Nonnegative Matrix Factorization Benchmark

Vincent-Cuaz Cédric, Le Pelletier de Woillemont Pierre

Guillaume Lecué – ENSAE ParisTech

29/03/2019

Introduction

NMF – Definition

Definition

Given a non-negative matrix $X \in \mathbb{R}_{+}^{f*n}$, algorithms aim to find non-negative matrices $F \in \mathbb{R}_{+}^{f*r}$ and $W \in \mathbb{R}_{+}{}^{r*n}$, such as $X \approx FW$

$$\min_{F,W \geq 0} G(F,W) = \min_{F,W \geq 0} \frac{1}{2}||X - FW||_F^2 \qquad (1)$$

First-order optimality conditions (FOOC)

$$F \geq 0, \quad \nabla_F G = FWW^T - XW^T \geq 0, \quad F \circ \nabla_F G = 0$$
$$W \geq 0, \quad \nabla_W G = F^T FW - F^T X \geq 0, \quad W \circ \nabla_W G = 0$$

Multiplicative Updates – Theory

Theorem (Lee & Seung- 2001) The Euclidean distance is nonincreasing under the update rules:

$$W_{(t+1)} \leftarrow W_{(t)} * \frac{F_{(t)}^T . X}{F_{(t)}^T . F_{(t)} . W_{(t)}} \qquad F_{(t+1)} \leftarrow F_{(t)} * \frac{X . W_{(t+1)}^T}{F_{(t)} . W_{(t+1)} . W_{(t+1)}^T}$$

The Euclidean distance is invariant under these updates iif F and W are at a stationary point of the distance.
Theoretical problem: If an entry (of W or F) is null, these updates cannot modify it but it is possible that its partial derivative is negative, which implies contradiction with FOOC.

Gillis & Glineur – Theory

Gillis & Glineur (GG) suggested a solution to solve this problem.

Theorem (Gillis & Glineur-2008)

For any constant $\delta > 0$, $X \geq 0$ and any $(F, W) \geq \delta$, $||X - FW||_F$ is nonincreasing under

$$F \leftarrow max(\delta, F \circ \frac{XW^T}{FWW^T}), \qquad W \leftarrow max(\delta, W \circ \frac{F^T X}{F^T FW})$$

LS-MU and GG-MU – Results

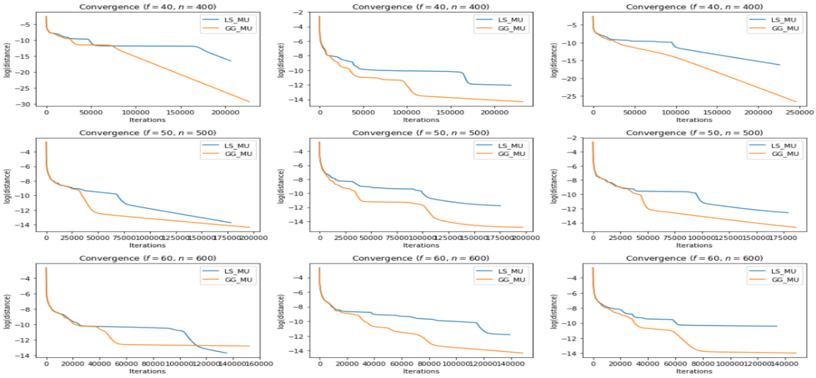


Figure 1: Convergence Profiles – MU and GG

- Convergence depends on initialization
- In most cases, modification provided by GG allows a faster convergence

Hierarchical Alternating Least Squares – Theory

HALS considers another optimization scheme considering successively each rank-one factor $F_{:,k} W_{k,:}$ while fixing the rest of the variables

$$X \approx F_{:,k} W_{k,:} + \sum_{i \neq k} F_{:,i} W_{i,:} \Leftrightarrow F_{:,k} W_{k,:} \approx X - \sum_{i \neq k} F_{:,i} W_{i,:} = R_k$$

Then, for same problems than ill-posed Multiplicative Update algorithm, Gillis & Glineur suggested modified closed-form update rules for HALS:

$$F^{\star}_{:,k} = argmin_{F_{:,k} \geq \delta} ||R_k - F_{:,k} W_{k,:}||_F^2 = \max(\delta, \frac{R_k W_{k,:}^T}{||W_{k,:}||_2^2})$$

$$W^{\star}_{k,:} = argmin_{W_{k,:} \geq 0} ||R_k - F_{:,k} W_{k,:}||_F^2 = \max(\delta, \frac{F_{:,k}^T R_k}{||F_{:,k}||_2^2})$$

They proved that this variant of the algorithm is now well-defined in all situations and converges to a stationary point.

Acceleration Methods – Theory

In their 2011 paper, N.Gillis and F.Glineur discussed two different strategies to choose the number of inner iterations while doing asymmetric quantities of updates:

- A fixed number of inner iterations determined by the flop counts
- A dynamic stopping criterion that checks the difference between two consecutive iterates :

Noting $F^{(k,l)}$ the iterate after l updates of $F^{(k)}$, they stopped inner iterations as soon as

$$||F^{(k,l+1)} - F^{(k,l)}||_F \leq \epsilon ||F^{(k,1)} - F^{(k,0)}||_F$$

Therefore, by combining both, they introduce a new method which finally turns out to be the fastest.

Acceleration Methods – Results

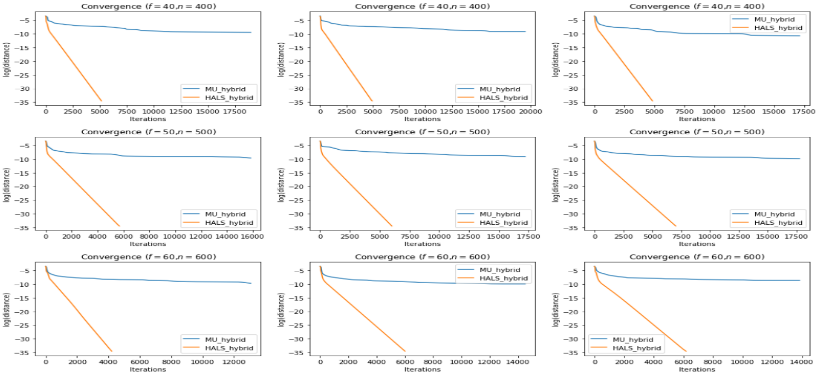


Figure 2: Convergence Profiles – Acceleration Methods

- Still sensitive to initializations
- Significative improvement of convergence speeds, especially with HALS

Separability or Near-Separability assumption

Definition A set of vectors $\{\mathbf{x}_1, ..., \mathbf{x}_r\} \subset \mathbb{R}^d$ is simplicial if no vector $\mathbf{x}_i$ lies in the convex hull of $\{\mathbf{x}_j : i \neq j\}$. The set of vectors is α-robust simplicial if, for each i, the l_1 distance from $\mathbf{x}_i$ to the convex hull $\{\mathbf{x}_j : i \neq j\}$ is at least α.

Definition A NMF $\mathbf{X} = \mathbf{F}\mathbf{W}$ is called separable if the rows of $\mathbf{W}$ are simplicial and there is a permutation matrix Π such that :

$$\Pi\,\mathbf{F} = \begin{bmatrix} I_r \\ M \end{bmatrix}$$

Definition A NMF $\mathbf{X} = \mathbf{Y} + \Delta = \mathbf{F}\mathbf{W}$ is called near- separable when $\mathbf{Y}$ admits a separable NMF and Δ is bounded(by $\epsilon > 0$).

AGKM – Theory

Requirements Columns of X normalized; knowledge of α and ϵ.

Algorithm

1: Initialize $R = \emptyset$.
2: Compute the $f \times f$ matrix $\boldsymbol{D}$ with $D_{ij} = \|\boldsymbol{X}_{i\cdot} - \boldsymbol{X}_{j\cdot}\|_1$.
3: **for** $k = 1, \dots f$ **do**
4:   Find the set $\mathcal{N}_k$ of rows that are at least $5\epsilon/\alpha + 2\epsilon$ away from $\boldsymbol{X}_{k\cdot}$.
5:   Compute the distance δ_k of $\boldsymbol{X}_{k\cdot}$ from $\mathrm{conv}(\{\boldsymbol{X}_{j\cdot} \,:\, j \in \mathcal{N}_k\})$.
6:   **if** $\delta_k > 2\epsilon$, add k to the set R.
7: **end for**
8: Cluster the rows in R as follows: j and k are in the same cluster if $D_{jk} \leq 10\epsilon/\alpha + 6\epsilon$.
9: Choose one element from each cluster to yield $\boldsymbol{W}$.
10: $\boldsymbol{F} = \arg\min_{\boldsymbol{Z} \in \mathbb{R}^{f \times r}} \|\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}\|_{\infty,1}$

Theorem Supposing $\epsilon \leq \frac{\alpha^2}{20+13\alpha}$, $||\Delta||_{\infty,1} \leq \epsilon$. Then AGKM algorithm finds a rank-r NMF $\hat{F}\hat{W}$ that satisfies the error bound $||X - \hat{F}\hat{W}||_{\infty,1} \leq \frac{10\epsilon}{\alpha} + 7\epsilon$

XRAY algorithms

Requirements: Columns of X normalized or not (depending on algorithm); factorization rank r.

Algorithms intuition

- The goal in exact NMF is to find a matrix W such that the cone generated by its columns (ie their non-negative linear combinations) contains all columns of X
- Under separability assumption the columns of matrix W are to be picked directly form X
- These algorithms proposed by A.Kumar build the cone incrementally (r iterations) by picking an anchor column from X in every iteration $\rightarrow$ furthest point from current cone $\rightarrow$ projection and residuals updates

Near Separable – Results

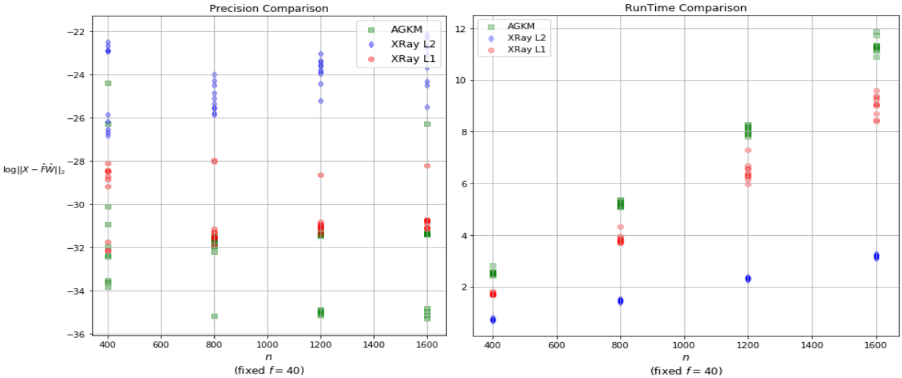


Figure 3: Performances of Near Separable Methods

- AGKM is the slowest but achieve the best accuracy
- Xray L2 is by far the fastest but also the less accurate
- Xray L1 seems to be a good trade-off between speed and accuracy

Noisy Input

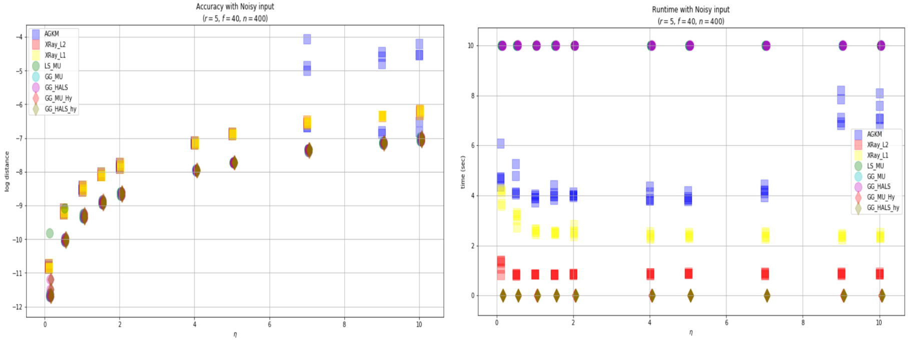


Figure 4: Benchmark based on noise intensity

- Gradient methods are the slowest
- Except for hybrid acceleration methods that are not impacted by noise
- AGKM is the most affected by noise intensity

Noisy Input With duplicates

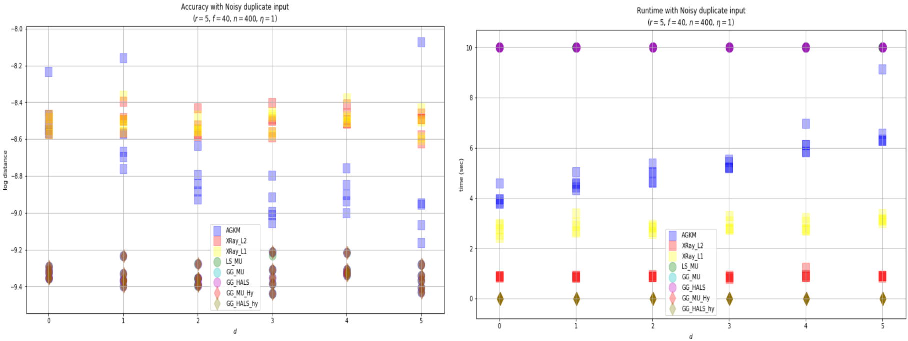
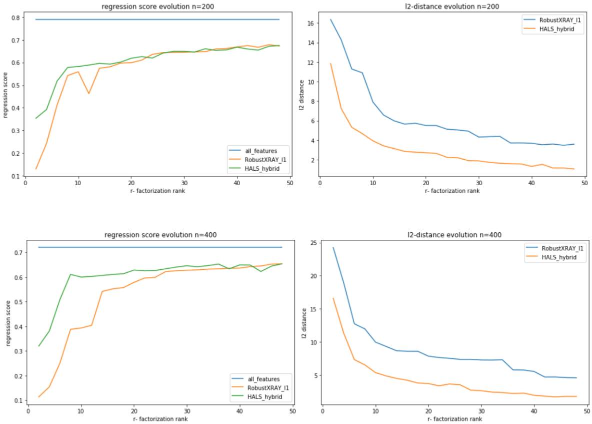


Figure 5: Benchmark based on noise intensity

- Hybrid acceleration methods are not impacted by duplicates in the hott topics
- Duplicates make Xrays methods more volatile

Regression- Communities and Crimes Dataset

Dataset: (1994*101) 100 real features to explain total number of violent crimes in USA.



Conclusion

- HALS with hybrid criterion proposed by Gillis and Glineur and XRAY proposed by Kumar have best trade-off (accuracy/time) amoung studied algorithms.
- Efficient features selection (XRAY) or dimension reduction(TBCD) for regression or classification tasks
- Sparser representations of NMF can be enforced through regularizations for clustering and recommandation system.