Nonnegative **u**nimodal **M**atrix **F**actorization

- Andersen Ang (Dept. Combinatorics and Optimization, UWaterloo, Canada)
  Homepage: angms.science

- Paper: **A**, Gillis, Vandaele and De Sterck, "Nonnegative Unimodal Matrix Factorization"

- Content: Nu
  - What is it? Introduction
  - ~~Why?~~ ~~Motivation~~
  - How to solve? Algorithm
  - ~~What is known about Nu?~~ ~~Theory~~

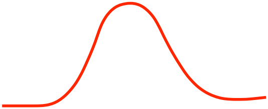
East Coast Optimization Meeting, 02 April 2021

Setting

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{C} \qquad \text{OR} \qquad \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| + \lambda c(\mathbf{x}).$$

Common $\mathcal{C}$ and c

- ℓ_2 norm
- ℓ_1 norm, ℓ_0 norm, sparsity
- ℓ_p norm
- TV-norm, smoothness
- Nonnegativity
- Cone
- ...

This talk: **unimodal** structure. 

(Not the **unimodular** structure in combinatorial optimization.)

Structural factorization

- Factorize data it into (low-rank) factors with **structural constraints**.

Matrix Factorization $\boxed{M \in \mathbb{R}^{m \times n}} \approx \boxed{W}\,\boxed{H}$

- **Nonnegative unimodal structure**

$$\underbrace{0 \leq a_1 \leq \cdots \leq a_{p-1} \leq}_{\text{increasing head}} \quad a_p \quad \underbrace{\geq a_{p+1} \geq \cdots \geq a_m \geq 0}_{\text{decreasing tail}} . \qquad \text{(Nu)}$$

Example of Nu vectors

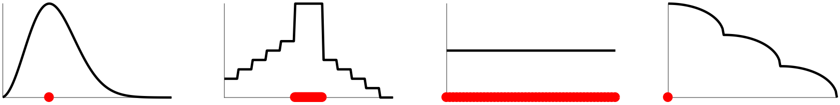


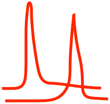
Figure: Four Nu vectors. Black curve: the plot of the sequence. Red dots: the locations of p.

Characterizing the Nu set

$\{1,2,\cdots,m\}$

$$\mathbf{x}\in\mathbb{R}^m \text{ is Nu } \iff \underbrace{\exists p\in[m] \text{ s.t. } 0\le x_1\le\cdots\le x_p\ge\cdots\ge x_m\ge 0}_{\mathbf{x}\in\mathcal{U}_+^{m,p}}.$$

- Notations: $\mathbf{x}\in\mathcal{U}_+^m$ means $\mathbf{x}\in\mathbb{R}^m$ is Nu but p unknown.
- Facts
  - $\mathcal{U}_+^{m,p}$ is cvx
  - $\mathcal{U}_+^m=\bigcup_k\mathcal{U}_+^{m,k}$ is **non**cvx



  - The set $\mathcal{U}_+^{m,p}\cup\mathcal{U}_+^{m,p+1}$ is cvx.

$$\mathbf{x}\in\mathbb{R}^m \text{ is Nu } \iff \exists p\in[m] \text{ s.t. } \mathbf{x}\in\mathcal{U}_+^{m,p}\cup\mathcal{U}_+^{m,p+1}$$

Characterizing Nu membership by a system of inequalities

$$\mathbf{x} \in \mathbb{R}^m \text{ is Nu} \quad \Longleftrightarrow \quad \exists p \in [m] \text{ s.t. } \mathbf{x} \in \underbrace{\mathcal{U}_+^{m,p} \cup \mathcal{U}_+^{m,p+1}}_{\text{convex}}$$

$$\Longleftrightarrow \quad \begin{cases} 0 & \leq & x_1 \\ x_1 & \leq & x_2 \\ & \vdots & \\ x_{p-1} & \leq & x_p \\ x_{p+1} & \geq & x_{p+2} \\ & \vdots & \\ x_{m-1} & \geq & x_m \\ x_m & \geq & 0 \end{cases}$$

$$\Longleftrightarrow \quad \mathbf{U}_p \mathbf{x} \underset{\text{elementwise}}{\geq} \mathbf{0}, \quad \mathbf{U}_p = \begin{pmatrix} \underbrace{\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}_{p \times p}}_{\mathbf{D}_{p\times p}} & \mathbf{0}_{p\times(m-p)} \\ \mathbf{0}_{(m-p)\times p} & \mathbf{D}^\top_{(m-p)\times(m-p)} \end{pmatrix}.$$

NuMF

- GIVEN $\mathbf{M} \in \mathbb{R}_+^{m \times n}$ and $r \in \mathbb{N}$,
  FIND $\mathbf{W} \in \mathbb{R}^{m \times r}$ and $\mathbf{H} \in \mathbb{R}^{r \times n}$ by solving

$$
\begin{aligned}
\min \ \tfrac{1}{2}\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F^2 \quad \text{s.t.} \quad & \mathbf{H} \geq \mathbf{0}, \\
& \mathbf{w}_j \in \mathcal{U}_+^m \text{ for all } j \in [r], \\
& \mathbf{w}_j^\top \mathbf{1}_m = 1 \text{ for all } j \in [r],
\end{aligned}
$$

- Apply the Nu characterization:

$$
\mathbf{w}_j \in \mathcal{U}_+^m \quad \to \quad \mathbf{U}_{p_j} \mathbf{w}_j \geq \mathbf{0},
$$

  where integers $p_1, p_2, \dots, p_r$ are unknown.

- How to solve: BCD.
  - Subproblem on $\mathbf{H}$ is simple.
  - Main difficulty: subproblem on $\mathbf{W}$.

HALS: Column-wise block coordinate descent

$$
\begin{aligned}
\frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2 &= \frac{1}{2}\left\|\mathbf{M}-\sum_{j=1}^{r}\mathbf{w}_j\mathbf{h}^j\right\|_F^2 \\
&= \frac{1}{2}\left\|\underbrace{\mathbf{M}-\sum_{j\neq i}^{r}\mathbf{w}_j\mathbf{h}^j}_{:=\mathbf{M}_i}-\mathbf{w}_i\mathbf{h}^i\right\|_F^2 \\
&= \frac{1}{2}\|\mathbf{M}_i-\mathbf{w}_i\mathbf{h}^i\|_F^2 \\
&= \text{a quadratic function on } \mathbf{w}_i
\end{aligned}
$$

The $\mathbf{w}_j$-subproblem

- A linearly-constrained quadratic program, **convex**:

$$\min_{\mathbf{w}_i} \frac{\|\mathbf{h}^i\|_2^2}{2}\|\mathbf{w}_i\|_2^2 - \langle \mathbf{M}_i \mathbf{h}^{i\top}, \mathbf{w}_i \rangle \ \text{ s.t. } \underbrace{\mathbf{U}_{p_i}\mathbf{w}_i \geq \mathbf{0}}_{\mathbf{w}_i \in \mathcal{U}_+^{m,p_i}}, \ \mathbf{w}_i^\top \mathbf{1} = 1, \quad (*)$$

- Involves integer variables, **nonconvex**:

$$\min_{\mathbf{w}_i, p_i} \frac{\|\mathbf{h}^i\|_2^2}{2}\|\mathbf{w}_i\|_2^2 - \langle \mathbf{M}_i \mathbf{h}^{i\top}, \mathbf{w}_i \rangle \ \text{ s.t. } \mathbf{w}_i \in \mathcal{U}_+^m, \ \mathbf{w}_i^\top \mathbf{1} = 1, \quad (**)$$

- Brute-force: solve (*) on all p, pick the best one as p_i to solve (**).

- Directly solving (**) by proximal gradient is not scalable ($\$\$\$$).
  - Proximal gradient = a 2-branch isotonic projection.
  - Isotonic projection: $\mathbf{x} \leq \mathbf{y} \implies \mathcal{P}_\mathcal{K}\mathbf{x} \leq \mathcal{P}_\mathcal{K}\mathbf{y}$.

Speed up the brute-force algorithm for large m

$$\min_{\mathbf{w}_i} \frac{\|\mathbf{h}^i\|_2^2}{2} \|\mathbf{w}_i\|_2^2 - \langle \mathbf{M}_i \mathbf{h}^{i\top}, \mathbf{w}_i \rangle \;\text{ s.t. }\; \mathbf{U}_{p_i} \mathbf{w}_i \geq \mathbf{0},\; \mathbf{w}_i^\top \mathbf{1} = 1, \qquad (*)$$

$$\min_{\mathbf{w}_i, \color{red}{p_i}} \frac{\|\mathbf{h}^i\|_2^2}{2} \|\mathbf{w}_i\|_2^2 - \langle \mathbf{M}_i \mathbf{h}^{i\top}, \mathbf{w}_i \rangle \;\text{ s.t. }\; \mathbf{U}_{p_i} \mathbf{w}_i \geq \mathbf{0},\; \mathbf{w}_i^\top \mathbf{1} = 1, \qquad (**)$$

- Brute-force on p in $[m]$ ok if m small.
- Speed up 1: solve (*) by accelerated projected gradient.
- Speed up 2: reduce the search space for p_i in (**)
  - By guessing the location of p_i's
  - By dimension reduction: **multi-grid method**
    - Multi-grid preserves Nu: a theorem with proof in 3 sentences!
    - Other dimension reduction techniques such as PCA or Gaussian sampling do not work here as they destroy the Nu.

APG: Accelerated Projected Gradient

$$\min_{\mathbf{w}_i} \frac{\|\mathbf{h}^i\|_2^2}{2}\|\mathbf{w}_i\|_2^2 - \langle \mathbf{M}_i \mathbf{h}^{i\top}, \mathbf{w}_i \rangle \ \text{ s.t. } \underbrace{\mathbf{U}_{p_i}\mathbf{w}_i \geq \mathbf{0},\ \mathbf{w}_i^\top \mathbf{1} = 1}_{\text{hard to project}}. \qquad (*)$$

- Transform (*) via $\mathbf{y} = \mathbf{U}\mathbf{w}$:

$$\min_{\mathbf{y}} \frac{1}{2}\Big\langle \|\mathbf{h}^i\|_2^2 \mathbf{U}_{p_i}^{-\top}\mathbf{y},\ \mathbf{y}\Big\rangle - \Big\langle \mathbf{U}_{p_i}^{-\top}\mathbf{M}_i\mathbf{h}^{i\top},\ \mathbf{y}\Big\rangle \text{ s.t. } \mathbf{y} \geq \mathbf{0},\ \mathbf{y}^\top \mathbf{U}_{p_i}^{-\top}\mathbf{1} = 1$$

  or equivalently

$$\min_{\mathbf{y}} \frac{1}{2}\langle \mathbf{Q}\mathbf{y}, \mathbf{y}\rangle - \langle \mathbf{p}, \mathbf{y}\rangle \ \text{ s.t. } \mathbf{y} \geq \mathbf{0},\ \mathbf{y}^\top \mathbf{b} = 1. \qquad (*')$$

- $\mathbf{y}^*$ solves (*′) gives $\mathbf{w}_i^*$ that solves (*) by $\mathbf{y} = \mathbf{U}\mathbf{w}$.

APG on solving $\mathbf{y}$

$$\min_{\mathbf{y}} \frac{1}{2}\langle \mathbf{Q}\mathbf{y}, \mathbf{y}\rangle - \langle \mathbf{p}, \mathbf{y}\rangle \;\text{ s.t. }\; \mathbf{y} \geq \mathbf{0},\; \mathbf{y}^\top \mathbf{b} = 1. \tag{*$'$}$$

- Projection: $P(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \mathbf{z}\|_2^2 \text{ s.t. } \mathbf{y} \geq \mathbf{0},\; \mathbf{y}^\top \mathbf{b} = 1.$
- Optimal sol. by partial Lagrangian

$$\mathbf{y}^* \overset{(*)}{=} \min_{\mathbf{y}\geq\mathbf{0}} \max_{\nu} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{z}\|_2^2 + \nu(\mathbf{y}^\top \mathbf{b} - 1)}_{L(\mathbf{y},\nu)} = \underbrace{[\mathbf{z} - \nu^* \mathbf{b}]_+}_{\text{soft-thresholding}},$$

where Lagrangian multiplier ν^* is the root of a piece-wise linear eqn.

$$\sum_{i=1}^{m} \max\left\{0, z_i - \nu b_i\right\} b_i - 1 = 0,$$

which costs $\mathcal{O}(m)$ to $\mathcal{O}(m \log m)$ to solve by sorting the break points $\frac{z_i}{b_i}$. After sorting, the magical-one-line-code that no one can read is

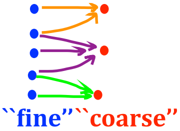
```
nu = max((cumsum(z.*b)-1)./(cumsum(b.*b)));
```

(*): The problem satisfies the Slater's condition which guarantees strong duality.

Multi-grid

- Idea: instead of working on $\mathbf{w}$, work on $\mathbf{R}_N \ldots \mathbf{R}_1 \mathbf{w}$ with smaller search space of p.

- Restriction $\mathbf{R} \in \mathbb{R}_+^{m_1 \times m}$ changes $\mathbf{x} \in \mathbb{R}_+^m$ to $\mathbf{R}\mathbf{x} \in \mathbb{R}_+^{m_1}$ with $m_1 < m$.



$$\mathbf{R}(a,b) = \begin{bmatrix} a & b & & & & & \\ & b & a & b & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b & a & b & \\ & & & & & b & a \end{bmatrix}, \quad \begin{matrix} a > 0, b > 0, \\ a + 2b = 1. \end{matrix}$$

- **Theorem (if $\mathbf{x}$ is NU, then $\mathbf{R}\mathbf{x}$ is Nu)** Let $\mathbf{x} \in \mathcal{U}_+^{m,p}$ with p is even[1] and $\mathbf{R} \in \mathbb{R}^{m_1 \times m}$. Then $\mathbf{y} = \mathbf{R}\mathbf{x} \in \mathcal{N}_+^{m_1, p_y}$ with $\mathcal{N}_+^{m,p} = \mathcal{U}_+^{m,p} \cup \mathcal{U}_+^{m,p+1}$ and $p_y \in \{\lfloor \frac{p}{2} + 1 \rfloor, \lfloor \frac{p}{2} \rfloor\}$.

[1] If p is odd, by considering the vector $[0, \mathbf{x}]$ does not change the unimodality and increases p by one.

The whole algorithm (in words) for NuMF$(\mathbf{M}, r)$

Steps:

1. Restrict: $\mathbf{M}^{[N]} = \mathbf{R}_N \ldots \mathbf{R}_1 \mathbf{M}$ and $\mathbf{W}_0^{[N]} = \mathbf{R}_N \ldots \mathbf{R}_1 \mathbf{W}_0$.
2. Solve coarse problem: brute-force and APG on
$$[\mathbf{W}_*^{[N]}, \mathbf{H}_*, \mathbf{p}_*^{[N]}] \leftarrow \mathrm{NuMF}(\mathbf{M}^{[N]}, \mathbf{W}_0^{[N]}, \mathbf{H}_0).$$
3. Interpolate: $[\mathbf{W}_0, \mathbf{p}_0] \leftarrow \mathrm{Interpolate}(\mathbf{W}_*^{[N]}, \mathbf{p}_*^{[N]})$.
4. Solve the original fine problem:
$$[\mathbf{W}_*, \mathbf{H}_*, \mathbf{p}_*] \leftarrow \mathrm{NuMF}(\mathbf{M}, \mathbf{W}_0, \mathbf{H}_0, \mathbf{p}_0).$$

   no brute-forcing!

Convergence

- Optimization: local convergence.
- Linear Algebra: global convergence.
  Identifiability – when does solving NuMF give a unique sol?

Fancy picture: multi-grid saves 75% time with 2-layer

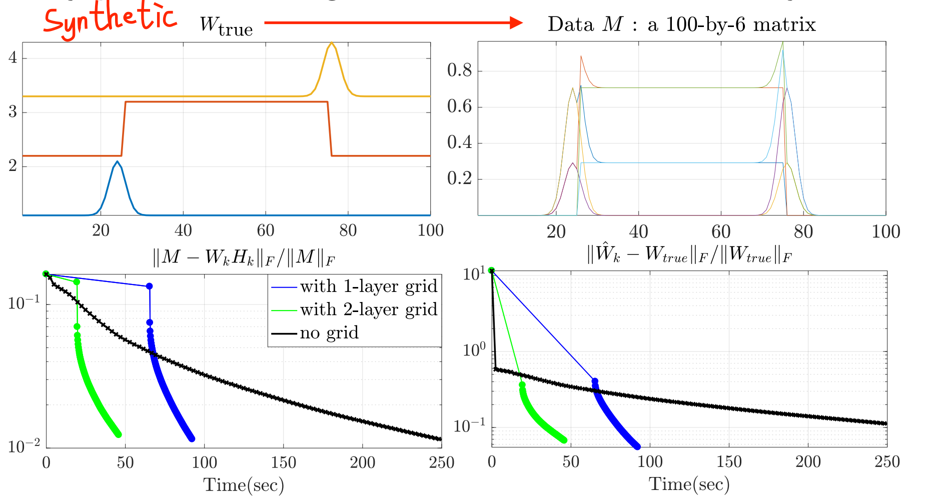
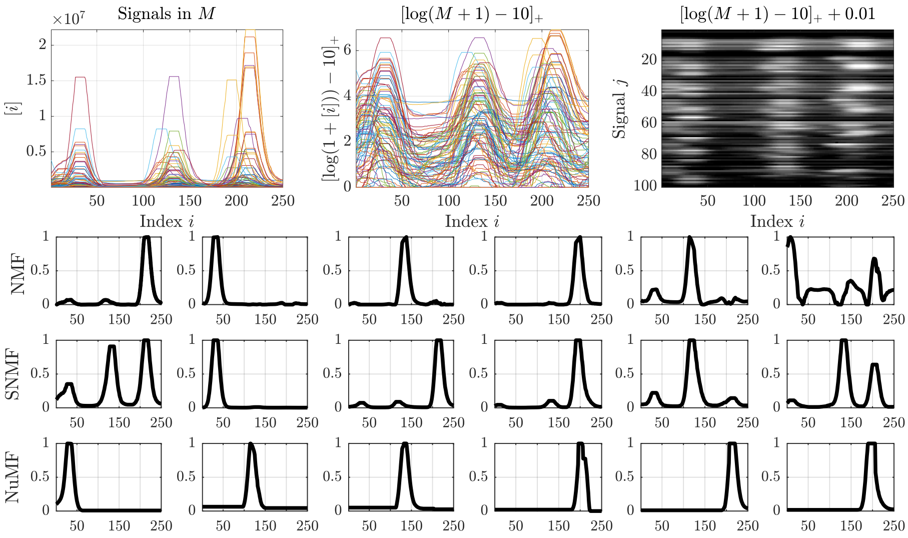
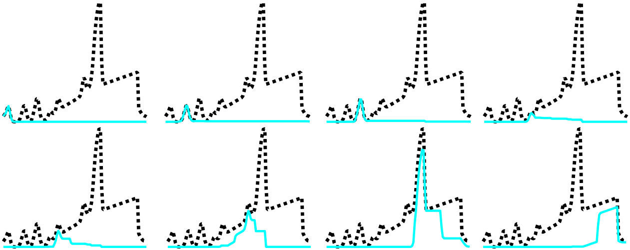


Figure: Experiment on a toy example. All algo. run 100 iterations with same initialization. For algo. with MG, the computational time taken on the coarse grid are also taken into account, as reflected by the time gap between time 0 and the first dot in the curves.

Fancy picture: on Belgian beers



Fancy picture: on $r > n$



- On a data vector in $\mathbb{R}_+^{947}$ (black curve) with $r = 8 > 1 = n$.
- Cyan curves are the components $\mathbf{w}_i h_i$.
- Relative error $\|\mathbf{M} - \mathbf{WH}\|_F / \|\mathbf{M}\|_F = 10^{-8}$.
- The first two peaks in the data satisfy an identifiability Theorem, NuNMF identifies them perfectly.
- For the other peaks: supports overlap, decomposition not unique.

Last page - summary

- NuMF problem: nonconvex and block-nonconvex.
- Nu characterization and brute-force
- Acceleration by APG and MG
- Not discussed
  - The log-concavity
  - Guessing location of p
  - Identifiability of NuMF.
  - The traditional non-NuMF approach used in analytical chemistry
- References
  - Chapter 5 of my thesis "Nonnegative Matrix and Tensor Factorizations: Models, Algorithms and Applications".
  - **A**, Gillis, Vandaele and De Sterck, "Nonnegative Unimodal Matrix Factorization", to be presented in ICASSP21.
- Slide, paper, code at angms.science