_q$ and $\mathbf{B}_{i,q}$. Then $\mathbf{S}_j$ can be represented by the weighted sum of $\mathbf{D}_{p,q}$. Suppose that $\widetilde{\mathbf{S}}_j=\sum_{p=1}^{n}\sum_{q=1}^{n}\widetilde{\mathbf{X}}_{p,q,j}\mathbf{D}_{p,q}$ is an approximation of $\mathbf{S}_j$ ($\widetilde{\mathbf{X}}_{p,q,j}$ is an entry of the 3-dimensional matrix $\widetilde{\mathbf{X}}$. p, q and j are indexes of the three dimensions respectively), $\mathbf{X}_j\in\mathbb{R}^{n\times n}$ can be obtained

$$\mathbf{X}_j=\underset{\mathbf{X}_j}{\arg\min}\ \|\mathbf{S}_j-\widetilde{\mathbf{S}}_j\|_F^2=\mathbf{H}^{-1}g \tag{5}$$

where

$$\mathbf{H}=\begin{pmatrix}tr(\mathbf{D}_{p_1,q_1}\mathbf{D}_{p_1,q_1}^T) & \cdots & tr(\mathbf{D}_{p_1,q_1}\mathbf{D}_{p_l,q_l}^T)\\ \vdots & & \vdots\\ tr(\mathbf{D}_{p_l,q_l}\mathbf{D}_{p_1,q_1}^T) & \cdots & tr(\mathbf{D}_{p_l,q_l}\mathbf{D}_{p_l,q_l}^T)\end{pmatrix},$$

$g=[tr(\mathbf{S}_j,\mathbf{D}_{p_1,q_1}^T),\ldots,tr(\mathbf{S}_j,\mathbf{D}_{p_l,q_l}^T)]^T$. l is the number of the selected atoms in the current iteration. For the details of 2D-OMP, please refer to [19].

2.2. Dictionary optimization

In this phase, we will optimize the low dimensional dictionaries $\mathbf{A}$ and $\mathbf{B}$. Different from [7], we add two log-barrier functions to the objective function to emphasize the full rank and the incoherence constraints, as follows.