Learning with Nonnegative Matrix Factorization

Nicolas Gillis[⋆], Kejun Huang[†], and Xiao Fu[‡]

[⋆]Department of Math & OR, Université de Mons
[†]Department of CISE, University of Florida
[‡]School of EECS, Oregon State University

Mini-Tutorial at SIAM LA 2021, Virtual Conference

May 25, 2021

Nonnegative Matrix Factorization (NMF)

PART I – Application-oriented aspects

- Sec. 1: Applications of NMF
  - Data analysis: blind hyperspectral image unmixing, image feature extraction, topic modeling, audio source separation
  - Statistical machine learning: crowdsourcing, joint probability estimation, Hidden Markov models, and community detection.
- Sec. 2: NMF Algorithms
  - MU
  - Nonconvex Optimization
  - Non-Euclidean cost

PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

Nonnegative Matrix Factorization (NMF)

PART I – Application-oriented aspects

- Sec. 1: Applications of NMF
  - **Data analysis: blind hyperspectral image unmixing, image feature extraction, topic modeling, audio source separation**
  - Statistical machine learning: crowdsourcing, joint probability estimation, clustering, Hidden Markov models, and community detection.
- Sec. 2: NMF Algorithms
  - MU
  - Nonconvex Optimization
  - Non-Euclidean cost

PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

Nonnegative Matrix Factorization (NMF)

Given a matrix $X \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $W \in \mathbb{R}^{p \times r}$ and $H \in \mathbb{R}^{r \times n}$ such that

$$\min_{W \geq 0, H \geq 0} D(X||WH), \qquad \text{(NMF)}$$

e.g., $D(X||WH) = ||X - WH||_F^2 = \sum_{i,j}(X - WH)_{ij}^2$.

Nonnegative Matrix Factorization (NMF)

Given a matrix $X \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $W \in \mathbb{R}^{p \times r}$ and $H \in \mathbb{R}^{r \times n}$ such that

$$\min_{W \geq 0, H \geq 0} D(X||WH), \qquad \text{(NMF)}$$

e.g., $D(X||WH) = ||X - WH||_F^2 = \sum_{i,j}(X - WH)_{ij}^2$.

NMF is a linear dimensionality reduction technique :

$$\underbrace{M(:,i)}_{\geq 0} \approx \sum_{k=1}^{r} \underbrace{U(:,k)}_{\geq 0} \underbrace{V(k,i)}_{\geq 0} \qquad \text{for all } i.$$

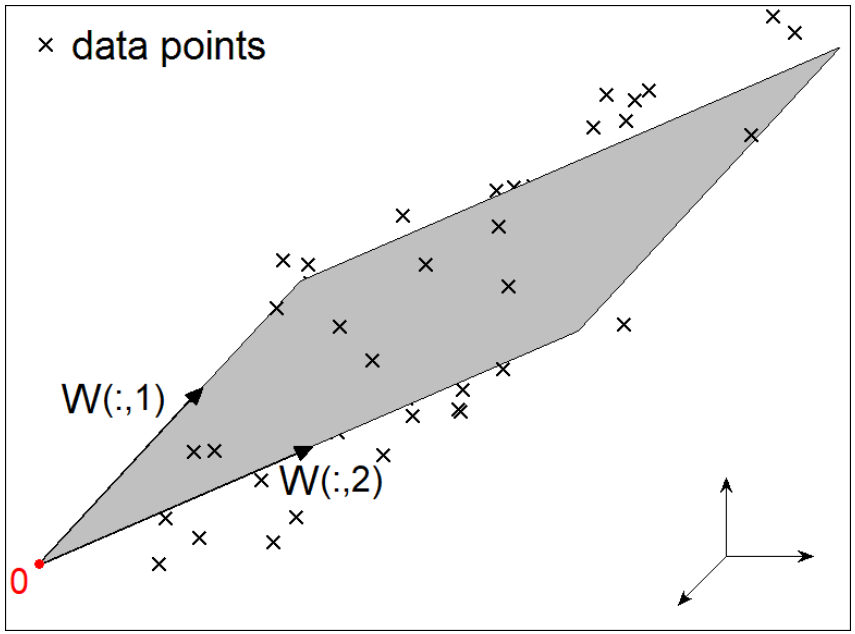
Nonnegative Matrix Factorization (NMF)

Given a matrix $X \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $W \in \mathbb{R}^{p \times r}$ and $H \in \mathbb{R}^{r \times n}$ such that

$$\min_{W \geq 0, H \geq 0} D(X||WH), \qquad \text{(NMF)}$$

e.g., $D(X||WH) = ||X - WH||_F^2 = \sum_{i,j}(X - WH)_{ij}^2$.

NMF is a linear dimensionality reduction technique :



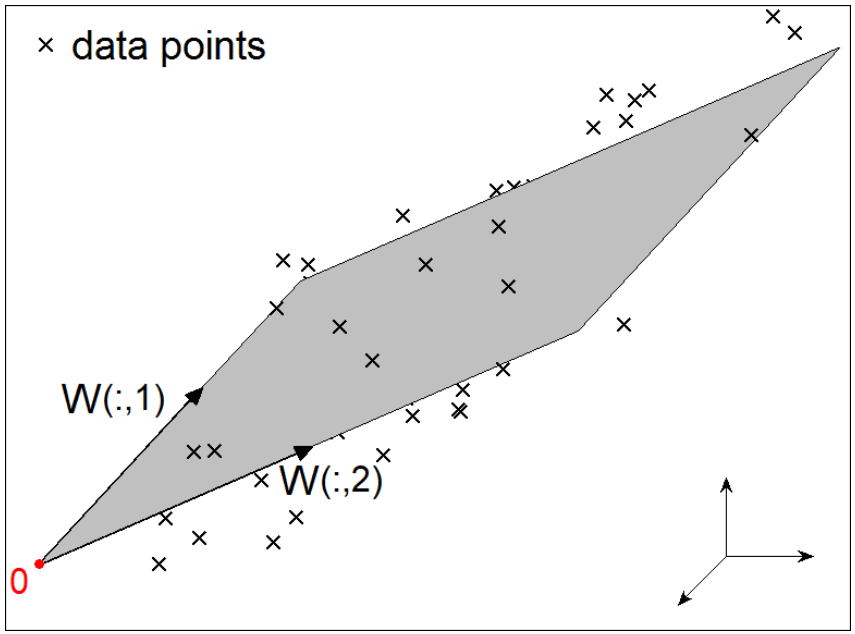
Nonnegative Matrix Factorization (NMF)

Given a matrix $X \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $W \in \mathbb{R}^{p \times r}$ and $H \in \mathbb{R}^{r \times n}$ such that

$$\min_{W \geq 0, H \geq 0} D(X||WH), \qquad \text{(NMF)}$$

e.g., $D(X||WH) = ||X - WH||_F^2 = \sum_{i,j}(X - WH)_{ij}^2$.

NMF is a linear dimensionality reduction technique :



Why nonnegativity?

Blind hyperspectral unmixing

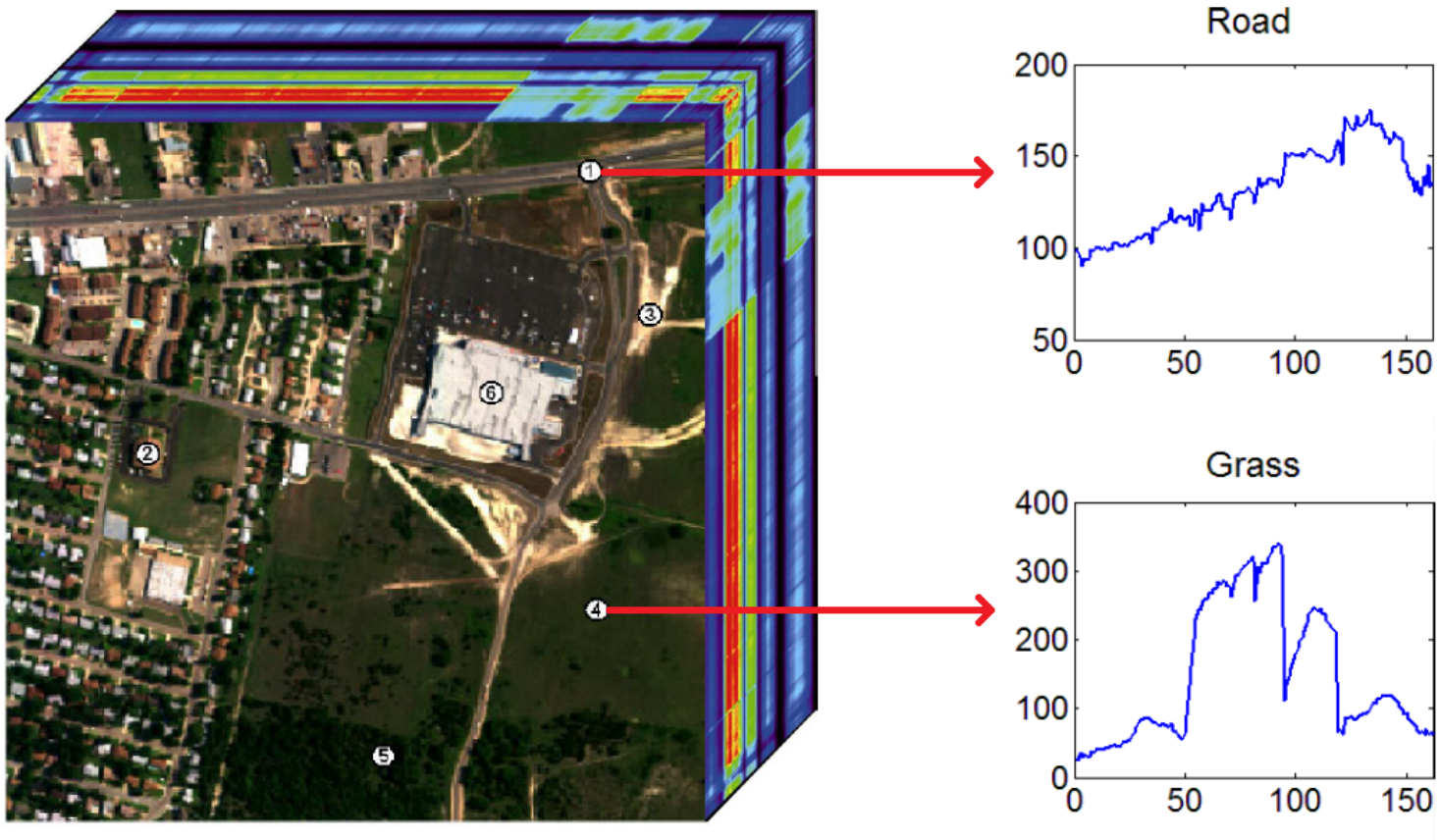
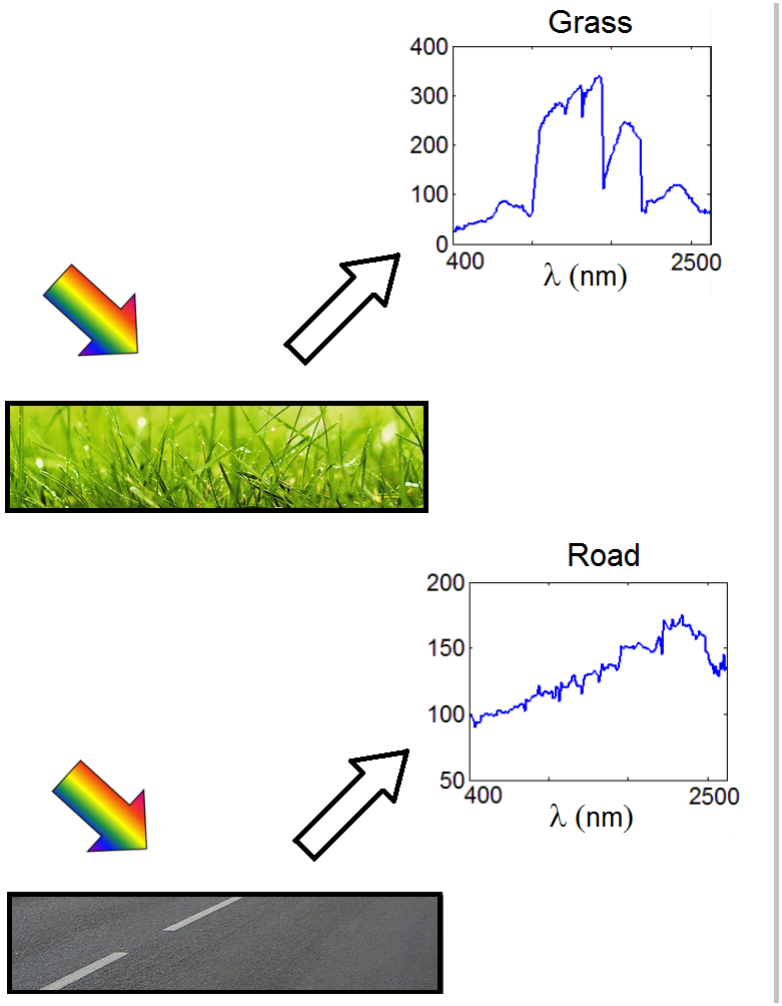
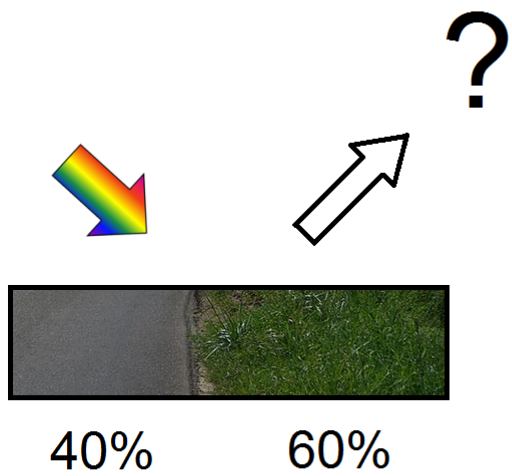


Figure 1: Urban hyperspectral image, 162 spectral bands and 307-by-307 pixels.

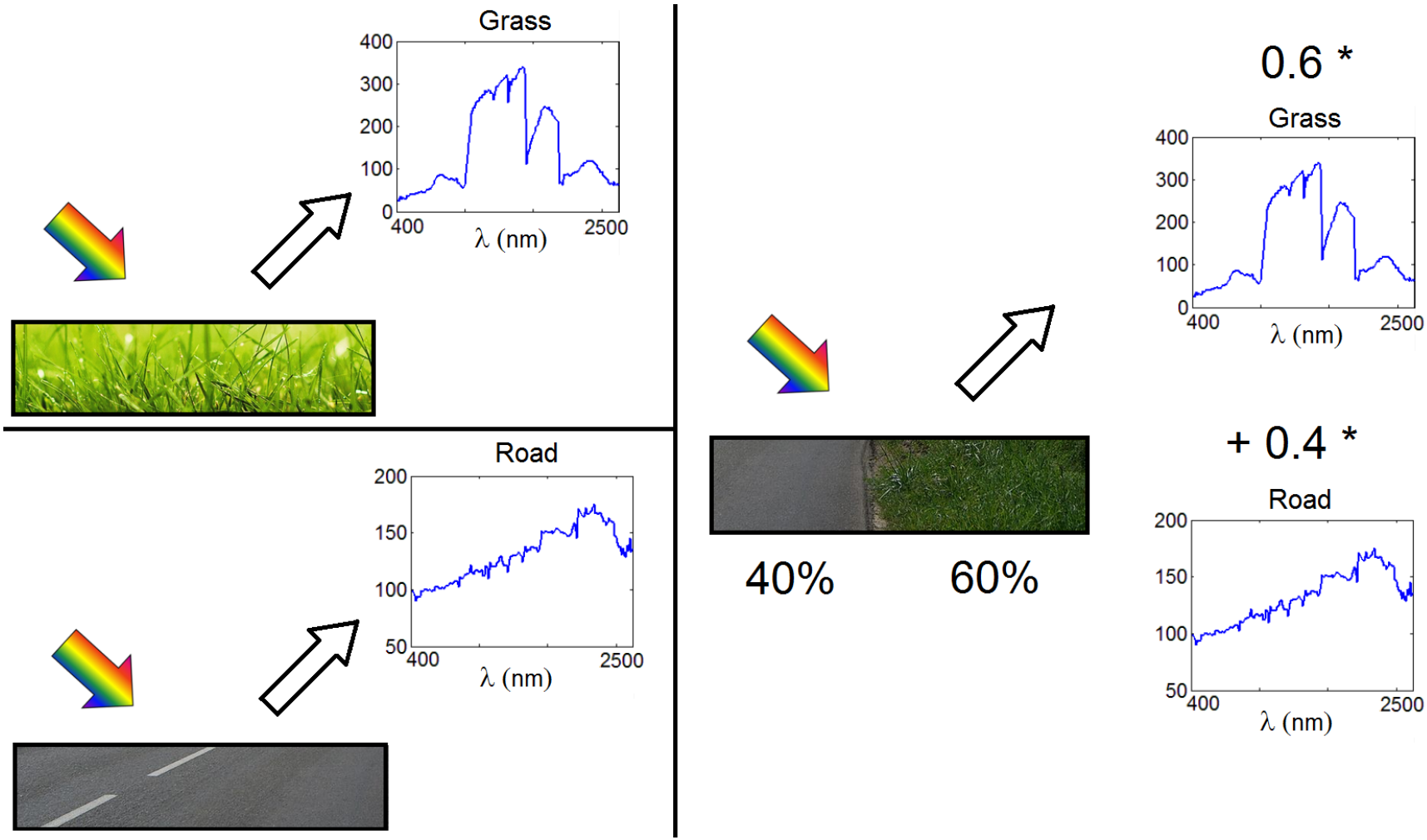
Problem. Identify the materials and classify the pixels.

Linear mixing model

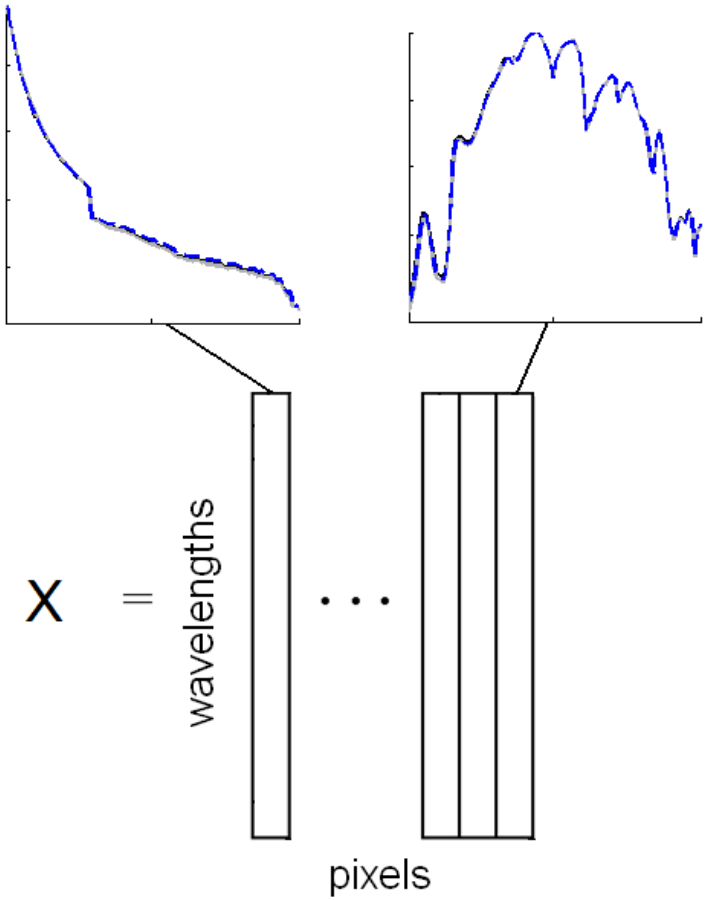




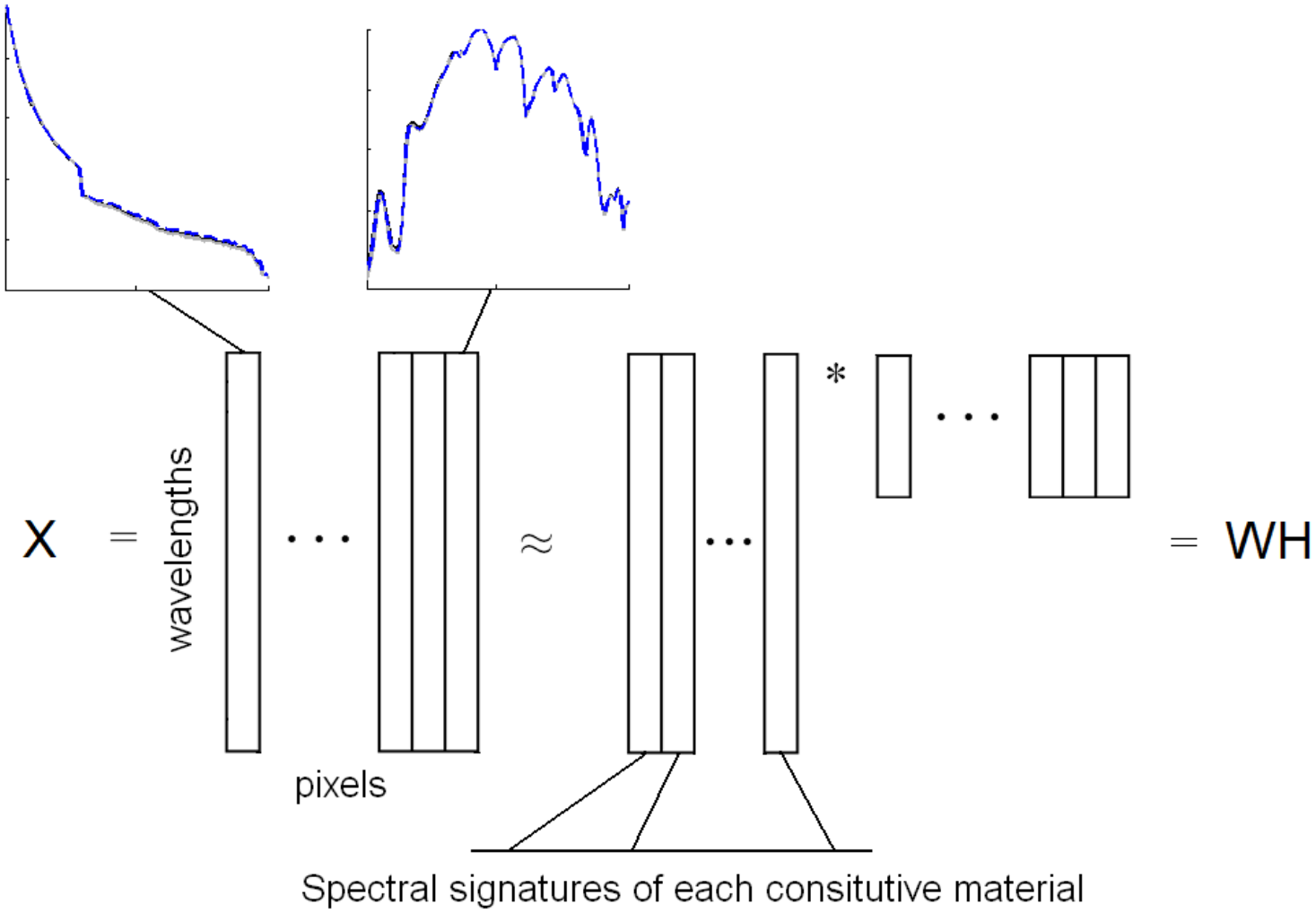
Linear mixing model



Blind hyperspectral unmixing with NMF

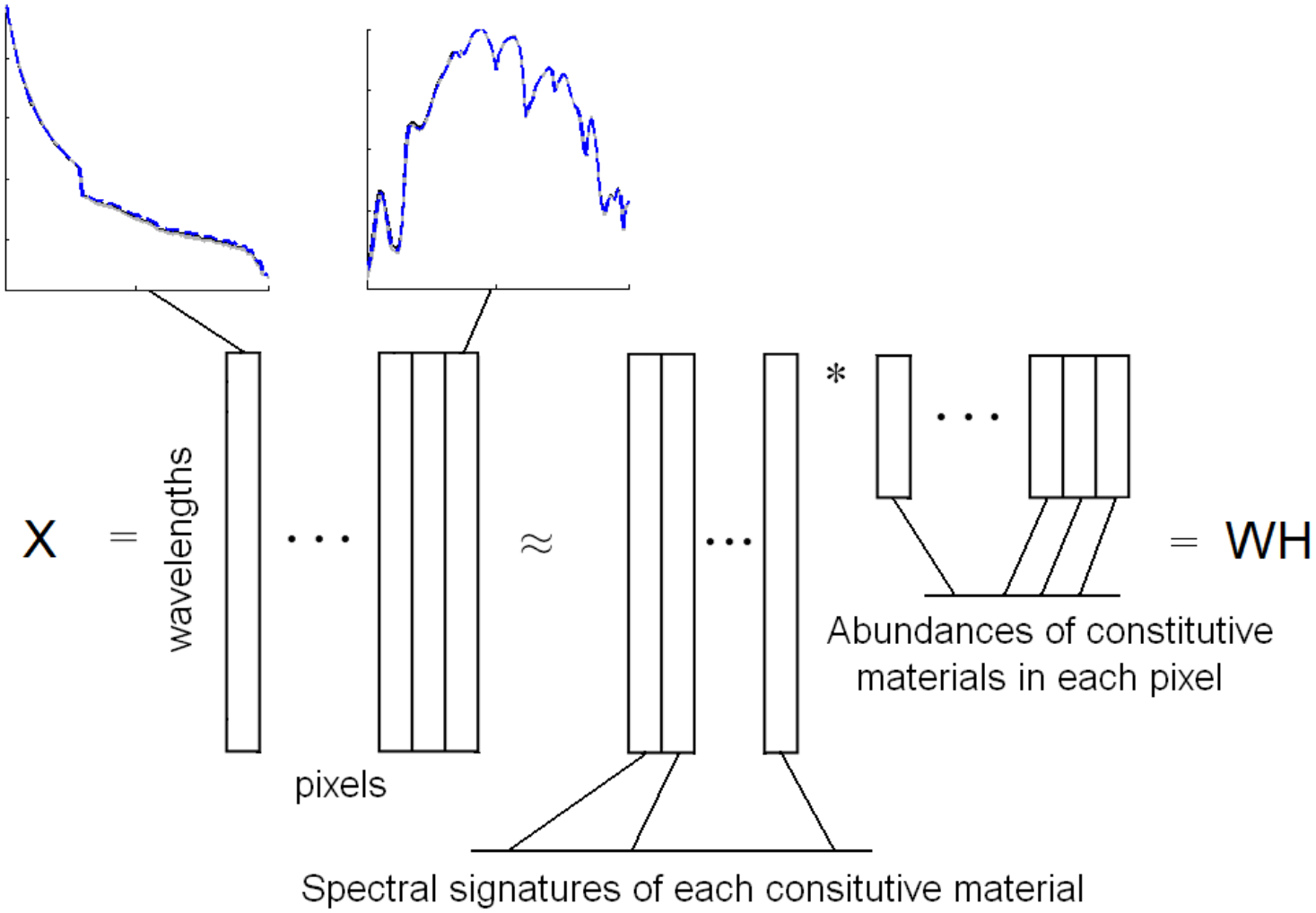


Blind hyperspectral unmixing with NMF



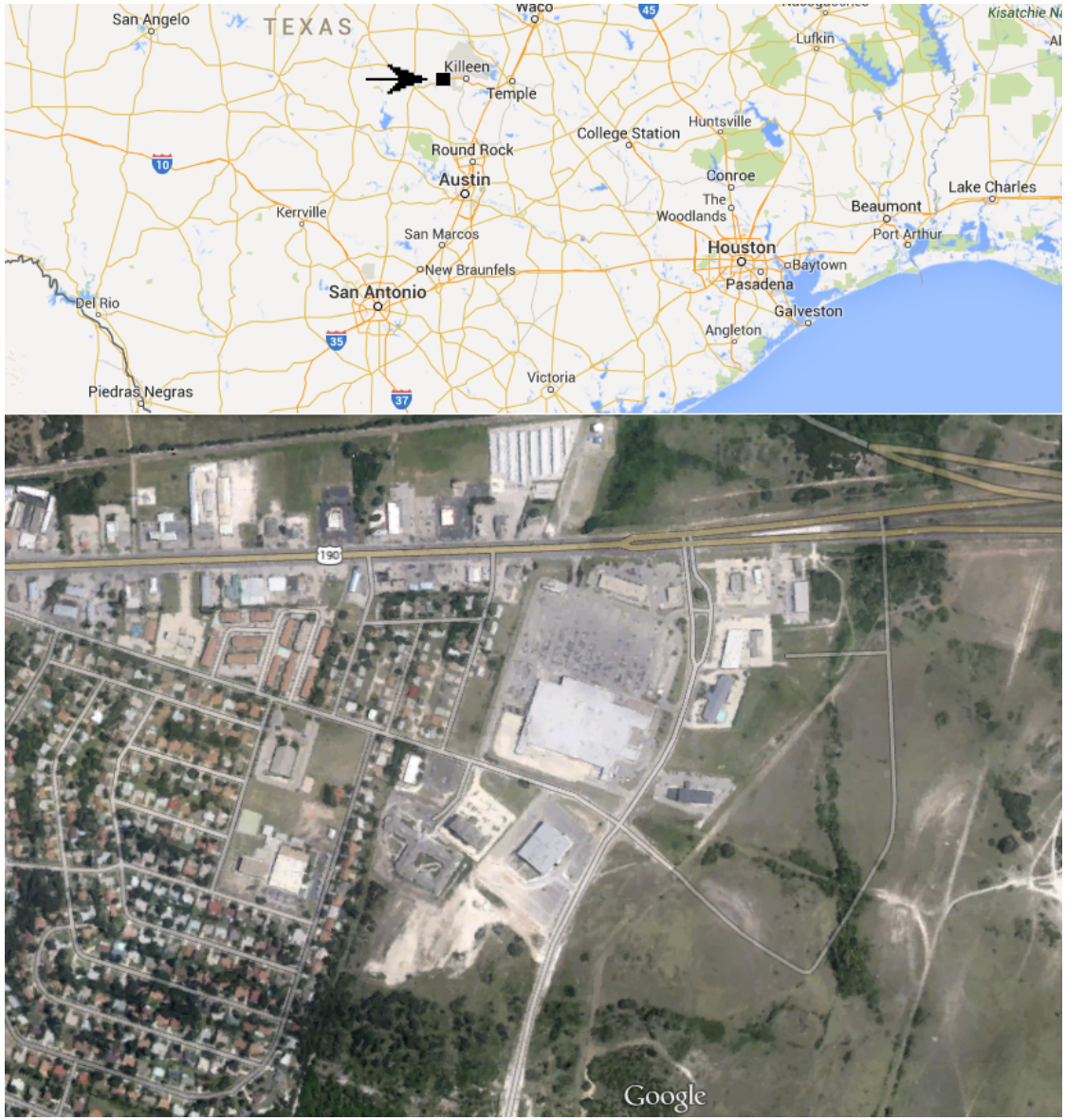
- Basis elements recover the different endmembers: $W \geq 0$;

Blind hyperspectral unmixing with NMF



- Basis elements recover the different endmembers: $W \geq 0$;
- Abundances of the endmembers in each pixel: $H \geq 0$.

Urban hyperspectral image



Urban hyperspectral image

$$\underbrace{\mathsf{X}(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathsf{W}(:,k)}_{} \underbrace{\mathsf{H}(k,j)}_{} .$$

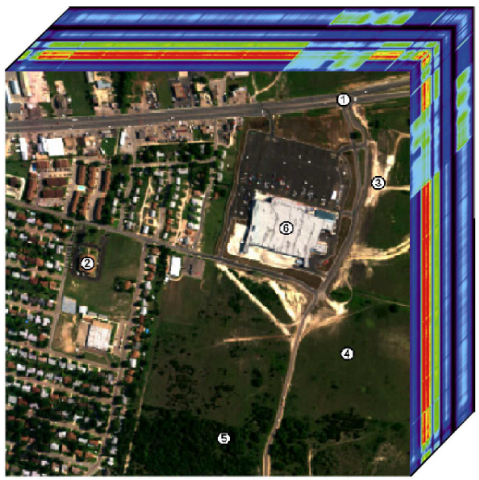
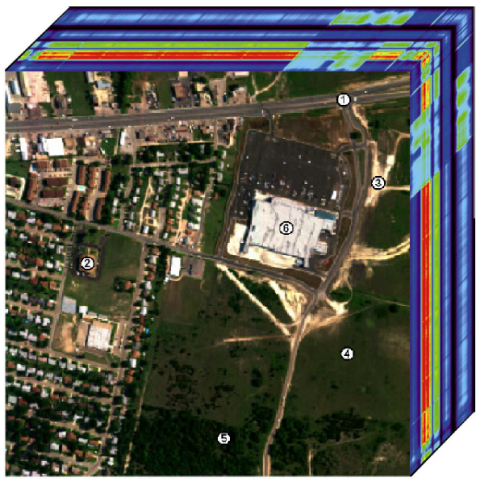


Figure 2: Decomposition of the Urban dataset.

Urban hyperspectral image

$$\underbrace{\mathsf{X}(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathsf{W}(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{\mathsf{H}(k,j)}_{\text{}} .$$



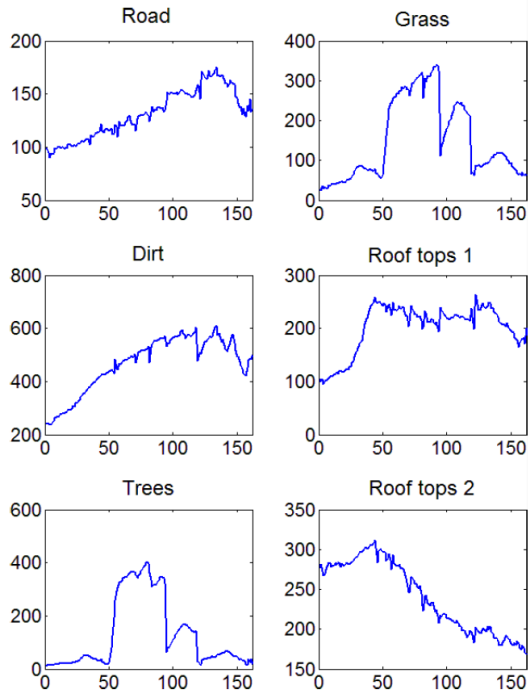
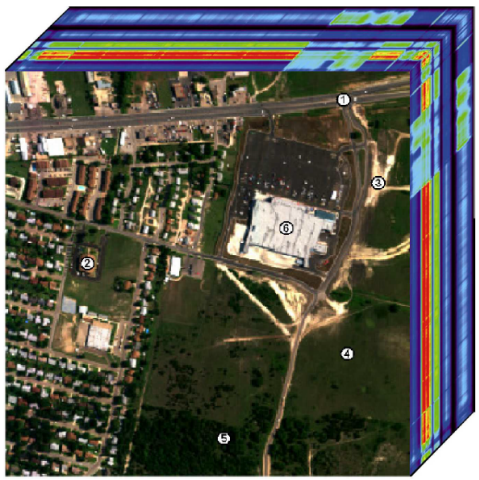
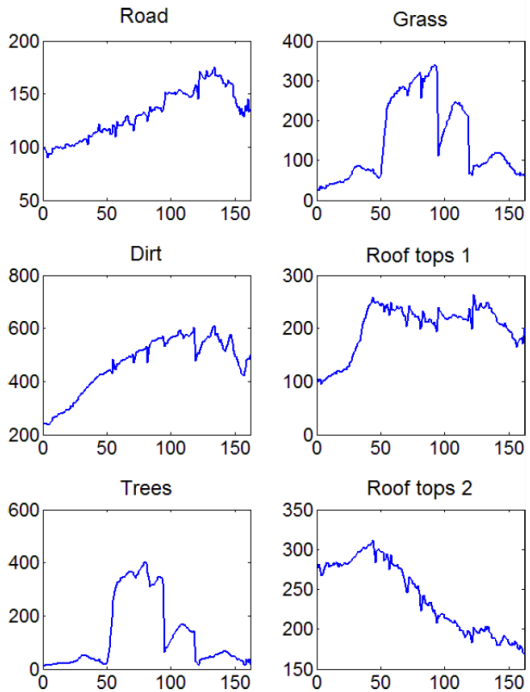


Figure 2: Decomposition of the Urban dataset.

Urban hyperspectral image

$$\underbrace{\mathsf{X}(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathsf{W}(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{\mathsf{H}(k,j)}_{\substack{\text{abundance of } k\text{th endmember}\\ \text{in } j\text{th pixel}}} .$$





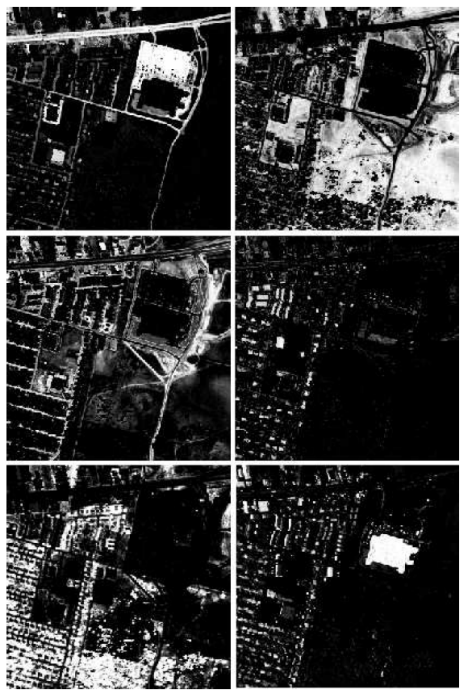
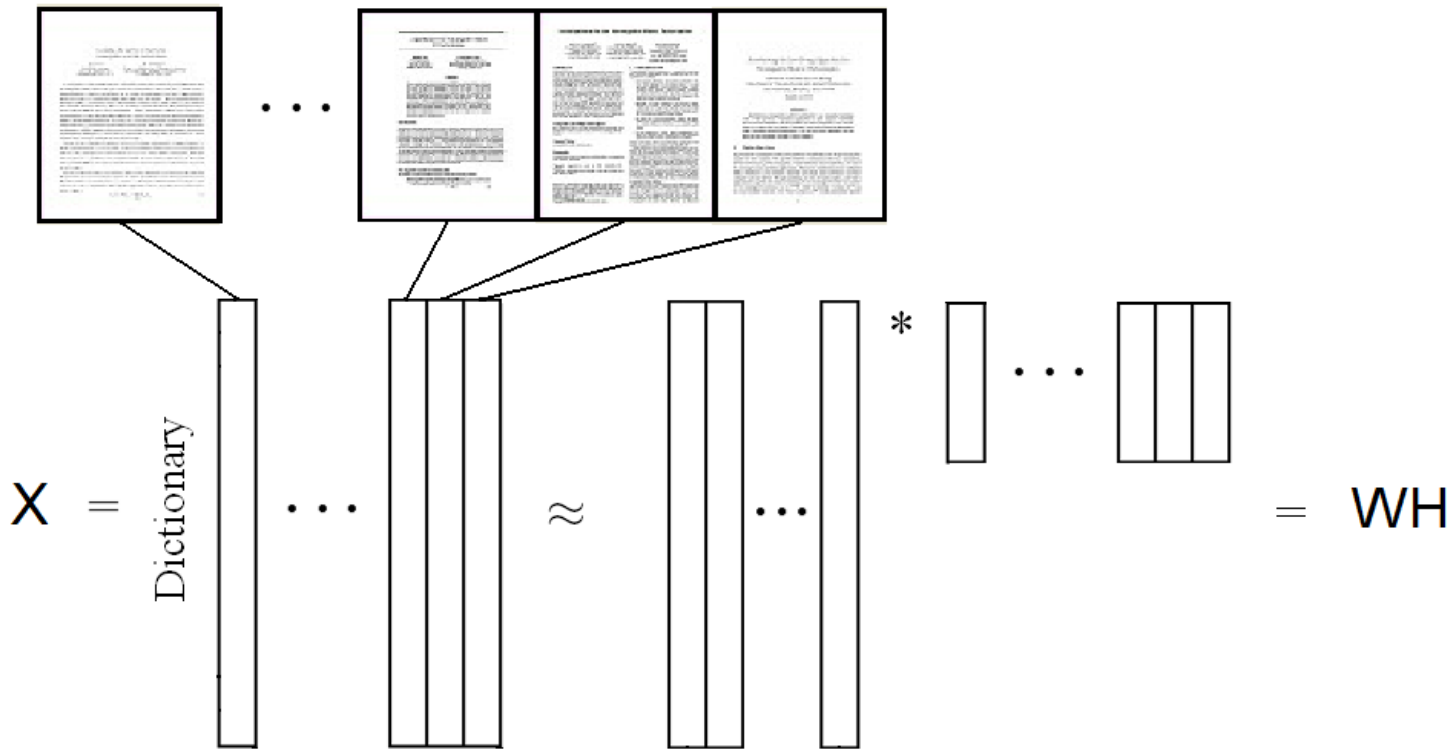
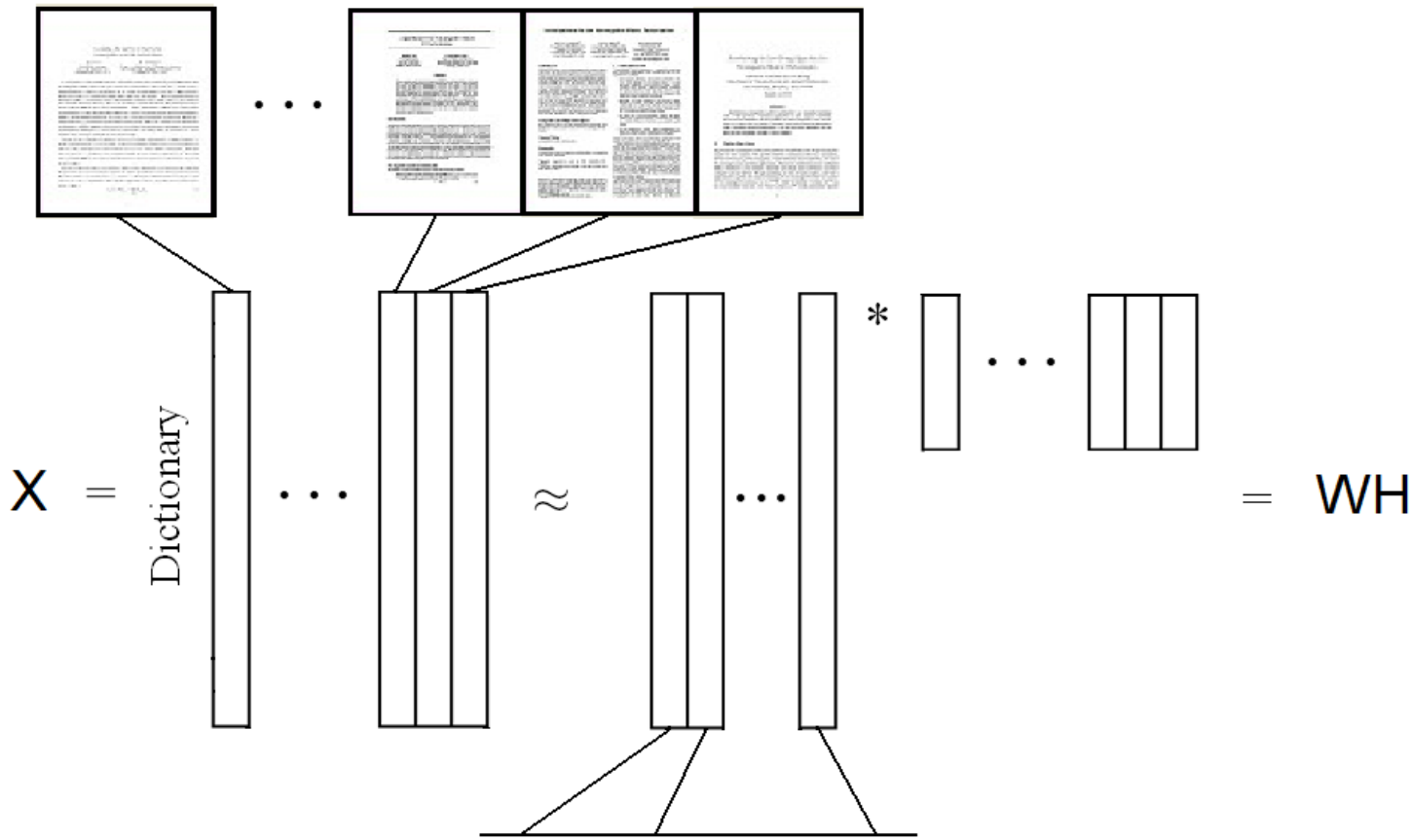


Figure 2: Decomposition of the Urban dataset.

Topic recovery and document classification

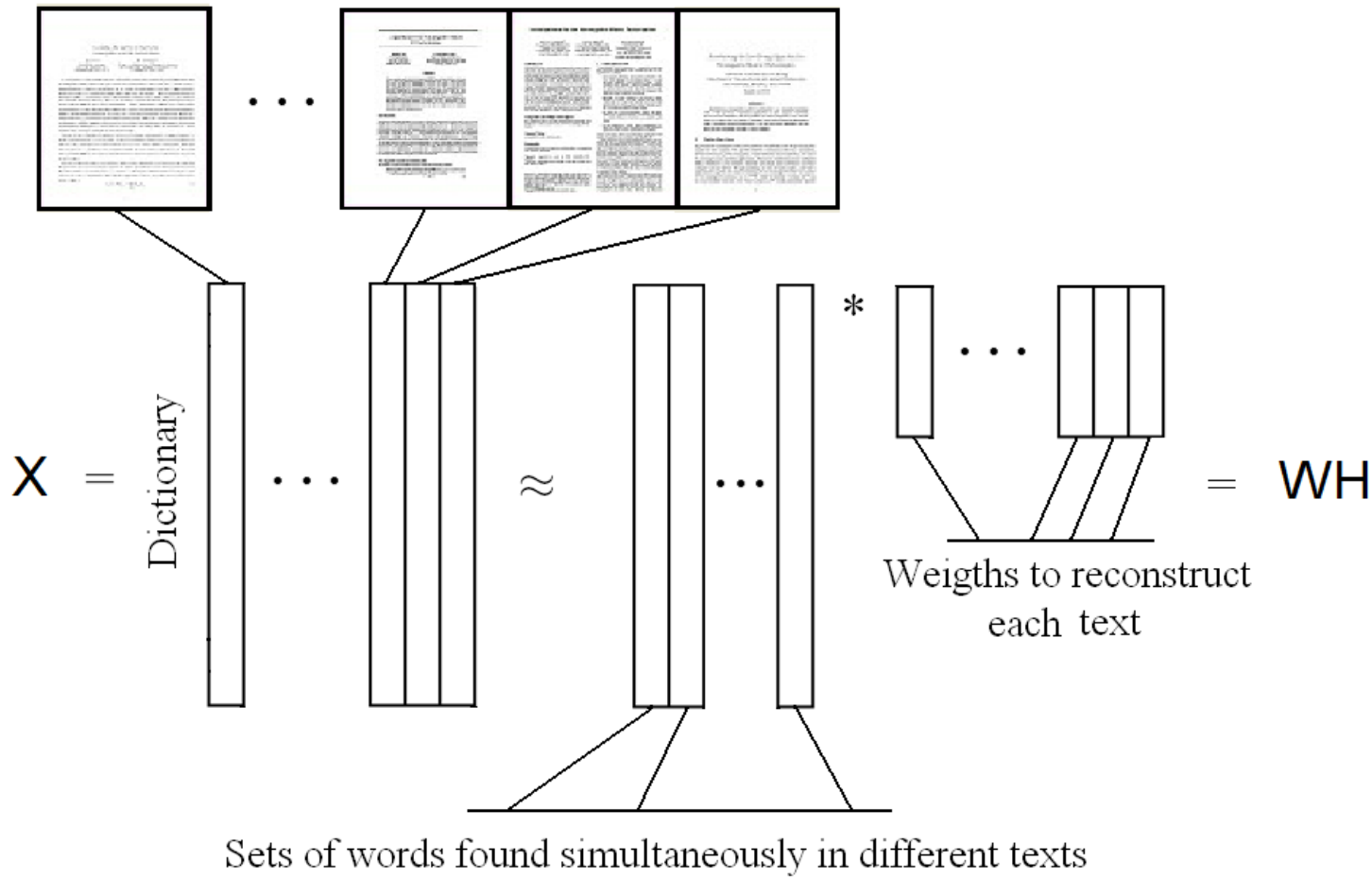


Topic recovery and document classification



- Basis elements allow to recover the different topics;

Topic recovery and document classification



- Basis elements allow to recover the different topics;
- Weights allow to assign each text to its corresponding topics.

Topic recovery and document classification

→ MATLAB demo.

tdt2_top30: The TDT2 corpus (Nist Topic Detection and Tracking corpus) consists of data collected during the first half of 1998 and taken from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). Only the largest 30 categories are kept, thus leaving us with 9,394 documents in total.

Toolbox: https://gitlab.com/ngillis/nmfbook/

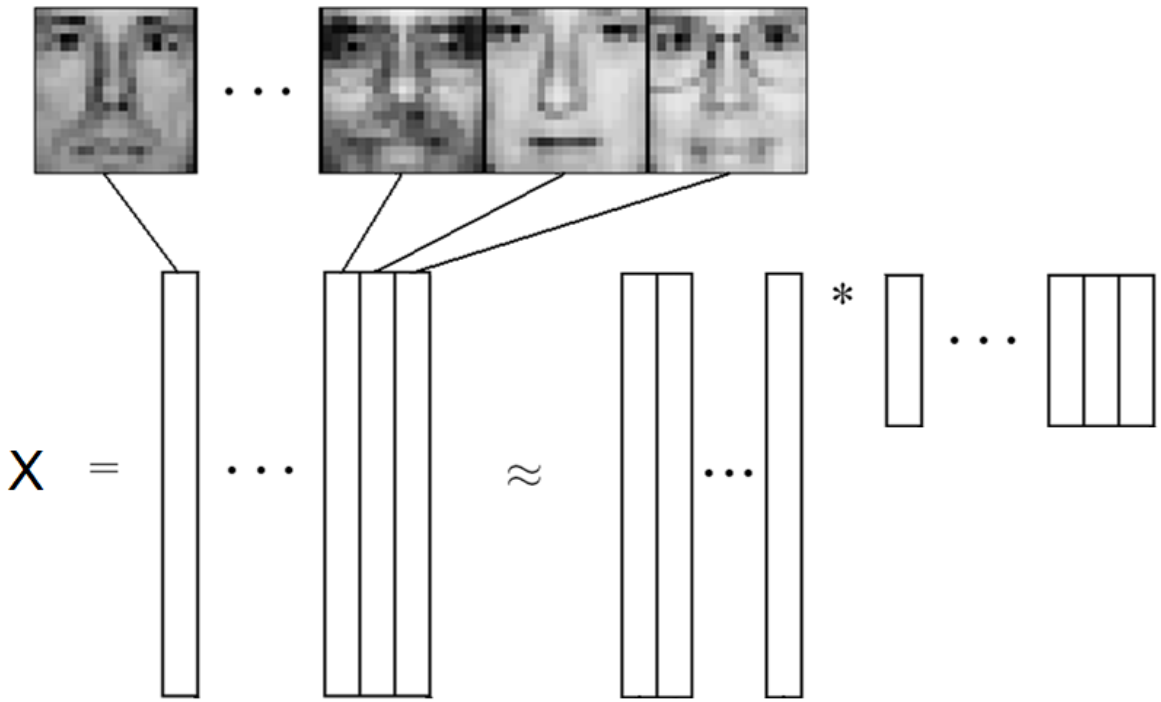
Topic recovery and document classification

Five of the topics extracted by NMF on tdt2_top30 (top 10 words):

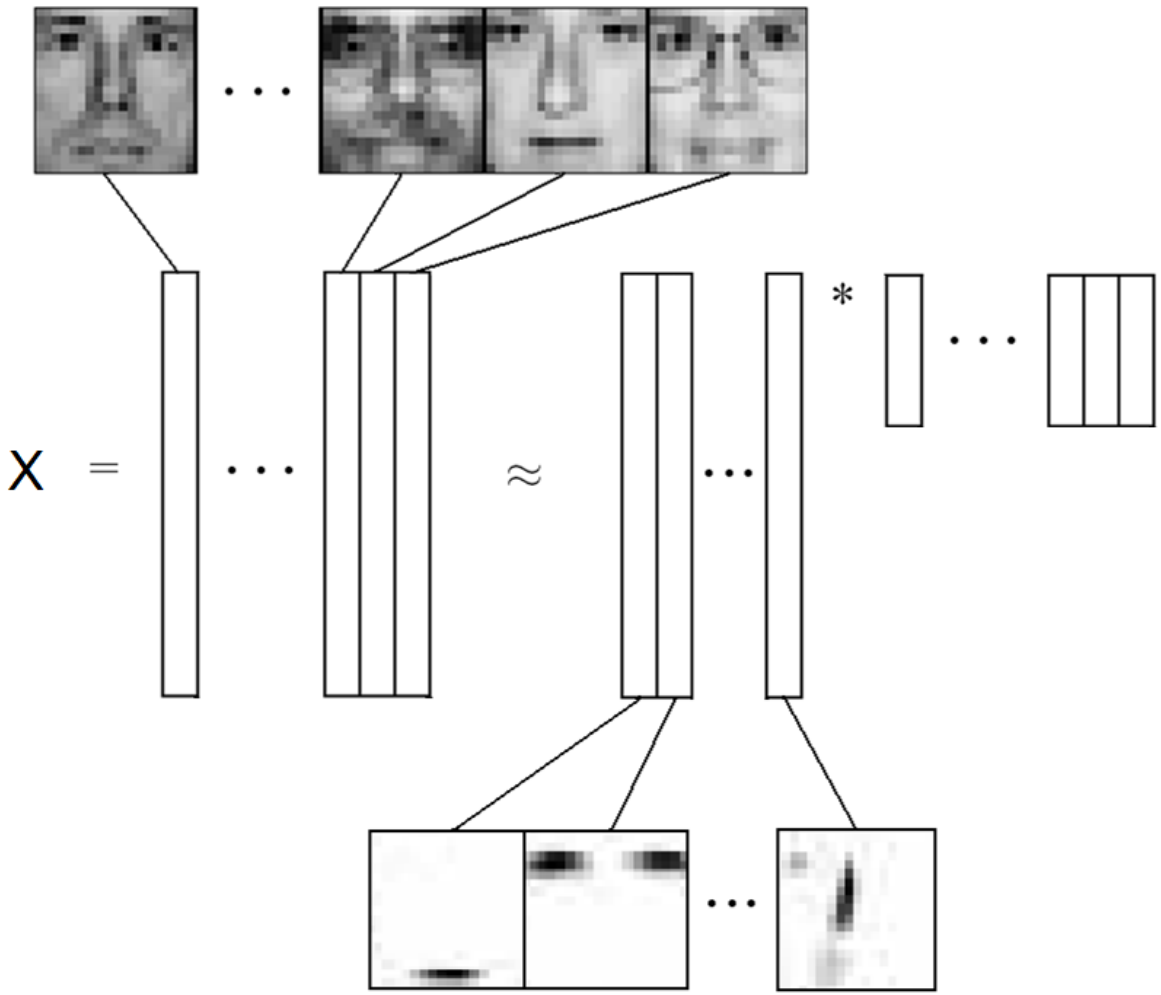
Lewinsky scandal	**Israeli-Palestinian conflict**	**Stock Market**	**Winter olympics in Nagano**	**Sports**
lewinsky	israel	percent	olympic	game
mrs	israeli	stock	games	denver
jones*	netanyahu	market	olympics	team
lawyers	palestinian	stocks	nagano	super
clinton	peace	points	gold	bowl
president	palestinians	investors	medal	packers
sexual	arafat	prices	team	jordan
jordan**	bank	index	japan	play
relationship	minister	companies	winter	green
told	talks	quarter	won	bulls

*Paula Jones sued Bill Clinton for an earlier sexual harassment affair.
**Vernon Jordan, a friend and political adviser to Bill Clinton, helped Monica Lewinsky after she left the White House.

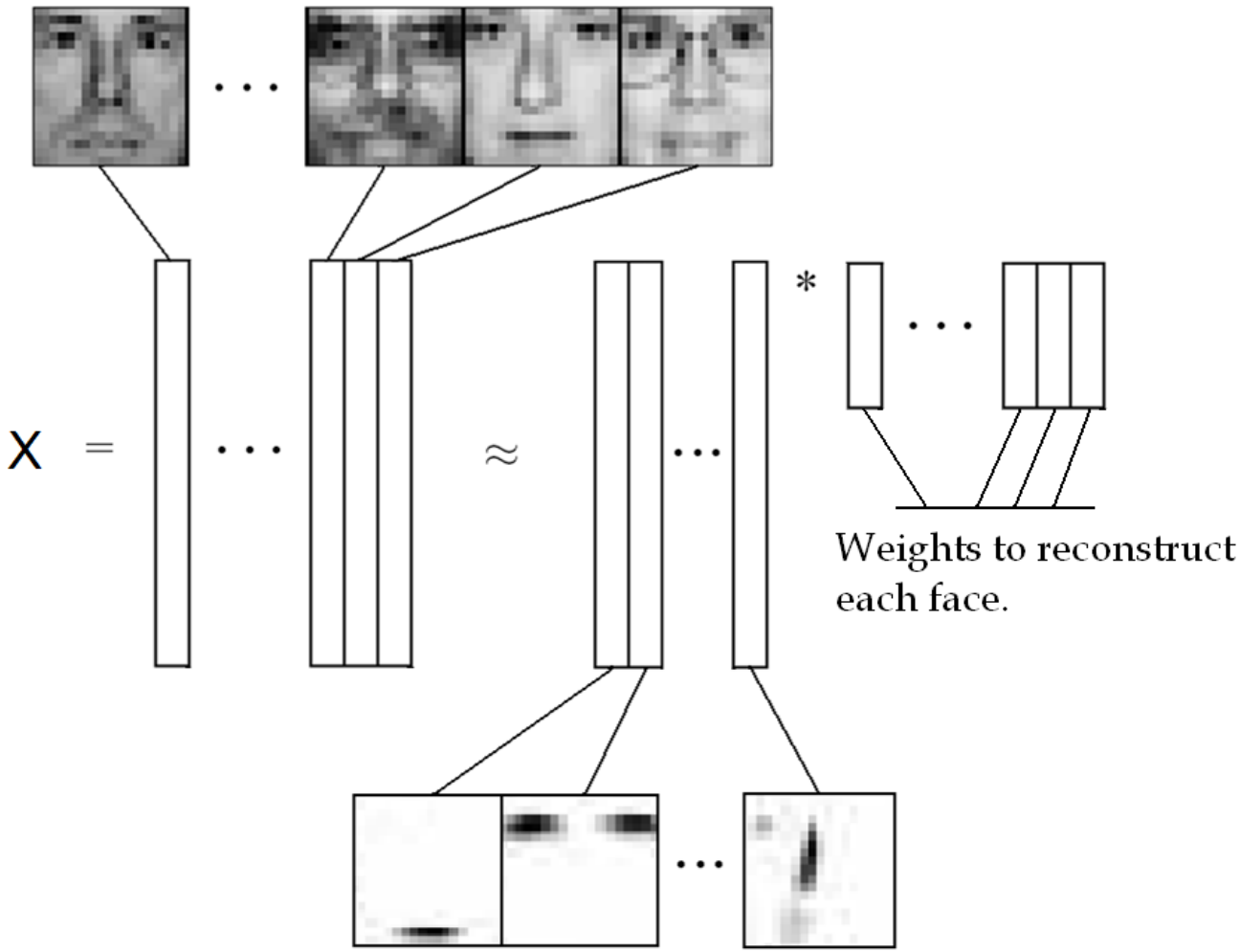
Facial feature extraction and classification



Facial feature extraction and classification

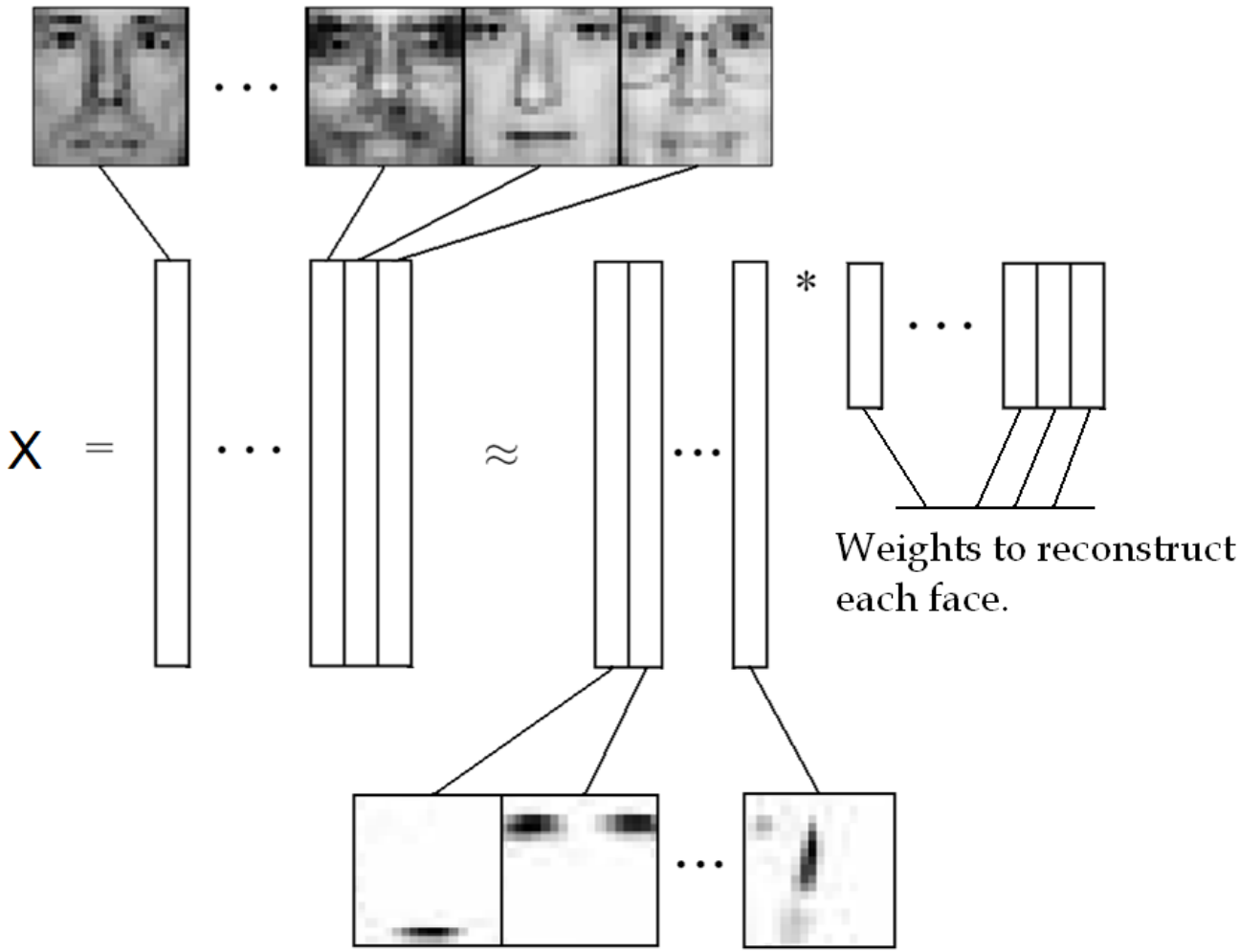


Facial feature extraction and classification



The basis elements extract facial features such as eyes, nose and lips.

Facial feature extraction and classification



The basis elements extract facial features such as eyes, nose and lips. $\rightarrow$ MATLAB demo.

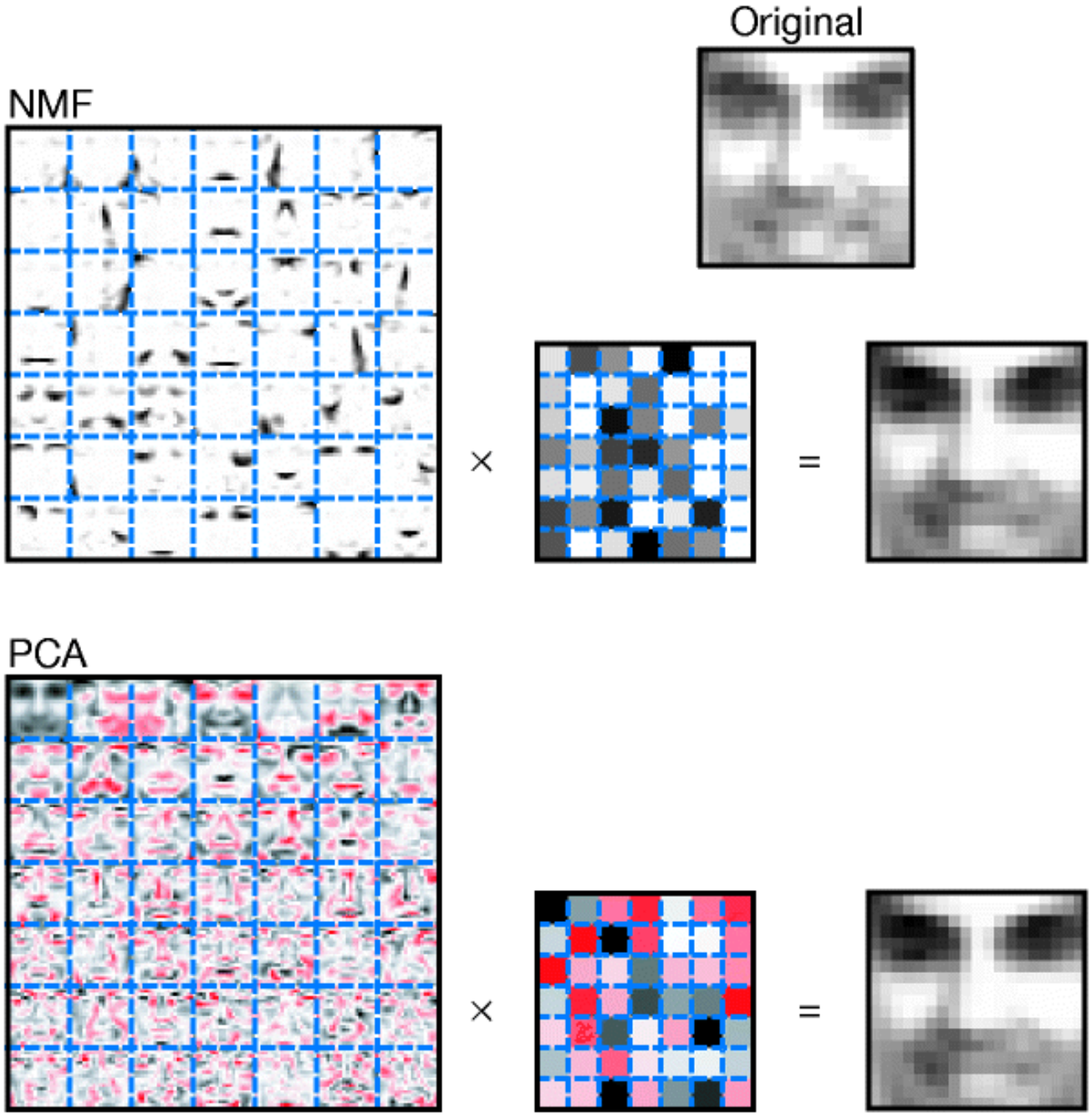
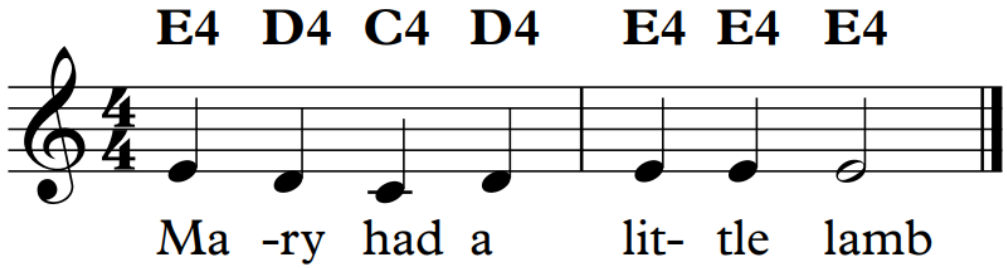
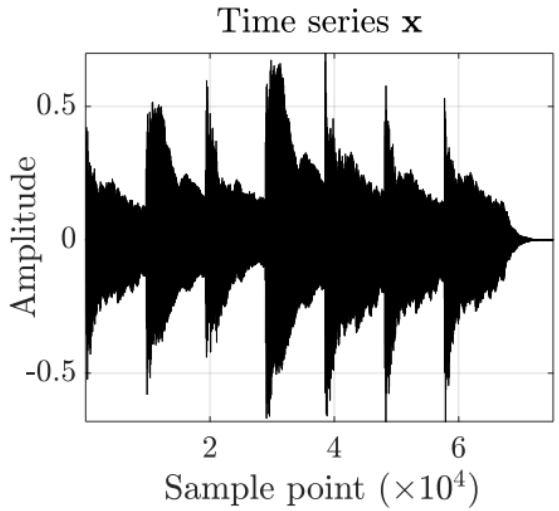


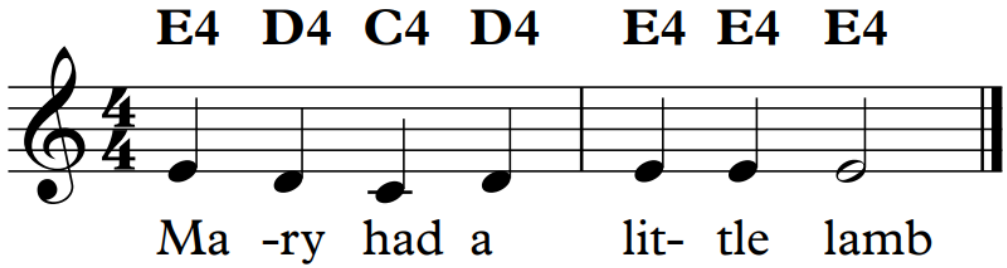
Figure 3: Experiment on facial images [LS99].

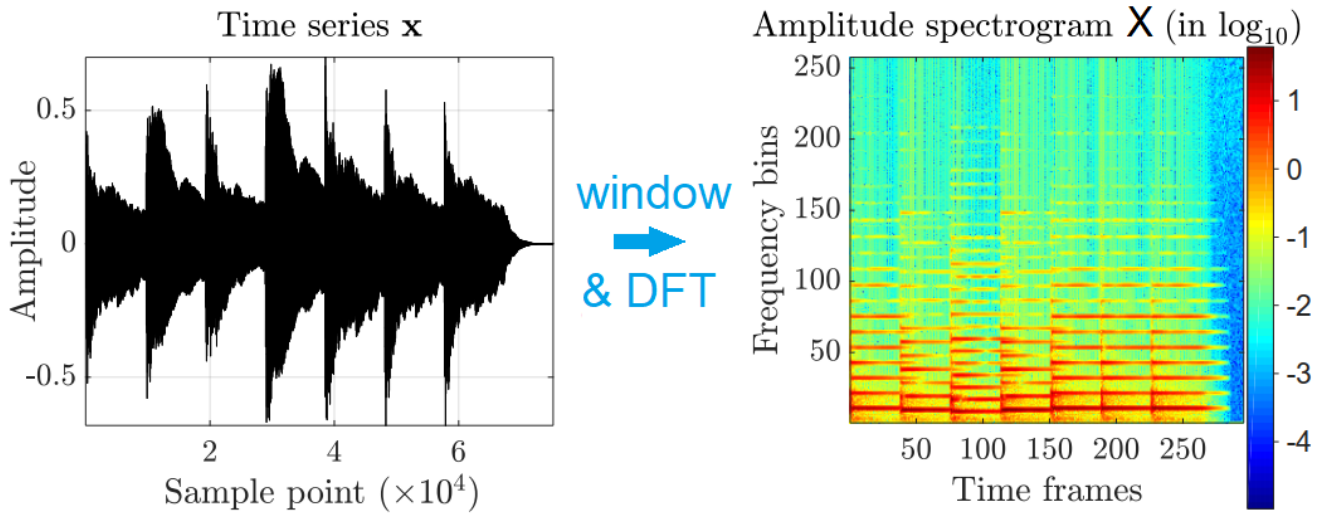
Audio source separation



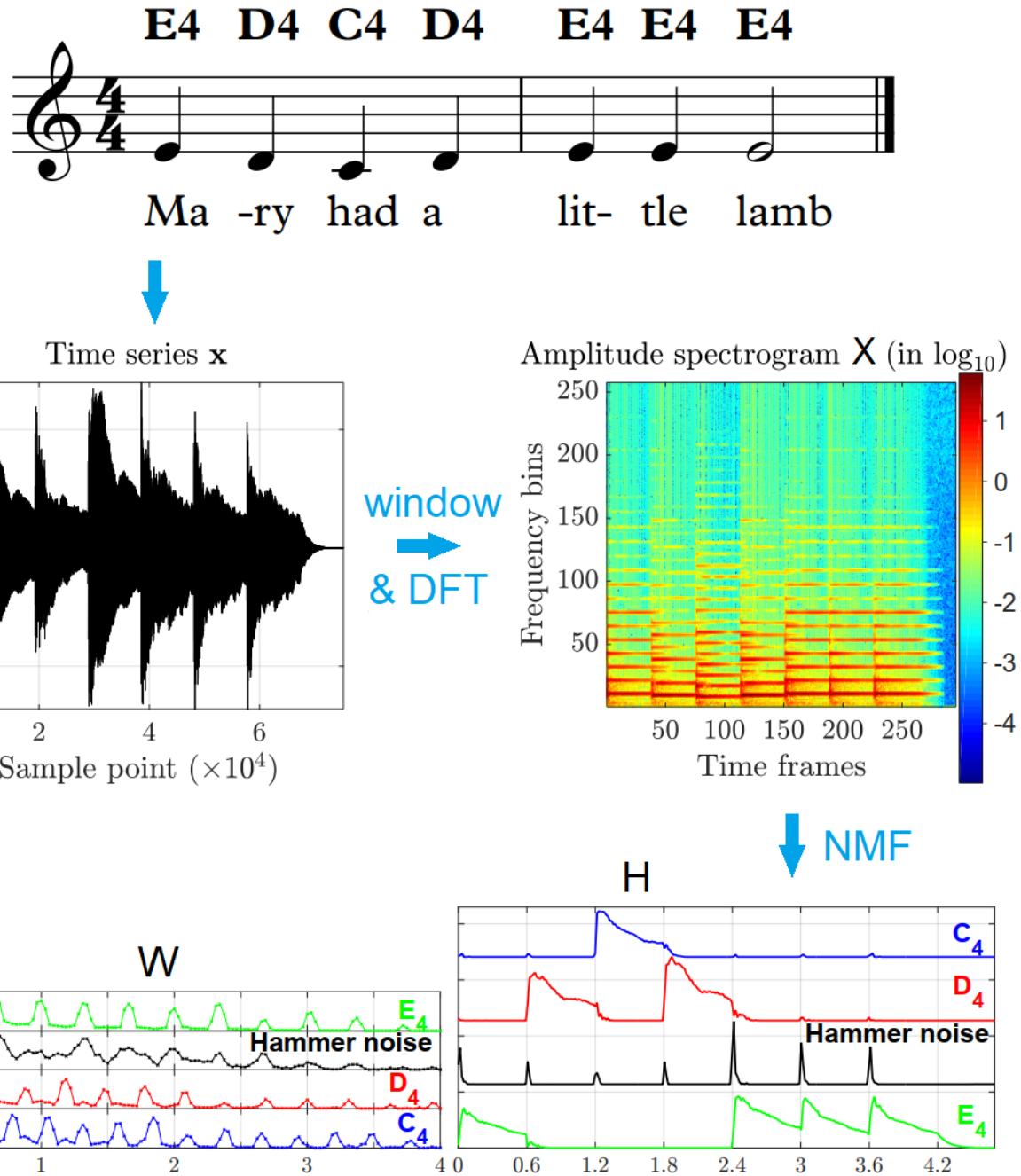


Audio source separation





Audio source separation



Nonnegative Matrix Factorization (NMF)

PART I – Application-oriented aspects

- Sec. 1: Applications of NMF
  - Data analysis: blind hyperspectral image unmixing, image feature extraction, topic modeling, audio source separation
  - **Statistical machine learning: crowdsourcing, joint probability estimation, Hidden Markov models, and community detection.**
- Sec. 2: NMF Algorithms
  - MU
  - Nonconvex Optimization
  - Non-Euclidean cost

PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

Statistical Unsupervised Machine Learning Problems

Problems that NMF can handle (incomplete list):

- Topic Modeling [AGH+13, HFS16]
- Crowdsourced Data Labeling [IFKH19]
- Learning Joint Probability from Marginals [IF20]
- Hidden Markov Model Identification [HFS18, LR10]
- Community Detection [HF19a, MSC17, PSU17]

- **Many are less obvious as NMF models** – but can be shown through careful constructions.
- **Warning**: I will use boldface $\boldsymbol{X}$ to represent matrices and X random variables.

Crowdsourced Data Labeling

Data Labeling, Deep Learning, and Crowdsourcing

- How many labels used in DL?
- Labeling millions of samples is costly
- Data labeling - an emerging business
  - *Forbes* in Sep 2019: 2 of "AI 50: America's Most Promising Artificial Intelligence Companies" are data labeling companies.
  - *Financial Times* in July 2019: the market for data labeling was already $150 million in 2018;
  - Market projected $\geq$$1B by 2023.

Amazon team taps millions of Alexa interactions to reduce NLP error rate

KYLE WIGGERS @KYLE_L_WIGGERS

JANUARY 22, 2019 6:59 AM

Amazon VP of devices David Limp at a September 2018 event at Amazon headquarters in Seattle, Washington.

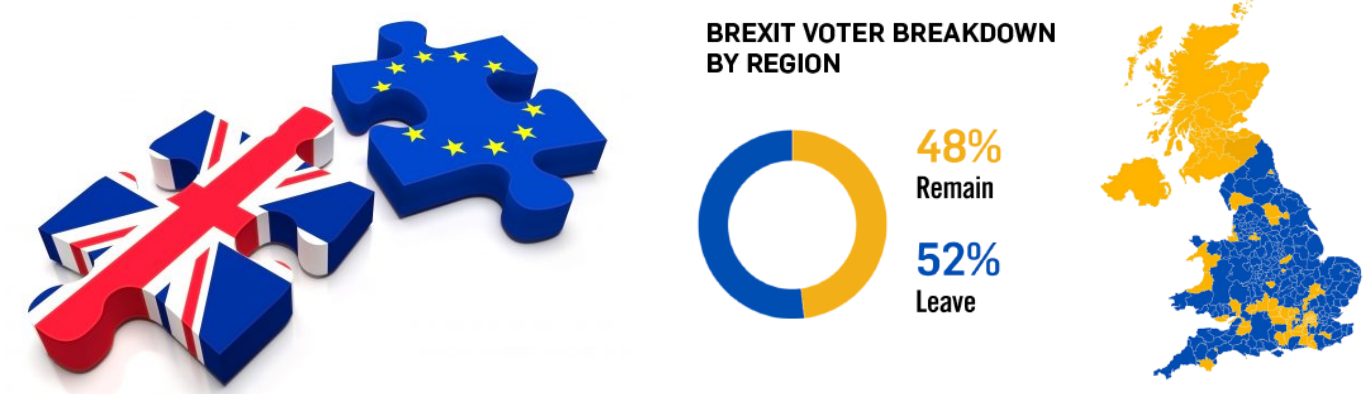
Image Credit: Khari Johnson / VentureBeat

- Labeling data items is nontrivial
- Any individual annotator could be easily **confused**

sources: left: markelbroch.com, right: Joel Sartore, Nat Geo Image Collection

- **Crowdsourcing**: use a group of annotators to label data - and then **integrate** the result
  - A naive integration approach is **majority voting**.

- **Majority voting** may or may not be optimal



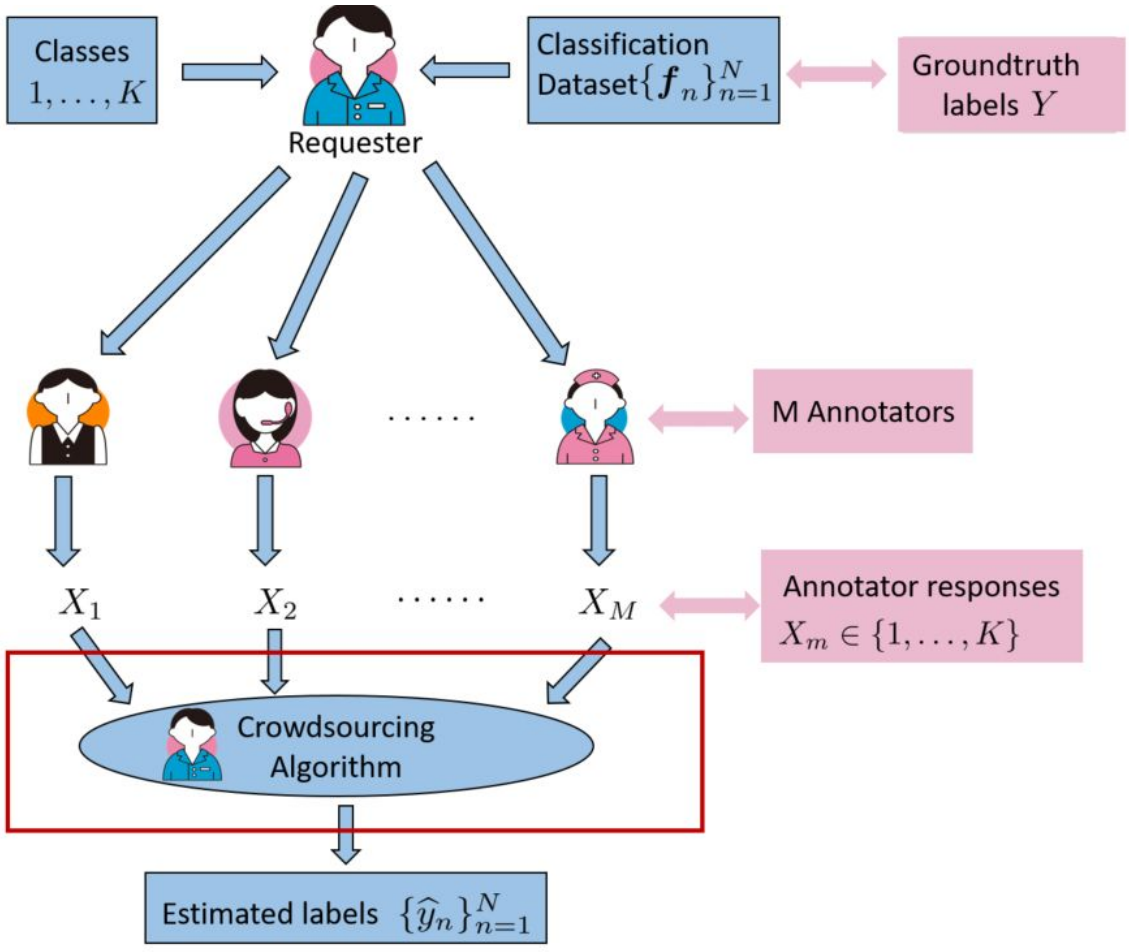
picture source: left: https://www.outsidethebeltway.com. right: CTVNews

- Many annotators may not have been well trained (self-registered)

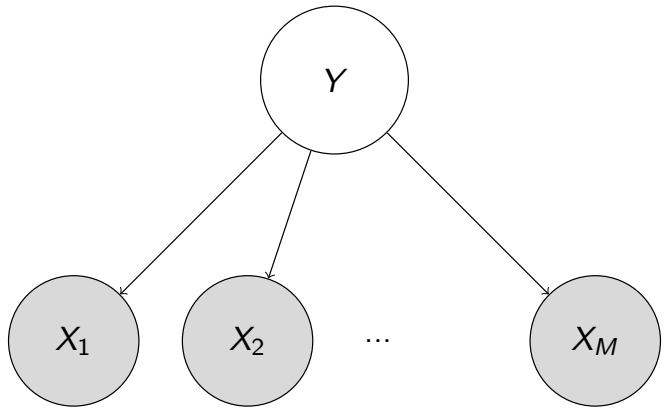
amazon mechanical turk

- Not all the annotators work on all the data (for cost control)

- **Problem setup**



- The **Dawid-Skene** model [DS79]



- A naive Bayes model
- (perhaps) the canonical model in statistical crowdsourcing [ZCZJ14a, ZCZJ14b, TPZG18, TG18, TG19, RYZ+10, SOJN08, SFB+95, ZBMP12].
- simple and effective

- A closer look

$$\Pr(X_1 = k_1, \ldots, X_M = k_M)$$
$$= \sum_{k=1}^{K} \underbrace{\Pr(Y = k)}_{\text{label prior}} \prod_{m=1}^{M} \underbrace{\Pr(X_m = k_m | Y = k)}_{\text{confusion prob of annot. m}}$$

- Define

$$\boldsymbol{d}(k) = \Pr(Y = k), \ \boldsymbol{A}_m(k_m, k) = \Pr(X_m = k_m | Y = k)$$

- $\boldsymbol{d} \in \mathbb{R}^K$ is the prior PMF
- $\boldsymbol{A}_m \in \mathbb{R}^{K \times K}$ is the **confusion matrix** of annotator m
- **Our goal:** identify $\boldsymbol{d}$ and $\boldsymbol{A}_m$ for $m = 1, \ldots, M$

- Consider **second-order** statistics (co-occurrences):

$$\boldsymbol{R}_{m,\ell}(k_m, k_\ell) = \Pr(X_m = k_m, X_\ell = k_\ell)$$
$$= \sum_{k=1}^{K} \underbrace{\Pr(Y = k)}_{\boldsymbol{d}(k)} \underbrace{\Pr(X_m = k_m | Y = k)}_{\boldsymbol{A}_m(k_m,k)} \underbrace{\Pr(X_\ell = k_\ell | Y = k)}_{\boldsymbol{A}_\ell(k_\ell,k)}$$

- In matrix form:

$$\boldsymbol{R}_{m,\ell} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_\ell^\top,$$

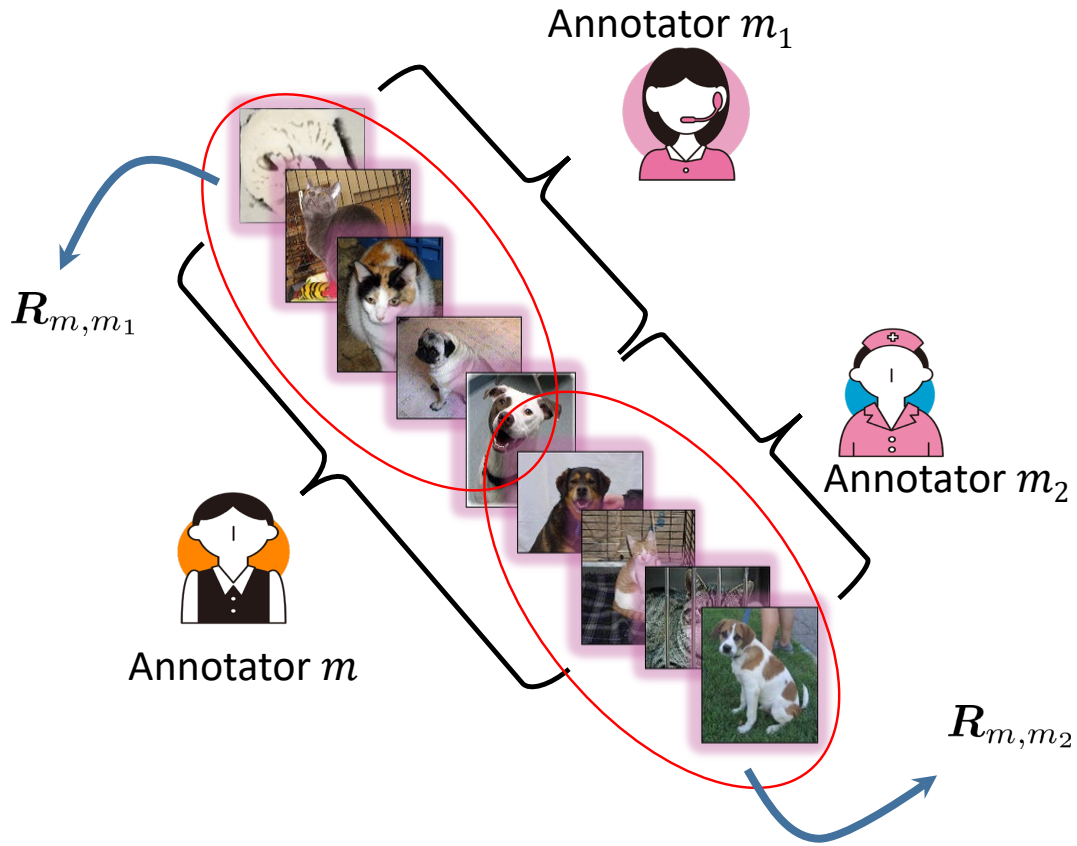
  where $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{d})$, $\boldsymbol{R}_{m,\ell} \in \mathbb{R}^{K \times K}$

- Already an NMF model:

$$\boldsymbol{R}_{m,\ell} = \underbrace{\boldsymbol{A}_m \boldsymbol{D}}_{\boldsymbol{W}} \underbrace{\boldsymbol{A}_\ell^\top}_{\boldsymbol{H}}.$$

  - normally won't work if directly factoring $\boldsymbol{R}_{m,\ell}$ (see Sec. II).
  - we hope $\boldsymbol{H}$ to be a very fat matrix (roughly speaking)

- Annotator m may co-label data with annotators $m_1, \ldots, m_{T(m)}$.



- Assume $\boldsymbol{R}_{m,m_t}$ for $t \in \{1, \ldots, T(m)\}$ are available.

- Construct

$$
\begin{aligned}
\boldsymbol{Z}_m &= \left[\boldsymbol{R}_{m,m_1}, \boldsymbol{R}_{m,m_2}, \ldots, \boldsymbol{R}_{m,m_{T(m)}}\right] \\
&= \left[\boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_{m_1}^\top, \ldots, \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_{T(m)}^\top\right] \\
&= \boldsymbol{A}_m \underbrace{\left[\boldsymbol{D} \boldsymbol{A}_{m_1}^\top, \ldots, \boldsymbol{D} \boldsymbol{A}_{T(m)}^\top\right]}_{\boldsymbol{H}_m} \in \mathbb{R}^{K \times KT(m)} \\
&= \boldsymbol{W}_m \boldsymbol{H}_m
\end{aligned}
$$

- Why the trouble?
  - We hope $\boldsymbol{H}_m$ to be "fat" so that it has a better chance to attain the "separability" or "sufficiently scattered" conditions—which makes $\boldsymbol{W}_m$ and $\boldsymbol{H}_m$ identifiable.
  - See identifiability theory (Part II).

- The NMF procedure is denoted as MultiSPA [IFKH19]:

Classification Error (%) and Run-time (sec) : AMT Datasets

Algorithms	TREC		Bluebird		RTE	
	(%) Error	(sec) Time	(%) Error	(sec) Time	(%) Error	(sec) Time
`MultiSPA`	31.47	50.68	13.88	0.07	8.75	0.28
`Spectral-D&S`	29.58	919.98	12.03	1.97	7.12	6.40
`TensorADMM`	N/A	N/A	12.03	2.74	N/A	N/A
`Majority Voting`	34.85	N/A	21.29	N/A	10.31	N/A

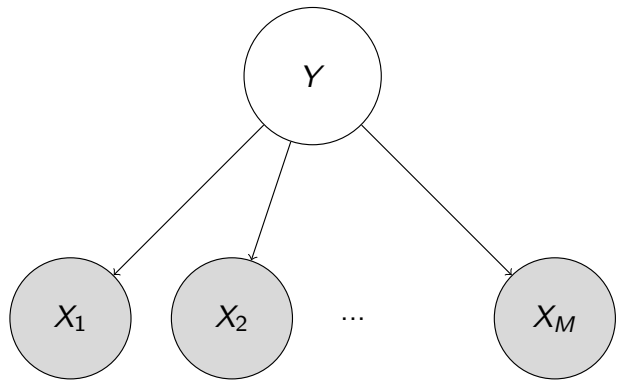
AMT Dataset description

Dataset	# classes	# items	# annotators	# annotator labels
Bluebird	2	108	30	3240
RTE	2	800	164	8,000
TREC	2	19,033	762	88,385

Joint Prob. Estimation from Marginals

A Bigger Picture - Joint PMF Learning from Marginals

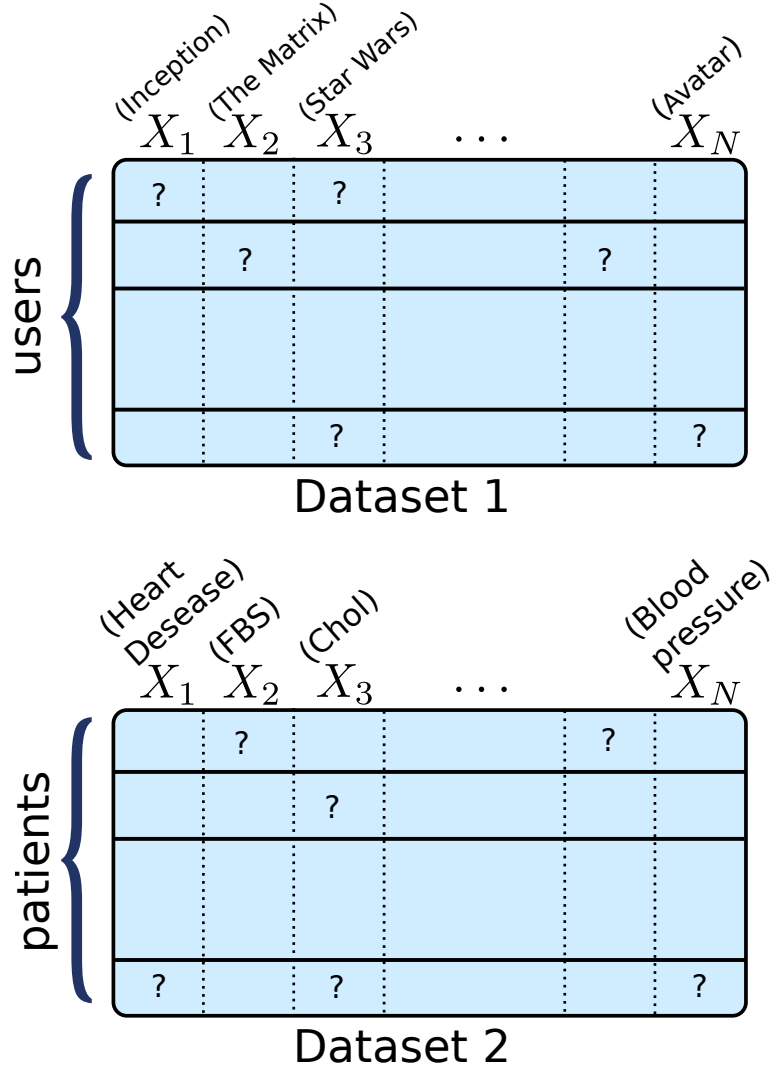
- Essentially, we have shown that $\Pr(X_1, \ldots, X_M)$ can be learned from $\Pr(X_i, X_j)$.
  - if $X_1, \ldots, X_M$ are conditionally independent given a hidden discrete variable Y.
- In other words, the **Naive Bayes structure** is the key.



- Can this be generalized to handle *any* $X_1, \ldots, X_M$?

Many ML Problems Boil Down to Joint PMF Learning

- Direct estimation for $\Pr(X_1, \ldots, X_N)$ suffers from the curse of dimensionality.
- #Sample $\gg \Omega(I^N)$, where I is the alphabet size of X_n.
- Est. $\Pr(X_i, X_j)$ is much easier.



One-slide Proof

- The joint PMF of $X_1, \ldots, X_N$ is an $I_1 \times I_2 \times \ldots \times I_N$ tensor, where I_m is the size of the alphabet of X_m.

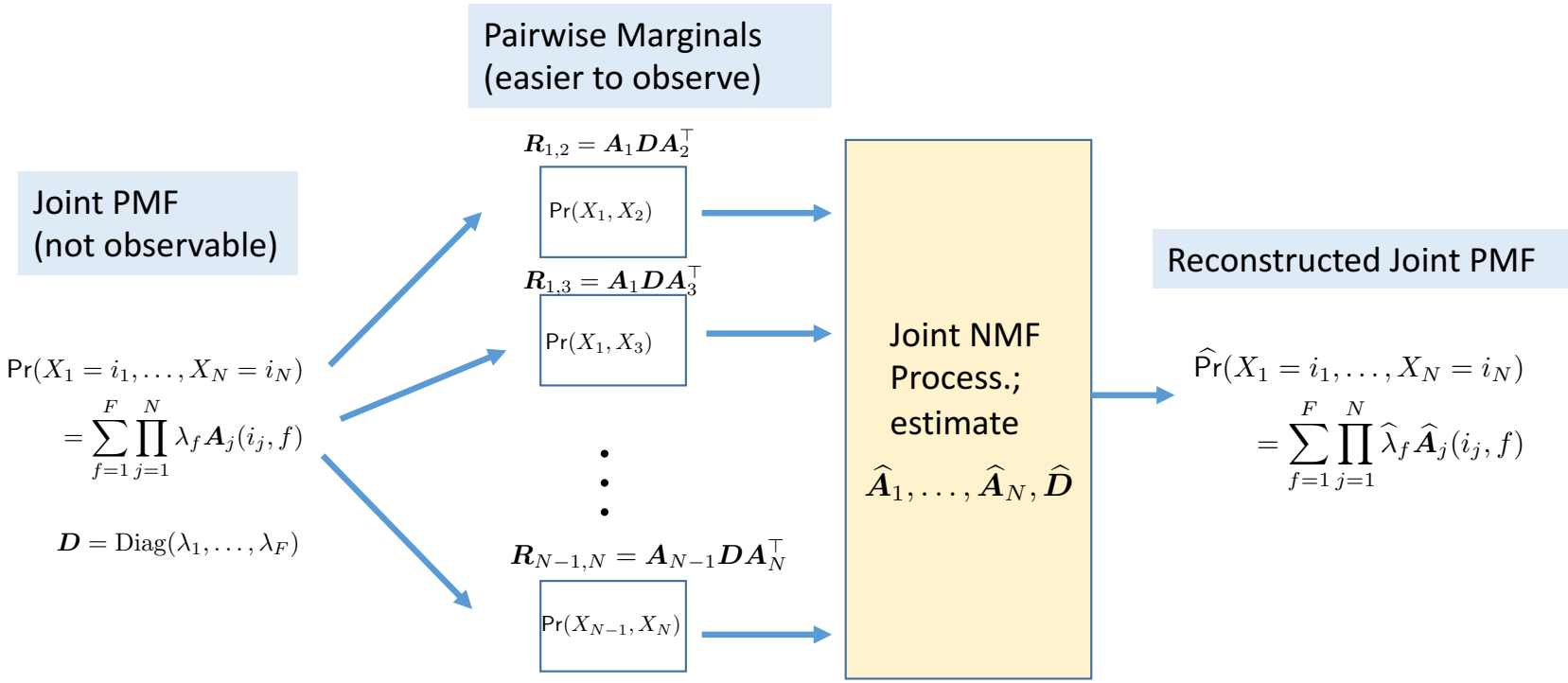
$$\Pr(i_1, \ldots, i_N) = \underline{\boldsymbol{X}}(i_1, \ldots, i_N)$$

- Every nonnegative tensor admits a nonnegative canonical polyadic decomposition (CPD) with a finite rank F [KSF18, SDLF+17]:

$$\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \sum_{f=}^{F} \prod_{j=1}^{N} \underbrace{\lambda_f}_{\Pr(Y=f)} \underbrace{\boldsymbol{A}(i_j, f)}_{\Pr(X_j=i_j \mid Y=f)}$$

- Hence, **every joint PMF admits a Naive Bayes representation**, with a latent Y that has a finite alphabet.
- When $X_1, \ldots, X_N$ are "reasonably dependent", F is not large [KSF18].

Schematic of The Approach



- Movie Recommendation

Movie Recommendation Results on the Animation set (MovieLens)

Algorithm	RMSE	MAE	Time (s)
CNMF-SPA [Proposed]	0.8705±0.0095	0.6798±0.0060	0.028
CNMF-OPT [Proposed]	**0.8124±0.0031**	**0.6241±0.0041**	61.018
CNMF-SPA-EM [Proposed]	0.8170±0.0075	0.6317±0.0086	2.424
CTD [Kargas *et al.*]	0.8300±0.0053	0.6335±0.0029	48.253
BMF	0.8408±0.0023	0.6553±0.0015	46.637
Global Average	0.9371±0.0021	0.7042±0.0014	–
User Average	0.8850±0.0009	0.6632±0.0011	–
Movie Average	0.9027±0.0019	0.6900±0.0013	–

- Estimate the rating of movie N by user k via computing the MMSE estimator:

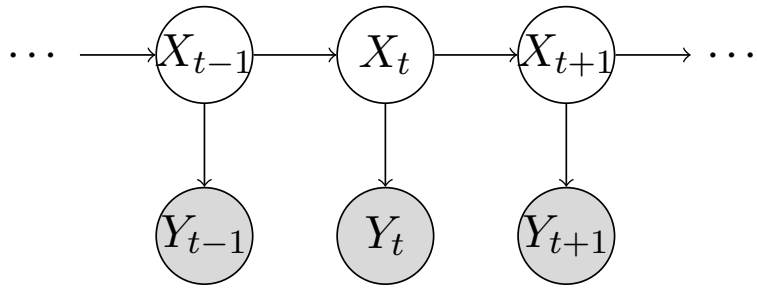
$$\mathbb{E}[r_k(N)|r_k(1),\ldots,r_k(N-1)],$$

where the joint PMF of a user rating N movies $\Pr(r(1),\ldots,r(N))$ is estimated.

- more in [IF20].

Learning Hidden Markov Models

Hidden Markov Model Learning



- HMMs are widely used in machine learning for modeling time series, e.g., speech and language data.
- a sequence of observable data $\{Y_t\}_{t=0}^{T}$ emitted from an underlying unobservable Markov chain $\{X_t\}_{t=0}^{T}$.
- The HMM estimation problem aims at identifying the **transition probability** $\Pr[X_{t+1}|X_t]$ and the **emission probability** $\Pr[Y_t|X_t]$.

- How is HMM related to NMF?

$$\Pr[Y_t, Y_{t+1}] = \sum_{k,j=1}^{R} \Pr[Y_t|X_t = x_k] \Pr[Y_{t+1}|X_{t+1} = x_j] \times \Pr[X_t = x_k, X_{t+1} = x_j]$$

- Define $\boldsymbol{\Omega} \in \mathbb{R}^{M \times M}$, $\boldsymbol{M} \in \mathbb{R}^{M \times R}$, and $\boldsymbol{\Theta} \in \mathbb{R}^{R \times R}$ such that

$$\begin{aligned} \boldsymbol{\Omega}(m, \ell) &= \Pr[Y_t = y_m, Y_{t+1} = y_\ell], \\ \boldsymbol{M}(m, r) &= \Pr[Y_t = y_m | X_t = x_r], \\ \boldsymbol{\Theta}(k, j) &= \Pr[X_t = x_k, X_{t+1} = x_j]. \end{aligned}$$

Here, M denotes the number of observable states and R the number of hidden states.

- We then have [HFS18, LR10]:

$$\boldsymbol{\Omega} = \boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^{\top}.$$

- Let $\boldsymbol{\Omega} = \boldsymbol{X}$, $\boldsymbol{W} = \boldsymbol{M}$ and $\boldsymbol{H} = \boldsymbol{\Theta}\boldsymbol{M}^{\top}$, we have an NMF model $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$.
  - columns of $\boldsymbol{M}$ are emission probabilities.
  - elements of $\boldsymbol{\Theta}$ are transition probabilities .
- Using HMM to handle topic modeling

china daily vermin eat pct grain stocks survey provinces and cities showed vermin consume and pct china grain stocks china daily each year mln tonnes pct china fruit output left rot and mln tonnes pct vegetables paper blamed waste inadequate storage and bad preservation methods government had launched national programme reduce waste calling for improved technology storage and preservation and greater production additives paper gave details

china daily vermin eat pct grain stocks survey provinces and cities showed vermin consume and pct china grain stocks china daily each year mln tonnes pct china fruit output left rot and mln tonnes pct vegetables paper blamed waste inadequate storage and bad preservation methods government had launched national programme reduce waste calling for improved technology storage and preservation and greater production additives paper gave details

Wrap-up

- Many statistical model learning problems can be converted to an NMF problem.
- Less obvious compared to examples like HU.
- Often need to construct some statistics from the data.
- Some more examples in
  - Community detection [HF19a, MSC17, PSU17].

Nonnegative Matrix Factorization (NMF)

PART I – Application-oriented aspects

- Sec. 1: Applications of NMF
  - Data analysis: blind hyperspectral image unmixing, image feature extraction, topic modeling, audio source separation
  - Statistical machine learning: crowdsourcing, joint probability estimation, Hidden Markov models, and community detection.
- **Sec. 2: NMF Algorithms**
  - **MU**
  - **Nonconvex Optimization**
  - **Non-Euclidean cost**

PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

Algorithms

Problem formulation:

$$\underset{W,H\geq 0}{\text{minimize}} \quad \|X - WH^T\|^2$$

- nonconvex and NP-hard [Vav10]
- fix one factor (say W), convex w.r.t. the other (H)
- still, nonnegative least squares has no closed-form solution

convergence at-a-glance:

- block coordinate descent (BCD) requires each update to be unique [Ber99], which is hard to guarantee
- block successive upperbound minimization (BSUM) similar [RHL13]
- BCD with 2 blocks: every limit point is a limit point (without uniqueness of the minimizer) [GS00]

Multiplicative Update (MU)

MU for least squares loss [LS01]

$$H \leftarrow H \circledast X^T W \oslash W H^T H \qquad W \leftarrow W \circledast X H \oslash H W^T W$$

- an instance of BSUM
- upperbound function is a quadratic with diagonal Hessian
- minimizer is unique if $W, H > 0$ strictly
- if initialized dense, will remain dense (thus converges)

Hierarchical alternating least squares (HALS)

column-wise BCD [CP09]

$$w_j \leftarrow \left[w_j + (Xh_j - WH^T h_j)/h_j^T h_j\right]_+ \qquad j = 1, \ldots, k$$

$$h_j \leftarrow \left[h_j + (X^T w_j - HW^T w_j)/w_j^T w_j\right]_+ \qquad j = 1, \ldots, k$$

- an instance of BCD with $2k$ blocks
- not guaranteed to converge in general

Alternating projected gradient (APG)

APG works better than direct PG [Lin07]

$$H \leftarrow [H - \gamma(HW^T W - X^T W)]_+ \qquad W \leftarrow [W - \gamma(WH^T H - XH)]_+$$

- every limit point is a stationary point (with appropriate γ)
- could be accelerated by extrapolation [XYWZ12]

Exact 2-block BCD

Can we apply exact 2-block BCD?

- use a double-loop algorithm
- many choices for the inner-loop

alternating optimization (AO) variants:

- accelerated MU and HALS [GG12]
- (accelerated) gradient descient [Lin07, GTLY12]
- active set [KP08]
- block principal pivoting [KP11]
- alternating direction method of multipliers [HSL16]

ALS revisited

$$\underset{W,H}{\text{minimize}} \quad \|X - WH^T\|_F^2$$

- Fix W, update H: $H^T = (W^T W)^{-1} W^T X$.
  - $W^T X$: $\mathcal{O}(\text{nnz}(X)k)$;
  - Cholesky decomposition of $W^T W = LL^T$: $\mathcal{O}(mk^2 + k^3)$;
  - $L^{-T} L^{-1} (W^T X)$ using forward/backward substitution: $\mathcal{O}(nk^2)$.
- Then fix H and update W similarly.

AO sub-problem

Now we focus on how to efficiently solve

$$\underset{H}{\text{minimize}} \quad \ell(X - WH^T) + r(H).$$

- For matrix factorization, this is just the update of the right factor;
- We propose to solve it using alternating direction method of multipliers (ADMM).

ADMM: scaled form

Consider a convex optimization problem in the following form.

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & f(x) + g(z) \\ & Ax + Bz = c, \end{aligned}$$

The alternating direction method of multipliers (ADMM) [GM75], [GM76], [BPC+11]:

$$\begin{aligned} x &\leftarrow \arg\min_x f(x) + (\rho/2)\|Ax + Bz - c + u\|_2^2, \\ z &\leftarrow \arg\min_z g(z) + (\rho/2)\|Ax + Bz - c + u\|_2^2, \\ u &\leftarrow u + (Ax + Bz - c), \qquad \text{dual update} \end{aligned}$$

For convex problems, converges to the global solution.

Least-squares loss

Adopting the alternating optimization framework, reformulate the sub-problem for H in each AO iteration as

$$\begin{aligned} \underset{H,\tilde{H}}{\text{minimize}} \quad & \|X - W\tilde{H}\|_F^2 + r(H) \\ & H = \tilde{H}^T. \end{aligned}$$

ADMM iterates:

$$\begin{aligned} \tilde{H} &\leftarrow (W^T W + I)^{-1}(W^T X + \rho(H + U)^T), \\ H &\leftarrow \arg\min_H r(H) + \frac{\rho}{2}\|H - \tilde{H}^T + U\|_F^2, \\ U &\leftarrow U + H - \tilde{H}^T. \end{aligned}$$

Inner convergence

- Linear convergence. Theoretically, best convergence rate given by $\rho = \sigma_{\max}(W)\sigma_{\min}(W)$;
- Empirically, $\rho = \text{trace}(W^T W)/k$ works almost as good (and much easier to obtain);
- Initialize H and $\boldsymbol{U}$ from the previous AO outer-loop, then optimality gap is bounded by the per-iteration improvement of an AO step;
- After ~ 10 AO outer-iterations, ADMM converges in only one inner-iteration.

Implementation: $\tilde{H}$

$$\tilde{H} \leftarrow (W^T W + \rho I)^{-1}(W^T X + \rho(H + U)^T)$$

- $W^T X$ and $W^T W$ only need to be computed once, with complexity $\mathcal{O}(\text{nnz}(X)k)$ and $\mathcal{O}(mk^2)$ respectively;
- Can cache the Cholesky decomposition ($\mathcal{O}(k^3)$) of $W^T W + \rho I = LL^T$;
- Within one ADMM iteration, $\tilde{H}$ is obtained from one forward substitution and one backward substitution, with complexity $\mathcal{O}(nk^2)$.

Implementation: H

$$H \leftarrow \arg\min_H r_H(H) + \frac{\rho}{2}\|H - \tilde{H}^T + U\|_F^2$$

so-called proximity operator, in a lot of cases can be efficiently evaluated

- non-negative: $H \geq 0$;
- l_1 regularization: $\lambda\|H\|_1$ (soft thresholding);
- sum to one: $H1 = 1$ or $H^T 1 = 1$;
- smoothness regularization: $\lambda\|TH\|_F^2$ where T is tri-diagonal (linear complexity using banded-system solver).

All with $\mathcal{O}(nk)$ complexity.

ADMM inner-loop: least-squares loss

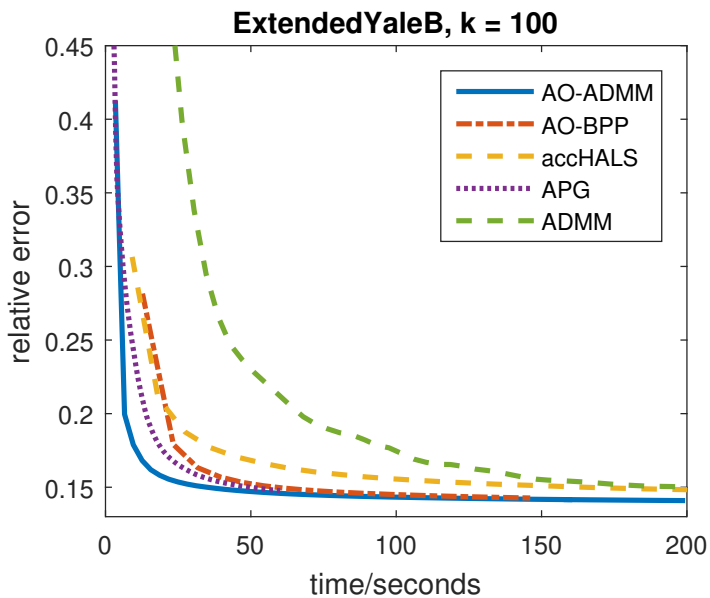
1: Initialize H and $\boldsymbol{U}$ from previous AO iteration
2: $G = W^T W$, $F = W^T X$
3: $\rho = \text{trace}(G)/k$
4: Cholesky decomposition: $G + \rho I = LL^T$
5: **repeat**
6: $\quad \tilde{H} \leftarrow L^{-T} L^{-1}(F + \rho(H + U))$ $\quad$ by substitution
7: $\quad H \leftarrow \arg\min_H r_H(H) + \frac{\rho}{2}\|H - \tilde{H}^T + U\|_F^2$
8: $\quad U \leftarrow U + H - \tilde{H}$
9: **until** convergence
10: **return** H and U.

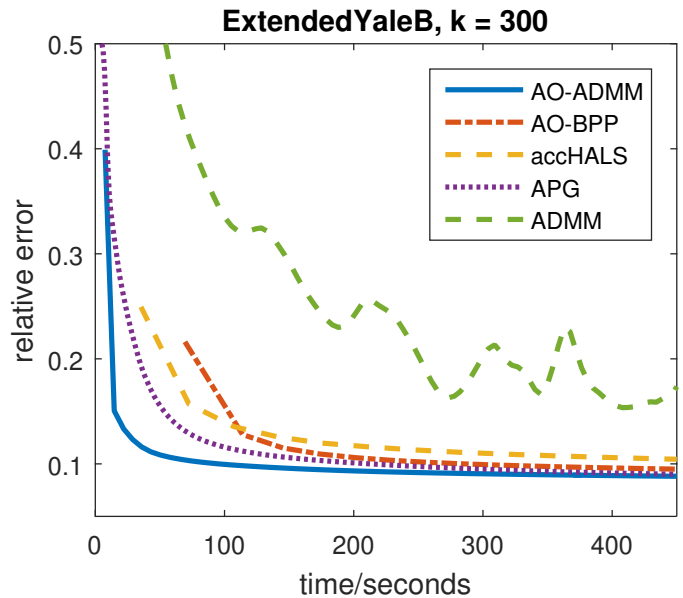
- Most of the computations are done at line 2 and line 4;
- One iteration of ADMM has complexity = one ALS update;
- Only one matrix inversion is required (line 4).

NMF

Extended Yale Face Database B

- A dense image data set;
- 32256×1932.

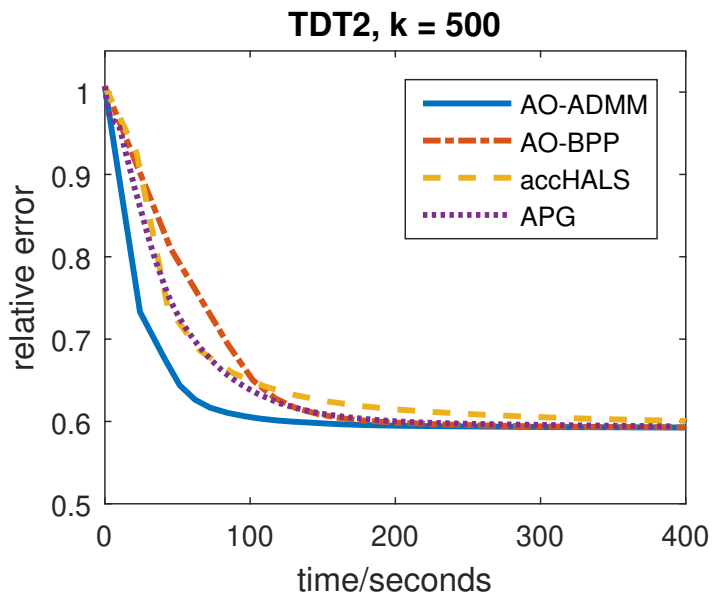


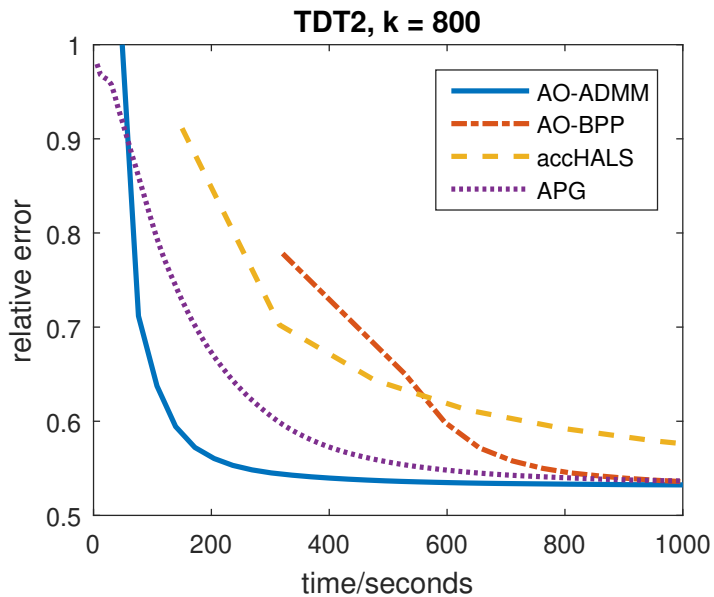


NMF

Topic Detection and Tracking 2 (TDT2):

- A sparse text corpus;
- 10212×36771.





Extensions

all-at-once updates:

- proximal Gauss-Newton [HF19b]

stochastic algorithms:

- block-randomized stochastic proximal gradient [FGWH20]

non-Euclidean losses:

- generalized Kullback-Liebler divergence [HG20]
- β-divergence [FBD09]
- α-divergence [CLKC08]

Nonnegative Matrix Factorization (NMF)

PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

- Sec. 1: Some theory of NMF
  - Geometric intepretation
  - Nonnegative rank
  - Computational complexity
- Sec. 2: Identifiability
  When are the factors W and H unique, up to permutations and scalings, in an exact NMF model $X = WH$?

Geometric interpretation

Geometric interpretation I: nested cones

Given $X = WH$ with $(W, H) \geq 0$,

Geometric interpretation I: nested cones

Given $X = WH$ with $(W, H) \geq 0$, for all j,

$$X(:, j) = WH(:, j) \ \subseteq \ \text{cone}\,(W) = \{x \mid x = Wh, h \geq 0\} \ \subseteq \ \mathbb{R}^r_+.$$

Equivalently,

$$\text{cone}\,(X) \ \subseteq \ \text{cone}\,(W) \ \subseteq \ \mathbb{R}^r_+.$$

Geometric interpretation I: nested cones

Given $X = WH$ with $(W, H) \geq 0$, for all j,

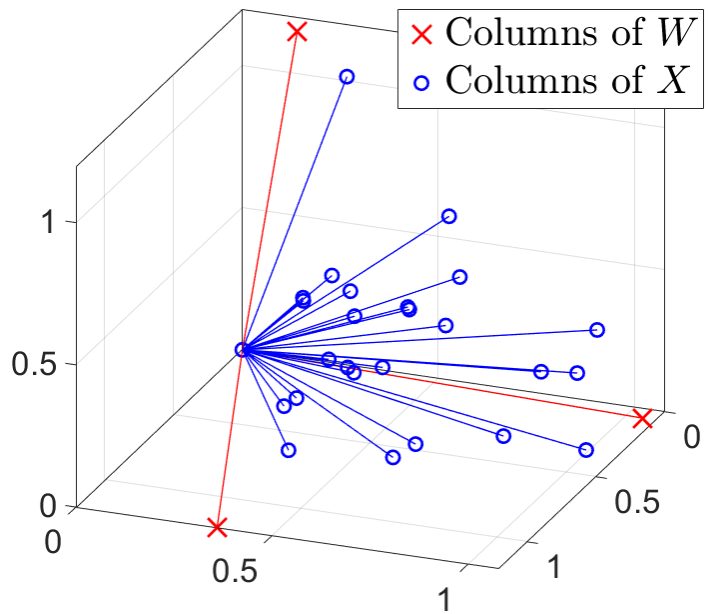
$$X(:, j) = WH(:, j) \ \subseteq \ \text{cone}\,(W) = \{x \mid x = Wh, h \geq 0\} \ \subseteq \ \mathbb{R}^r_+.$$

Equivalently,

$$\text{cone}\,(X) \ \subseteq \ \text{cone}\,(W) \ \subseteq \ \mathbb{R}^r_+.$$

Exact NMF is a nested cone problem:

Given two cones, $\text{cone}\,(X) \subseteq \mathbb{R}^r_+$, find a nested cone, $\text{cone}\,(W)$, with minimum number of extreme rays.



Geometric interpretation II: nested polytopes

Given $X = WH$, let us scale X and W to be column stochastic: $e^\top X = e^\top$ and $e^\top W = e^\top$.

Geometric interpretation II: nested polytopes

Given $X = WH$, let us scale X and W to be column stochastic: $e^\top X = e^\top$ and $e^\top W = e^\top$. This implies

$$e^\top \;=\; e^\top X \;=\; e^\top WH$$

Geometric interpretation II: nested polytopes

Given $X = WH$, let us scale X and W to be column stochastic: $e^\top X = e^\top$ and $e^\top W = e^\top$. This implies

$$e^\top \;=\; e^\top X \;=\; e^\top WH \;=\; e^\top H,$$

and hence H is column stochastic.

Geometric interpretation II: nested polytopes

Given $X = WH$, let us scale X and W to be column stochastic: $e^\top X = e^\top$ and $e^\top W = e^\top$. This implies

$$e^\top \;=\; e^\top X \;=\; e^\top WH \;=\; e^\top H,$$

and hence H is column stochastic.
The columns of X are convex combinations of the columns of W: for all j

$$X(:,j) = \sum_{k=1}^{r} W(:,k)\, H(k,j) \quad \text{with} \quad e^\top H(:,j) = 1 \text{ and } H(:,j) \geq 0.$$

Geometric interpretation II: nested polytopes

Given $X = WH$, let us scale X and W to be column stochastic: $e^\top X = e^\top$ and $e^\top W = e^\top$. This implies

$$e^\top \;=\; e^\top X \;=\; e^\top WH \;=\; e^\top H,$$

and hence H is column stochastic.
The columns of X are convex combinations of the columns of W: for all j

$$X(:,j) = \sum_{k=1}^{r} W(:,k)\, H(k,j) \quad \text{with} \quad e^\top H(:,j) = 1 \text{ and } H(:,j) \geq 0.$$

In other terms,

$$\text{conv}\,(X) \quad \subseteq \quad \text{conv}\,(W) \quad \subseteq \quad \Delta^p,$$

where $\text{conv}\,(X)$ is the convex hull of the columns of X, and $\Delta^p = \{x \in \mathbb{R}^p \,|x \geq 0, e^\top x = 1\,\}$ is the unit simplex.

Geometric interpretation of Exact NMF

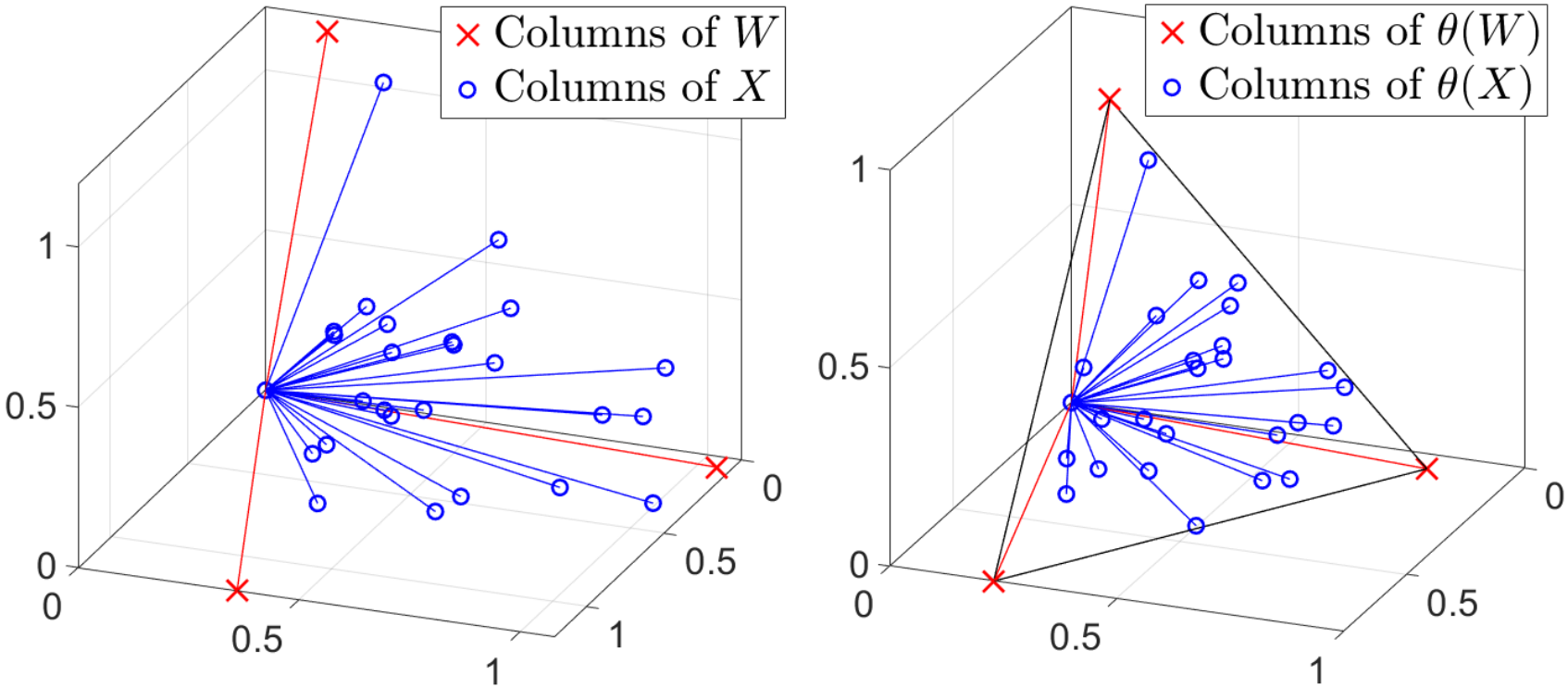


Figure 4: Left: nested convex cones, Right: nested convex hull on scaled data.

Geometric interpretation of NMF

Some observations:

- For $\operatorname{rank}(W) = r$ and $W^\top e = e$,
  $\operatorname{conv}(W)$ is an $(r-1)$-dimensional polytope.

Geometric interpretation of NMF

Some observations:

- For $\text{rank}(W) = r$ and $W^\top e = e$,
  $\text{conv}(W)$ is an $(r-1)$-dimensional polytope.

- If $\text{rank}(X) = r$, then $\text{rank}(W) = r$ and $\text{col}(W) = \text{col}(X)$,

$$\text{conv}(X) \quad \subseteq \quad \text{conv}(W) \quad \subseteq \quad \Delta^p \cap \text{col}(X),$$

  so that $\text{conv}(X)$ and $\text{conv}(W)$ have the same dimension, namely $\text{rank}(X) - 1$.

  This is called the nested polytope problem in computational geometry.

Nonnegative Rank

What is the nonnegative rank?

The nonnegative rank of a nonnegative matrix $X \in \mathbb{R}_+^{p \times n}$ is the minimum r such that an exact NMF exists:

$$X = WH = \sum_{k=1}^{r} W(:,k)H(k,:), \quad W \geq 0, H \geq 0.$$

What is the nonnegative rank?

The nonnegative rank of a nonnegative matrix $X \in \mathbb{R}_+^{p \times n}$ is the minimum r such that an exact NMF exists:

$$X = WH = \sum_{k=1}^{r} W(:,k)H(k,:), \quad W \geq 0, H \geq 0.$$

Geometrically: minimum number of vertices of a polytope nested between $\text{conv}(X)$ and Δ^p (where X is normalized).

What is the nonnegative rank?

The nonnegative rank of a nonnegative matrix $X \in \mathbb{R}_+^{p \times n}$ is the minimum r such that an exact NMF exists:

$$X = WH = \sum_{k=1}^{r} W(:,k)H(k,:), \quad W \geq 0, H \geq 0.$$

Geometrically: minimum number of vertices of a polytope nested between $\text{conv}(X)$ and Δ^p (where X is normalized).

The nonnegative rank of X is denoted $\text{rank}_+(X)$. Clearly,

$$\text{rank}(X) \quad \leq \quad \text{rank}_+(X) \quad \leq \quad \min(p, n).$$

Example on nested hexagons

Example. Let

$$X_a = \frac{1}{6a}\begin{pmatrix} 1 & a & 2a-1 & 2a-1 & a & 1 \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \\ 2a-1 & a & 1 & 1 & a & 2a-1 \\ 2a-1 & 2a-1 & a & 1 & 1 & a \\ a & 2a-1 & 2a-1 & a & 1 & 1 \end{pmatrix},$$

$$= \frac{1}{6a}\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}\begin{pmatrix} -1 & a-2 & -1 & 1-2a & 2-3a & 1-2a \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \end{pmatrix},$$

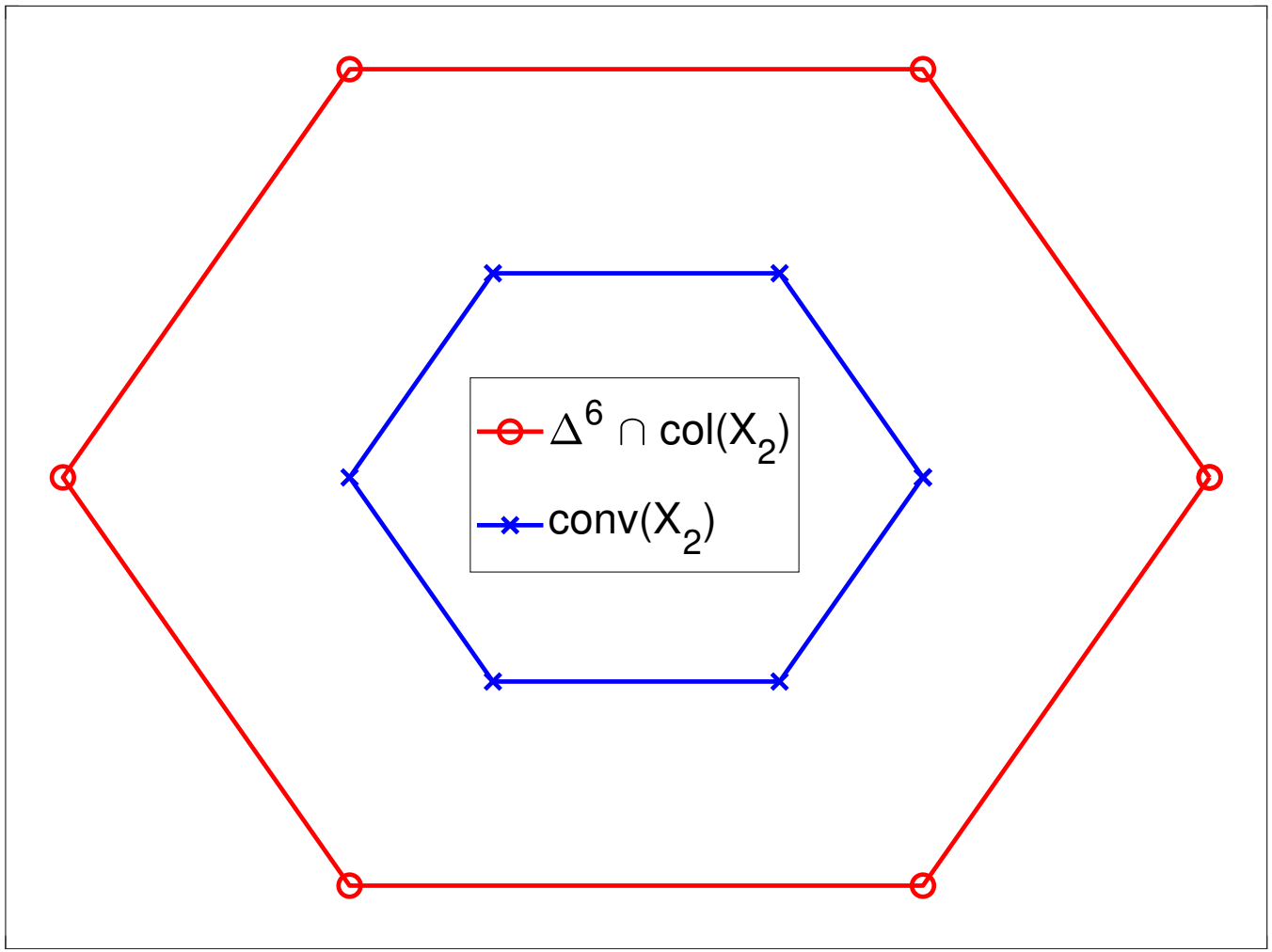
hence $\operatorname{rank}(X_a) = 3$ for all $a > 1$, so that $\operatorname{conv}(X_a)$ is a polygon.

In fact, it is an hexagon.

Example on nested hexagons

Example. Case 1: $a = 2$,

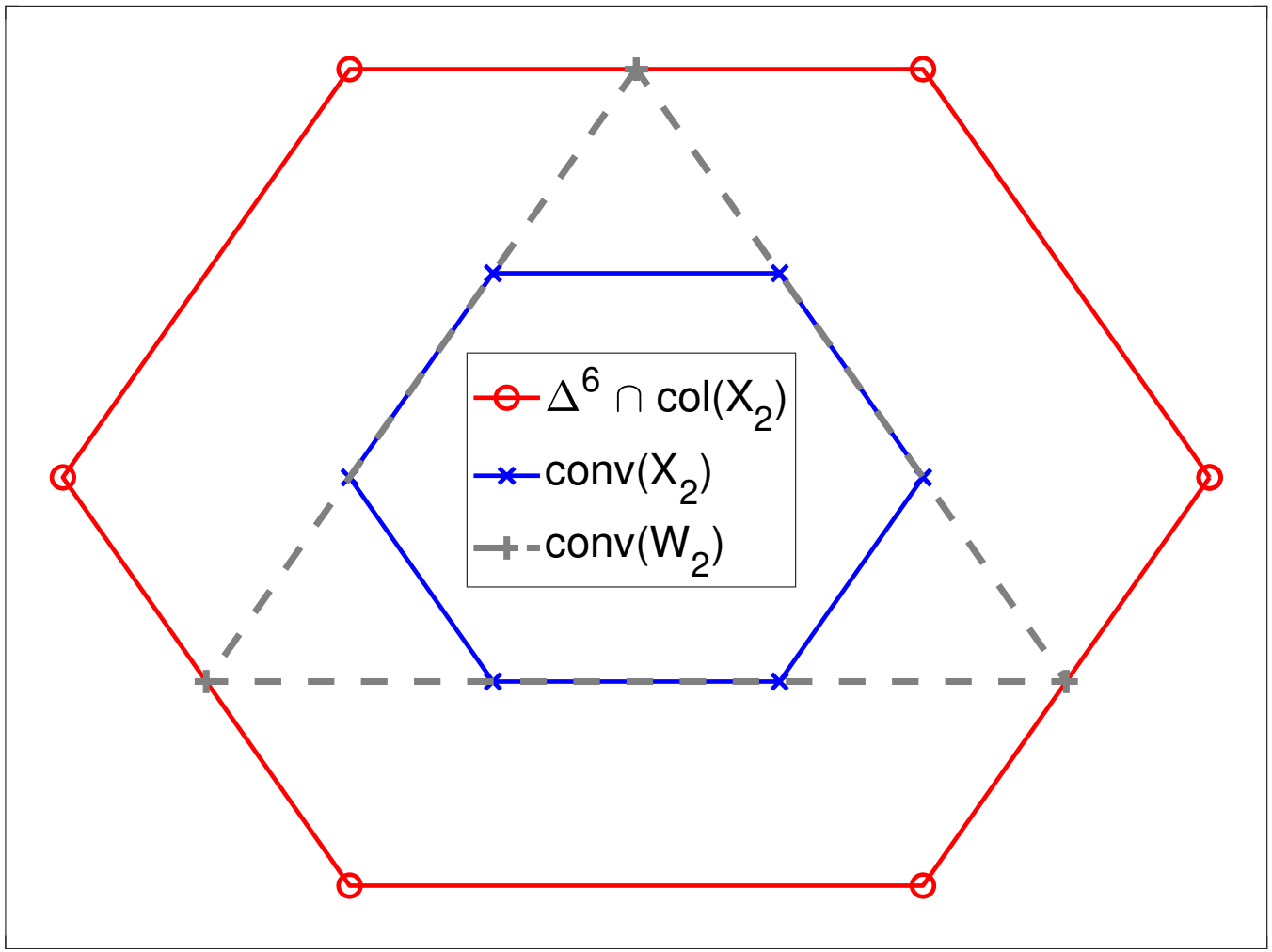
$$\mathrm{conv}\,(X) \subseteq \quad ? \quad \subseteq \mathrm{col}(X) \cap \Delta^p.$$



Example on nested hexagons

Example. Case 1: $a = 2$, we have $rank_+(X_a) = 3$, with

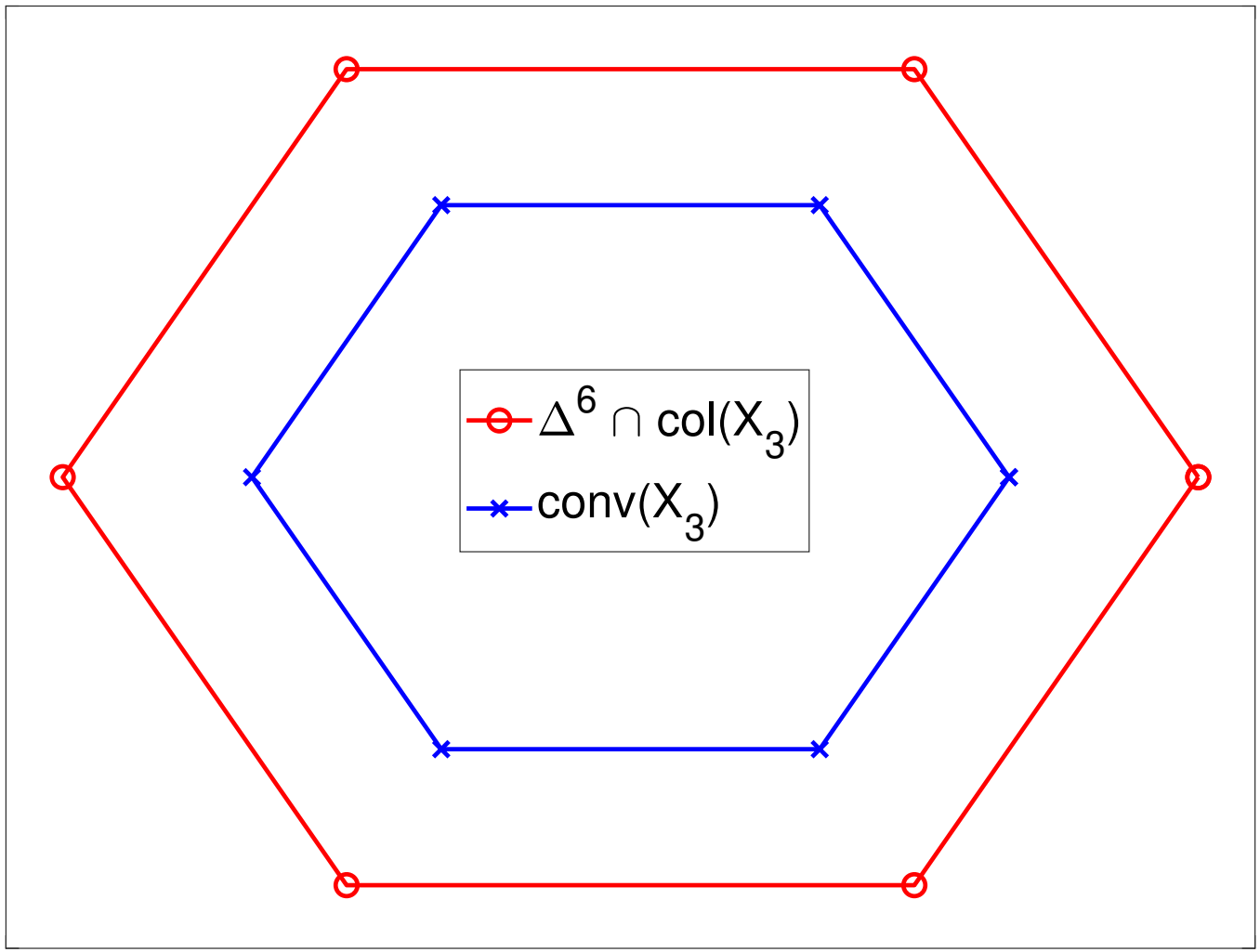
$$\mathrm{conv}(X) \subseteq \mathrm{conv}(W) \subseteq \mathrm{col}(X) \cap \Delta^p.$$



Example on nested hexagons

Case 2: $a = 3$,

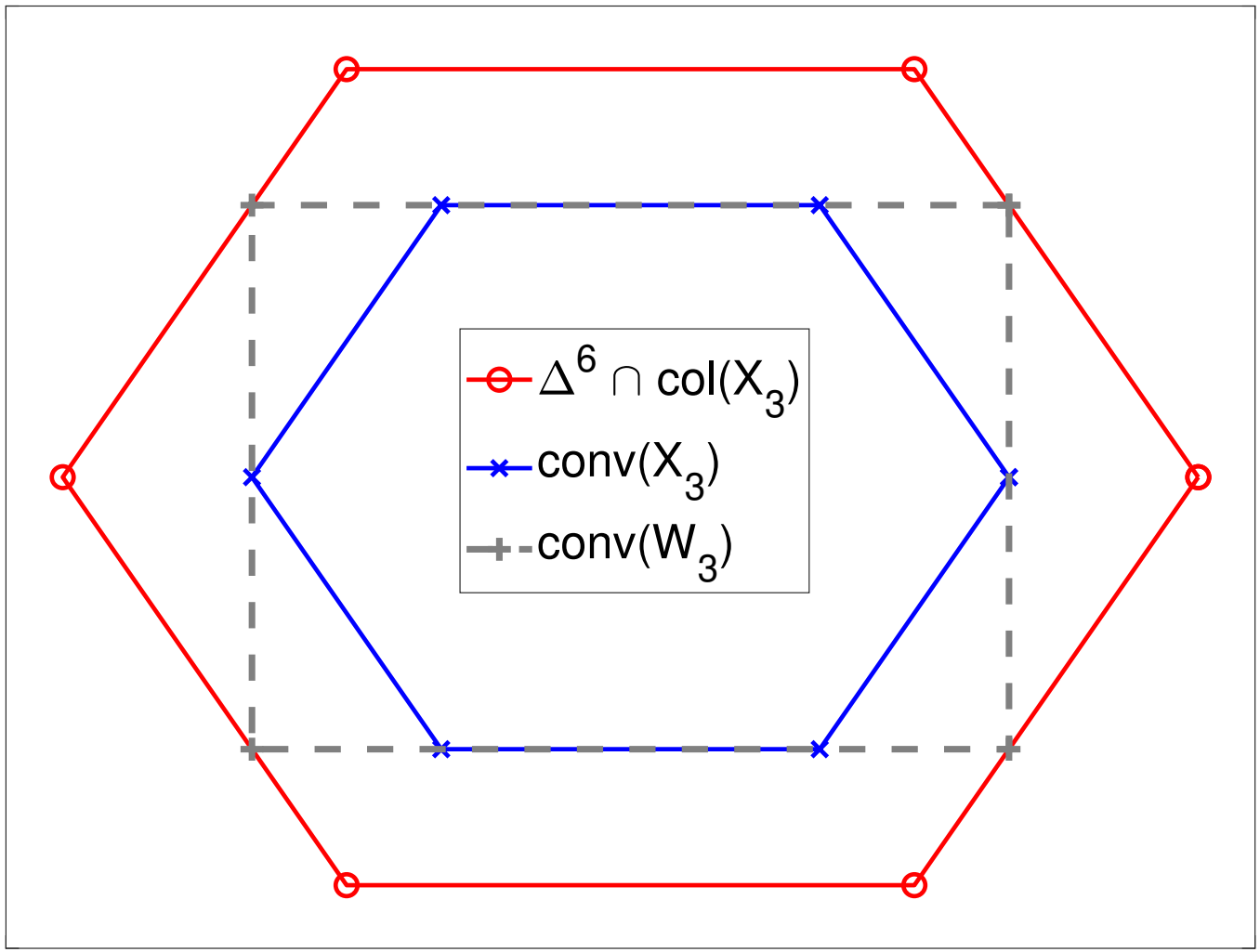
$$\mathrm{col}(X) \subseteq \quad ? \quad \subseteq \mathrm{col}(X) \cap \Delta^p.$$



Example on nested hexagons

Case 2: $a = 3$, we have $rank_+(X_a) = 4$, with

$$\mathrm{col}(X) \subseteq \mathrm{conv}(W) \subseteq \mathrm{col}(X) \cap \Delta^p.$$



Example on nested hexagons

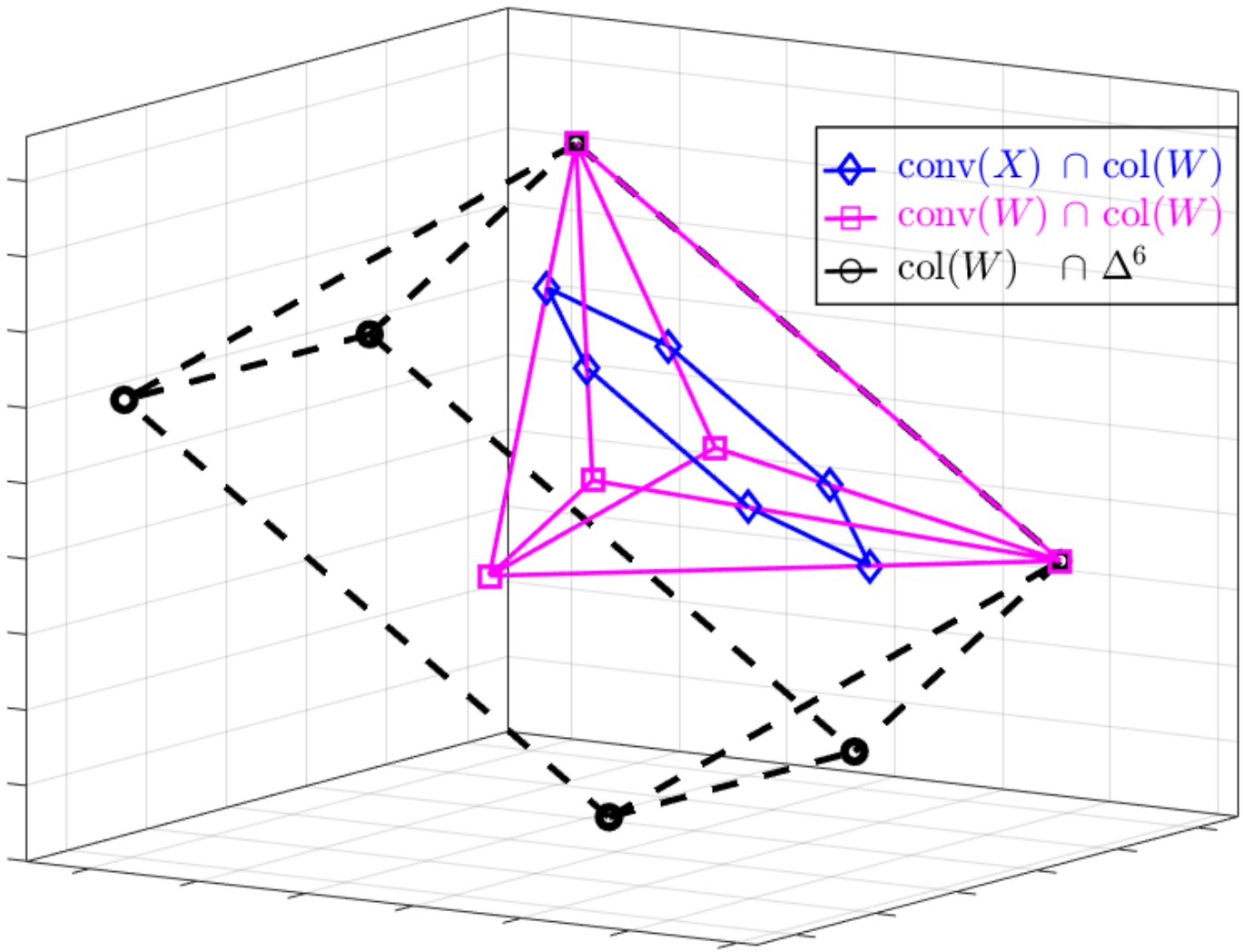
Example. Case 3: $a \to +\infty$, the inner and outer polygons coincide, no nested polygon with less than 6 vertices...

$$X_{+\infty} = \frac{1}{6} \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}.$$

What is the nonnegative rank of X?

Example on nested hexagons

Example. Case 3: $a \to +\infty$.
However, $\text{rank}_+(X_a) = 5$, because there exists $W \geq 0$ with $\text{rank}(W) = 4 > 3 = \text{rank}(X)$ such that $X = WH$, $H \geq 0$.



Factorization of the nested hexagon matrix

$$X_{+\infty} = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $4 = \text{rank}(W) > \text{rank}(X) = 3$, so that
$\text{conv}\,(W)$ has dimension 3, while $\text{conv}\,(X)$ has dimension 2.

An amazing result: NMF and extended formulations

Let $\mathcal{P}$ be a polytope

$$\mathcal{P} = \{x \in \mathbb{R}^k \mid b_i - A(i,:)x \geq 0 \text{ for } 1 \leq i \leq p\},$$

and let v_j's $(1 \leq j \leq n)$ be its vertices.

An amazing result: NMF and extended formulations

Let $\mathcal{P}$ be a polytope

$$\mathcal{P} = \{x \in \mathbb{R}^k \mid b_i - A(i,:)x \geq 0 \text{ for } 1 \leq i \leq p\},$$

and let v_j's $(1 \leq j \leq n)$ be its vertices.

We define the p-by-n slack matrix $S_{\mathcal{P}}$ of P as follows:

$$S_{\mathcal{P}}(i,j) = b_i - A(i,:)v_j \geq 0 \qquad 1 \leq i \leq p, 1 \leq j \leq n.$$

An amazing result: NMF and extended formulations

Let $\mathcal{P}$ be a polytope

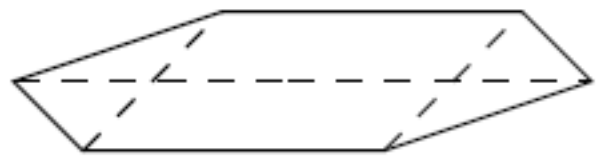
$$\mathcal{P} = \{x \in \mathbb{R}^k \mid b_i - A(i,:)x \geq 0 \text{ for } 1 \leq i \leq p\},$$

and let v_j's $(1 \leq j \leq n)$ be its vertices.

We define the p-by-n slack matrix $S_{\mathcal{P}}$ of P as follows:

$$S_{\mathcal{P}}(i,j) = b_i - A(i,:)v_j \geq 0 \qquad 1 \leq i \leq p, 1 \leq j \leq n.$$

The hexagon:



$$S_{\mathcal{P}} = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}$$

An amazing result: NMF and extended formulations

An extended formulation of $\mathcal{P}$ is higher dimensional polyhedron $Q \subseteq \mathbb{R}^{k+p}$ that (linearly) projects onto P. The minimum number of facets of such a polytope is called the extension complexity $\text{xp}(\mathcal{P})$ of $\mathcal{P}$.

An amazing result: NMF and extended formulations

An extended formulation of $\mathcal{P}$ is higher dimensional polyhedron $Q \subseteq \mathbb{R}^{k+p}$ that (linearly) projects onto P. The minimum number of facets of such a polytope is called the extension complexity $\mathrm{xp}(\mathcal{P})$ of $\mathcal{P}$.

Theorem [Yan91]

$$\mathrm{rank}_+(S_{\mathcal{P}}) = \mathrm{xp}(\mathcal{P}).$$

An amazing result: NMF and extended formulations

An extended formulation of $\mathcal{P}$ is higher dimensional polyhedron $Q \subseteq \mathbb{R}^{k+p}$ that (linearly) projects onto P. The minimum number of facets of such a polytope is called the extension complexity $\text{xp}(\mathcal{P})$ of $\mathcal{P}$.

Theorem [Yan91]

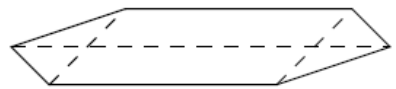
$$\text{rank}_+(S_{\mathcal{P}}) = \text{xp}(\mathcal{P}).$$

Application: limits of LP for solving combinatorial problems: given a polytope, what is the most compact way to represent it?
Its extension complexity = nonnegative rank of its slack matrix.

Key tool: lower bound techniques for the nonnegative rank.

Ex. The matching problem cannot be solved via a polynomial-size LP [Rot14]. (Fulkerson prize 2018)

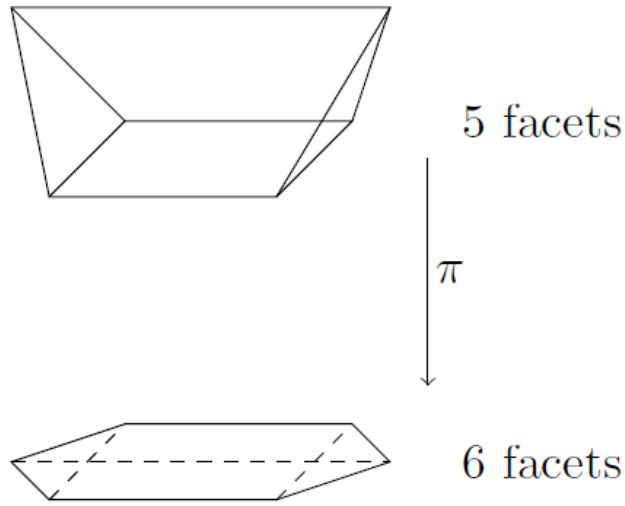
The Hexagon



6 facets

$$S_{\mathcal{P}} = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}$$

The Hexagon



$$S_{\mathcal{P}} = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 \end{pmatrix},$$

with

$$\text{rank}(S_{\mathcal{P}}) = 3 \quad \leq \quad \text{rank}_+(S_{\mathcal{P}}) = 5 \quad \leq \quad \min(m, n) = 6.$$

Bounding the nonnegative rank

Providing lower bounds for the nonnegative rank has been an important direction of research.

Bounding the nonnegative rank

Providing lower bounds for the nonnegative rank has been an important direction of research. Two simple examples:

- If the columns of X have different supports

$$\text{rank}_+(X) \geq \log_2(n). \qquad \text{(Goemans, 2008)}$$

Bounding the nonnegative rank

Providing lower bounds for the nonnegative rank has been an important direction of research. Two simple examples:

- If the columns of X have different supports

$$\text{rank}_+(X) \geq \log_2(n). \qquad \text{(Goemans, 2008)}$$

Example: For $X(i,j) = (i-j)^2$ for $i,j = 1,2,\ldots,n$, $\text{rank}(X) = 3$ while $\text{rank}_+(X) \geq \log_2(n)$.

Bounding the nonnegative rank

Providing lower bounds for the nonnegative rank has been an important direction of research. Two simple examples:

- If the columns of X have different supports

$$\text{rank}_+(X) \geq \log_2(n). \qquad \text{(Goemans, 2008)}$$

  Example: For $X(i,j) = (i-j)^2$ for $i,j = 1,2,\ldots,n$, $\text{rank}(X) = 3$ while $\text{rank}_+(X) \geq \log_2(n)$.

- The nonnegative rank is larger than the minimum number of binary rectangles needed to cover X; see [FKPT13].

Bounding the nonnegative rank

Providing lower bounds for the nonnegative rank has been an important direction of research. Two simple examples:

- If the columns of X have different supports

$$\operatorname{rank}_+(X) \geq \log_2(n). \qquad \text{(Goemans, 2008)}$$

  Example: For $X(i,j) = (i-j)^2$ for $i,j = 1,2,\ldots,n$, $\operatorname{rank}(X) = 3$ while $\operatorname{rank}_+(X) \geq \log_2(n)$.

- The nonnegative rank is larger than the minimum number of binary rectangles needed to cover X; see [FKPT13].

  Example (continued): for $n = 6$, $\operatorname{rank}_+(X) \geq 4$ since

$$\begin{pmatrix} 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 1&1&1&0&0&0\\ 1&1&1&0&0&0\\ 1&1&1&0&0&0 \end{pmatrix} + \begin{pmatrix} 0&0&0&0&0&0\\ 1&0&0&1&0&1\\ 1&0&0&1&0&1\\ 0&0&0&0&0&0\\ 1&0&0&1&0&1\\ 0&0&0&0&0&0 \end{pmatrix} + \begin{pmatrix} 0&1&0&0&1&1\\ 0&0&0&0&0&0\\ 0&1&0&0&1&1\\ 0&1&0&0&1&1\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0 \end{pmatrix}$$
$$+ \begin{pmatrix} 0&0&1&1&1&0\\ 0&0&1&1&1&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&1&1&1&0 \end{pmatrix} = \begin{pmatrix} 0&1&1&1&2&1\\ 1&0&1&2&1&1\\ 1&1&0&1&1&2\\ 1&2&1&0&1&1\\ 2&1&1&1&0&1\\ 1&1&2&1&1&0 \end{pmatrix},$$

Computational Complexity

Complexity of NMF

$$\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} ||M - WH||_F^2 \quad \text{such that } W \geq 0, H \geq 0.$$

- For $r = 1$, Eckart-Young and Perron-Frobenius theorems.

Complexity of NMF

$$\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} ||M - WH||_F^2 \quad \text{such that } W \geq 0, H \geq 0.$$

- For $r = 1$, Eckart-Young and Perron-Frobenius theorems. Polynomial-time algorithms for the Kullback-Leibler divergence [HVD08], for ℓ_∞ [GS19].

Complexity of NMF

$$\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} ||M - WH||_F^2 \quad \text{such that } W \geq 0, H \geq 0.$$

- For $r = 1$, Eckart-Young and Perron-Frobenius theorems. Polynomial-time algorithms for the Kullback-Leibler divergence [HVD08], for ℓ_∞ [GS19].
- Checking whether there exists an exact factorization $X = WH$:
  - NP-hard where p, n and r are not fixed [Vav10].

Complexity of NMF

$$\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} ||M - WH||_F^2 \quad \text{such that } W \geq 0, H \geq 0.$$

- For $r = 1$, Eckart-Young and Perron-Frobenius theorems. Polynomial-time algorithms for the Kullback-Leibler divergence [HVD08], for ℓ_∞ [GS19].
- Checking whether there exists an exact factorization $X = WH$:
  - NP-hard where p, n and r are not fixed [Vav10].
  - Using quantifier elimination:
    - [CR93]: $(pn)^{O(pr+nr)}$, non-polynomial
    - [AGKM12]: $(pn)^{O(2^r)}$, polynomial in p and n
    - [Moi13]: $(pn)^{O(r^2)}$, polynomial in p and n
      $\rightarrow$ not really useful in practice ...

Complexity of NMF

$$\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} ||M - WH||_F^2 \quad \text{such that } W \geq 0, H \geq 0.$$

- For $r = 1$, Eckart-Young and Perron-Frobenius theorems. Polynomial-time algorithms for the Kullback-Leibler divergence [HVD08], for ℓ_∞ [GS19].
- Checking whether there exists an exact factorization $X = WH$:
  - NP-hard where p, n and r are not fixed [Vav10].
  - Using quantifier elimination:
    - [CR93]: $(pn)^{O(pr+nr)}$, non-polynomial
    - [AGKM12]: $(pn)^{O(2^r)}$, polynomial in p and n
    - [Moi13]: $(pn)^{O(r^2)}$, polynomial in p and n
      $\rightarrow$ not really useful in practice ...
- Does not imply that rank_+ can be computed in polynomial time for fixed rank, because there are no upper bound on rank_+ that depend on the rank.

Complexity for other norms

$$\min_{w \in \mathbb{R}^p, h \in \mathbb{R}^n} \quad ||X - wh^T||_1 = \sum_{i,j} |X_{ij} - w_i h_j| \,. \qquad (\ell_1 \text{ norm})$$

Complexity for other norms

$$\min_{w \in \mathbb{R}^p, h \in \mathbb{R}^n} \quad ||X - wh^T||_1 = \sum_{i,j} |X_{ij} - w_i h_j| \,. \qquad (\ell_1 \text{ norm})$$

If X is binary, $X \in \{0,1\}^{p \times n}$, any optimal solution (w^*, h^*) can be assumed to be binary, that is, $(w^*, h^*) \in \{0,1\}^p \times \{0,1\}^n$ [GV18].

Complexity for other norms

$$\min_{w\in\mathbb{R}^p, h\in\mathbb{R}^n} \quad ||X - wh^T||_1 = \sum_{i,j} |X_{ij} - w_i h_j| \,. \qquad (\ell_1 \text{ norm})$$

If X is binary, $X \in \{0,1\}^{p\times n}$, any optimal solution (w^*, h^*) can be assumed to be binary, that is, $(w^*, h^*) \in \{0,1\}^p \times \{0,1\}^n$ [GV18].

$$\min_{w\in\mathbb{R}^p, h\in\mathbb{R}^n} \; ||X - wh^T||_P^2 = \sum_{i,j} P_{ij}(X - wh^T)_{ij}^2, \; \text{(weighted } \ell_2 \text{ norm)}$$

where P is a nonnegative weight matrix. This model can be used when

- data is missing ($P_{ij} = 0$ for missing entries),
- entries have different variances ($P_{ij} = 1/\sigma_{ij}^2$).

Complexity for other norms

$$\min_{w\in\mathbb{R}^p, h\in\mathbb{R}^n} \quad ||X - wh^T||_1 = \sum_{i,j} |X_{ij} - w_i h_j| \,. \qquad (\ell_1 \text{ norm})$$

If X is binary, $X \in \{0,1\}^{p\times n}$, any optimal solution (w^*, h^*) can be assumed to be binary, that is, $(w^*, h^*) \in \{0,1\}^p \times \{0,1\}^n$ [GV18].

$$\min_{w\in\mathbb{R}^p, h\in\mathbb{R}^n} \; ||X - wh^T||_P^2 = \sum_{i,j} P_{ij}(X - wh^T)_{ij}^2, \; \text{(weighted } \ell_2 \text{ norm)}$$

where P is a nonnegative weight matrix. This model can be used when

- data is missing ($P_{ij} = 0$ for missing entries),
- entries have different variances ($P_{ij} = 1/\sigma_{ij}^2$).

The ℓ_1 and weighted rank-one approximation problems are NP-hard [GV18, GG11].

Nonnegative Matrix Factorization (NMF)

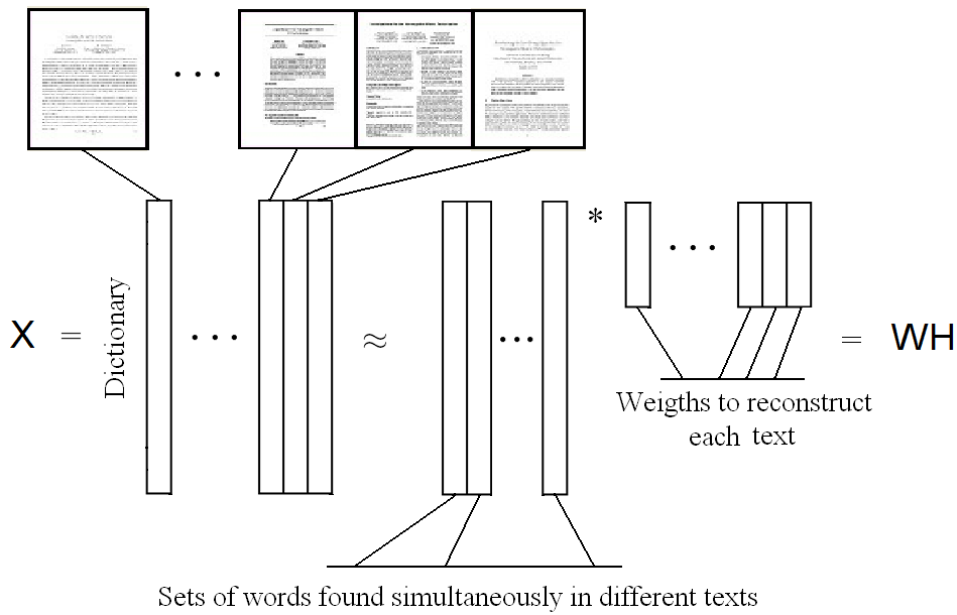
PART II – Theoretical aspects : understanding NMF, and using NMF meaningfully

- Sec. 1: Some theory of NMF
  - Geometric intepretation
  - Computational complexity
  - Nonnegative rank
- **Sec. 2: Identifiability**
  When are the factors W and H unique, up to permutations and scalings, in an exact NMF model $X = WH$?

Identifiability of NMF

Why care about identifiability/uniqueness of the factor matrices?

- In general, $X = WH$ is not unique.
  - $\widetilde{W} = WQ \geq 0$ and $\widetilde{H} = Q^{-1}H \geq 0$ give the same $X = \widetilde{W}\widetilde{H}$, where Q is any nonsingular matrix.
- Many problems are latent factor identification problems.



- If an NMF algorithm outputs WQ, it is not meaningful.

Identifiability

Definition of NMF Identifiability:

Consider a data matrix generated from the model

$$X = W_\natural H_\natural.$$

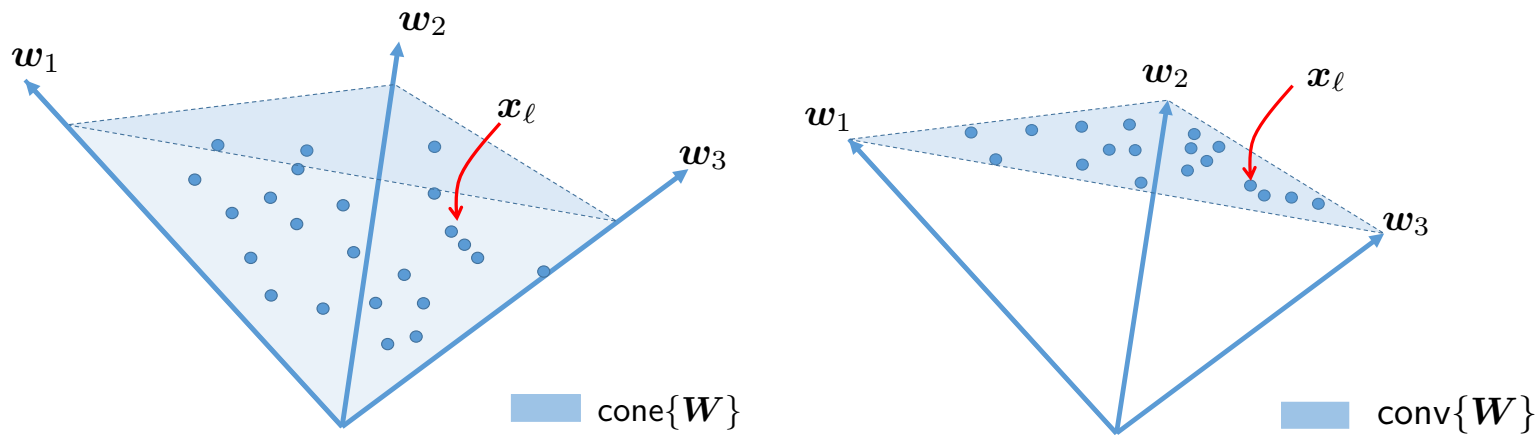
Suppose that $W_\natural$ and $H_\natural$ satisfy **a certain condition**. Let $(W_\star, H_\star)$ be obtained from an **identification method**. If

$$W_\star = W_\natural \Pi D, \quad H_\star = \Pi^T D^{-1} H_\natural,$$

where Π is a perm. matrix and D is a full-rank diagonal matrix, then the NMF model is **identifiable**.

- scaling and permutation are intrinsic.

Geometric View Revisited



- The dots are x_ℓ's, i.e., the columns of X.
- NMF identifiability = a geometric problem:

> Given the dots, can we recover the extremes rays (vertices) of the convex cone (convex hull) that enclose all the dots?

- Apparently no in general.

Separability-based NMF

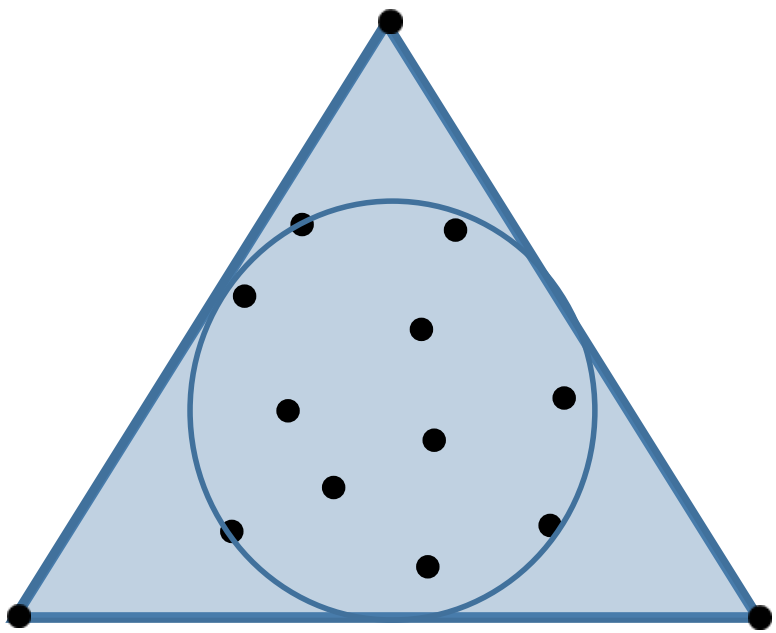
A Key Assumption

- **Separability** [DS03]

  For every $k = 1, ..., r$, there exists a row index ℓ_k such that $H(:, \ell_k) = h_{\ell_k} = \alpha_k e_k$, where $\alpha_k > 0$ is a scalar and e_k is the kth coordinate vector in $\mathbb{R}^r$.

  (There exists Λ such that $H(:, \Lambda) = \Sigma = Diag(\alpha_1, \ldots, \alpha_r)$.)

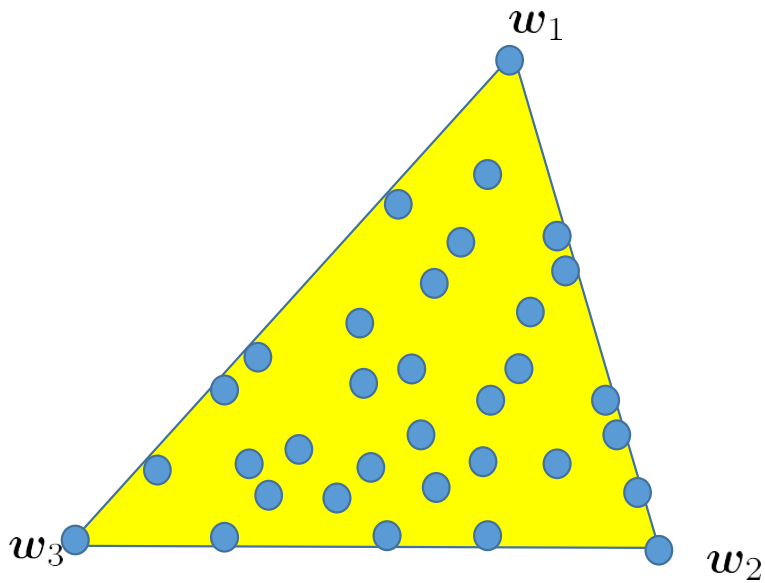
- Also means $\text{cone}(H) = \text{cone}\{e_1, \ldots, e_r\}$.



Separability-Based NMF

Under separability, the task of NMF is significantly simplified.

- Consider the case where $1^T H = 1^T$.
- Note that if $x_{\ell_k} = WH(:, \ell_k) = Wh_{\ell_k} = We_k = w_k$.

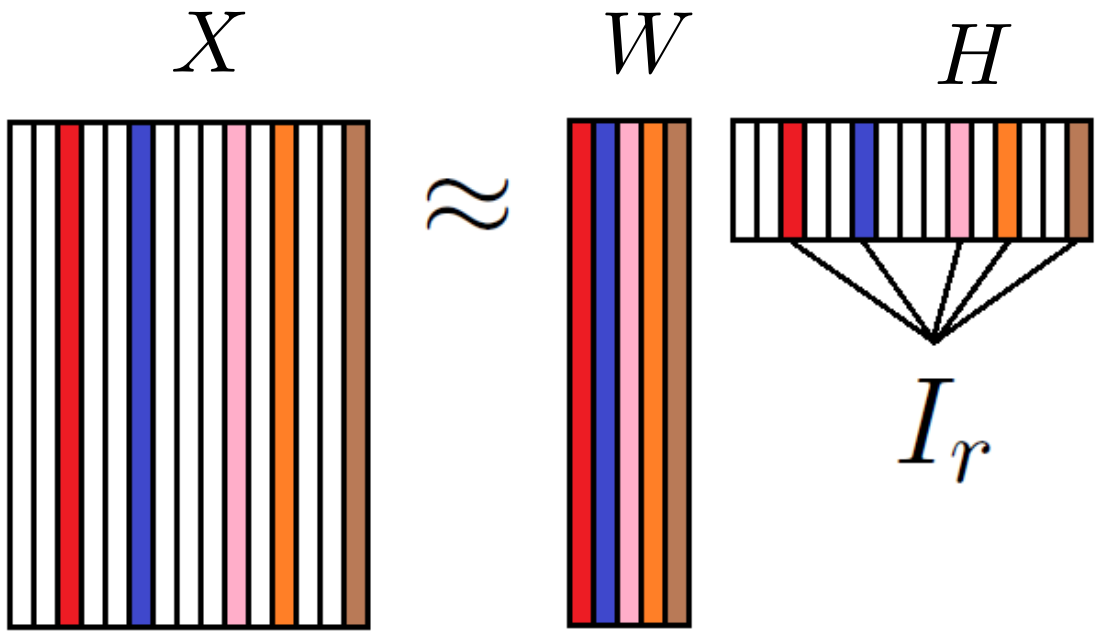


- Instead of estimating W and H from nowhere, we can simply find x_ℓ's that are vertices of the convex hull and let

$$\widehat{W} = [x_{\ell_1}, \ldots, x_{\ell_r}].$$

Separability-Based NMF

- Another view of separability.



A Gram-Schmidt-like Algorithm

- Assume $1^T H = 1^T$ and $H \geq 0$. First vertex of $\mathrm{conv}\,(W)$ can be found by

$$\widehat{\ell}_1 = \arg \max_{\ell=1,\ldots,n} \|x_\ell\|_2, \qquad \widehat{w}_1 = x_{\widehat{\ell}_1}$$

Proof:

$$\begin{aligned} \|x_\ell\|_2 &= \left\| \sum_{i=1}^{r} w_i h_{i,\ell} \right\|_2 \leq \sum_{i=1}^{r} \|w_i h_{i,\ell}\|_2 \\ &= \sum_{i=1}^{r} h_{i,\ell} \|w_i\|_2 \leq \max_{i=1,\ldots,r} \|w_i\|_2 \end{aligned}$$

where the equalities hold if and only if $h_\ell = e_i$ for some i.

- Assume that there are k ($k < r$) vertices have been found; we now wish to find the next one.
- First form $\widehat{W}_{1:k} = [\widehat{w}_1, \ldots, \widehat{w}_k]$.

$$\widehat{\ell}_{k+1} = \arg \max_{\ell=1,\ldots,n} \left\| P^{\perp}_{\widehat{W}_{1:k}} x_\ell \right\|_2 .$$

Proof:

$$\left\| P^{\perp}_{\widehat{W}_{1:k}} x_\ell \right\|_2 = \left\| \sum_{i=1}^{r} P^{\perp}_{\widehat{W}_{1:k}} w_i h_{i,\ell} \right\|_2 \leq \sum_{i=1}^{r} \left\| P^{\perp}_{\widehat{W}_{1:k}} w_i h_{i,\ell} \right\|_2$$

$$= \sum_{i=1}^{r} h_{i,\ell} \left\| P^{\perp}_{\widehat{W}_{1:k}} w_i \right\|_2 \leq \max_{i=k+1,\ldots,r} \left\| P^{\perp}_{\widehat{W}_{1:k}} w_i \right\|_2$$

- Called **successive projection algorithm** (SPA) [ASGa+01].
- Provably robust to bounded noise [GV14].

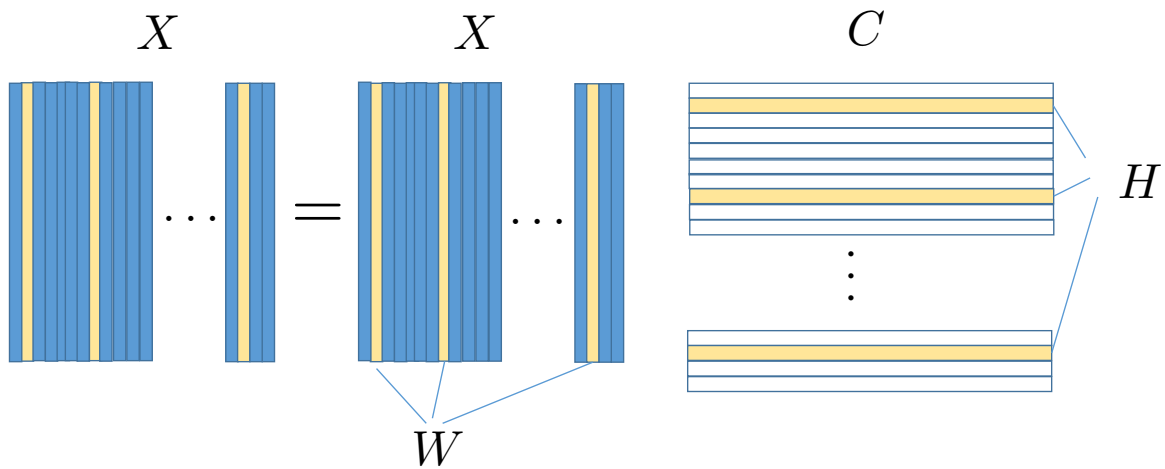
Other Separable NMF Methods

Other Separability-Based Approach (incomplete list)

- Sparse Optimization

$$\min_{C \in \mathbb{R}^{n \times n}} \|X - XC\|_F^2 + \lambda \text{RowSparse}(C)$$
$$s.t.\ C \geq 0;$$

see convex relaxation in [FM16, RRTB12, GL14].

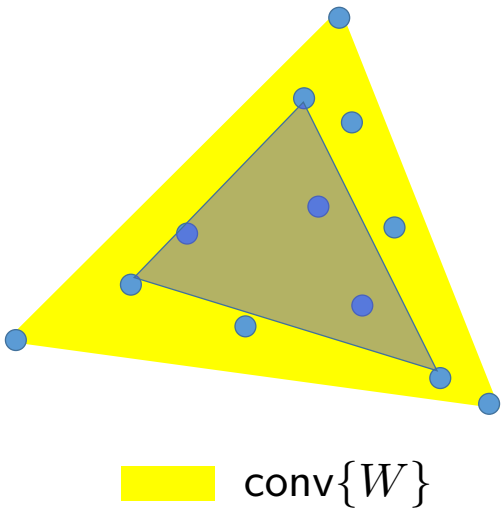


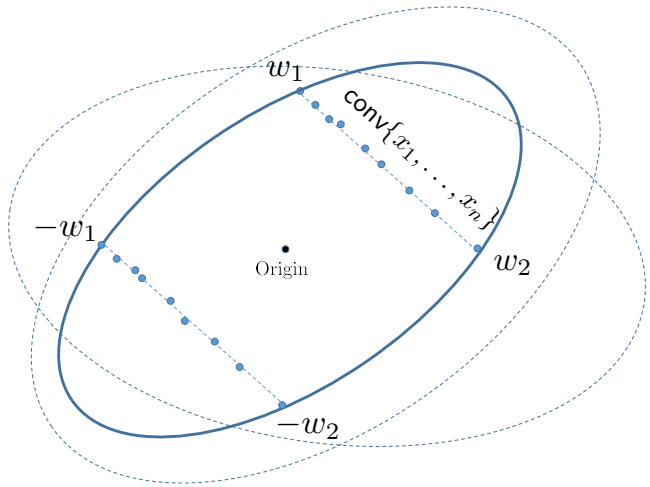
$$X = XC = WH$$

Other Separable NMF Methods

Other Separability-Based Approach (incomplete list)

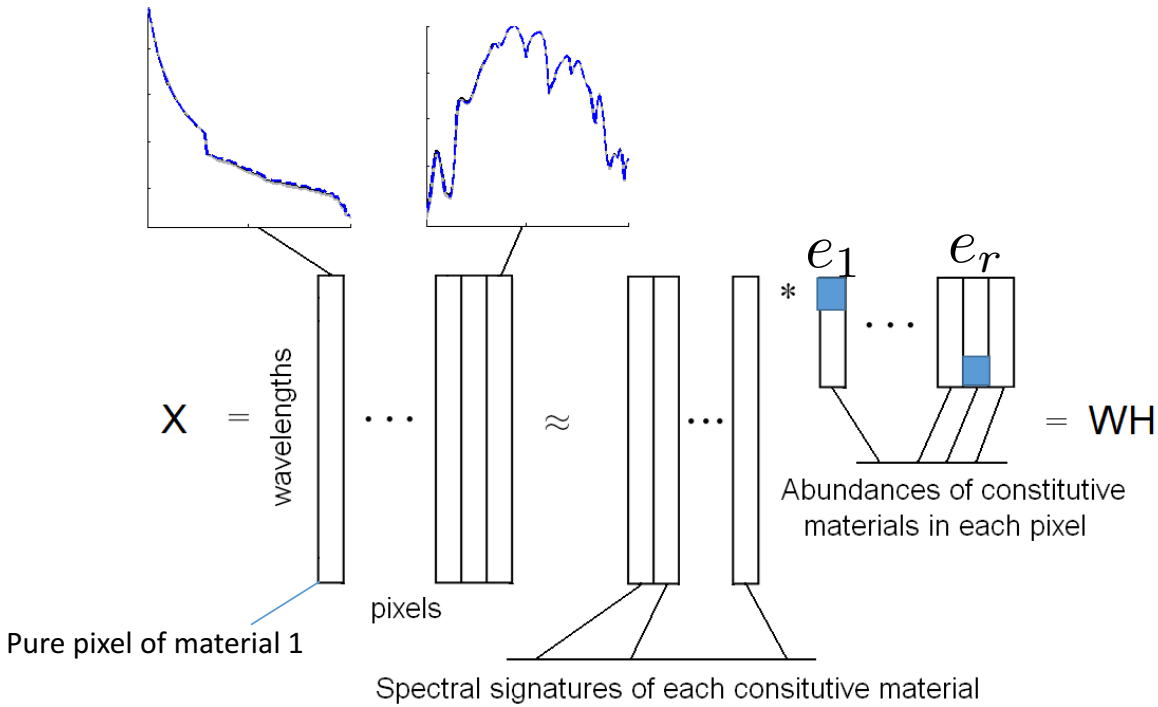
- Simplex-volume Maximization [Win99, CMAC11].
- Minimum-volume Enclosing Ellipsoid [Miz14, GV15].





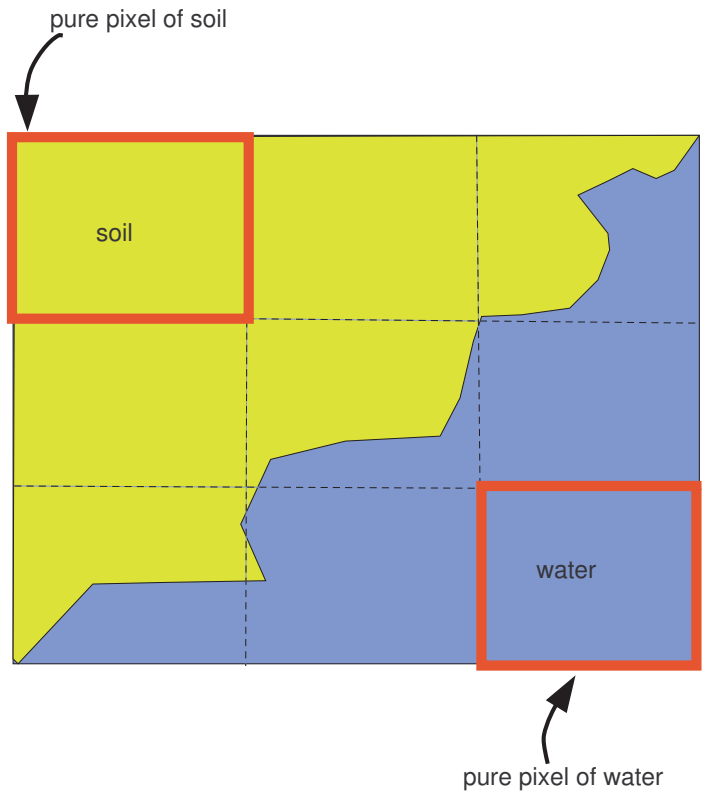
Separability and Applications

In HU, separability = pure pixels [BDPD+12, MBDC+14].



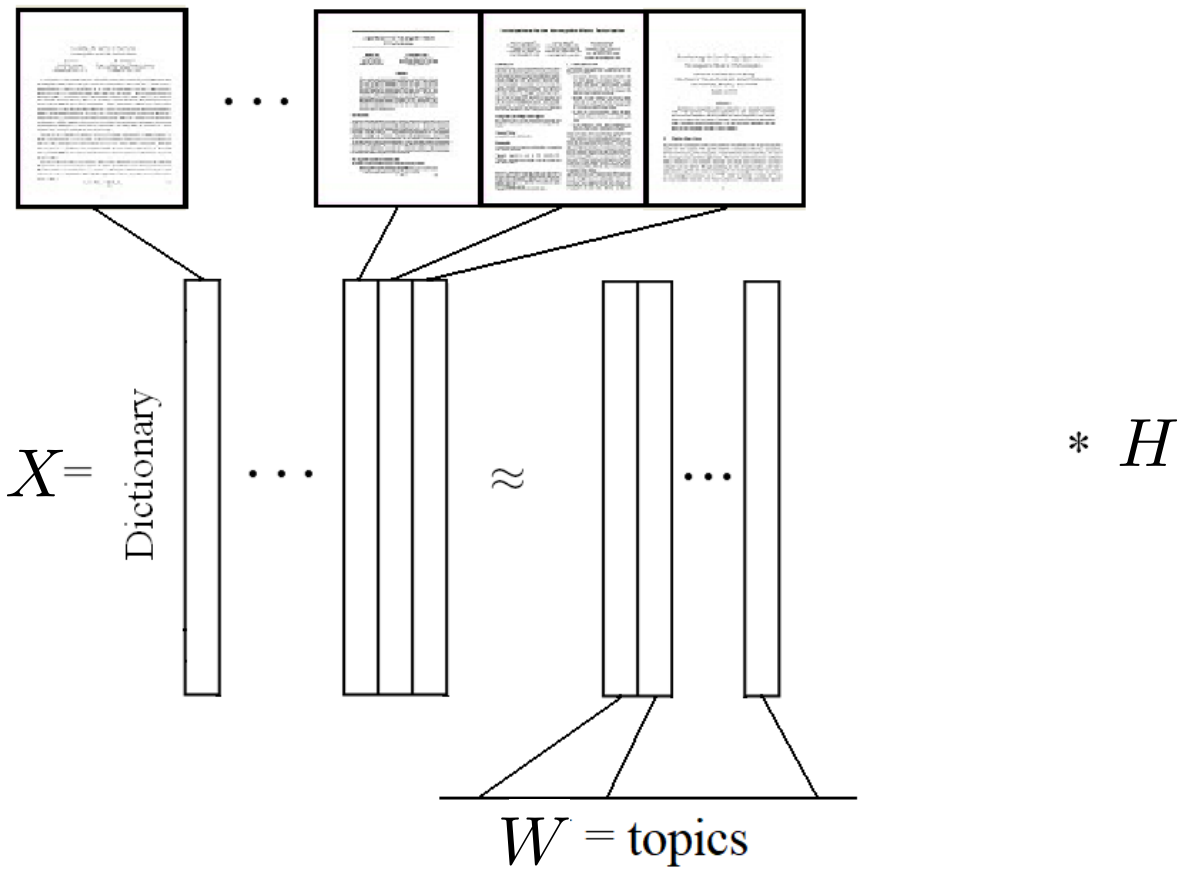
Separability and Applications

The remote sensing community started using "pure pixels" earlier (see [ASGa+01]) than "separability" formally defined by Donoho and Stodden in 2004 [DS03].



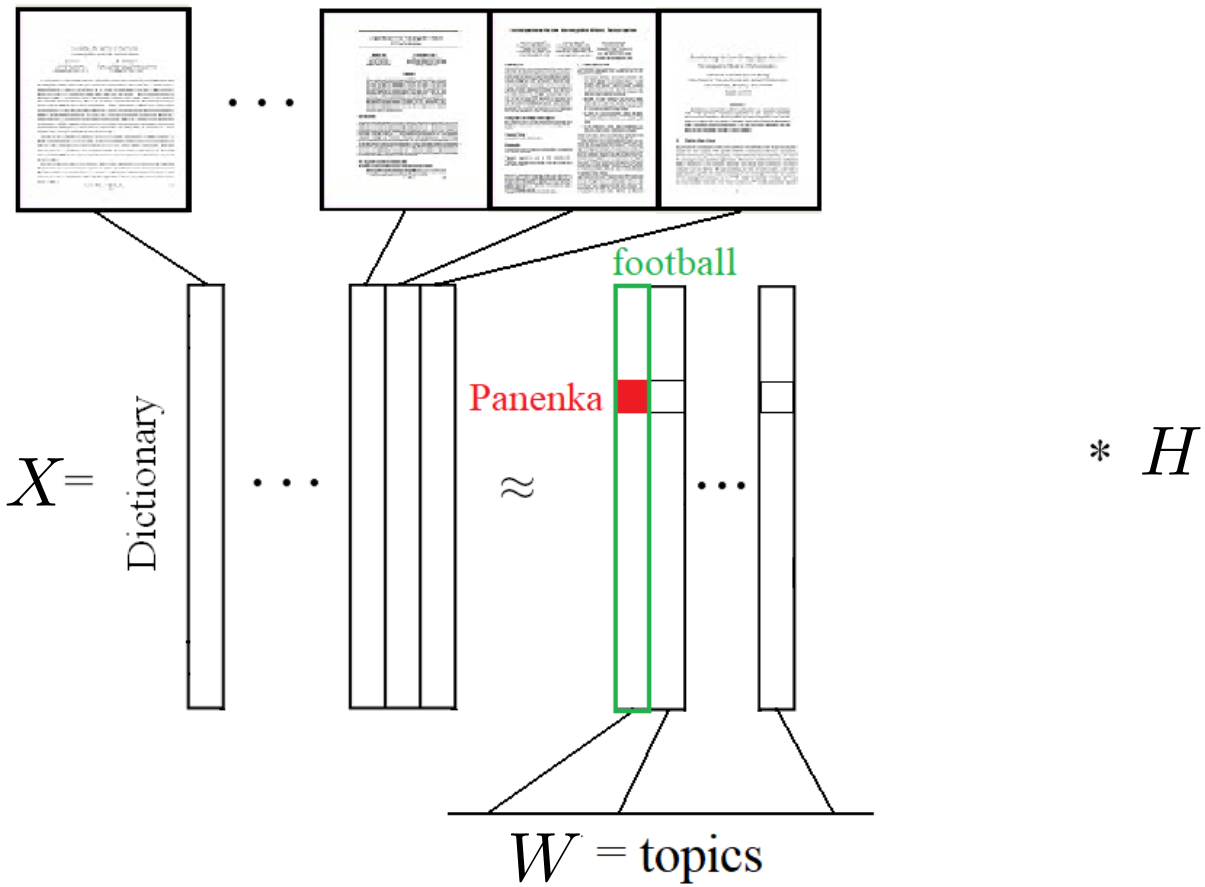
Separability and Applications

In topic modeling, separability = anchor words
[AGM12, AGH+13, RRTB12]



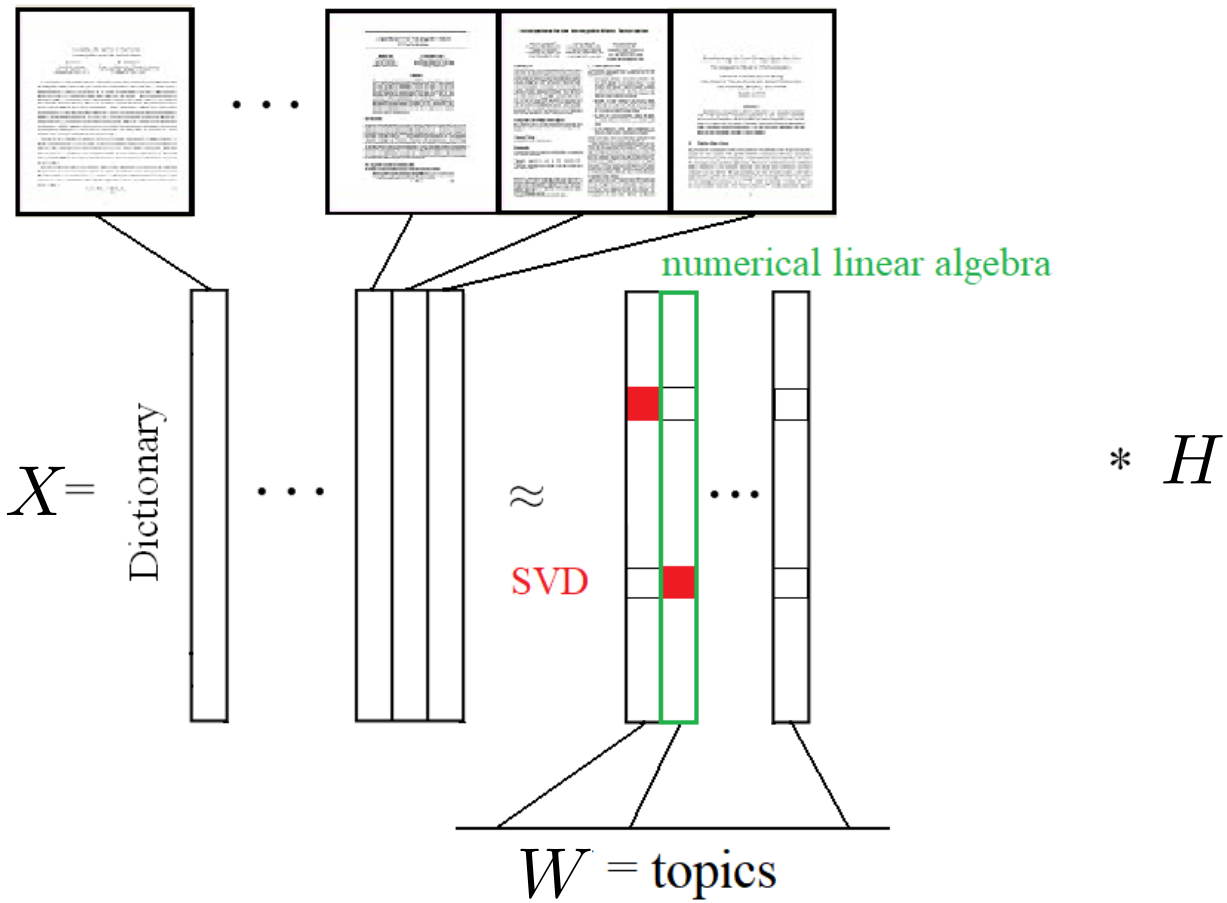
Separability and Applications

In topic modeling, separability = anchor words
[AGM12, AGH+13, RRTB12]



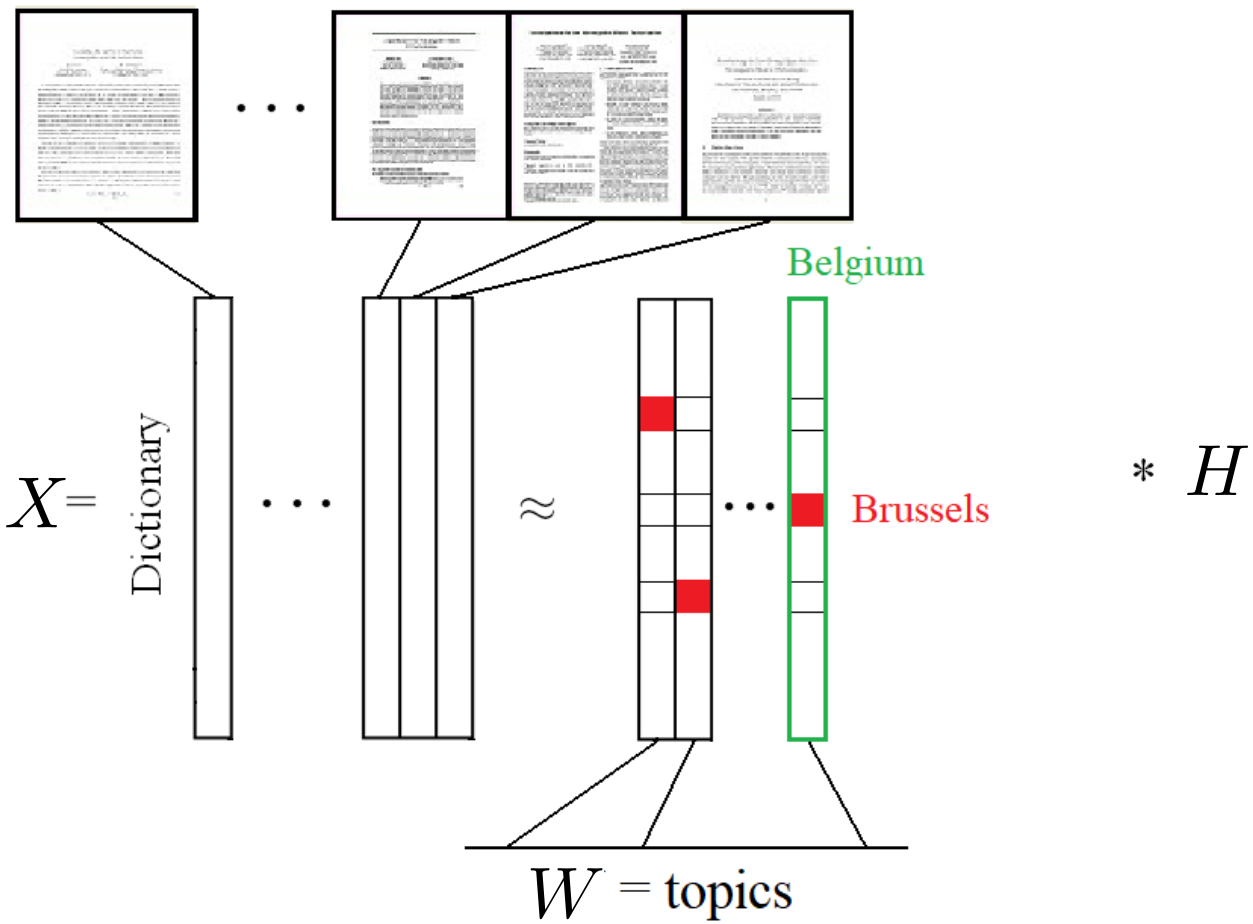
Separability and Applications

In topic modeling, separability = anchor words
[AGM12, AGH+13, RRTB12]



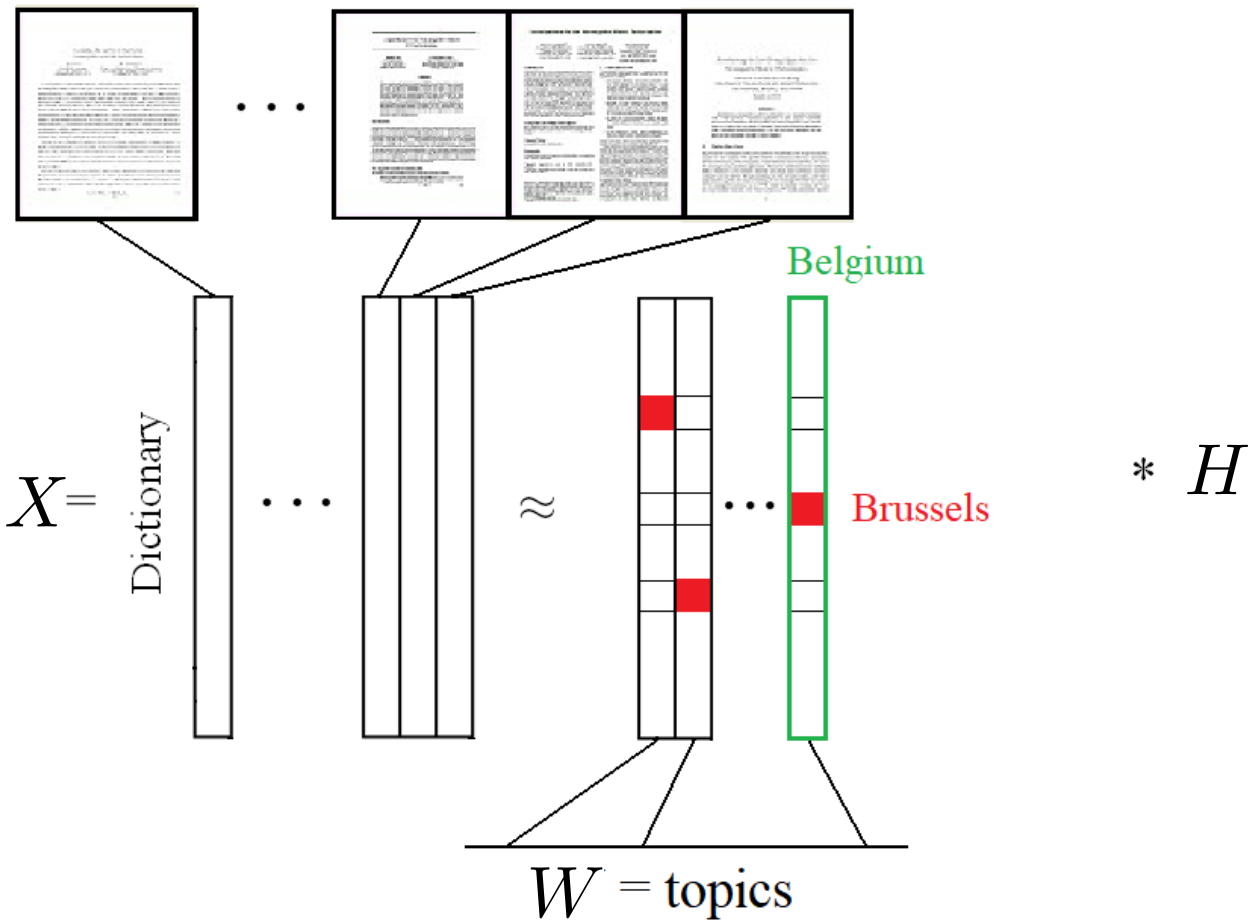
Separability and Applications

In topic modeling, separability = anchor words
[AGM12, AGH+13, RRTB12]



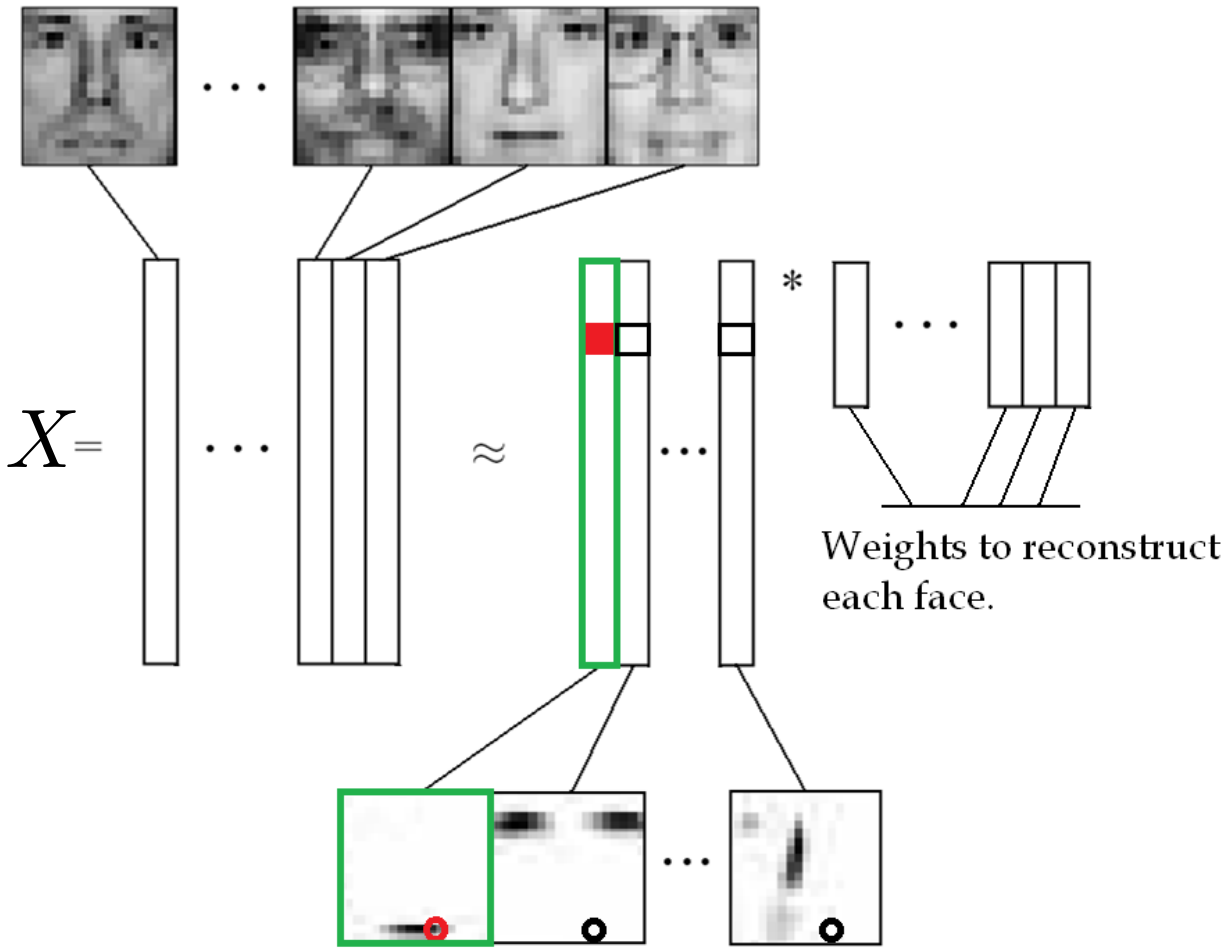
Separability and Applications

In image rep. learning, separability = local dominance



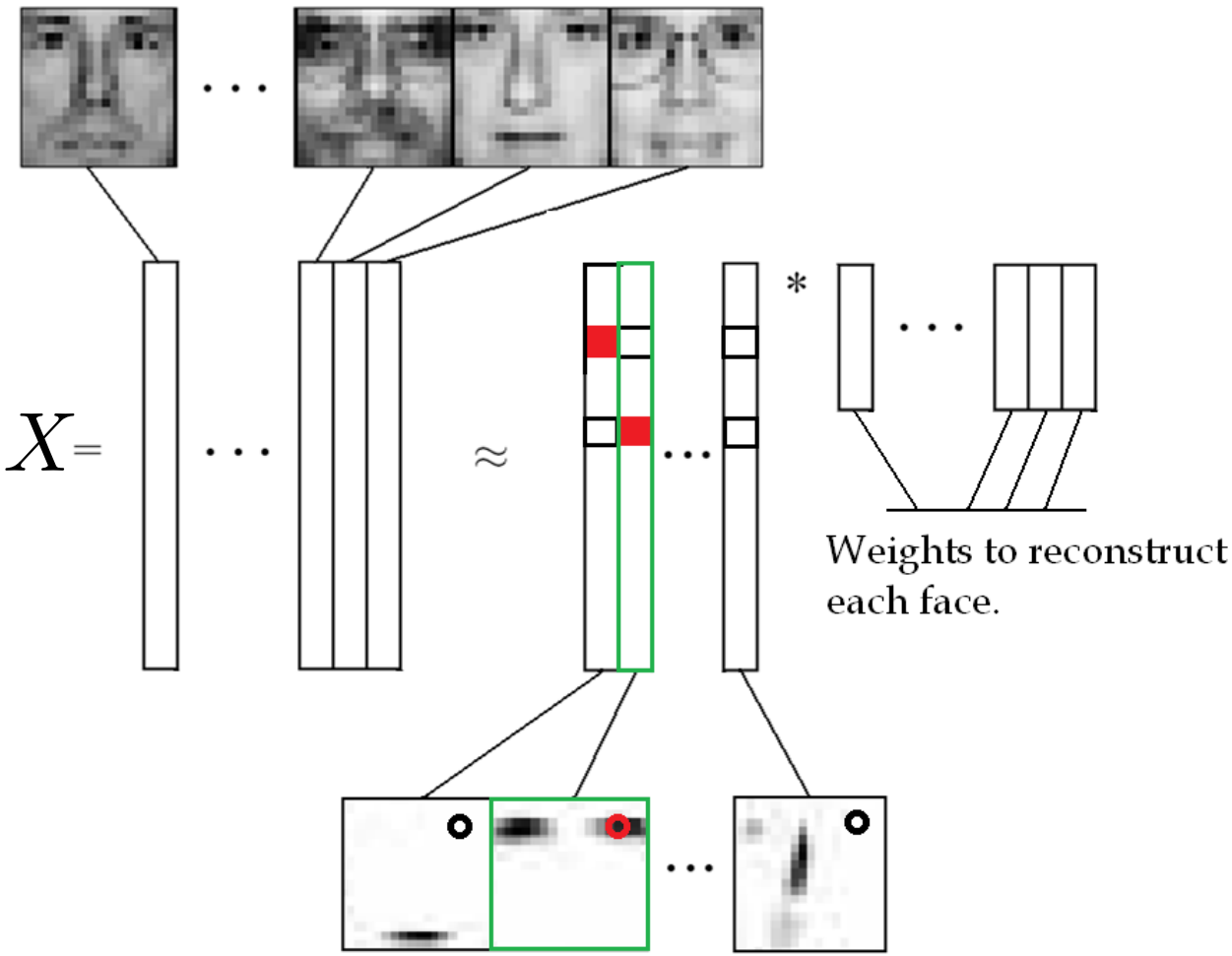
Separability and Applications

In image rep. learning, separability = local dominance



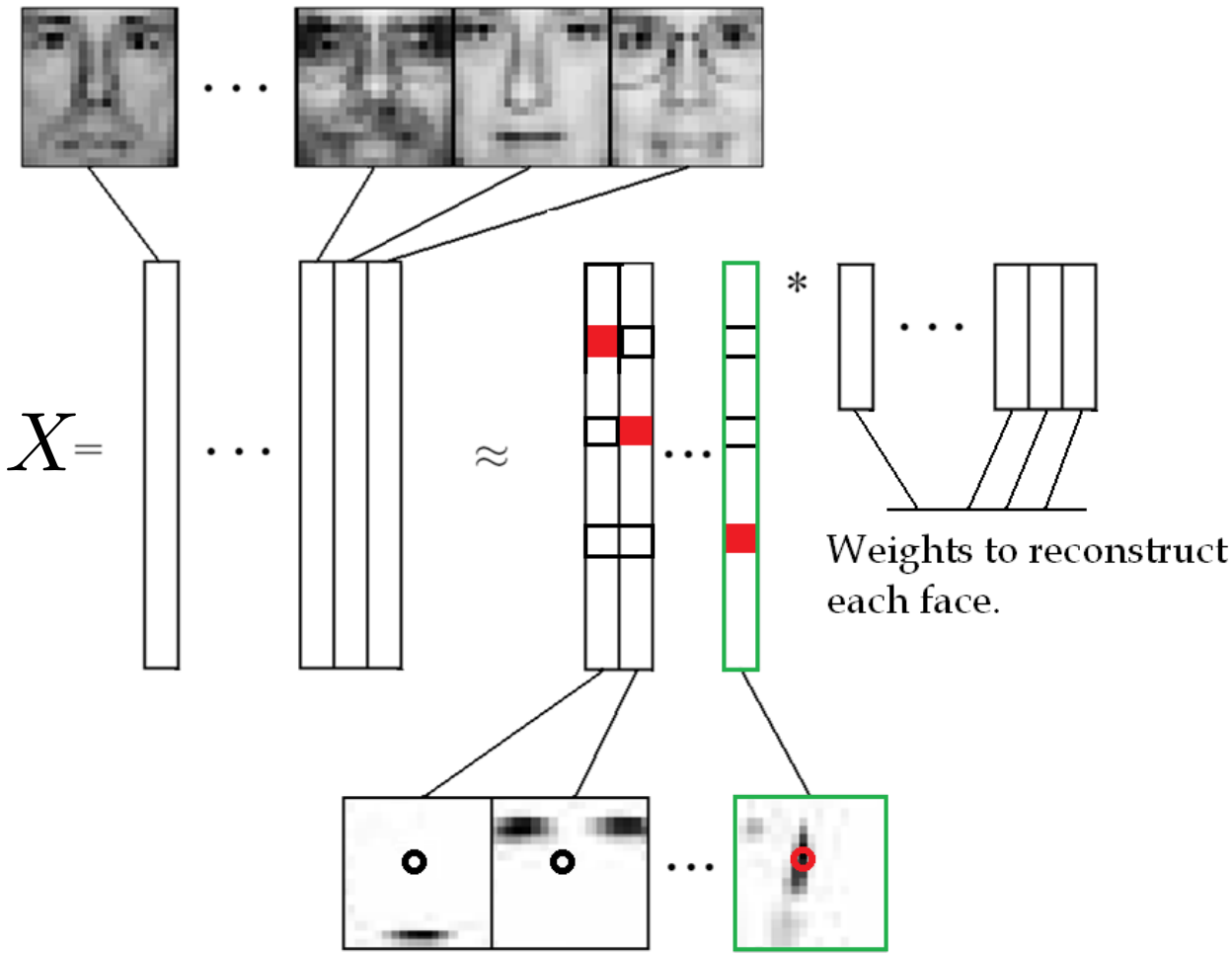
Separability and Applications

In image rep. learning, separability = local dominance



Separability and Applications

In image rep. learning, separability = local dominance



Separability and Applications

More in the literature ...

- In speech separation, separability could be induced by pauses between utterances [FMHS15].
- In crowdsourcing, separability = existence of class specialists [IFKH19].
- In community detection, separability = existence of anchor nodes [MSC17, PSU17].
- In spectrum sensing (comm.), separability = existence of designated carrier freq. [FMS16].
- ...

Separability and H's Size

In the crowdsourcing example, we mentioned that we hope H to be "fat" to attain separability.

- Assume $H(:,\ell) \in \mathbb{R}^r$ for $\ell = 1, \ldots, n$ are drawn from the prob. simplex uniformly at random.

Proposition [IFKH19] Let $\rho > 0, \varepsilon > 0$. If

$$n = \Omega\left(\frac{\varepsilon^{-2(r-1)}}{r} \log\left(\frac{r}{\rho}\right)\right),$$

then, with prob. of at least $1 - \rho$, there exist $H(:,\ell_k)$'s such that

$$\|H(\ell_k,:) - e_k^T\|_2 \leq \varepsilon, \ k = 1, \ldots, r.$$

Separability-free NMF

Plain Vanilla Fitting

- Most popular criterion:

$$\min_{W,H} \ \|X - WH\|_F^2$$
$$\text{s.t. } W \geq 0, \ H \geq 0.$$

- The above was the original formulation used in Lee and Seung's *Nature* article.
- If one finds the optimal solution $(W_\star, H_\star)$, do we have

$$W_\star = W_\natural \Pi D, \ H_\star = \Pi^T D^{-1} H_\natural$$

hold for all possible optimal solutions?

Beyond Separability

- **Sufficiently Scattered Condition (SSC)**

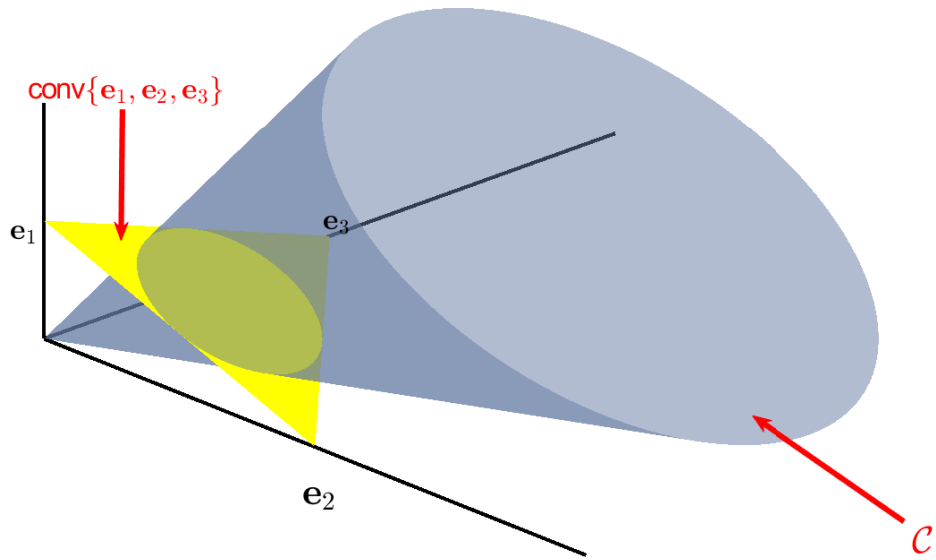
  A nonnegative matrix $H \in \mathbb{R}^{k \times n}$ is said to be sufficiently scattered if the following two conditions are satisfied:

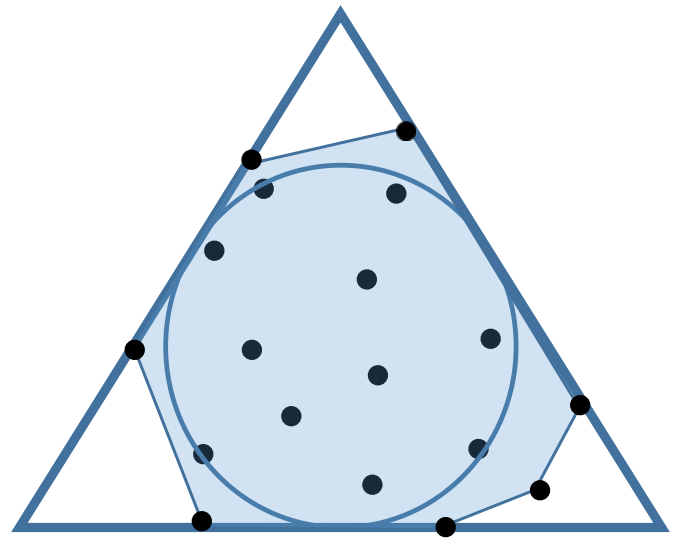
  1. the column of H are spread enough so that

     $$\mathcal{C} \subseteq \text{cone}\,(H)$$

     where $\mathcal{C}$ is a second-order cone defined as $\mathcal{C} = \{x \in \mathbb{R}^k \mid x^\top 1 \geq \sqrt{k-1}\|x\|_2\}$.
  2. $\text{cone}\{H\} \subseteq \text{cone}\{Q\}$ does not hold for any orthonormal Q except the permutation matrices.

A Closer Look at $\mathcal{C}$





Fitting-Based NMF

- [HSS14] proved the following theorem

> *If both $W_\natural^T$ and $H_\natural$ are sufficiently scattered, then the optimal solutions of the fitting criterion are $W_\star = W_\natural \Pi D$ and $H_\star = D^{-1}\Pi^T H_\natural$, where Π and D are defined as before.*

- **Notable**: the generative model does not need to satisfy the separability condition.
- **Remarks**:
  - unlike separability-based methods that come with tractable algorithms, the identifiability result of fitting-based NMF does not tell us how to really solve the fitting problem.
  - the fitting problem is in fact NP-hard.
  - necessary condition for identifiabity of fitting based NMF: both $W_\natural$ and $H_\natural$ have some zeros.

Sufficiently Scattered Condition

Proposition [HSS14]

If H is sufficiently scattered, each row of H contains at least $k-1$ zeros.

Random W and H with $m = 200$, $n = 250$, and $k = 30$.

Algorithm: AO-ADMM [HSL16]

density	max $\|\widehat{W} - W\|_F$	max $\|\widehat{H} - H\|_F$
0.5	0.0070×10^{-7}	0.0083×10^{-5}
0.6	0.0030×10^{-7}	0.0067×10^{-5}
0.7	0.0184×10^{-7}	0.0302×10^{-5}
0.8	0.1154×10^{-7}	0.1991×10^{-5}
0.9	0.4	78.2

Casual Claim

NMF $X = WH$ is unique if both W and H are sparse.

Sufficiently scattered on **one** of the factors

Theorem [FHS18]

Suppose $H_\natural$ is sufficiently scattered, and $\operatorname{rank}(W_\natural) = k$. Denote $(W_\star, H_\star)$ as an optimal solution of

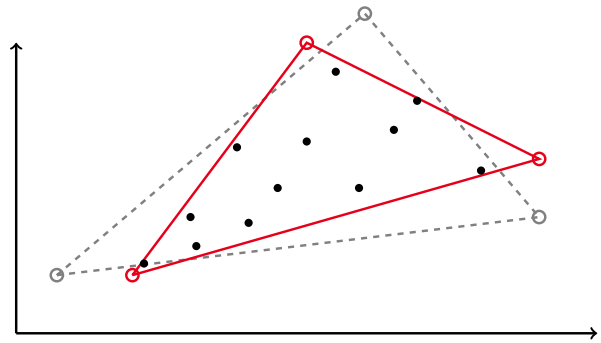
$$\begin{aligned} \underset{W,H}{\text{minimize}} \quad & \det(W^T W) \\ \text{subject to} \quad & X = WH, H \geq 0, H1 = 1 \end{aligned}$$

Then there exists a permutation matrix Π and a diagonal matrix D such that $W_\natural = W_\star \Pi D$ and $H_\natural = D^{-1} \Pi^T H_\star$.

- NMF is identifiable if only one of the factor is "sparse"
- W does not even have to be ≥ 0
- Identifiability can be enforced by a *det* criterion
- A slight variation resolves the long-standing *Craig's conjecture* in remote sensing

Geometric Interpretation

$$x_i = Wh_i = \sum_{j=1}^{k} w_j h_{ji} \qquad h_i \in \Delta$$



- x_i is a *convex combination* of $w_1, \ldots, w_k$
- x_i belongs to the *convex hull* of $w_1, \ldots, w_k$
- There are infinitely many enclosing simplexes

★ Intuition: Find the one with *minimum volume*

$$\underset{W,H}{\text{minimize}} \quad \frac{1}{(k-1)!} \left| \det \begin{bmatrix} w_1 - w_k & \cdots & w_{k-1} - w_k \end{bmatrix} \right|$$

$$\text{subject to} \quad X = WH, \quad H \geq 0, \quad 1^T H = 1.$$

Algorithm

$$\begin{aligned} \underset{W,H}{\text{minimize}} \quad & \det(W^T W) \\ \text{subject to} \quad & X = WH, H \geq 0, H1 = 1 \end{aligned}$$

- assume W is invertible (apply PCA first)
- change of variable $M = W^{-1}$
- eliminate H

$$\Downarrow$$

$$\begin{aligned} \underset{M}{\text{maximize}} \quad & |\det M| \\ \text{subject to} \quad & MX \geq 0, MX1 = 1 \end{aligned}$$

- convex (linear) constraints, disjoint on each row
- nonconvex objective

Cyclic Row Update

- let m_i denote the ith row of M
- Laplace's formula: $\det M = f^T m_i$ where

  $f_j = (-1)^{i+j} \det(M_{ij}), \quad M_{ij}$ is M without its ith row and jth col.

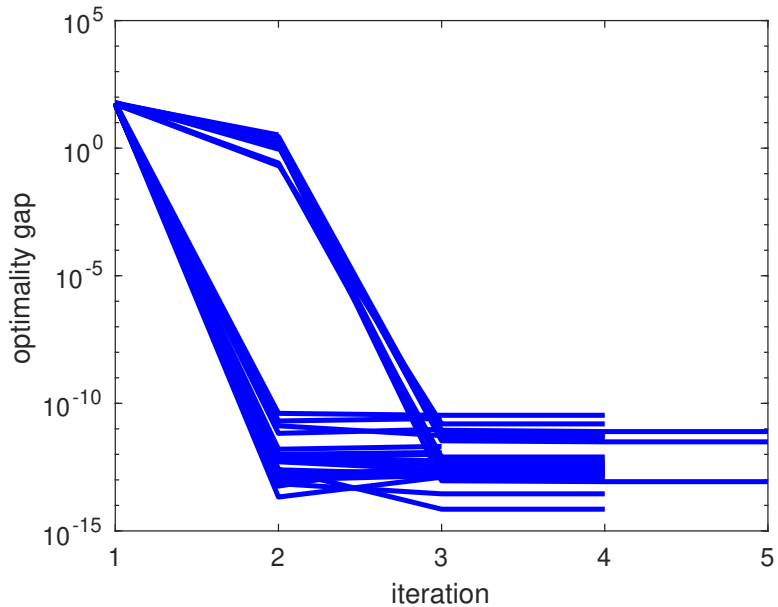
- in fact $f = (\det M) M^{-1} e_i$ per Cramer's rule
- cyclic row update [HF19a]

1: **repeat**
2: **for** $i = 1, \ldots, k$ **do**
3: $f \leftarrow M^{-1} e_i$
4: $z \leftarrow \arg\max_z f^T z$ s.t. $z^T X \geq 0, z^T X 1 = 1$
5: replace ith row of M with z
6: **end for**
7: **until** convergence

- monotonically increases the objective
- converges to a stationary point

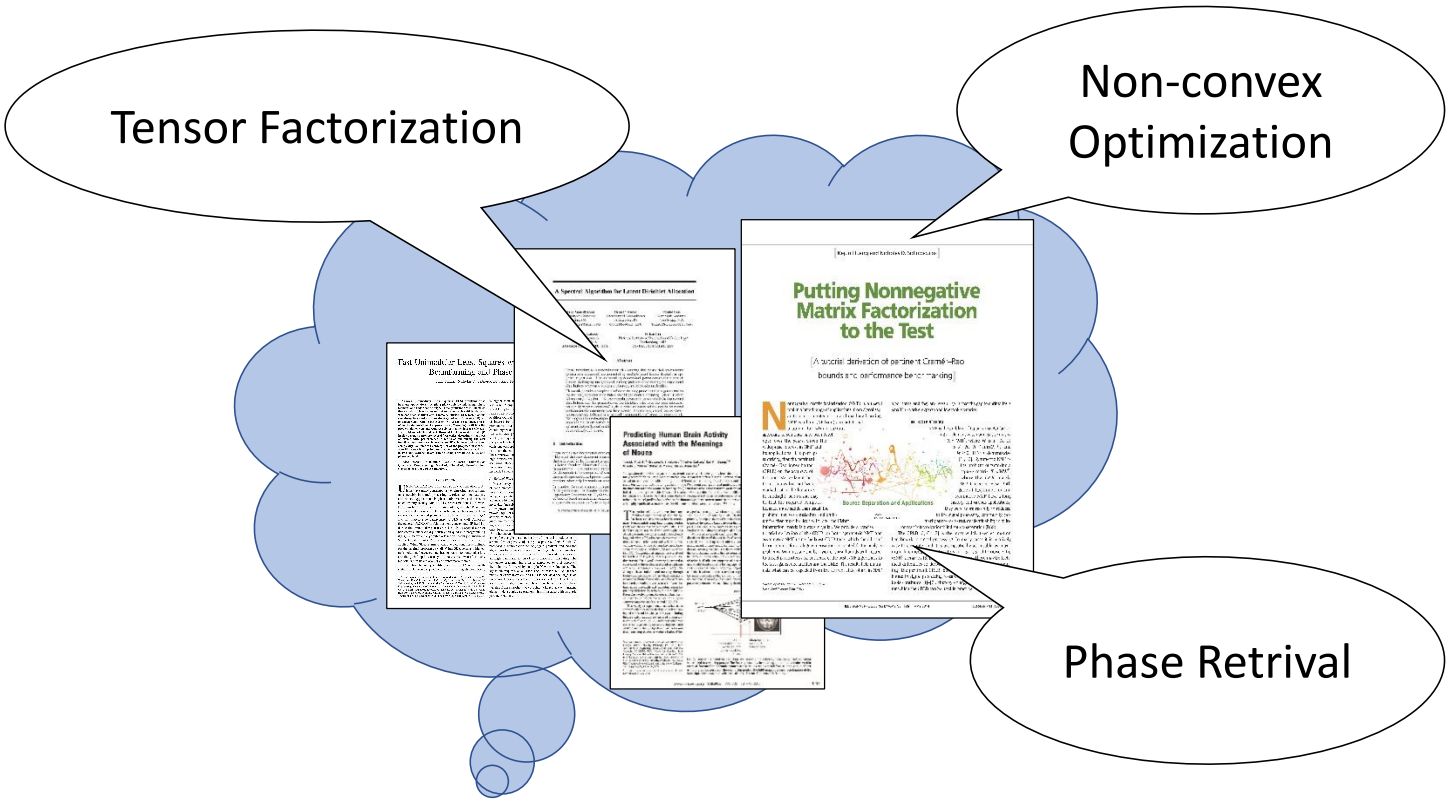
Synthetic Experiment

- $H_\natural \in \mathbb{R}^{20\times 1000}$ is 50% sparse (w.h.p. *sufficiently scattered*)
- $W_\natural \in \mathbb{R}^{20\times 20}$ is i.i.d. normal
- $X = W_\natural H_\natural$ is fed into the algorithm
- we know the optimal value is $(|\det W_\natural|)^{-1}$
  and the solution is unique (up to permutation)
- 50 random trials, optimality gap goes to zero in all cases



Application: Topic Modeling

Mining topics from a corpus [BNJ03]



5 topics in TDT2

Latent Dirichlet Allocation [BNJ03]

lewinsky	spkr	school	clinton	jordan
starr	news	gm	president	time
president	people	workers	house	game
white	monica	arkansas	white	bulls
house	story	people	clintons	night
ms	voice	children	public	team
grand	reporter	students	allegations	left
lawyers	washington	strike	presidents	i'm
jury	dont	jonesboro	political	chicago
jones	media	flint	bill	jazz

5 topics in TDT2

AnchorFree [HFS16]

lewinsky	gm	shuttle	bulls	jonesboro
monica	motors	space	jazz	arkansas
starr	plants	columbia	nba	school
grand	flint	astronauts	chicago	shooting
white	workers	nasa	game	boys
jury	michigan	crew	utah	teacher
house	auto	experiments	finals	students
clinton	plant	rats	jordan	westside
counsel	strikes	mission	malone	middle
intern	gms	nervous	michael	11year

Application: Community Detection

Detecting communities from a social network [ABFX08]

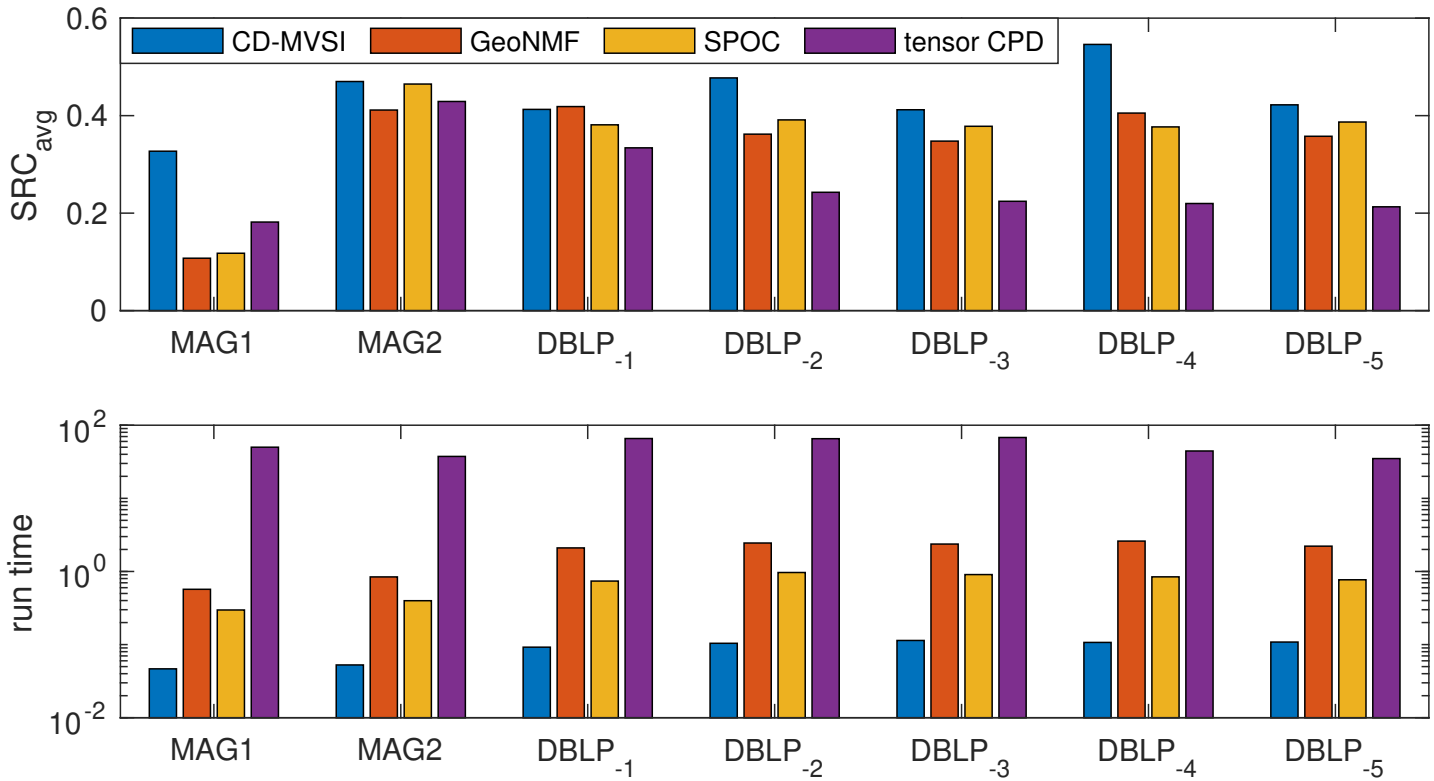
- 518×518 co-authorship network [HSS14]

courtesy: A. Swami @ARL

G. B. Giannakis	Lang Tong	Mariusz A. Fecko
Shengli Zhou	Ananthram Swami	Sunil Samtani
Xiaoli Ma	Qing Zhao	M. Umit Uyar
Pengfei Xia	Brian M. Sadler	Ibrahim Hokelek
Xiaodong Cai	Yunxia Chen	Jianping Zou
Tairan Wang	Min Dong	Jianliang Zheng
Qingwen Liu	Youngchul Sung	Myung Jong Lee
Xing Wang	Ting He	Tarek N. Saadawi
Zhengdao Wang	G. B. Giannakis	Ulas C. Kozat
Alfonso Cano	P. Venkitasubramaniam	Phillip T. Conrad

Community Detection from Co-authorship Networks

CD-MVSI [HF19a]



Concluding Remarks

1. NMF is a useful and widely used linear model in data analysis and machine learning.

Concluding Remarks

1. NMF is a useful and widely used linear model in data analysis and machine learning.

2. NMF is difficult (NP-hard) and non-unique (identifiability issue).

Concluding Remarks

1. NMF is a useful and widely used linear model in data analysis and machine learning.

2. NMF is difficult (NP-hard) and non-unique (identifiability issue).

3. However, there are several venues to circumvent these issues:
   - Separable NMF, which relies on a rather strong assumption, makes NMF identifiable and solvable in polynomial time, even in the presence of noise.

Concluding Remarks

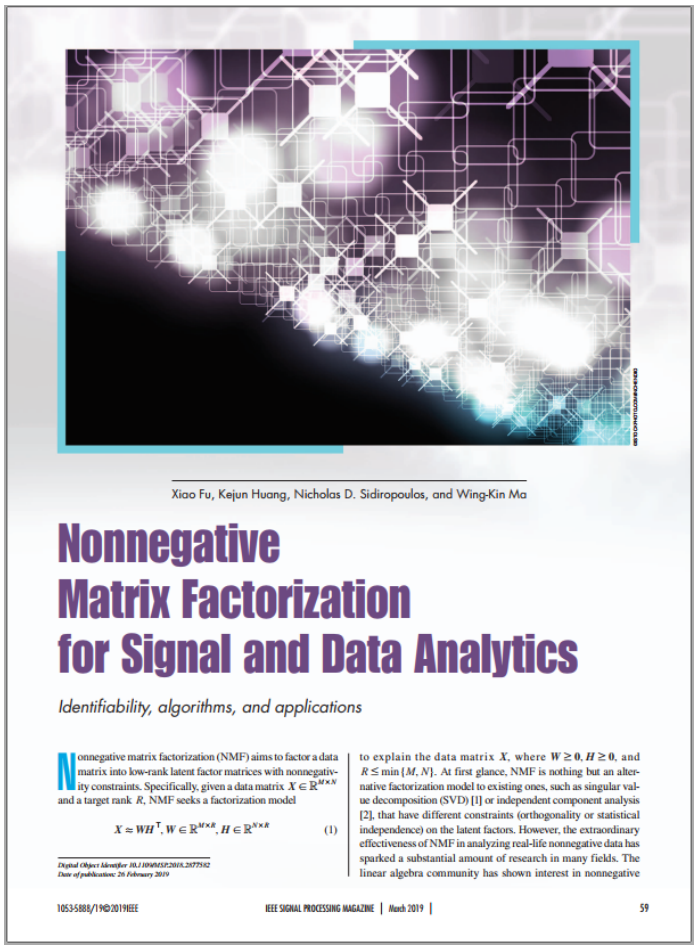
1. NMF is a useful and widely used linear model in data analysis and machine learning.

2. NMF is difficult (NP-hard) and non-unique (identifiability issue).

3. However, there are several venues to circumvent these issues:
   - Separable NMF, which relies on a rather strong assumption, makes NMF identifiable and solvable in polynomial time, even in the presence of noise.
   - Identifiability via minimum volume or sparsity but, as opposed to separability, no polynomial-time algorithm yet. This is an important direction of research, along with their robustness to noise.

Main references

Survey on NMF (2019)
IEEE Signal Process. Mag.



Xiao Fu, Kejun Huang, Nicholas D. Sidiropoulos, and Wing-Kin Ma

Nonnegative Matrix Factorization for Signal and Data Analytics

Identifiability, algorithms, and applications

Nonnegative matrix factorization (NMF) aims to factor a data matrix into low-rank latent factor matrices with nonnegativity constraints. Specifically, given a data matrix $X \in \mathbb{R}^{M \times N}$ and a target rank R, NMF seeks a factorization model

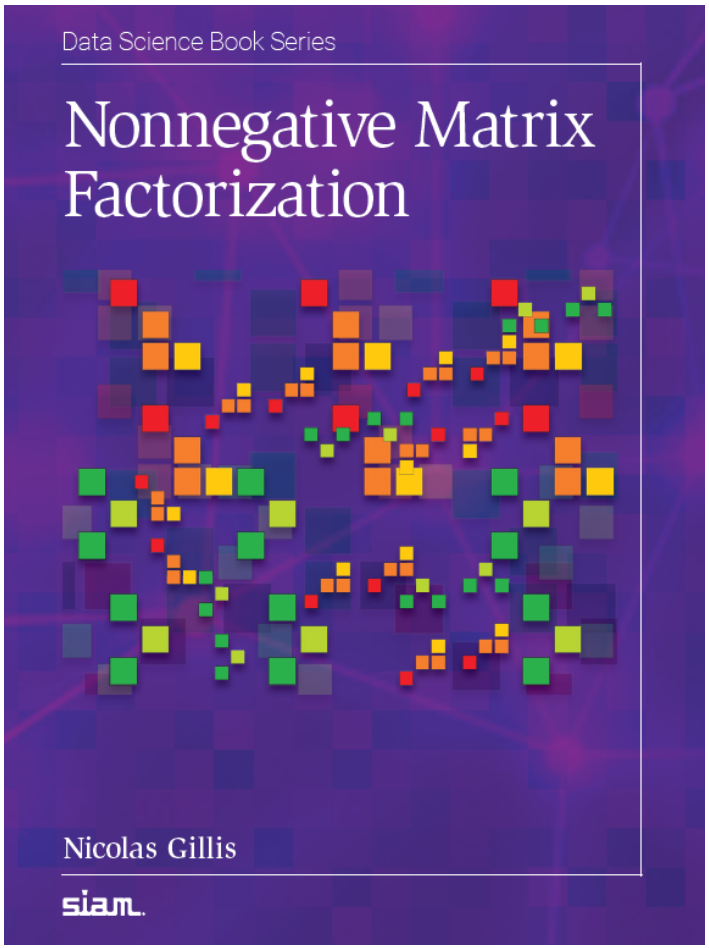
$$X \approx WH^T, W \in \mathbb{R}^{M \times R}, H \in \mathbb{R}^{N \times R} \quad (1)$$

to explain the data matrix X, where $W \geq 0, H \geq 0$, and $R \leq \min\{M, N\}$. At first glance, NMF is nothing but an alternative factorization model to existing ones, such as singular value decomposition (SVD) [1] or independent component analysis [2], that have different constraints (orthogonality or statistical independence) on the latent factors. However, the extraordinary effectiveness of NMF in analyzing real-life nonnegative data has sparked a substantial amount of research in many fields. The linear algebra community has shown interest in nonnegative

Digital Object Identifier 10.1109/MSP.2018.2877582
Date of publication: 26 February 2019

1053-5888/19©2019IEEE

Book on NMF (2020)
SIAM, bit.ly/NMFbook



References I

[ABFX08] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[AGH+13] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of ICML 2013*, 2013.

[AGKM12] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proc. of the 44th Symp. on Theory of Computing (STOC '12)*, pages 145–162, 2012.

[AGM12] S. Arora, R. Ge, and A. Moitra. Learning topic models–going beyond SVD. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1–10. IEEE, 2012.

[ASGa+01] U.M.C. Araújo, B.T.C. Saldanha, R.K.H Galvão, T. Yoneyama, H.C. Chame, and V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[BDPD+12] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE J. Sel. Topics Appl. Earth Observ.*, 5(2):354–379, 2012.

References II

[Ber99] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[CLKC08] Andrzej Cichocki, Hyekyoung Lee, Yong-Deok Kim, and Seungjin Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440, 2008.

[CMAC11] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi. A simplex volume maximization framework for hyperspectral endmember extraction. *IEEE Trans. Geosci. Remote Sens.*, 49(11):4177 –4193, Nov. 2011.

[CP09] A. Cichocki and A.-H. Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, 92(3):708–721, 2009.

[CR93] J.E. Cohen and U.G. Rothblum. Nonnegative ranks, decompositions and factorization of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

References III

[DS79] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pages 20–28, 1979.

[DS03] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, volume 16, 2003.

[FBD09] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[FGWH20] Xiao Fu, Cheng Gao, Hoi-To Wai, and Kejun Huang. Block-randomized stochastic proximal gradient for low-rank tensor factorization. *IEEE Trans. Signal Process.*, 68:2170–2185, 2020.

[FHS18] Xiao Fu, Kejun Huang, and Nicholas D Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.

[FKPT13] Samuel Fiorini, Volker Kaibel, Kanstantsin Pashkovich, and Dirk Oliver Theis. Combinatorial bounds on nonnegative rank and extended formulations. *Discrete Mathematics*, 313(1):67–83, 2013.

[FM16] X. Fu and W.-K. Ma. Robustness analysis of structured matrix factorization via self-dictionary mixed-norm optimization. *IEEE Signal Process. Lett.*, 23(1):60–64, 2016.

References IV

[FMHS15] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Trans. Signal Process.*, 63(9):2306–2320, May 2015.

[FMS16] X. Fu, W.-K. Ma, and N.D. Sidiropoulos. Power spectra separation via structured matrix factorization. *IEEE Trans. Signal Process.*, 64(17):4592–4605, 2016.

[GG11] Nicolas Gillis and Fran$\boldsymbol{c}$*coisGlineur*. *Low* − *rankmatrixapproximationwithweightsormissingdataisNP* − *hard*. *SIAMJournalonMatrixAnalysisandApplications*, 32(4) : 1149 − −1165, 2011.

[GG12] N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012.

[GL14] Nicolas Gillis and Robert Luce. Robust near-separable nonnegative matrix factorization using linear optimization. *The Journal of Machine Learning Research*, 15(1):1249–1280, 2014.

[GM75] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

References V

[GM76] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[GS00] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

[GS19] Nicolas Gillis and Yaroslav Shitov. Low-rank matrix approximation in the infinity norm. *Linear Algebra and its Applications*, 581:367–382, 2019.

[GTLY12] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Nenmf: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

[GV14] N. Gillis and S.A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):698–714, April 2014.

[GV15] Nicolas Gillis and Stephen A Vavasis. Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. *SIAM Journal on Optimization*, 25(1):677–698, 2015.

[GV18] Nicolas Gillis and Stephen A Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Mathematics of Operations Research*, 43(4):1072–1084, 2018.

References VI

[HF19a] Kejun Huang and Xiao Fu. Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2859–2868, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[HF19b] Kejun Huang and Xiao Fu. Low-complexity proximal gauss-newton algorithm for nonnegative matrix factorization. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019.

[HFS16] Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pages 1786–1794, 2016.

[HFS18] Kejun Huang, Xiao Fu, and Nicholas D. Sidiropoulos. Learning hidden markov models from pairwise co-occurrences with applications to topic modeling. In *Proceedings of ICML 2018*, 2018.

[HG20] Le Thi Khanh Hien and Nicolas Gillis. Algorithms for nonnegative matrix factorization with the kullback-leibler divergence. *arXiv preprint arXiv:2010.01935*, 2020.

References VII

[HSL16] Kejun Huang, Nicholas D Sidiropoulos, and Athanasios P Liavas. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing*, 64(19):5052–5065, 2016.

[HSS14] Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2014.

[HVD08] Ngoc-Diep Ho and Paul Van Dooren. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra and its Applications*, 429(5-6):1020–1025, 2008.

[IF20] Shahana Ibrahim and Xiao Fu. Recovering joint probability of discrete random variables from pairwise marginals. *arXiv preprint arXiv:2006.16912*, 2020.

[IFKH19] Shahana Ibrahim, Xiao Fu, N. Kargas, and K. Huang. Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms. In *Advances in Neural Information Processing Systems*, pages 7847–7857, 2019.

[KP08] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

References VIII

[KP11] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[KSF18] Nikos Kargas, Nicholas D Sidiropoulos, and Xiao Fu. Tensors, learning, and 'Kolmogorov extension'for finite-alphabet random vectors. *IEEE Trans. Signal Process.*, 66(18):4854–4868, 2018.

[Lin07] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[LR10] Balaji Lakshminarayanan and Raviv Raich. Non-negative matrix factorization for parameter estimation in hidden markov models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 89–94. IEEE, 2010.

[LS99] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[LS01] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562, 2001.

[MBDC+14] W.-K. Ma, J.M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A.J. Plaza, A. Ambikapathi, and C.-Y. Chi. A signal processing perspective on hyperspectral unmixing. *IEEE Signal Process. Mag.*, 31(1):67–81, Jan 2014.

References IX

[Miz14] Tomohiko Mizutani. Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. *The Journal of Machine Learning Research*, 15(1):1011–1039, 2014.

[Moi13] A. Moitra. An almost optimal algorithm for computing nonnegative rank. In *Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA '13)*, pages 1454–1464, 2013.

[MSC17] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. On mixed memberships and symmetric nonnegative matrix factorizations. In *Proceedings of ICML 2017*, pages 2324–2333, 2017.

[PSU17] Maxim Panov, Konstantin Slavnov, and Roman Ushakov. Consistent estimation of mixed memberships with successive projections. In *International Conference on Complex Networks and their Applications*, pages 53–64. Springer, 2017.

[RHL13] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[Rot14] Thomas Rothvoss. The matching polytope has exponential extension complexity. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 263–272. ACM, 2014.

[RRTB12] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012.

References X

[RYZ+10] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[SDLF+17] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.*, 65(13):3551–3582, 2017.

[SFB+95] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. 1995.

[SOJN08] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[TG18] Panagiotis A Traganitis and Georgios B Giannakis. Blind multi-class ensemble learning with dependent classifiers. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2025–2029. IEEE, 2018.

[TG19] Panagiotis A Traganitis and Georgios B Giannakis. Unsupervised ensemble classification with dependent data. *arXiv preprint arXiv:1906.09356*, 2019.

References XI

[TPZG18] Panagiotis A Traganitis, Alba Pages-Zamora, and Georgios B Giannakis. Blind multiclass ensemble classification. *IEEE Trans. Signal Process.*, 66(18):4737–4752, 2018.

[Vav10] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

[Win99] Michael E Winter. N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Imaging Spectrometry V*, volume 3753, pages 266–275. International Society for Optics and Photonics, 1999.

[XYWZ12] Y. Xu, W. Yin, Z. Wen, and Y. Zhang. An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers of Mathematics in China*, 7(2):365–384, 2012.

[Yan91] Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441–466, 1991.

[ZBMP12] Dengyong Zhou, Sumit Basu, Yi Mao, and John C Platt. Learning from the wisdom of crowds by minimax entropy. In *Advances in neural information processing systems*, pages 2195–2203, 2012.

[ZCZJ14a] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. Advances in Neural Information Processing Systems, 2014.

References XII

[ZCZJ14b] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pages 1260–1268, 2014.