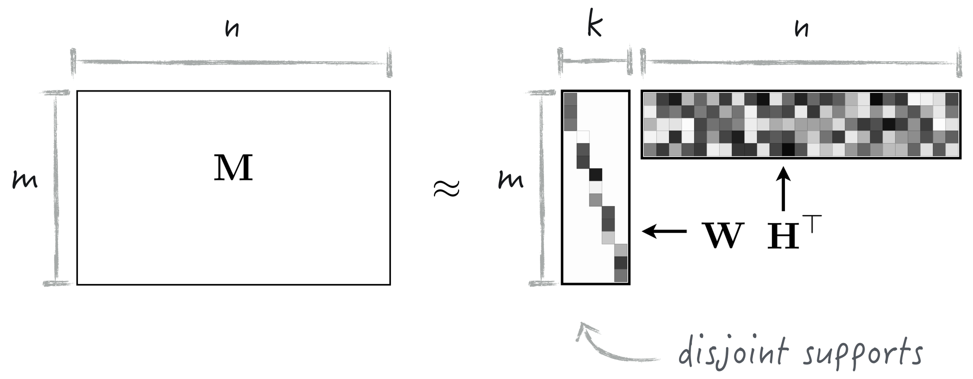
Orthogonal NMF through Subspace Exploration

Megasthenis Asteris, Dimitris Papailiopoulos, Alex Dimakis

TEXAS
The University of Texas at Austin

[Orthogonal NMF]

Given a nonnegative $m \times n$ matrix $\mathbf{M}$ and a target dimension $k \ll m, n$ approximate $\mathbf{M}$ by the product of an $m \times k$ nonnegative matrix $\mathbf{W}$ with orthogonal columns, and an $n \times k$ nonnegative matrix $\mathbf{H}$, i.e.,



As an optimization:

(ONMF) Input

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{M} - \mathbf{W}\mathbf{H}^\top\|_F^2$$

$$\text{subject to: } \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0}$$

$$\mathbf{W}^\top\mathbf{W} = \mathbf{I}_k$$

Orthogonal columns

Sparse, part-based representations.

- *Clustering*: **W** is as a cluster membership matrix, **H** corresponds to the k cluster centroids [Ding 2006, Li 2006, Pompili 2012]
- *Text analysis*: **M** is a words by documents matrix, and the orthogonal columns **W** can be interpreted as *topics* defined by disjoint subsets of words.

[Connections to Nonnegative PCA]

If an oracle reveals $\mathbf{W} \geq \mathbf{0}$ then $\mathbf{H} = \mathbf{M}^\top\mathbf{W} \geq \mathbf{0}$.

By assumption nonnegative

By construction: orthogonal columns + nonnegative

(ONMF) reduces to:

(NNPCA)

$$\max_{\mathbf{W} \in \mathcal{W}_k} \|\mathbf{M}^\top\mathbf{W}\|_F^2$$

$$\mathcal{W}_k = \{\mathbf{W} \in \mathbb{R}^{m \times k} :\ \mathbf{W} \geq \mathbf{0},\ \mathbf{W}^\top\mathbf{W} = \mathbf{I}_k\}$$

- Seek k orthogonal and nonnegative components that jointly maximize the explained variance on the (centered) data **M**.
- Defined for arbitrary matrix **M** (not necessarily nonnegative).
- NP hard even for k=1 (single component NNPCA).

[Contributions]

- Novel approximation algorithm for ONMF with provable guarantees.
- Strictly satisfies both *nonnegativity* and *orthogonality*.
- Relies on novel approximation algorithm for NNPCA.
  (No assumption on NNPCA input — not even nonnegativity).

[Our NNPCA Algorithm]

Input: $m \times n$ matrix $\overline{\mathbf{M}}$ with $\text{rank}(\overline{\mathbf{M}}) = r$ For now, think $r \ll m, n$

Objective: (approximately) solve

$$\overline{\mathbf{W}}_\star = \arg\max_{\mathbf{W} \in \mathcal{W}_k} \|\overline{\mathbf{M}}^\top\mathbf{W}\|. \qquad (*)$$

Observation I

Input matrix can be decomposed as $\overline{\mathbf{M}} = \overline{\mathbf{U}}\,\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{V}}^\top$ (r: rank) and

$$\|\overline{\mathbf{M}}^\top\mathbf{W}\|_F^2 = \|\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{U}}^\top\mathbf{W}\|_F^2 = \sum_{j=1}^{k}\|\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{U}}^\top\mathbf{w}_j\|_2^2 = \sum_{j=1}^{k}\max_{\mathbf{c}_j \in \mathbb{S}_2^{r-1}}\langle \mathbf{w}_j, \overline{\mathbf{U}}\overline{\boldsymbol{\Sigma}}\mathbf{c}_j\rangle^2$$

Cauchy-Schwarz

r-dimensional unit l_2 sphere

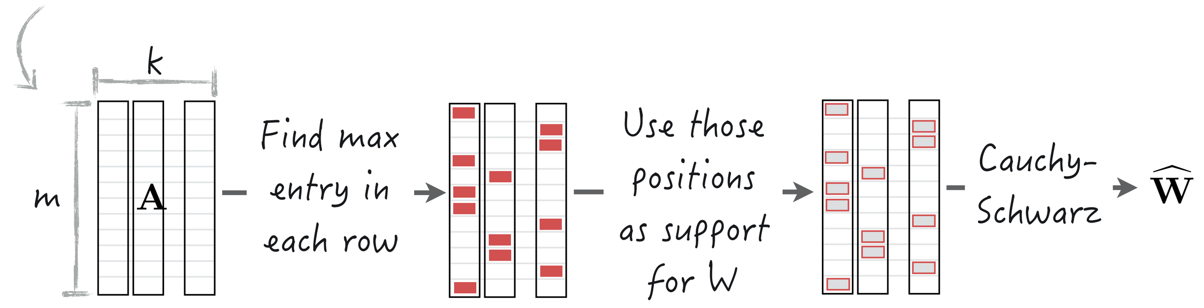
⇒ (∗) can be written as joint maximization over $\mathbf{W}$ and $\mathbf{C}$
m x k r x k

Observation II

Fix the value of the $r \times k$ variable $\mathbf{C}$. Let $\mathbf{A} \leftarrow \overline{\mathbf{U}}\overline{\boldsymbol{\Sigma}}\mathbf{C}$
m x k

Easy to solve. How? (almost)

$$\Big[\ \widehat{\mathbf{W}} = \arg\max_{\mathbf{W} \in \mathcal{W}_k}\sum_{j=1}^{k}\langle \mathbf{w}_j, \mathbf{a}_j\rangle^2\ \Big]$$



⇒ Low rank NNPCA reduces to finding the best $\mathbf{C}$ r x k variable. (smaller than W)

[Algorithm: Low Rank NNPCA]

Input: $m \times n$ rank-r matrix $\overline{\mathbf{M}}$

- Initialize empty collection $\mathcal{S}$
- $\overline{\mathbf{U}}, \overline{\boldsymbol{\Sigma}}, \overline{\mathbf{V}} \leftarrow \text{SVD}(\overline{\mathbf{M}})$ (Truncated SVD)
- For $i = 1 : O\left((4/\epsilon)^{r \cdot k}\right)$
  - Sample $\mathbf{C}$ (r x k variable. Each column is unit-norm)
  - Compute $\mathbf{A} \leftarrow \overline{\mathbf{U}}\overline{\boldsymbol{\Sigma}}\mathbf{C}$
  - Solve

$$\Big[\ \widehat{\mathbf{W}} = \arg\max_{\mathbf{W} \in \mathcal{W}_k}\sum_{j=1}^{k}\langle \mathbf{w}_j, \mathbf{a}_j\rangle^2\ \Big]$$

  - Add $\widehat{\mathbf{W}}$ to the collection $\mathcal{S}$.

Output: $\overline{\mathbf{W}}$ best solution in collection $\mathcal{S}$.

Theorem: Low Rank NNPCA

Input: *i)* $m \times n$ rank-r matrix $\overline{\mathbf{M}}$, *ii)* k: desired # of components, *iii)* accuracy $\epsilon \in (0, 1)$.

Output: $\overline{\mathbf{W}} \in \mathcal{W}_k$ such that

$$\|\overline{\mathbf{M}}\,\overline{\mathbf{W}}\|_F^2 \geq (1 - \epsilon) \cdot \mathsf{OPT},$$

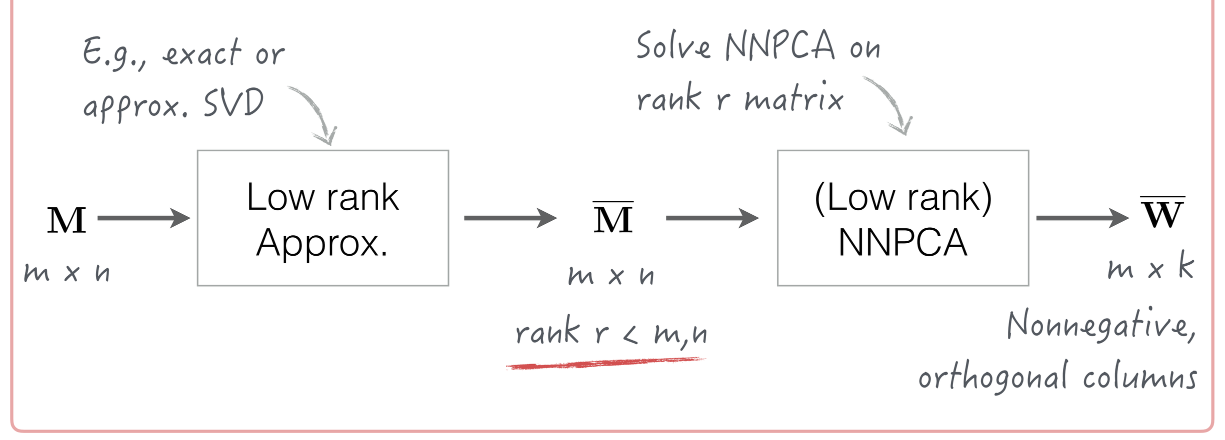
in time $T_{\mathsf{SVD}}(r) + O\left(\left(\frac{2}{\epsilon}\right)^{r \cdot k} \cdot k \cdot m\right)$.

- First algorithm for multi-component NNPCA.
- Provable approximation guarantees.
- Complexity:
  - Low-order polynomial in the ambient dimension (m, n)
  - Exponential in the intrinsic dimension r.

Too high?

Separates ambient and intrinsic dimension.

[Algorithm: NNPCA]



Theorem: NNPCA

Input: *i)* $m \times n$ matrix $\mathbf{M}$,
ii) k: desired # of components, *iii)* accuracy $\epsilon \in (0, 1)$.
iv) rank-r approximation $\overline{\mathbf{M}}$

Output: $\overline{\mathbf{W}} \in \mathcal{W}_k$ such that

$$\|\mathbf{M}\,\overline{\mathbf{W}}\|_F^2 \geq (1 - \epsilon) \cdot \mathsf{OPT} - k \cdot \|\mathbf{M} - \overline{\mathbf{M}}\|_2^2,$$

in time $T_{\mathsf{SVD}}(r) + O\left(\left(\frac{2}{\epsilon}\right)^{r \cdot k} \cdot k \cdot m\right)$.

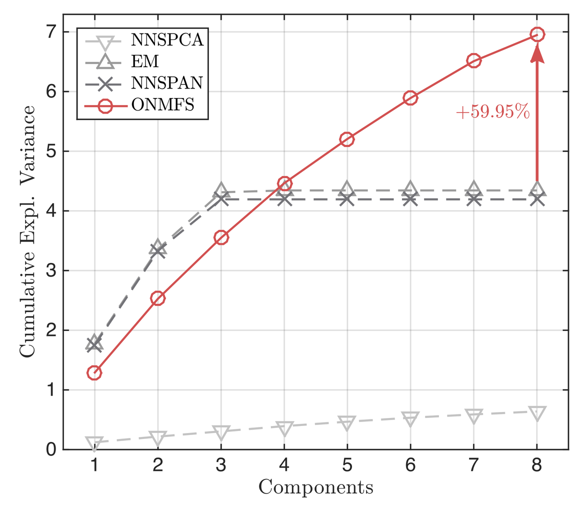
2 levels of approximation

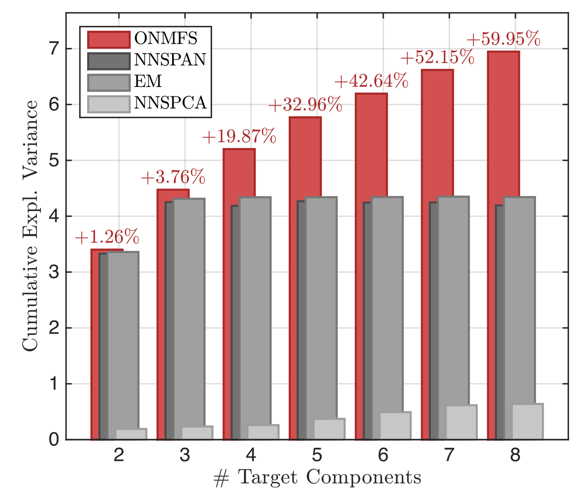
[In Practice]

- Taking too long?
- Run our algorithm and stop it any time.
  → Ignore the theoretical guarantees
  → Still finds solutions with higher explained variance, compared to *deflation* based methods.

Example: CBCL dataset [Sung1996}
2429 19 x19 pixel, gray scale face images.

- Computes k=8 components
- Compare to deflation methods using solvers for the single component NNPCA problem. [Zass:2007, Sigg 2008, Asteris 2014]



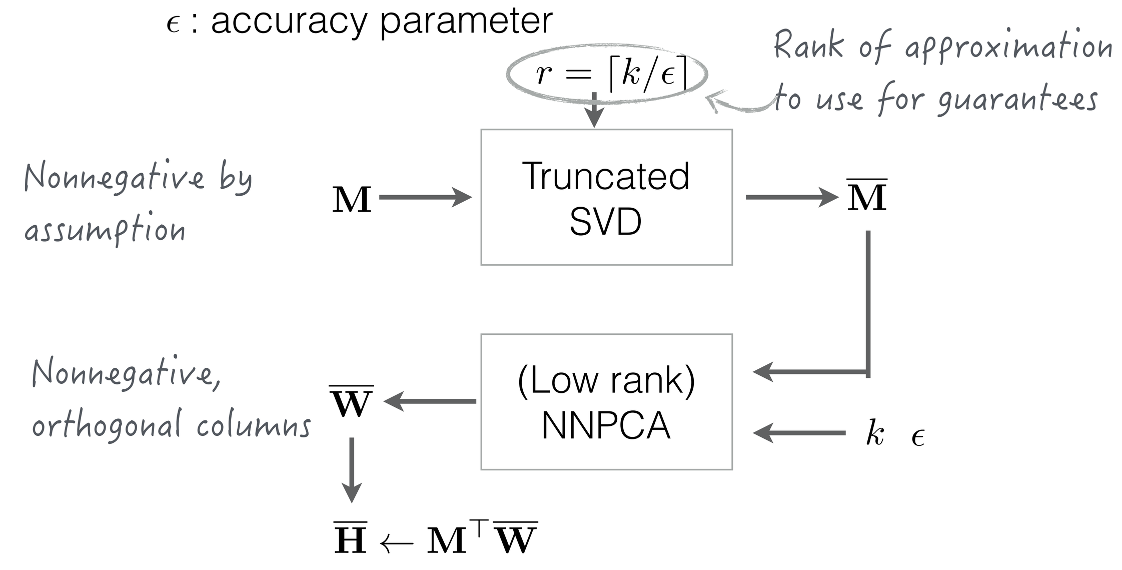


[Our ONMF Algorithm]

Use the NNPCA algorithm to approximately solve the ONMF problem.

[Algorithm: ONMF]

Input: $m \times n$ nonnegative matrix $\mathbf{M}$
k: inner dimension of factorization
ϵ: accuracy parameter



Output: Pair $\overline{\mathbf{W}}, \overline{\mathbf{H}}$

Theorem: ONMF

Our ONMF algorithm outputs an $m \times k$ matrix $\mathbf{W}$ (nonnegative, orthogonal), and an $n \times k$ matrix $\mathbf{H}$ (nonnegative), such that

$$\|\mathbf{M} - \mathbf{W}\mathbf{H}^\top\|_F^2 \leq \mathcal{E}_\star + \epsilon \cdot \|\mathbf{M}\|_F^2$$

in time $T_{\mathsf{SVD}}\left(\frac{k}{\epsilon}\right) + O\left(\left(\frac{1}{\epsilon}\right)^{k^2/\epsilon} \cdot (k \cdot m)\right)$.

Additive guarantee on relative frob. error

- First algorithm for ONMF with provable approximation guarantees.
- No assumption on input (beyond nonnegativity).
  No generative model.
- Output strictly satisfies requirements (nonnegativilty, orthogonality)
- Complexity: Low-order polynomial in the ambient dimension (m, n)
- For constant target dimension k: *additive Efficient Polynomial Approximation Scheme (EPTAS) on the relative Frobenius error.*

[In Practice]

(Similar to NNPCA…)

- Run the NNPCA solver on an arbitrary low rank approximation.
- Stop any time, ignoring the theoretical guarantees.

Example: Synthetic

- Generate nonnegative data ~ close to ONMF + noise.
- Compare with several algos: O-PNMF [Yang 2010], ONP-MF, EM-ONMF [Pompili 2012, 2013], k-means, spherical k-means