Signal Processing 120 (2016) 620–626

Contents lists available at ScienceDirect

Signal Processing

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

Fast communication

Discriminative separable nonnegative matrix factorization by structured sparse regularization

Shengzheng Wang*, Jing Peng, Wei Liu

Merchant Marine College, Shanghai Maritime University, Shanghai 201306, PR China

ARTICLE INFO

Article history:
Received 18 July 2015
Received in revised form
24 September 2015
Accepted 21 October 2015
Available online 29 October 2015

Keywords:
Non-negative matrix factorization
Structured sparse regularization
Separability
Discriminative learning

ABSTRACT

Non-negative matrix factorization (NMF) is one of the most important models for learning compact representations of high-dimensional data. With the separability condition, separable NMF further enjoys a global optimal solution. However, separable NMF is unable to make use of data label information and thus unfavourable for supervised learning problems. In this paper, we propose discriminative separable NMF (DS-NMF), which extends separable NMF by encoding data label information into data representations. Assuming that each conical basis vector under the separability condition is only contributable to representing data from a few classes, DS-NMF exploits a structured sparse regularization to learning a sparse data representation and provides higher discrimination power than the standard separable NMF. Empirical evaluations on face recognition and scene classification problems confirm the effectiveness of DS-NMF and its superiority to separable NMF.

1. Introduction

Given a data matrix $X=[x_1,x_2,\ldots,x_n]\in\mathbb{R}_+^{p\times n}$, containing n nonnegative examples of p-dimensional vector space, non-negative matrix factorisation (NMF) [22,23,21] finds the a pair of nonnegative matrices $A\in\mathbb{R}_+^{r\times n}$ and $B\in\mathbb{R}_+^{p\times r}$, such that

$$X\approx BA. \quad (1)$$

The columns of B consists of a basis for the representation of X, while the columns of A store the coefficients of each data example under such a basis. In general, the column size r of B, i.e., the rank of the basis, is much less than the original data dimensionality p. Therefore, NMF leads to compact data representation and data compression. In addition, the non-negativity of A and B generally gives rise to more natural and interpretable data representations than other matrix factorization methods [17,9], which makes NMF a favourable model for a wide-range of applications, from text topic modelling, signal separation, social networks, collaborative filtering, dimension reduction, sparse coding and feature selection.

Different metrics can be used to measure the approximation residual between X and BA, such as matrix norms or information-theoretic quantities (e.g., divergences), up to the intention of modelling data properties. In this paper, we use the Frobenius matrix norm to measure the approximation residual, i.e., we optimize A and B by

$$\min_{A\in\mathbb{R}_+^{r\times n},B\in\mathbb{R}_+^{p\times r}} \|X-BA\|_F^2. \quad (2)$$

However, it is worth emphasizing that the results of this study is readily extendable to NMFs with other metrics.

1.1. Separable nonnegative matrix factorization

Although NMF provides a favourable approach for find compact data representations, the computation of the global optimal solution of (2) or forms with other metrics for the approximation residuals is intractable. Most

* Corresponding author. Tel.: +86 21 38282928.
E-mail address: szwang.smu@gmail.com (S. Wang).

http://dx.doi.org/10.1016/j.sigpro.2015.10.021

practical NMF algorithms [24,18,20,8,16] solve (2) with local optima by alternating minimisation over A and B using different heuristics. It has been proved that NMF is generally NP-hard [31]. In addition, the non-convexity of NMF makes it difficulty to find a unique and globally optimal factorization. To overcome the above-mentioned drawbacks of NMF, additional assumptions on the data matrix can be used to transfer the original NMF into more amiable problems. In [9], the authors proposed the separability assumption on the data matrix X and showed that this ensures NMF to have a unique factorisation. The separability assumes that there exits a subset of columns $X_{\mathcal{I}}$ of X such that the rest of columns can be represented by a nonnegative combination of $X_{\mathcal{I}}$. Therefore, NMF with the separable assumption reduces to finding the subset index $\mathcal{I}$ and the coefficients of representing X under $X_{\mathcal{I}}$.

Definition 1.1 (*Separability*). A data matrix X is called separable if there is a subset index $\mathcal{I} \subset [n]$, $|\mathcal{I}| = k$, and a permutation matrix Π such that

$$X\Pi = X_{\mathcal{I}}A, \quad \text{with } A = \begin{bmatrix} I_k & A' \end{bmatrix} \tag{3}$$

where I_k is the identity matrix of size k and $A' \in \mathbb{R}_+^{k\times(n-k)}$.

Geometrically, the separability of X can be interpreted as: the data examples in $X_{\mathcal{I}}$ generates a convex cone cone $(X_{\mathcal{I}})$, and the other data examples in X are located within cone $(X_{\mathcal{I}})$. Since a finitely generated convex cone has a unique set of extreme rays from its generators, NMF with the separable assumption is unique. Further, if we can allow approximate data representation or noise contamination, separable NMF can be formulated into

$$\begin{aligned} &\min_{\mathcal{I},\Pi,A} \|X\Pi - X_{\mathcal{I}}A\|_F^2 \\ &\quad \text{subject to} \quad \mathcal{I} \subset [n], |\mathcal{I}| = k \\ &\quad \Pi \text{ is a permutation matrix} \\ &\quad A = \begin{bmatrix} I_k & A' \end{bmatrix}, \ A' \in \mathbb{R}_+^{k\times(n-k)} \end{aligned} \tag{4}$$

In addition, under the mild regularity condition that $X_{\mathcal{I}}$ cannot be represented by combinations of the rest of the examples in X, it is possible to get rid of the permutation matrix Π by using the following equivalent form of (4)

$$\begin{aligned} &\min_{\mathcal{I},A} \|X - X_{\mathcal{I}}A\|_F^2 \\ &\quad \text{subject to } \mathcal{I} \subset [n], |\mathcal{I}| = k, A \in \mathbb{R}_+^{k\times n} \end{aligned} \tag{5}$$

Several algorithms have been developed to solve the separable NMF Eq. (4) or (9.) [3,5,12,14], which are commonly motivated by the geometric interpretation of the separability of X. Specifically, these algorithms apply linear programming (LP) to detect the extreme rays or generators $X_{\mathcal{I}}$ of the convex cone $cone(X_{\mathcal{I}})$ and to find the combination weights of the rest of examples in X. Very efficient algorithms for separable NMF have also been proposed by using recursive projections [13] and randomised methods [33]. In addition, by using the idea of group sparsity, the separable NMF problem (5) can be relaxed into

$$\min_{W \geq 0} \|X - XW\|_F^2 + \varrho \sum_{i=1}^{n} \|W(i,:)\|_2, \tag{6}$$

from which the index $\mathcal{I}$ can be recovered by the nonzero rows of the optimal W and the coefficient matrix A can be obtained by $A = W(\mathcal{I},:)$, i.e., the nonzero rows. Such formulation has been used for unmixing hyperspectral images in a blind and fully constrained manner [1].

1.2. Notations

Throughout this paper, we use the following notations. Upper letter A denotes a matrix. $A_{\mathcal{I}}$ or $A(:,\mathcal{I})$ denotes a sub-matrix of A, where $\mathcal{I}$ is an index variable and $A_{\mathcal{I}}$ is composed by the corresponding columns of A indexed by $\mathcal{I}$. $A(i,:)$ denotes the i-th row of A. Lower letter a denotes a vector or a scalar. $a(\mathcal{I})$ denotes as sub-vector of a, indexed by $\mathcal{I}$. $[n]$ denotes the set $\{1,2,\ldots,n\}$. $\|A\|_F$ denotes the Frobenius norm of matrix A. $\|a\|_2$ denotes the ℓ_2 norm of vector a. $\nabla\mathcal{L}(\cdot)$ denotes the gradient of the loss function $\mathcal{L}(\cdot)$. $\mathcal{I}_1 \cap \mathcal{I}_2$ denotes the intersection between index sets $\mathcal{I}_1$ and $\mathcal{I}_2$. $\overline{\mathcal{I}}$ is the complementary of index set $\mathcal{I}$ with respect to $[n]$.

2. Discriminative separable nonnegative matrix factorisation

In the separable NMF (5), the coefficient matrix A provides a compact representation of original high-dimensional example in the data matrix X. However, such representation does not encode any discriminative information, if we know the labels of the data examples. To address this limitation of separable NMF in the supervised setting, we propose to exploit structured sparse regularization to construct a discriminative separable NMF (DS-NMF), so that the obtained low-dimensional representation A is more favourable for classification.

2.1. Discriminative separable NMF by structured sparse regularisation

Suppose the data examples in X are categorized into m classes. DS-NMF is built upon the assumption that each example X_i in the conical basis $X_{\mathcal{I}}$, $i \in \mathcal{I}$, only responds to a small number of the m classes, i.e., the corresponding row $A(i,:)$ of the coefficient matrix A is sparse in terms of the class distribution of the data examples. This motivates us to propose the following structured sparse regularization:

$$\mathcal{R}(A) = \sum_{i=1}^{k} \mathcal{R}(A(i,:)) = \sum_{i=1}^{k}\sum_{c=1}^{m} \|A(i,\mathcal{I}_c)\|_2, \tag{7}$$

where $\mathcal{I}_c$ is the index set of the c-th class, i.e., $\mathcal{I}_c = \{j : x_j \in X \text{ is in the } c - th \text{ class}\}$. For each row of A, $\sum_{c=1}^{m} \|A(i,\mathcal{I}_c)\|_2$ is actually a mixed ℓ_2/ℓ_1 norm over $A(i,:)$. It has been shown in recent studies on sparse learning that such mixed norm is powerful in encouraging group sparsity. In particular, if $\|A(i,\mathcal{I}_c)\|_2 = 0$ for a certain class c, it implies that $X_i \in X_{\mathcal{I}}$ does not contribute to the representation of the examples from that class.

Based upon separable NMF (5) and the structured sparse regularization $\mathcal{R}(A)$ in (7), we define DS-NMF as the