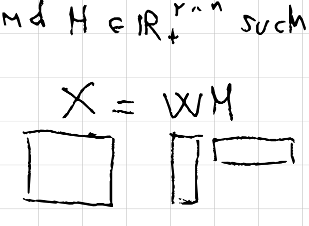
Lesson 4 ~> start with slider app.

[3] Nonnegative Matrix Factorization (NMF)

Given $X \in \mathbb{R}_+^{d\times n}$ and $r \le \min(m,n)$, the NMF problem is to find $W \in \mathbb{R}_+^{d\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ such that

$$X = WH$$



Main keyword: Interpretability

or that minimize

$$\min_{W\in\mathbb{R}_+^{d\times r},\ H\in\mathbb{R}_+^{r\times n}} \|X - WH\|_F^2$$

Preprocessing: remove zero columns from X, $0 = W\cdot 0$. Now everything can be normalized with respect to $\|\cdot\|_1$ vector norm:

$$X = WH \iff XD = W\cdot(HD) \iff XD = (W\tilde D)(\tilde D^{-1} H D)$$

For any $D, \tilde D$ diagonal invertible matrices. We can take

$$D = \mathrm{diag}(\|X_{:,i}\|_1^{-1}),\quad \tilde D = \mathrm{diag}(\|W_{:,i}\|_1^{-1}) \text{ to obtain}$$

$$\tilde X = \tilde W \tilde H,\quad e^T\tilde X = e^T,\quad e^T\tilde W = e^T,\quad e^T\tilde H = e^T\tilde W\tilde H = e^T$$

~> we obtain an equivalent NMF $X = WH$ where $X, W, H \ge 0$ and H is column-stochastic

$$X_{:,i} = h_{1,i} W_{:,1} + h_{2,i} W_{:,2} + \dots + h_{r,i} W_{:,r}$$

$\Rightarrow$ each data $x_i = X_{:,i}$ is now a convex combination of the columns of W with weights given by $H_{:,i}$

We can thus interpret each x_i as a superposition of "features" w_j where $h_{j,i}$ says what % of w_j is inside x_i

~> in case of clustering, we can take the biggest $h_{j,i}$ to choose which w_j represents x_i better

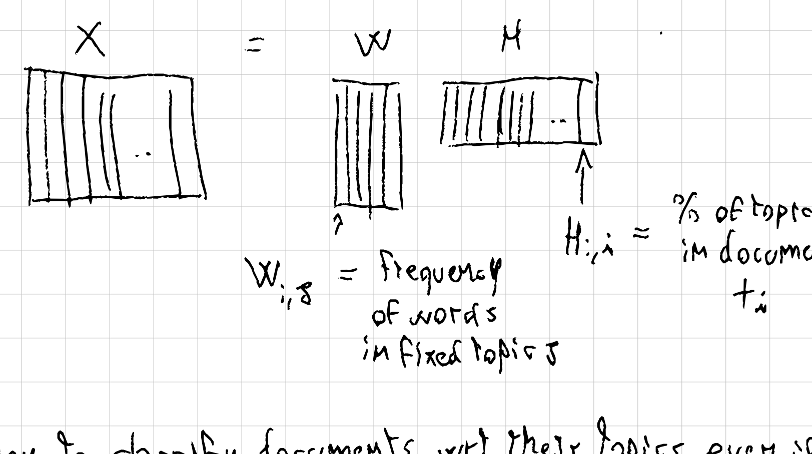
Example with Text Mining: Suppose we have n documents written over with a dictionary of d words. $X_{j,i}$ will be the frequence of word d_j in document t_i

$$X_{j,i} = \frac{\#\{\text{occurrence of word } d_j \text{ in } t_i\}}{\#\{\text{words in } t_i\}}$$

A document may cover different topics. The idea is that each topic is identified by a set of word that may intersect, and has some unique ones.

sports ~ score, play, goal; religion ~ church, priest, mass; politics ~ minister, government, leadership

document covering more topics will have a mix of those defining words inside



$W_{:,j}$ = frequency of words in fixed topic j; $H_{j,i}$ = % of topics in document t_i

This is a way to classify documents with their topics even if they contain more than one topic or also a way to identify the main topic of each text.

In the case of Exact NMF $X = WH$, we call

Nonnegative Rank: $\mathrm{Rank}_+(X) = \min\{r \mid W \in \mathbb{R}_+^{d\times r}, H \in \mathbb{R}_+^{r\times n}\}$

where obviously $\mathrm{rank}_+(X) \ge \mathrm{rank}(X)$. There is a big literature about the NNR, but here we give only some defining properties

- If $\mathrm{rank}(X) \le 2$, then $\mathrm{rank}(X) = \mathrm{rank}_+(X)$. For $\mathrm{rank}(X) = 3$, there are examples for $\mathrm{rank}_+(X) = \Omega(\sqrt{d})$
- $\mathrm{Rank}_+(X) \le \min\{d, n\}$, since $X = I_d X = X I_n$
- Up to a traslation of data $X \mapsto X + ve^T$, $v \ge 0$ we have $\mathrm{rank}(X + ve^T) = \mathrm{rank}_+(X + ve^T)$
- It is NP-Hard to check if $\mathrm{rank}(X) = \mathrm{rank}_+(X)$
- If $\overset{d\times r}{W}, \overset{r\times n}{H} \ge 0$ are randomly generated with continuous distr., then with prob. 1 $\mathrm{rank}_+(X) = \mathrm{rank}(X) = r$

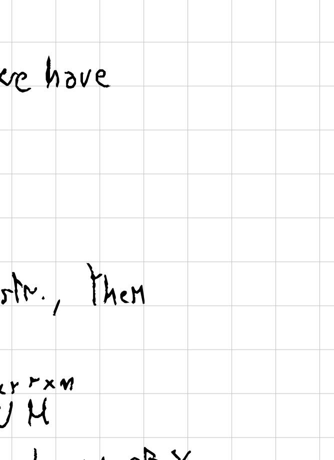
~> In applications, it is usually supposed that $X \sim \overset{d\times r}{W}\overset{r\times n}{H}$ where $r = \mathrm{rank}(W) = \mathrm{rank}(H)$, and r is a numerical rank of X.

Geometric Interpretation

$X = WH$, ≥ 0, ℓ_1-norm.

Since $x_i = \overset{d\times r}{W}\overset{r\times n}{h_i}$ and h_i are convex combinations, then

$$\mathrm{Conv}(X) \subseteq \mathrm{Conv}(W) \subseteq \Delta^r$$



where

$$\Delta^r = \{x \in \mathbb{R}_+^r \mid e^T x = 1\},\quad \mathrm{Conv}(A) = \{Av \mid v \in \Delta^r\}.$$

We thus want the convex polytope $\mathrm{Conv}(W)$ with less vertices possible ($\mathrm{rank}_+(X)$) that contains $\mathrm{Conv}(X)$ and it is contained in Δ^r.

~> This is called "Nested Polytope"

Notice: $\mathrm{Conv}(W) = \Delta^r$ coincides with $W = I$ so $r = d$
$\mathrm{Conv}(W) = \mathrm{Conv}(X)$ " " $W = X$ so $r = n$

~> we are looking for $r \ll \min\{d, n\}$

NPP: Nested Polytope Problem

Given $\mathcal{A} \subseteq \mathcal{B} \subset \mathbb{R}^m$ polytopes full-dimensional with $\mathcal{B}$ bounded, and given $r > 0$ find a polytope $\mathcal{E}$ with r vertices and

$$\mathcal{A} \subseteq \mathcal{E} \subseteq \mathcal{B}$$

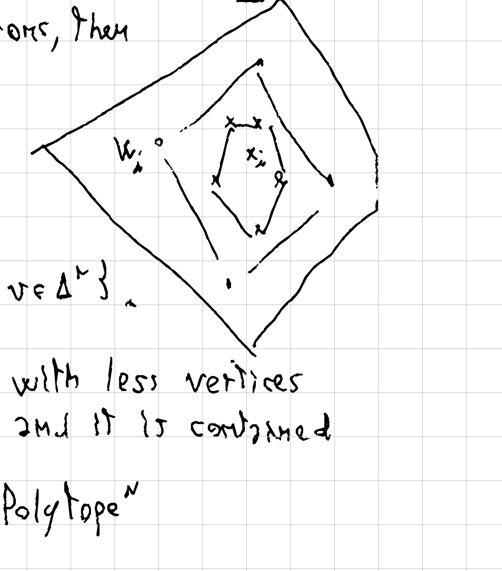
Theorem 11.1 If $m = 2$, NPP is in P [11e]

- Start from a point $P_1 \in \partial\mathcal{B}$
- Compute $P_2 = f(P_1)$, i.e. from P_1 draw the tangent to $\mathcal{A}$ that intersect $\mathcal{B}$ in P_2 and $P_3 = f(P_2)$.

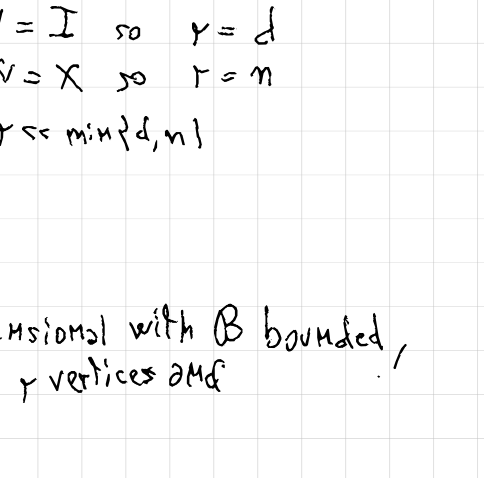
Clockwise

- If $\overline{P_1P_3}$ does not intersect $\mathcal{A}^\circ$, close the polytope, otherwise compute $P_{k+1} = f(P_k)$ until $\overline{P_kP_1}$ does that.
- Identify the "contact change points" between P_1 and P_2:
  - The vertices of $\mathcal{B}$ between P_1, P_2
  - intersection between $\partial\mathcal{B}$ and an edge (the line spanned) of $\mathcal{A}$

  Call them c_i
- The optimal $\mathcal{E}$ is one of the polytopes generated with f starting from one of the c_i



Theorem 11.2 NPP with r vertices ($m = r-1$) and NMF with r, $X = WH$ and $\mathrm{rank}(X) = \mathrm{rank}(W) = r$ are equivalent: there exist polynomial reductions between the two algorithms. In particular, they have the same complexity.



Theorem 11.3 If $X \in \mathbb{R}_+^{d\times n}$ admits $X = WH$, $W, H \ge 0$ with $\mathrm{rank}(W) = \mathrm{rank}(X) \le 3$, then it is possible to recover W, H in poly time

Proof ($r=3$) ℓ_1-normalize X and remove zero columns. Let U be a subset of r indep. columns of X, so that $X = UV$, $U \ge 0$ and ℓ_1-norm. This means $e^T = e^T X = e^T UV = e^T V \Rightarrow V$ is ℓ_1-norm. but may be negative. Call $\bar V$ the matrix V without the last row and $\mathcal{A} = \mathrm{conv}(\bar V)$. Since $r = 3$, $\mathcal{A} \subseteq \mathbb{R}^2$. Moreover if $\bar U$ is U without last column,

$$\mathcal{B} := \{x \in \mathbb{R}^2 \mid \bar U x + u_3(1 - x_1 - x_2) \ge 0\}.$$

Notice that $\mathcal{A} \subseteq \mathcal{B}$ because if $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$ is a column of V, then

$$x_i = Uv \ge 0,\quad Uv = \bar U\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + u_3(1 - v_1 - v_2).$$

Since V has rank 3, $\mathcal{A}$ has dimension 2, and so has $\mathcal{B}$. Notice also

$$\bar U x + u_3(1 - x_1 - x_2) = (u_1 - u_3)x_1 + (u_2 - u_3)x_2 + u_3 = Fx + u_3$$

where F is full rank since U was, and $e^T F = 0$. If $\mathcal{B}$ is unbounded then $\exists\, x, q$ s.t. for every $\lambda > 0$ $F(x + \lambda q) + u_3 \ge 0 \Rightarrow Fq \ge 0$, where $q \ne 0$. $Fq \ne 0$ since F is full rank and $e^T Fq = 0$, that is impossible since $Fq > 0$.

Now we have $X = UV = WH$, with $\mathrm{rank}(U) = \mathrm{rank}(W) = 3$, so they span the same column space $\Rightarrow W = U\cdot C \ge 0$, $UV = UCH \Rightarrow V = CH$ $\Rightarrow \bar V = \bar C H \Rightarrow \mathcal{A} = \mathrm{conv}(\bar V) \subseteq \mathrm{Conv}(\bar C)$ and $e^T = e^T W = e^T UC$

$\Rightarrow e^T = e^T C$, $UC \ge 0 \Rightarrow$ every $\bar c_i$ column of $\bar C$ belongs to $\mathcal{B}$
$\Rightarrow \mathcal{A} \subseteq \mathrm{Conv}(\bar C) \subseteq \mathcal{B}$ ~> it is a solution to NPP(3).

Vice versa, suppose $\mathcal{A} \subseteq \mathcal{E} \subseteq \mathcal{B} \subseteq \mathbb{R}^2$ is a solution of NPP(3) with vertices $[\bar c_1, \bar c_2, \bar c_3] = \bar C$. Complete to C such that $e^T C = e^T$. Since $\mathcal{E} \subseteq \mathcal{B}$, we have $U\cdot C \ge 0$ and we take it as $W = U\cdot C$. Moreover, $\mathcal{A} \subseteq \mathcal{E} \Rightarrow \bar V = \bar C H$ for some $H \ge 0$, $e^T H = e^T$, and $V = CH$ since $e^T V = e^T CH = e^T H = e^T$ $\Rightarrow X = UV = UCH = WH$. Moreover, V is full rank, so C is full rank and as a consequence $\mathrm{rank}(W) = \mathrm{rank}(V) = 3$ ☒

- For $\mathrm{rank}(X) \ge 4$ finding $\min\{r : X = \overset{d\times r}{W}\overset{r\times n}{H},\ W, H \ge 0,\ \mathrm{rk}(W) = \mathrm{rk}(X)\}$ is NP-Hard
- If r is fixed, solving $X = \overset{d\times r}{W}\overset{r\times n}{H}$, $\mathrm{rk}(W) = \mathrm{rk}(X)$ takes $O((dn)^{c\cdot r^2})$ for some constant c, that is way too large.
- As of now, there is no algorithm to solve NMF in $(mn)^{o(r)}$

↪ we solve instead $\min_{W,H \ge 0} \|X - WH\|_F^2$ for fixed r, that takes usually $O(dnr)$ per iteration with iterative solvers, but may converge to local minima.

- Another problem is that even if X has rational entries, its NMF W, H may have irrational entries, not approximable by a machine

Notice that NMF with $r = 1$ is solved easily: take $X \sim uv^T$ best approx. by SVD. Then, since $X \ge 0$

$$\|X - |u||v^T|\|_F^2 \ge \|X - uv^T\|_F^2 \ge \|X - |u||v^T|\|_F^2$$

$\Rightarrow X \sim |u|\cdot|v^T|$ is the sol. to 1-NMF (akin to Perron-Frob.) but if $X \not\ge 0$ then even 1-NMF is NP-Hard.

Identifiability

The NMF $X = WH$ is very not unique. In fact, for every invertible diagonal D with positive diag. entries, and any permut. P

$$X = WH = (WDP)(P^T D^{-1} H)$$

but in application we do not care about the order of the features in W and we can always renormalize them to, for example, unit ℓ_1 norm. As a consequence we define

Uniqueness: We say that the NMF $X = WH$ is unique when $X = \tilde{W}\tilde{H}$ differ by scaling D and permutation P

Scaling and Permutation, in fact, are not the only problem. Even when $\mathrm{rank}(W) = r$, there may be an invertible Q s.t.

$$X = WH = (WQ)(Q^{-1}H), \quad WQ \geq 0, \quad Q^{-1}H \geq 0$$

For example, $X = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} = X \cdot I = (X \cdot X^{-1})(X \cdot I) = IX$.
Uniqueness is important for the application and for stability since

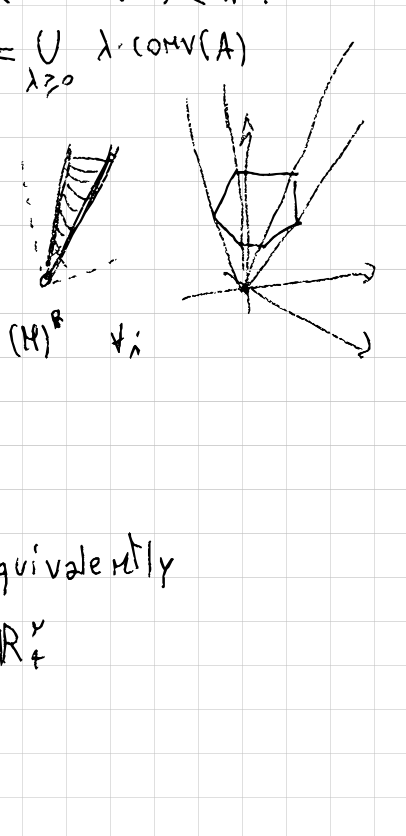
Theorem 12.1 If $X = WH$ NMF is unique, then for $\hat{X}$ in a neigh. of X, $\hat{X} = \hat{W}\hat{H}$ NMF is unique and $(\hat{W}, W)$, $(\hat{H}, H)$ are close up to scaling and permutations.

For $\mathrm{rank}(X) \leq 2$, it is easy:

Theorem 12.2 For $\mathrm{rank}(X) \leq 2$, then $X = WH$ of size 2 always exists. It is unique iff X has a pd diagonal 2x2 submatrix.

To identify the rk 2 NMF it is enough to
- Normalize X with ℓ_1 norm (after removing 0-columns)
- Take $y_1 = x_i$ with greatest $\|\cdot\|$ norm
- Take $y_2 = x_i$ furthest from y_1, i.e. $\max \|x_i - y_1\|$
- identify j s.t. $y_{1,j} \neq y_{2,j}$
- For every i, solve $\begin{bmatrix} y_{1,j} & y_{2,j} \\ 1 & 1 \end{bmatrix} h_i = \begin{bmatrix} x_{i,j} \\ 1 \end{bmatrix}$
- $W = [y_1, y_2]$, $H = [h_i]$

Let's go back to $X = WH \iff \overset{\mathrm{conv}}{\mathrm{Cone}}(X) \subseteq \overset{\mathrm{conv}}{\mathrm{Cone}}(W) \subseteq \overset{\Delta^r}{\mathbb{R}_+^r}$
where $\mathrm{Cone}(A) = \{Ax \mid x \in \mathbb{R}_+^r\} = \bigcup_{\lambda \geq 0} \lambda \cdot \mathrm{conv}(A)$



Dual Cone: Given a cone C, then

$$C^* = \{x \mid y^T x \geq 0 \ \forall y \in C\}$$

Notice that $X = WH \geq 0 \Rightarrow w_i^T \in \mathrm{Cone}(H)^* \ \forall i$

$\Rightarrow \mathrm{Cone}(W^T) \subseteq \mathrm{Cone}(H)^*$

Separability $\overset{r \times m}{A} \geq 0$ is separable if equivalently
- $\mathrm{Cone}(A) = \mathbb{R}_+^r$ • $\mathrm{Cone}(A)^* = \mathbb{R}_+^r$
- $\exists$ subm. of A that is $\overset{r \times r}{D \cdot P}$

Separable: We say that a NMF $X = WH$ is separable if H is sep.

Notice that H separable $\iff DPH$ separable, so we can work with H ℓ_1-norm. In this case, H separable $\iff I_r$ is a subm of H up to permutation, or also said, if W is composed by col. of X.

In applications this has different names. For example in hyperspectral imaging, $X = WH$ is separable if for every material W there exists a pixel made only of that material (pure-pixel assumption). In text mining, sep. of W it means that each topic has a word that is exclusive of that topic (anchor word). In source sep. is that each note must be played alone in a short span time.

Theorem 12.3 If $X = WH$, where $X \in \mathbb{R}^{d \times m}$, $W \in \mathbb{R}^{d \times r}$, $\mathrm{rank}(X) = r$, $H \in \mathbb{R}_+^{r \times m}$ separable, then this is the unique separable MF.

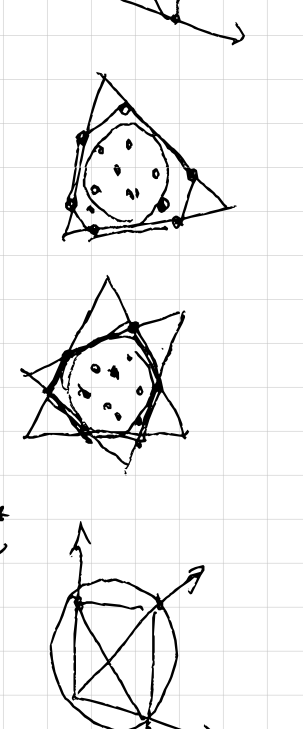
Proof If $X = WH = \hat{W}\hat{H}$ with $H, \hat{H}$ separable, then W and $\hat{W}$ are made of columns of X, but then $\mathrm{Cone}(X) \supseteq \mathrm{Cone}(W) \supseteq \mathrm{Cone}(X)$ and the same with $\hat{W}$. Since $W, \hat{W}$ are full rank, their columns are the vertices of $\mathrm{Cone}(X)$ up to scaling and permutation. Since W is full rank, then H is also unique ∎

⤳ This has the advantage that retrieving a separable MF is doable in poly time or terminating
⤳ In some application like face features extraction, is W to keep. They have the same properties.
The drawback is that separability is very strong and does not hold in many applications, so we need something weaker

Sufficiently Scattered Condition $A \in \mathbb{R}_+^{r \times m}$ is SSC if

SSC1 • $\mathcal{C} := \{x \in \mathbb{R}_+^r \mid e^T x \geq \sqrt{r-1}\,\|x\|\} \subseteq \mathrm{Cone}(H)$

SSC2 • If Q orth. and $\mathrm{Cone}(H) \subseteq \mathrm{Cone}(Q)$, then Q is a perm.



Geometrically $\mathcal{C}$ is the cone of C_r, the circle in Δ^r tangent to $\mathbb{R}_+^r$. In particular, $\mathcal{C} \subseteq \mathbb{R}_+^r$

The idea is that for separability, $\mathrm{Cone}(H) = \mathbb{R}_+^r$, that is, it must contain $\{e_1, e_2, \dots, e_r\}$ among its columns. With SSC, it must contain a smaller $\mathcal{C}$.

SSC2 is saying that there is no other "rotated" Δ^r containing $\mathrm{Cone}(H)$ except for Δ itself. This is satisfied for example when $\mathcal{C}$ is contained in the interior part of $\mathrm{Cone}(H)$.

SSC is a sufficient condition for identifiability for certain problems with minvol, but to prove it we need some properties of dual cones:
- given two convex cones, $A \subseteq B \Rightarrow B^* \subseteq A^*$
- $\mathcal{C}^* = \{y \in \mathbb{R}^r \mid e^T y \geq \|y\|\}$ is the circular cone circumscribed to Δ^r, and it contains also $y \notin \mathbb{R}_+^r$.

⤳ bad news: checking SSC is NP-hard
⤳ good news: If H is randomly generated and each row has at least $r-1$ zeros, then it is SSC with high prob.

⤳ ? news: Uniqueness of NMF do not require SSC of W, H necessarily

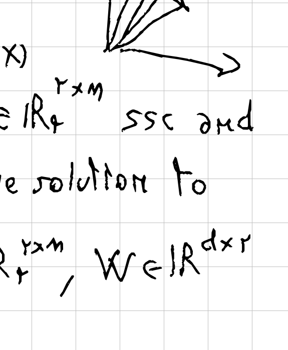
Minimum Volume Recall that when $X = WH$ ℓ_1-norm, then

$$\mathrm{conv}(X) \subseteq \mathrm{Conv}(W) \subseteq \Delta^r$$

but when the decomposition is not unique, one may try to find the W with the minimum volume that contains $\mathrm{Conv}(X)$.

The volume of $\mathrm{Conv}(W \cup \{0\})$ when $W \in \mathbb{R}^{d \times r}$, $d \geq r$ is

$$\mathrm{Vol}(W) = \frac{1}{r!}\sqrt{\det(W^T W)}$$



Theorem 12.4 Let $X = W^\# H^\#$ where $e^T H^\# = e^T$, $H^\# \in \mathbb{R}_+^{r \times m}$ SSC and $X \in \mathbb{R}^{d \times m}$, $W^\# \in \mathbb{R}^{d \times r}$, $r = \mathrm{rank}(X)$. Then (W, H) is the unique solution to

$$\min_{W,H} \mathrm{Vol}(W) : X = WH, \ e^T H = e^T, \ H \in \mathbb{R}_+^{r \times m}, \ W \in \mathbb{R}^{d \times r}$$

Proof Let $X = W^\# H^\# = WH$ where (W, H) is also feasible. Since $r = \mathrm{rank}(X)$, then $W = W^\# Q^{-1}$, $H = QH^\#$. Since $H^\#$ is full rank and $e^T H^\# = e^T$, then $e^T = e^T H = e^T Q H^\# \Rightarrow e^T Q = e^T$. Moreover, $H = QH^\# \geq 0$
$\Rightarrow \mathrm{Cone}(Q^T) \subseteq \mathrm{Cone}(H^\#)^* \subseteq \mathcal{C}^*$ (SSC1) $= \{y \mid y^T e \geq \|y\|\}$
$\Rightarrow q_j$ rows of Q $q_j^T e \geq \|q_j\| \ \forall j$. Then

$$|\det Q|^2 = \prod \|q_j\|^2 \,|\det \overset{\text{row norm}}{D^{-1}} Q|^2 = \prod \|q_j\|^2 \left|\det \overset{\text{SPD}}{\begin{bmatrix} 1 & * \\ * & 1 \end{bmatrix}}\right| \overset{\text{AM-GM}}{\leq} \prod \|q_j\|^2$$

$$|\det Q| \leq \prod \|q_j\| \leq \prod (q_j^T e) \overset{\text{AM-GM}}{\leq} \left(\frac{\sum q_j^T e}{r}\right)^r = \left(\frac{e^T Q e}{r}\right)^r = 1$$

But since $W^\#$ has the least volume, $\det(W^{\#T} W^\#) \leq \det(W^T W) =$

$$= \det(Q^{-T} W^{\#T} W^\# Q^{-1}) = \det(W^{\#T} W^\#) \cdot \det(Q)^{-2} \Rightarrow \det(Q) \leq 1$$

so $|\det Q| = 1$ and by AM-GM above, $q_j^T e = 1 = \|q_j\| \ \forall j$ and moreover $D = I$, $Q^T Q = I \Rightarrow Q^T$ orth., but $\mathrm{Cone}(Q^T)^* = \{y \mid Qy \geq 0\} = \{Q^T z \mid z \geq 0\} = \mathrm{Cone}(Q^T)$, so

$$\mathrm{Cone}(Q^T) \subseteq \mathrm{Cone}(H^\#)^* \Rightarrow \mathrm{Cone}(H^\#) \subseteq \mathrm{Cone}(Q^T) \overset{\text{SSC2}}{\Rightarrow} Q^T \text{ permutation}$$ ∎

⤳ Open Question: minvol(W) is poly? How does it fare with noise?

Possible solution: • maximum ellipsoid inscribed in W (exp)
• max vol of dual polar

Problem: $e^T H = e^T$ might not be a good condition when there are low norm x_i with loose connection with the features W. A better condition that is also more robust wrt noise is $e^T W = e^T$. Theorem 12.4 holds even for the modified min vol problem

$$\min_{W,H} \det(W^T W) : X = WH, \ e^T W = e^r, \ W \in \mathbb{R}^{d \times r}, \ H \in \mathbb{R}_+^{r \times m}$$

Recall that k-means can be recast as

$$\min_{C,H} \|X - CH\|_F^2 \;:\; C \in \mathbb{R}^{d\times r},\; H \in \{0,1\}^{r\times m},\; e^T H = e^T$$

Notice that if $X \geq 0$, then $C \geq 0$ from k-means, since they are averaged points in X. Since $H \geq 0$, this is nothing else than an NMF. We can always bring X to positive by translation and if we normalize them in $\|\cdot\|_1$ norm, we obtain the classic

$$X = WH, \quad X, W, H \geq 0, \quad e^T X = e^T H = e^T, \quad e^T W = e^T$$

$\leadsto$ since in NMF H are not boolean, general NMF is much more flexible to different kind of representations.

A variant of k-means, called Spherical k-means, looks for directions W_j and clusters the data x_i whose angle with W_j are the least. In formulae, it is

$$\arg\max_{W \in \mathbb{R}^{d\times k}} \sum_i \max_{1\leq l \leq k} \frac{W_{:,l}^T x_i}{\|W_{:,l}\| \cdot \|x_i\|} = SK(X)$$

so if X and W are normalized in $\|\cdot\|$, this is

$$\sum_i x_i^T W \overset{h_i}{\overset{\shortparallel}{e_{l(i)}}} = Tr(X^T WH) = -\frac{1}{2}\Big[\|X - WH\|_F^2 - \overset{n}{\overset{\shortparallel}{\|X\|_F^2}} - \overset{n}{\overset{\shortparallel}{\|WH\|_F^2}}\Big]$$

$$\leadsto SK(X) = \arg\min_{H,\, W \in \mathbb{R}^{d\times k}} \|X - WH\|_F^2 \;:\; H \in \{0,1\}^{k\times n},\; e^T H = e^T,\; \|W_{:,j}\| = 1 \;\forall j$$

$$\xrightarrow[X \geq 0]{\text{Relax}} \arg\min_{H,\, W \in \mathbb{R}_+^{d\times k}} \|X - WH\|_F^2 \;:\; H \in \mathbb{R}_+^{k\times m},\; HH^T = I$$

Notice: each col. of H has still only 1 entry > 0

<u>ONMF</u> : Assures that $supp(H_{i,:})$ are distinct and

$$\arg\min_{W \geq 0} \|X - WH\|_F^2 = XH^T$$

so it is equivalent to $\displaystyle \arg\min_{\substack{H \in \mathbb{R}_+^{k\times m} \\ HH^T = I}} \|X - XH^T H\|_F^2 = \arg\max_{\substack{H \in \mathbb{R}_+^{k\times m} \\ HH^T = I}} \|XH^T\|_F^2$

The property HH^T diag. assures that $supp(H_{i,:})$ are disjoint, so this is in itself a clustering algorithm. Moreover it tells us that $H^T H$ is a ~~projector~~ orth. projector to some subspace when $H^T H = I$.

— o — o —

If we now reconsider SBM, we see that $\overset{1/0}{A_n} \sim \mathbb{E}[A_n]$ where

$$\mathbb{E}[A_n] = \begin{bmatrix} P_{11} & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & P_{rr} \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} [P_{ij}]_{ij} \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} = ZPZ^T$$

and $Z, P \geq 0$ $\leadsto$ This in a <u>Tri-Sym NMF</u>

$$\overset{n\times n}{X} \sim \overset{n\times r}{W}\,\overset{r\times r}{S}\,\overset{r\times m}{W^T}, \quad W, S \geq 0 \quad (S \text{ sym if } X \text{ sym})$$

Recall that for Spectral Clustering we had $X \sim U \Lambda U^T$ where we needed to cluster the rows of U. Here instead $W \geq 0$ is more interpretable. In general clustering, W is the "belonging to cluster" matrix and S is instead the relations between different clusters.

Let's see an example of Topic Modeling:

In text mining, we had A matrix of frequency of words in document, and $A = WH$ where W are the words in topics and H is the topic-in-documents matrix. The problem here is that if for example the documents are short (see Twitter or Facebook or X) then even if a post is discussing a topic, it won't use most of the words associated with that. For this reason, we form the co-occurrence matrix $X = AA^T$ that says how common is it that two words are in the same document. Now

$$X \sim WSW^T$$

has the usual word-to-topic W and also a topic-to-topic correlation S

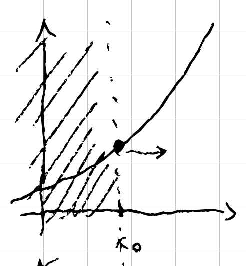
KKT Conditions for NMF

$$\min_{\substack{W \in \mathbb{R}_+^{d\times r} \\ H \in \mathbb{R}_+^{r\times n}}} D(X, WH) \qquad \text{where } X \in \mathbb{R}_+^{d\times n}$$

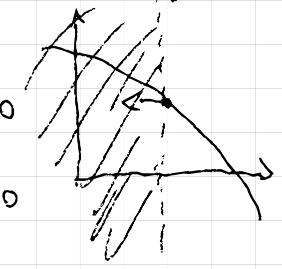
in unconstraint optimization, the 1° order conditions are

$$\nabla_W D(X, WH) = 0, \quad \nabla_H D(X, WH) = 0$$

but with inequalities constraints we need something different. The idea is that if $W_{ij} = 0$, then we don't need that $\partial/\partial w_{ij} D(X, WH) = 0$: we just need to be sure that $D(X, WH)$ does not decrease if we increase w_{ij}, i.e. we just need $\partial/\partial w_{ij} D(X, WH) \geq 0$.



If instead $W_{ij} > 0$, we actually need the gradient $= 0$.
As a consequence, this is written as



$$\begin{cases} W \geq 0, & \nabla_W D(X, WH) \geq 0, \quad W .* \nabla_W D(X, WH) = 0 \\ H \geq 0, & \nabla_H D(X, WH) \geq 0, \quad H .* \nabla_H D(X, WH) = 0 \end{cases}$$

↑ Karush-Kuhn-Tucker conditions

For $D(X, WH) = \|X - WH\|_F^2$, we get for example

$$\nabla_W \|X - WH\|_F^2 = -2XH^T + 2WHH^T \rightsquigarrow (WH - X)H^T \geq 0, \quad W .* (\ ) = 0$$

and similar with H. Notice that $W = 0, H = 0$ is always a solution.

Another one is the Kullback-Leibler (KL) Divergence, i.e.

$$D(A \| B) = \sum_{i,j} A_{ij} \log\left(A_{ij}/B_{ij}\right) - A_{ij} + B_{ij}, \qquad \begin{array}{l} D(x\|0) = \infty \ (x>0) \\ D(0\|x) = x \\ D(0\|0) = 0 \end{array}$$

$$\frac{\partial D(x\|y)}{\partial y} = 1 - x/y \quad \text{so we can compute the grad. of the error :}$$

$$\nabla_W D(X \| WH) = \nabla_W \sum_{i,j} D(x_{ij} \| (WH)_{ij}) = \left[\sum_{i,j}\left(1 - \frac{X_{ij}}{(WH)_{ij}}\right)\frac{\partial (WH)_{ij}}{\partial W_{kh}}\right]_{kh}$$

$$= \left[\sum_j \left(1 - \frac{X_{kj}}{(WH)_{kj}}\right) H_{hj}\right]_{k,h} = ee^T H^T - (X ./ WH) H^T = \left(\frac{WH - X}{WH}\right) H^T$$

$$\nabla_H D(X \| WH) = \left[\sum_i \left(1 - \frac{X_{ih}}{(WH)_{ih}}\right) W_{ik}\right]_{k,h} = \overset{W^T\left(\frac{WH - X}{WH}\right)}{W^T ee^T - W^T (X ./ WH)}$$

The KKT conditions are thus

$$\begin{cases} \frac{WH - X}{WH} H^T \geq 0, & W \odot \left(\frac{WH - X}{WH} \cdot H^T\right) = 0 \\ W^T \frac{WH - X}{WH} \geq 0, & H \odot \left(W^T \cdot \frac{WH - X}{WH}\right) = 0 \end{cases}$$

Again, $(0,0)$ is a solution.

$\rightsquigarrow$ In general any rank-deficient stationary (W, H) are <u>saddle</u> points.
so they are not stable for most algorithms

Depending on the model of our noise/perturbation N s.t. $X = WH + N$, it is opportune to choose different $D(\cdot, \cdot)$.

Noise	$\mathcal{N}(0, \Sigma)$	Uniform	Poisson
Distance	$\sum \frac{1}{\Sigma_{ij}^2}(X_{ij} - (WH)_{ij})^2$	$\max_{i,j} \lvert X_{ij} - (WH)_{ij} \rvert$	$D(X \| WH)$

(we always suppose the entries of N are indep.)
Gaussian or Uniform noise may not be useful when X is sparse, since the noise may generate many negative entries. In these cases, it is better to use the divergence KL. In particular, images tend to be dense, so audio/text $\rightsquigarrow$ KL, Images $\rightsquigarrow$ Frob. or ℓ^p. Notice that

Theorem 13.1 Given $X \in \mathbb{R}_+^{d\times n}$ and $r > 0$, any stationary point $(\bar{W}, \bar{H})$ of $\min_{W \geq 0, H \geq 0} D(X \| WH)$ satisfies

$$Xe = WHe, \quad e^T X = e^T WH$$

<u>Proof</u> a stationary point satisfies $\bar{W} \circ \nabla_W D(X \| \bar{W}\bar{H}) = 0$ since if $\bar{W}_{ij} \neq 0$, then the derivative must be zero (KKT condition) so

$$\left(\bar{W} \circ \frac{\bar{W}\bar{H} - X}{\bar{W}\bar{H}} H^T\right) e = 0 \iff \left(\bar{W} \circ ee^T \bar{H}^T\right) e = \left(\bar{W} \circ \frac{X}{\bar{W}\bar{H}} \bar{H}^T\right) e$$

$$\iff \operatorname{diag}\left(\bar{W} (ee^T \bar{H}^T)^T\right) = \operatorname{diag}\left(\bar{W} \bar{H} \left(\frac{X}{\bar{W}\bar{H}}\right)^T\right)$$

$$\iff \bar{W}\bar{H}e = \left(\bar{W}\bar{H} \frac{X}{\bar{W}\bar{H}}\right) e = Xe, \quad \text{and the same with } H \ \blacksquare$$

Lemma 13.2 Given $X \in \mathbb{R}_+^{d\times n}$, $X \neq 0$, the unique solution to $r = 1$ of $\min D(X \| wh^T)$ is

$$w = Xe, \quad h^T = e^T X / e^T X e \quad \text{up to scaling}.$$

<u>Proof</u> By Th. 13.1, a solution (w, h) satisfies $Xe = w(h^T e)$ and $e^T X = (e^T w) h^T$ so we can take $w = Xe$, $h^T = e^T X / e^T X e$, and it is the unique solution up to scaling. $\blacksquare$

<u>Problem</u> : For every $D(\cdot, \cdot)$ "famous" i.e. KL, $\|\cdot\|_F$, $\|\cdot\|_1$, $\|\cdot\|_\infty$, etc. NMF is NP-Hard, so we need good algorithm for approx. in $r \geq 3$. ℓ_1-norm is also NP-Hard for $r = 1$.

Other results : For stationary points (W, H) in $\|\cdot\|_F$ it holds that

$$\|X - WH\|_F^2 = \|X\|_F^2 - \|WH\|_F^2$$

so $\|X\|_F^2 \geq \|WH\|_F^2$

Lessor 9

SPA : Successive Projection Algorithm

Suppose $\underset{d\times n}{X} = \underset{d\times r}{W}\,\underset{r\times n}{H}$, where H is separable and substochastic, i.e. $H \geq 0$, $e^T H \leq e^T$, $I_r = H(:,J)$ for $|J| = r$. Suppose moreover that W is full rank. Notice that $W = X(:,J)$ and thus
$$\mathrm{conv}(X) \subseteq \mathrm{conv}(W) \subseteq \mathrm{conv}(X) \implies \mathrm{conv}(W) = \mathrm{conv}(X)$$
and the columns of W are the vertices of $\mathrm{conv}(X)$.

SPA recovers the vertices of $\mathrm{conv}(X)$ sequentially. In fact notice that the max ℓ_2-norm column of X must be one of such vertices since the norm is convex, and when we consider it on $\mathrm{conv}(X)$, it takes maximum only on the vertices. In formulae,
$$\|X(:,j)\| = \|WH(:,j)\| \leq \sum_i H(i,j)\cdot\|W(:,i)\| \leq \left(\max_i \|W(:,i)\|\right)\left(\sum_i H(i,j)\right)$$
$$\leq \max_i \|W(:,i)\| \leq \max_j \|X(:,j)\|.$$ For a max-ℓ_2 norm $X(:,j)$ all of these are equalities, but the first is tr. ineq. that is $=$ only when the $W(:,k)$ for which $H(k,j) \neq 0$ are collinear and W is full rank, so that's possible only when there's at most one coeff. $H(k,j) \neq 0$. The 2° ineq. is $=$ only when $H(:,j)^T e = 1$, meaning that $H(k,j) = 1$ and so $X(:,j) = W(:,k)$.

After recovering k vertices $\{W(:,1), \dots, W(:,k)\}$, SPA projects over the orthogonal space to all of these vertices, called V_k. Then it repeats the same procedure, finding the k-th index since $P_{V_k} X = (P_{V_k} W)\cdot H$ and we can remove the 0 columns from W and the corresponding rows from H to obtain a separable $\tilde{H}$ and a rank $(r-k)$ $P_{V_k} W$.

SPA

> Given $X \in \mathbb{R}^{d\times n}$ and $r > 0$, let $J = \emptyset$
> while $|J| < r$
>   Find the max $\|X(:,k)\| = \max_i \|X(:,i)\| = v$
> $J = J \cup \{k\}$
> $X = (I - vv^T/\|v\|^2)\, X$
> end

Notice that X, W does not need to be ≥ 0, but $X \geq 0 \implies W \geq 0$, so it is an NMF. The complexity is $O(dnr)$; and if X is sparse it is $O(r\cdot nnz(X))$

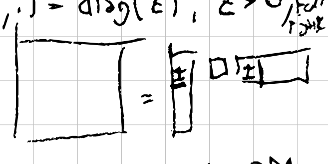
This is also robust to noise in the sense that given $X = WH + N$ where H is separable and substochastic, W is full rank and
$$\varepsilon = \max_j \|N(:,j)\| = O\left(\frac{\sigma_{min}(W)}{\sqrt{r}\,\kappa^2(W)}\right)$$
Then SPA can identify W with an ℓ_2 error up to $O(\varepsilon\,\kappa^2(W))$.
With further improvements (Fast Anchor Words / preconditioning through Min. Vol. Ellipsoid) one gets
$$\varepsilon = O\left(\frac{\sigma_{min}(W)}{\sqrt{r}\,\kappa(W)}\right), \qquad \text{error} = O(\varepsilon\,\kappa(W))$$
$\rightsquigarrow$ we get anyway an error proportional to $\sigma_{min}(W)/\sqrt{r}$.

Notice that if W_1 is very close to the span of the rest $W_2, \dots, W_r$ then after the proj. wrt. $\{W_2, \dots, W_r\}^\perp$, the projected W_1 will have very small norm, thus a random perturbation can easily prevent us to find W_1. Notice that
$$\min_k \min_\alpha \|W_k - W_{\bar{k}}\cdot\alpha\| = \min_{k,\alpha} \left\|W\begin{bmatrix}\alpha\\1\\\alpha\end{bmatrix}\right\| \leq \min_{\|x\|\geq 1} \|Wx\| = \sigma_r(W)$$
This is the reason why ε must be bounded by $\sigma_{min}(W)$.

$\rightsquigarrow$ It is also very sensitive to outliers. Smoothed/random version exist for this reason [13]

$\rightsquigarrow$ It can be proved that SPA is equivalent to $\max_{|J|=r} \mathrm{Vol}(X(:,J))$

Tri-NMF Suppose now we want the Tri-SymNMF $\underset{n\times n}{A} = \underset{n\times r}{W}\,\underset{r\times r}{S}\,\underset{r\times n}{W^T}$ in the case W is separable, i.e. $W(J,:) = \mathrm{diag}(z)$, $z > 0$, S full rank and everything is nonnegative, and A, S are sym. In this case there is an exact poly algorithm.



In fact, $A = (WS)W^T$ is a separable NMF for which we can apply SPA to find J: $A(:,J) = (WS)\cdot\mathrm{diag}(z)$. Moreover,
$$A(J,J) = \mathrm{diag}(z)\,S\,\mathrm{diag}(z) \implies S = \mathrm{diag}(z)^{-1} A(J,J)\,\mathrm{diag}(z)^{-1} \qquad (1)$$
and since we can always suppose $e^T W = e^T$, then
$$A(J,:)e = \mathrm{diag}(z)\cdot S\cdot W^T e = A(J,J)\,\mathrm{diag}(z)^{-1} e = A(J,J)\,[1/z_i]_i$$
$\implies$ we can compute z from $[1/z_i]_i = A(J,J)^{-1} A(J,:)e$ and then compute S by (1) and $A(J,:) = \mathrm{diag}(z) S W^T \implies W^T = S^{-1}\mathrm{diag}(z)^{-1} A(J,:)$.

Tri-symNMF

> Given $A \in \mathbb{R}_+^{n\times n}$ symmetric and $r > 0$
> Compute the index set J from SPA(A, r)
> Compute $y = A(J,J)^{-1} A(J,:)e$
> Compute $S = \mathrm{diag}(y)\cdot A(J,J)\cdot\mathrm{diag}(y)$
> Compute $W^T = S^{-1}\mathrm{diag}(y)\,A(J,:)$

$\rightsquigarrow$ one could compute once $B = A(J,J)^{-1}A(J,:)$, so that $y = Be$ and $W^T = \mathrm{diag}(y)^{-1} B$. If there's noise, instead of inverting $A(J,J)$ one solves
$$y = \arg\min_x \|A(J,:)e - A(J,J)x\|, \quad W^T = \arg\min_M \|A(J,J)\,\mathrm{diag}(y)\,M - A(J,:)\|$$
or with KL-div instead of $\|\cdot\|$.

$\rightsquigarrow$ In case one has SSC on W, one need to solve a min-vol TriNMF

In general $\min_{\substack{W\geq 0\\ H\geq 0}} D(X, WH)$ is solved with alternating method i.e. optimize over W with H fixed and viceversa. So from now on we focus into solving

$$\min_{H\geq 0} D(\overset{d\times n}{X}, \overset{d\times r\ \ r\times n}{WH}) \quad : \quad W, X \geq 0$$

Active Set

Both for KL divergence and $\|\cdot\|_F^2$, the $D(\cdot,\cdot)$ can be decomposed as

$$D(X; WH) = \sum_i D(x_i, Wh_i)$$

so we can analyze separately $\min_{h\geq 0} D(x, Wh)$. The KKT conditions

$$h \geq 0 \ , \ \nabla_h D(x, Wh) \geq 0 \ , \ h \odot \nabla_h D(x, Wh) = 0$$

in this case are Necessary and Sufficient for global minimality. We can thus call $I(h) = \{i : h_i > 0\}$ and notice that for the solution h^* we need

$$\nabla_h D(x, Wh^*)\big|_{I(h^*)} = 0$$

- In case of $\|\cdot\|_F^2$, we get $[W^T(Wh-x)]_I = 0$, i.e.,

$$W(:,I)^T W(:,I) h(I) = W(:,I)^T x$$

- In case of KL, we get $[W^T[(Wh-x) \oslash Wh]]_I = 0$ i.e.

$$W(:,I)^T e = W(:,I)^T \operatorname{diag}(x) \cdot \frac{1}{W(:,I)h(I)}$$

$\rightsquigarrow$ In both cases, $h(I)$ and thus h^* is solvable as long as we know $I(h^*)$. So we can try to find I, that are finite but in theory exponential in r, so they tend to be slow even if accurate. The usual methods tend to update I to find the optimal one while looking for indices to add/remove in order to lower the error as much as possible.

MU

One of the, if not the, first algorithm proposed for NMF. Recall the KKT conditions for $\|\cdot\|_F^2$

$$\nabla_H \|X - WH\|_F^2 = 2W^T(WH - X) \geq 0 \ , \quad H \odot (W^T(WH-X)) = 0$$

so if $(\nabla_H)_{i,j} > 0$ and $H_{i,j} > 0$, then we have to decrease $H_{i,j}$ and we know that $(\nabla_H)_{i,j} > 0 \Rightarrow (W^TWH)_{ij} > W^TX$

$\Rightarrow (W^TX)_{ij} / (W^TWH)_{ij} < 1$. Viceversa, $(\nabla_H)_{ij} < 0$ then we have to increase H_{ij} and $(\nabla_H)_{ij} < 0 \Rightarrow (W^TX)_{ij} / (W^TWH)_{ij} > 1$, so

$$H \rightsquigarrow H \circ \frac{W^TX}{W^TWH} \quad \text{and} \quad W \rightsquigarrow W \circ \frac{XH^T}{WHH^T}$$

MU

Given $X \in \mathbb{R}_+^{d\times n}$ and $r > 0$ initialize $W, H > 0$
Repeat until convergence

$$\Big| \ H \rightsquigarrow H \circ \frac{W^TX}{W^TWH + \varepsilon \cdot E} \ , \quad W \rightsquigarrow W \circ \frac{XH^T}{WHH^T + \varepsilon E}$$

Here $\varepsilon > 0$ is a small enough constant to avoid dividing by zero, and E is the all-ones matrix.

There is also the version for the KL-div.

MU-KL

Given $X \in \mathbb{R}_+^{d\times n}$ and $r > 0$ initialize $W, H > 0$
Repeat until convergence

$$\Big| \ H \rightsquigarrow H \circ \frac{W^T X / WH}{W^TE + \varepsilon E} \ , \quad W \rightsquigarrow W \circ \frac{\frac{X}{WH} H^T}{EH^T + \varepsilon E}$$

Theorem 14.1 MU makes $D(X, WH)$ decrease at each step both for $\|\cdot\|_F^2$ and $D(\cdot\|\cdot)$

$\rightsquigarrow$ Notice that if W or H have zero entries at some point, then they will continue having zeros forever on those entries. This says that MU generates sparse solutions, but at the same time, it tends to get stuck in local minima because it cannot get out of the zero entries, so it has convergence issues. $O(mnr)$ comp. cost for dense, or $O(r \cdot nnz(X))$ for sparse

Coordinate Gradient descend

Let's split the problem even more.

$$\min_{H\geq 0} \|X - WH\|_F^2 = \min_{H\geq 0} \sum_i \|x_i - Wh_i\|^2$$

$$= \min_{H\geq 0} \sum_i \| \underbrace{x_i - \overline{W}\,\overline{h}_i}_{v} - \underbrace{h_{j,i}}_{\alpha} \underbrace{w_j}_{z} \|^2$$

where $\overline{W}$ is the submatrix of W where we removed column j and in $\overline{h}_i$ we removed entry j. If we now fix everything except for $h_{j,i}$, we see that it becomes an easy problem of the kind

$$\arg\min_{\alpha\geq 0} \|v - \alpha z\|^2 = \arg\min_{\alpha\geq 0} \alpha^2 \|z\|^2 - 2\alpha \langle v, z\rangle$$

$$= \max\left\{0, \frac{\langle v, z\rangle}{\|z\|^2}\right\}$$

$$\text{so } h_{j,i}^* = \max\left\{0, \frac{w_j^T x_i - w_j^T \overline{W}\,\overline{h}_i}{\|w_j\|^2}\right\}$$

$$\rightsquigarrow H(j,:)^* = \max\left\{0, \frac{1}{\|w_j\|^2}\left(w_j^T X - w_j^T \overline{W}\,\overline{H}\right)\right\}$$

HALS

Given $X \in \mathbb{R}_+^{d\times n}$ $r > 0$. Initialize $W, H \geq 0$
Repeat until convergence
  Compute XH^T and HH^T
  for $k = 1, \dots, r$

$$W(:,k) = \max\left\{0, \frac{(XH^T)(:,k) - \sum_{\ell\neq r} W(:,\ell) \cdot (HH^T)(\ell,k)}{(HH^T)_{k,k}}\right\}$$

  Compute W^TX and W^TW
  For $k = 1, \dots, r$

$$H(k,:) = \max\left\{0, \frac{(W^TX)(k,:) - \sum_{\ell\neq r} H(\ell,:)(W^TW)(k,\ell)}{(W^TW)_{k,k}}\right\}$$

One can accelerate the method by repeating the internal loop of update of W, H multiple time, since the computations of XH^T, HH^T, W^TX, W^TW are the actual bottleneck. This is also highly parallelizable. $O(mnr) / O(r\, nnz(X))$

$\rightsquigarrow$ This is an effective Alternating Method, that solves perfectly each internal convex problem, so it is convergent by theorem ...

It is moreover one of the most performing algorithms known nowadays and it guarantees the error to decrease at each step.

There are many other : Fast Pros. Gradient Method, or ADMM on

$$\min_{\substack{h\in\mathbb{R}^r\\ y\in\mathbb{R}_+^r}} \|x - Wh\|^2 \ ; \quad y = h$$

or ADMM directly on $\min_{\substack{W,H\\ U,V}} \|X - WH\|, \ W = U \geq 0, \ H = V \geq 0$

but all of these do not guarantee the error to decrease at each step