From Data Mining using Tensor Factorizations to Multimodal Data Mining using Coupled Matrix/Tensor Factorizations

Evrim Acar

Simula Metropolitan Center for Digital Engineering

Oslo, Norway

Nordic Probabilistic AI School

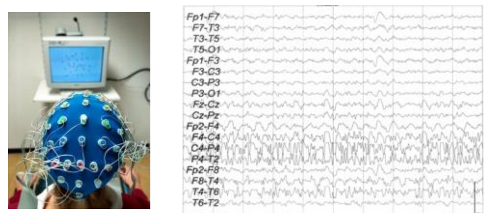
June 16, 2021

simula

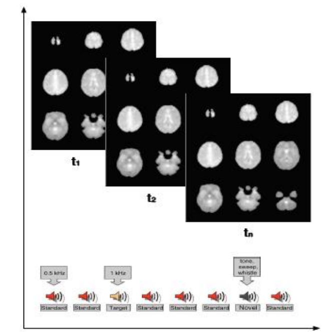
Motivation$_1$: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery

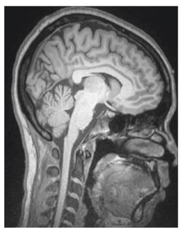
Neuroscience: Due to the complexity of the brain, we often need to use **multiple neuroimaging techniques** to better **understand neural activities**. Neuroimaging techniques provide **complementary** information at different scales.

EEG (Electroencephalography)

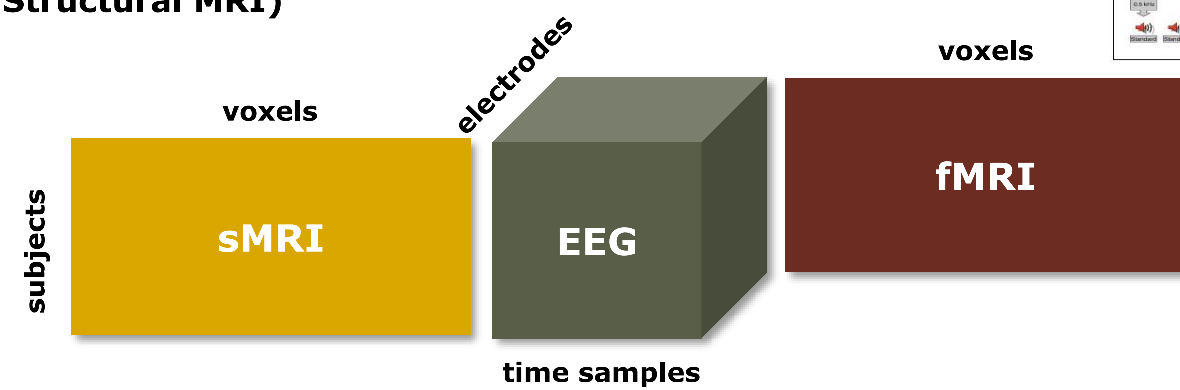


fMRI (Functional Magnetic Resonance Imaging)





sMRI (Structural MRI)

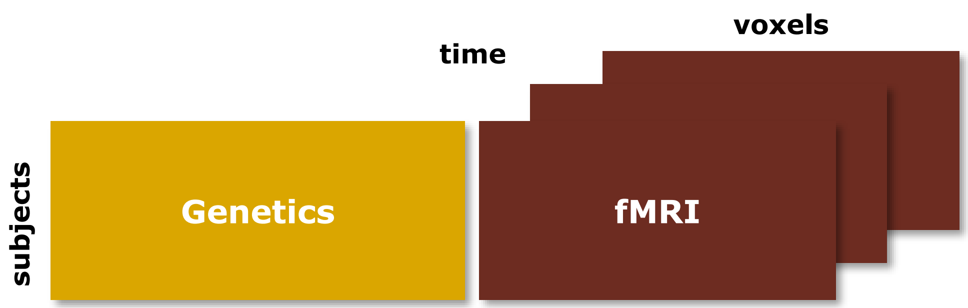


Can we jointly analyze these data sets and capture functional/structural patterns that differ between healthy controls and patients suffering from a psychiatric disorder?

Challenges: (i) Data sets in the form of matrices and higher-order arrays
(ii) Common patterns as well as patterns visible only in one modality
(iii) Need for interpretable and robust patterns

Motivation$_2$: Understanding dynamic systems requires joint analysis of both static and time-evolving data sets

Neuroscience: By jointly analyzing **time-evolving** neuroimaging signals as well as **static** data sets, we can better characterize **how brain networks evolve in time** and for different psychiatric disorders.

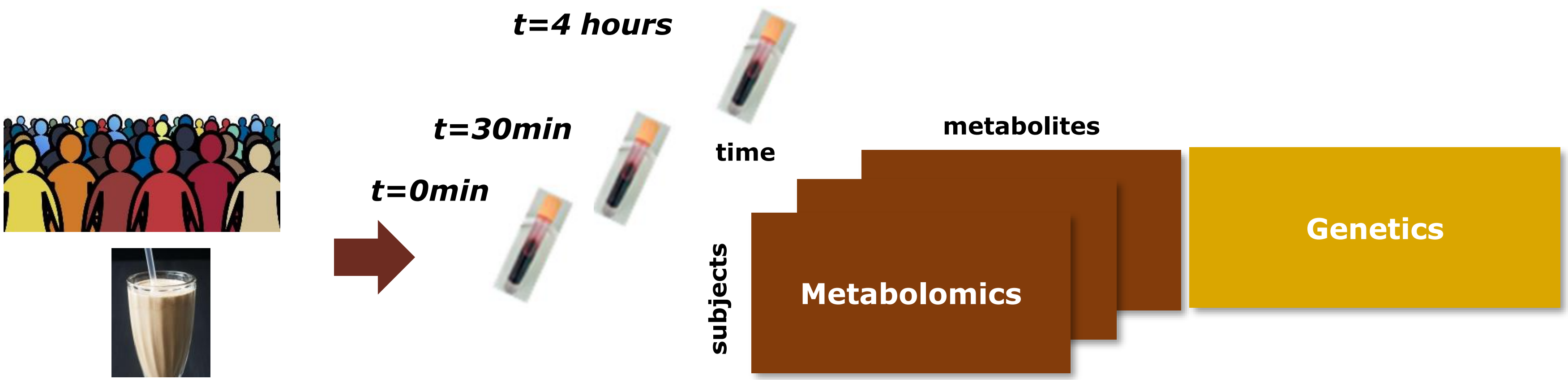


How can we jointly analyze data from these static and dynamic modalities and capture underlying patterns, e.g., brain networks, as well as how those networks evolve in time?

Challenges: ***(i) Data sets in the form of matrices and higher-order arrays***
(ii) Common patterns as well as patterns visible only in one modality
(iii) Need for interpretable and robust patterns
(iv) Fusion models need to account for the temporal evolution of the patterns

Motivation$_2$: Understanding dynamic systems requires joint analysis of both static and time-evolving data sets

Omics: Metabolomics is significant in terms of revealing the link between genotypes and phenotypes, and understanding **changes in metabolism** will provide better understanding of many diseases.

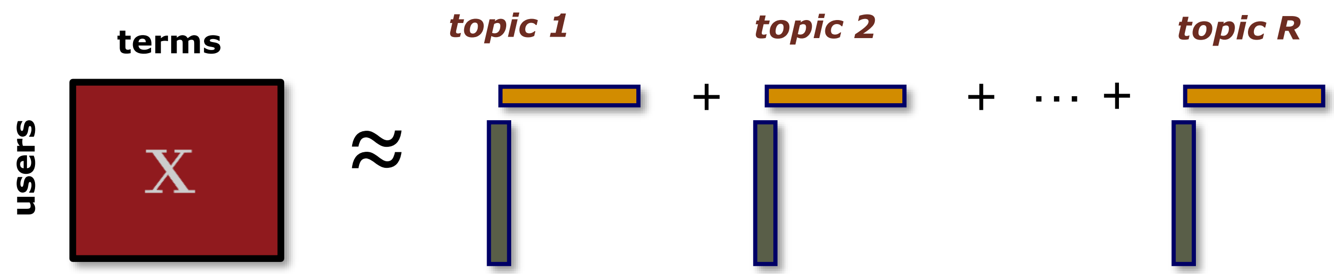


How can we capture metabolic networks and their temporal evolution?
Do the dynamics of metabolic networks differ across individuals or various subgroups?
Can we use the captured temporally evolving patterns for early diagnosis of diseases?

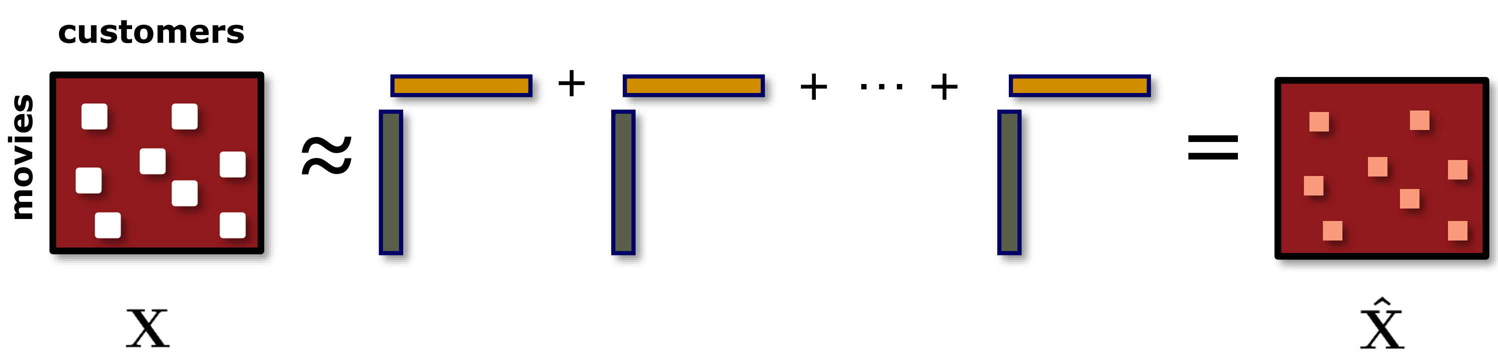
Challenges: ***(i) Data sets in the form of matrices and higher-order arrays***
(ii) Common patterns as well as patterns visible only in one modality
(iii) Need for interpretable and robust patterns
(iv) Fusion models need to account for the temporal evolution of the patterns

Matrix Factorizations in Data Mining

Capturing underlying factors/patterns

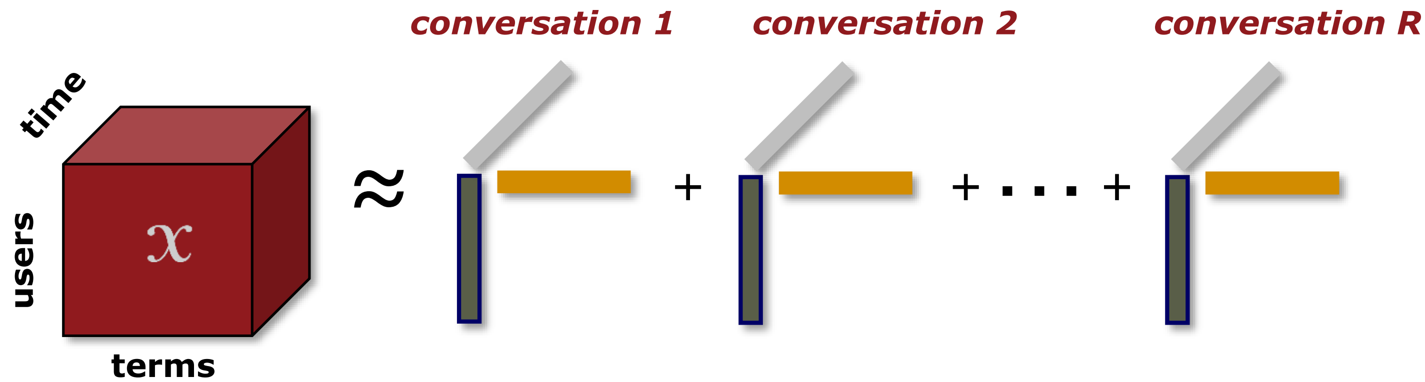


Incomplete Matrix Factorization/ Matrix Completion

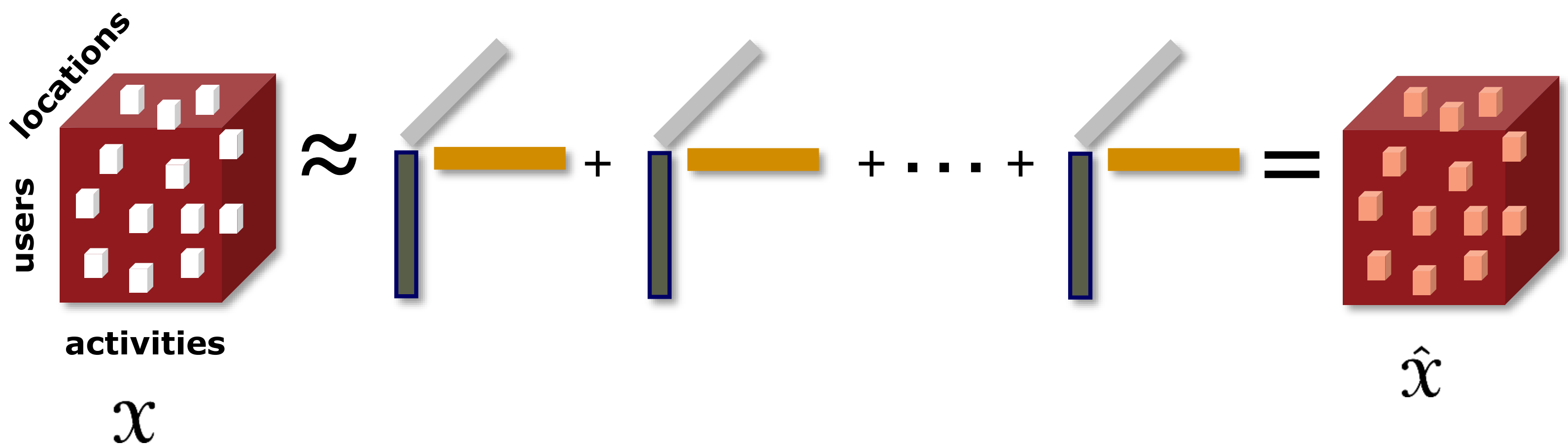


Data is often multi-way!

Capturing underlying factors/patterns

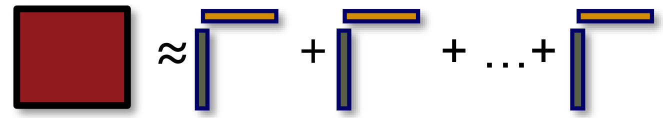


Incomplete Tensor Factorization/ Tensor Completion

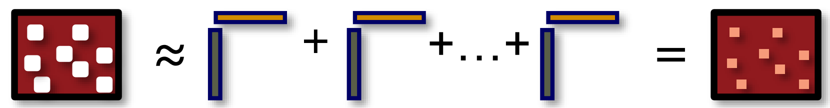


Matrix Factorizations in Data Mining

Matrix Factorizations

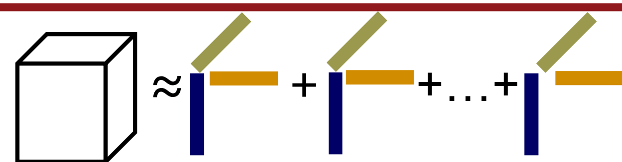


Matrix Completion

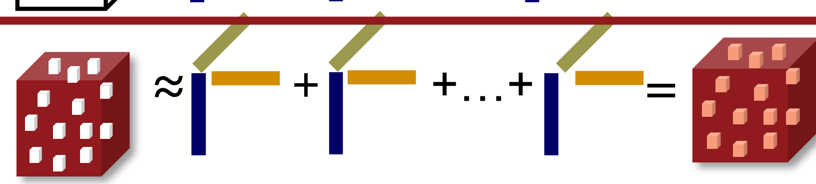


Data sets are often multi-way: Tensor Factorizations

Tensor Factorizations



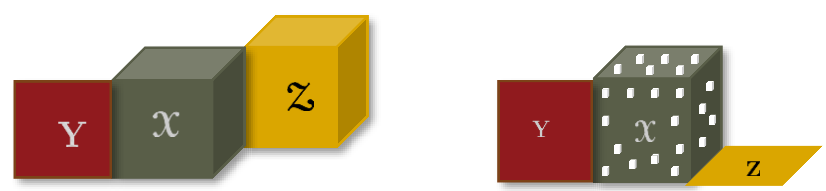
Tensor Completion



Data sets often come from multiple sources: Data Fusion

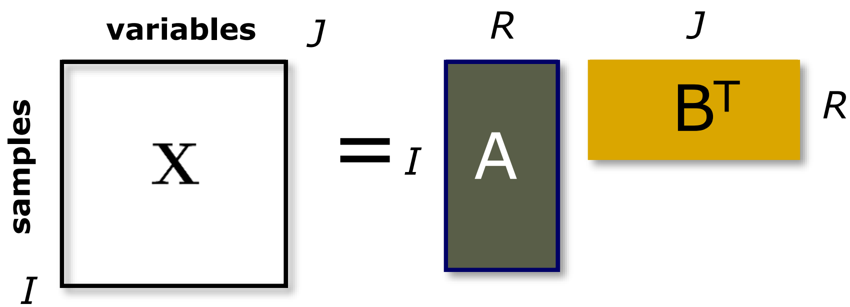
Coupled Matrix Factorizations

Coupled Tensor Factorizations



Matrix Factorizations

Matrix Factorizations, e.g., Singular Value Decomposition (SVD), Nonnegative Matrix Factorization (NMF), Independent Component Analysis (ICA), are commonly used in data mining to capture the underlying structures in data sets.



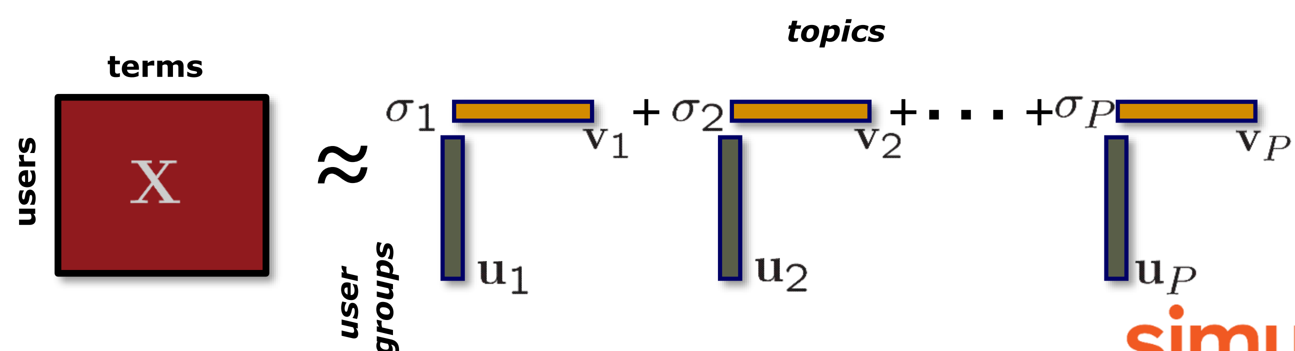
$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{a}_r \mathbf{b}_r^{\mathsf{T}} = \mathbf{A}\mathbf{B}^{\mathsf{T}}$$

$$\mathbf{A} \in \mathbb{R}^{I \times R} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_R \end{bmatrix}$$

$$\mathbf{B} \in \mathbb{R}^{J \times R} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_R \end{bmatrix}$$

For instance, we may compute truncated SVD of a *users* by *terms* matrix and identify the topics and user groups talking about those topics.

$$\mathbf{X} \approx \sum_{r=1}^{P} \sigma_r \mathbf{u}_r \mathbf{v}_r^{\mathsf{T}} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$$



simula

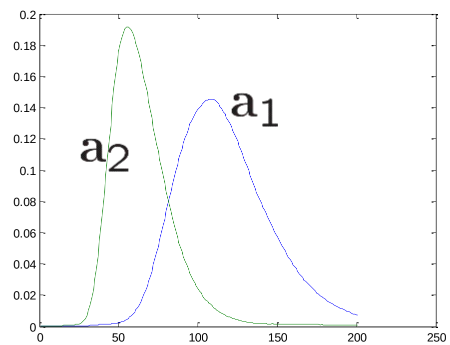
Uniqueness is an issue:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathsf{T}} = \mathbf{A}\mathbf{M}\mathbf{M}^{-1}\mathbf{B}^{\mathsf{T}} = \bar{\mathbf{A}}\bar{\mathbf{B}}^{\mathsf{T}}$$

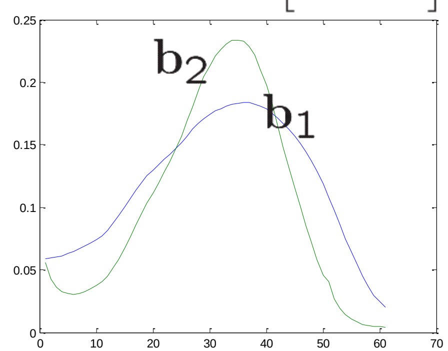
Constraints are used to deal with the uniqueness problem, e.g., SVD. However, factorizations with constraints may not be meaningful in terms of the application.

True factors

$$\mathbf{A} \in \mathbb{R}^{201\times 2} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{bmatrix}$$

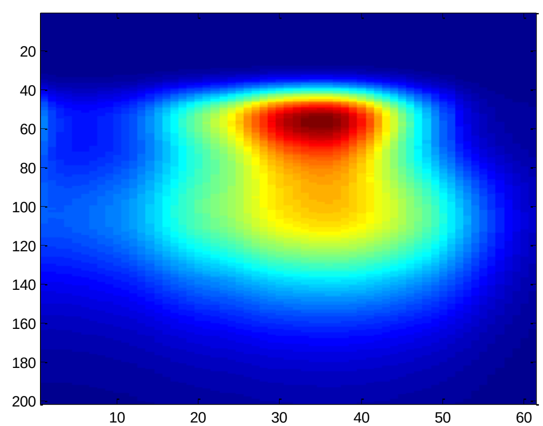


$$\mathbf{B} \in \mathbb{R}^{61\times 2} = \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}$$



Data matrix

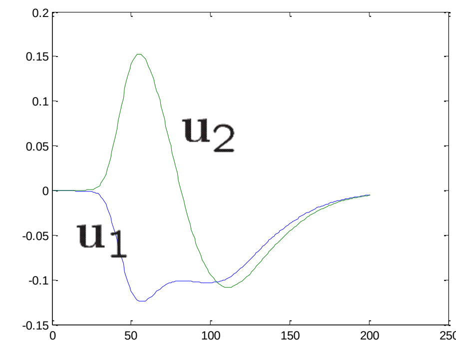
$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathsf{T}}$$

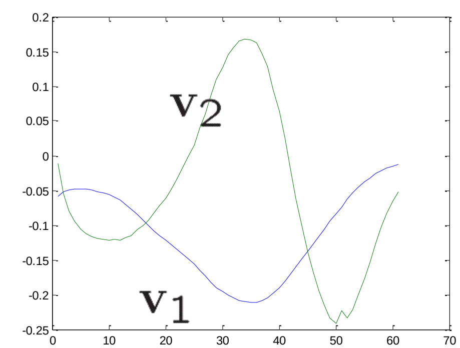


Given X, can we recover the true factors?

SVD captures...

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$$





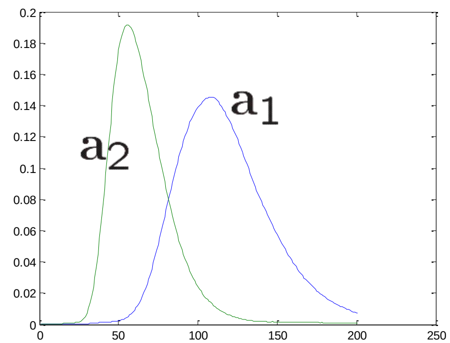
Uniqueness is an issue:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\mathsf{T} = \mathbf{A}\mathbf{M}\mathbf{M}^{-1}\mathbf{B}^\mathsf{T} = \bar{\mathbf{A}}\bar{\mathbf{B}}^\mathsf{T}$$

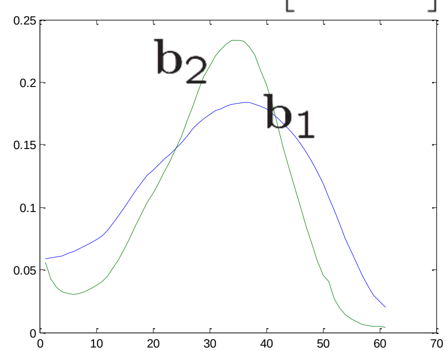
Constraints are used to deal with the uniqueness problem, e.g., SVD. However, factorizations with constraints may not be meaningful in terms of the application.

True factors

$$\mathbf{A} \in \mathbb{R}^{201\times 2} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{bmatrix}$$

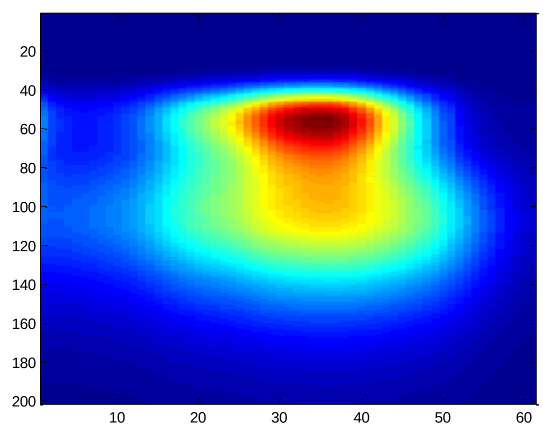


$$\mathbf{B} \in \mathbb{R}^{61\times 2} = \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}$$



Data matrix

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\mathsf{T}$$

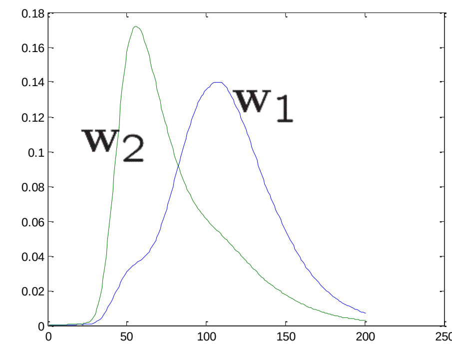


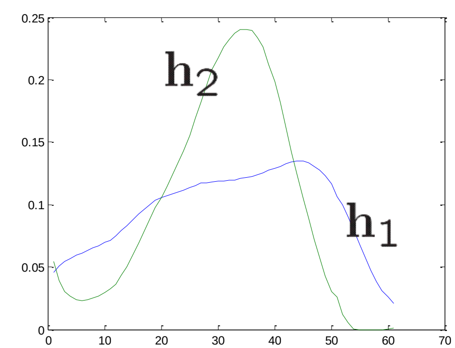
Given X, can we recover the true factors?

NMF captures...

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{W}\mathbf{H}^\mathsf{T}$$

$$w_{ir}, h_{jr} \geq 0$$





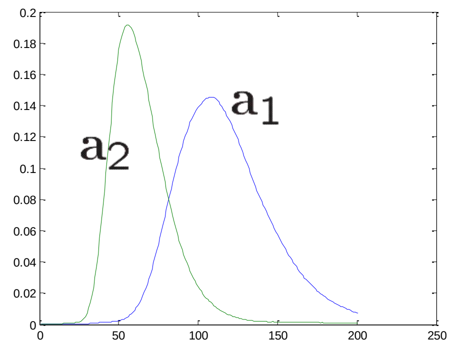
Uniqueness is an issue:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\top = \mathbf{A}\mathbf{M}\mathbf{M}^{-1}\mathbf{B}^\top = \bar{\mathbf{A}}\bar{\mathbf{B}}^\top$$

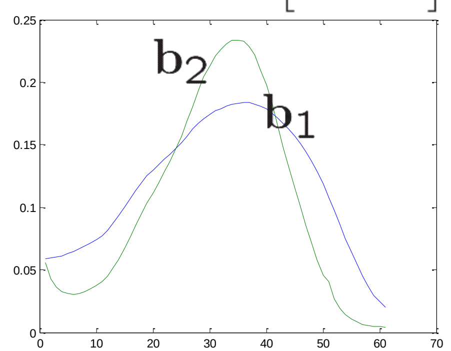
Constraints are used to deal with the uniqueness problem, e.g., SVD. However, factorizations with constraints may not be meaningful in terms of the application.

True factors

$$\mathbf{A} \in \mathbb{R}^{201\times 2} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{bmatrix}$$

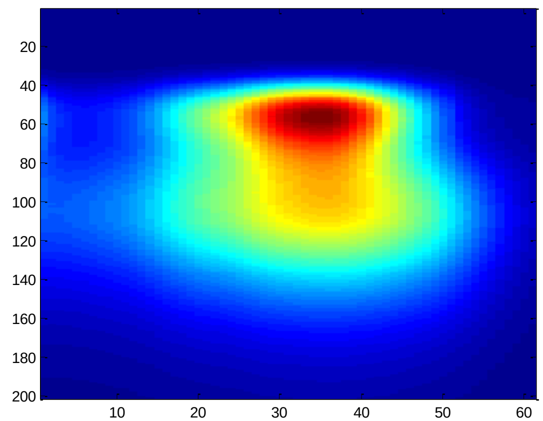


$$\mathbf{B} \in \mathbb{R}^{61\times 2} = \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}$$



Data matrix

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\top$$

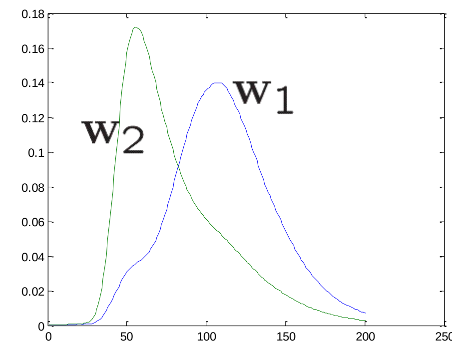


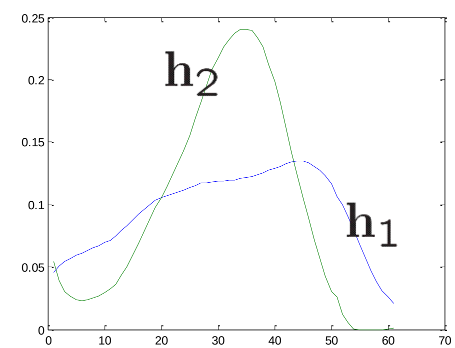
Given X, can we recover the true factors?

NMF captures…

$$\mathbf{X} \approx \hat{\mathbf{X}} = \underbrace{\mathbf{W}\mathbf{M}}_{\bar{\mathbf{W}}}\underbrace{\mathbf{M}^{-1}\mathbf{H}^\top}_{\bar{\mathbf{H}}^\top}$$

$$\bar{w}_{ir}, \bar{h}_{jr} \geq 0$$

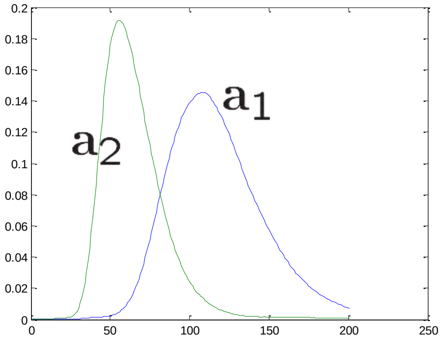




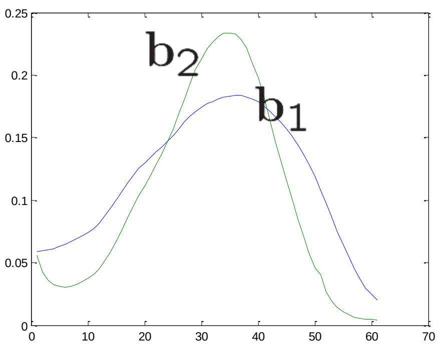
What if we have multiple matrices with the same underlying factors but in different proportions...

True factors

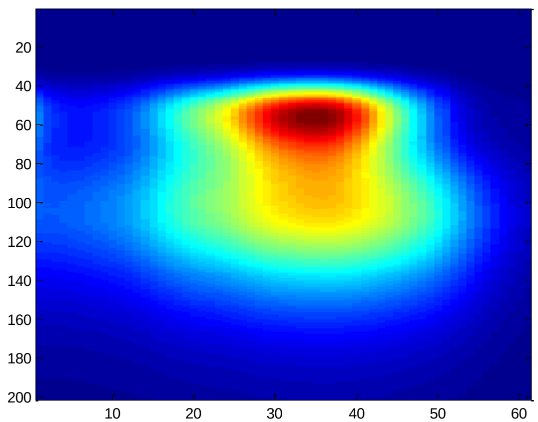
$$\mathbf{A} \in \mathbb{R}^{201 \times 2} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix}$$



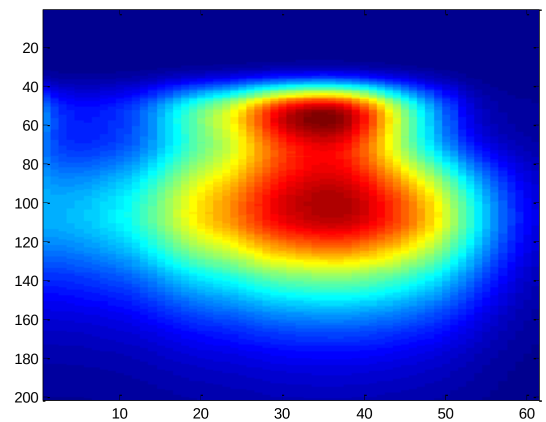
$$\mathbf{B} \in \mathbb{R}^{61 \times 2} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 \end{bmatrix}$$



$$\mathbf{X}_1 = \mathbf{A} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{B}^{\mathsf{T}}$$



$$\mathbf{X}_2 = \mathbf{A} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{B}^{\mathsf{T}}$$



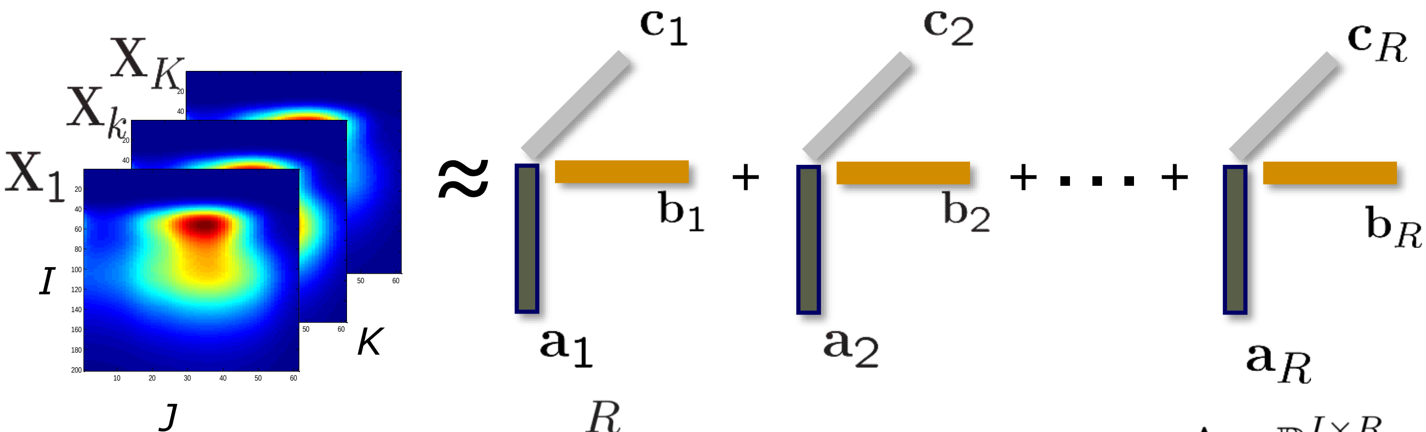
We can recover the true factors uniquely up to trivial indeterminacies, i.e., scaling and permutation.

simula

Tensor Factorizations: CANDECOMP/PARAFAC (CP)

[Hitchcock, 1927; Harshman, 1970; Carroll & Chang, 1970]

As an extension of matrix factorizations to higher-order tensors (multi-way arrays), tensor factorizations are used to extract the underlying factors in higher-order data sets. In particular, we are interested in the CP model, which represents a tensor as a sum of rank-one tensors:



$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\mathbf{X}_k \approx \mathbf{A} \begin{bmatrix} c_{k1} & & 0 \\ & \ddots & \\ 0 & & c_{kR} \end{bmatrix} \mathbf{B}^{\mathsf{T}}$$

$$\mathbf{A} \in \mathbb{R}^{I \times R} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_R \end{bmatrix}$$

$$\mathbf{B} \in \mathbb{R}^{J \times R} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_R \end{bmatrix}$$

$$\mathbf{C} \in \mathbb{R}^{K \times R} = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_R \end{bmatrix}$$

N-mode vector outer product

$$\mathcal{X} \in \mathbb{R}^{I \times J \times K}, \mathbf{a} \in \mathbb{R}^{I}, \mathbf{b} \in \mathbb{R}^{J}, \mathbf{c} \in \mathbb{R}^{K}$$

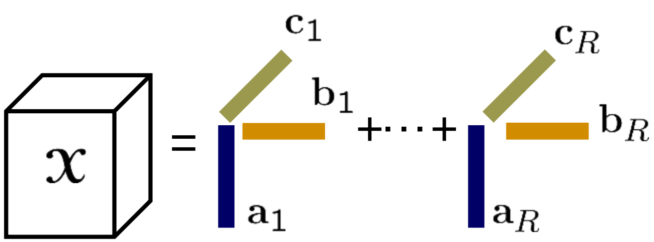
$$\mathcal{X} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \text{ iff } x_{ijk} = a_i b_j c_k$$

CP is unique under certain (mild) conditions.

Well-known sufficient conditions by [Kruskal, 1977] and [Sidiropoulos and Bro, 2000]

simula

Algorithms for fitting the CP model



The goal is to find matrices **A**, **B**, **C** that solve the following optimization problem:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2$$

while "not converged" **do**

Solve for **A** (with fixed **B**, **C**)

$$\min_{\mathbf{A}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2$$

Solve for **B** (with fixed **A** and **C**)

$$\min_{\mathbf{B}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2$$

Solve for **C** (with fixed **A** and **B**)

$$\min_{\mathbf{C}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2$$

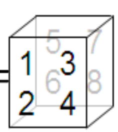
end while

Traditional approach: Alternating Least Squares (ALS)

[Harshman, 1970; Carroll & Chang, 1970]

$$\begin{aligned}\mathbf{A} &= \mathbf{X}_{(1)}(\mathbf{C}\odot\mathbf{B})((\mathbf{C}\odot\mathbf{B})^{\mathsf{T}}(\mathbf{C}\odot\mathbf{B}))^{\dagger}\\ &= \underbrace{\mathbf{X}_{(1)}(\mathbf{C}\odot\mathbf{B})}_{\text{Matricized Tensor Times Khatri-Rao Product (MTTKRP)}}(\mathbf{C}^{\mathsf{T}}\mathbf{C} * \mathbf{B}^{\mathsf{T}}\mathbf{B})^{\dagger}\end{aligned}$$

Matricization

$\mathcal{X} =$ 

$$\mathbf{X}_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$$

Khatri-Rao Product

$$\mathbf{A}\odot\mathbf{B} = \begin{bmatrix} \mathbf{a}_1\otimes\mathbf{b}_1 & \cdots & \mathbf{a}_R\otimes\mathbf{b}_R \end{bmatrix}$$

All-at-once approaches

- **Gradient-based/Quasi-Newton methods:** [Paatero, 1999; Acar et al., 2011]
- **Gauss-Newton-based approaches:** [Paatero, 1997; Tomasi and Bro, 2006; Phan et al., 2013; Sorber et al., 2013]

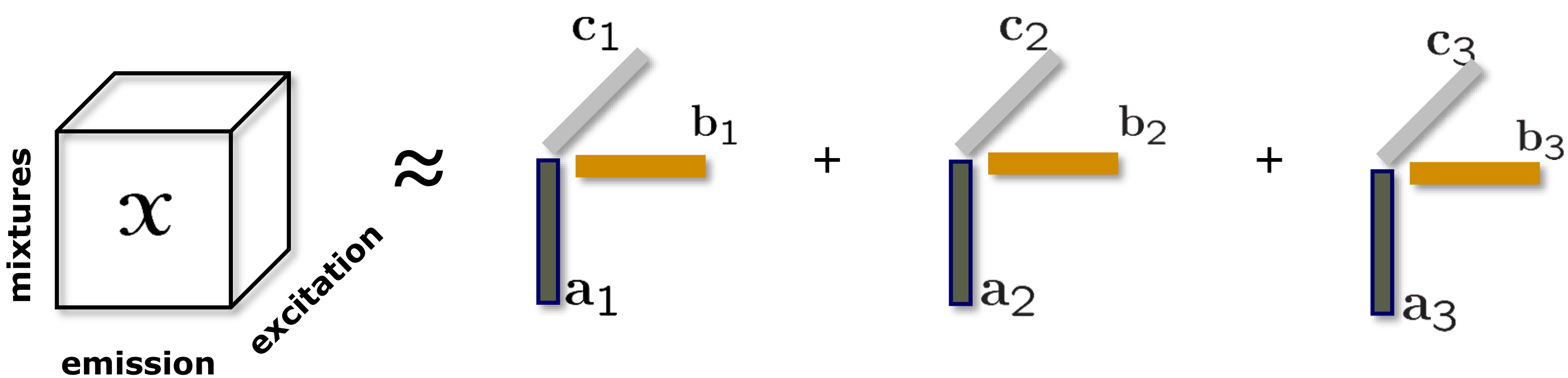
Recent advances

- **Generalized CP for different loss functions** [Hong et al., 2020]
- **Randomization and/or parallelization:** [Battaglino et al., 2018; Yang et al., 2018; Larsen and Kolda, 2020]

Chemometrics: CP can separate the chemicals in mixtures

[Andersen and Bro, *Journal of Chemometrics*, 2003]

A popular application of the CP model is the separation of individual chemicals from mixtures of chemicals measured using fluorescence spectroscopy.

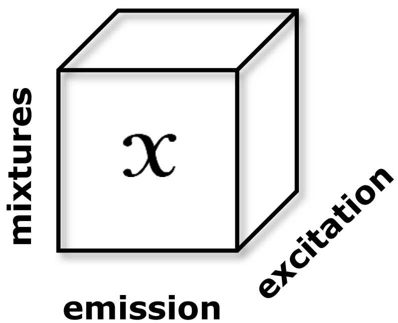


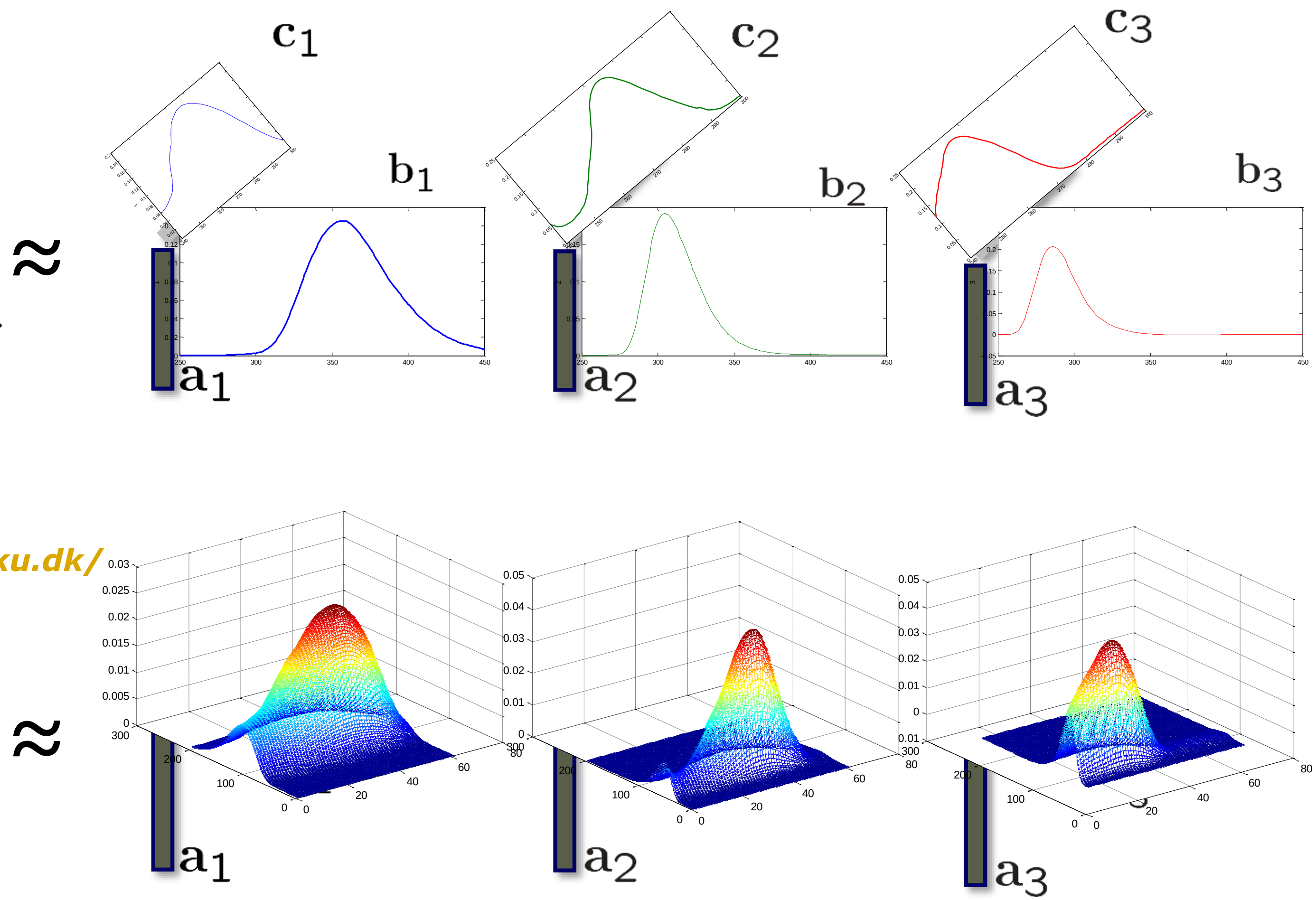
Amino Acid Data
http://www.models.life.ku.dk/

Chemometrics: CP can separate the chemicals in mixtures

[Andersen and Bro, *Journal of Chemometrics*, 2003]

A popular application of the CP model is the separation of individual chemicals from mixtures of chemicals measured using fluorescence spectroscopy.

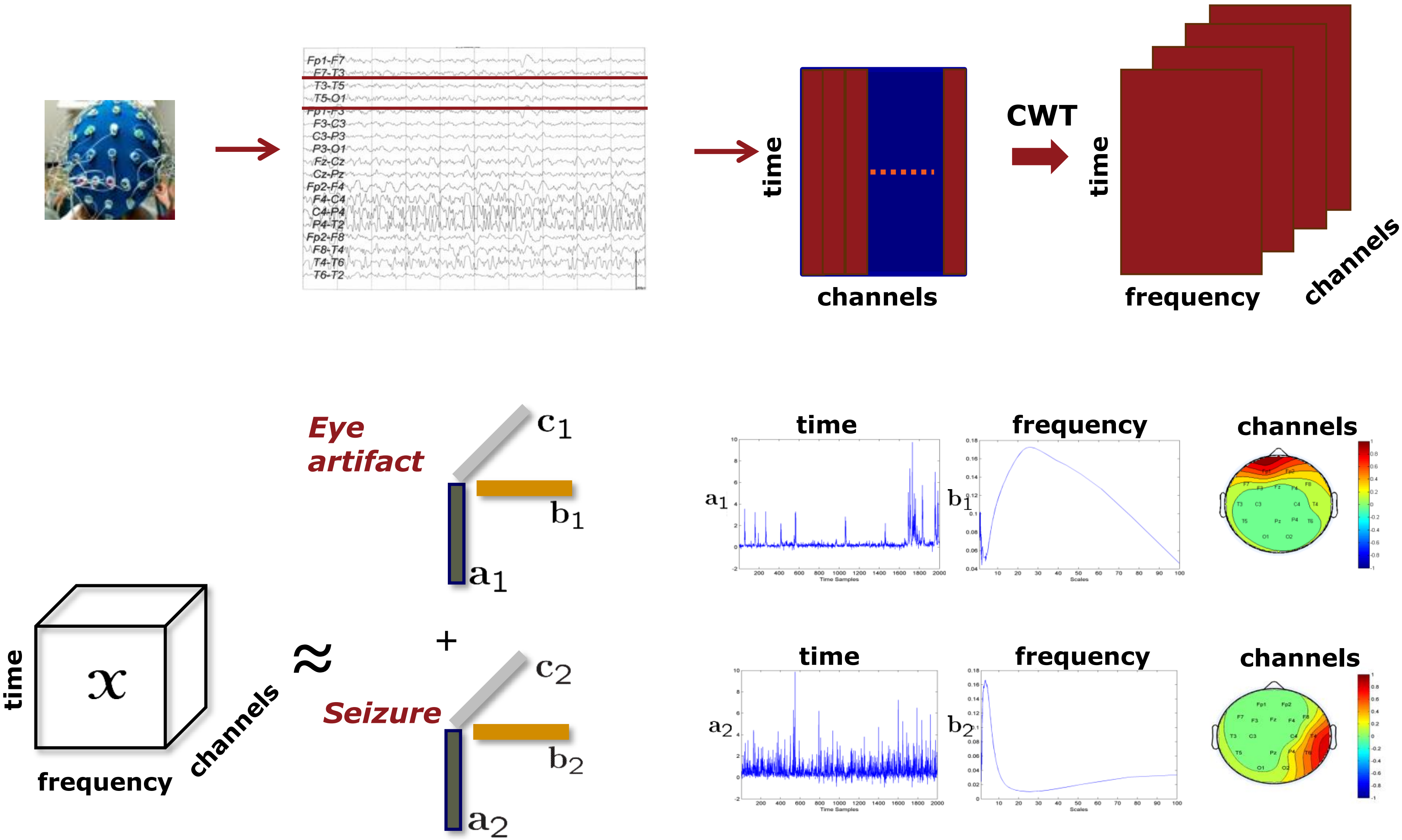




Amino Acid Data
http://www.models.life.ku.dk/

Neuroscience: CP components have been shown to localize epileptic seizures

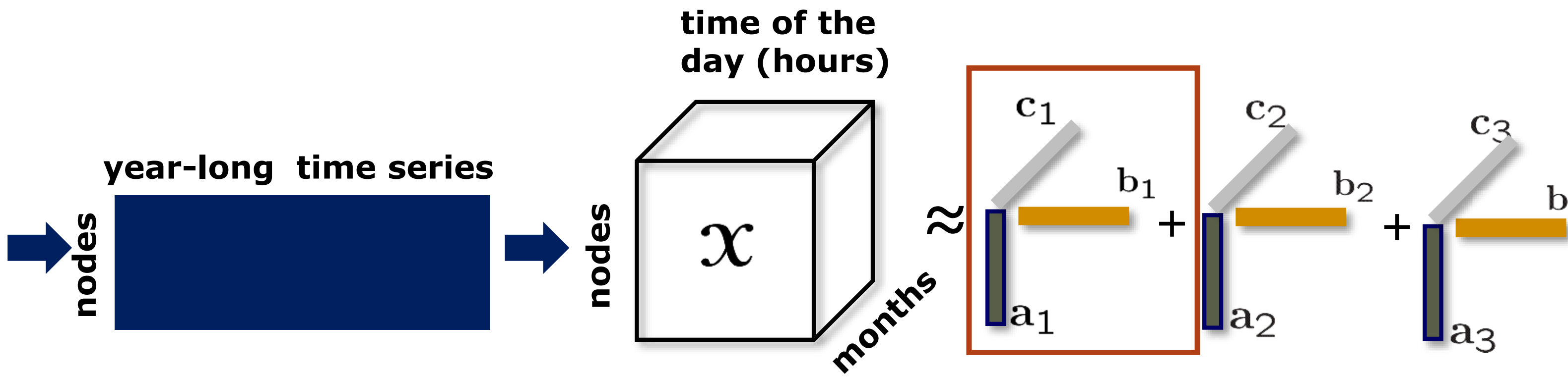
[Acar et al., *Bioinformatics*, 2007; De Vos et al., *Neuroimage*, 2007]



Communication Networks: CP reveals outage patterns in mobile broadband networks

[Fida et al., *ACM IMC*, 2019]

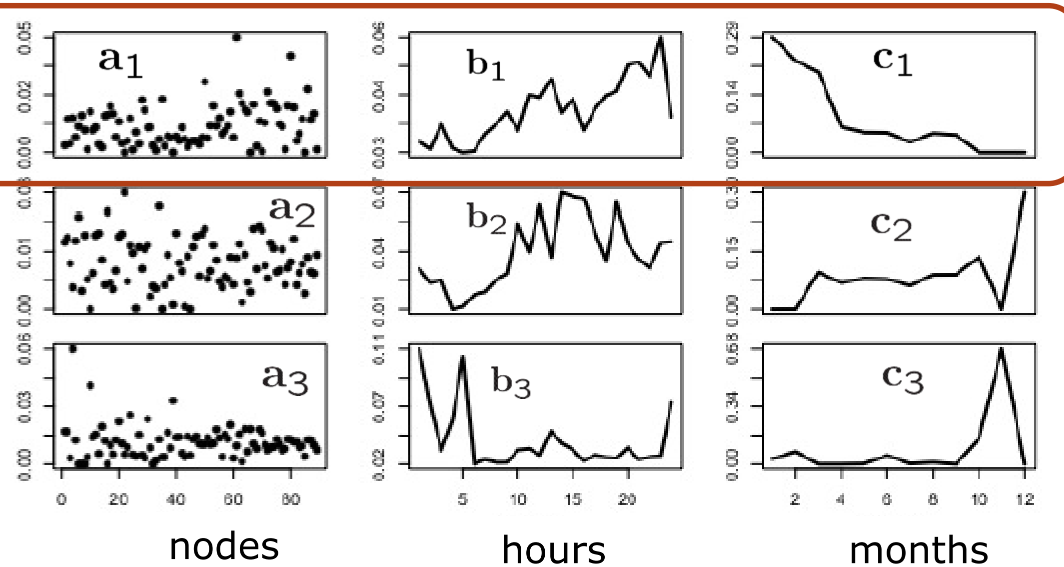
Mobile broadband networks are considered as **critical infrastructures**. There is a need for a better understanding of **their performance and reliability**. Network performance can degrade due to various reasons such as planned maintenance, equipment failure, congestion, poor coverage or failed handovers.



Being stuck in 3G; Active-mode handover not implemented until June

Failure at a central location in December, repeated over days around 3pm

Routine maintenance at night primarily in November

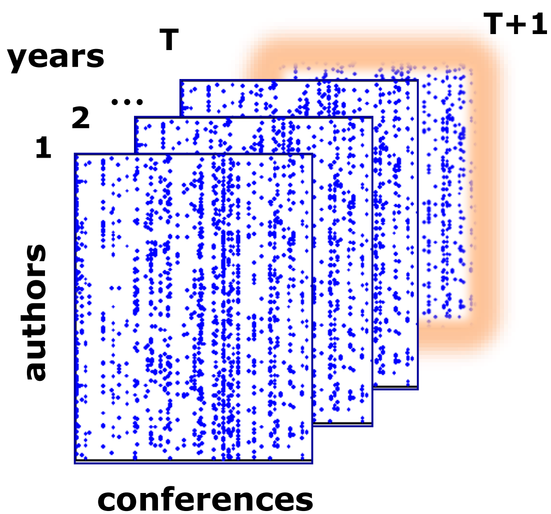


Recommender Systems: CP can capture temporal patterns useful for link prediction

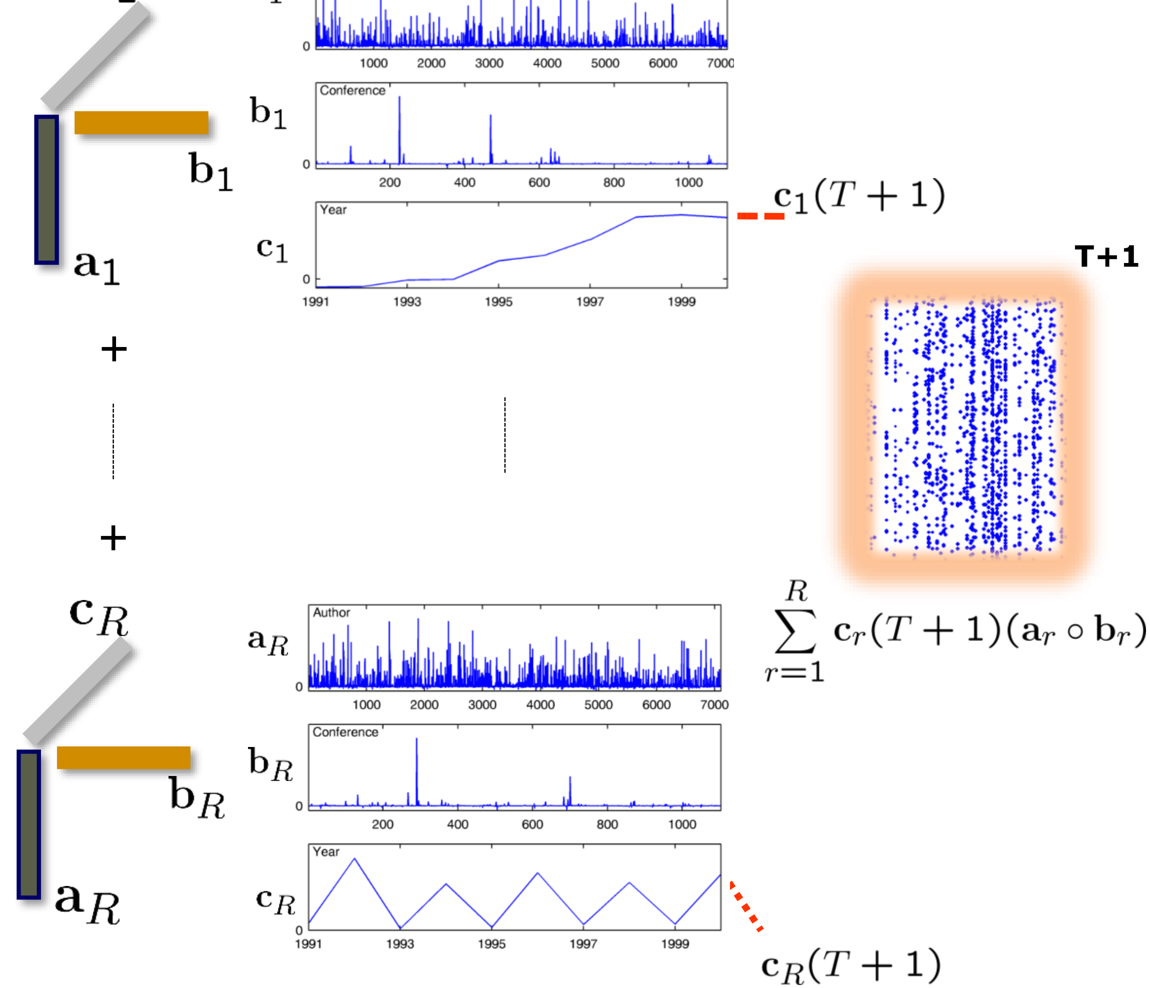
[Dunlavy et al., *ACM TKDD,* 2011]

Temporal Link Prediction

Which customers will buy which products?
Which webpages will users visit?



$\approx$

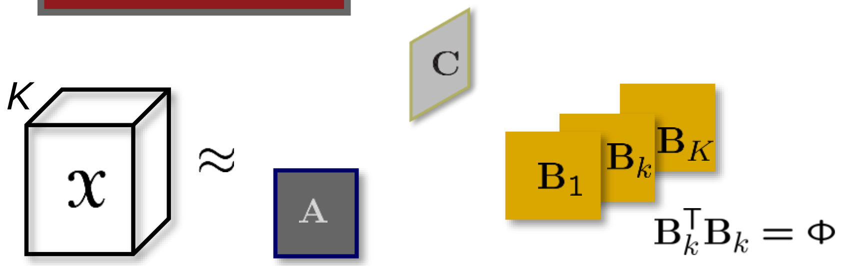


In addition to CP, there are other tensor factorization models

Mainly for applications with uniqueness & interpretability concerns

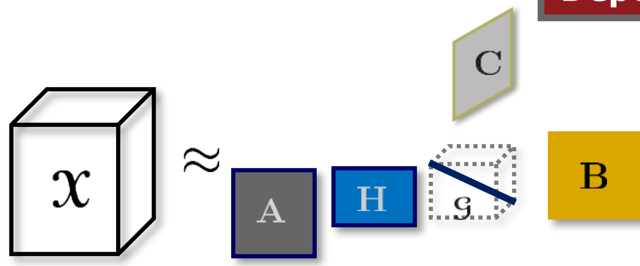
PARAFAC2

[Harshman, 1972]



PARALIND (Parallel profiles with Linear Dependences)

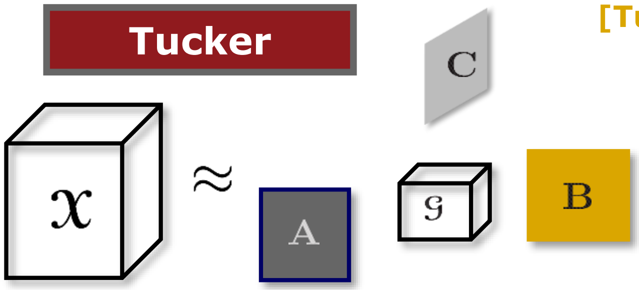
[Bro et al., 2009]



Mainly for compression & reconstruction applications

Tucker

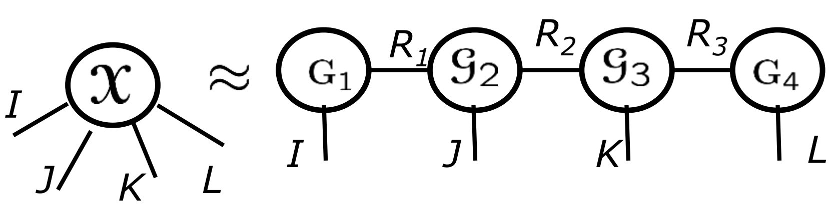
[Tucker, 1966]



$$\mathcal{X} \approx [\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Tensor Networks (TN) , Tensor Trains (TT)

[Oseledets, 2011]

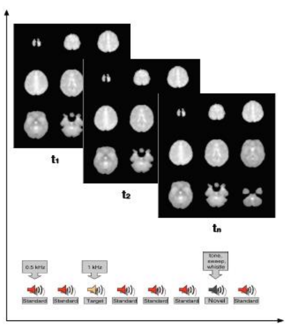


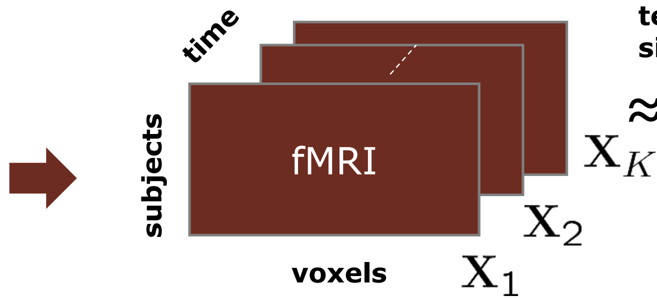
With applications in tensorizing neural networks [Novikov et al., 2015; Yang et al., 2017]

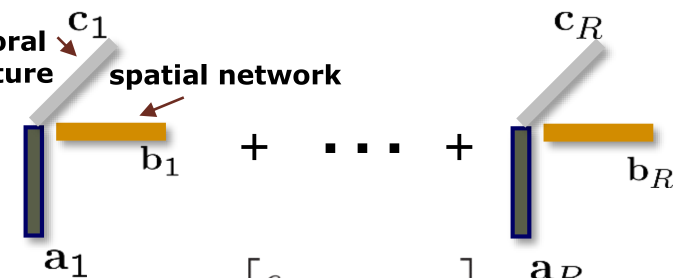
Time-evolving data analysis: Can PARAFAC2 be used to trace evolving networks?

Higher-order tensor representation is a natural way to **represent temporal data.** While the CP model is limited in terms of understanding how networks evolve in time, PARAFAC2 model can reveal both the underlying **networks** and their **evolution.**

CANDECOMP/PARAFAC (CP)





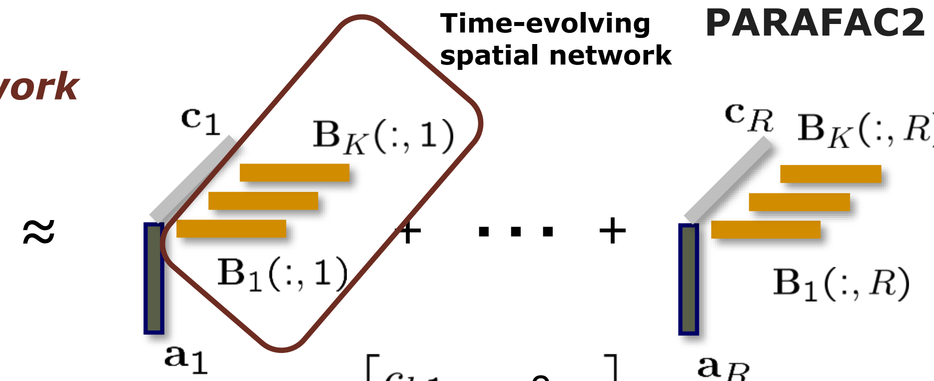


$$X_k \approx A \begin{bmatrix} c_{k1} & & 0 \\ & \ddots & \\ 0 & & c_{kR} \end{bmatrix} B^{\top}$$

Static network

PARAFAC2

Evolving network

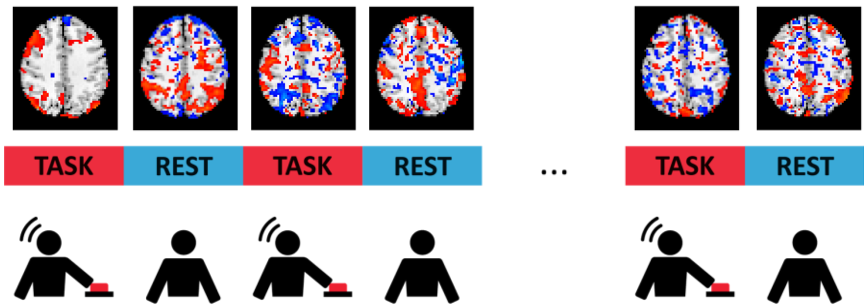


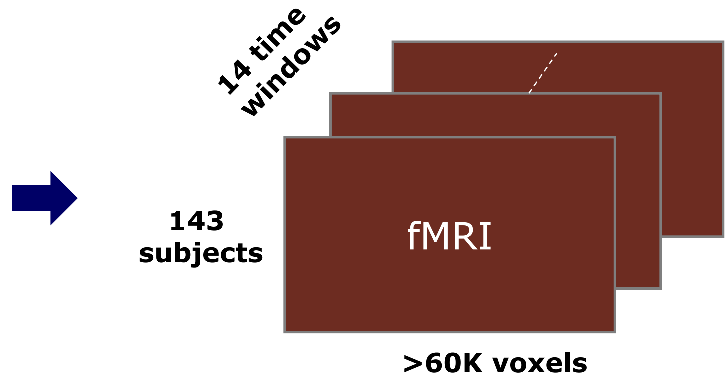
$$B_k^{\top} B_k = \Phi, \text{for } k = 1, 2, ..K$$

$$X_k \approx A \begin{bmatrix} c_{k1} & & 0 \\ & \ddots & \\ 0 & & c_{kR} \end{bmatrix} B_k^{\top}$$

PARAFAC2 has a promising performance in terms of capturing task-related functional networks and their evolution [Roald et al., *ICASSP*, 2020]

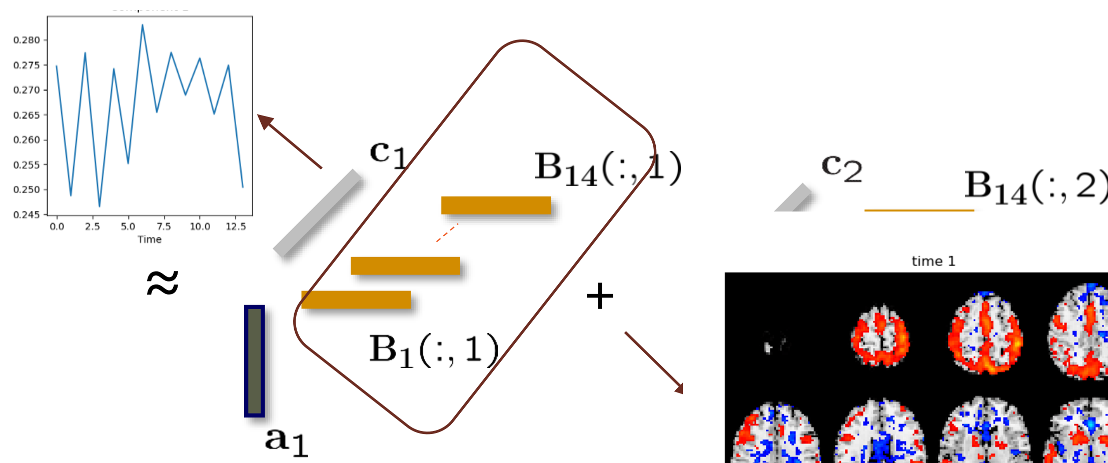
Connectivity in the brain, also termed functional connectivity, is often simplified using **static functional networks**. There is a lack of methods that can reveal the **functional networks** as well as **their temporal evolution** [Yuan et al., 2018].

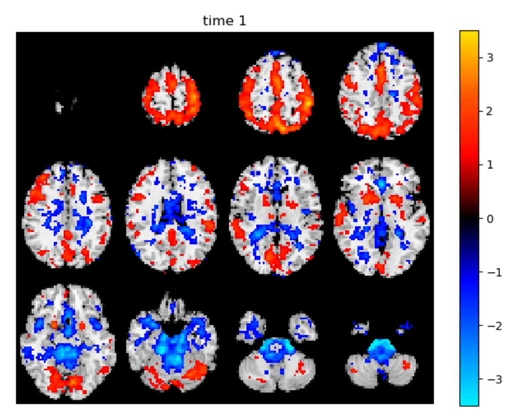




90 Healthy controls and

53 Patients with schizophrenia during a sensory motor task





This component shows statistical significance in terms of differentiating between healthy controls and patients

There are probabilistic tensor factorization approaches as well

Many studies focusing on the CP model:

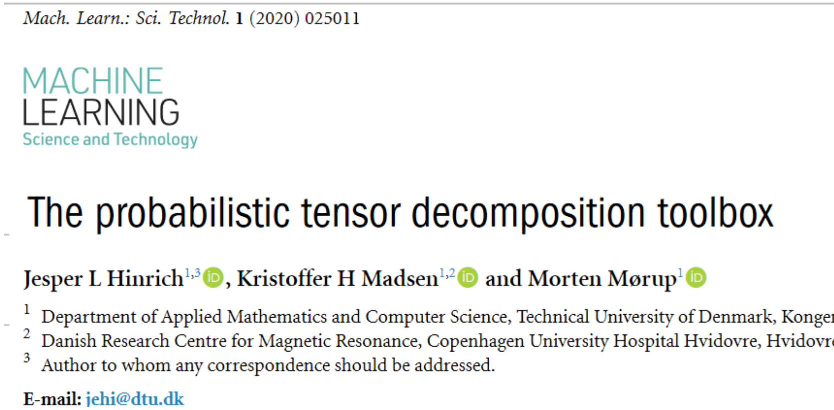
Bayesian approaches using variational inference or Gibbs sampling **[Nielsen, 2004; Shan et al., 2012; Hoff, 2011; Zhao et al., 2015]**

Also other tensor methods:

General probabilistic framework for tensor factorizations (**CP, Tucker,..**) **[Yilmaz and Cemgil, 2010]**

Tucker: relying on maximum-a-posteriori estimation **[Mørup and Hansen, 2009]**; variational inference **[Xu et al., 2012]**; Gibbs sampling **[Hoff, 2016]**

PARAFAC2, Tensor Train...

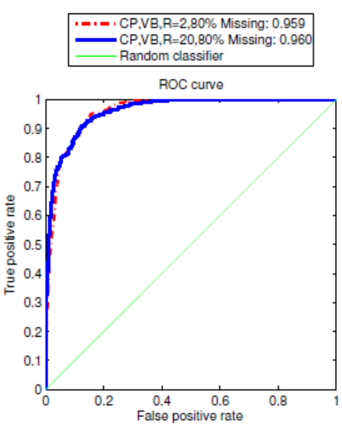

Mach. Learn.: Sci. Technol. 1 (2020) 025011

MACHINE LEARNING
Science and Technology

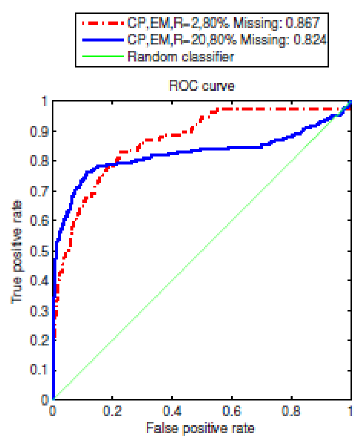
The probabilistic tensor decomposition toolbox

Jesper L Hinrich[1,3], Kristoffer H Madsen[1,2] and Morten Mørup[1]
1 Department of Applied Mathematics and Computer Science, Technical University of Denmark, Konge
2 Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Hvidovr
3 Author to whom any correspondence should be addressed.
E-mail: jehi@dtu.dk

Extensions to different priors on the factors, likelihoods other than Gaussian also exist. For a short review, see **[Hinrich et al., 2020]**

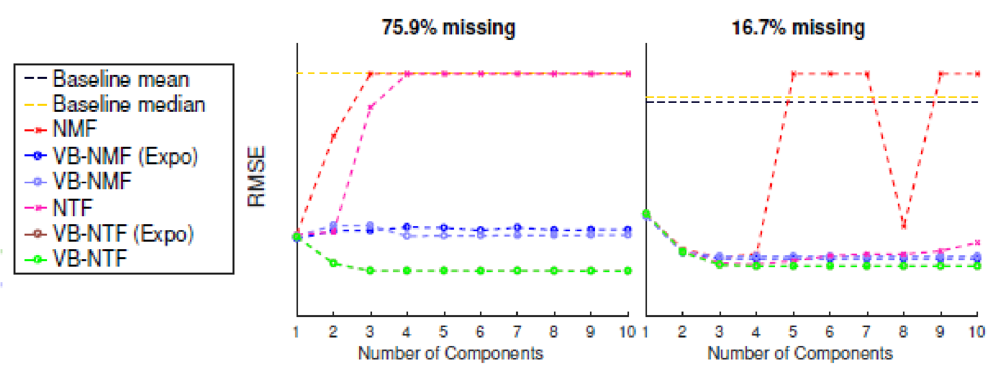


(c) PLTF-VB, 80%missing



(c) PLTF-EM, 80%missing

[Ermis et al., 2014]



[Hinrich et al., 2018]

With Bayesian approaches, we observe better missing data estimation performance with misspecified number of components and/or high amount of missing data

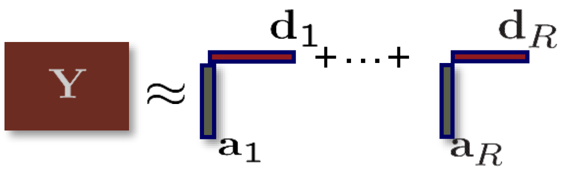
Multimodal Data Mining using Coupled Matrix/Tensor Factorizations

Coupled Matrix and Tensor Factorizations (CMTF)

[Banerjee et al., *SDM*, 2007; Acar et al., *KDD Workshop MLG*, 2011]

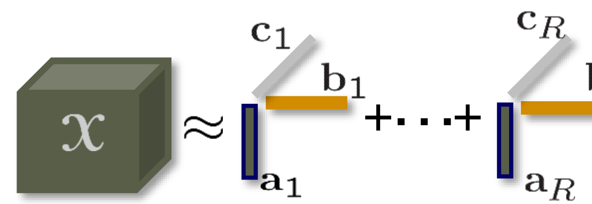
Joint analysis of heterogeneous data, i.e., data sets in the form of matrices and higher-order tensors, from multiple sources can be formulated as a coupled matrix and tensor factorization problem.

Matrix Factorization:

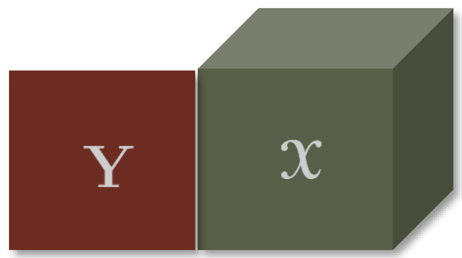


$$\mathbf{Y} \approx \sum_{r=1}^{R} \mathbf{a}_r \mathbf{d}_r^{\mathsf{T}} \approx \mathbf{A}\mathbf{D}^{\mathsf{T}}$$

Tensor Factorization: CANDECOMP/PARAFAC(CP)



$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

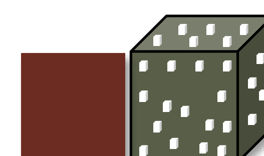


$$\mathbf{Y} \approx \boxed{\mathbf{A}}\mathbf{D}^{\mathsf{T}} \quad \mathcal{X} \approx [\![\boxed{\mathbf{A}}, \mathbf{B}, \mathbf{C}]\!]$$

The problem can be formulated as:

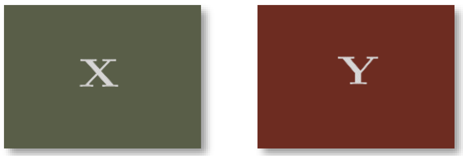
$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \| \mathcal{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \|^2 + \| \mathbf{Y} - \mathbf{A}\mathbf{D}^{\mathsf{T}} \|^2$$

Many applications, particularly in recommender systems **[Zheng et al., 2010; Ermis et al., 2015; Bhargava et al., 2015]**



Challenges: (i) Data sets in the form of matrices and higher-order arrays

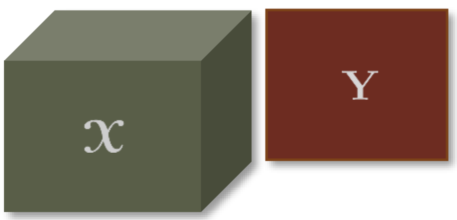
Data Fusion based on Coupled Factorizations



$$\min_{\mathbf{U},\mathbf{V},\mathbf{W}} \| \mathbf{X} - \mathbf{U}\mathbf{V}^\mathsf{T} \|^2 + \| \mathbf{Y} - \mathbf{U}\mathbf{W}^\mathsf{T} \|^2$$

psychometrics, chemometrics, bioinformatics, signal processing, data mining, …

Cannot handle joint analysis of matrices and higher-order tensors!



$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2 + \| \mathbf{Y} - \mathbf{A}\mathbf{D}^\mathsf{T} \|^2$$

Psychometrics: Linked-mode PARAFAC **[Harshman and Lundy, 1984]**

Chemometrics: Multi-way Multi-block component models **[Smilde et al., 2000]**

Bioinformatics: Coupled analysis of in vitro and histology tissue samples **[Acar et al., 2012]**

Signal Processing: Joint analysis of a covariance matrix and a cumulant tensor **[De Lathauwer and Vandewalle, 2004; Comon, 2004]**; Generalized Coupled Tensor Factorizations **[Yilmaz et al., 2011]**; Structured Data Fusion **[Sorber et al., 2015]**; Multimodal fusion of EEG and fMRI data from epileptic patients **[Hunyadi et al., 2016]**; Joint analysis of EEG and Development Scores **[Kinney-Lang et al., 2019]**

Data Mining: Multi-way clustering **[Banerjee et al., 2007]**; Community detection **[Lin et al., 2009]**; Missing value estimation **[Zheng et al., 2010]**; Missing link prediction **[Ermis et al., 2012 & 2015]**; Temporal link prediction **[Araujo et al., 2017]**; Modeling co-evolution **[Yu et al., 2018]**; Phenotyping using electronic health records **[Afshar et al., 2020]**; Scalable approaches (sampling–based **[Papalexakis et al., 2014]**, distributed **[Beutel et al., 2014; Jeon et al., 2016; Gudibanda et al., 2018]**)

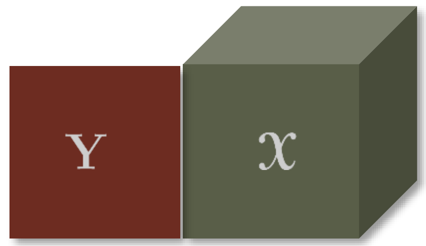
Assumption of all shared components!

ACMTF (Advanced CMTF): Structure-revealing CMTF

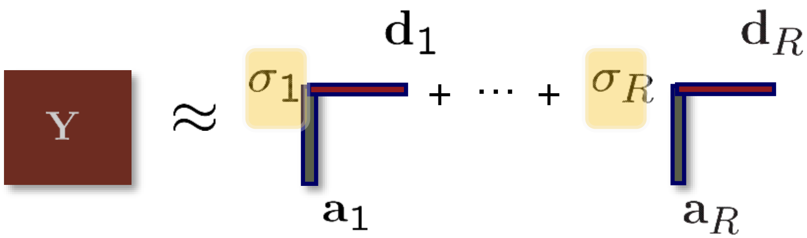
[Acar et al., *BMC Bioinformatics*, 2014; Rivet et al., *IEEE EMBC*, 2015]

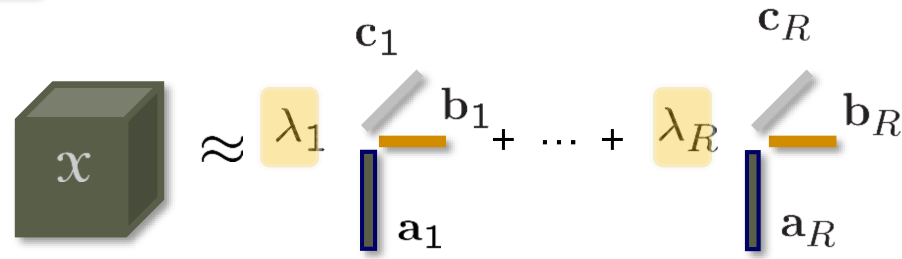
We reformulate the coupled matrix and tensor factorization problem by having factor matrices with unit norm columns and explicitly representing the weights of rank-one components in the formulation. Through modeling constraints/penalties, we let the model identify shared and unshared components.

$\mathbf{Y} \approx \mathbf{A}\boldsymbol{\Sigma}\mathbf{D}^{\mathsf{T}}$



$\mathcal{X} \approx [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$





l1-norm
$\Vert \mathbf{x} \Vert_1 = \sum \vert x_i \vert$

Structure-revealing CMTF model:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\boldsymbol{\Sigma},\lambda} \Vert \mathcal{X} - [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \Vert^2 + \Vert \mathbf{Y} - \mathbf{A}\boldsymbol{\Sigma}\mathbf{D}^{\mathsf{T}} \Vert^2 + \beta \Vert \lambda \Vert_1 + \beta \Vert \sigma \Vert_1$$

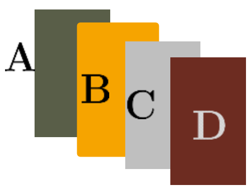
$$\text{s.t. } \Vert \mathbf{a}_r \Vert_2 = \Vert \mathbf{b}_r \Vert_2 = \Vert \mathbf{c}_r \Vert_2 = \Vert \mathbf{d}_r \Vert_2 = 1, \text{ for } r = 1, \ldots, R.$$

Original CMTF
$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \Vert \mathcal{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \Vert^2 + \Vert \mathbf{Y} - \mathbf{A}\mathbf{D}^{\mathsf{T}} \Vert^2$

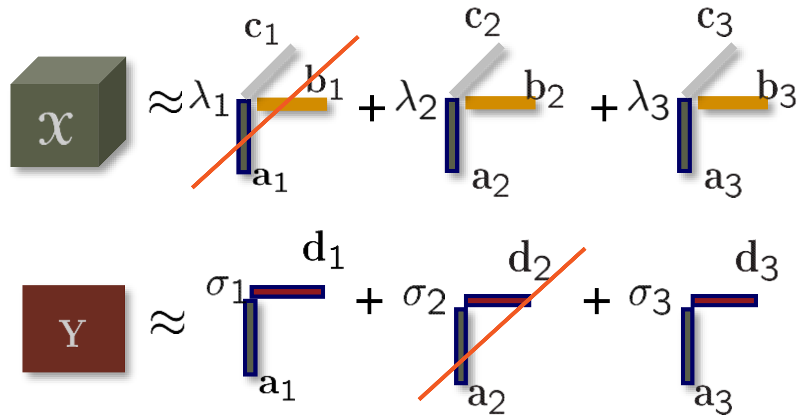
simula

Structure-revealing CMTF reveals the true structure

Generate factor matrices



Construct coupled data sets



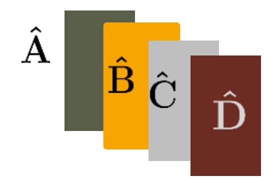
One shared and one unshared component in each data set:

$$[\lambda_1 \ \lambda_2 \ \lambda_3] = [0 \ 1 \ 1]$$
$$[\sigma_1 \ \sigma_2 \ \sigma_3] = [1 \ 0 \ 1]$$

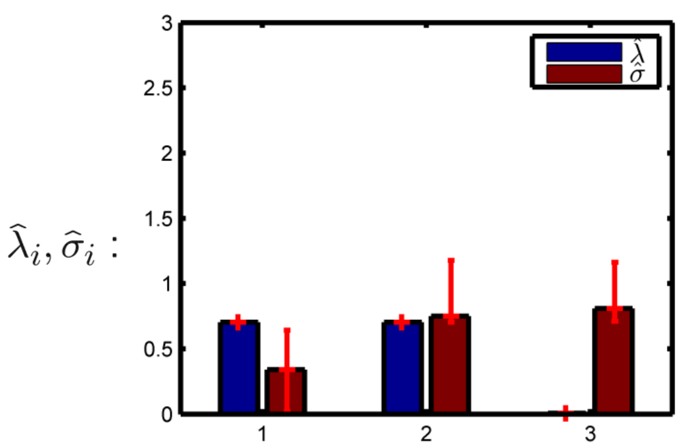
Minimize

CMTF: $f(\hat{A}, \hat{B}, \hat{C}, \hat{D})$

ACMTF: $f(\hat{\lambda}, \hat{\Sigma}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$

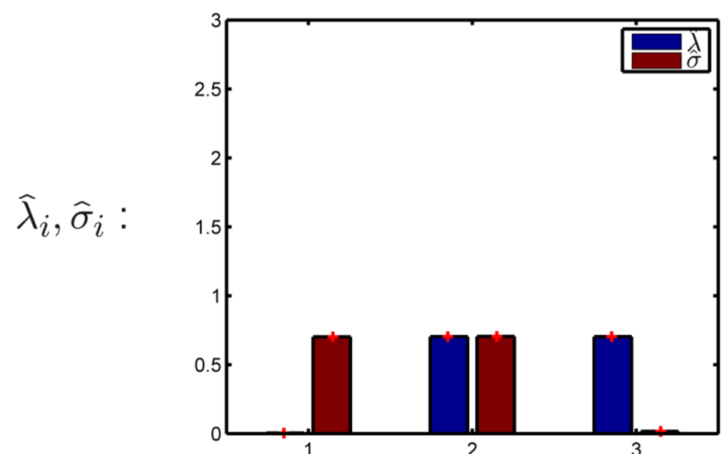


CMTF



Fails to identify shared and unshared components!

ACMTF

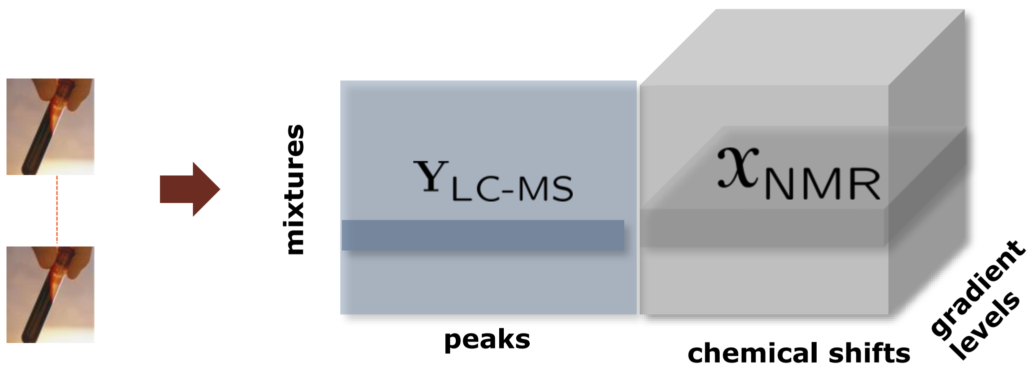


True structure captured!

simula

Chemometrics application: Exploiting the low-rank structure of tensors reveals the true structures in the data!

[Acar et al., *BMC Bioinformatics*, 2014]

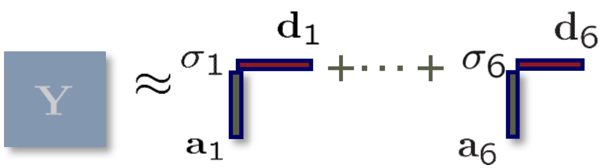


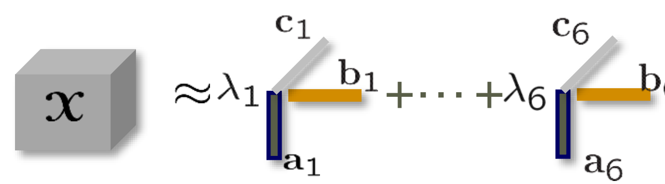
Mixtures prepared using five chemicals:

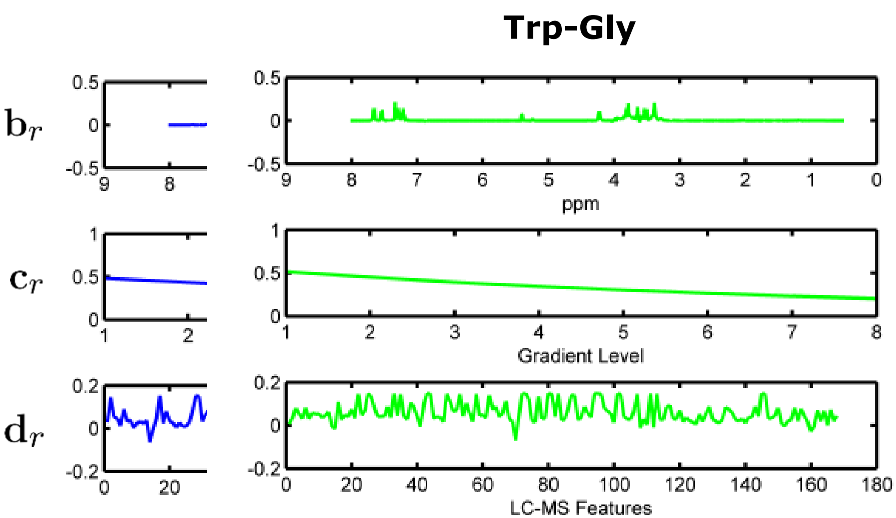
- Val-Try-Val
- Trp – Gly
- Phe
- Maltoheptaose
- Propanol

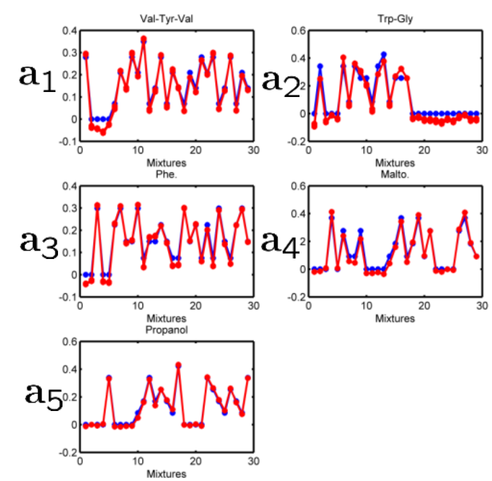
$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\Sigma,\lambda} \| \mathcal{X} - [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \|^2 + \left\| \mathbf{Y} - \mathbf{A}\Sigma\mathbf{D}^{\mathsf{T}} \right\|^2 + \beta \| \lambda \|_1 + \beta \| \sigma \|_1$$

$$\texttt{s.t.} ||\mathbf{a}_r||_2 = ||\mathbf{b}_r||_2 = ||\mathbf{c}_r||_2 = ||\mathbf{d}_r||_2 = 1, \text{for } r = 1, \ldots, R$$



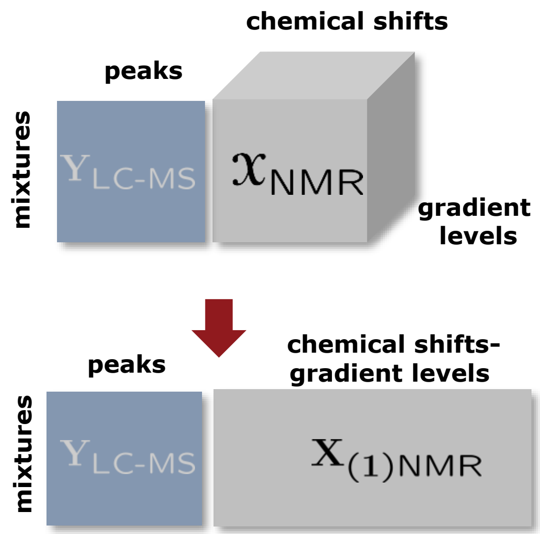




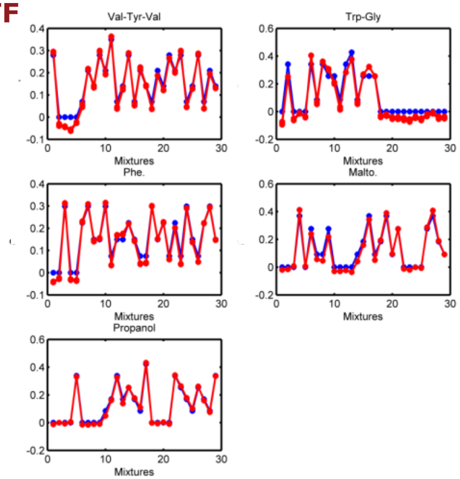


Matrix factorization-based data fusion models fail!

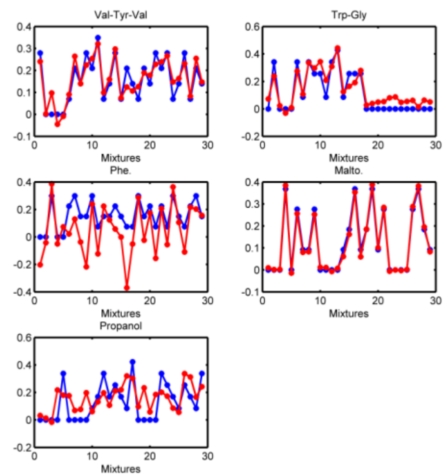
[Acar, Bro, Smilde, *Proceedings of the IEEE*, 2015]



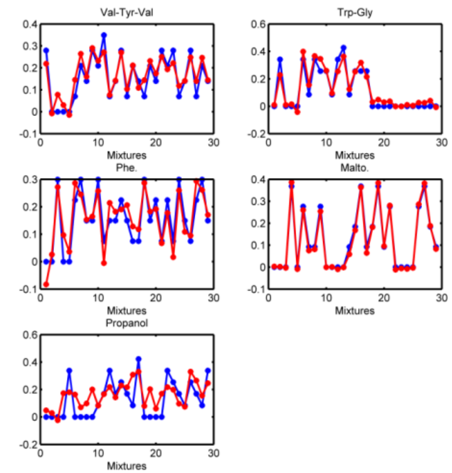
ACMTF



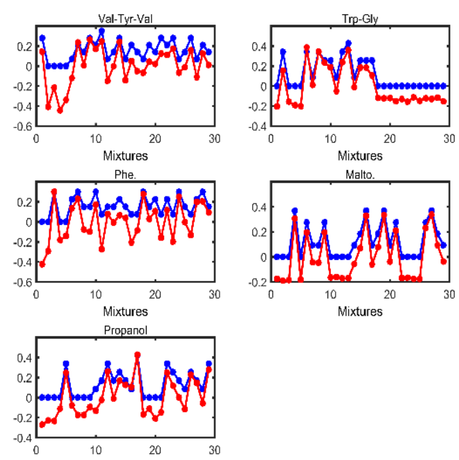
GSVD (Generalized Singular Value Decomposition) [Van Loan,1976; Paige & Saunders, 1981]



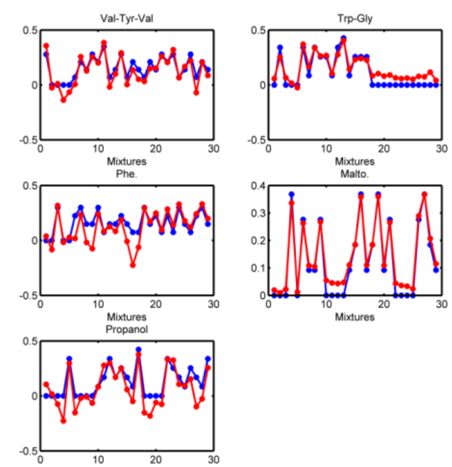
Adapted GSVD [Van Deun et al., 2012]



JIVE (Joint and Individual Variation Explained) [Lock et al., 2013]



BIBFA (Bayesian Interbattery Factor Analysis) [Klami et al., 2013]

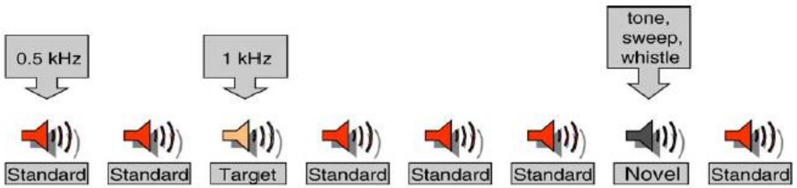


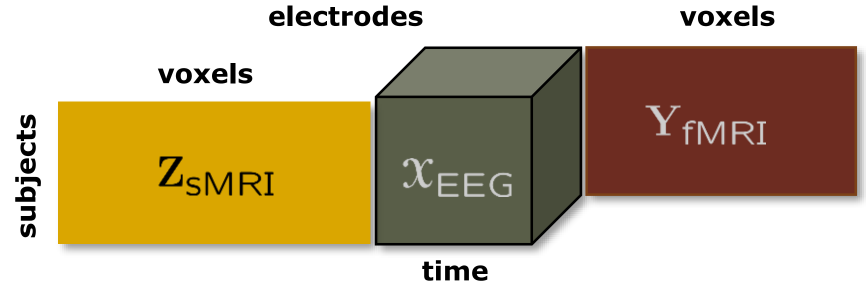
More studies on structure-revealing CMTF models, i.e., Khan et al., 2016; Farias et al., 2016.

Neuroscience application: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery

Healthy controls and **Patients** with schizophrenia from an auditory oddball task (AOD)

[Acar et al., *Frontiers in Neuroscience*, 2019]



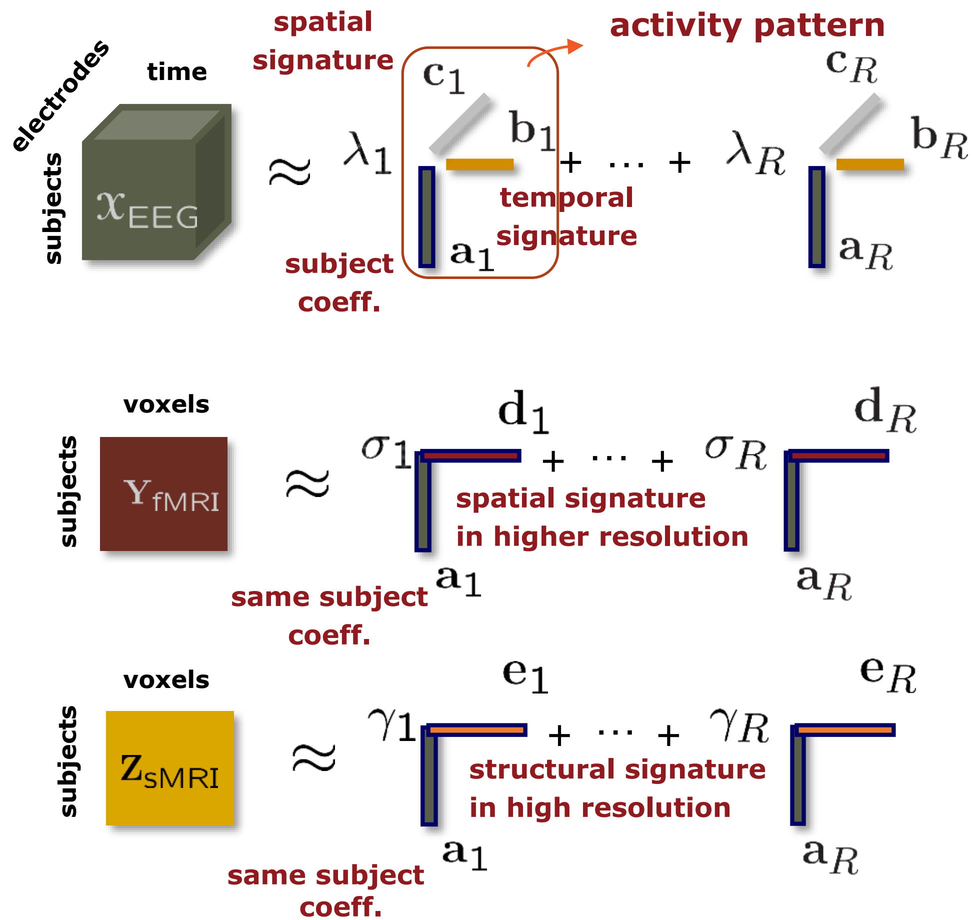


Model: Structure-revealing CMTF

$$\mathcal{X} \approx [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

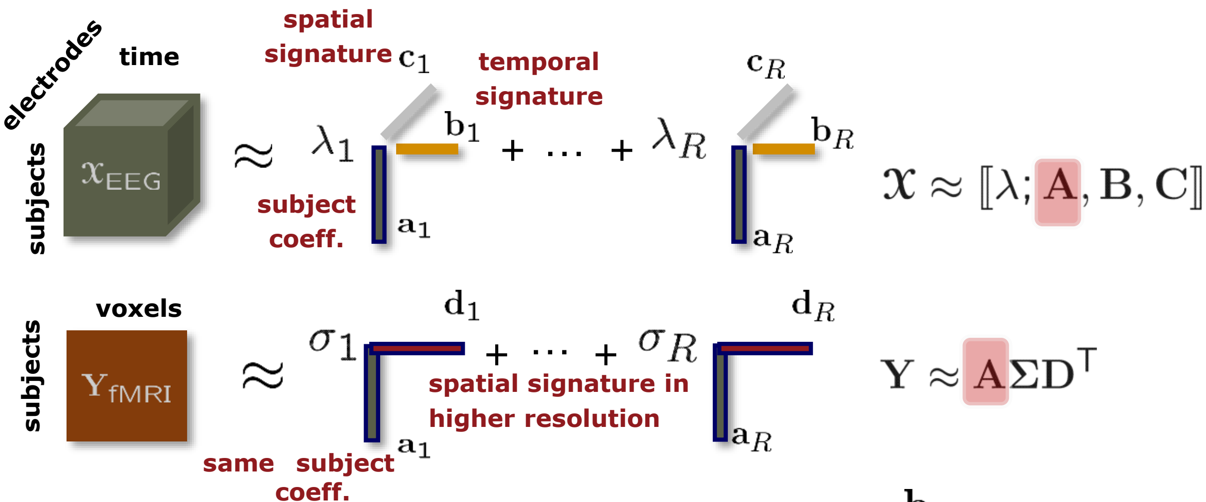
$$\mathbf{Y} \approx \mathbf{A}\Sigma\mathbf{D}^{\mathsf{T}}$$

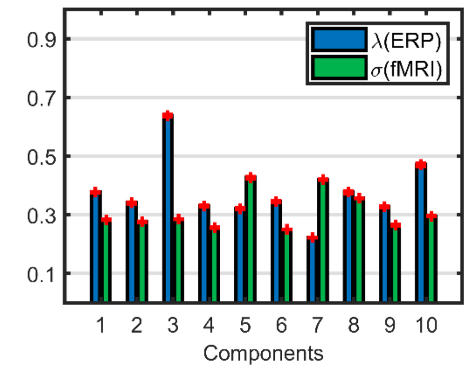
$$\mathbf{Z} \approx \mathbf{A}\Gamma\mathbf{E}^{\mathsf{T}}$$



Modeling Assumption: Subject coefficients are the same in all modalities!

EEG - fMRI: Structure-revealing CMTF captures meaningful temporal & spatial patterns with high resolution

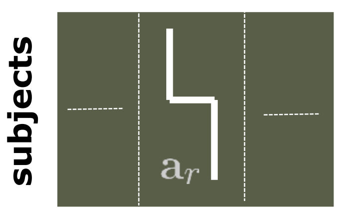


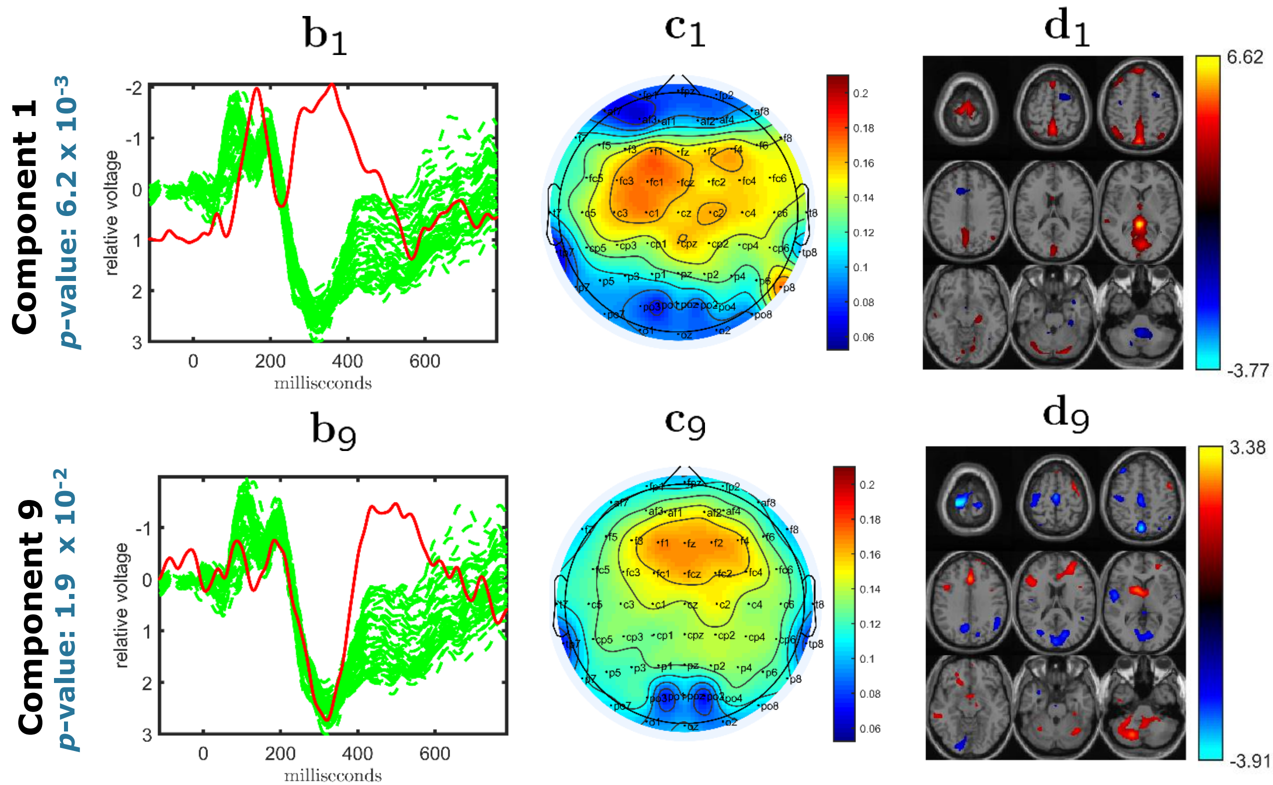


Significant components identified using a two-sample *t*-test on the columns of A

21 healthy

11 patients with schizophrenia





Structure-revealing CMTF improves the clustering performance!

$$\mathcal{X} \approx [\![\lambda; \mathbf{A}_{cp}, \mathbf{B}, \mathbf{C}]\!]$$

$$\mathcal{X} \approx [\![\lambda; \mathbf{A}_{acmtf}, \mathbf{B}, \mathbf{C}]\!] \qquad [\mathbf{X}_{(1)}\ \mathbf{Y}] \approx \mathbf{A}_{jica}\mathbf{S}$$

$$\mathbf{Y} \approx \mathbf{A}_{acmtf}\mathbf{\Sigma}\mathbf{D}^{\top}$$

$$\mathbf{Z} \approx \mathbf{A}_{acmtf}\mathbf{\Gamma}\mathbf{E}^{\top}$$

Added value of fMRI:
Improved clustering performance
Higher spatial resolution

Added value of sMRI:
No improvement in clustering
Structure information

				Clustering Performance	
	R	Centering	Number of Electrodes	Accuracy (%)	F_1-score
EEG (CP)	3	No	11	78	0.76
	3	No	62	81	0.79
EEG - fMRI (ACMTF)	10	No	11	91	0.87
	10	No	62	88	0.82
	11	No	62	88	0.80
	12	No	62	91	0.86
	9	Yes	62	88	0.78
	10	Yes	62	91	0.88
	11	Yes	62	91	0.87
	12	Yes	62	88	0.78
EEG - fMRI (jICA)	10	No	62	84	0.74
	15	No	62	82	0.70
	20	No	62	91	0.84
EEG - fMRI -sMRI (ACMTF)	10	No	62	84	0.71
	10	Yes	62	91	0.87
	15	Yes	62	91	0.86

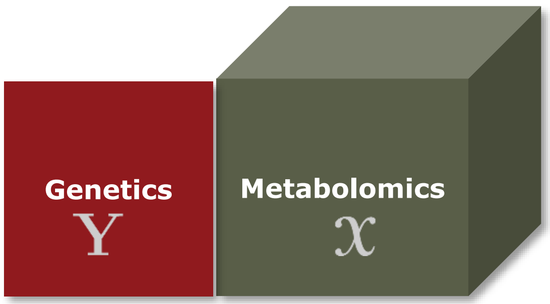
Leaving out one subject at a time:
Captured patterns are robust!

FMS				Clustering Performance	
Component A		Component B		Accuracy (%)	F_1-score
EEG	fMRI	EEG	fMRI		
0.98 (0.05)	0.95 (0.09)	0.92 (0.16)	0.90 (0.16)	87.7 (2.4)	0.81 (0.04)

Factor Match Score (FMS)

$$\mathrm{FMS}_k(\hat{\mathcal{X}}^{(1)}, \hat{\mathcal{X}}^{(2)}) = \frac{|\mathbf{a}_k^{(1)\top}\mathbf{a}_k^{(2)}|}{\left\|\mathbf{a}_k^{(1)}\right\|\left\|\mathbf{a}_k^{(2)}\right\|} \times \frac{|\mathbf{b}_k^{(1)\top}\mathbf{b}_k^{(2)}|}{\left\|\mathbf{b}_k^{(1)}\right\|\left\|\mathbf{b}_k^{(2)}\right\|} \times \frac{|\mathbf{c}_k^{(1)\top}\mathbf{c}_k^{(2)}|}{\left\|\mathbf{c}_k^{(1)}\right\|\left\|\mathbf{c}_k^{(2)}\right\|}$$

There is a need for a flexible algorithmic framework for data fusion



$$\mathrm{Y} \approx \mathrm{A}\mathrm{D}^{\top} \quad \mathcal{X} \approx [\![\mathrm{A}, \mathrm{B}, \mathrm{C}]\!]$$

Data with different distributions (e.g., count data, binary data, real entries)➔ different loss functions

Various constraints on the factors

Various types of couplings between data sets

$$\min_{\mathrm{A,B,C,D}} \| \mathcal{X} - [\![\mathrm{A}, \mathrm{B}, \mathrm{C}]\!] \|^2 + \| \mathrm{Y} - \mathrm{A}\mathrm{D}^{\top} \|^2 \Rightarrow \min_{\mathrm{A_1,A_2,B,C,D}} L_x(\mathcal{X}, [\![\mathrm{A_1}, \mathrm{B}, \mathrm{C}]\!]) + L_y(\mathrm{Y}, \mathrm{A_2}\mathrm{D}^{\top}) + g(\mathrm{B})$$
$$\text{s.t. } \mathbf{H}\mathbf{A}_1 = \mathbf{A}_2$$

State-of–the art in terms of CMTF algorithms

Alternating least squares (ALS)-based approaches **[Wilderjans et al, 2009; Bahargam and Papalexakis, 2019]**

Limited to Frobenius norm

--- with linear coupling **[Farias et al., 2016; Kanatsoulis et al., 2018]**

Limited in terms of constraints

All-at-once optimization

- Unconstrained using gradient-based approaches **[Acar et al., 2011]**
- Nonlinear least squares **[Sorber et al., 2015; Vervliet et al., 2016]**

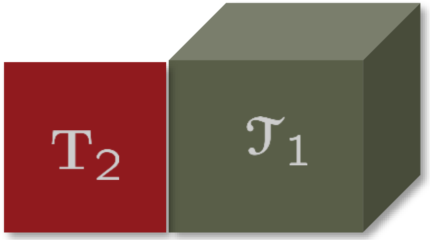
  Limited to Frobenius norm

  --- with linear coupling and various constraints
- Constrained optimization using a general–purpose optimization solver **[Acar et al., 2014]**

There is a need for a flexible algorithmic framework that can handle various loss functions, incorporate different type of constraints and couplings.

Alternating Optimization (AO) –Alternating Direction Method of Multipliers (ADMM) for Regularized CMTF with Linear Couplings

[Schenker et al., *IEEE JSTSP*, 2021]



Previously, AO-ADMM has shown promising flexibility for constrained tensor factorizations [Huang et al., 2016]. We extend this framework to coupled matrix/tensor factorizations to incorporate various **constraints**, **loss functions** and **linear couplings.**

$$\min_{\{\mathbf{C}_{i,d},\boldsymbol{\Delta}_d\}_{d\le D_i, i\le N}} \mathcal{L}_1(\mathcal{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]) + \mathcal{L}_2(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}) + \sum_{i=1}^{N}\sum_{d=1}^{D_i} g_{i,d}\left(\mathbf{C}_{i,d}\right)$$
$$\text{s.t.} \quad \mathbf{H}_{i,d}\mathbf{C}_{i,d} = \boldsymbol{\Delta}_d$$

Fix all other modes, and solve for one mode using an alternating scheme (AO).

Example: Coupling only in the first mode

while convergence criterion is not met **do**

mode 1

$$\min_{\{\mathbf{C}_{i,1}\}_{i\le N},\boldsymbol{\Delta}_1} \mathcal{L}_1(\mathcal{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]) + \mathcal{L}_2(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}) + \sum_{i=1}^{N} g_{i,1}\left(\mathbf{C}_{i,1}\right)$$
$$\text{s.t.} \quad \mathbf{H}_{i,1}\mathbf{C}_{i,1} = \boldsymbol{\Delta}_1$$

mode 2

$$\min_{\{\mathbf{C}_{i,2}\}_{i\le N}} \mathcal{L}_1(\mathcal{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]) + \mathcal{L}_2(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}) + \sum_{i=1}^{N} g_{i,2}\left(\mathbf{C}_{i,2}\right)$$

end while

ADMM
Alternating Direction Method of Multipliers

$$\min_{\mathbf{x},\mathbf{z}} \quad f(\mathbf{x}) + g(\mathbf{z})$$
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}$$

Algorithm 1 Skeleton of scaled-form ADMM

while convergence criterion is not met **do**

$$x^{(k+1)} = \underset{x}{\text{argmin}}\ f(x) + \tfrac{\rho}{2}\|Ax + Bz^{(k)} - c + \mu^{(k)}\|_2^2$$
$$z^{(k+1)} = \underset{z}{\text{argmin}}\ g(z) + \tfrac{\rho}{2}\|Ax^{(k+1)} + Bz - c + \mu^{(k)}\|_2^2$$
$$\mu^{(k+1)} = \mu^{(k)} + Ax^{(k+1)} + Bz^{(k+1)} - c$$
$$k = k + 1$$

end while

ADMM subproblem for Regularized CMTF with linear couplings (mode 1)

Optimization Problem

$$\min_{\{\mathbf{C}_{i,1}\}_{i\leq N},\boldsymbol{\Delta}_1} \mathcal{L}_1(\mathfrak{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]) + \mathcal{L}_2(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}) + \sum_{i=1}^{N} g_{i,1}\left(\mathbf{C}_{i,1}\right)$$
$$\text{s.t.} \quad \mathbf{H}_{i,1}\mathbf{C}_{i,1} = \boldsymbol{\Delta}_1$$

Introduce variable $\mathbf{Z}_{i,1}$ to separate the regularization from the factorization

$$\min_{\{\mathbf{C}_{i,1},\mathbf{Z}_{i,1}\}_{i\leq N},\boldsymbol{\Delta}_1} \mathcal{L}_1(\mathfrak{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]) + \mathcal{L}_2(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}) + \sum_{i=1}^{N} g_{i,1}\left(\mathbf{Z}_{i,1}\right)$$
$$\text{s.t.} \quad \mathbf{H}_{i,1}\mathbf{C}_{i,1} = \boldsymbol{\Delta}_1$$
$$\mathbf{C}_{i,1} = \mathbf{Z}_{i,1}$$

Introduce dual variables and formulate the augmented Lagrangian

$$L(\mathbf{C}_{i,1},\mathbf{Z}_{i,1},\boldsymbol{\Delta}_1,\mu_{i,1_{(z)}},\mu_{i,1_{(\delta)}}) = \mathcal{L}_1\left(\mathfrak{T}_1, [\![\mathbf{C}_{1,1},\mathbf{C}_{1,2},\mathbf{C}_{1,3}]\!]\right) + \mathcal{L}_2\left(\mathbf{T}_2, \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}}\right)$$
$$+ \sum_{i=1}^{N}\left[g_{i,1}(\mathbf{Z}_{i,1}) + \frac{\rho}{2}\left\| \mathbf{C}_{i,1} - \mathbf{Z}_{i,1} + \mu_{i,1_{(z)}} \right\|_F^2 + \frac{\rho}{2}\left\| \mathbf{H}_{i,1}\mathbf{C}_{i,1} - \boldsymbol{\Delta}_1 + \mu_{i,1_{(\delta)}} \right\|_2^2\right]$$

Using alternating optimization, solve for $\{\mathbf{C}_{i,1}\}_{i\leq N}, \{\mathbf{Z}_{i,1}\}_{i\leq N}, \boldsymbol{\Delta}_1$ followed by dual updates.

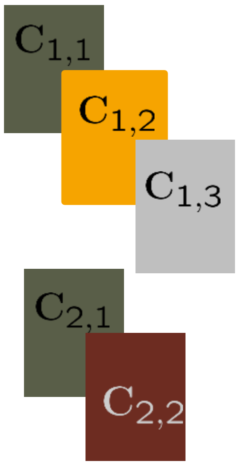
In case of Frobenius norm-based loss:
Solution of a linear least squares problem/Sylvester eqn.

Other differentiable losses:
Numerical optimization using LBFGS-B

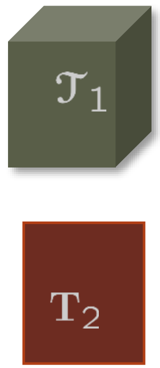
Proximal operators
(Proximity Operator Repository)
http://proximity-operator.net/proximityoperator.html

Exact coupling: AO-ADMM framework is accurate and efficient!

Generate factor matrices



Construct coupled data sets



$$= [\![\mathbf{C}_{1,1}, \mathbf{C}_{1,2}, \mathbf{C}_{1,3}]\!] + \mathcal{N}$$

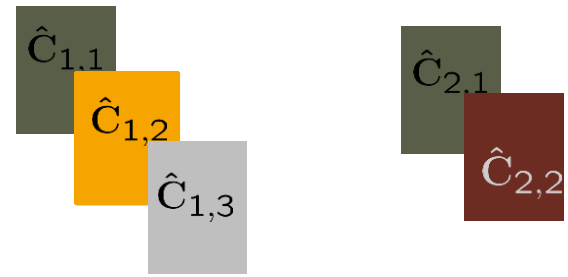
$$= \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}} + \mathbf{N}$$

Solve using AO-ADMM

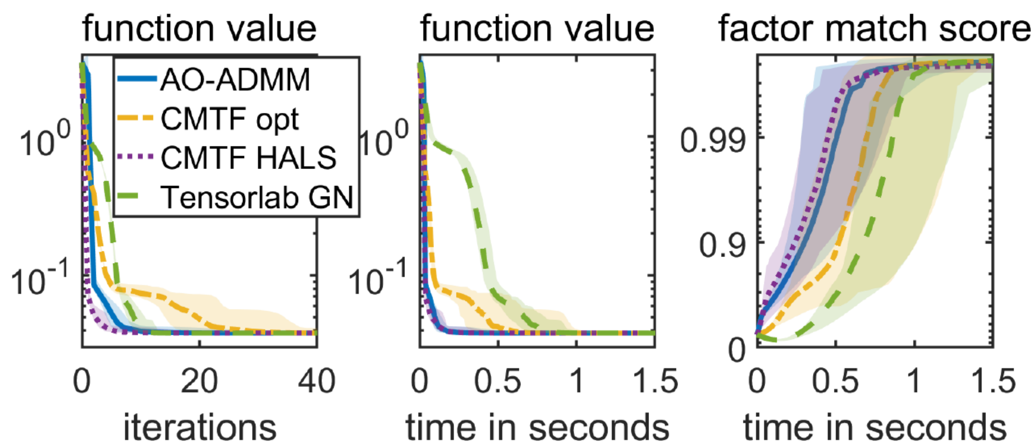
$$\min_{\{\mathbf{C}_{i,d}\}_{d \le D_i, i \le N}, \boldsymbol{\Delta}_1} \left\| \mathcal{T}_1 - [\![\mathbf{C}_{1,1}, \mathbf{C}_{1,2}, \mathbf{C}_{1,3}]\!] \right\|_F^2 + \left\| \mathbf{T}_2 - \mathbf{C}_{2,1}\mathbf{C}_{2,2}^{\mathsf{T}} \right\|_F^2$$

$$\text{s.t.} \quad \mathbf{C}_{i,1} = \boldsymbol{\Delta}_1$$

$$\mathbf{C}_{i,d} \ge 0, i \le N, d \le D_i$$



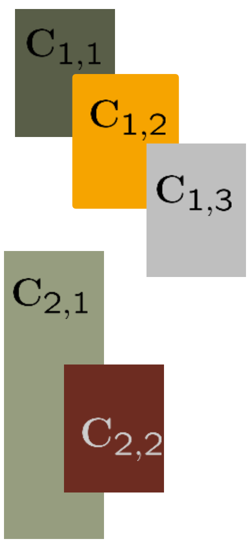
$$\mathbf{C}_{1,1} = \mathbf{C}_{2,1} = \boldsymbol{\Delta}_1$$



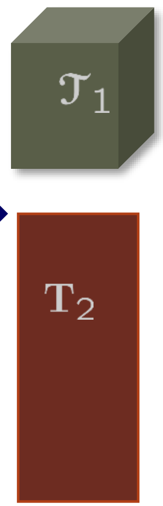
Comparable performance in terms of accuracy and computational efficiency!

Linear coupling: AO-ADMM framework is accurate and efficient!

Generate factor matrices



Construct coupled data sets



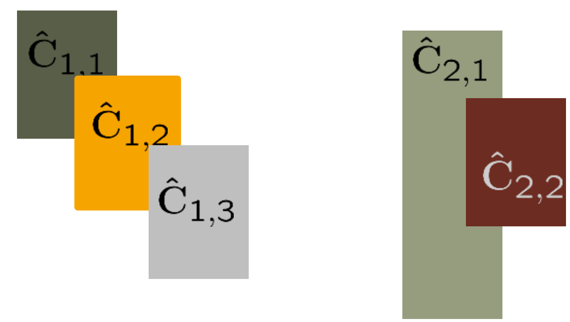
$$= [\![C_{1,1}, C_{1,2}, C_{1,3}]\!] + \mathcal{N}$$

$$= C_{2,1} C_{2,2}^T + N$$

Solving regularized CMTF with linear couplings using AO-ADMM

$$\min_{\{C_{i,d}\}_{d \le D_i, i \le N}, \Delta_1} \left\| \mathcal{T}_1 - [\![C_{1,1}, C_{1,2}, C_{1,3}]\!] \right\|_F^2 + \left\| T_2 - C_{2,1} C_{2,2}^T \right\|_F^2$$

$$\text{s.t.} \quad H_{i,1} C_{i,1} = \Delta_1$$



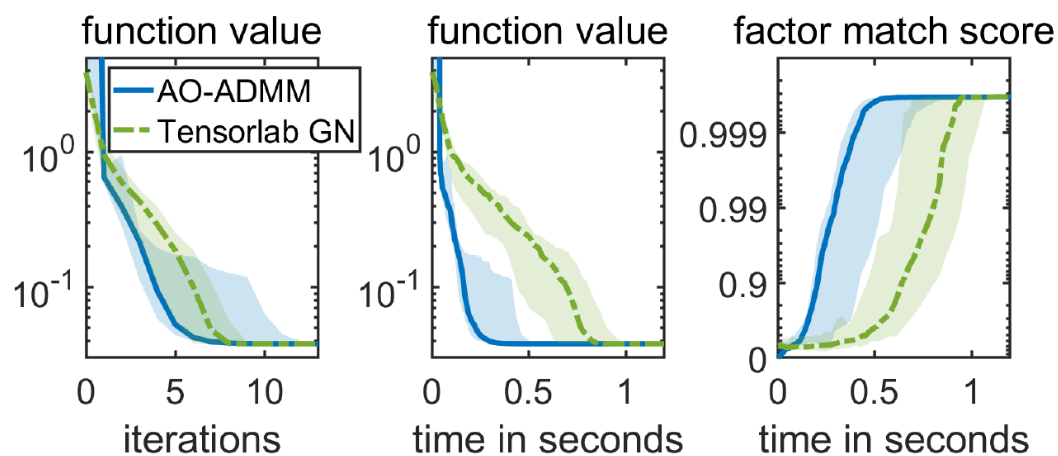
$$H_{1,1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$C_{1,1}$

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

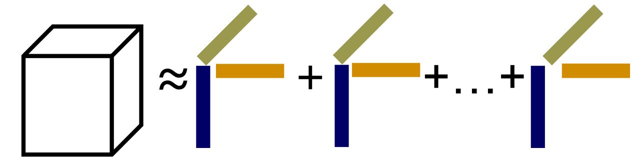
$C_{2,1}$

$$\left.\begin{matrix} H_{1,1} C_{1,1} \\ H_{2,1} C_{2,1} \end{matrix}\right\} = \Delta_1$$



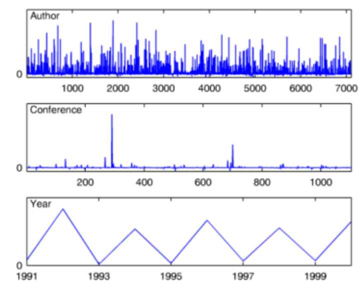
AO-ADMM is computationally competitive!

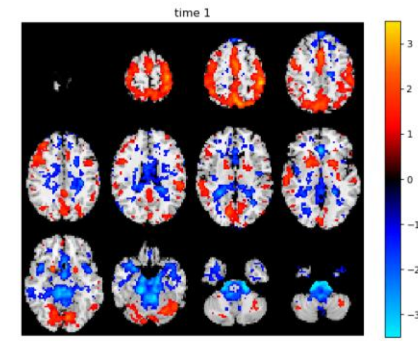
Summary



Tensor factorizations are promising tools in terms of revealing the **underlying patterns** in complex data in many applications in data mining, chemometrics, signal processing, neuroscience, e.g.,

- (Temporal) Link prediction
- Unmixing brain activities in neuroscience
- Revealing **time-evolving** spatial networks from task fMRI data
- ….

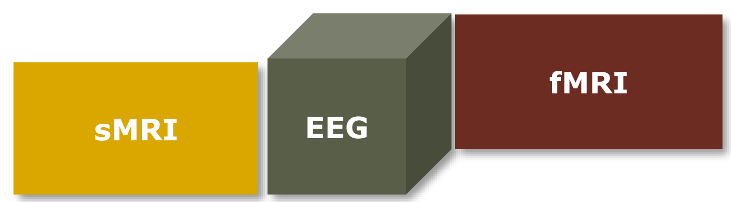




Also for **data completion, compression** …

Coupled matrix/tensor factorizations have shown promise in terms of jointly analyzing data from multiple sources in many disciplines.

- Different formulations to identify **shared/unshared** factors
- Flexible **algorithmic** approaches to impose constraints, incorporate different loss functions and types of couplings



Many challenges remain…

Our Tensor/Fusion Team @ Machine Intelligence Department

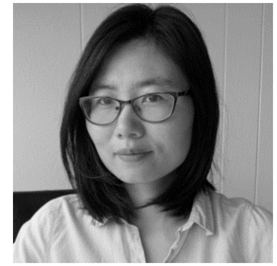
Available PhD & Postdoc Positions

Evrim Acar

Carla Schenker
PhD Student

Marie Roald
PhD Student

Age Smilde
University of Amsterdam
(Adjunct @Simula)



Lu Li
Postdoc

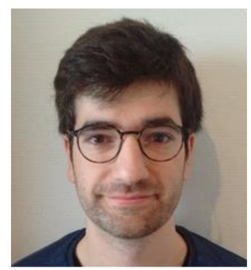
Florian Becker
PhD Student

Shi Yan
Postdoc

Collaborators

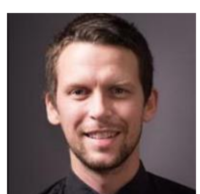
Multimodal Neuroimaging Data Analysis

Tülay Adalı, University of Maryland Baltimore County, MD



Flexible Algorithmic Approaches for Coupled Factorizations

Jeremy E. Cohen, INRIA, CNRS IRISA, France



Tensor Factorizations & Omics fusion

Rasmus Bro, University of Copenhagen, Denmark

Morten A. Rasmussen, University of Copenhagen, Denmark

simula

Surveys on Tensor Factorizations

T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications, *SIAM Review*, 51(3):455-500, 2009

E. Acar and B. Yener. Unsupervised Multiway Data Analysis: A Literature Survey, *IEEE Transactions on Knowledge and Data Engineering*, 21(1): 6-20, 2009

E. Papalexakis et al., Tensors for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms, *ACM Transactions on Intelligent Systems and Technology*, 8(2): Article 16, 2016

Coupled Matrix/Tensor Factorizations

E. Acar, R. Bro and A. K. Smilde. Data Fusion in Metabolomics based on Coupled Matrix and Tensor Factorizations. *Proceedings of the IEEE*, 103: 1602-1620, 2015

A. K. Smilde and I. Van Mechelen. A Framework for Low-Level Data Fusion. *In Data Fusion Methodology and Applications,* 31: 27-50, 2019

TrACEr

Time-Aware ConstrainEd Multimodal Data Fusion

https://tracer.simulamet.no

Evrim Acar

evrim@simula.no

novo nordisk fonden

The Research Council of Norway

simula