MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Sparse and Shift-Invariant Feature Extraction From Non-Negative Data

Paris Smaragdis, Bhiksha Raj, Madhusudana Shashanka

TR2008-013 August 2008

Abstract

In this paper we describe a technique that allows the extraction of multiple local shift-invariant features from analysis of non-negative data of arbitrary dimensionality. Our approach employs a probabilistic latent variable model with sparsity constraints. We demonstrate its utility by performing feature extraction in a variety of domains ranging from audio to images and video.

ICASSP 2008

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2008
201 Broadway, Cambridge, Massachusetts 02139

SPARSE AND SHIFT-INVARIANT FEATURE EXTRACTION FROM NON-NEGATIVE DATA

Paris Smaragdis
Adobe Systems
Newton, MA, USA

Bhiksha Raj
Mitsubishi Electric Research Laboratories
Cambridge, MA, USA

Madhusudana Shashanka
Mars Inc.
Hackettstown, NJ, USA

ABSTRACT

In this paper we describe a technique that allows the extraction of multiple local shift-invariant features from analysis of non-negative data of arbitrary dimensionality. Our approach employs a probabilistic latent variable model with sparsity constraints. We demonstrate its utility by performing feature extraction in a variety of domains ranging from audio to images and video.

Index Terms— Feature extraction, Unsupervised learning

1. INTRODUCTION

The extraction of thematic components from data has long been a topic of active research. Typically, this has been viewed as a problem of deriving bases that compose the data, with the bases themselves representing the underlying components. Techniques such as Principal or Independent Component Analysis, or, in the case of non-negative data, Non-negative Matrix Factorization, excel at the discovery of such bases and have found wide use in multiple applications.

In this paper we focus on the discovery of components that exhibit the properties of being multidimensional, local, and shift-invariant in arbitrary dimensions. We consider local instead of global features, i.e. features that have a small support as compared to the input and individually only describe a limited section of it. Extracting such local features also necessitates the use of shift-invariance, the ability to have such features appear in arbitrary locations throughout the input. An example case where such flexibility is required is on images of text where the features (letters in this case) are local and appear with arbitrary shifting. Additionally we consider the case or arbitrary rank for both input and feature data, i.e. we consider inputs which are not just 2-D structures like images or matrices, but higher rank objects which can be composed from features of arbitrary rank. Using traditional decomposition techniques, such as those mentioned above, obtaining this flexibility is a complex, if possible, process. Each of these properties have been individually addressed in the past, but not in an unified and extensible manner.

In this paper we focus on resolving these problems and enabling more intuitive feature extraction for a wide range of inputs. We specifically concentrate on non-negative data which are often encountered when dealing with representations of audio and visual data. We use a probabilistic interpretation of non-negative data in order to derive flexible learning algorithms which are able to efficiently extract shift-invariant features in arbitrary dimensionalities. An important component of our approach is imposing sparsity in order to extract meaningful components. Due to the flexibility of our approach we will also introduce an entropic prior which can directly optimize the sparsity of any estimated parameter in our model (be it a component, or its weight). Finally we will show how we can discover patterns in diverse data such as audio, images and video, and also apply the technique to deconvolution of positive-only data.

2. SPARSE SHIFT-INVARIANT PROBABILISTIC LATENT COMPONENT ANALYSIS

We will begin by assuming without loss of generality that any N-dimensional non-negative data is actually a scaled N-dimensional distribution. To adapt the data to this assumption, we normalize it to sum to unity. The scaling factor can be multiplied back into the discovered patterns later if desired, or incorporated in the estimation procedures thereby resolving any quantization issues. We now describe our algorithm sequentially by first introducing the basic probabilistic latent component analysis model, extending it to include shift-invariance, and finally by imposing sparsity on it.

2.1. Probabilistic Latent Component Analysis

The Probabilistic Latent Component Analysis (PLCA) model, which is an extension of Probabilistic Latent Semantic Indexing (PLSI) [1] to multi-dimensional data, models any distribution $P(\mathbf{x})$ over an N-dimensional random variable $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ as the sum of a number of *latent* N-dimensional distributions that in turn are completely specified by their marginal distributions:

$$P(\mathbf{x}) = \sum_z P(z) \prod_{j=1}^N P(x_j|z) \tag{1}$$

Here z is a latent variable that indexes the latent component distributions, and $P(x_j|z)$ is the marginal distribution of x_i within the z^{th} component distributions, i.e. the *conditional* marginal distribution conditioned on z. The objective of the decomposition is to discover the most appropriate marginal distributions $P(x_j|z)$.

The estimation of the marginals $P(x_j|z)$ is performed using a variant of the EM algorithm. In the expectation step we estimate the 'contribution' of the latent variable z:

$$R(\mathbf{x}, z) = \frac{P(z) \prod_{j=1}^N P(x_j|z)}{\sum_{z'} P(z') \prod_{j=1}^N P(x_j|z')} \tag{2}$$

and in a maximization step we re-estimate the marginals using the above weighting to obtain a new and more accurate estimate:

$$P^*(z) = \int P(\mathbf{x}) R(\mathbf{x}, z) d\mathbf{x} \tag{3}$$

$$P^*(x_j|z) = \frac{\int \cdots \int P(\mathbf{x}) R(\mathbf{x}, z) dx_k, \forall k \neq j}{P^*(z)} \tag{4}$$

Estimates of all conditional marginals and the mixture weights $P(z)$ are obtained by iterating the above equations to convergence. Note that when the x_i are discrete (such as the X and Y indices of pixels in an image), the integrals in the above equations become summations over all values of the variable. An example of a PLCA analysis is shown in figure 1. The only parameter that needs to be defined by the user is the number of states that the latent variable assumes. In this example case z assumes three states.

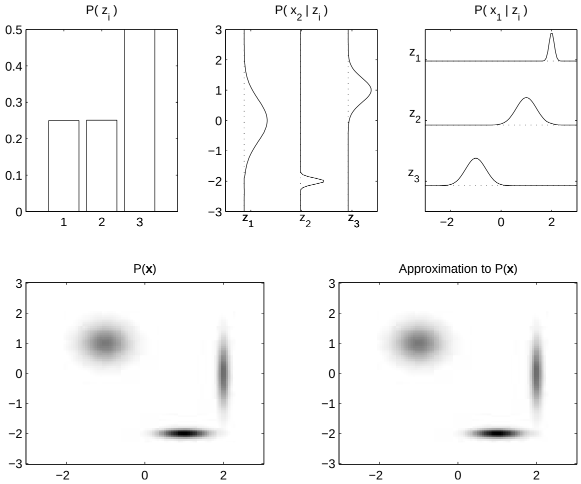

Fig. 1. Illustration of PLCA. The 2-D distribution $P(\mathbf{x})$ in the lower left plot is composed of three Gaussians. Upon applying PLCA we see that the three Gaussians have been properly described by their marginals, and the latent priors reflect their correct mixing weights as show in the top row. The lower right plot shows the approximation of $P(\mathbf{x})$ using the weighted sum of the discovered marginal products. Note that for readability on paper the colormap of some figures in this paper is inverted so that large values are dark and zero is whlte.

2.2. Shift-Invariance

The basic PLCA model describes distribution in terms of marginals and hence cannot detect shifted patterns. We now extend it to deal with the latter. We modify the PLCA model of Equation 1 to:

$$P(\mathbf{x}) = \sum_z \left(P(z) \int P(\mathbf{w}, \boldsymbol{\tau}|z) P(\mathbf{h} - \boldsymbol{\tau}|z) d\boldsymbol{\tau} \right) \tag{5}$$

where $\mathbf{w}$ and $\mathbf{h}$ are mutually exclusive subsets of components, $\mathbf{w} = \{x_i\}$, $\mathbf{h} = \{x_i\}$, such that $\mathbf{x} = \{\mathbf{w}, \mathbf{h}\}$. $\boldsymbol{\tau}$ is a random variable that is defined over the same domain as $\mathbf{h}$. This decomposition uses a set of *kernel distributions* $P(\mathbf{w}, \boldsymbol{\tau}|z)$ which when convolved with their corresponding *impulse distributions* $P(\mathbf{h} - \boldsymbol{\tau}|z)$ and appropriately weighed and summed by $P(z)$ approximate the input $P(\mathbf{x})$. The exact set of components x_i that go into $\mathbf{w}$ and $\mathbf{h}$ must be specified.

Estimation of $P(z)$, $P(\mathbf{w}, \boldsymbol{\tau}|z)$ and $P(\mathbf{h}|z)$ is once again done using Expectation-Maximization. The expectation step is:

$$R(\mathbf{x}, \boldsymbol{\tau}, z) = \frac{P(z) P(\mathbf{w}, \boldsymbol{\tau}|z) P(\mathbf{h} - \boldsymbol{\tau}|z)}{\sum_{z'} P(z') \int P(\mathbf{w}, \boldsymbol{\tau}'|z') P(\mathbf{h} - \boldsymbol{\tau}'|z') d\boldsymbol{\tau}'} \tag{6}$$

And the parameter updates in the maximization step are:

$$P^*(z) = \int R(\mathbf{x}, z) d\mathbf{x} \tag{7}$$

$$P^*(\mathbf{w}, \boldsymbol{\tau}|z) = \frac{\int P(\mathbf{x}) R(\mathbf{x}, \boldsymbol{\tau}, z) d\mathbf{h}}{P^*(z)} \tag{8}$$

$$P^*(\mathbf{h}|z) = \frac{\int P(\mathbf{w}, \mathbf{h} + \boldsymbol{\tau}) R(\mathbf{w}, \mathbf{h} + \boldsymbol{\tau}, \boldsymbol{\tau}, z) d\mathbf{w} d\boldsymbol{\tau}}{\int P(\mathbf{w}, \mathbf{h}' + \boldsymbol{\tau}) R(\mathbf{w}, \mathbf{h}' + \boldsymbol{\tau}, \boldsymbol{\tau}, z) d\mathbf{h}' d\mathbf{w} d\boldsymbol{\tau}} \tag{9}$$

The above equations are iterated to convergence.

Figure 2 illustrates the effect of shift-invariant PLCA of a distribution. We note that both component distributions in the figure are discovered. As in PLCA, the number of components must be specified. In addition the desired size of the kernel distributions (i.e. the area of their support) must also be specified.

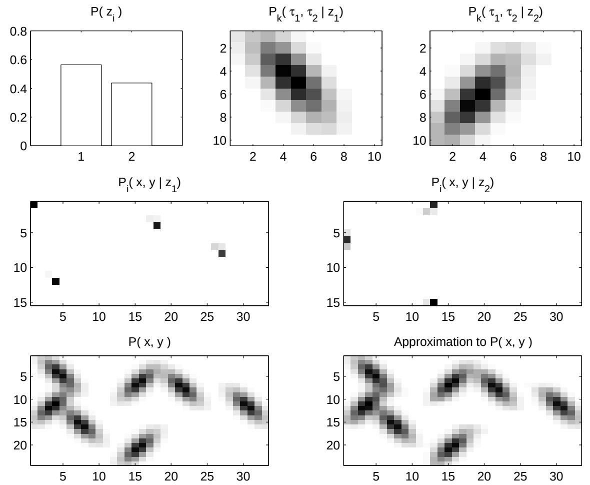

Fig. 2. Illustration of shift-invariant PLCA. The top left plot displays the latent variable priors, whereas the remaining two top plots display the two kernel distributions we extracted. The second row of plots display the impulse distributions, whereas the bottom row displays the original input distribution at the left and the model approximation at the right.

2.3. Sparsity Constraints

The shift-invariant decomposition of the previous section is by nature indeterminate. The kernels and impulse distributions are interchangeable – there is no means of explicitly specifying what must be the kernel and what the impulse. To overcome this indeterminacy, we now impose a restriction of *sparsity*. We do so by explicitly imposing an *entropic* a-priori distribution on some of the component distributions to minimize their entropy. To do so we use the methodology introduced by Brand [3].

Let $\boldsymbol{\theta}$ be any distribution in our model whose entropy we wish to bias during training. We specify the *a priori* distribution of $\boldsymbol{\theta}$ as $P(\boldsymbol{\theta}) = e^{-\beta \mathcal{H}(\boldsymbol{\theta})}$, where $\mathcal{H}(\boldsymbol{\theta})$ is the entropy of $\boldsymbol{\theta}$. The parameter β is used to indicate the severity of the prior, and can also assume negative values to encourage high entropy estimates.

The imposition of the entropic prior does not modify the update rules of shift-invariant PLCA. It only introduces two additional steps that are used in a two iteration loop to refine the estimate of $\boldsymbol{\theta}$:

$$\frac{\omega}{\theta_i} + \beta + \beta log \theta_i + \lambda = 0 \tag{10}$$

$$\boldsymbol{\theta} = \frac{-\omega/\beta}{\mathcal{W}(-\omega e^{1+\lambda/\beta}/\beta)} \tag{11}$$

where $\mathcal{W}(\cdot)$ is Lambert's function. If $\boldsymbol{\theta} = P(x_j|z)$ then ω is given by:

$$\omega = \int \cdots \int P(\mathbf{x}) R(\mathbf{x}, z) dx_k, \forall k \neq j \tag{12}$$

where R is given by Equation 6. A more detailed description of the entropic prior for PLCA appears in [4].

Note that $\boldsymbol{\theta}$ can be any of the distributions in our model, i.e., we could impose sparsity on either the kernel distributions or the impulse distributions (and even the latent variable priors). The effect of this manipulation is illustrated in figure 3 where, for the same input, pairs of kernel and impulse distributions with varying entropic priors are shown. Even though all three cases result into qualitatively similarly good explanations of the input, clearly they all result in different analyses. As may be inferred from this figure, the most helpful

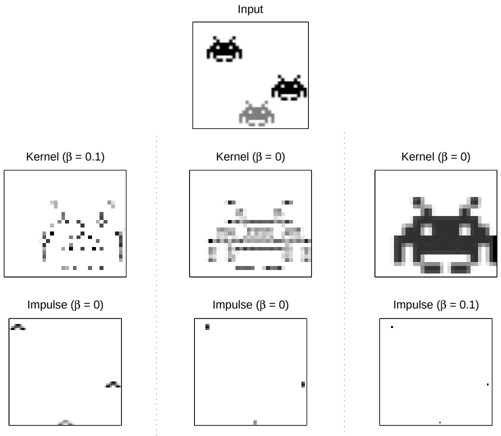

Fig. 3. Illustration of the effects of sparse shift-invariant PLCA. The top plot is the input $P(\mathbf{x})$. The leftmost column shows the kernel (top) and impulse (bottom) distributions obtained when the kernel is made sparse. The center column is obtained when no sparsity is imposed. In the right column, the impulse distribution has been made sparse.

use of the entropic prior is when we employ it to enforce on the impulse distributions, permitting the kernel distributions themselves to be most informative.

3. APPLICATIONS OF SPARSE SHIFT-INVARIANT PLCA

3.1. Audio Examples

Patterns in audio signals are often discernible in their time-frequency representations, such as spectrograms. The magnitude spectrogram (the magnitude of the STFT) of an audio signal is inherently non-negative and amenable to analysis by our algorithm. We illustrate the use of PLCA on spectrograms with the example in figure 4. The input time-frequency distribution in this figure represents a passage of a few piano notes from a real piano recording. The specific time-frequency distribution used is constant-Q [5]. In this particular type of distribution notes appear as as a series of energy peaks across frequency (each peak corresponding to a harmonic). Different notes will be appropriately shifted vertically according to their fundamental frequency. Aside from this shifting, the shape of the harmonic peaks will remain the same. Likewise shifting in the time (horizontal) axis denotes time delay. In a musical piece we would expect to find the same harmonic pattern shifted across both dimensions so as to represent each played note.

As shown in figure 4 applying PLCA on this input results in a very concise description of its content. We automatically obtain a kernel representing the harmonic series of a piano note. The peaks in the impulse distribution represent the fundamental frequency of the note and its location in time. We thus effectively discover the building blocks of this music passage and perform a rough transcription in an unsupervised manner.

3.2. Applications in Images and Video

Color images are normally represented as a three dimensional structure spanning two spatial and one color dimension. In the context of this paper an image can be represented as a distribution $P(x, y, c)$, where x and y are the two spatial dimensions and c is a 3-valued index for each of the primary colors (red, green, blue). Consider then the color image in figure 5. This image is composed of three distinct handwritten characters each associated with a color (an “α”, a “γ” and “1”). Due to the handwriting different instances of the same characters are not identical. Figure 5 also displays the results of sparse shift-invariant PLCA analysis of the image. We have estimated three kernel functions. The estimated kernels are found to represent the three characters, and the corresponding impulse distributions represent their locations in the image.

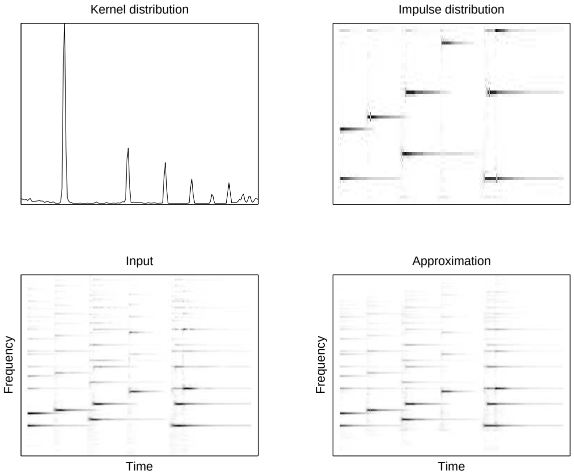

Fig. 4. An input constant-Q spectrogram of a few overlapping piano notes is shown in the bottom left figure. The top left plot displays the kernel distribution across the frequency (horizontal) axis, which we can see being a harmonic series describing the generic structure of a musical note. The impulse distribution shown in top right, identifies the locations in time and frequency of the basic note. As before, the reconstruction of the input is shown in the bottom right plot.

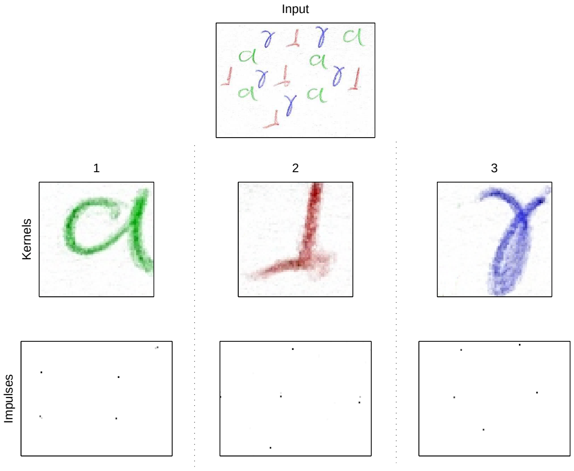

Fig. 5. Analysis of a color image. The input is shown in the top panel and the three columns of plots under it display the pairs of kernel and impulse distributions that were extracted.

This method is also applicable to video streams which can be viewed as four-dimensional data (two spatial dimensions, one color dimension and time). Results from a video analysis are shown in figure 6.

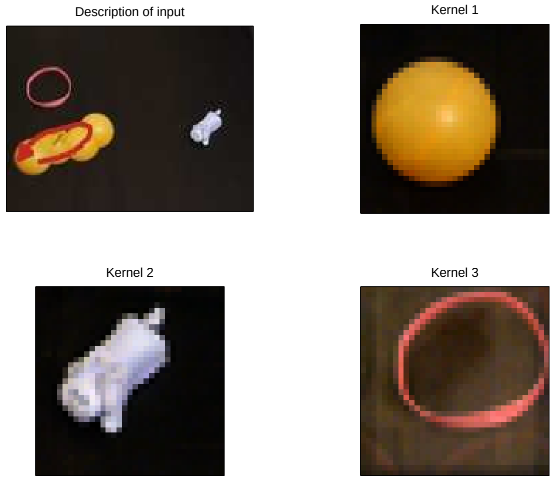

Fig. 6. Application of PLCA on video data. In the top left plot we see a time-lapse description of a video stream which was represented as a four dimensional input {x,y,color,time}. An arrow is superimposed on the figure to denote the path of the ball. The remaining plots display the three components that were extracted from the video (the impulse distributions are not displayed being 3-dimensional structures).

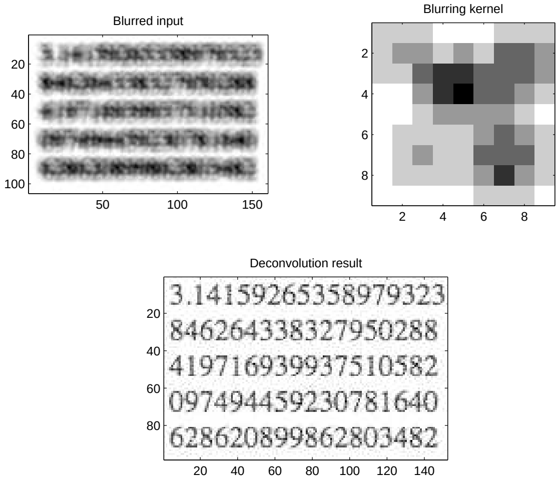

Fig. 7. An example of positive deconvolution. The top left image is an image of numbers that have been convolved with the top right image. Performing PLCA with the known blurring function as a kenrl and estimating only the impulse distribution we can recover the image before the blurring.

3.3. Positive Deconvolution

Shift-invariant PLCA can also be used for positive deconvolution in arbitrary dimensions. So far the model we have been describing is a generalization of a multidimensional sum of convolutions where neither the filters, nor the impulses are known. If we simplify this model to only use a single convolution (i.e. having z assume only one state), and assume that the kernel distribution is known, we then obtain the classical definition of a positive-only deconvolution problem. The only differences in this particular case is that instead of optimizing with respect to an MSE error, we are doing so over a Kullback-Leibler distance between the input and the estimate. In terms of the perceived quality of the results there is no significant difference however. To illustrate this application consider the data in figure 7. In the top left image we have a series of numbers which have been convolved with the kernel shown in the top right. PLCA decomposition was performed on this data, fixing the kernel to the value shown in the top right. The estimated impulse distribution is shown at the bottom of figure 7, and as is clear it has removed the effects of the convolution and has recovered the digits before the blurring without presenting any problems by recovering any negative values.

4. DISCUSSION

We note from the examples that the sparse shift-invariant decomposition is able to extract patterns that adequately describe the input in semantically rich terms from various forms of non-negative data. We have also demonstrated that the model can be used or positive data deconvolution.

As mentioned earlier, the basic foundation of our method is a generalization of the Probabilistic Latent Semantic Indexing (PLSI) [1] model. Interestingly, in the absence of shift-invariance and sparsity constraints, it can also be shown through algebraic manipulation that for 2-D data the model is, in fact, also identical numerically to non-negative matrix factorization. The shift-invariant form of PLCA is also very similar to the convolutive NMF model [2]. The fundamental contribution in this work is the extension of these models to shift-invariance in multiple dimensions, and the enforcement of sparsity that enables us to extract semantically meaningful patterns in various data.

The formulation we have used in the paper is extensible. For instance, the model can also allow for the kernel to undergo transforms such as scaling, rotation, shearing etc. It is also amenable to the application of a variety of priors. Discussion of these topics is outside the scope of this paper, but their implementation is very similar to what we have presented so far.

5. REFERENCES

[1] Hofmann, T., “Probabilistic Latent Semantic Indexing” in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (SIGIR’99).

[2] Smaragdis, P., “Convolutive Speech Bases and their Application to Supervised Speech Separation”, *IEEE Transaction on Audio, Speech and Language Processing*, January 2007.

[3] Brand, M.E., “Structure Learning in Conditional Probability Models via an Eutropic Prior and Parameter Extinction”, *Neural Computation Journal*, Vol. 11, No. 5, pp. 1155-1182, July 1999.

[4] Shashanka, M.V.S., “Latent Variable Framework for Modeling and Separating Single Channel Acoustic Sources”, Ph.D. Thesis, Department of Cognitive and Neural Systems. Boston University, Boston 2007.

[5] Brown, J.C., “Calculation of a Constant Q Spectral Transform,”, *Journal of Acoustical Society of America*, vol. 89(1), January 1991.

[6] Lee, D.D., Seung, S.H., “Algorithms for Non-negative Matrix Factorization”, *Advances in Neural Information Processing Systems*, Vol. 13 (2000), pp. 556-562.