Operations Research Letters 43 (2015) 423–426

Contents lists available at ScienceDirect

Operations Research Letters

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

Parameter selection for nonnegative l_1 matrix/tensor sparse decomposition

Yiju Wang [a,*], Wanquan Liu [b], Louis Caccetta [c], Guanglu Zhou [c]

[a] School of Management Science, Qufu Normal University, Rizhao Shandong, China
[b] Department of Computing, Curtin University, Perth, Western Australia, Australia
[c] Department of Mathematics & Statistics, Curtin University, Perth, Western Australia, Australia

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form
26 January 2015
Accepted 8 June 2015
Available online 15 June 2015

Keywords:
Sparse decomposition
Regularization parameter
Global optimal solution
Threshold bound
Zero solution

ABSTRACT

For the nonnegative l_1 matrix/tensor sparse decomposition problem, we derive a threshold bound for the parameters beyond which all the decomposition factors are zero. The obtained result provides a guideline on selection for l_1 regularization parameters and extends the corresponding result on Lasso optimization problem.

1. Introduction

Obtaining a low-rank matrix from a given multi-dimensional data is a classical feature extraction process in data mining, which is usually formulated as a low-rank matrix decomposition(approximation) problem [4]. For example, for a set of n-dimensional observations, principal component analysis (PCA) amounts to computing the singular decomposition of the data matrix and projecting the n-dimensional data along several principal orthogonal eigenvectors [9]. While, the low-rank tensor decomposition(approximation) based on multi-linear algebra like CANDECOMP/PARAFAC (CP) and Tucker models [4] provide a unified framework for higher-order data analysis [4,7].

It should be noted that the density of the latent factors in matrix/tensor low-rank decomposition may destroy the supporting information behind the data and hence the decomposition cannot provide sufficient information [15]. For example, for a gene expression data set with 5000 genes for cancer patients, PCA can give a low dimensional representation which can help cluster cancer versus healthy patients [8]. However, in reality, we do not know in advance which genes should be expressed and hence the dense factors cannot provide sufficient information. To search the interpretable factors, sparsity needs to be imposed on the latent factors so that one can associate cancer versus no cancer with a small group of genes, and this results in matrix/tensor sparse decompositions [5]. Now, the sparsity strategy is widely used in signal processing, biostatistics, etc. [13–15].

Another example of applications in matrix/tensor sparse decompositions is clustering, which is widely used in information retrieval, databases, text and data mining, bioinformatics, market-basket analysis, and so on [2,10]. By grouping the data, the clustering can identify distinctive “checkerboard” patterns for a given data and hence some useful features can be extracted from the mass data. In essence, clustering is to group a set of objects (represented typically by a set of feature vectors) into distinct classes which can be modeled as partitioning the data set into clusters such that the feature vectors falling in the same cluster are close to each other and the vectors in different clusters are far away from each other [10,12]. The clustering problem can also be formulated as matrix/tensor sparse decompositions [12].

It is well known that l_1 regularization is an efficient way to control the sparsity of latent factors in sparse optimization and the strategy is widely used in signal processing [3,14] and biostatistics [15]. In traditional l_1 regularization problem, such as compressed sensing [16], speech emotion recognition [17], the regularization parameter selection is investigated theoretically [6,11]. However, for l_1 regularized matrix/tensor nonnegative sparse decompositions, the regularization parameter selection has not been inves-

[*] Corresponding author. Tel.: +86 86 633 3980468.
E-mail addresses: wang-yiju@163.com (Y. Wang), w.liu@curtin.edu.au (W. Liu), caccetta@maths.curtin.edu.au (L. Caccetta), g.zhou@curtin.edu.au (G. Zhou).

http://dx.doi.org/10.1016/j.orl.2015.06.005
0167-6377/© 2015 Elsevier B.V. All rights reserved.

tigated systemically in the literature. Based on this observation, we consider the l_1 regularization parameter selection for the matrix/tensor nonnegative sparse decomposition in this paper. More precisely, we provide a threshold bound of the regularization parameters beyond which all the optimal decomposition factors are zero. The obtained result provides a guideline on selection for l_1 regularization parameters. Furthermore, the result improves the norm balancing property for l_1 regularization parameters [12] and extends the corresponding results on the original Lasso function [11].

To end this section, we present some notations used in this paper. Throughout this paper, we use a lowercase letter, say x, to denote scalars, bold lowercase letter, say $\mathbf{x}$, to represent vectors, and $\mathbf{x}_i$ to denote ith entry of vector $\mathbf{x}$. We use a bold uppercase letter, say $\mathbf{A}$, to denote a matrix, and use $\mathbf{A}_{ij}$ to denote the ijth entry of matrix $\mathbf{A}$. We use a bold and calligraphic letter, say $\mathcal{A}$, to denote a higher order tensor. We use $\mathbf{x}^{(1)} \circ \mathbf{x}^{(2)} \circ \cdots \circ \mathbf{x}^{(m)}$ or $\circ_{i=1}^m \mathbf{x}^{(i)}$ to denote the out-product of vectors $\mathbf{x}^{(1)} \in \mathbb{R}^{n_1}, \mathbf{x}^{(2)} \in \mathbb{R}^{n_2}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^{n_m}$ with $i_1 i_2 \cdots i_m$th entry $\mathbf{x}^{(1)}_{i_1}\mathbf{x}^{(2)}_{i_2} \cdots \mathbf{x}^{(m)}_{i_m}$, $1 \le i_j \le n_j, j = 1, 2, \ldots, m$. We use $\|\cdot\|_1$ to denote l_1-norm of a vector or a matrix which refers to the sum of the absolute value of the entries, use $\|\cdot\|_2$ to denote the l_2-norm of a vector, and use $\|\cdot\|_\infty$ to denote the maximum of the absolute value of the vector entries. The inner product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ or two higher-order tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n_1\times\cdots\times n_m}$ is defined as

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i,j=1}^{m,n} \mathbf{A}_{ij}\mathbf{B}_{ij}, \quad \langle \mathcal{A}, \mathcal{B}\rangle = \sum_{i_1,\ldots,i_m=1}^{n_1,\ldots,n_m} \mathcal{A}_{i_1 i_2\cdots i_m}\mathcal{B}_{i_1 i_2\cdots i_m}.$$

The Frobenius norms of matrix $\mathbf{A}$ and tensor $\mathcal{A}$ are respectively defined as

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle}, \qquad \|\mathcal{A}\|_{\mathrm{F}} = \sqrt{\langle \mathcal{A}, \mathcal{A}\rangle}.$$

2. Sparsity analysis of latent factors in matrix/tensor decomposition

As a powerful tool in data analysis with multiple arrays, tensor has received much attention of researchers and from its similarities to matrix, tensor decomposition is proposed by exploring their multilinear algebra properties [4]. As massive amounts of data often lead to limitations and challenges in analysis, sparsity is often imposed on the latent factors to improve the analysis and inference learning [13]. Mathematically, the nonnegative sparse tensor decomposition is formulated as follows [12],

$$\begin{aligned} &\min \left\| \mathcal{A} - \sum_{j=1}^{K} \left(\circ_{i=1}^m \mathbf{x}^{(i,j)}\right) \right\|_{\mathrm{F}}^2, \\ &\text{s.t.} \quad \mathbf{x}^{(i,j)} \in R^{n_i} \text{ is nonnegative and sparse}, \\ &\quad i = 1, \ldots, m; j = 1, \ldots, K, \end{aligned}$$

where tensor $\mathcal{A} \in \mathbb{R}^{n_1\times\cdots n_m}$ is the given data which is often nonnegative. As the problem is NP-hard [1], it can be relaxed as follows by introducing l_1 regularization into the model,

$$\begin{aligned} &\min_{\mathbf{x}^{(i,j)} \ge \mathbf{0}, i=1,\ldots,m; j=1,2,\ldots,K} \left\| \mathcal{A} - \sum_{j=1}^{K} \left(\circ_{i=1}^m \mathbf{x}^{(i,j)}\right) \right\|_{\mathrm{F}}^2 \\ &\quad + \sum_{i=1}^{m} \lambda^{(i)} \|\mathbf{X}^{(i)}\|_1, \end{aligned} \tag{2.1}$$

where positive numbers $\lambda^{(i)}, i = 1, 2, \ldots, m$ are regularization parameters used to control the sparsity of latent factors $\mathbf{X}^{(i)} = (\mathbf{x}^{(i,1)}, \mathbf{x}^{(i,2)}, \ldots, \mathbf{x}^{(i,K)}), i = 1, 2, \ldots, m$.

If $K = 1$, then problem (2.1) reduces to

$$\min_{\mathbf{x}^{(i)} \ge \mathbf{0}, i=1,\ldots,m} \|\mathcal{A} - \circ_{i=1}^m \mathbf{x}^{(i)}\|_{\mathrm{F}}^2 + \sum_{i=1}^{m} \lambda^{(i)} \|\mathbf{x}^{(i)}\|_1. \tag{2.2}$$

If the tensor order is 2, then tensor $\mathcal{A}$ reduces to matrix $\mathbf{A}$ and problem (2.1) reduces to the following nonnegative matrix sparse decomposition arising in data mining such as PCA [5,8] and co-clustering [12],

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{Y} \ge \mathbf{0}} \quad \|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_{\mathrm{F}}^2, \\ &\text{s.t.} \quad \mathbf{X} \in \mathbb{R}^{I\times K}, \mathbf{Y} \in \mathbb{R}^{J\times K} \text{ are both sparse}. \end{aligned}$$

Correspondingly, its l_1 relaxed form is as follows

$$\min_{\mathbf{X},\mathbf{Y} \ge \mathbf{0}} \|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_{\mathrm{F}}^2 + \lambda^x \|\mathbf{X}\|_1 + \lambda^y \|\mathbf{Y}\|_1, \tag{2.3}$$

where $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(K)}) \in R^{I\times K}$ and $\mathbf{Y} = (\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(K)}) \in R^{J\times K}$. Furthermore, if matrix $\mathbf{X}\mathbf{Y}^\top$ is of rank 1, i.e., $K = 1$, then matrices $\mathbf{X}, \mathbf{Y}$ reduce to vectors $\mathbf{x} \in R^I, \mathbf{y} \in R^J$ and problem (2.3) reduces to

$$\min_{\mathbf{x},\mathbf{y} \ge \mathbf{0}} \|\mathbf{A} - \mathbf{x}\mathbf{y}^\top\|_{\mathrm{F}}^2 + \lambda^x \|\mathbf{x}\|_1 + \lambda^y \|\mathbf{y}\|_1, \tag{2.4}$$

which is reminiscent of the regularization form of the Lasso model in compressed sensing [3,6],

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1.$$

In the Lasso model, the Lagrangian multiplier λ is served as the regularization factor to control the sparsity of latent vector $\mathbf{x}$ and it was shown that when the regularization factor $\lambda > 0$ is sufficiently large, or more precisely, if $\lambda \ge 2\|\mathbf{A}^\top \mathbf{b}\|_\infty$, then its solution is a zero vector [11]. A similar question for models (2.1) and (2.3) is proposed naturally: Do the sufficient large values of the regularization factors guarantee that the optimality solutions of problems (2.1) and (2.3) are zero? What is the relation of the regularization parameters in controlling the sparsity of the latent factors?

To investigate these problems, we first present the following norm-balancing property of problem (2.1) established in [12] and the assumptions needed in subsequent analysis.

Lemma 2.1. *For any optimal solution* $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)})$ *of problem* (2.1) *with regularization factor* $(\lambda^{(1)}, \ldots, \lambda^{(m)})$, *it holds that* $\lambda^{(1)}\|\mathbf{X}^{(1)}\|_1 = \cdots = \lambda^{(m)}\|\mathbf{X}^{(m)}\|_1$.

Assumption 2.1. For any fixed positive regularization factor $(\lambda^{(1)}, \ldots, \lambda^{(m)})$, the optimization problems (2.1) and (2.2) both have a unique solution.

In the following analysis, we will first consider problem (2.2) and then extend the obtained results to problem (2.1).

Lemma 2.2. *Suppose Assumption 2.1 holds. Let* $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$ *and* $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)})$ *be the optimal solutions of problem* (2.2) *respectively with regularization factors* $(\lambda_x^{(1)}, \ldots, \lambda_x^{(m)})$ *and* $(\lambda_y^{(1)}, \ldots, \lambda_y^{(m)})$ *such that* $\prod_{i=1}^m \lambda_x^{(i)} = \prod_{i=1}^m \lambda_y^{(i)}$. *Then* $\|\mathcal{A} - \circ_{i=1}^m \mathbf{x}^{(i)}\|_{\mathrm{F}}^2 = \|\mathcal{A} - \circ_{i=1}^m \mathbf{y}^{(i)}\|_{\mathrm{F}}^2$ *and problem* (2.2) *has the same optimal objective function value.*

Proof. For given regularization factors $(\lambda_x^{(1)}, \lambda_x^{(2)}, \ldots, \lambda_x^{(m)})$ and $(\lambda_y^{(1)}, \lambda_y^{(2)}, \ldots, \lambda_y^{(m)})$, from the assumption on them, there exist $t_i, i = 1, 2, \ldots, m$ such that $\lambda_x^{(i)} = t_i \lambda_y^{(i)}$ and $\prod_{i=1}^m t_i = 1$. By the assumption,

$$\begin{aligned} &\|\mathcal{A} - \circ_{i=1}^m \mathbf{x}^{(i)}\|_{\mathrm{F}}^2 + \sum_{i=1}^{m} \lambda_x^{(i)} \|\mathbf{x}^{(i)}\|_1 \\ &\quad = \|\mathcal{A} - \circ_{i=1}^m t_i \mathbf{x}^{(i)}\|_{\mathrm{F}}^2 + \sum_{i=1}^{m} \lambda_y^{(i)} \|t_i \mathbf{x}^{(i)}\|_1 \\ &\quad \ge \|\mathcal{A} - \circ_{i=1}^m \mathbf{y}^{(i)}\|_{\mathrm{F}}^2 + \sum_{i=1}^{m} \lambda_y^{(i)} \|\mathbf{y}^{(i)}\|_1. \end{aligned}$$