Approximate L0 constrained Non-negative Matrix and Tensor Factorization

Mørup, Morten; Madsen, Kristoffer Hougaard; Hansen, Lars Kai

Published in:
2008 IEEE International Symposium on Circuits and Systems

Link to article, DOI:
10.1109/ISCAS.2008.4541671

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Approximate $L_0$ constrained Non-negative Matrix and Tensor Factorization

Morten Mørup, Kristoffer Hougaard Madsen and Lars Kai Hansen
Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs Lyngby, Denmark
Email: {mm.khm,lkh}@imm.dtu.dk

Abstract—Non-negative matrix factorization (NMF), i.e. $V \approx WH$ where both $V$, $W$ and $H$ are non-negative has become a widely used blind source separation technique due to its part based representation. The NMF decomposition is not in general unique and a part based representation not guaranteed. However, imposing sparseness both improves the uniqueness of the decomposition and favors part based representation. Sparseness in the form of attaining as many zero elements in the solution as possible is appealing from a conceptional point of view and corresponds to minimizing reconstruction error with an $L_0$ norm constraint. In general, solving for a given $L_0$ norm is an NP hard problem thus convex relaxation to regularization by the $L_1$ norm is often considered, i.e., minimizing $(\frac{1}{2}\|V-WH\|_F^2 + \lambda\|H\|_1)$. An open problem is to control the degree of sparsity $\lambda$ imposed. Here we demonstrate that a full regularization path for the $L_1$ norm regularized least squares NMF for fixed $W$ can be calculated at the cost of an ordinary least squares solution based on a modification of the Least Angle Regression and Selection (LARS) algorithm forming a non-negativity constrained LARS (NLARS). With the full regularization path, the $L_1$ regularization strength $\lambda$ that best approximates a given $L_0$ can be directly accessed and in effect used to control the sparsity of $H$. The MATLAB code for the NLARS algorithm is available for download.

I. INTRODUCTION

Non-negative matrix factorization (NMF) is a blind source separation algorithm (BSS) given by the decomposition

$$V_{n,m} \approx \sum_d W_{n,d} H_{d,m},$$  

(1)

where $V \in \mathbb{R}^{N \times M}$, $W \in \mathbb{R}^{N \times D}$ and $H \in \mathbb{R}^{D \times M}$, i.e. such that the variables $V$, $W$ and $H$ are non-negative. The decomposition is useful because it results in easy interpretable part based representations [1]. Non-negative decompositions is also named positive matrix factorization [2] but was popularized by Lee and Seung due to a simple algorithmic procedure based on multiplicative updates [3]. The decomposition has proven useful for a wide range of data where non-negativity is a natural constraint. These encompass data for text-mining based on word counts, image data, biomedical data, and spectral data. The algorithm can also be useful even when the data in itself is negative by considering, say, the amplitude of a spectral representation [4]. The model is often estimated alternatingly solving for $W$ and $H$. While the estimation of $H$ for fixed $W$ is a convex problem, the joint problem solving for $W$ and $H$ is not convex and as such a global minimum is not guaranteed. For data with modality more than two, i.e., tensors, purely non-negative decompositions are also relevant. The two most widely used decompositions for tensors are the PARAFAC model independently proposed in [5], [6] and the Tucker model [7]. The two model read

\[
\begin{align*}
V_{i_1,i_2,...,i_N} & \approx \sum_d A_{i_1,d}^{(1)} A_{i_2,d}^{(2)} \cdots A_{i_n,d}^{(n)} \quad \text{PARAFAC:} \\
V_{i_1,i_2,...,i_N} & \approx \sum_{j_1,j_2,...,j_N} g_{j_1,j_2,...,j_N} A_{i_1,j_1}^{(1)} A_{i_2,j_2}^{(2)} \cdots A_{i_N,j_N}^{(n)} \quad \text{Tucker:}
\end{align*}
\]

In [8] it was demonstrated how NMF generalizes well to the PARAFAC model and in [9] how NMF also generalizes to the Tucker model forming fully non-negative decompositions as for regular NMF resulting in part based representation. The generalization of NMF to the two models are based on the observation that by matricizing the array, i.e. $V_{(n)} \in \mathbb{R}^{l_n \times l_1l_2 \cdots l_{n-1}l_{n+1} \cdots l_N}$ the optimization of the loading of each mode can be restated as regular NMF problem, for details on this see [9]. As such, the PARAFAC and Tucker models can be estimated through algorithms for regular matrix decomposition alternating solutions for each mode keeping the other modes fixed. We will here only treat the matrix case keeping in mind that the approach generalizes to tensors.

The non-negative decomposition is in general not unique [10] which hampers interpretation of the decomposition. Furthermore, the representation is not guaranteed to be part-based for instance if the data resides well inside the positive orthant. To improve the uniqueness of the decomposition as well as enforcing a part based representation sparseness constraints have been suggested for the NMF decomposition. This has been implemented by penalizing with the $L_1$ norm [11], [12], i.e.

$$C_{\text{Sparse NMF}} = \frac{1}{2}\|V - WH\|_F^2 + \lambda\|H\|_1,$$  

(2)

where $\|H\|_p$ is the $L_p$-norm of $H$ given by $L_p(H) = \sqrt[\lambda]{\sum_{d,m} |H_{d,m}|^p}$ (Notice, for the $L_0$ norm we use $L_0(H) = \sum_{d,m} |H_{d,m}|^0$). The $L_1$ norm is often used as a penalty for sparsity as it is known to mimic the behavior of the $L_0$ norm [13] while keeping the alternating optimization problem estimating $H$ for fixed $W$ convex. Thus, although often the aim is to find a sparse solution with respect to the $L_0$-norm,
in practice the NP-hard optimization problem is sought solved with the \( L_1 \)-norm as a proxy to avoid local minima in the optimization problem.

In [14] the sparseness of the decomposition was controlled by the sparseness measure

\[
s_{PH-L1/L2}(H_{d,:}) = \frac{1}{\sqrt{M} - 1} \left( \sqrt{M} - \frac{\|H_{d,:}\|_1}{\|H_{d,:}\|_2} \right),
\]

where \( H_{d,:} \) denotes the \( d \)-th row of \( H \). Notice, \( s_{PH-L1/L2} \) takes values between 0 and 1. This was achieved solving a constrained minimization problem forming the Non-negative matrix factorization with Sparseness Constraint (NMFSC) algorithm described in [14]. Solving for a specific sparsity level for each component is a difficult problem. However, solving for a specific sparsity on the full matrix \( H \) mounts to controlling the single parameter \( \lambda \) which we presently demonstrate has a simple solution.

Consider the unconstrained Least squares minimization with \( L_1 \)-norm constraint also referred to as the Least Absolute Shrinkage and Selection Operator (LASSO) [15] or Basis Pursuit De-noising (BPD) [16], i.e., the problem

\[
C_{\text{LASSO}} = \frac{1}{2} \| y - X\beta \|_F^2 + \lambda \| \beta \|_1.
\]

From the Least Angle Regression and Selection algorithm (LARS) [17] or the equivalent homotopy algorithm [18] the entire regularization path, i.e., solution for all values of \( \lambda \) can be achieved for the cost of an ordinary least squares (OLS) solution. We will in the present paper modify the LARS algorithm such that the entire regularization path can be found for the Sparse NMF decomposition. With this result the sparsity can be directly controlled for the following two new types of sparsity measures on the full matrix \( H \)

\[
s_{PL0}(H) = 1 - \frac{\|H\|_0}{DM},
\]

\[
s_{PL1/L2}(H) = \frac{1}{\sqrt{DM} - 1} \left( \sqrt{DM} - \frac{\|H\|_1}{\|H\|_2} \right).
\]

The measure \( s_{PL1/L2}(H) \) is equivalent to the measure proposed in [14], but for the full matrix rather than for each row of \( H \) and is included to give a qualitative comparison to the measure given in equation 3. \( s_{PL0} \) on the other hand gives direct control of the percentage of active elements in \( H \), i.e., the \( L_0 \) norm of \( H \). Both measures take values between 0 and 1, for instance \( s_{PL0}(H) = 1 \) implies that all elements of \( H \) are turned off while \( s_{PL0}(H) = 0 \) that all elements are nonzero.

## II. Method

In the paper on Least Angle Regression and Selection (LARS) [17] an algorithm was proposed for the LASSO by considering normalized data \( X \) and \( y \) formed by an active set algorithm. Let \( A \) be the indices of the active set and \( I \) the indices of the inactive set. Then the LARS algorithm solves for the LASSO as

\[
\text{NLARS}
\]

while \( \exists j \in I : |c_j| > 0 \)

\[
c = X^\top (y - X\beta)
\]

\[
j = \text{arg max}(|c_j|)
\]

\[
A = [A \ j]
\]

\[
I = I \setminus j
\]

\[
\beta_A = \beta_A + \mu(X^\top X)^{-1}A A^\top \text{sign}(c_A)
\]

\[
\mu = \text{arg min}_\mu \begin{cases} 
\exists \beta_A = 0 \text{ then } I = [I \ A_k], A = A \setminus A_k \\
\exists \mu \in I : |c| = |c|_A \\
\hat{c}_A = 0
\end{cases}
\]

where \( \hat{c} \) is the gradient and \( \beta \) the value of \( \beta \) at the step of size \( \mu \). We make the following observations

1. \( c = X^\top (y - X\beta) \) is identical to the negative gradient of the un-regularized problem.
2. Since the inverse Hessian is given by \((X^\top X)^{-1}\) the step \((X^\top X)^{-1}A A^\top \text{sign}(c_A)\) will be in the direction such that the amplitude of the gradients in the active set are identical, i.e. \(|c_A| = |c_A| = \ldots = |c_A|\).
3. Since the amplitude of the gradients of the active set are identical the update correspond to the Newton-Raphson step \( \beta_A = \beta_A + \gamma (X^\top X)^{-1}A A^\top c_A \) where \( \gamma = \frac{\mu}{\mu A^\top c_A} \).
4. The gradient of the \( L_1 \) penalty is \( \lambda \text{sign}(\beta) \) and as such thresholds the gradient of the un-regularized problem such that the solution for a given regularization \( \lambda \) is found when \(|c_A| = \lambda|\).

From observation one there is no reason to normalize the data. This serves only the purpose of interpreting the selection by angles, however, in general the selection is according to the largest absolute gradient. Add 2 and 3 it follows that the update is along the line where the gradients \( c_A \) in the active set are identical and corresponds to a Newton-Raphson step in the active variables. Notice, for least squares problems the inverse Hessian is a constant within each active set forming the piecewise linearity of the LARS path. Since the gradient of the \( L_1 \) constraint is \( \lambda \text{sign}(\beta) \), \( \lambda \) serves as a threshold of the gradient of the LASSO problem given by \( g = -c + \lambda \text{sign}(\beta) \). Thus, does not allow for additional elements to enter the active set when \(|c_j| < \lambda|\). By simply replacing the criterion \( j = \text{argmax}(|c_j|) \) with \( j = \text{argmax}(|c_j|) \) for \( c_j > 0 \), i.e. considering only negative gradients of the least squares objective - variables are only introduced if they reduce the objective by becoming positive. Once a variable is updated to zero it is taken out of the active set. Thus, the following modification of the above LARS algorithm forms the non-negativity constrained LARS (NLARS) (i.e., for \( \beta \geq 0 \)). For more details on the algorithm see the MATLAB code available from [19].
now be separated into $N$ LASSO problems where the $n^{th}$ problem is given by
\[
C_{\text{SparseNMF}} = \frac{1}{2} \| V_n - WH_n \|_F^2 + \lambda \| h_n \|_1
\] (7)

Notice, the algorithm works even when $W$ and $V$ are unconstrained. Having solved for each column of $H$ we need to order the given LASSO solutions. However, this is easy since we can directly calculate the value of $\lambda$ corresponding to each part of the regularization path of each given non-negative LASSO solution as
\[
\lambda = (W_A^T v_m - (W^T W)_{A,A} h_{A,m}).
\] (8)

Notice, all entries in $\lambda$ take the same value $\lambda$. Hence, the solutions of each column of $H$ can be sorted according to this value. The sparseness measures given in equation 5 can now be controlled by evaluating the sparseness degree over the entire regularization path. Where several values of $\lambda$ result in same sparseness degree, the smallest value of $\lambda$ was chosen as regularization strength. Let $\cdot$ and $\cdot \cdot$ denote element wise division and multiplication, respectively. Then, $W$ was updated according to the normalization invariant multiplicative updates proposed in [11] given by
\[
W \leftarrow W \cdot \frac{VH + W \text{diag}(1(WH + W) \cdot W))}{WH + W \text{diag}(1(W) \cdot W))},
\] (9)
such that $W_d$ is constrained to have unit $L_2$ norm, i.e., $W_{n,d} = \frac{W_{n,d}}{\| W_{n,d} \|_2}$. We solve for $H$ alternatingly according to the NLARS and return the $H$ corresponding to the correct degree of the $sp_{PL}$ criterion and update $W$ as given above. The algorithm was stopped after 100 iterations or when the relative change in $\lambda$ was less than $10^{-9}$.

III. RESULTS

We tested our $sp_{PL}$ constrained NMF algorithm on the 2429 training faces of size 361 pixels obtained from the MIT face database [20] (see figure 1) as well as the US Postal Service (USPS) database of 7291 handwritten digit each of size 256 pixels - the training set (see figure 2). We compared our sparseness measure given in equation 5 to the corresponding decomposition obtained from the same degree of sparseness within each components and as such each component has different degree of sparsity. Thus, component 16 is completely turned off when $sp_{PL} = 0.96$.

IV. DISCUSSION

We have proposed an algorithm for non-negative matrix factorization with an $L_0$ norm constraint based on convex relaxation, i.e. invoking an $L_1$ norm as a proxy. This has the benefit that the NP-hard problem of constraining $H$ by a given $L_0$-norm, $sp_{PL}$, for a fixed $W$ can be approximated by a convex problem which is efficiently solved through the also proposed NLARS algorithm. The main benefit is that the entire regularization path can be found for the cost of an ordinary least squares solution. By evaluating the $L_0$ norm of the solution throughout the path, we can find the appropriate strength of the $L_1$ regularization term $\lambda$ that will provide a specific value of the $L_0$ norm.

In [14] sparseness was controlled by constraining each component to a given degree of sparseness. The present measures of sparseness given by controlling the degree of sparsity on the full matrix $H$ rather than each row $H_d$, has the benefit that the control of regularization reduces to the single parameter $\lambda$. This is a much simpler problem in which the entire regularization path formed by the NLARS algorithm can be used to directly control this value. In general, controlling the sparseness componentwise versus controlling...
the sparsity of the full matrix as here have different properties. It was demonstrated for the MIT face data that sparseness imposed on the full matrix enabled some components to be less sparse than others and also for components to be turned off completely whereas controlling the degree of sparseness of each component separately enabled us to fix the number of components used to represent the data. Both for the MIT face as well as the USPS data the results obtained with the proposed algorithm corresponded well to the results obtained for the corresponding degree of sparsity using the NMFSC algorithm. Thus, controlling the sparseness of the full matrix $H$ through the NLARS is a viable approach to control the sparsity of the solutions.

Both the present NLARS algorithm as well as the fNNLS algorithm given in [21] for non-negativity constrained least squares minimization are based on an active set approach. As seen from figure 3 the two algorithms had about the same computational cost. Thus, the present NLARS algorithm is comparable to ’state of the art’ non-negative least squares algorithms such as the fNNLS while also allowing for finding the full regularization path solution.

The sparsity measures proposed can also be used to control the degree of sparsity for sparse coding [22], i.e., unconstrained decompositions with $L_1$-norm regularization by the regular LARS algorithm rather than the non-negative extension presently derived. Finally, it is worth emphasizing that solving for the $sp_{PLD}$ problem can be done at the computational cost. Thus, the present NLARS algorithm is seen from figure 3 the two algorithms had about the same computational cost. Hence the full regularization path can be calculated at the cost of the ordinary LS solution given by fNNLS.

**REFERENCES**