Pattern Discovery in Time-Course Omics Data Using Non-Negative CP tensor decomposition (NCPD)

Shoaib Bin Masud [1] Anna Konstorum [2] Misha Kilmer [3] Shuchin Aeron [1]

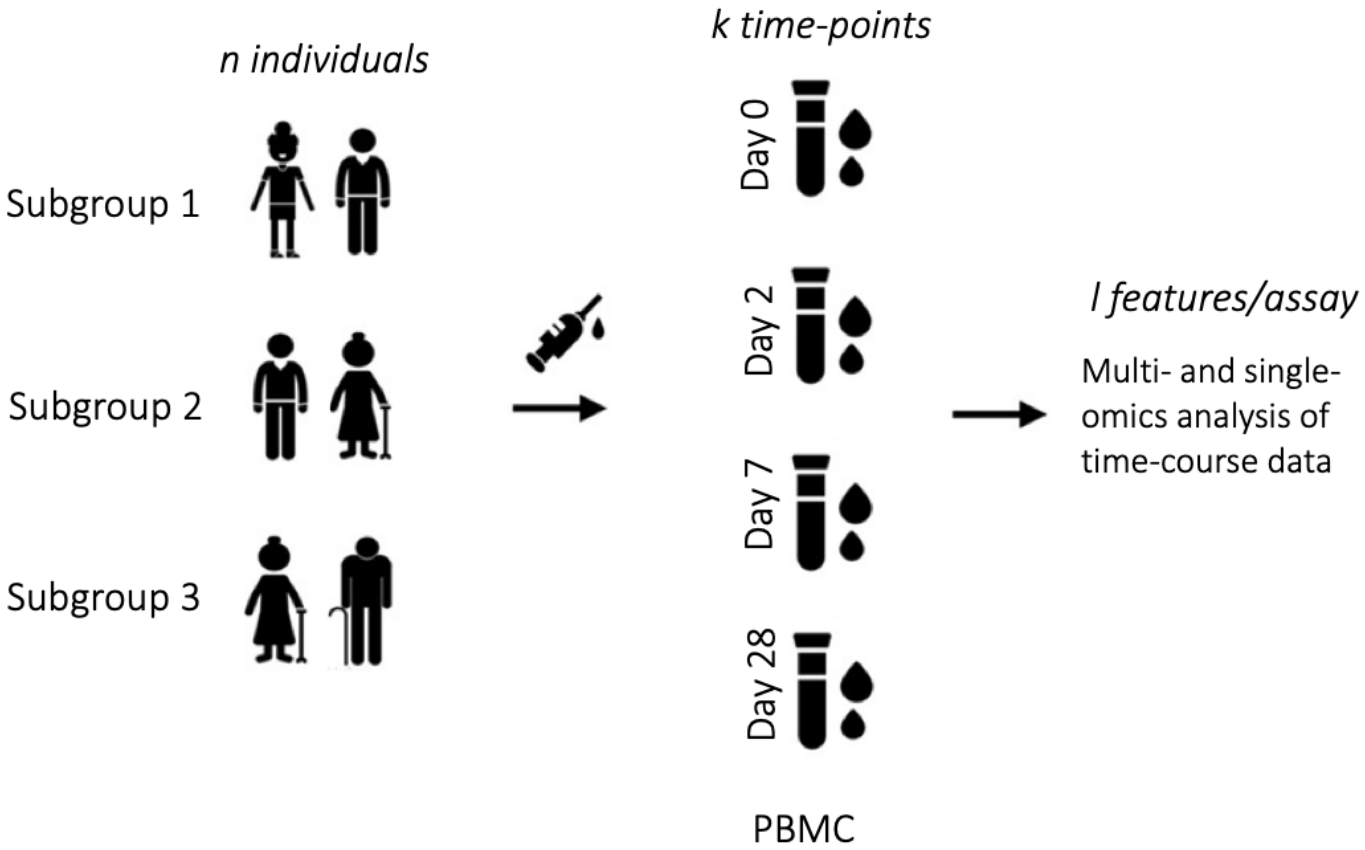
[1] Department of ECE, Tufts University

[2] Department of Pathology, Yale School of Medicine

[3] Department of Mathematics, Tufts University

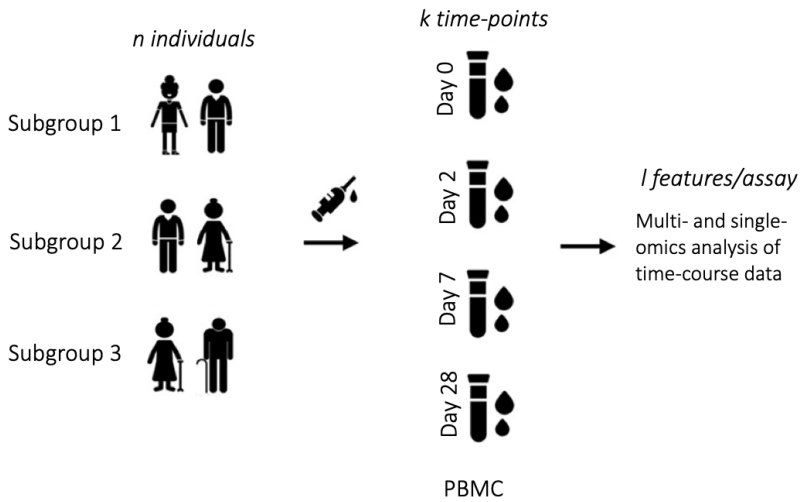
February 26, 2023

Vaccination response data structure



- Subgroups of individuals of interest to researchers (e.g. age, sex, living-condition) are vaccinated at day 0, and samples are collected pre-vaccination (day 0), and at time-points post-vaccination.
- Total of $l \times n \times k$ data points

Vaccination response data structure



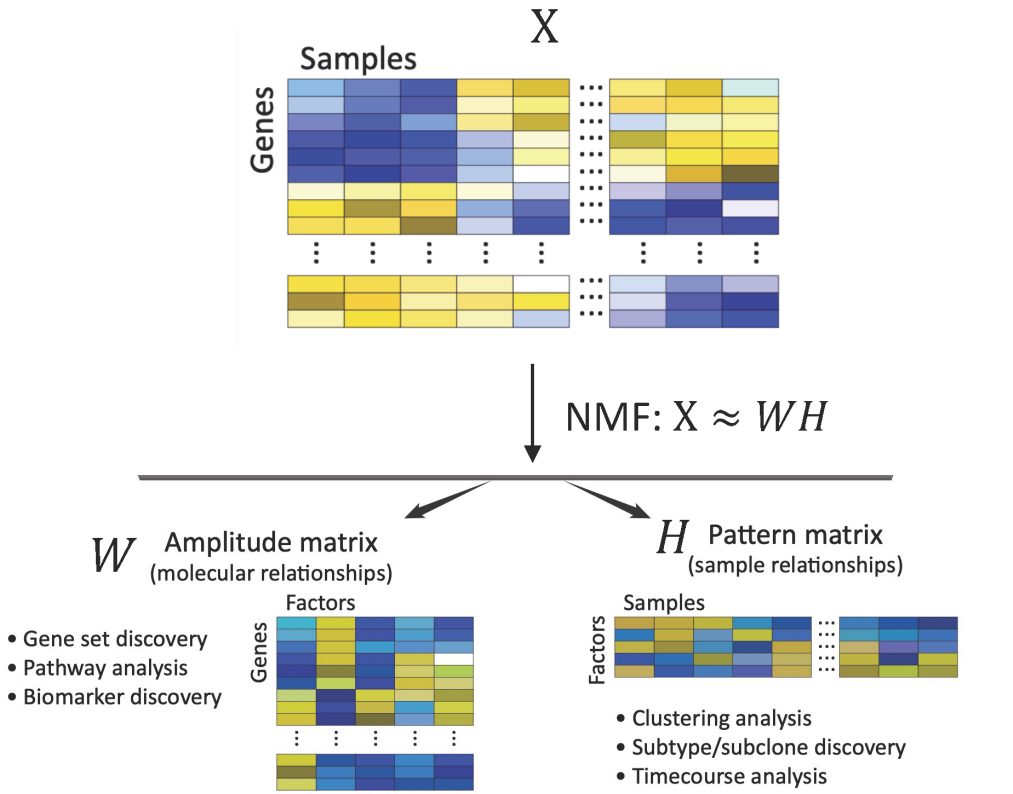
- Total of $n \times k \times l$ data points
- Want to answer questions including:
  - Are there different subgroups of individuals that respond differently to the vaccine (either the pre-determined subgroups, or other clusters)?
  - Which genes are most active in the response?
  - Are there additional covariates that mitigate response signature? (e.g. demographics, experiment,...)

Non-negative Matrix Factorization (NMF)

Recall that for a non-negative matrix, $\mathbf{X}_{n\times m}$, non-negative matrix factorization (NMF) finds a rank R decomposition,

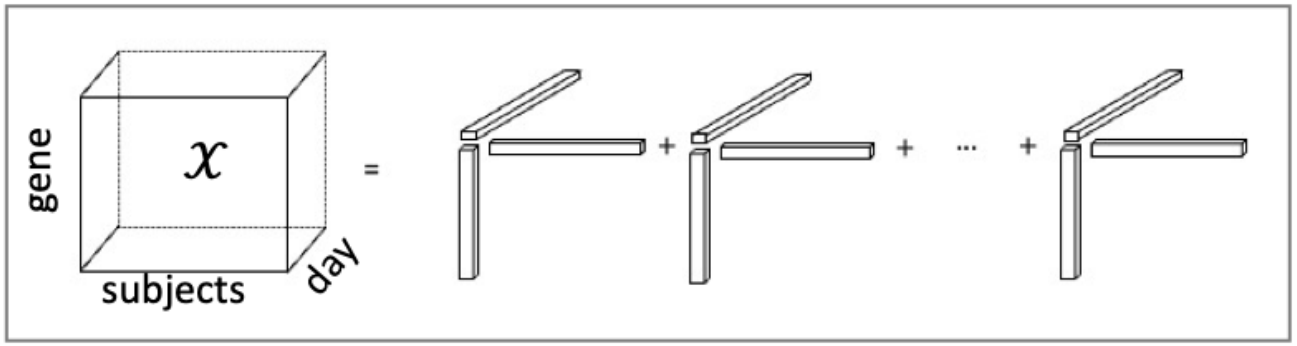
$$\mathbf{X}_{n\times m} \approx \mathbf{W}_{n\times R}\mathbf{H}_{R\times m}$$

- If X is a gene by sample matrix for a given time-point, then each column of W represents a particular co-expressed pattern of gene activation, and each column of H represents the strength of presence of that pattern in a given subject.



(Stein-O'Brien et al., 2018)

Non-negative CP decomposition (NCPD)



- Non-negative CP decomposition (NCPD) extends the concept of representing a dataset as the sum of rank-one components to ≥ 2 dimensions.
- For a tensor $\mathcal{X}$ of gene expression x subject x time data, we can represent $\mathcal{X}$ as

$$\mathcal{X} \approx [[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)}, \tag{1}$$

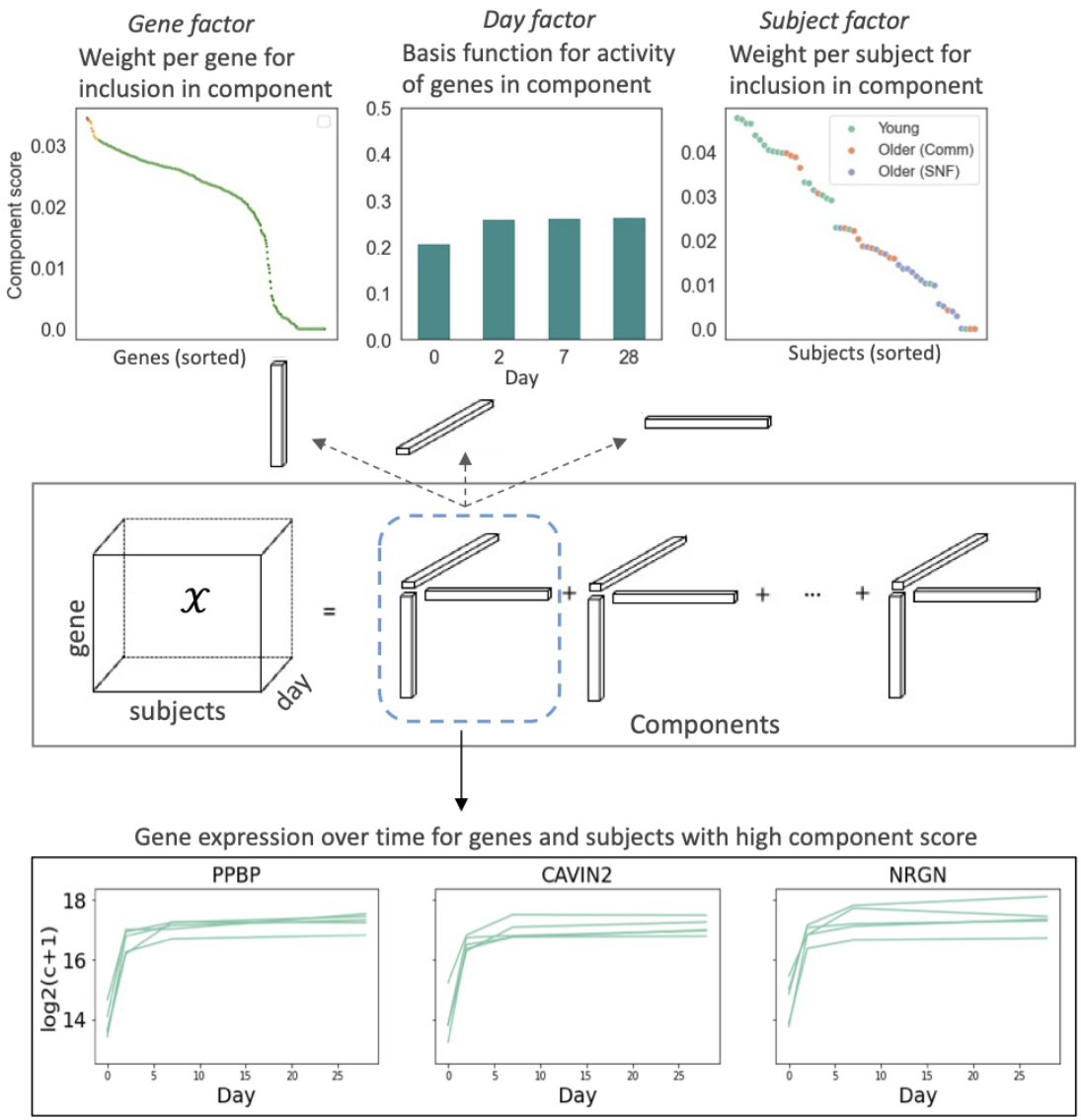
  where $\mathbf{a}_r^{(i)} \geq 0$ and $||\mathbf{a}_r^{(i)}||_2 = 1$ for $i = \{1, 2, 3\}$, and λ is the normalization constant.
- Denoting $\hat{\mathcal{X}} = [[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]]$, the objective to minimize for tensor decomposition:

$$\min_{\lambda, A^{(1)}, A^{(2)}, A^{(3)} \geq 0} \phi(\mathcal{X}, \hat{\mathcal{X}}), \tag{2}$$

- $\phi(\mathcal{X}, \hat{\mathcal{X}})$ is the loss function e.g., squared Frobenius norm[1] $\phi(\mathcal{X}, \hat{\mathcal{X}}) = \frac{1}{2}\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2$. We denote this decomposition as NCPD-F.

[1] Used in CP-OPT/CP-ALS library in Tensor Toolbox

Non-negative CP decomposition (NCPD): Interpretation



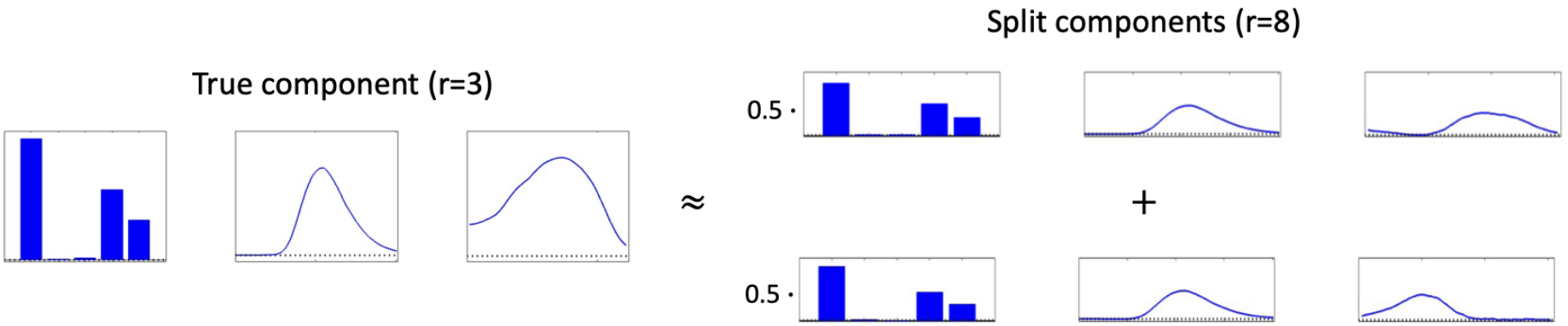
- Performed using CP-OPT library in Tensor Toolbox.

Non-negative CP decomposition (NCPD): choosing rank

How to choose the number of rank R for decomposition?

- If R is less than the true rank, we risk not being able to detect all patterns in the dataset.
- By setting the number of ranks too high, component splitting may occur. In this case, decomposition will split one pattern into two or more- rendering the analysis of each of these not only uninformative but also misleading.

- A well-known amino acids data set from Andersson and Bro [2].
- Contains fluorescence measurements of 5 samples containing 3 amino acids: Tryptophan, Tyrosine, and Phenylalanine.
- Each amino acid corresponds to a rank-one component.
- The tensor is of size 5 x 51 x 201 from 5 samples, 51 excitations, and 201 emissions.
- When the decomposition rank is set to be *higher* than the true rank, component splitting can occur. Here, a 'true' component from a rank three decomposition is split on the third mode.



[2] Rasmus Bro, PARAFAC: Tutorial and applications, Chemometrics and Intelligent Laboratory Systems, 1997, 38, 149-171

- We choose the rank R of the model based on the following considerations:
  - **Relative Frobenius error**

$$\frac{\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2}{\|\mathcal{X}\|_F^2}$$

  - **Similarity score:** For two mode-3 models of same rank R, $\hat{\mathcal{X}}_1 = [[\lambda^1; A^{(1)}, A^{(2)}, A^{(3)}]]$ and $\hat{\mathcal{X}}_2 = [[\lambda^2; B^{(1)}, B^{(2)}, B^{(3)}]]$, similarity score is defined as:

$$S(\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2) = \max_{\omega \in \Omega} \frac{1}{R} \sum_{i=1}^{R} \left(1 - \frac{|\lambda_r^1 - \lambda_{\omega(r)}^2|}{\max(\lambda_r^1, \lambda_{\omega(r)}^2)}\right) \prod_{i=1}^{3} {a_r^{(i)}}^\top b_{\omega(r)}^{(i)}. \tag{3}$$

    - Ω denotes the set of all permutations of the factors, and ω is a particular permutation.
    - Similarity score measures robustness of a decomposition across varying initial conditions.
  - **max internal n-similarity (mINS)**: A new metric to determine whether component splitting has occurred. [3]

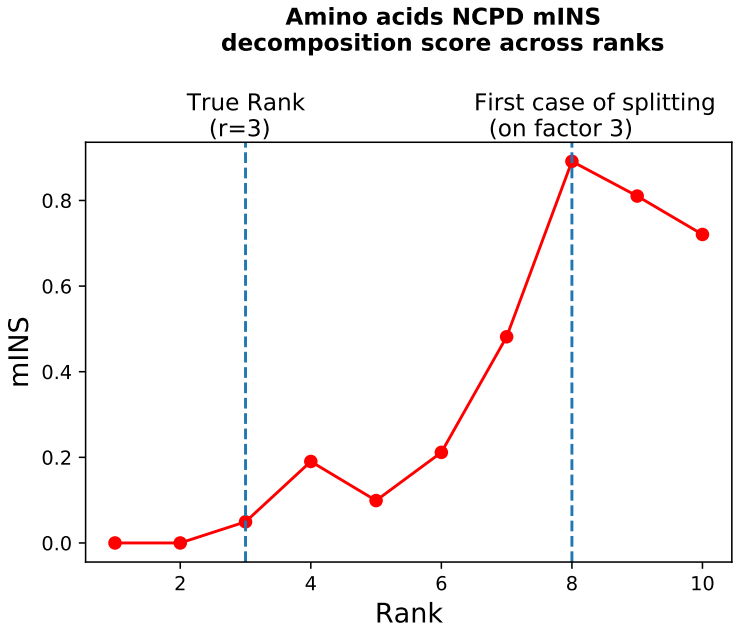
[3] Developed by Konstorum et. al [1] Kleinstein Lab, Yale University. Manuscript preparation in progress.

Non-negative CP decomposition (NCPD): choosing rank

- The *maximum internal n-Similarity* (mINS) is used to assess whether component splitting has occurred in an NCPD.
  - For a mode-3 model of rank R, $\hat{\mathcal{X}} = [[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]]$, the mINS is defined as

$$mINS = \max_{i,j,i \neq j} \left(g_{i,j}\left(A^{(1)}\right) \cdot g_{i,j}\left(A^{(2)}\right) \right), \tag{4}$$

where $g_{i,j}(A) = S(a_i, a_j)$ is a similarity measure (such cosine or correlation) between columns i and j in matrix A.



- The plot is generated using CP-OPT library in Tensor Toolbox where *the objective used for the decomposition is the Frobenius norm (NCPD-F).*

Motivation behind Optimal transport-based Tensor Factorization

- We are thus motivated to modify NCPD so that splitting does not occur before all patterns are identified.

- This is where using the recently proposed optimal transport based tensor factorization namely Wasserstein tensor factorization [4] [2] may prove useful.
  - Loss function based optimal transport theory incorporates the underlying geometry of the data, thus is able to recover all components without distortions.

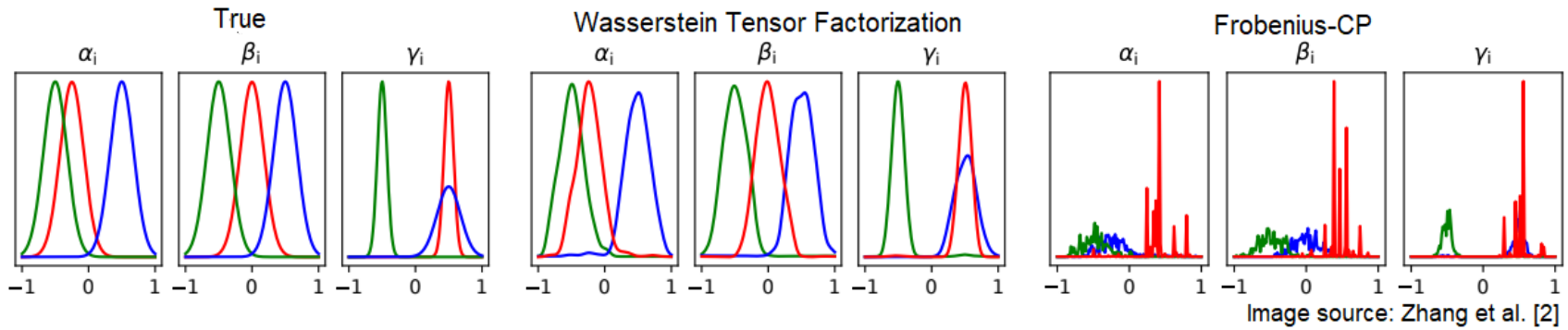


Figure: $\mathcal{X}_{\text{True}} = \sum_{i=1}^{3} \alpha_i \circ \beta_i \circ \gamma_i$, where $\{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{3}$ are univariate Gaussian. $\circ$ denotes the outer product of vectors.

[4] Zhang et. al A unified framework for non-negative matrix and tensor factorisations with a smoothed wasserstein loss, https://arxiv.org/abs/2104.01708

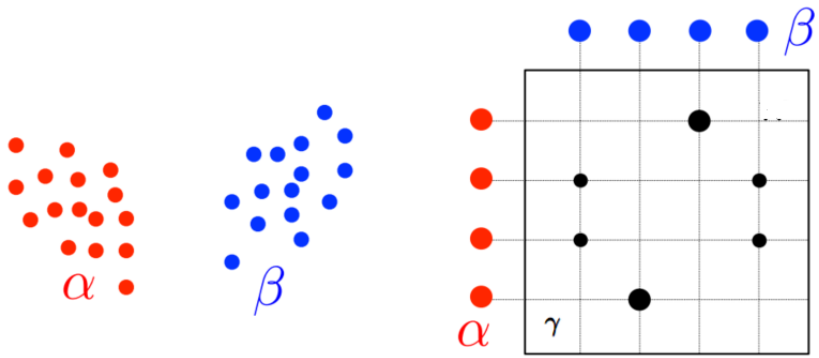


image source: Computational OT by Cuturi and Peyre [2]

- Given α and β are probability distribution supported on a finite set Ω, with Cardinality$(\Omega) = n$.
- Let $C \in \mathbb{R}^{n \times n}$ be the ground cost matrix with C_{ij} denoting the cost of transporting mass from point i to j. The optimal transport distance is defined via

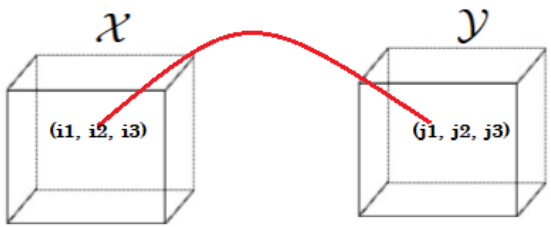
$$\mathrm{OT}(\alpha, \beta) \overset{\text{def}}{=} \min_{\gamma \in \Gamma(\alpha,\beta)} \langle C, \gamma \rangle \overset{\text{def}}{=} \min_{\gamma \in \Gamma(\alpha,\beta)} \sum_{i,j}^{n,n} C_{ij}\gamma_{ij} \tag{5}$$

  where $\Gamma(\alpha, \beta) = \{\gamma \in \mathbb{R}^{n \times n}_{\geq 0} : \gamma\, \mathbf{1} = \alpha, \gamma^\top\, \mathbf{1} = \beta\}$ is the set of all couplings of (α, β), and γ is the optimal coupling.
- For ground cost C_{ij}, one may choose squared of Euclidean distance between the points in support of α and β (this is also known Wasserstein-2 distance).
- **Drawbacks**: Computationally expensive $\mathcal{O}(n^3)$.

- Entropy regularized optimal transport [3]

$$\mathrm{OT}_\varepsilon(\alpha, \beta) := \inf_{\gamma \in \Gamma(\alpha,\beta)} \langle C, \gamma \rangle + \varepsilon E(\gamma), \tag{6}$$

  - $E(\gamma) = \langle \gamma, \log \gamma - 1 \rangle$ is the entropy of the coupling matrix.
  - $\varepsilon > 0$ is the entropy regularizer.
  - As $\varepsilon \to 0$, $\mathrm{OT}_\varepsilon \to \mathrm{OT}$.
  - Advantages
    - Smooth and strongly convex.
    - Computationally cheaper, e.g. Sinkhorn algorithm has complexity $\mathcal{O}(n^2 \log n\varepsilon^{-2})$ [4].



- Given two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. Consider $\mathcal{X}, \mathcal{Y} \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the set of all probability distributions over the multi-index support set Ω. Optimal transport distance between X and Y is,

$$\mathrm{OT}(\mathcal{X}, \mathcal{Y}) := \inf_{\gamma \in \Gamma(\mathcal{X}, \mathcal{Y})} \langle C, \gamma \rangle. \tag{7}$$

- $\Gamma(\mathcal{X}, \mathcal{Y})$ is the set of all possible couplings between $\mathcal{X}$ and $\mathcal{Y}$ and C is the ground cost tensor where $C_{(i_1 i_2 i_3),(j1,j2,j3)}$ denotes the cost to couple point (i_1, i_2, i_3) to (j_1, j_2, j_3).

$$C_{(i_1 i_2 i_3),(j_1 j_2 j_3)} = \sum_{k=1}^{3} C^{(k)}_{i_k, j_k}, \tag{8}$$

  where $C^{(k)}$ is the cost matrix for the *kth* mode of tensor $\mathcal{X}$.
- *Entropic OT between $\mathcal{X}$ and $\mathcal{Y}$*

$$\mathrm{OT}_\varepsilon(\mathcal{X}, \mathcal{Y}) := \inf_{\gamma \in \Gamma(\mathcal{X}, \mathcal{Y})} \langle C, \gamma \rangle + \varepsilon E(\gamma). \tag{9}$$

Zhang et al. A unified framework for non-negative matrix and tensor factorisations with a smoothed wasserstein loss, https://arxiv.org/abs/2104.01708

- Decomposition of tensor $\mathcal{X}_{\geq 0} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ using entropy regularized optimal transport [2]

$$\min_{\lambda, A^{(1)}, A^{(2)}, A^{(3)} \geq 0} \mathrm{OT}_{\varepsilon}(\mathcal{X}, [[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]]), \tag{10}$$

  where $A^{(i)} \in \mathbb{R}^{n_i \times R}_{\geq 0}$, $i = 1, 2, 3$, $[[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)}$.

- [2] solves the following to decompose $\mathcal{X}$

$$\min_{\lambda, A^{(1)}, A^{(2)}, A^{(3)}} \mathrm{OT}_{\varepsilon}(\mathcal{X}, [[\lambda; A^{(1)}, A^{(2)}, A^{(3)}]]) + \sum_{i=1}^{3} \rho_i E_{\Sigma_i}(A^{(i)}) \tag{11}$$

  where $E_{\Sigma_i}(A^i) = E(A^{(i)}) + \ell(A^{(i)} \in \Sigma_i)$, Σ_i is the constraint set of $A^{(i)}$ and $\ell(\cdot)$

$$\ell(x \in A) = \begin{cases} 0, \text{ if } x \in A \\ +\infty \text{ otherwise.} \end{cases}$$

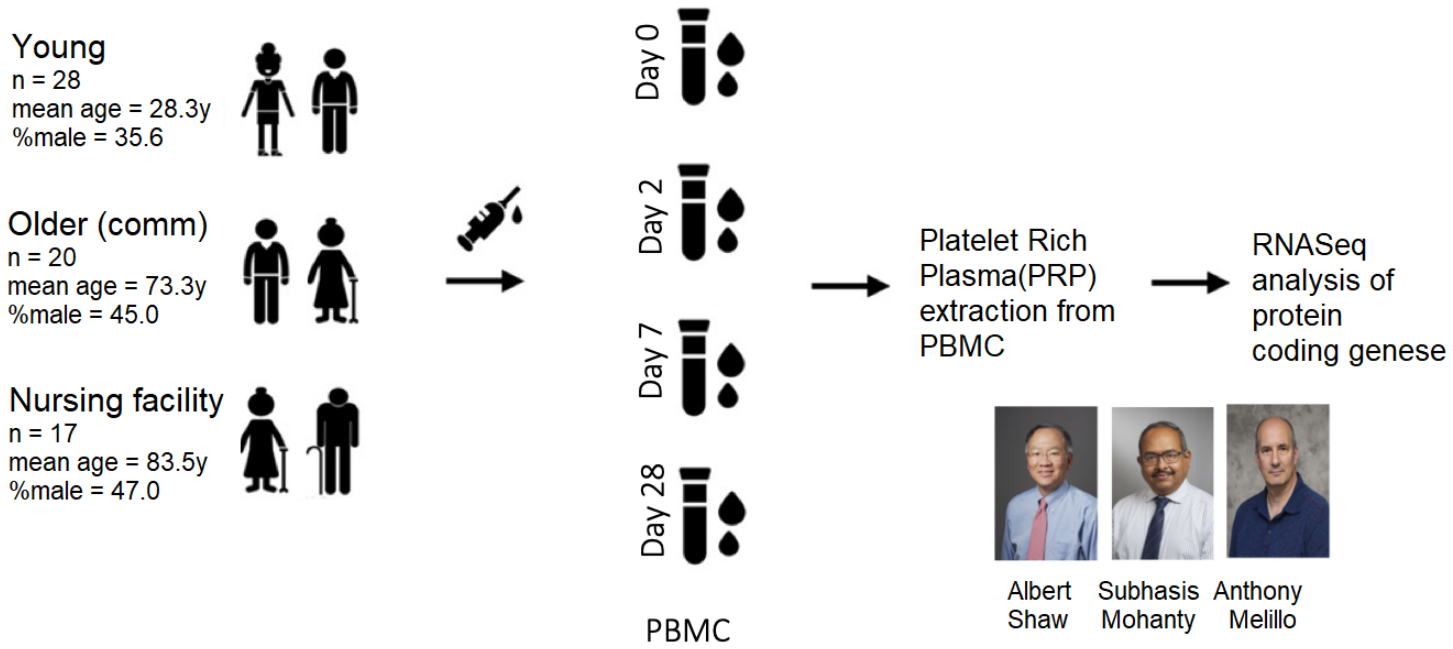
  is the indicator function.

- $E(\cdot)$ relaxes the non-negativity constraint.
- Equation 9 is solved by performing *block coordinate descent algorithm in each of the factor matrices individually* [2].
- We denote this decomposition as NCPD-W.

Zhang et al. A unified framework for non-negative matrix and tensor factorisations with a smoothed wasserstein loss, https://arxiv.org/abs/2104.01708

Platelet RNASeq vaccination time-course

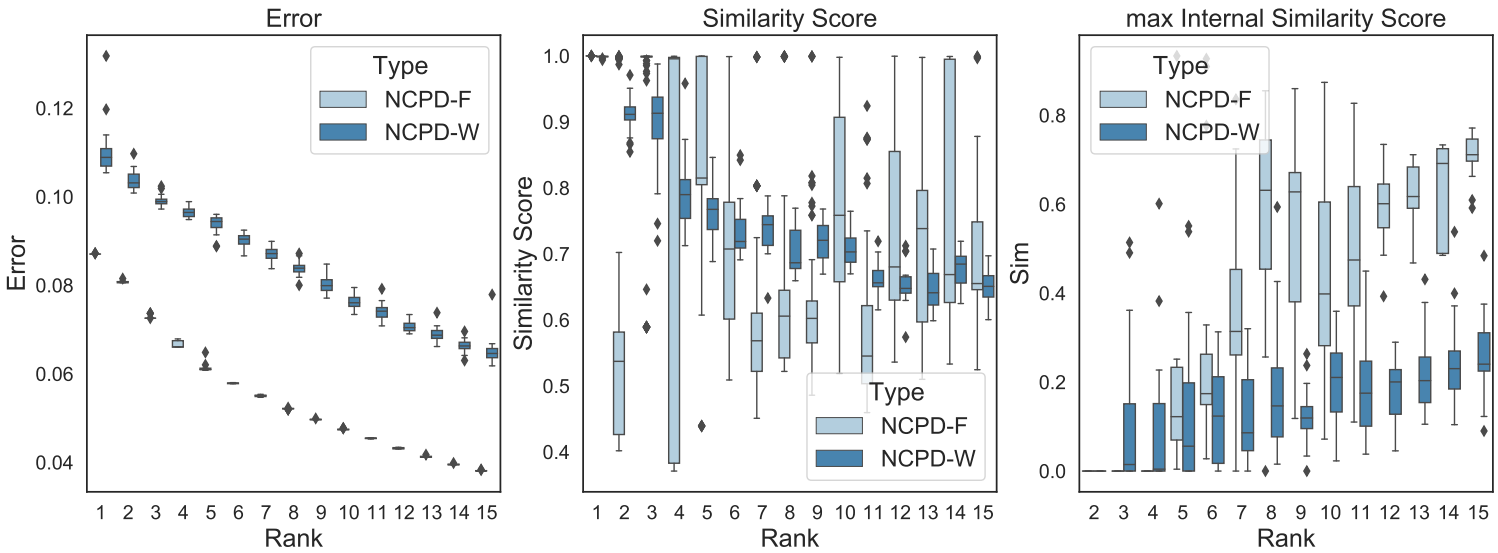
Recruitment: subjects getting the flu vaccine (2018) from three cohorts



- A tensor framework of genes $\times$ subject $\times$ day is used to store the data.
- The final dimensions of the tensor is 500 x 54 x 4. (500 most variable genes, 54 subjects with data for all days.)
- Hyperparameters used in NCPD-W.

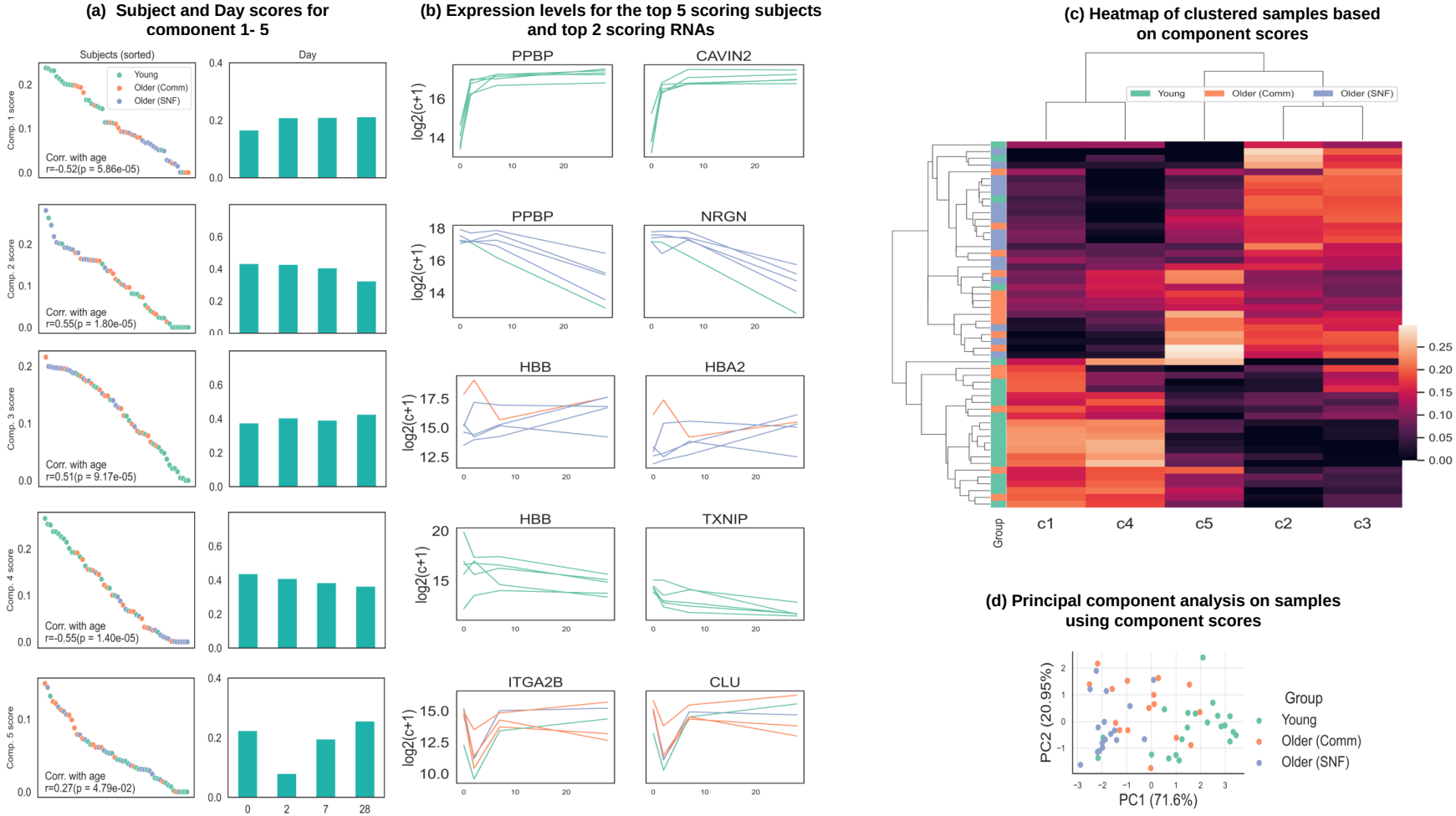
Hyperparameter	Value
ε	0.1
ρ	0.01
learning rate	1

- Comparison between NCPD-F and NCPD-W in terms of normalized Frobenius error, similarity score and mINS.



- **Observations**:
  - Error decreases as rank increases for both NCPD-F and NCPD-W. However, for any given rank R, NCPD-W has slightly higher error compared to NCPD-F.
  - Similarity score decreases almost monotonically as rank R increases for NCPD-W.
  - As rank (R) increases, mINS scores increases for both NCPD-F and NCPD-W.
  - mINS scores for NCPD-F are higher than NCPD-W when $R > 5$, which may indicate component splitting for NCPD-F.

Tensor components related to platelet activation and age group (NCPD-F, Rank = 5)

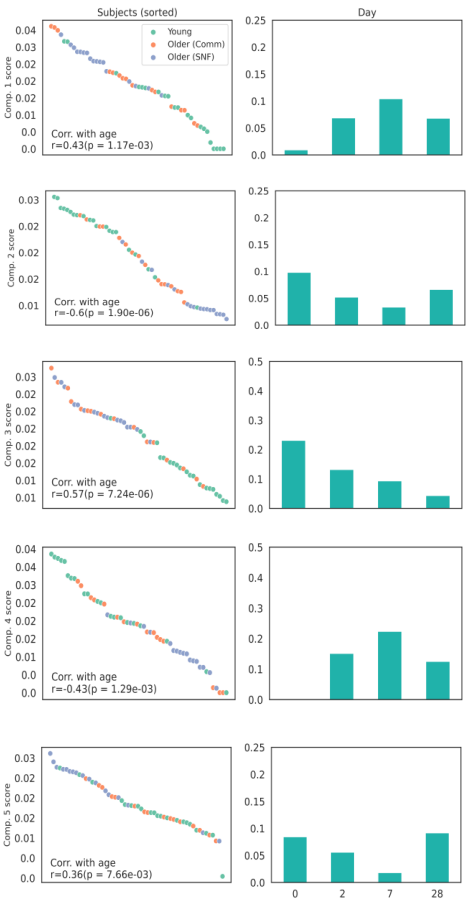


- For $R = 5$, NCPD-F successfully discovers all 5 components with no splitting.
- Subject component scores show a strong association with the three groups (Figure (c) and (d)).

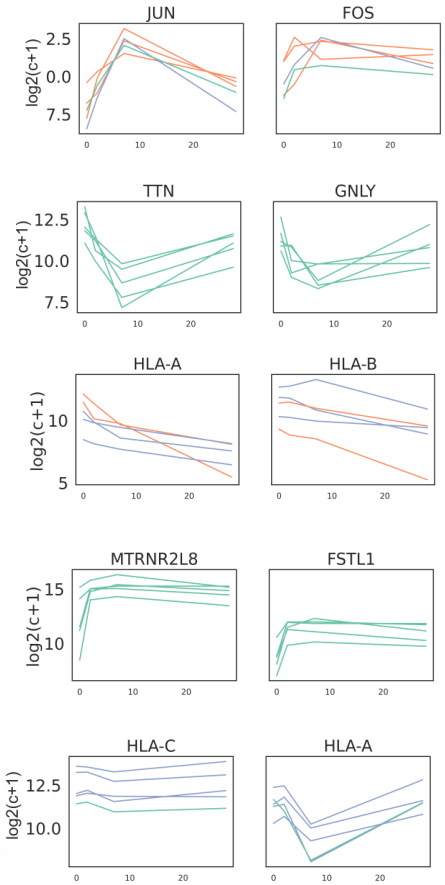
Konstorum et al. Platelet response to influenza vaccination reflects effects of aging, bioRxiv (2022) 2022.04.06.487196

Tensor components related to platelet activation and age group (NCPD-W, Rank = 5)

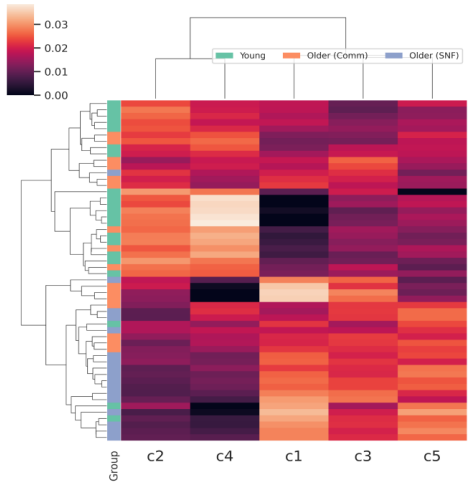
(a) Subject and Day scores for component 1- 5



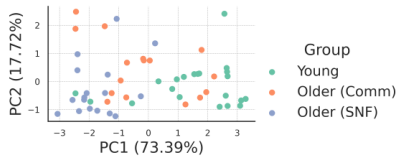
(b) Expression levels for the top 5 scoring subjects and top 2 scoring RNAs



(c) Heatmap of clustered samples based on component scores



(d) Principal component analysis on samples using component scores



(f) Pathways (only found with NCPD-W)

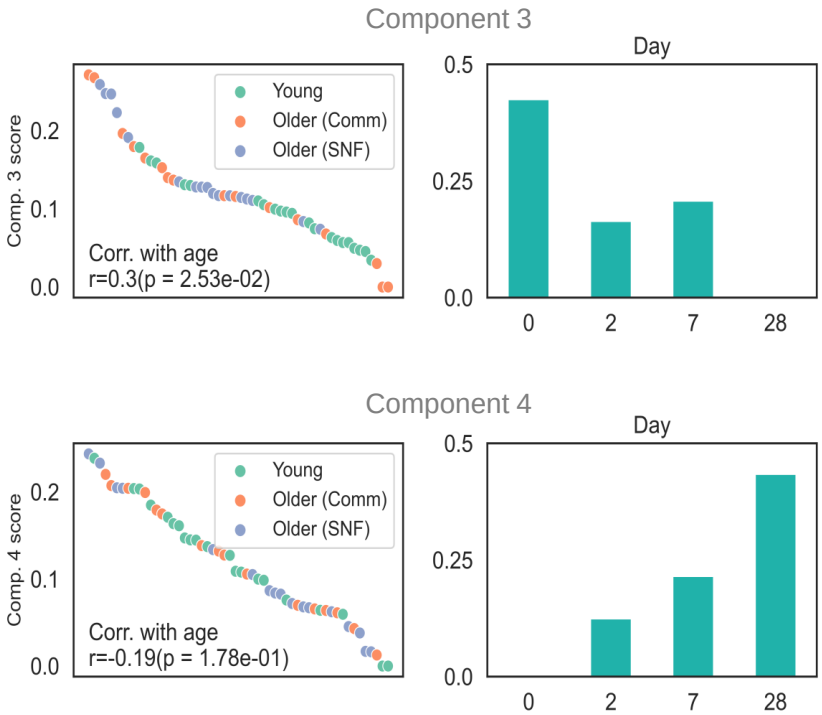
Pathway (component no.)	Adjusted p-value
Endosomal/Vacuolar pathway (1)	0.000236
Integrin cell surface interactions (2)	0.003791
Antigen processing-Cross presentation (2)	0.009523
Formation of Fibrin Clot (Clotting Cascade) (2)	0.012693

- For $R = 5$, NCPD-W discovers few new patterns ones e.g., 1,2, 4.
- Day scores closely follows the original expression levels for top scoring genes.
- Subject component scores show a strong association with the three groups (Figure (c) and (d)).
- NCPD-W components found association with few new pathways which were not discovered with NCPD-F.

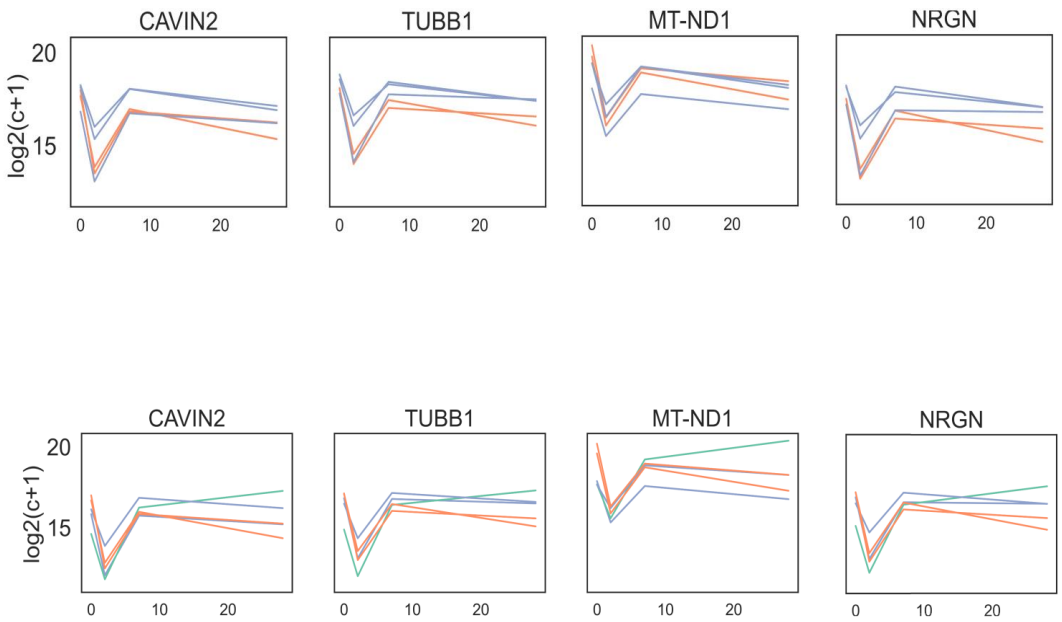
- Effect of component splitting when NCPD-F decomposition with higher rank.

Tensor components related to platelet activation and age group (NCPD-F, Rank = 9)

(a) Subject and day scores for component 3 and 4



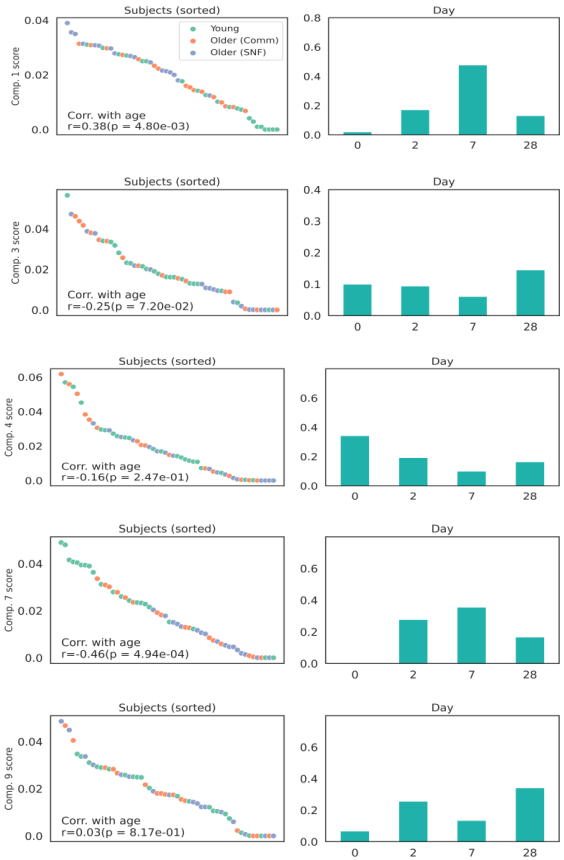
(b) Expression levels for top 5 scoring subjects and top few scoring RNAs



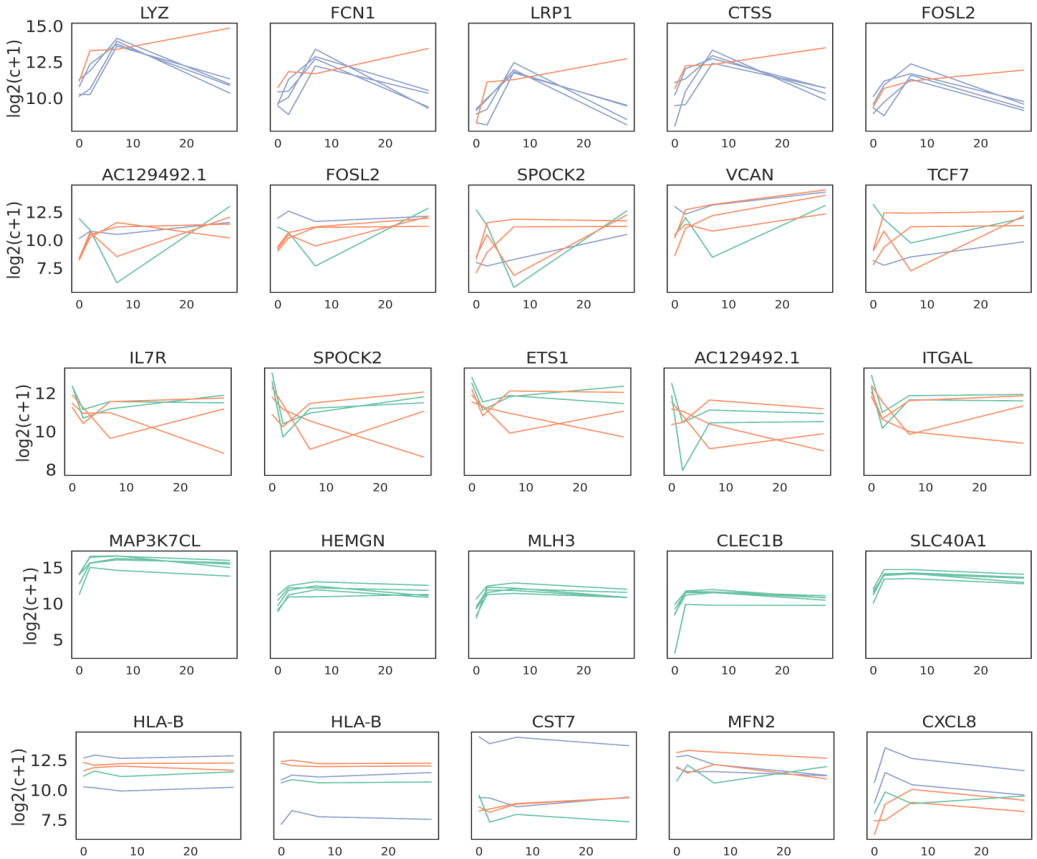
- Day patterns associated with component 3 and 4 do not follow the original gene expression individually.
- Added components will follow the gene expression, which clearly indicates the NCPD-F suffers from component splitting.

Tensor components related to platelet activation and age group (NCPD-W, Rank = 9)

(a) Subject and day scores for component 1,3, 4, 5, 9



(b) Expression levels for the top 5 scoring subjects and few scoring RNAs



- Day pattern follow the trend of the expression of most of the top scoring genes.
- NCPD-W is able to avoid component splitting which may help to discover new patterns.

Summary of contributions

- Propose using optimal transport based tensor decomposition (NCPD-W) for the omics-time course data.
- Shown that NCPD-W can discover patterns which are not discovered by NCPD-F for the rank 5 decomposition.
- Shown that For higher rank decomposition, NCPD-W shows less component splitting compared to NCPD-F.

Next steps

- Finding the appropriate set of hyperparameters for NCPD-W in order to improve the decomposition quality.
- Understanding why some genes (like the HLAs in this study), do not follow patterns.
- For NCPD-W, search for another ground cost function which can act as prior in order to reduce component splitting.
- Adding regularization in the tensor decomposition objective function in order to reduce component splitting.

Acknowledgement

Yale University
Yale School of Medicine

Department of Pathology
Steven Kleinstein
Anna Konstorum

Tufts University

Dept. of Mathematics
Misha Kilmer

Dept. Electrical & Computer Engineering
Shuchin Aeron

Thank you!

Questions?

Shoaib Bin Masud
supported by
U.S. Army DEVCOM Soldier Center
Cooperative Agreement Number
W911QY-19-2-0003.

References I

[1] A Konstorum, S Mohanty, Y Zhao, A Melillo, B Vander Wyk, A Nelson, S Tsang, TP Blevins, RB Belshe, DG Chawla, et al.
Platelet response to influenza vaccination reflects effects of aging.
bioRxiv, 2022.

[2] Stephen Y Zhang.
A unified framework for non-negative matrix and tensor factorisations with a smoothed wasserstein loss.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4203, 2021.

[3] Marco Cuturi.
Sinkhorn distances: Lightspeed computation of optimal transport.
Advances in neural information processing systems, 26, 2013.

[4] Gabriel Peyré, Marco Cuturi, et al.
Computational optimal transport: With applications to data science.
Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.