MF TUTORIAL PART 2

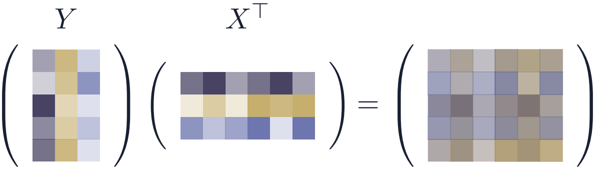
Nonnegative Matrix Factorization

Sibylle Hess and Michiel Hochstenbach

We have seen that we cannot get a better low rank approximation than with **SVD**.

What if we want to impose **CONSTRAINTS** on the factor matrices?

NONNEGATIVE MATRIX FACTORIZATION



$$\min_{Y,X} \; \|D - YX^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}_+^{n\times k},\; Y \in \mathbb{R}_+^{m\times k}$$

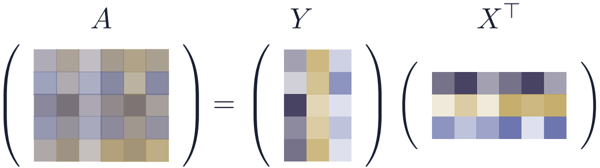
The nonnegative constraints make the polynomially solvable task of SVD into the **NP-HARD** problem of **NMF**.

So, why would I use nonnegative constraints?

Parts-Based Representation

The basis-coefficient interpretation of MF

If we have a factorization of the following form:



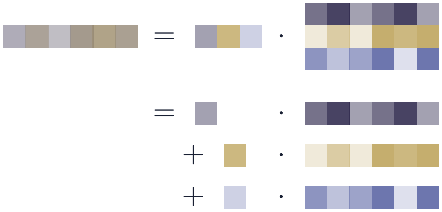
Then the approximation of the jth row is given by

$$D_{j\cdot} \approx Y_{j\cdot} X^\top = \sum_{c=1}^{k} Y_{jc} X_{\cdot c}^\top,$$

The Basis-Coefficient Interpretation of MF

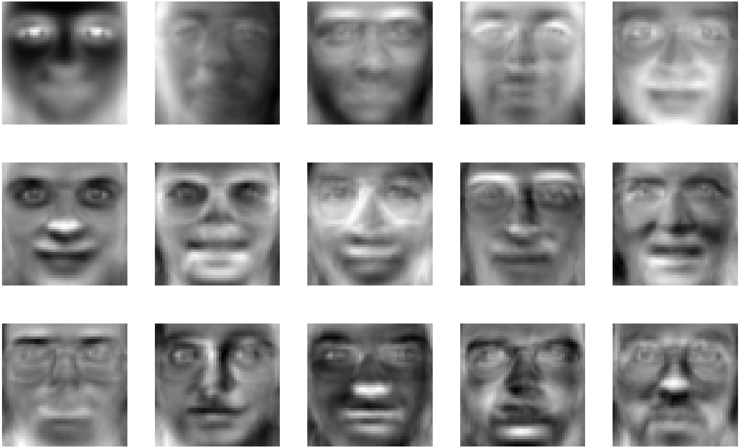
$$D_{j\cdot} \approx Y_{j\cdot} X^\top = \sum_{c=1}^{k} Y_{jc} X_{\cdot c}^\top ,$$

$$A_{j\cdot} \quad = Y_{j\cdot} \qquad X^\top$$



Let us have a look at the visualization of found feature/basis vectors of **SVD** and **NMF**

SVD FACTORIZATIONS OF FACES



NMF Factorizations of Faces

Optimization

OBJECTIVE PROPERTIES

$$\min_{Y,X} \|D - YX^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}_+^{n\times k},\ Y \in \mathbb{R}_+^{m\times k}$$

The NMF objective is **NONCONVEX** in (X, Y) but convex in X if Y is fixed and vice versa.

In principle, we could do alternating minimization:

$$X_{t+1} \in \arg\min_X F(X, Y_t) \tag{1}$$

$$Y_{t+1} \in \arg\min_Y F(X_{t+1}, Y), \tag{2}$$

but the solutions to Eqs. (1) and (2) are **NOT KNOWN**.

Grippo & Sciandrone 2000

Most popular NMF optimizations are **LAZY** versions of **ALTERNATING MINIMIZATION**.

MULTIPLICATIVE UPDATES

$$X_{ic} \leftarrow X_{ic} \frac{D_{\cdot i}^{\top} Y_{\cdot c}}{X_{i\cdot} Y^{\top} Y_{\cdot c}}$$
$$Y_{jc} \leftarrow Y_{jc} \frac{D_{j\cdot} X_{\cdot c}}{Y_{j\cdot} X^{\top} X_{\cdot c}}.$$

Multiplicative updates can be seen as GD with a small enough step-size such that constraints are not violated.

PROS:

- iterates stay nonnegative
- other constraints are usually easy to integrate
- no hyperparameters except for rank

CONS:

- slow convergence (if it converges)
- once an element X_{ic} or Y_{jc} is zero, it stays zero

Lee & Seung 2001, Gillis & Glineur 2012

PROJECTED GRADIENT DESCENT[1]

$$X_{t+1} \leftarrow [X_t - \alpha_t \nabla_X F(X_t, Y_t)]_+$$
$$Y_{t+1} \leftarrow [Y_t - \beta_t \nabla_Y F(X_{t+1}, Y_t))]_+$$

Pros:

- faster convergence than MU

Cons:

- sensitive to stepsize
- unclear how to integrate other constraints

There has been little theory about PGD until the Proximal Alternating Linearized Minimization has been proposed.

Lin 2007

Proximal Alternating Linearized Minimization (PALM)

Given the an objective of the following form:

$$\min_{X,Y} \|D - YX^\top\|^2 + \phi(X) + \phi(Y)$$

The PALM updates are defined as follows:

$$X_{t+1} = \text{prox}_{\alpha_t \phi}\left(X_t - \alpha_t \nabla_X F(X_t, Y_t)\right)$$
$$Y_{t+1} = \text{prox}_{\beta_t \phi}\left(Y_t - \beta_t \nabla_Y F(X_{t+1}, Y_t)\right)$$

The proximal mapping is an optimization problem itself:

$$\text{prox}_\phi(X) \in \arg\min_{X^\star} \left\{ \frac{1}{2}\|X - X^\star\|^2 + \phi(X^\star) \right\}$$

Bolte et al. 2014

PALM NMF

$$X_{t+1} \leftarrow [X_t - \alpha_t \nabla_X F(X_t, Y_t)]_+ \qquad \alpha_t < \frac{1}{\|Y_t^\top Y_t\|}$$

$$Y_{t+1} \leftarrow [Y_t - \beta_t \nabla_Y F(X_{t+1}, Y_t))]_+ \qquad \beta_t < \frac{1}{\|X_{t+1}^\top X_{t+1}\|}$$

Pros:

- guaranteed convergence to a stationary point
- rich theory, improvements are actively researched
- no hyperparameters except for rank

Cons:

- feasibility of additional constraints depends on availability of prox-operator

Pock & Sabach 2016, Driggs et al. 2020

Conclusions

DISCUSSION

- NMF optimization is an old, but ongoing topic of research
- NMF and other MF objectives need GPUs for fast updates. Implementing SOTA optimization methods e.g., in PyTorch, is not un-tricky.
- Nonnegative constraints are also discovered in deep learning and the synergy between the effects would be interesting to explore Sivaprasad et al. 2021

Some References Part II

Parikh & Boyd 2014	Survey proximal Methods
Wang & Zhang 2012	Survey NMF
Bolte et al. 2014	PALM
Udell 2015	Survey low-rank models