Non-negative Matrix Factorization (Paatero & Tapper, Lee & Seung)

Setting: Given N observations of P non-negative examples in data matrix

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_N^T \end{bmatrix}$$

Note: $x_{ij} \geq 0$ is motivated by experimental data

- Biology/Genetics: cell counts, protein expression
- Imaging: pixel intensity $0 \mapsto 1$ (white $\to$ black)
  same for RGB in color images
- Topic Modeling: count or frequency of terms appearing in a document

Goal: Find a matrix $\hat{X} \in \mathbb{R}^{N \times P}$ with $\text{rank}(\hat{X}) \ll \text{rank}(\hat{X})$ that minimizes the loss of information contained in X.

Question 1: How do we measure information loss?

1) Froebinus norm $\quad \|X - \hat{X}\|_F = \sqrt{\sum_{n=1}^{N} \sum_{j=1}^{P} |X_{nj} - \hat{X}_{nj}|^2}$

2) Divergence $\quad D(X \| \hat{X}) = \sum_{n=1}^{N} \sum_{j=1}^{P} \left[X_{ij} \log \frac{X_{ij}}{\hat{X}_{ij}} + \hat{X}_{ij} - X_{ij} \right]$

   i) In the case $\sum \sum X = \sum \sum \hat{X} = 1$, $D(X \| \hat{X})$ is called the Kullback-Liebler Divergence, or relative entropy of X from $\hat{X}$.

   ii) $D(X \| \hat{X}) = -\underbrace{\sum \sum \left(X_{ij} \log(\hat{X}_{ij}) - \hat{X}_{ij} \right)}_{\substack{\text{LOG-LIKELIHOOD OF} \\ X_{ij} \sim \text{Poisson}(\hat{X}_{ij})}} + \text{"}X_{ij}\text{ terms"}$

THM: LET X BE A DATA MATRIX W/ SVD

$$X = \underbrace{\left[\vec{u}_1, \ldots, \vec{u}_N\right]}_{U} \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \\ & 0 & \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} \vec{v}_1^T \\ \vdots \\ \vec{v}_p^T \end{bmatrix}}_{V^T}$$

FIX $1 \leq k < p$ AND SET

$$\hat{X}_k = U \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_k & & \\ & 0 & & 0 & \\ & & & & \ddots \; 0 \\ & & 0 & & \end{bmatrix} V^T$$

THEN

1) $\hat{X}_k$ IS RANK k AND

2) $\hat{X}_k = \underset{\substack{Y \in \mathbb{R}^{N \times p} \\ \text{RANK}(Y) = k}}{\operatorname{argmin}} \|X - Y\|_F$

NOTE:

1) THE ABOVE THM. SAYS THAT PCA GIVES THE ANSWER WHEN WE USE $\|\cdot\|_F$.

2) THIS HAS AN INTERPRETABILITY PROBLEM! RECALL

$$\hat{X}_k = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_N^T \end{bmatrix} \left[\vec{v}_1 \; \ldots \; \vec{v}_k\right] \begin{bmatrix} \vec{v}_1^T \\ \vdots \\ \vec{v}_k^T \end{bmatrix}$$

$\vec{v}_1 \ldots \vec{v}_k$ ARE FIRST k PRIN COMP LOADINGS

$$\Rightarrow \vec{x}_i^{(k)T} = \sum_{s=1}^{k} \underbrace{\vec{x}_i^T \vec{v}_s}_{y_{is}} \vec{v}_s^T$$

$$\Rightarrow x_{ij} \approx \sum_{s=1}^{k} y_{is} v_{sj}$$

RECALL, $x_{ij} \geq 0$ IN THIS SETTING BUT

PCA MAKES NO SUCH CONSTRAINT ON SCORES OR LOADINGS!

$$\vec{x}_i \approx \sum_{s=1}^{k} \underbrace{y_{is}}_{\text{WEIGHTS}} \underbrace{\vec{v}_s}_{\text{FEATURES / COMPONENTS}}$$

FROM THIS VIEW

1) EACH $\vec{x}_i$ IS BUILT FROM THE SAME LIBRARY OF FEATURES $\vec{v}_1 \ldots \vec{v}_k$

2) IF $y_{is} < 0$ WE ARE 'SUBTRACTING' FEATURE $\vec{v}_s$

PHYSICAL INTERPRETATION?

3) SUPPOSE $p = 2$

$$x_{ij} = \sum_{s=1}^{2} y_{is} v_{sj}$$

- $y_{i1} v_{1j} \approx 0 \approx y_{i2} v_{2j} \Rightarrow x_{ij} \approx 0$
- $y_{i1} w_{1j} \gg 0$, $y_{i2} w_{2j} \ll 0 \Rightarrow x_{ij} \approx 0$

HOW DO WE DIFFERENTIATE BETWEEN

a) ABSENCE OF FEATURE $w_{1j} = w_{2j} = 0$ AND b) CANCELLATION OF FEATURE $w_{1j} = -w_{2j}$?

$\Rightarrow$ FEATURES $\vec{w}_1, \ldots \vec{w}_k$ CANNOT BE EXPECTED TO HAVE A MEANINGFUL PHYSICAL INTERPRETATION AS "BUILDING BLOCKS" OF OUR DATA!

$\Rightarrow$ ENFORCE NON-NEGATIVITY OF ELEMENTS OF FACTORIZATION

$$X \approx \hat{X} = \cancel{U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ 0 & & \sigma_k \end{bmatrix} V^T = Y V^T} \quad \text{(PCA)}$$

$$= WH = \begin{bmatrix} w_1^T \\ \vdots \\ w_N^T \end{bmatrix} \begin{bmatrix} \vec{h}_1^T \\ \vdots \\ \vec{h}_k^T \end{bmatrix}$$

$w_1, \ldots, w_N \in \mathbb{R}^{N \times k}$ WEIGHTS

$\vec{h}_1, \ldots, \vec{h}_k \in \mathbb{R}^{p \times k}$ FEATURES

$$\boxed{w_{ij}, h_{ij} \geq 0}$$

SO THAT

$$\vec{x}_i \approx \sum_{s=1}^{k} w_{is} \vec{h}_s$$

IS A SUPERIMPOSITION OF FEATURES $\vec{h}_1, \ldots \vec{h}_s$.

GEOMETRIC INTERPRETATION

EX: $\vec{x}_1, \ldots \vec{x}_N \in \mathbb{R}^3$ AND THERE EXISTS A $W \in \mathbb{R}^{N\times 2}$; $H \in \mathbb{R}^{2\times 3}$

S.T. $X = WH$

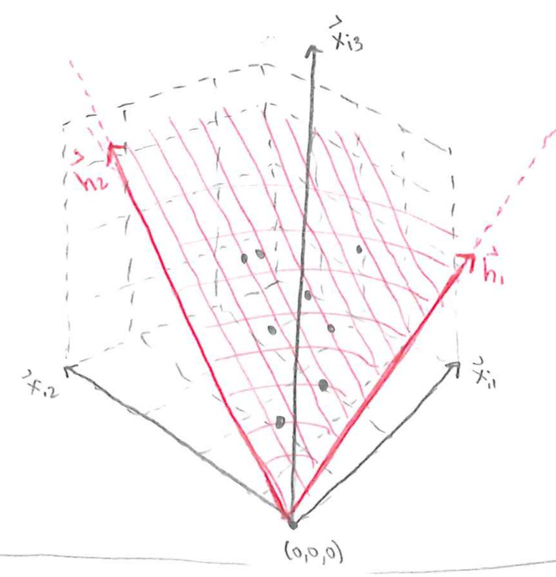
1) NON-NEGATIVITY OF x_{ij} IMPLIES $\vec{x}_1, \ldots \vec{x}_N$ LIE IN THE POSITIVE ORTHANT OF $\mathbb{R}^3$

$$P_3 = \{ \vec{x} \in \mathbb{R}^3 \ , \ x_j \geq 0, j = 1, \ldots, 3 \}$$

2) EQUALITY $WH = X$ IMPLIES

$$\vec{x}_i = \sum_{s=1}^{2} w_{is} \vec{h}_s \in \text{SPAN}\{\vec{h}_1, \vec{h}_2\}$$

i.e. DATA RESIDE ON A 2-D SUBSPACE



DEF: THE SIMPLICIAL CONE GENERATED BY THE SET OF VECTORS $\{\vec{h}_1, \ldots \vec{h}_s\}$ IN $\mathbb{R}^p$ IS THE SET

$$\Gamma(\{\vec{h}_1, \ldots \vec{h}_s\}) = \left\{ \vec{v} \in \mathbb{R}^p \mid \vec{v} = \sum_{j=1}^{s} a_j \vec{h}_j \quad a_1, \ldots a_s \geq 0 \right\}$$

ALSO CALLED THE POSITIVE SPAN OF $\{\vec{h}_1, \ldots \vec{h}_s\}$

CONICAL HULL OF $\{\vec{h}_1, \ldots \vec{h}_s\}$

Note: WE CAN CHOOSE DIFFERENT STEP SIZES FOR EACH COORDINATE

for $(j = 1, \ldots p)$ {

$$x_j^{(i+1)} = x_j^{(i)} - \Delta_{x_j} \nabla f \left(\begin{bmatrix} x_1^{(i+1)} \\ \vdots \\ x_{j-1}^{(i+1)} \\ x_j^{(i)} \\ \vdots \\ x_p^{(i)} \end{bmatrix} \right)$$

}

AND ITERATE $i = 1 \ldots$ UNTIL CONVERGENCE

Ex: NMF & DIVERGENCE

$$\text{ERROR} = D(X \| WH) = \sum_{s=1}^{N} \sum_{t=1}^{p} \left[X_{st} \log \frac{X_{st}}{(WH)_{st}} + (WH)_{st} - X_{st} \right]$$

ASSUME WE HAVE STEP SIZES $\Delta^W_{s\ell}, \Delta^H_{\ell t}$ $\quad 1 \le s \le N$, $1 \le \ell \le k$, $1 \le t \le p$

1) FIND $\frac{\partial}{\partial W_{ij}} D(X \| WH)$

DOES NOT DEPEND ON W OR H!

$$\frac{\partial}{\partial W_{ij}} D(X\|WH) = \frac{\partial}{\partial W_{ij}} \left(\sum_{st} \left[-X_{st} \log (WH)_{st} + (WH)_{st} \right] + \cancel{\sum_{st} \left[X_{st} \log X_{st} - X_{st} \right]} \right)$$

$$(WH)_{st} = \sum_{\ell=1}^{k} W_{s\ell} H_{\ell t}$$

$$= \sum_{st} \left[-X_{st} \frac{\partial}{\partial W_{ij}} \log (WH)_{st} + \frac{\partial}{\partial W_{ij}} (WH)_{st} \right]$$

$$= \sum_{t} \left[\left(-X_{it} \frac{1}{(WH)_{it}} + 1 \right) \cdot \frac{\partial}{\partial W_{ij}} (WH)_{it} \right]$$

$$= -\sum_{t} \left[\left(\frac{X_{it}}{(WH)_{it}} - 1 \right) H_{jt} \right]$$

2) FIND $\frac{\partial}{\partial H_{ij}} D(X \| WH)$

$$\frac{\partial}{\partial H_{ij}} D(X\|WH) = \sum_{st} \left[-X_{st} \frac{\partial}{\partial H_{ij}} \log (WH)_{st} + \frac{\partial}{\partial H_{ij}} (WH)_{st} \right]$$

$$= \sum_{s} \left[\left(-X_{sj} \frac{1}{(WH)_{sj}} + 1 \right) \frac{\partial}{\partial H_{ij}} (WH)_{sj} \right]$$

$$= - \sum_{s} \left[\left(\frac{X_{sj}}{(WH)_{sj}} - 1 \right) W_{si} \right]$$

3) ITERATIVE GRADIENT DESCENT

$$W_{ij} \longleftarrow W_{ij} + \Delta_{ij}^{W} \left[\sum_t \frac{H_{jt} X_{it}}{(WH)_{it}} - \sum_t H_{jt} \right]$$

$$H_{j\ell} \longleftarrow H_{j\ell} + \Delta_{j\ell}^{H} \left[\sum_s \frac{W_{sj} X_{sj}}{(WH)_{sj}} - \sum_s W_{sj} \right]$$

$i = 1, \ldots N$
$j = 1, \ldots k$
$\ell = 1 \ldots p$

* NOTHING PROHIBITS NEGATIVE VALUES OF W_{ij} OR $H_{j\ell}$

4) LEE & SEUNG (2001)

1) SET $\Delta_{ij}^{W} = \frac{W_{ij}}{\sum_t H_{jt}}$, $\Delta_{j\ell}^{H} = \frac{H_{j\ell}}{\sum_s W_{sj}}$

$$W_{ij} \longleftarrow W_{ij} \left[\frac{\sum_t \left(H_{jt} X_{it} / (WH)_{it} \right)}{\sum_t H_{jt}} \right]$$

$$H_{j\ell} \longleftarrow H_{j\ell} \left[\frac{\sum_s \left(W_{sj} X_{s\ell} / (WH)_{s\ell} \right)}{\sum_s W_{sj}} \right]$$

<u>NOTE:</u>

1) MULTIPLICATIVE CONSTANTS ≥ 0 IMPLIES ITERATES REMAIN ≥ 0

2) IF $X = WH$ THEN MULTIPLICATIVE CONSTANTS $= 1$

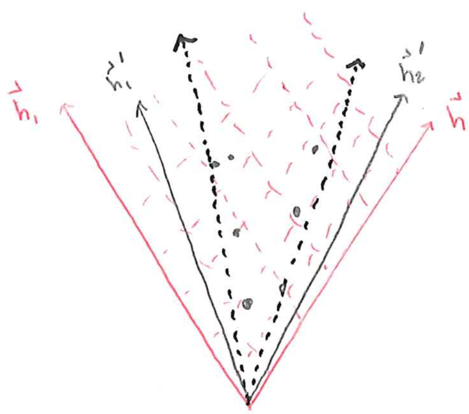
3) IMPLEMENTED IN PACKAGE NMF

IN THIS EXAMPLE $\vec{x}_1, \ldots \vec{x}_N$ LINE IN THE SIMPLICIAL CONE OF $\{\vec{h}_1, \vec{h}_2\}$

NOTE: $\vec{x}_1, \ldots \vec{x}_N$ ALWAYS RESIDE IN $\Gamma(\{\vec{x}_1, \ldots \vec{x}_N\})$, THE CONICAL HULL $\{\vec{x}_1, \ldots \vec{x}_N\}$, TRIVIALLY

QUESTION: CAN WE FIND TWO VECTORS $\vec{h}_1' \neq \vec{h}_1$, $\vec{h}_2' \neq \vec{h}_2$ SUCH THAT

$$\{\vec{x}_1, \ldots \vec{x}_N\} \subseteq \Gamma(\{\vec{h}_1', \vec{h}_2'\}) \text{ ?}$$



TRIVIAL CHOICES

- $\vec{h}_1 \rightleftarrows \vec{h}_2$, $w_{i1} \rightleftarrows w_{i2}$

  SWAP FEATURES & WEIGHTS

- $\vec{h}_1 \mapsto c\vec{h}_1$, $w_{i1} \mapsto \frac{1}{c} w_{i1}$

  $c>0$

  RESCALE FEATURES/WEIGHTS

- EXCLUDE THESE CASES

YES! INFINITE NUMBER OF CHOICES

- NEW FEATURES $\vec{h}_1', \vec{h}_2'$ $\Longrightarrow$ NEW WEIGHTS w_{i1}' w_{i2}'

  $\Longrightarrow$ NEW MATRICES W', H'

$$X = WH = W'H'$$

- EVEN WHEN THERE IS A LOWER-DIMENSIONAL REPRESENTATION OF X, NO GUARANTEE THAT THE SOLUTION $X=WH$ IS UNIQUE!

QUESTION: IGNORING TRIVIAL CHOICES, IF $X=WH$, WHEN ARE W, H UNIQUE?

ANSWER: 1) CAN PLACE DATA ON THE BOUNDARY OF P!

IN THE PICTURE ABOVE, NO GAP BETWEEN $\vec{h}_1, \vec{h}_2$ AND DATA!

2) OTHER CASES (SEE DONOHO & STODDEN, LAURBERG ET AL.) BUT DIFFICULT TO VERIFY W/ DATA

Estimating a NMF

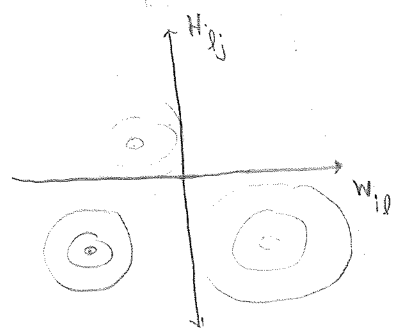
Given $X \in \mathbb{R}^{N \times p}$ and $k < \min\{N, p\}$ we want to minimize

$$\|X - WH\|_F \quad \text{or} \quad D(X \| WH) \quad \text{or general error } E(X, WH)$$

for $W \in \mathbb{R}^{N \times k}$ $H \in \mathbb{R}^{k \times p}$

1) No closed form expression for W or H in general

2) Non-convex $\longrightarrow$ many local minima



* Difficult to find <u>the</u> global minimizer !

$\Rightarrow$ use gradient descent (or variants) but try <u>many</u> random initial conditions then pick best

Gradient Descent

Goal: minimize $f(\vec{x})$ $\vec{x} \in \mathbb{R}^p$.

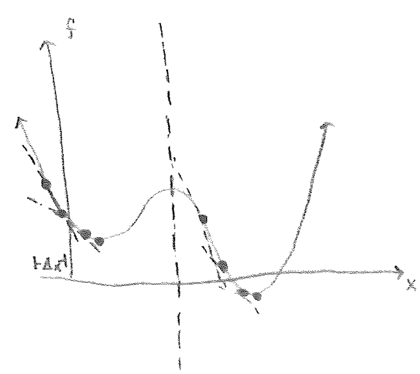
i) Pick small step size Δx

ii) Find $\nabla f = \begin{bmatrix} \partial_{x_1} f \\ \vdots \\ \partial_{x_p} f \end{bmatrix}$

iii) Iterate

$$\vec{x}_{i+1} = \underbrace{\vec{x}_i - \Delta_x \nabla f(\vec{x}_i)}_{\text{"take a step downhill"}}$$

until sequence converges !



Ex: Handwritten 4s

* See markdown document on Canvas > Files > Numerical Examples
* Note differences from PCA, output and # of data points

NMF in practice

0) (Optional) Rescale data (don't standardize which gives <0 entries)

1) If prior knowledge suggests a specific rank k, use it. Otherwise, choose a suitable range of ranks.

2) For each rank, generate multiple fits w/ different initialization of the algorithm

   i) Choose best at each rank

   ii) Compare residuals at different ranks and make choice of k based on results

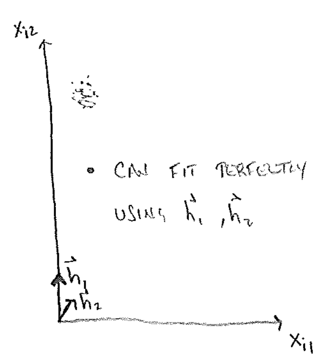
   * Balance overfitting w/ low rank

3) Can also consider

   - Different choices of error $\|\cdot\|_F$, $D(X\|WH)$ or penalties to impose sparsity in WH (Hoyer 2004)
   - Fixing/initializing W or H based on prior knowledge

Archetype Analysis

Default NMF may not give features H which are representative of data



- Can fit perfectly using $\vec{h}_1, \vec{h}_2$
- $\vec{h}_1, \vec{h}_2$ look nothing like the data!

* REQUIRE $\vec{h}_1, \ldots \vec{h}_k$ TO BE <u>CONVEX COMBINATIONS</u> OF THE DATA

$$\vec{h}_k^T = \sum_{n=1}^{N} \pi_{kn} \vec{x}_n^T \qquad \pi_{kn} \geq 0 \quad \sum_n \pi_{kn} = 1$$

$$\Rightarrow \quad H = \begin{bmatrix} \vec{h}_1^T \\ \vdots \\ \vec{h}_k^T \end{bmatrix} = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1n} \\ & & \\ \pi_{k1} & \cdots & \pi_{kn} \end{bmatrix} X$$

<u>Ex:</u> CONSIDER THE DATA BELOW. IDENTIFY THE SMALLEST REGION CONTAINING <u>ALL</u> POSSIBLE CONVEX COMBOS OF THE DATA POINTS, THE CONVEX HULL OF THE DATA!

i.e. THROW A LOOP AROUND POINTS AND PULL IT TIGHT!

<u>ARCHETYPAL ANALYSIS</u>

- MINIMIZE $\|X - WBX\|_F$ SUBJECT TO CONSTRAINTS

$X \in \mathbb{R}^{N \times p}$ i) $W_{ij} \geq 0$

$W \in \mathbb{R}^{N \times k}$ ii) $B_{ij} \geq 0$

$B \in \mathbb{R}^{k \times N}$ iii) $B\mathbb{1} = \mathbb{1}$, $\mathbb{1}$ = VECTOR OF ALL ONES

- THE ROWS OF BX ARE CALLED ARCHETYPES!

<u>NOTE:</u>
- IN NMF, WE REQUIRE $k \leq \min\{N, p\}$
- IN ARCHETYPE ANALYSIS, WE CAN LET $k > p$ TO SEARCH FOR k CLUSTERS/ARCHETYPES/PROTOTYPES IN OUR DATA.
  * RELATED TO k-MEANS CLUSTERING!

THE NUMERICAL NMF EXAMPLE CONTAINS AN ANALYSIS USING ARCH. ANALYSIS!

DIMENSION REDUCTION : NMF v. ARCH. ANALYSIS

GOALS OF TWO METHODS ARE SIMILAR BUT DIFFERENT FOCUS

1) IN NMF, LOW DIMENSIONAL REPRESENTATION USUALLY FROM WEIGHTS IN W

i.e. $X \in \mathbb{R}^{N \times p}$, $W \in \mathbb{R}^{N \times k}$, $H \in \mathbb{R}^{k \times p}$ $\boxed{k < \min\{N, p\}}$

REDUCE FROM p DIMENSIONS TO k DIMENSIONS "$\mathbb{R}^{N \times p} \longmapsto \mathbb{R}^{N \times k}$"

2) IN ARCH. ANALYSIS, ARCHETYPE ARE OF GREATER EMPHASIS, CAN HAVE MORE THAN p!

i.e. $X \in \mathbb{R}^{N \times p}$, $W \in \mathbb{R}^{N \times k}$ $B \in \mathbb{R}^{k \times N}$ $k \ll N$

REDUCE FROM p DIMENSIONS TO k LABELS OR ARCHETYPES

"$\mathbb{R}^{N \times p} \to \{1, .. k\}^N$"

- $\vec{x}_i$ IS LIKE ARCHETYPE j !
- CLUSTER POINTS BASED ON MOST SIMILAR ($w_{ij} \approx 1$) ARCHETYPE !

$\vec{x}_1, \ldots \vec{x}_N$

$\downarrow$

ARCHETYPES

$$\begin{bmatrix} \vec{y}_1^T \\ \vdots \\ \vec{y}_k^T \end{bmatrix} = BX$$

"$\mathbb{R}^{N \times p} \longmapsto \mathbb{R}^{k \times p}$"

* IN PSET 4 TAKE NOTE OF LOCATION OF ARCHETYPES FOR INCREASING k. CONSIDER CONNECTIONS W/ CONVEX HULLS AND POSITIVE SPANS !