Chapter 2
Nonnegative Matrix Factorizations for Intelligent Data Analysis

G. Casalino, N. Del Buono and C. Mencar

Abstract We discuss nonnegative matrix factorization (NMF) techniques from the point of view of intelligent data analysis (IDA), i.e., the intelligent application of human expertize and computational models for advanced data analysis. As IDA requires human involvement in the analysis process, the understandability of the results coming from computational models has a prominent importance. We therefore review the latest developments of NMF that try to fulfill the understandability requirement in several ways. We also describe a novel method to decompose data into user-defined—hence understandable—parts by means of a mask on the feature matrix, and show the method's effectiveness through some numerical examples.

Keywords Nonnegative matrix factorization · Intelligent data analysis

2.1 Introduction

The amount of available data has grown dramatically over the past 50 years. Every year more than 200 Exabytes of data are generated on Internet.[1] Huge quantities of digital data are produced daily from different sources: numerical data from satellites or sensors, textual data (both structured and unstructured) from websites, emails,

[1]Source: Cisco® Visual Networking Index (VNI) Forecast (2010–2015).

G. Casalino (✉) · C. Mencar
Department of Informatics University of Bari, University Campus "Ernesto Quagliariello", Via E. Orabona, 4, 70125 Bari, Italy
e-mail: gabriella.casalino@uniba.it

C. Mencar
e-mail: corrado.mencar@uniba.it

N. Del Buono
Department of Mathematics University of Bari, University Campus "Ernesto Quagliariello", Via E. Orabona, 4, 70125 Bari, Italy
e-mail: nicoletta.delbuono@uniba.it

G.R. Naik (ed.), *Non-negative Matrix Factorization Techniques*,
Signals and Communication Technology, DOI 10.1007/978-3-662-48331-2_2

forums, newsgroups, public and private digital archives, images, and videos, are just some examples. Data overload is a fact of life for all of us in the information era.

Although this profusion of information potentially allows to satisfy all information needs, it also presents some limits; the larger is the amount of data the fewer are the possibilities to capture, discover, and understand useful knowledge to guide action or decision making [8, 76].

Clearly human capabilities prove to be unsuitable to process big amounts of data, therefore automatic mechanisms, which are able to assist humans in extracting useful information and knowledge from rapidly growing volumes of digital data are indispensable and an extensive effort of research in this direction has been made in the last years.

Intelligent data analysis (IDA) aims to the intelligent application of human expertise and computational models for advanced data analysis. Automatic tools, which strive for involving the analyst in the process of data analysis and extracting useful patterns from big data, can be enumerated among IDA methods. In this scenario, techniques coming from different areas (such as statistics, artificial intelligence, data mining, machine learning, optimization, dynamic programming), which favor the interaction with users and produce understandable knowledge, could be favorably exploited in IDA.

Nonnegative matrix factorizations (NMF) are powerful techniques recently proposed to uncover latent low-dimensional structures intrinsic in high-dimensional data and provide a nonnegative, part-based, representation of data [5, 25, 40, 69, 70, 115]. Nonnegativity enhances meaningful interpretations of mined information and distinguishes NMF from other traditional dimensionality reduction algorithms, such as principal component analysis (PCA) [65] or singular value decomposition (SVD) [45].

However, the understandability of the results, obtained by applying classical NMF, is not guaranteed a priori, as they often do not correspond with the intuitive notions of parts in the original data. Several variants of constraints and various regularization terms have been proposed to improve NMF capabilities so as to make the extracted parts easier to understand by the data analyst.

This chapter aims to review such techniques from the point of view of IDA, by stressing on their understandability capabilities and usefulness as tools for IDA.

Along with the review of NMF techniques suited for IDA, we describe an approach for injecting user knowledge in the factorization process, by masking the factor matrix (one of the products of NMF) [13]. Masking enables the decomposition of data into user-defined parts, which are consequently easy to understand by the analyst. The results of Masked NMF enables the analyst to understand which subset of the available data are best represented by the specified parts, thus extracting potentially useful knowledge from large quantities of data.

In the next section, we give an overview of IDA and its objective, while we focus on NMF techniques in Sects. 2.3 and 2.4. In Sect. 2.4.3.1, we describe Masked NMF along with some numerical examples to show the effectiveness of the method. In Sect. 2.5, the use of the NMF algorithms for intelligently analyze educational data is illustrated. Future perspectives are sketched in the conclusive section.

2.2 Intelligent Data Analysis

Data are collections of values or measurements. They can be numbers, words, observations, or even descriptions of things. In this chapter, we will simply refer to data as a collection of numerical values recording the magnitude of different attributes and/or features that describe the problem under study.

Hand writes *"data analysis is what we do when we turn data into information"* [50]. Intelligent data analysis is the intelligent way to do it. Moreover, he gives a definition of information: *"It is what we extract from data when we attempt to answer some questions. Before extracting information which can shed light on a question, one must be clear about what that question is"*. This is a crucial point in IDA: the analysis is driven by the questions that the analyst wants to answer to, otherwise it would be *unintelligent*.

IDA is an iterative process that enables the combination of human expertise and computational models to automatically extract useful patterns, event correlations and in general, understandable knowledge which would otherwise remain hidden in the data under consideration [6]. A data analyst could be interested in describing data by finding patterns and anomalies, or just by summarizing them; in this case, the term *exploratory data analysis* is used. Contrariwise, when the analyst is interested in verifying some hypotheses about the structure in data, e.g., differences among groups of data, evolution of the attribute values, etc., the term *confirmatory data analysis* is used. IDA is a multidisciplinary discipline that comes from the intersection of several research fields, the most important ones are statistics and machine learning.

IDA and knowledge discovery from data (KDD) are tightly correlated, yet with some noteworthy differences. Both are aimed at identifying valid, novel, potentially useful, and ultimately understandable patterns in data [36]; however, IDA emphasizes the importance of the prior knowledge possessed by human experts that intelligently guide the analysis process in an interactive and iterative way [7]. Data mining is one step of the KDD process and refers to the set of tools that allow to *automatically* extract knowledge from large amounts of data [36]. However, a full automatization of the data analysis process is impossible [7], for this reason IDA is focused on the human contribution to the analysis process.

Holmes and Peek categorize IDA methods in three main classes: data exploration, classification and prediction, and dimensionality reduction [54]. Data exploration plays a fundamental role in data analysis. Analysts look at data for discovering relations among features, trends, anomalies or outliers, relations among features, classes, etc. Most of these techniques rely on visual tools to represent information. IDA-based approaches for data exploration integrate automatic techniques with a priori user knowledge in the exploration process, thus enabling user interaction. Classification and prediction methods are used in several domains dealing with real data. Machine learning literature provides many different techniques for classification (both supervised, semi-supervised, or unsupervised) and prediction. Most of them are based on some automatic learning tools to acquire knowledge that can be used for classifying (or predicting) unobserved data. However, only few of them are

capable of yielding knowledge that is intelligible to users (e.g., knowledge expressed in form of rules), a mandatory requirement for their use within IDA. Learning interpretable knowledge from data is a topic of current research in Machine Learning and Computational Intelligence. In this context are located dimensionality reduction techniques that represent data in a reduced space through feature selection and extraction. This facilitates to manage, understand, and visualize data. Because of their tight relationship with NMF, a brief overview of such techniques is outlined in the following subsection.

2.2.1 Dimensionality Reduction Techniques

Often, in high-dimensional data not all the measured variables are "important" for understanding the underlying phenomena of interest. Hence, mechanisms that transform data and reduce the number of original variables are frequently used.

Let $X \in \mathbb{R}^{n \times m}$ be the observation data matrix, where each columns vector is composed by n observations for each of the m-dimensional variable in $x = (\mathbf{x_1}, \ldots, \mathbf{x_m})^\top$. In this formalization, the dimension of data is meant the number of variables that are measured on each observation, while the term dimensionality of X indicates the number m of original features. A dimensionality reduction method is a transformation of a given data matrix X into a meaningful representation $S \in \mathbb{R}^{n \times k}$ of reduced dimensionality $k \leq m$ [108]. The low-dimensional vectors $s = (\mathbf{s_1}, \ldots, \mathbf{s_k})^\top$, with $k \leq m$ capture information in the original data, according to some particular criteria. The components of s are called "hidden components" or "latent factors," while—depending on the particular research context one is working with— the m multivariate vectors are alternatively named "variables," "attributes" or "features." Dimensionality reduction methods mitigate *the curse of dimensionality* [1], which refers to difficulties related to data analysis when data dimensionality increases; these methods are able to overcome problems coming from data sparseness and noise, and can be adopted as a visualization tool to show multivariate data in a human intelligible form.

Dimensionality reduction techniques can be categorized in two classes: (i) *feature selection* and (ii) *feature extraction*. A feature selection method is a process that selects a subset of k original (and supposed relevant) features for spanning a reduced space that may better describe the phenomena of interest. Feature selection mechanisms reduce the computational costs, but a good trade-off between accuracy of the results and efficiency is needed.

On the other hand, feature extraction methods try to capture hidden properties of data and discover the minimum number of uncorrelated or lowly correlated factors that can be used to better describe the phenomena of interest. It is accomplished by the creation of new features obtained as functions of the original data. Reduction of the computational complexity of data both in time (for elaboration) and in space (for storage) and the discovery of latent structure hidden in data, (meaningful structures and/or unexpected relationships among variables) are some of the advantages resulting from feature extraction methods.

The simplest dimensionality reduction methods are linear and derive each of the $k \leq m$ components of the new variables in S as a linear combination of the original variables:

$$S = XA, \tag{2.1}$$

or equivalently

$$X = SB, \tag{2.2}$$

being $A \in \mathbb{R}^{m \times k}$ and $B \in \mathbb{R}^{k \times m}$ appropriate linear transformation weight matrices. Equation (2.2) makes clear the motivation why the new variables in S are called hidden or latent factors. (PCA) [56, 65, 93], factor analysis (FA) [102], independent component analysis (ICA) [60], linear discriminant analysis (LDA) [38], and CUR decomposition [85] are all well-known linear dimensionality reduction techniques used for analyzing multivariate data. Among linear dimensionality reduction methods, the most widely used in the context of IDA is PCA.

2.2.1.1 Principal Component Analysis

Principal component analysis is the best, in the least square error sense, linear dimensionality reduction technique [62, 65]. It is based on the covariance matrix of the variables and seeks to reduce the dimensionality of data matrix X by finding few orthogonal linear combinations (the principal components—PCs) of the original variables with the largest variance. The first PC is the linear combination of the original data with the largest variance; the second PC is the linear combination with the second largest variance and orthogonal to the first PC, and so on. The principal components are given by

$$Y = XU, \tag{2.3}$$

where $U \in \mathbb{R}^{m \times m}$ is an orthogonal weight matrix computed as the orthogonal factor of the spectral decomposition of the covariance matrix $X^\top X$ of the standardized data matrix X.[2] Therefore, the columns of the matrix U are the eigenvectors of the covariance matrix. These eigenvectors (principal axes) map a data vector from the original space of m variables to a new space of k variables which are uncorrelated over the dataset. Hence keeping only the first $k < m$ principal components a dimensional reduction on k-dimensional subspace of the original data is derived.

Moreover, it is proven that the transformed data matrix, obtained by only considering the first $k < m$ principal components, is the best least squares k-approximation of the original data X (this result is known as the *Eckart–Young–Mirsky theorem* [46]).

[2]Since the values of the variance of data depends on the scale of the variables, usually the original data contained in X are subject to a standardization process so that each variable has mean zero and standard deviation one.

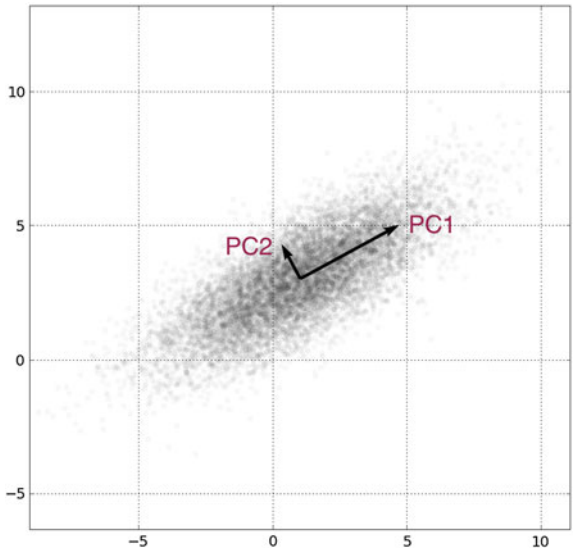

Fig. 2.1 Graphical illustration of PCA. From Wikipedia, the free encyclopedia

Figure 2.1 shows the behavior of PCA of a data matrix collecting points that belong to a bivariate Gaussian distribution centered in the coordinates (1, 3). Standard deviation of data is 3 in the direction (0.878, 0.478) and 1 in the orthogonal direction. The first principal component ($PC1$) captures information in the direction of the maximum variability in data; instead, the second principal component ($PC2$) is orthogonal to the first one and captures information in the second most variable direction. The principal axes are therefore the bases of the rotated space and are centered in the center of the points. This is a simple example where the dimensionality of the original data space and that of the transformed one are the same. As an example of dimensionality reduction, one can represent the same points using the first principal axis only; in this case, a one-dimensional space is obtained where data points are projected onto.

In many applications, the most of data variance can be captured by the first two (or three) PCs; this makes the PCA a widely used visualization tool in IDA. However, even though the PCs are uncorrelated variables constructed as linear combinations (with mixed signs) of the original variables, and have some desirable properties (they are orthogonal and ordered in a decreasing manner w.r.t. the variance of original data), they do not necessarily correspond to meaningful physical quantities. Hence, a clear interpretation of the results provided by PCA is sometimes difficult to be derived.

To clarify this point consider the computer vision problem of human face recognition, where PCA has been largely adopted to obtain a set of basis images—*the eigenfaces*—that can be linearly combined to reconstruct images in the original dataset of face [107]. As it can be observed by Fig. 2.3 (left panel), eigenfaces are not physically intuitive and far to correspond to what humans use to explain why a face is a face. In particular, because of the presence of negative signs in the components of

principal axes, PCA reconstructs the original data adding up some basis images and subtracting others; this may not make sense in some applications. A simply question can be posed:"What does it mean to subtract a face basis?"

These considerations can be extended to documents, genes, preferences, questionaries and to all nonnegative data. In the following Sect. 2.3, a review of Nonnegative Matrix Factorization is given. It is able to represent original data by only additive, not subtractive, combinations of some basis vectors. This characteristic of parts-based representation is appealing because it reflects the intuitive notion of combining parts to form a whole providing more distinct and clearer dimensionality reduction results and a easier understandability of the obtained results.

2.3 Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a computational technique for linear dimensionality reduction of a given data matrix X, which is able to explain data in terms of additive combination of nonnegative factors that represent realistic *building blocks* for the original data (provided that data are nonnegative too) [25, 40, 69, 70, 115].

The nonnegativity constraint is useful for learning part-based representations and has a twofold motivation. First in many applications one knows that the quantities involved cannot be negative (for example by the rules of physics). Second, intuitively parts are generally combined additively (and not subtracted) to form a whole and physiological principles assume that humans learn objects as part-based [70]. Hence, nonnegativity potentially enhances meaningful interpretations of information mined from a given data matrix, allowing to a better understanding of the results obtained by the analysis process; this makes NMF a suitable computational models for IDA.

2.3.1 NMF Mathematical Formulation

Formally, given a nonnegative data matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}_+^{n \times m}$, where $\mathbf{x}_i \in \mathbb{R}_+^n$ are n-dimensional column vectors representing samples,[3] NMF aims to approximate X into the product of two lower rank nonnegative matrices—a *basis matrix* $W = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k] \in \mathbb{R}_+^{n \times k}$ and an *encoding matrix* $H = (h_{ij}) \in \mathbb{R}_+^{k \times m}$—such that

$$X \approx WH, \tag{2.4}$$

[3] Henceforth a matrix is denoted with an uppercase letter, e.g., X, its elements with the corresponding lowercase letter, e.g., x_{ij}, a column vector in lowercase boldface, e.g., $\mathbf{x}_i$.

or, equivalently,

$$\mathbf{x}_j \approx \sum_{i=1}^{k} \mathbf{w}_i h_{ij}. \tag{2.5}$$

where W and H both have nonnegative elements (namely, $W \geq 0$ and $H \geq 0$) and the product matrix (WH) is of rank k with $(n+m)k \leq nm$.

To compute a nonnegative matrix factorization (2.4) of a given data matrix X, some quality measures have to be taken into account to evaluate how well the product (WH) approximates the data matrix X. Particularly, some divergence function

$$D : \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{r\times m} \to \mathbb{R}_+,$$

can be adopted. It should be observed that the divergence D is a function of the factor matrices W and H, but it is also parametrized by the input data matrix X. This dependence can be expressed by writing $D(X; W, H)$ [104]. Using the previous formalization, the NMF problem may be rewritten as a nonlinear constrained optimization problem over the divergence D, that is,

$$\min_{W\geq 0, H\geq 0} D(X; W, H). \tag{2.6}$$

The most frequently adopted instance of (2.6) leads to the minimization of

$$\min_{W\geq 0, H\geq 0} D(X; W, H) = \|X - WH\|_F^2, \tag{2.7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Many other divergence measures have also been used, the interested reader can refer to [30].

The NMF performs a conical coordinate transformation; indeed, geometrically the basis vectors generate a simplicial cone which contains the original data and which is contained in the positive orthant [24, 33, 59].

It should be pointed out that the value of the parameter k (the *rank* of the factorization) is problem dependent and user-specified. It identifies the number of factors to be used to explain data and plays a fundamental role in the factorization process. In fact, different values of k lead to different factorization results.

2.3.2 Interpretation of the Basis and Encoding Matrices

The results of NMF applied to a data matrix X have an immediate geometrical interpretation. According to Eq. (2.5), the columns of the matrix W are basis vectors spanning a subspace in $k \leq n$ dimensions, called NMF-subspace, while each column of the encoding matrix H represents the new coordinates of the corresponding data sample in the NMF-subspace. From a numerical point of view, each data sample

is approximated by a linear combination of vectors in W, where the linear coefficients are grouped in a column of H corresponding to the data sample. Therefore, the elements h_{ij} codify the amount of the factors (i.e., the columns of W) used to reconstruct each sample of X in the NMF-subspace.

The coefficients h_{ij} in each column of H define the importance of each basis vector in approximating the data sample; if a coefficient is very small, then the corresponding basis vector is useless in approximating the sample. Under some hypotheses,[4] the basis vectors can be interpreted as prototypes of data clusters. In this case, the coefficients h_{ij} can be easily interpreted as membership degrees of each sample to each cluster.

Examples of successful applications of NMF are: basic student skills describing student questionnaire results in educational data mining [27]; topics represented as bag of words in text mining [14, 101]; anatomic parts of images describing human faces in face identification problems [49, 103]; part-based representation of digital characters for object recognition [48, 80]; community categories extracted to describe users networks [110]; diversified portfolio describing trends in stock markets in financial data mining [34, 99]; topics used to clusterize social tags data [18]; users–items relations in recommender systems [47]; chemical constituents in air pollution revelations [55, 66]; musical instrument frequencies for music classification [2–4]; endmembers of constituent materials of hyperspectral images [10, 42, 44, 63, 84, 92]; genes in microarray data [9, 12, 29, 32, 67, 86]. Other successful applications of NMF, where interpretability is a key requirement, belong to molecular pattern discovery [39, 67] and object detection [15].

A key aspect of NMF, that is advantageous for its application in IDA, stands in the possibility of approximating data samples as linear combination of factors, where the factors are subsets of the same features used to represent data samples. Therefore, unlike other low-rank approximation techniques, NMF allows to represent data as composition of parts, being each part expressed with the same features used in data. This makes the results of NMF easily interpretable for the analyst, who can intelligently guide the factorization process, in order to achieve results that are interesting and useful for understanding the problem at hand.

2.3.3 Comparison of NMF and PCA

As stated before, PCA can be used as a tool in IDA because of its dimensionality reduction and visualization capability. However, it presents some drawbacks (such as the presence of mixed sign values) and several research papers demonstrated that it is outperformed by NMF in many applications such as face recognition [25, 48]. In the following, some of the differences among these two techniques are briefly highlighted [116].

[4]The use of NMF in clustering applications will be detailed in Sect. 2.4.2.

Uniqueness. PCA is able to find the global minimum of the optimization problem, while NMF is usually trapped into local minima; this implies that the set of principal components is unique, while NMF has multiple solutions (in terms of basis and encoding matrices).

To overcome NMF nonuniqueness problem, bootstrapping techniques can be used; several executions of the factorization are performed and the most frequent solutions selected.

Ranking. Principal components are naturally ranked accordingly to the quantity of variance they explain. On the contrary, factors in NMF have no ordering and are all equally important. This causes a problem to appropriate choose the value of the rank parameter k. When PCA is applied, no specification of the value k is provided; all the eigenpairs are computed and then the most important components are selected according to the proportion of variance that one wants to preserve. Instead, when NMF is applied, the parameter k has to be specified (by user) as input parameter for the factorization. The choice of the rank value is problem dependent; usually, different factorizations are performed with different rank values and then the results are evaluated accordingly to the target of the analysis.

Orthogonality. Principal components are orthogonal directions which capture the variance in data. On the other hand, factors obtained by NMF are positive vectors that better approximate data, but they are not necessarily orthogonal. They are the bases of the hypercone containing all data and are able to preserve local data structure in this subspace. Figure 2.2 shows the principal components and the factors returned by PCA and NMF (left and right panels, respectively) when applied to nonnegative two-dimensional data matrix.

The orthogonality constraint is a desiderable property; however, this implies the presence of some negative values in the elements of principal components that, as previously highlighted, does not make sense in some contexts. The nonnegativity constraint is always violated by PCA, even when it is applied to nonnegative data. Hence, the interpretability of data is lost when moving from original data space to the reduced low-dimensionality subspace. From Fig. 2.2 (left panel) it can be observed

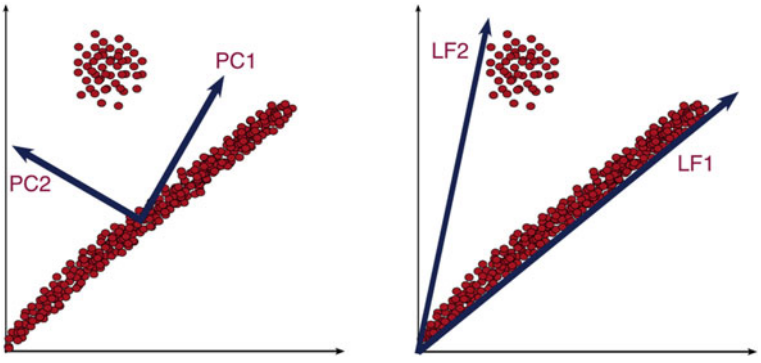

Fig. 2.2 Comparison between principal factors (*left panel*) and NMF latent factors (*right panel*)

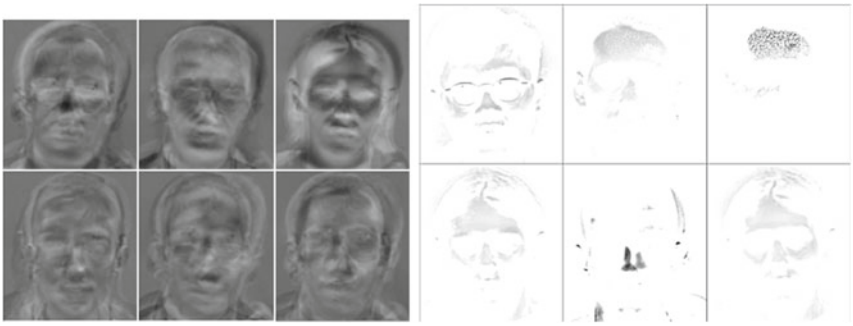

Fig. 2.3 Comparison between bases extracted with PCA (*left panel*) and NMF (*right panel*)

that, starting from samples in the positive orthant, after transforming them by PCA, samples belonging to the line assume negative values. On the contrary (right panel), NMF preserves the nonnegativity of data that leads to a part-based representation.

The interpretability of the factors is one of the strength point in NMF. The parts-based representation obtained by NMF is more intuitive and human-understandable than the holistic results of PCA. A clear example is illustrated in Fig. 2.3 in the context of facial image recognition problem [75]. PCA provides for the eigenfaces that are prototypical faces containing all kinds of facial traits (left panel), while NMF basis vectors represent particular facial traits; different kinds of eyes, noses, and mouths (right panel).

It is worth mentioning that nonnegative variants of PCA and ICA have been developed in literature [90, 95–97, 114] to overcome the difficulties of interpreting nonnegative data derived when standard PCA or ICA are used. However, these constrained variants are based on the same statistical hypotheses on the initial data as their unconstrained versions, and therefore they are applied in more specific contexts than NMF [79, 88, 89].

2.4 Constrained NMF

The key feature of NMF is to decompose the original data as combinations of parts. However, without any constraint the resulting parts could not be as intuitive as to help the analyst in a clear understanding of data. In order to be easy to understand, parts should be composed by a small number of features; however, this structural requirement must be imposed in the factorization process. This can be achieved through different possible variants of NMF which have been proposed in literature.

More specifically, the objective function

$$f(W, H) = \|X - WH\|_F^2, \tag{2.8}$$

that is minimized by the NMF factorization process[5] can be modified in several ways in order to introduce additional properties on the resulting matrices. For example, a penalty term could be added to $f(W, H)$ in order to enforce sparseness [57] as well as to enhance smoothness [35] or to improve clustering ability of NMF [31, 73, 74]. Hence, a more general objective function can be formulated

$$f(W, H) = \|X - WH\|_F^2 + \alpha J_1(W) + \beta J_2(H), \tag{2.9}$$

where the penalty terms $J_1(W)$ and $J_2(H)$ add constraints to the original problem, while the regularization parameters α and β balance the trade-off between the approximation error and additional constraints.

Penalization terms are used in order to constrain the factorization process to yield more interpretable results, so as to be more suitable for IDA. In the following, constrained variants of NMF have been reviewed.

2.4.1 Sparse NMF

Sparseness is a quality that *"refers to a representational scheme where only a few units (out of a large population) are effectively used to represent typical data vectors"* [57]. Sparse representation of hidden factors makes them easier to be interpreted because the resulting parts are structurally simple.

In fact, NMF naturally promotes a sparse representation of data. The matrices W and H describe the relationships among the original features and the latent factors, and among the latent factors and the samples, respectively. Thus, there will be many zero-entries in these matrices where such relationships are not present in data.

When the basis matrix W is sparse, basis vectors representing data subspace are sparse; thus only few features are used to describe the latent factors. This enables a part-based representation where each part is very simple and, therefore, easy to understand by the analyst. Similarly, when the encoding matrix H is sparse, then each sample is described by few (or just one) latent factors. This means that it is possible to easily explain data samples as a composition of few parts.

Sparseness is desirable because it enhances interpretability; however, it could negatively affect the accuracy of the approximation. Thus sparseness should be regulated, but this is not possible in standard NMF unless some additional constraints are added. In [57, 58], the classical NMF optimization algorithm has been modified

[5] In this chapter, we mainly consider NMF based on the error function described in (2.8), but other divergence measures could be used (e.g., generalized Kullback–Leibler divergence, α-divergence). Anyway, technical details apart, the general ideas described in the section still hold.

to include the sparseness constraint. The basic idea is to introduce a measure of sparseness of a k-dimensional vector $\mathbf{x}$ as follows[6]:

$$\text{sparseness}(x) = \frac{\sqrt{k} - (\|\mathbf{x}\|_1)/(\|\mathbf{x}\|_2)}{\sqrt{k} - 1}. \tag{2.10}$$

This measure is then used to design a projected gradient descent algorithm that controls both sparseness and nonnegativity. In essence, this algorithm essentially takes a step in the direction of the negative gradient of the cost function (2.8), and subsequently projects the solution onto the constraint space, that is the cone of nonnegative matrices with a prescribed degree of sparseness ensured by imposing the degree of sparseness to s_W and s_H for the matrices W and H, respectively.

Depending on the specific application of NMF, a desired degree of sparseness for W and H can be imposed. For example, when data samples represent images, high sparseness in both the encoding and the bases matrices is convenient. This allows to generate small *pieces* (factors) of the whole images, and few of them are used to describe each image. Differently, in a medical application where each data sample represents the symptoms of a patient and latent factors are diseases, we should expect to have a sparse encoding matrix H (because we expect patients have one or few more diseases) but W could be dense (since each disease could cause a large number of symptoms).

The prominent role of the data analyst to intelligently guide the factorization process is clear from these simple examples. Based on the questions the analyst wants to ask, and depending on the problem she needs to solve, the NMF process is modified by tuning the sparseness degree of its factors. Many variants of sparse NMF have been proposed subsequently to Hoyer's paper [58]. Some examples are sparse nonnegative matrix factorization, SNMF [39, 67, 81, 92], nonsmooth nonnegative matrix factorization [91], localNMF [37, 72], nonnegative matrix underapproximation (NMU) [41].

2.4.1.1 Nonnegative Matrix Underapproximation—NMU

More recently, Gillis and Glineur proposed the nonnegative matrix underapproximation (NMU) technique, which returns sparser representations than those obtained with classical NMF [41]. NMU is a recursive modification of the NMF algorithm obtained by imposing the upper bound constraint $WH \leq X$ to the factor matrices.

[6]The function in (2.10) yields values in the interval [0,1], where 0 indicates the minimum degree of sparseness obtained when all the elements x_i have the same absolute value, while 1 indicates the maximum degree of sparseness, which is reached when only one component of the vector x is different from zero.

Formally, given a data matrix $X \in \mathbb{R}^{m \times n}$ and a rank $1 \leq k \leq \min(m, n)$, NMU solves the following optimization problem:

$$\begin{aligned} \text{minimize}: \quad & \|X - WH\|_F^2 \\ \text{subject to}: \quad & WH \leq X \\ & W \geq 0, H \geq 0 \end{aligned} \tag{2.11}$$

with $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$.

The basic idea is to recursively identify an optimal rank-one NMF solution $(\mathbf{w}_i, \mathbf{h}_i)$, which is easy to compute, and then apply the same technique to the residual matrix $R_{i+1} = R_i - \mathbf{w}_i \mathbf{h}_i^\top$ (being $R_1 = X$).

More precisely, we suppose that[7] $(W_{:1}, H_{1:}) = (\mathbf{w}_i, \mathbf{h}_i)$ is the solution to the underapproximation problem for X after the first iteration. Since it is an optimal rank-one NMF solution, then $W_{:1} H_{1:} \approx X \, (= R_1)$ and $W_{:1} H_{1:} \leq X$. At the second iterate, the residual matrix is $R_2 = X - W_{:1}, H_{1:}$ which is nonnegative, so it can be underapproximated by $W_{:2} H_{2:} \leq R_2$ leading to $R_3 = R_2 - W_{:2}, H_{2:}$. After k iterates

$$\begin{aligned} X &\geq W_{:1} H_{1:} + W_{:2} H_{2:} + \cdots + W_{:k} H_{k:} \\ &= [W_{:1} W_{:2} \ldots W_{:k}] \, [H_{1:}; H_{2:}; \ldots; H_{k:}] \\ &= WH \end{aligned} \tag{2.12}$$

This method has been tested on several images datasets and demonstrated that it is able to derive sparser solutions than classical NMF, thus improving the part-based representation of factors. In fact, the basis describes approximately disjoint parts of the input data. Also, differently from the classical NMF algorithms, it is possible to choose the factorization rank k during the computation; this enables the data analyst to guide the number of iterations according to the quality of the results. Finally, the optimal solution is, under some assumptions, unique [43]. An extension of NMU, which further emphasizes sparseness, has been proposed to analyze hyperspectral images [44].

2.4.2 Orthogonal NMF and Clustering Capabilities

Dimensionality reduction can be exploited for endowing NMF with clustering capabilities. The theoretical relationship between NMF (with additional orthogonal constraints on its factors), k-means, and spectral-based clustering was demonstrated [31], while the mathematical equivalence between orthogonal NMF and a weighted variant of spherical k-means was proved together with some indications about the cases in which orthogonal NMF should be preferred over k-means and spherical k-means [98].

[7] The symbol $W_{:i}$ denotes the ith column of W. The same applies for H.

Clustering is one of the most useful tools in IDA, since it produces a summarized view of data that helps the analyst to understand data by means of compact and informative representations of large collections of samples [7]. Many different clustering methods exist in literature, like hierarchical clustering, prototype-based clustering, and density-based clustering (just to cite the most important ones). Hierarchical clustering yields a collection of nests groups of data, while in prototype-based clustering groups are represented in a compressed form through a prototype, i.e., an element belonging to the same domain of data. Finally in density-based clustering, groups are formed in regions of data space where data are more crowded. The choice of the most appropriate method is up to the data analyst.

In the case that prototype-based clustering is a convenient method for the problem at hand, NMF could be a valid tool. NMF has been widely used in clustering applications [101, 112] where the factors W and H have been interpreted in terms of cluster centroids and cluster membership, respectively.

From a geometric point of view, columns of W are the axes of the sub-dimensional space where samples are spanned. They represent latent feature extracted from data. Vector samples are clustered according to their closeness to these basis vectors.

NMF without constraints finds a convex hull containing data points. However, [31] pointed out that adding orthogonality constraint to NMF algorithms is necessary to improve their clustering capabilities. In fact, the bases obtained from orthogonal NMF tend to point to the center of the clusters. The minimization problem in (2.8) has been modified imposing orthogonality constraint to the rows of the encoding matrix H as follows:

$$\min_{W\geq 0,\ H\geq 0} \|X - WH\|_F^2, \quad \text{s.t.} \quad HH^\top = I. \tag{2.13}$$

Orthogonality constraint on the matrix H forces samples belonging to the same cluster to be closer to same bases. In the same manner, a feature clustering can be achieved by imposing the orthogonality constraint on the columns of the basis matrix W (i.e., $W^\top W = I$).

As a natural consequence, [31] proposed a new minimization problem. Simultaneously, clustering of both features and objects (i.e., co-clustering) has been archived imposing orthonormality constraints on both columns of W and rows of H.

$$\min_{W\geq 0,\ H\geq 0} \|X - WH\|_F^2, \quad \text{s.t. } W^{\mathrm{T}}W = I,\ HH^{\mathrm{T}} = I. \tag{2.14}$$

In this representation, the matrix W is the clustering indicator matrix, and the rows of the matrix H are the cluster centers for the features clustering problem; while the matrix H is the clustering indicator matrix, and the columns of the matrix W are the cluster centers for the objects clustering problem. However, this double orthogonality constraint is very restrictive and it leads to a rather poor matrix low-rank approximation. Different multiplicative updates for NMF preserving orthogonality were recently proposed [23, 26, 61]. To overcome the limits of the two factor orthogonal NMF, tri-factors NMF–TNMF has been proposed. Particularly, TNMF adds an

extra factor to absorb the different scales of X,W,H and to allow different number of clusters for features and objects, that is

$$X \approx USV, \tag{2.15}$$

being $X \in \mathbb{R}_+^{n \times m}$, $U \in \mathbb{R}_+^{n \times k}$, $S \in \mathbb{R}_+^{k \times l}$, $V \in \mathbb{R}_+^{l \times m}$, where the number of rows in S correspond to the number of feature-clusters k, while the number of columns to the number of objects-clusters l.

The interested reader can find a deep investigation about NMF algorithms with orthogonality constraint and their application in clustering on [68, 73, 74, 87].

2.4.3 Semi-Supervised NMF

NMF is an unsupervised machine learning algorithm, in fact it allows to *automatically* extract human-significative feature from data and to reduce the dimensionality of data. As it has been shown in the previous paragraph, classical NMF algorithms, and constrained ones, are widely used in clustering applications. They group data in a unsupervised way, but without taking in account any prior information of data. However, when class labels are available, this knowledge could be injected in the factorization process, to improve the quality of clustering. Labeling dataset could be difficult, expensive, or time consuming, and often incomplete labels are available. Semi-supervised learning methods use a large amount of unlabeled data, together with labeled data, to train the process [94].

Different algorithms have been also proposed in the context of NMF to inject a priori knowledge. This can be done extending the objective function in (2.8) to include extra terms containing the available a priori knowledge (that could be class labels associated to the samples or pairwise constraints provided by the user, which indicate data to be clustered together—*must link*—and data that have not to be clustered together—*cannot link*). Research on NMF is going in the direction of considering it an interactive tool, instead of a black box. Semi-supervised NMFs allows to modify the factorization process taking in account the knowledge of the analyst. Some examples are [11, 19–22, 51–53, 64, 71, 77, 78, 83, 109, 111, 113].

With the idea of injecting a priori knowledge into NMF process, a new algorithm have been proposed to represent data subspace by user-defined basis as prescribed by IDA [16]. This novel masked nonnegative matrix factorization (MNMF) algorithm could be used either to explain data as a composition of interpretable parts (which are actually hidden in them) and to introduce knowledge in the factorization process as it is briefly described in the following paragraph.

2.4.3.1 Masked NMF

In MNMF the structure of the basis matrix W is defined by a user-provided mask matrix. The analyst specifies the parts she is interested to discover in data and the MNMF technique extracts the subset of data that are actually represented by those parts [13].

From the vector representation of data, it is possible to observe that each sample is represented by a vector $x \in \mathbb{R}_+^n$ of n features $\{f_1, \ldots, f_n\}$. A part p is defined as a sparse vector in $\mathbb{R}^n$ where at least two components are nonzero. A feature belongs to a part iff its value is nonzero. In this way the factorization process is constrained to describe data as a linear correlation of different parts, whose features are linearly correlated among them. The structure of the part (i.e., the features set to zero, thus excluded by the part), as well as the number of parts, constitutes the a priori knowledge and is user-defined.

To obtain basis factors that are able to extract parts, the columns $\mathbf{w}_k$ in W are constrained to contain only few nonzero elements. Factors possessing this type of structure enable the elicitation of local linear relationships in subsets of data and therefore it is very useful for IDA.

A binary matrix $P \in \{0, 1\}^{n \times k}$, with the same dimensions of the basis matrix W is used as mask for the NMF problem. Particularly, the mask matrix P is used to identify the parts that the analyst would like to extract from data. This is accomplished by defining P as a set of k column vectors, where each element in a column is 1 if the corresponding feature has to be selected, 0 if it has not be considered.

To incorporate the additional constraint described above, the NMF minimization problem (2.7) has been extended to automatically impose the structure of the mask P to the basis matrix W:

$$\min_{W \geq 0,\ H \geq 0} \frac{1}{2} \|X - (P \odot W)H\|_F^2 + \frac{1}{2}\lambda \left\| P \odot \tilde{W} \right\|_F^2, \tag{2.16}$$

where $\tilde{w}_{ij} = \exp\left(-w_{ij}\right)$ and $P \in \{0, 1\}^{n \times k}$ and $\lambda \geq 0$ is a regularization parameter.

The objective function in (2.16) is composed of two terms: the first one represents a weighted modification of the classical NMF problem where the mask matrix P is used to fix the structure the basis matrix W has to possess. The second term is a penalty term used to enhance the elements w_{ij} corresponding to elements $p_{ij} = 1$. For this purpose, the exponential function has been chosen; when the value of an entry w_{ij} of W is small, it is increased by the penalty term, when it is high the penalty tends to zero. The choice of the exponential function allow us to prevent that zero values correspond to features that we want to include in the parts. The regularization parameter $\lambda \geq 0$ is used to balance the influence of the two terms.

Two updating formulas for the factors W and H have been derived as a modification of the standard multiplicative update rules of Lee and Seung, taking into account the mask constraint [16].

A query-based MNMF algorithm is used to select the parts in the query that are actually represented by data samples. However, the selected parts are generally

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

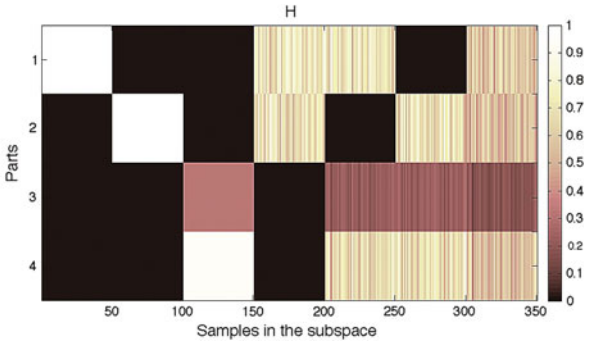

Fig. 2.4 Matrix H obtained with MNMF and P_1

represented by a subset of data only. The algorithm uses the information in the encoding matrix H to understand if samples have been correctly reconstructed by the selected parts. Particularly, the elements of each columns of the encoding matrix H codify the information needed to identify the factors (columns of W) used to reconstruct each sample in the low-rank subspace. They codify the importance of each basis vector in approximating data sample; if a coefficient is very small, then the corresponding basis vector is useless in approximating the sample; as a consequence, data sample does not contain the part represented by this basis vector.

A user-defined threshold is used to quantify the goodness of the reconstruction.

Figure 2.4 shows a simple encoding matrix H obtained with a synthetic dataset X where linear relations among feature are present. On the rows there are the parts and on the columns the samples. Colors in the image highlight the parts that have been used to reconstruct the samples. Different colors indicate how good the reconstruction have been. For instance, samples from 1 to 50 have been perfectly reconstructed using only the first part P_1. This means that, in this subset of data, a local linear relationship between the first and the second features is present. In fact, the first part, corresponding to the first column of P, selects the first and the second features.

An automatic procedure is used to extract the subset of data that are actually represented by the parts, discarding those data in the matrix X that do not find a clear representation by the parts and returning the subset of samples that contains the selected parts.

2.5 An Illustrative Example: NMF for Educational Data Mining

In order to illustrate the use of NMF for intelligently analyzing real-world data, an application of some NMF variants is discussed in the realm of educational data mining.

Educational data mining (EDM) aims at extracting useful knowledge from data coming from e-learning scenarios [100]. The different methods in EDM are designed to collect, store, and analyze data coming from learning and evaluation processes of students, in order to detect conceptual categories that are not directly observable, such as attitudes, interests, values, personality, cognitive abilities, etc.

Roughly speaking, two fundamental theories influence the choice of the appropriate method for analyzing e-learning data: classical test theory (CTT) [106] and item response theory (IRT) [82]. Both are based on the assumption that the answers to specific tests can be considered as manifestations of some skills, which are not immediately observable but can be indirectly derived from the answers.

In this context, NMF can be used to analyze data coming from student questionnaires, in order to identify the latent factors involved in the learning process. Particularly, given a data matrix X (*score matrix*) representing the answers of the students to the questions (*items*) in a test, NMF decomposes it in two factor matrices W and H, such that W encodes the relations between the latent factors (*skills*) and the questions, while H describes the abilities of the students with respect to these latent factors. The latent factors are represented with nonnegative vectors of skill, such that the skills of each student can be defined as a linear combination of these vectors.

The extracted information provides the building blocks of a learning cognitive model, which corresponds to a particular matrix, the so-called question matrix (*Q-matrix*) describing the student necessary skills to adequately answer the questionnaires [105]. Since the skills do not occur explicitly (the actual presence of a particular skill can be only hypothesized on subjective bases), the construction of a Q-matrix is a nontrivial process. To overcome this difficulties, the NMF could be used to automatically extract a Q-matrix (W) [17, 28].

In order to show the application of NMF for extracting a Q-matrix, the *SAT* dataset has been used.[8] The dataset reports the answers of 296 students to 40 questions (items) on four subjects: Mathematics (items 1–10), Biology (item 11–20), History (items 21–30), and French (items 31–40) [27] from published study guides for the SAT Subject Tests.[9]

The left panel of Fig. 2.5 represents the Q-matrix W obtained applying the standard NMF to SAT. Each row represents a skill, while the columns identify the items. The gray shade of each cell indicates the weight of each skill in characterizing the corresponding item; the lightest shades indicate the heaviest weights. As it can be observed, the first and the second skills are mainly composed by contiguous groups of items, while the last two skills are described by scattered items in the dataset.

Since the items correspond to contiguous questions on the same subject (in blocks of ten), it is possible to conclude that the skills number one and two are semantically aligned with the subjects related to the two groups of contiguous items (specifically, Mathematics, and French), while the third and fourth skills are defined by a composition of the remaining subjects (Biology and History). The latter could represent

[8] The dataset is available at http://alumni.cs.ucr.edu/~titus/ (accessed: March 25th 2015).

[9] http://sat.collegeboard.org.

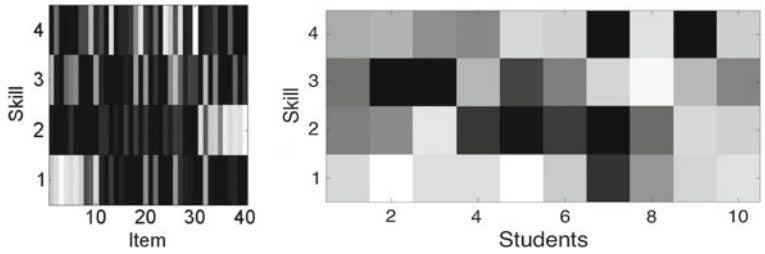

Fig. 2.5 W and H matrices obtained with the standard NMF on the SAT score matrix

"mixed abilities" that cannot be semantically framed in one of the a prior known subjects. These results could suggest a reorganization of the questions to adapt the arguments to the skills, or on the contrary they could suggest more interdisciplinary skills.

The right panel of the Fig. 2.5 illustrates the first ten columns of the matrix H resulting after factorization. This matrix identifies the degrees to which a student has acquired a particular skill. Information in H allows to highlight the skills in which each student is more or less experienced and to group the students according to their abilities, for example, to organize remedial courses.

Constrained versions of NMF could be used to add extra knowledge to the Q-matrix. For instance, by adding the sparsity constraint to NMF, the resulting Q-Matrix is more sparse, i.e., each skill is described by few items. Thus, the sparsity constraint allows to restrict the influence the items have on the skills. The most semantically relevant items can be automatically extracted, thus facilitating the analysis when the number of items is very high. On the other hand, by adding the orthogonal constraint, NMF produces a Q-matrix where skills are well separated, i.e., few items simultaneously belong to different skills. Finally, if the analyst possesses a priori information on the skills required to solve each question, she could impose these relationships by using the masked version of NMF (MNMF).

On the overall, this example gives an idea of the possibilities of NMF as an intelligent analysis tool. The different constraints can be used to reflect the needs of an analyst in the peculiar educational context.

2.6 Conclusions

In this chapter several variants of NMF algorithms have been analyzed in order to highlight their usefulness in the context of Intelligent Data Analysis.

Particularly, it has been pointed out how sparseness and orthogonality constraints have been used to modify classical NMF algorithms in order to overcome their limitations as holistic bases not corresponding with the part-based expected, and not unique decomposition. Moreover, it has been shown how to incorporate a priori knowledge

in the factorization process. Semi-supervised NMF algorithm which inject meta-information about samples by adding extra matrices are also briefly introduced. A novel algorithm used to impose user-defined structure to the reduced space in which data are approximated has been also presented in some details.

The potentialities of NMF as a promising tool for IDA have been stressed: dimensionality reduction, clustering, and part-based representation. However, this chapter could be considered as a starting point in the research panorama in which NMF is no more a black box, but an interactive tool that can be driven by data analyst capabilities.

References

1. R.E. Bellman, *Adaptive Control Processes—A Guided Tour* (Princeton University Press, Princeton, 1961)
2. E. Benetos, M. Kotti, C. Kotropoulos, Applying supervised classifiers based on non-negative matrix factorization to musical instrument classification, in ICME (IEEE, 2006), pp. 2105–2108
3. E. Benetos, M. Kotti, C. Kotropoulos, Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection, in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), vol. V (2006), pp. 221–224
4. E. Benetos et al., Comparison of subspace analysis-based and statistical model-based algorithms for musical instrument classification, in 2nd Workshop on Immersive Communication and Broadcast Systems (ICOB'05), (Berlin, Germany, 2005)
5. M. Berry et al., Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. **52**(1), 155–173 (2007)
6. M. Berthold, D.J. Hand (eds.), *Intelligent Data Analysis: An Introduction*, 1st edn. (Springer, New York, 1999)
7. M.R. Berthold et al., *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, 1st edn. (Springer, Incorporated, London, 2010)
8. R. Bierig et al., Conquering data: the state of play in intelligent data analytics (2015)
9. J.P. Brunet et al., Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl Acad. Sci. 101(12), 4164–4169 (2004). doi:10.1073/pnas.0308531101. ISSN: 1091-6490
10. J.E. Burger, P.L.M. Geladi, Hyperspectral image data conditioning and regression analysis, *Techniques and Applications of Hyperspectral Image Analysis* (Wiley, Chichester, 2007)
11. D. Cai et al., Locality preserving nonnegative matrix factorization, in Proceedings of 2009 International Joint Conference on Artificial Intelligence (IJCAI'09) (2009)
12. P. Carmona-Saez et al., Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinform. **7**, 78 (2006)
13. G. Casalino, Non-negative factorization methods for extracting semantically relevant features in Intelligent Data Analysis. Ph.D. thesis, Dipartimento di Informatica, Università degli Studi di Bari (2015)
14. G. Casalino, N. Del Buono, C. Mencar, Subtractive clustering for seeding non-negative matrix factorizations. Inf. Sci. **257**, 369–387 (2014). doi:10.1016/j.ins.2013.05.038. ISSN: 0020-0255
15. G. Casalino, N. Del Buono, M. Minervini, Nonnegative matrix factorizations performing object detection and localization. Appl. Comput. Intell. Soft Comput. 2012, 15:15–15:15 (2012). doi:10.1155/2012/781987. ISSN: 1687-9724

16. G. Casalino, N. Del Buono, C. Mencar, Part-based data analysis with masked non-negative matrix factorization, in Computational Science and Its Applications–ICCSA 2014–14th International Conference, Guimarães, Portugal, 30 June-3 July 2014, Proceedings, Part VI, ed. by B. Murgante, S. Misra, A. Maria, A.C. Rocha, C. Maria Torre, J. Gustavo Rocha, M. Irene Falcão, D. Taniar, B.O. Apduhan, O. Gervasi. Lecture Notes in Computer Science, vol. 8584 (Springer, 2014), pp. 440–454. doi:10.1007/978-3-319-09153-2_33
17. G. Casalino et al., Fattorizzazioni matriciali non negative per l'analisi dei dati nell'educational data mining, in DIDAMATICA2012 (2012)
18. J. Chen, S. Feng, J. Liu, Topic sense induction from social tags based on non-negative matrix factorization. Inf. Sci. **280**, 16–25 (2014). doi:10.1016/j.ins.2014.04.048. ISSN: 0020-0255
19. Y. Chen et al., Non-negative matrix factorization for semi-supervised data clustering. Knowl. Inf. Syst. **17**(3), 355–379 (2008). doi:10.1007/s10115-008-0134-6. ISSN: 0219-1377
20. Y. Chen et al., Non-negative matrix factorization for semisupervised heterogeneous data coclustering. IEEE Trans. Knowl. Data Eng. 22(10), 1459–1474 (2010). doi:10.1109/TKDE.2009.169. ISSN: 1041-4347
21. Y. Chen et al., Incorporating user provided constraints into document clustering, in Seventh IEEE International Conference on Data Mining, ICDM 2007 (2007), pp. 103–112. doi:10.1109/ICDM.2007.67
22. Y. Cho, L.K. Saul, Nonnegative matrix factorization for semi-supervised dimensionality reduction, in CoRR (2011). abs/1112.3714
23. S. Choi, Algorithms for orthogonal nonnegative matrix factorization, in IEEE International Joint Conference on Neural Networks, 2008. IJCNN. IEEE World Congress on Computational Intelligence (IEEE, 2008), pp. 1828–1832
24. M. Chu et al., Optimality, computation, and interpretation of nonnegative matrix factorizations. SIAM J. Matrix Anal. 4–8030 (2004)
25. A. Cichocki et al., *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation* (Wiley, Chichester, 2009). ISBN 0470746661, 9780470746660
26. N. Del Buono, A penalty function for computing orthogonal non-negative matrix factorizations, in ISDA (IEEE Computer Society, 2009), pp. 1001–1005. ISBN: 978-0-7695-3872-3
27. M. Desmarais, Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. best paper award, in EDM, ed. by M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, J.C. Stamper (2011), pp. 41–50. ISBN: 978-90-386-2537-9
28. M.C. Desmarais, B. Beheshti, R. Naceur, Item to skills mapping: deriving a conjunctive Q-matrix from data, in Intelligent Tutoring Systems (2012), pp. 454–463
29. K. Devarajan, Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput. Biol. **4**(7), e1000029 (2008)
30. I.S. Dhillon, S. Sra, Generalized nonnegative matrix approximations with Bregman divergences, in Proceeding of Neural Information Processing Systems (Curran Associates Inc., 2005), pp. 283–290
31. C. Ding, X. He, H.D. Simon, On the equivalence of nonnegative matrix factorization and k-means–spectral clustering, in Proceedings of the SIAM Data Mining Conference (SIAM, 2005), pp. 606–610
32. C. Ding et al., Orthogonal nonnegative matrix tri-factorizations for clustering, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, 2006), pp. 126–135
33. D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts? in Advances in Neural Information Processing Systems, vol. 16, ed. by S. Thrun, L. Saul, B. Schölkopf (MIT Press, Cambridge, 2004)
34. K. Drakakis et al., Analysis of financial data using non-negative matrix factorization. Int. Math. Forum **3**(38), 1853–1870 (2008)
35. S. Essid, C. Févotte, Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. IEEE Trans. Multimed. **15**(2), 415–425 (2013). doi:10.1109/TMM.2012.2228474

36. U.M. Fayyad et al. (eds.), Advances in Knowledge Discovery and Data Mining (American Association for Artificial Intelligence, 1996). Chap. From data mining to knowledge discovery: an overview, pp. 1–34. ISBN: 0-262-56097-6
37. T. Feng et al., Local non-negative matrix factorization as a visual representation, in Proceedings of the 2nd International Conference on Development and Learning, ICDL'02 (IEEE Computer Society, 2002), p. 178
38. R.A. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(7), 179–188 (1936)
39. Y. Gao, G. Church, Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics **21**(21), 3970–3975 (2005)
40. N. Gillis, The why and how of nonnegative matrix factorization, in Regularization, Optimization, Kernels, and Support Vector Machines, ed. by M. Signoretto, J.A.K. Suykens, A. Argyriou. Machine Learning and Pattern Recognition Series (Chapman and Hall/CRC, Boca Raton, 2014)
41. N. Gillis, F. Glineur, Using underapproximations for sparse nonnegative matrix factorization. Pattern Recognit **43**(4), 1676–1687 (2010). doi:10.1016/j.patcog.2009.11.013. ISSN: 0031-3203
42. N. Gillis, D. Kuang, H. Park, Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization, in CoRR (2013). abs/1310.7441
43. N. Gillis, R.J. Plemmons, Dimensionality reduction, classification, and spectral mixture analysis using non-negative underapproximation. Opt. Eng. **50**(2), 027001 (2011). doi:10.1117/1.3533025
44. N. Gillis, R.J. Plemmons, Sparse nonnegative matrix underapproximation and its application to hyperspectral image analysis. Linear Algebra Appl. **438**(10), 3991–4007 (2013). doi:10.1016/j.laa.2012.04.033. ISSN: 0024-3795
45. G.H. Golub, C.F. Van Loan, *Matrix Computations*, 3rd edn. (The Johns Hopkins University Press, Baltimore, 2001)
46. G.H. Golub, A. Hoffman, G.W. Stewart, A generalization of the Eckart-Young-Mirsky matrix approximation theorem. Linear Algebra Appl. **88——89**(0), 317–327 (1987). doi:10.1016/0024-3795(87)90114-5. ISSN: 0024-3795
47. Q. Gu, J. Zhou, C.H.Q. Ding, Collaborative filtering: weighted nonnegative matrix factorization incorporating user and item graphs, in SDM (SIAM, 2010), pp. 199–210
48. D. Guillamet, J. Vitriá, Evaluation of distance metrics for recognition based on non-negative matrix factorization. Pattern Recognit. Lett. **24**(9–10), 1599–1605 (2003). doi:10.1016/S0167-8655(02)00399-9. ISSN: 0167-8655
49. D. Guillamet, J. Vitriá, Non-negative matrix factorization for face recognition, in CCIA'02: Proceedings of the 5th Catalonian Conference on AI (Springer, New York, 2002), pp. 336–344
50. D.J. Hand, Intelligent data analysis: issues and opportunities, in IDA, ed. by X. Liu, P.R. Cohen, M.R. Berthold. Lecture Notes in Computer Science, vol. 1280 (Springer, New York, 1997), pp. 1–14
51. Y. He, H. Lu, S. Xie, Semi-supervised non-negative matrix factorization for image clustering with graph Laplacian. Multimed. Tools Appl. **72**(2), 1441–1463 (2014). doi:10.1007/s11042-013-1465-1. ISSN: 1380-7501
52. Y. He et al., Non-negative matrix factorization with pairwise constraints and graph Laplacian. Neural Process. Lett. pp. 1–19 (2014). doi:10.1007/s11063-014-9350-0. ISSN: 1370-4621
53. M. Heiler, C. Schnörr, Learning sparse representations by non-negative matrix factorization and sequential cone programming. J. Mach. Learn. Res. **7**, 1385–1407 (2006). ISSN: 1532-4435
54. J.H. Holmes, N. Peek, Intelligent data analysis in biomedicine. J. Biomed. Inform. **40**(6), 605–608 (2007)
55. P.K. Hopke, *Receptor Modeling in Environmental Chemistry* (Wiley, New York, 1985)
56. H. Hotelling, Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417–441 (1933)

57. P.O. Hoyer, Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004). ISSN: 1532-4435
58. P.O. Hoyer, Non-negative sparse coding, in Neural Networks for Signal processing XII (Proceedings of IEEE Workshop on Neural Networks for Signal Processing) (2002), pp. 557–565
59. K. Huang, N.D. Sidiropoulos, A. Swami, Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition. IEEE Trans. Signal Process. (TSP) **62**(1), 211–224 (2014)
60. A. Hyvärinen, Survey on Independent component analysis. Neural Comput. Surv. **2**, 94–128 (1999)
61. J. Yoo, S. Choi, Orthogonal nonnegative matrix tri-factorization for co-clustering: multiplicative updates on Stiefel manifolds. Inf. Process. Manag. **46**, 559–570 (2010)
62. J.E. Jackson, A User's Guide to Principal Components. Wiley Series in Probability and Statistics (Wiley-Interscience, Hoboken, 2003). ISBN: 0471471348
63. S. Jia, Y. Qian. A complexity constrained nonnegative matrix factorization for hyperspectral unmixing, in ICA, ed. by M.E. Davies et al. Lecture Notes in Computer Science, vol. 4666 (Springer, New York, 2007), pp. 268–276. ISBN: 978-3-540-74493-1
64. L. Jing et al., Semi-supervised clustering via constrained symmetric non-negative matrix factorization, in Brain Informatics, ed. by F. Zanzotto, et al. Lecture Notes in Computer Science, vol. 7670 (Springer, Berlin, 2012), pp. 309–319. ISBN: 978-3-642-35138-9. doi:10.1007/978-3-642-35139-6_29
65. I.T. Jolliffe, *Principal Component Analysis* (Springer, New York, 1986)
66. E. Kim, P.K. Hopke, E.S. Edgerton, Source identification of Atlanta aerosol by positive matrix facorization. J. Air Waste Manag. Assoc. 733–739 (2003)
67. H. Kim, H. Park, Sparse Non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007). doi:10.1093/bioinformatics/btm134. ISSN: 1367-4803
68. C. Lazar, A. Doncescu, Non negative matrix factorization clustering capabilities; application on multivariate image segmentation, in CISIS, ed. by L. Barolli, F. Xhafa, H.-H. Hsu (IEEE Computer Society, 2010), pp. 924–929. ISBN: 978-0-7695- 3575-3
69. D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing Systems*, vol. 13, ed. by T.K. Leen, T.G. Dietterich, V. Tresp (MIT Press, Cambridge, 2001), pp. 556–562
70. D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999). doi:10.1038/44565. ISSN: 0028-0836
71. H. Lee, J. Yoo, S. Choi, Semi-supervised nonnegative matrix factorization. IEEE Signal Process. Lett. **17**(1), 4–7 (2010). doi:10.1109/LSP.2009.2027163. ISSN: 1070-9908
72. S.Z. Li et al., Learning spatially localized, parts-based representation. Comput. Vis. Pattern Recognit. **1**, 207–212 (2001). doi:10.1109/CVPR.2001.990477
73. T. Li, C. Ding, The relationships among various nonnegative matrix factorization methods for clustering, in Proceedings of the Sixth International Conference on Data Mining, ICDM'06 (IEEE Computer Society, Washington, 2006), pp. 362–371. ISBN: 0-7695-2701-9
74. T. Li, C.H.Q. Ding, Nonnegative matrix factorizations for clustering: a survey, *Data Clustering: Algorithms and Applications* (CRC Press, Boca Raton, 2013)
75. Z. Lihong, G. Zhuang, X. Xu, Facial expression recognition based on PCA and NMF, in 7th World Congress on Intelligent Control and Automation, WCICA 2008 (2008), pp. 6826–6829. doi:10.1109/WCICA.2008.4593968
76. H. Liu, H. Motoda, Computational Methods of Feature Selection. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. (Chapman & Hall/CRC, Boca Raton, 2007). ISBN: 1584888784
77. H. Liu, Z. Wu, Non-negative matrix factorization with constraints, in AAAI, ed. by M. Fox, D. Poole (AAAI Press, 2010)
78. H. Liu et al., Constrained nonnegative matrix factorization for image representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1299–1311 (2012)

79. P. Liu et al., The application of principal component analysis and nonnegative matrix factorization to analyze time-resolved optical waveguide absorption spectroscopy data. Anal. Methods **5**(17), 4454–4459 (2013)
80. W. Liu, N. Zheng, Non-negative matrix factorization based methods for object recognition. Pattern Recognit. Lett. **25**(8), 893–897 (2004). doi:10.1016/j.patrec.2004.02.002. ISSN: 0167-8655
81. W. Liu, N. Zheng, X. Lu, Non-negative matrix factorization for visual coding, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), vol. 3 (2003), pp. 293–296
82. F.M. Lord, A Theory of Test Scores (1952)
83. N. Lyubimov, M. Kotov, Non-negative matrix factorization with linear constraints for single-channel speech enhancement, in INTERSPEECH, ed. by F. Bimbot et al. (ISCA, 2013), pp. 446–450
84. W.K. Ma et al., A signal processing perspective on hyperspectral unmixing: insights from remote sensing. IEEE Signal Process. Mag. **31**(1), 67–81 (2014). doi:10.1109/MSP.2013.2279731
85. M.W. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. **106**(3), 697–702 (2009). doi:10.1073/pnas.0803205106
86. E. Mejía-Roa et al., BioNMF: a web-based tool for nonnegative matrix factorization in biology. Nucleic Acids Res. **36**, 523–528 (2008)
87. A. Mirzal, Clustering and latent semantic indexing aspects of the nonnegative matrix factorization, arXiv preprint arXiv:1112.4020 (2011), pp. 1–28
88. A. Montanari, E. Richard. Non-negative principal component analysis: message passing algorithms and sharp asymptotics, in CoRR (2014). abs/1406.4775
89. B. Ng, R. Abugharbieh, M.J. McKeown, Functional segmentation of fMRI data using adaptive non-negative sparse PCA (ANSPCA), in Medical Image Computing and Computer-Assisted Intervention (MICCAI) (Springer, London, 2009), pp. 490–497
90. E. Oja, M. Plumbley, Blind separation of positive sources using non-negative PCA, in Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003) (2003), pp. 11–16
91. A. Pascual-Montano et al., Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 403–415 (2006). doi:10.1109/TPAMI.2006.60. ISSN: 0162-8828
92. V.P. Pauca, J. Piper, R.J. Plemmons, Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl. **416**(1), 29–47 (2006). doi:10.1016/j.laa.2005.06.025. ISSN: 0024-3795
93. K. Pearson, On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**(6), 559–572 (1901)
94. N.N. Pise, P. Kulkarni, A survey of semi-supervised learning methods, in Computational Intelligence and Security (IEEE Computer Society, 2008), pp. 30–34. ISBN: 978-0-7695-3508-1
95. M. Plumbley, Algorithms for non-negative independent component analysis. IEEE Trans. Neural Netw. **14**(3), 534–543 (2003)
96. M. Plumbley, Conditions for non-negative independent component analysis. IEEE Signal Process. Lett. **9**(6), 177–180 (2002)
97. M.D. Plumbley, E. Oja, A nonnegative PCA algorithm for independent component analysis. IEEE Trans. Neural Netw. **15**(1), 66–76 (2004)
98. F. Pompili et al., Two algorithms for orthogonal nonnegative matrix factorization with application to clustering, in CoRR (2012). abs/1201.0901
99. B. Ribeiro et al., Extracting discriminative features using nonnegative matrix factorization in financial distress data, in ICANNGA, ed. by M. Kolehmainen, P.J. Toivanen, B. Beliczynski. Lecture Notes in Computer Science, vol. 5495 (Springer, New York, 2009), pp. 537–547. ISBN: 978-3-642-04920-0

100. C. Romero, S. Ventura, Educational data mining: a review of the state of the art. Trans. Syst. Man Cybern. Part C **40**(6), 601–618 (2010). ISSN: 1094-6977
101. F. Shahnaz et al., Document clustering using nonnegative matrix factorization. Inf. Process. Manag. **42**(2), 373–386 (2006). doi:10.1016/j.ipm.2004.11.005. ISSN: 0306-4573
102. C. Spearman, General intelligence, objectively determined and measured. Am. J. Psychol. **15**, 201–293 (1904)
103. X. Sun, Q. Zhang, Z. Wang, Face recognition based on NMF and SVM. Electron. Commer. Secur. Int. Symp. **1**, 616–619 (2009). doi:10.1109/ISECS.2009.98
104. R. Tandon, S. Sra, Sparse nonnegative matrix approximation: new formulations and algorithms. Technical report, MPI Technical report (2010)
105. K.K. Tatsuoka, Rule space: an approach for dealing with misconceptions based on item response theory. J. Educ. Meas. (1983)
106. Theory of Mental Tests. Wiley Publications in Psychology (Wiley, New York, 1950)
107. M. Turk, A. Pentland, Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991). doi:10.1162/jocn.1991.3.1.71. ISSN: 0898-929X
108. L.J.P. Van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality reduction: a comparative review (2008)
109. C. Wang et al., Non-negative semi-supervised learning, in AISTATS, JMLR Proceedings, vol. 5, ed. by D.A. Van Dyk, M. Welling, JMLR (2009), pp. 575–582
110. F. Wang et al., Community discovery using nonnegative matrix factorization. Data Min. Knowl. Discov. **22**(3), 493–521 (2011). doi:10.1007/s10618-010-0181-y. ISSN: 1384-5810
111. Y. Wang et al, Fisher non-negative matrix factorization for learning local features, in Asian Conference on Computer Vision (2004)
112. W. Xu, X. Liu, Y. Gong, Document clustering based on nonnegative matrix factorization, in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'03 (ACM, New York, 2003), pp. 267–273. ISBN: 1-58113-646-3
113. Y. Yang, B.-G. Hu, Pairwise constraints-guided non-negative matrix factorization for document clustering, in IEEE/WIC/ACM International Conference on Web Intelligence (2007), pp. 250–256. doi:10.1109/WI.2007.66
114. R. Zass, A. Shashua, Nonnegative sparse PCA, in Neural Information Processing Systems (2007)
115. Z. Zhang, Nonnegative matrix factorization: models, algorithms and applications, in DATA MINING: Foundations and Intelligent Paradigms, vol. 2, ed. by D.E. Holmes, L.C. Jain (Springer, Berlin, 2011), pp. 99–134
116. A. Zinovyev et al. Blind source separation methods for deconvolution of complex signals in cancer biology, in CoRR (2013). abs/1301.2634

http://www.springer.com/978-3-662-48330-5

Non-negative Matrix Factorization Techniques
Advances in Theory and Applications
Naik, G.R. (Ed.)
2016, VII, 194 p., Hardcover
ISBN: 978-3-662-48330-5