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 [1], and includes selected experiments implemented in Python to illustrate key ideas.

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:

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

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

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 WH of X, which can be written as:

$$\begin{bmatrix} x_1 \cdots x_n \end{bmatrix} \approx \begin{bmatrix} w_1 \cdots w_r \end{bmatrix} \begin{bmatrix} h_1 \cdots h_n \end{bmatrix}.$$

Date: June 4, 2025.

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 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)*.

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.

NMF Problem. Given a nonnegative matrix $X \in \mathbb{R}^{m \times n}$, a factorization rank r, and a distance measure between matrices $d(\cdot, \cdot)$, solve

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

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

2. Application on Facial Feature Extraction

...

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 optimization problem

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

where $f(W, H) = \frac{1}{2}\|X - WH\|_F^2$. This problem is equivalent to (1) for the case where $d(X, WH) = \|X - WH\|_F$.

Appendix A. Lagrange multipliers

One equality constrain. Consider the following optimization problem with one equality constrain

$$\min_{x,y} f(x,y) \quad \text{subject to } g(x,y) = 0. \tag{3}$$

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 (3), 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 (3) 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

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

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 (3), as discussed above.

To solve the original constrained optimization problem (3), we look for points (x, y, λ) 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, λ_0) of this system provides a candidate for a constrained extremum of f subject to $g(x, y) = 0$.

Multiple equality constraints. The method described above naturally extends to optimization problems with multiple equality constraints. Suppose we have M constraints $g_i(x, y) = 0$, for $i = 1, \ldots, M$. We define the Lagrange function as

$$\mathcal{L}(x, y, \lambda_1, \ldots, \lambda_M) = f(x, y) - \sum_{i=1}^{M} \lambda_i \cdot g_i(x, y).$$

To find candidate solutions, we again look for points such that $\nabla \mathcal{L}(x, y, \lambda_1, \ldots, \lambda_M) = 0$.

One inequality constrain. If we now modify the constraint in (3) to an inequality constraint, namely $g(x, y) \leq 0$, a similar principle applies. However, since the feasible set may include boundary and interior points, we must refine the conditions under which a point can be optimal. We now look for points (x, y, λ) such that

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

These are known as the *Karush-Kuhn-Tucker (KKT)* conditions for a problem with a single inequality constraint. The condition $\lambda \cdot g(x, y) = 0$ ensures that either $g(x, y) = 0$ and λ can be positive, or $g(x, y) < 0$, in which case the corresponding multiplier must be zero. In the latter case, the condition $\nabla f(x, y) - \lambda \cdot \nabla g(x, y) = 0$ reduces to $\nabla f(x, y) = 0$, indicating that the point is a stationary point of the objective function in the interior of the feasible region.

Multiple inequality constraints. To finish, we extend to problems involving multiple inequality constraints. Suppose we want to minimize $f(x, y)$ subject to $g_i(x, y) \leq 0$, for $i = 1, \ldots, M$. We then look for points $(x, y, \lambda_1, \ldots, \lambda_M)$ such that

$$\begin{cases} \nabla f(x, y) - \sum_{i=1}^{M} \lambda_i \cdot \nabla g_i(x, y) = 0, \\ g_i(x, y) \leq 0 \quad \text{for all } i, \\ \lambda_i \geq 0 \quad \text{for all } i, \\ \lambda_i \cdot g(x, y) = 0 \quad \text{for all } i. \end{cases}$$

These conditions provide a system of equations and inequalities whose solutions are candidates for constrained local minima or maxima of f.

References

[1] Gillis, N. (2021). Nonnegative matrix factorization. Society for Industrial and Applied Mathematics. https://lccn.loc.gov/2020042037