

# NONNEGATIVE MATRIX FACTORIZATION

SOFÍA LLAVAYOL

ABSTRACT. This document presents the final project for the course *Numerical Linear Algebra for Statistical Learning* at Universidad de la República, Uruguay. It outlines the fundamental concepts of Nonnegative Matrix Factorization based on the reference [2], and includes selected experiments implemented in Python to illustrate key ideas.

## CONTENTS

1. Introduction	1
1.1. LDR techniques for Data Analysis	1
1.2. NMF, an LDR technique for nonnegative data	2
2. Application on Facial Feature Extraction	2
3. Algorithm with multiplicative updates	3
3.1. KKT conditions	3
3.2. Multiplicative updates	4
Appendix A. Constrained optimization methods	5
A.1. Lagrange multipliers	5
A.2. Karush-Kuhn-Tucker (KKT) conditions	6
References	6

## 1. INTRODUCTION

*Nonnegative matrix factorization (NMF)* is an easily interpretable *linear dimensionality reduction (LDR)* technique for nonnegative data. We first introduce the general concept of LDR, followed by a more detailed discussion of NMF.

**1.1. LDR techniques for Data Analysis.** Extracting the underlying structure within data sets is one of the central problems in data science, and numerous techniques exist to perform this task. One of the oldest approaches is LDR. The idea of LDR is to represent each data point as a linear combination of a small number of basis elements.

Mathematically, given a dataset of  $n$  data points  $x_1, \dots, x_n \in \mathbb{R}^m$ , LDR looks for  $r \ll \min\{m, n\}$  basis vectors  $w_1, \dots, w_r \in \mathbb{R}^m$  such that each data point  $x_j$  is well-approximated by a linear combination of these basis vectors:

$$x_j \approx w_1 \cdot h_{1j} + \dots + w_r \cdot h_{rj} = [w_1 \cdots w_r] \begin{bmatrix} h_{1j} \\ \vdots \\ h_{rj} \end{bmatrix} = Wh_j,$$

for some  $h_j = [h_{1j}, \dots, h_{rj}]^T \in \mathbb{R}^r$ .

Note that this is equivalent to *low-rank matrix approximation (LRMA)* –that is, expressing  $X \approx WH$  where

- each column of  $X \in \mathbb{R}^{m \times n}$  is a data point,  $X(:, j) = x_j$ ;
- each column of  $W \in \mathbb{R}^{m \times r}$  is a basis element,  $W(:, j) = w_j$ ;
- each column of  $H \in \mathbb{R}^{r \times n}$  contains the coordinates of a data point  $x_j$  in the basis  $W$ ,  $H(:, j) = h_j$ .

Hence LDR provides a rank- $r$  approximation  $X \approx WH$ , which can be written as:

$$[x_1 \cdots x_n] \approx [w_1 \cdots w_r] [h_1 \cdots h_n].$$

In order to compute  $W$  and  $H$  given  $X$  and  $r$ , one needs to define an error measure. For example, when  $(W, H)$  minimizes the Frobenius (or Euclidean) norm

$$\|X - WH\|_F^2 = \sum_{i,j} (X - WH)_{ij}^2,$$

then LRMA is equivalent to *principal component analysis (PCA)*, which can be computed via the *singular value decomposition (SVD)*. In PCA, the optimal rank- $r$  approximation of  $X$  that minimizes the Frobenius norm is given by the truncated SVD of  $X$ , obtained by keeping the top  $r$  singular values and the corresponding singular vectors.

LRMA models are used to compress the data, filter the noise, reduce the computational effort for further manipulation of the data, or to directly identify hidden structure in a data set. Many variants of LRMA have been developed, and they differ in two key aspects: (1) the error measure can vary and should be chosen depending on the noise statistic assumed on the data, (2) different constraints can be imposed on the factors  $W$  and  $H$ .

**1.2. NMF, an LDR technique for nonnegative data.** Among LRMA models, nonnegative matrix factorization requires the factor matrices  $W$  and  $H$  to be componentwise nonnegative, which we denote  $W \geq 0$  and  $H \geq 0$ . In Section 2, we discuss an application where these nonnegativity constraints are natural and meaningful.

Formally, the NMF problem is defined as follows: find matrices  $(W, H)$  that minimize

$$(1) \quad \min_{\substack{W \in \mathbb{R}^{m \times r} \\ H \in \mathbb{R}^{r \times n}}} d(X, WH) \quad \text{subject to } W \geq 0 \text{ and } H \geq 0,$$

where  $d(\cdot, \cdot)$  is a measure of approximation error (e.g., the Frobenius norm).

In Section 3, we discuss an algorithm to approximately solve this problem when the distance measure is given by the Frobenius norm. An application of this algorithm to image processing is presented in Section 2.

## 2. APPLICATION ON FACIAL FEATURE EXTRACTION

...



FIGURE 1. Visualization of selected components from the NMF decomposition of facial images. Left: components from a rank-50 factorization. Right: components from a rank-250 factorization. Each component is reshaped into the original image size of (50, 37).

### 3. ALGORITHM WITH MULTIPLICATIVE UPDATES

Let  $X \in \mathbb{R}^{m \times n}$  and  $r \ll \min\{m, n\}$  be given. In this section, we focus on the following constrained optimization problem

$$(2) \quad \min_{\substack{W \in \mathbb{R}^{m \times r} \\ H \in \mathbb{R}^{r \times n}} f(W, H) \quad \text{subject to } W \geq 0 \text{ and } H \geq 0,$$

where  $f(W, H) = \frac{1}{2} \|X - WH\|_F^2$ . Note that this problem is equivalent to (1) for the case where  $d(X, WH) = \|X - WH\|_F$  is the Euclidean distance. We will apply the KKT conditions, as described in Appendix A.

**3.1. KKT conditions.** We will start by applying the conditions in (6) for the variables  $W = (W_{ik})$ . The constraints are given by  $g_{ik}(W) = -W_{ik} \leq 0$ , so the stationarity condition reads

$$0 = \frac{\partial f}{\partial W_{ik}} - \sum \lambda_{rs} \cdot \frac{\partial g_{rs}}{\partial W_{ik}} = \frac{\partial f}{\partial W_{ik}} - \lambda_{ik} \cdot (-1) = (\nabla_W f)_{ik} + \lambda_{ik}$$

for all  $i, k$ . Hence, writing  $\Lambda_W = (\lambda_{ik})$  and computing the gradient  $\nabla_W f$ , we obtain

$$\Lambda_W = -\nabla_W f = -(X - WH)H^T = WHH^T - XH^T.$$

Also, the complementary slackness conditions  $\lambda_{ik} \cdot g_{ik}(W) = 0$  in matrix form reads

$$0 = \Lambda_W \circ W = -\nabla_W f \circ W$$

where  $\circ$  is the component-wise product of two matrices.

```

1 import numpy as np
2
3 def multiplicative_updates(X, r, max_iter=2000, random_state=0):
4     m, n = X.shape
5
6     rng = np.random.default_rng(random_state) # Initialize local random generator
7
8     # Initialize W and H with values in [0, 1)
9     W = rng.random((m, r))
10    H = rng.random((r, n))
11
12    eps = 1e-10 # Small constant to avoid division by zero
13
14    for i in range(max_iter):
15        # Update rules for W and H
16        W *= (X @ H.T) / np.maximum(W @ H @ H.T, eps)
17        H *= (W.T @ X) / np.maximum(W.T @ W @ H, eps)
18
19    return W, H

```

LISTING 1. Algorithm with multiplicative updates.

Similarly, for the variables  $H = (H_{kj})$ , the constraints are  $-H_{kj} \leq 0$ , and applying the same reasoning yields

$$\begin{aligned}\Lambda_H &= -\nabla_H f = -W^T(X - WH) = W^TWH - W^TX, \\ 0 &= \nabla_H f \circ H.\end{aligned}$$

To finish, adding the feasibility conditions  $\Lambda_W \geq 0$  and  $\Lambda_H \geq 0$ , the KKT conditions read

$$(3) \quad \begin{cases} W \geq 0, & \nabla_W f = WHH^T - XH^T \geq 0, & W \circ \nabla_W f = 0, \\ H \geq 0, & \nabla_H f = W^TWH - W^TX \geq 0, & H \circ \nabla_H f = 0. \end{cases}$$

These conditions characterize first-order optimality for the constrained problem (2), and are satisfied at any local minimum.

**3.2. Multiplicative updates.** Given  $X$ ,  $W$  and  $H$ , the goal is to iteratively update  $W$  and  $H$  according to the rules in (4). Listing 1 shows the multiplicative update algorithm used in this work, implemented in Python.

The multiplicative updates (MU) modify  $W$  and  $H$  as follows:

$$(4) \quad W \leftarrow W \circ \frac{[XH^T]}{[WHH^T]} \quad \text{and} \quad H \leftarrow H \circ \frac{[W^TX]}{[W^TWH]},$$

where  $\frac{[ ]}{[ ]}$  denotes the componentwise division between two matrices.

It is straightforward to verify that if a point  $(W, H)$  satisfies the first-order optimality conditions (3), then the update rules (4) do not alter the values of  $W$  and  $H$ .

In Theorem 1 of [3], the authors prove the following:

**Theorem 1.** *The Euclidean distance  $\|X - WH\|_F$  is nonincreasing under the update rules (4).*

Hence, the MU lead to an algorithm for which  $f$  does not increase.

An intuitive interpretation of the update rules in (4) is as follow. Observe that from the gradient expression  $\nabla_W f = WHH^T - XH^T$ , we have that

$$\frac{(XH^T)_{ik}}{(WHH^T)_{ik}} \geq 1 \quad \Leftrightarrow \quad (\nabla_W f)_{ik} \leq 0.$$

Therefore, in order to look for matrices  $W$  and  $H$  that satisfy (3), for each entry of  $W$ , the multiplicative update behaves as follows: (i) increase the entry if its partial derivative is negative, (ii) decrease the entry if its partial derivative is positive, or (iii) leave the entry unchanged if its partial derivative is equal to zero. The same logic applies to the updates for  $H$ .

However, if an entry of  $W$  is zero, the MU cannot change it, regardless of the sign of the corresponding partial derivative. As a result, it is possible for an entry of  $W$  to be zero while its partial derivative is negative, which violates the optimality conditions in (3). Therefore, the MU are not guaranteed to converge to a point that satisfies the first-order optimality conditions.

## APPENDIX A. CONSTRAINED OPTIMIZATION METHODS

In many practical optimization problems, the solution is required to satisfy certain constraints. This section introduces two fundamental approaches for handling constraints: the method of Lagrange multipliers for equality constraints, and the *Karush-Kuhn-Tucker (KKT)* conditions for inequality constraints.

**A.1. Lagrange multipliers.** Consider the following optimization problem with one equality constraint

$$(5) \quad \min_{x,y} f(x,y) \quad \text{subject to } g(x,y) = 0.$$

We assume that  $f$  and  $g$  have continuous first partial derivatives.

Suppose that the point  $(x_0, y_0)$  satisfies the constraint  $g(x_0, y_0) = 0$  and that the gradient  $\nabla g(x_0, y_0) \neq 0$ . Recall that the gradient  $\nabla g(x_0, y_0)$  is orthogonal to the level set defined by  $g(x, y) = 0$ . Therefore, if  $f(x_0, y_0)$  is a minimum of the constrained problem (5), then the gradient  $\nabla f(x_0, y_0)$  must be parallel to  $\nabla g(x_0, y_0)$ . Otherwise, one could move along the constraint set  $g(x, y) = 0$  in a direction that decreases  $f$ , contradicting the minimality of  $f(x_0, y_0)$ .

In summary, if  $f(x_0, y_0)$  is a minimum of the constrained problem (5) and  $\nabla g(x_0, y_0) \neq 0$ , then there exists  $\lambda_0 \in \mathbb{R}$  such that

$$\nabla f(x_0, y_0) = \lambda_0 \cdot \nabla g(x_0, y_0).$$

Defining the *Lagrange function* as

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y).$$

Then, the gradient of  $\mathcal{L}$  is given by

$$\nabla \mathcal{L}(x, y, \lambda) = \left( \nabla f(x, y) - \lambda \cdot \nabla g(x, y), \quad -g(x, y) \right).$$

Thus, the condition  $\nabla \mathcal{L}(x_0, y_0, \lambda_0) = 0$  encodes the necessary conditions for  $(x_0, y_0)$  to be a solution of the constrained optimization problem (5), as discussed above.

To solve the original constrained optimization problem (5), we look for points  $(x, y, \lambda)$  such that  $\nabla \mathcal{L}(x, y, \lambda) = 0$ , that is to say

$$\begin{cases} \nabla f(x, y) - \lambda \cdot \nabla g(x, y) = 0 \\ g(x, y) = 0 \end{cases}$$

In other words, we reduce the problem to solving a system of equations given by the vanishing of the gradient of the Lagrange function. Any solution  $(x_0, y_0, \lambda_0)$  of this system provides a candidate for a constrained extremum of  $f$  subject to  $g(x, y) = 0$ .

**A.2. Karush-Kuhn-Tucker (KKT) conditions.** The method of Lagrange multipliers extends naturally to problems with multiple constraints—both equality and inequality. Suppose we aim to minimize  $f(x, y)$  subject to  $g_i(x, y) \leq 0$  and  $h_j(x, y) = 0$  for  $i = 1, \dots, M$  and  $j = 1, \dots, N$ . The KKT conditions generalize the necessary optimality conditions and state that, under regularity assumptions, a candidate solution  $(x, y)$  must satisfy:

$$(6) \quad \begin{cases} \nabla f(x, y) - \sum_{i=1}^M \lambda_i \nabla g_i(x, y) - \sum_{j=1}^N \mu_j \nabla h_j(x, y) = 0 & \text{(stationarity),} \\ g_i(x, y) \leq 0, \quad h_j(x, y) = 0 & \text{(primal feasibility),} \\ \lambda_i \geq 0 & \text{(dual feasibility),} \\ \lambda_i \cdot g_i(x, y) = 0 & \text{(complementary slackness).} \end{cases}$$

These conditions form a system whose solutions are candidates for local optima of the constrained optimization problem. For further details, see [1].

#### REFERENCES

- [1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [2] Gillis, N. (2021). Nonnegative matrix factorization. Society for Industrial and Applied Mathematics. <https://iccn.loc.gov/2020042037>
- [3] Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13, 556–562.