NON-NEGATIVE TENSOR FACTORIZATION USING ALPHA AND BETA DIVERGENCES

Andrzej CICHOCKI1,∗, Rafal ZDUNEK1†, Seungjin CHOI2, Robert PLEMMONS3 and Shun-ichi AMARI1

1 Brain Science Institute, RIKEN, Wako-shi, Saitama 351-0198, JAPAN,
2 Pohang University of Science and Technology, KOREA,
3 Wake Forest University, USA

ABSTRACT

In this paper we propose new algorithms for 3D tensor decomposition/factorization with many potential applications, especially in multi-way Blind Source Separation (BSS), multidimensional data analysis, and sparse signal/image representations. We derive, compare and implement in MATLAB NTFLAB Toolbox three classes of algorithms: Multiplicative, Fixed Point Alternating Least Squares (FPALS) and Alternating Interior-Point Gradient (AIPG) algorithms. Some of the proposed algorithms are characterized by improved robustness, efficiency and convergence rates and can be applied for various distributions of data and additive noise.

Index Terms— Algorithms, Learning systems, Linear approximation, Signal representations, Feature extraction.

1. MODELS AND PROBLEM FORMULATION

Tensors (also known as n-way arrays or multidimensional arrays) are used in a variety of applications ranging from neuroscience and psychometrics to chemometrics [6,8,9,17-19]. Nonnegative matrix factorization (NMF), Non-negative tensor factorization (NTF) and parallel factor analysis PARAFAC models with non-negativity constraints have been recently proposed as promising sparse and quite efficient representations of signals, images, or general data [2-7,10-13]. From a viewpoint of data analysis, NTF is very attractive because it takes into account spacial and temporal correlations between variables more accurately than 2D matrix factorizations, such as NMF, and it provides usually sparse common factors or hidden (latent) components with physiological meaning and interpretation [9,15]. In most applications, especially in neuroscience (EEG, fMRI), the standard NTF or PARAFAC models were used [15,16]. In this paper we consider more general model referred to as 3D NTF2 model (in analogy to the Parafac2 model [17]) (see Fig. 1). A given tensor \( \mathbf{X} \in \mathbb{R}^{I \times T \times K} \) is decomposed to a set of matrices \( \mathbf{S}, \mathbf{D} \) and \( \{ \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K \} \) with nonnegative entries. Here and elsewhere, \( \mathbb{R}_+ \) denotes the nonnegative orthant with appropriate dimensions. The three-way NTF2 model can be described as

\[
\mathbf{X}_k = \mathbf{A}_k \mathbf{D}_k \mathbf{S} + \mathbf{E}_k, \quad (k = 1, 2, \ldots, K)
\]

where \( \mathbf{X}_k = \mathbf{X}_{\mathbf{ik}} = [x_{ik}]_{I \times T} \in \mathbb{R}^{I \times T} \) are frontal slices of \( \mathbf{X} \in \mathbb{R}^{I \times T \times K} \), \( K \) is the number of frontal slices, \( \mathbf{A}_k = [a_{irk}]_{I \times R} \in \mathbb{R}_+^{I \times R} \) are the basis (mixing matrices), \( \mathbf{D}_k \in \mathbb{R}_+^{R \times T} \) is a diagonal matrix that holds the \( k \)-th row of the \( \mathbf{D} \in \mathbb{R}_+^{K \times R} \) in its main diagonal, and \( \mathbf{S} = [s_{rt}]_{R \times T} \in \mathbb{R}_+^{R \times T} \) is a matrix representing sources (or hidden components or common factors), and \( \mathbf{E}_k = \mathbf{E}_{\mathbf{ik}} = [e_{irk}]_{I \times T} \in \mathbb{R}_+^{I \times T} \) is the \( k \)-th frontal slice of a tensor \( \mathbf{E} \in \mathbb{R}_+^{I \times T \times K} \) representing error or noise depending upon the application. The objective is to estimate the set of matrices \( \{ \mathbf{A}_k \}, (k, K) \), \( \mathbf{D} \) and \( \mathbf{S} \), subject to some non-negativity constraints and other possible natural constraints such as sparseness and/or smoothness. Since the diagonal matrices \( \mathbf{D}_k \) are scaling matrices they can usually be absorbed by the matrices \( \mathbf{A}_k \) by introducing column-normalized matrices \( \mathbf{A}_k := \mathbf{A}_k / \mathbf{D}_k \), so usually in BSS applications the matrix \( \mathbf{S} \) and the set of scaled matrices \( \mathbf{A}_1, \ldots, \mathbf{A}_K \) need only to be estimated. However, in such a case we may loose the uniqueness of the NTF representation ignoring scaling and permutation ambiguities. The uniqueness still can be achieved by imposing nonnegativity, sparsity and other constraints. The above NTF2 model is similar to the well known PARAFAC2 model with non-negativity constraints and Tucker models [6,15,17]. In the special case, when all matrices \( \mathbf{A}_k \) are identical, the NTF2 model can be simplified to the ordinary PARAFAC model with the non-negativity constraints described \( \mathbf{X}_k = \mathbf{A}_k \mathbf{D}_k \mathbf{S} + \mathbf{E}_k, \quad k = 1, \ldots, K \) or equivalently \( x_{itk} = \sum a_{irk}s_{rt}d_{kr} + e_{itk} \) or \( \mathbf{X} = \sum a_{ir} \odot s_{rt} \odot d_{kr} + \mathbf{E} \), where \( s_{rs} \) are rows of \( \mathbf{S} \) and \( a_{ir}, d_{kr} \) are columns of \( \mathbf{A} \) and \( \mathbf{D} \), respectively and \( \odot \) means outer product of vectors [8,9]. Throughout this paper, we use the following notation: the \( rt \)-th element of the matrix \( \mathbf{S} \) is denoted by \( s_{rt} \), \( x_{itk} = [\mathbf{X}_k]_{it} \) means the \( it \)-th element of the \( k \)-th frontal slice matrix \( \mathbf{X}_k \), \( \mathbf{A} = [\mathbf{A}_1; \mathbf{A}_2; \ldots; \mathbf{A}_K] \in \mathbb{R}_+^{K \times I \times T} \) is a columnwise unfolded matrix of the slices \( \mathbf{A}_k, \mathbf{a}_{rpr} = [\mathbf{A}_k]_{rpr} \). Analogously, \( \mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_K] \in \mathbb{R}_+^{K \times I \times T} \) is the columnwise unfolded matrix of the slices \( \mathbf{X}_k \) and \( x_{pt} = [\mathbf{X}]_{pt} \).
2. ALPHA AND BETA DIVERGENCES AND ASSOCIATED ALGORITHMS

To deal with the model (1) efficiently we adopt several approaches from constrained optimization and multi-criteria optimization, where we minimize simultaneously several cost functions using alternating switching between sets of parameters. Alpha and Beta divergences are two complimentary generalized cost functions which can be applied for NMF and NTF [1,3,4,5,20].

2.1. Alpha Divergence

Let us consider a flexible and general class of the cost functions, called α-divergence [1,3,4]:

\[ D_{\alpha}^{(a)}(X||\hat{A}S) = \frac{\sum_{pt}(x_{pt}^{p}\hat{A}S_{pt}^{p})_{1-\alpha} - \alpha x_{pt} + (\alpha - 1)|\hat{A}S|_{pt}}{\alpha(\alpha - 1)} \]

\[ D_{\alpha}^{(a)}(X_k||\hat{A}_kS) = \frac{\sum_{it}(x_{itk}^{\alpha}\hat{A}_kS_{it}^{\alpha})_{1-\alpha} - \alpha x_{itk} + (\alpha - 1)|\hat{A}_kS|_{it}}{\alpha(\alpha - 1)} \]

We note that as special cases of α-divergence for \( \alpha = 2, 0.5, -1 \), we obtain the Pearson’s, Hellinger’s and Neyman’s chi-square distances, respectively, while for the cases \( \alpha = 1 \) and \( \alpha \to 0 \), respectively. When these limits are evaluated one obtains for \( \alpha \to 1 \) the generalized Kullback-Leibler divergence (I-divergence) and for \( \alpha \to 0 \) the dual generalized KL divergence [1,3,4].

Instead of applying the standard gradient descent method, we use the nonlinearly transformed gradient approach as generalization of the exponentiated gradient (EG)[4]:

\[ \Phi(a_{irk}) \leftarrow \Phi(a_{irk}) - \eta_{irk} \frac{\partial D_{\alpha_k}^{(a)}(X_k||\hat{A}_kS)}{\partial \Phi(a_{irk})}, \]  

\[ \Phi(s_{rt}) \leftarrow \Phi(s_{rt}) - \eta_{rt} \frac{\partial D_{\alpha}^{(a)}(X||\hat{A}S)}{\partial \Phi(s_{rt})}, \]

where \( \Phi(x) \) is a suitably chosen function.

It can be shown that such a nonlinear scaling or transformation provides a stable solution and the gradients are much better behaved in the space \( \Phi \). In our case, we employ \( \Phi(x) = x^\alpha \), which leads directly to the new learning algorithm (for \( \alpha \neq 0 \) ) (the rigorous proof of local convergence similar to this given by Lee and Seung [12] is omitted due to a lack of space):

\[ a_{irk} \leftarrow a_{irk} \left( \frac{\sum_{t=1}^{T}(x_{itk}/[\hat{A}_kS]_{it})^{\alpha}s_{rt}}{\sum_{t=1}^{T}s_{rt}} \right)^{1/\alpha} \],

\[ s_{rt} \leftarrow s_{rt} \left( \frac{\sum_{p=1}^{K}a_{pr}/[\hat{A}S]_{pt})^{\alpha}}{\sum_{p=1}^{K}a_{pr}} \right)^{1/\alpha} \).

2.2. Beta Divergence

Regularized beta divergence for the NTF2 problem can be defined as follows:

\[ D_{\beta}^{(s)}(X||\hat{A}S) = \sum_{pt}(x_{pt}/[\hat{A}S]_{pt})^{\beta} - \alpha x_{pt} + (\alpha - 1)|\hat{A}S|_{pt} \]  

\[ + [\hat{A}S]_{pt}^{\beta}([\hat{A}S]_{pt})^{\alpha} \]  

\[ D_{\beta}^{(k)}(X_k||\hat{A}_kS) = \sum_{it}(x_{itk}/[\hat{A}_kS]_{it})^{\beta} - \alpha x_{itk} + (\alpha - 1)|\hat{A}_kS|_{it} \]  

\[ + [\hat{A}_kS]_{it}^{\beta}([\hat{A}_kS]_{it})^{\alpha} \]  

\[ k = 1, \ldots, K, \quad t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, I, \]

where \( \alpha_S \) and \( \alpha_{A_k} \) are small positive regularization parameters which control the degree of sparseness of the matrices \( S \) and \( \hat{A}_k \), respectively, and the \( L_1 \)-norms defined as \( ||S||_{L_1} = \sum_{t \in T} ||S||_{L_1} \) and \( ||\hat{A}_k||_{L_1} = \sum_{i \in I} ||\hat{A}_k||_{L_1} \) are introduced to enforce sparse representations of the solutions. It is interesting to note that for \( \beta = 1 \), we obtain the squared Euclidean distances expressed by the Frobenius norms \( ||X_k - \hat{A}_kS||_F^2 \), while for the singular cases, \( \beta = 0 \) and \( \beta = -1 \), the beta divergence has to be defined as limiting cases as \( \beta \to 0 \) and \( \beta \to -1 \), respectively. When these limits are evaluated one gets for \( \beta \to 0 \) the generalized Kullback-Leibler divergence (called I-divergence) and for \( \beta \to -1 \) we obtain the Itakura-Saito distance. The choice of the parameter \( \beta \) depends on the statistical distribution of the data and the beta
divergence corresponds to the Tweedie models. For example, the optimal choice of the parameter for the normal distribution is \( \beta = 1 \), for the gamma distribution is \( \beta \to 1 \), for the Poisson distribution \( \beta = 0 \), and for the compound Poisson \( \beta \in (-1,0) \). By minimizing the beta divergence, we have derived various kinds of NTF algorithms: Multiplicative based on the standard gradient descent, Exponentiated Gradient (EG), Projected Gradient (PG), Alternating Interior-Point Gradient (AIPG), or Fixed Point Alternating Least Squares (FPALS) algorithms. For example, in order to derive a flexible multiplicative learning algorithm, we compute the gradient of (6)-(7) with respect to elements of matrices \( s_{rt} = s_r(t) = [S]_{rt} \) and \( a_{irk} = [A_k]_{ir} \) and performing simple mathematical manipulations we obtain the multiplicative update rules:

\[
\begin{align*}
\eta_{ir} &\leftarrow \eta_{ir} \frac{\sum_{t=1}^{T} [x_{rtk}] / \left[ A_k S_{st}^{1-\beta} \right] s_{rt} - \alpha A_k}{\sum_{t=1}^{T} \left[ A_k S_{st}^{1-\beta} \right] s_{rt} - \alpha A_k}, \\
\theta_{st} &\leftarrow \theta_{st} \frac{\sum_{p=1}^{K} a_{pir} \left[ x_{rtk} / \left[ A S_{st}^{1-\beta} \right] \right] / \alpha S_{st} - \alpha S_{st}}{\sum_{p=1}^{K} a_{pir} \left[ A S_{st}^{1-\beta} \right] / \alpha S_{st} - \alpha S_{st}},
\end{align*}
\]

where \( \lfloor x \rfloor = \max \{\varepsilon, x\} \) with a small \( \varepsilon = 10^{-16} \) is introduced in order to avoid zero and negative values.

In the special case for \( \beta = 1 \) we can derive an alternative algorithm, called FPALS (Fixed Point Alternating Least Squares) algorithm (see [5] for detail):

\[
\begin{align*}
A_k &\leftarrow \left( A_k S^T - \alpha A_k E \right) (S S^T)^{+}, \\
S &\leftarrow \left( \left(A^T A\right)^{+} (A^T X - \alpha S E S) \right)^{+},
\end{align*}
\]

where \( \lfloor A \rfloor^+ \) denotes Moore-Penrose pseudo-inverse of \( A \) and \( E \in \mathbb{R}^{I \times R} \), \( E \in \mathbb{R}^{R \times T} \) are matrices with all entries one. The above algorithm can be considered as a nonlinear projected Alternating Least Squares (ALS) or nonlinear extension of EM-PCA algorithm.

Furthermore, using the Alternating Interior-Point Gradient (AIPG) approach [14], another efficient algorithm has been developed and implemented [5]:

\[
\begin{align*}
A_k &\leftarrow A_k - \eta A_k P A_k, \\
S &\leftarrow S - \eta S P S,
\end{align*}
\]

where \( P A_k = \left( A_k \odot (A_k S S^T) \right) \odot \left( (A_k S - X_k) S^T \right) \), \( P S = \left( S \odot (A^T A S) \right) \odot \left( A^T A S - X \right) \) and operators \( \odot \) and \( \odot \) mean component-wise multiplication and division, respectively. The learning rates \( \eta A_k \) and \( \eta S \) are selected in this way to ensure the steepest descent, and on the other hand, to maintain non-negativity. Thus, \( \eta S = \min \{ \tau \eta S, \eta S \} \) and \( \eta A_k = \min \{ \tau \eta A_k, \eta A_k \} \), where \( \tau \in (0,1) \), \( \eta S = \{ \eta : S = \eta S \} \) and \( \eta A_k = \{ \eta : A_k = \eta A_k \} \) ensure non-negativity, and

\[
\begin{align*}
\eta^{*}_{A_k} &\leftarrow \frac{\text{vec}(P A_k)^T \text{vec}(A_k S S^T - X_k S^T)}{\text{vec}(P A_k S)^T \text{vec}(P A_k S)}, \\
\eta^{*}_{S} &\leftarrow \frac{\text{vec}(P S)^T \text{vec}(A^T A S - A^T X)}{\text{vec}(A_k P S)^T \text{vec}(A_k P S)}
\end{align*}
\]

are the adaptive steepest descent learning rates.

3. SIMULATION RESULTS

All the NMF algorithms discussed in this paper have been extensively tested for many difficult benchmarks for signals and images with various statistical distributions and also for real EEG data. We found the best performance can be obtained with the AIPG, FPALS and the algorithm (8)-(9) for \( \beta = 1 \).

Due to space limitation, we present here only one simulation Example: Six natural highly correlated images are mixed by randomly generated 3D tensor \( A \in \mathbb{R}^{12 \times 64 \times 25} \). The observed mixed data are collected in 3D tensor \( X \in \mathbb{R}^{12 \times 64 \times 25} \). The exemplary results are shown in Fig.2.

4. CONCLUSIONS AND DISCUSSION

In this paper we proposed generalized and flexible cost functions (controlled by a single parameter alpha or beta) that allows us to derive a family of robust and efficient NTF algorithms. The optimal choice of a free parameter of a specific cost function depends on a statistical distribution of data and additive noise, thus various criteria and algorithms (upgrading rules) should be applied for estimating the basis matrices \( A_k \) and the source matrix \( S \), depending on \( \text{a priori} \) knowledge about the statistics of noise or errors. It is worth to mention that we can use three different strategies to estimate common factors (the source matrix \( S \)). In the first approach, presented in this paper, we use two different cost functions: A global cost function (using unfolded column-wise matrices: \( X \), \( S \) for frontal slices of 3D tensors) to estimate the common factors \( S \), i.e., the source matrix \( S \); and local cost functions to estimate the slices \( A_k \), \( k = 1, 2, \ldots, K \). However, instead of using the unfolding matrices for the NTF2 model, in order to estimate \( S \), we can use, average matrices defined as \( \bar{X} = \sum_k X_k \in \mathbb{R}^{I \times T} \) and \( A = \sum_k A_k \in \mathbb{R}^{I \times T} \). Furthermore, it is also possible to apply a different approach by using only set of local cost functions, e.g., \( D_k(X_k||A_k S) = 0.5||X_k - A_k S||_F^2 \). In such a case, we estimate \( A_k \) and \( S \) cyclically by applying alternating minimization (similar to row-action projection of the Kaczmarz algorithm). We found that such approaches also work quite well for the NTF2 model. The advantage of the last approach is that the all updates learning rules are local (slice by slice) and algorithms are generally faster for large data, (especially, if \( K >> 1 \)).

Obviously, 3D NTF models can be transformed to a 2D non-negative matrix factorization (NMF) problem by unfold-
of a single unfolded 2-D matrix. The profiles of the stacked sub-matrices should not be considered as equal to a standard 2-way NMF and/or discover some inner structures in the data. Consequently, imposing additional, natural constraints such as smoothness, continuity, closure, unimodality, local rank, selectivity, etc.) to get full information about the available data and/or discover some inner structures in the data. Obviously, there are many challenging open issues remaining, such as global convergence, optimal choice of parameters and uniqueness of a solution when additional constraints are imposed.

5. REFERENCES