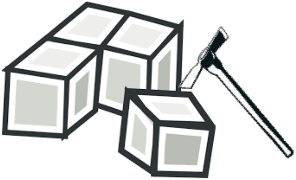
Environmental Multiway Data Mining

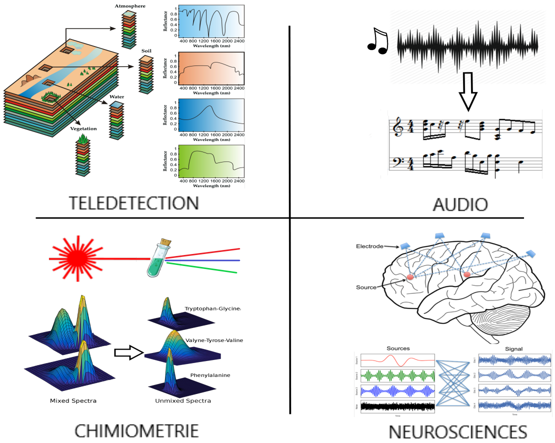


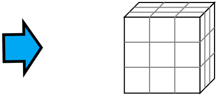
Jérémy E. Cohen, Nicolas Gillis

UMONS, FNRS

March 21, 2018

Tensors in Signal Processing





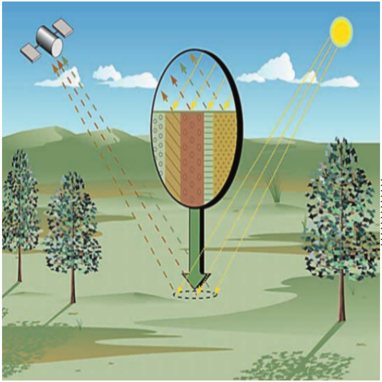
Data tensor

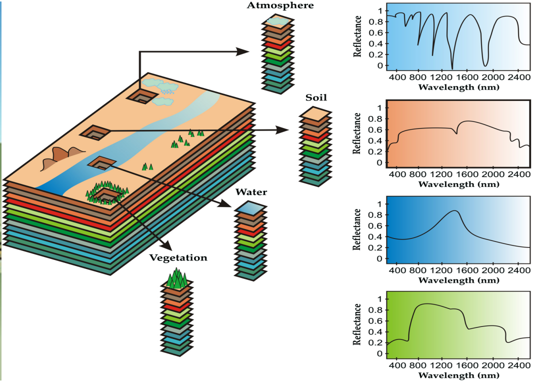
Ex : freq $\times$ time $\times$ sensor

Tensor model = Accounting for structure

Images : crédits à L. Korczowski et J.B-Dias

Hyperspectral imaging principle

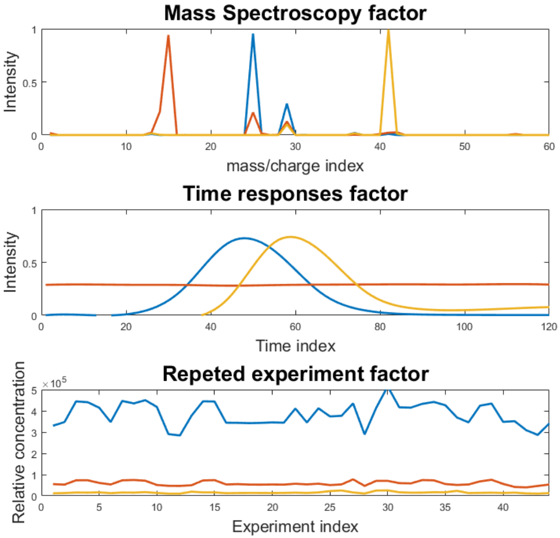




- Each image is a **linear mixture** of various spectral signatures.
- Each material has a unique spectral response.

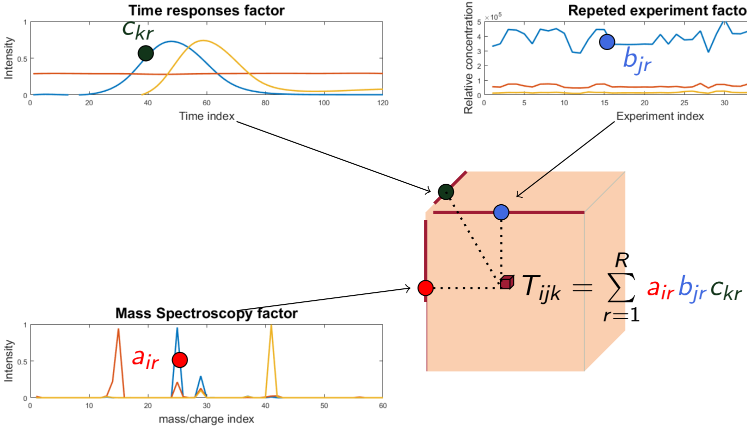
Credits for illustrations: Veganzones (left) and Bioucas (right)

Liquid Chromatography — Mass Spectroscopy component



Mass over charge intensities form a tensor

$$T(\lambda, t, k) = \sum_{r=1}^{R} a_r(\lambda) b_r(t) c_r(k)$$



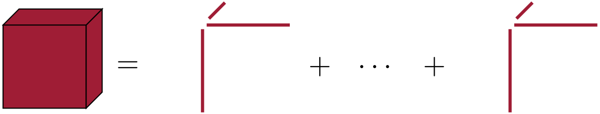
A tool for LRA: Canonical Polyadic Decomposition

Canonical Polyadic Decomposition [Hitchcock,1927] aims at extracting all R components.

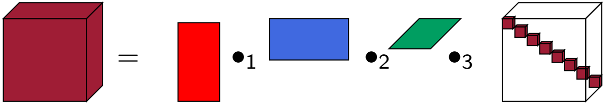
Tensor $=$ first component $+ \cdots +$ R*th* component

- Unmixing in theory does not require additional knowledge for order 3 and more.
- For matrices, not unique if $R > 1 \rightarrow$ SVD (orthogonality), NMF (non-negativity).

CPD



$$\mathcal{T} = \mathbf{a}_1 \otimes \mathbf{b}_1 \otimes \mathbf{c}_1 + \cdots + \mathbf{a}_R \otimes \mathbf{b}_R \otimes \mathbf{c}_R$$



$$\mathcal{T} = (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\ \mathcal{I}_R$$

$\mathcal{T}$ has sizes $K \times L \times M$ $\quad$ $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_R]$ has sizes $K \times R$

$\otimes$ is the tensor product $\quad$ $\bullet_i$ is the contraction on mode i

R is the rank of $\mathcal{T}$, *i.e.* smallest number of rank-one tensors spanning $\mathcal{T}$.

Tensor decomposition as an approximation problem

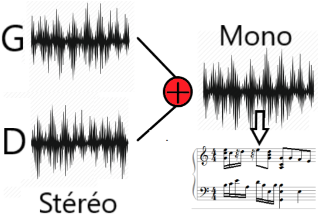
$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \quad & \|\mathcal{T} - (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\mathcal{I}_R\| \\ \text{sub. to} \quad & \mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathcal{C}_{A,B,C} \end{aligned}$$

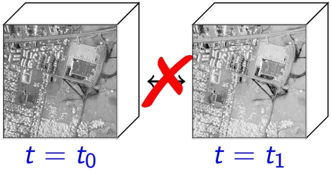
- Non-convex in the general case but convex with respect to each block $\mathbf{A}, \mathbf{B}, \mathbf{C}$.
- Example: Non-negative Matrix Factorization with Frobenius norm

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \quad & \|\mathbf{M} - \mathbf{A}\mathbf{B}^T\|_F^2 \\ \text{sub. to} \quad & \mathbf{A} \geq 0 \quad \mathbf{B} \geq 0 \end{aligned}$$

Challenges in tensor signal processing

Multidimensional structure not exploited !





Main issues:

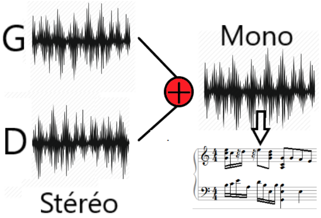
Interpretability

Tensor formalism

Constraints / Size

Challenges in tensor signal processing

Multidimensional structure not exploited !

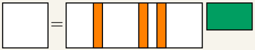


$t = t_0$ $t = t_1$

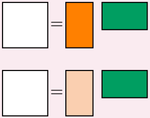
Main issues:

Interpretability	Tensor formalism	Constraints / Size

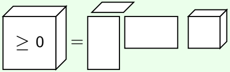
I. Informed decom.



II. Data fusion

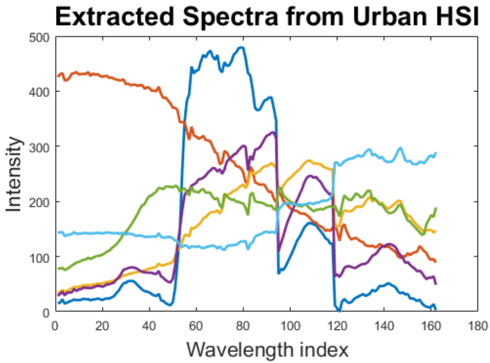


III. Optimisation

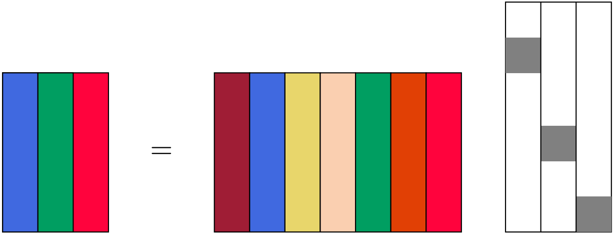


Challenge I: Identification.

Identification may be an issue



Let's choose **A** from a dictionary



$\mathbf{A} \quad = \quad \mathbf{D} \quad \mathbf{S}$

$$\mathbf{X} = \mathbf{DSB}^T = \mathbf{D}(:, \mathcal{K})\mathbf{B}^T \text{ or } \mathcal{T} = (\mathbf{DS} \otimes \mathbf{B} \otimes \mathbf{C})\mathcal{I}_R \text{ where } \|\mathbf{S}\|_{col,0} = 1$$

<u>I. Theorem:</u> If $spark(\mathbf{D}) > R$ and $\mathcal{K}$ has no repetition,

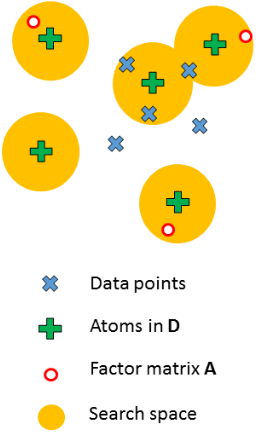
(i) if there exist $(\mathcal{K}, \mathbf{B})$ so that $\mathbf{M} = \mathbf{D}(:, \mathcal{K})\mathbf{B}$, then it is unique.

(ii) $\underset{\mathcal{K},\mathbf{B},\mathbf{C}}{\text{argmin}} \|\mathcal{T} - (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\mathcal{I}_R\|_F^2$ existe.

J.E.C. and N.Gillis, "Dictionary-based Tensor Canonical Polyadic Decomposition", IEEE Trans. on Signal Proc., 2018

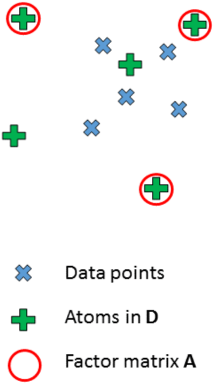
Flexibility and Separability

Flexibility



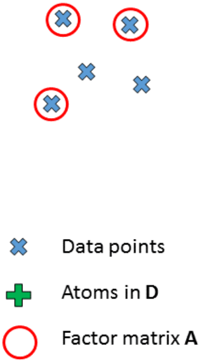
$$\mathbf{A} \approx \mathbf{DS}$$

Standard case



$$\mathbf{A} = \mathbf{DS}$$

Separability



$$\mathbf{A} = \mathbf{XS}$$

Matching Pursuit ALS (works for high-order tensors)

An alternating nonnegative least squares method where $\mathbf{A}$, $\mathcal{K}$ and $\mathbf{B}$ are estimated alternatively.

Input: Initial $\mathbf{A}$, $\mathbf{B}$, $\mathcal{K}$ (Using e.g. SPA, VCA...).
Run a few iterations of NMF.
while stopping criterion is not met,

- $\widehat{\mathbf{A}} = \underset{\mathbf{A}\geq 0}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{A}\mathbf{B}^T\|_F^2$
- $\widehat{\mathcal{K}}(i) = \underset{j}{\operatorname{argmax}}\ d_j^T \widehat{\mathbf{A}}_i \quad \forall i \in [R]$
- $\widehat{\mathbf{B}} = \underset{\mathbf{B}\geq 0}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{D}(:, \widehat{\mathcal{K}})\mathbf{B}^T\|_F^2$

Output: Selected atoms set $\mathcal{K}$ and abundances $\mathbf{B}$.

Pros and Cons

Pros

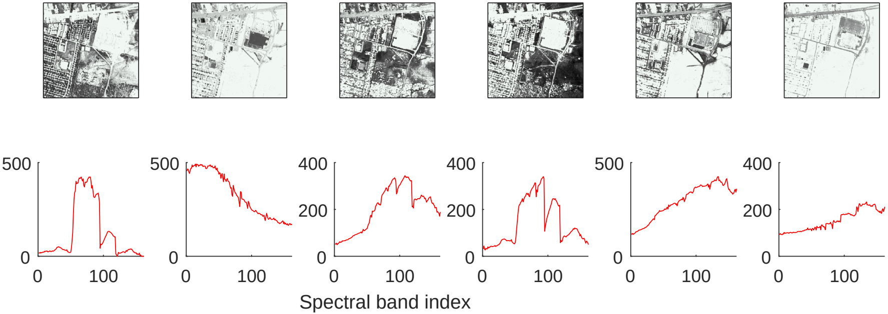
- ✓ Can be adapted to N-way arrays.
- ✓ Can be adapted for more complex estimation schemes of A and B.
- ✓ One iteration has the same complexity as geometric methods.
- ✓ Low memory requirements.
- ✓ Tries to minimize an explicit cost function.

Cons

- ✗ Very sensitive to initialization.
- ✗ No convergence proof.
- ✗ Requires the knowledge of R.

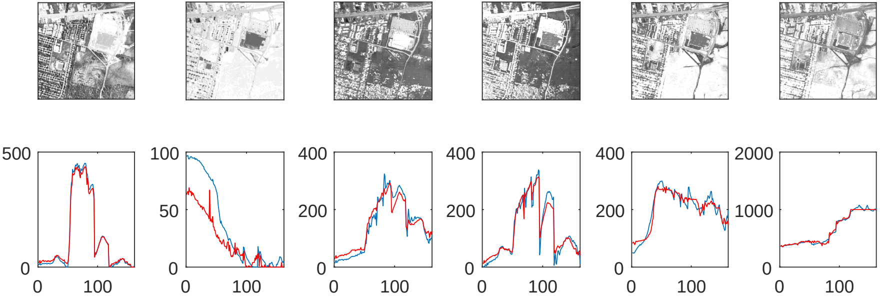
Application to Spectral Unmixing with Pure pixels

Spectra extracted exactly from the data (in red)



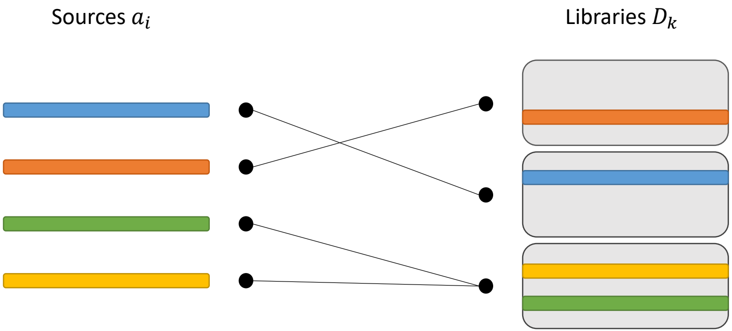
Application to Spectral Unmixing with Pure pixels

Spectra (in blue) close the data (in red)



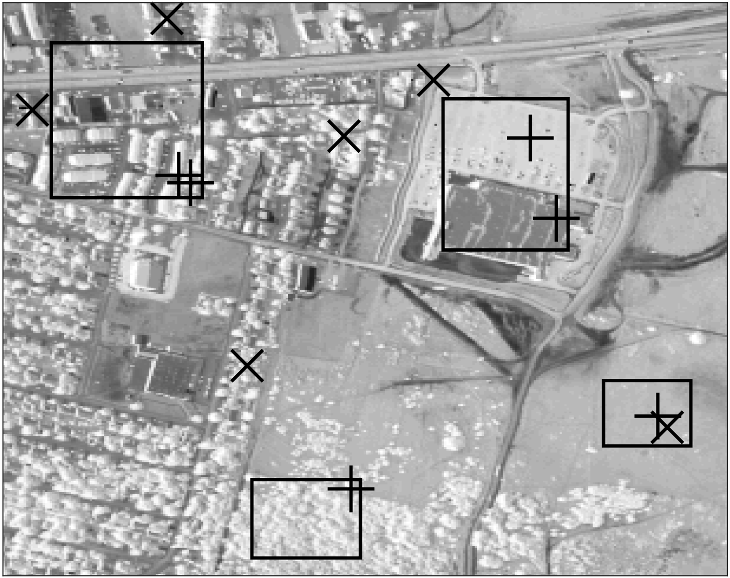
Toolbox available on my personal webpage `jeremy-e-cohen.jimdo.com`
[Cohen Gillis, 2017]

Multiple Dictionary for Hand-Picking Pure Pixels

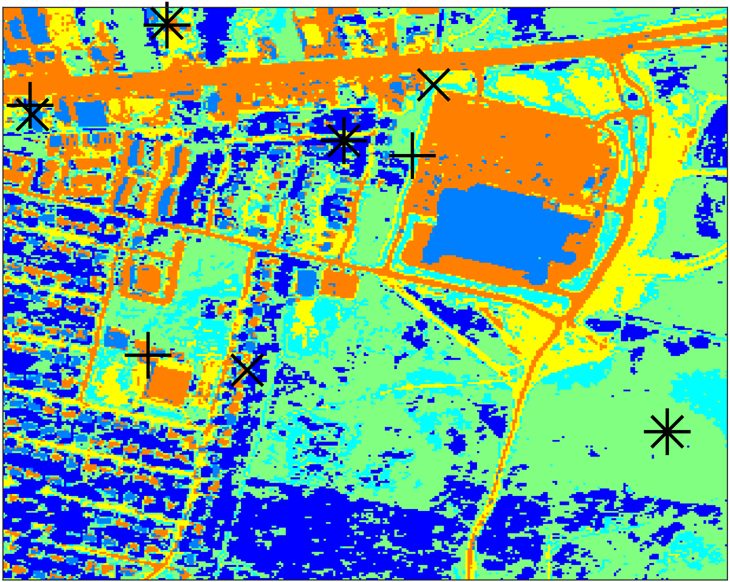


$$\mathbf{A} = [\mathbf{D}_1(:, \mathcal{K}_1), \ldots, \mathbf{D}_p(:, \mathcal{K}_p)]\mathbf{\Pi}$$

Example: Supervised Multiple Dictionary learning

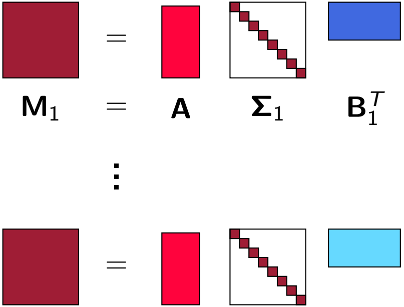


Example: Unsupervised version using segmentation



Challenge II: Subject Variability and Multimodality.

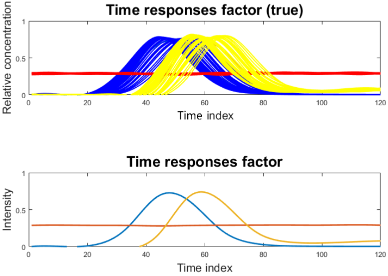
Subject Variability



$$\mathbf{M}_1 = \mathbf{A} \quad \boldsymbol{\Sigma}_1 \quad \mathbf{B}_1^T$$

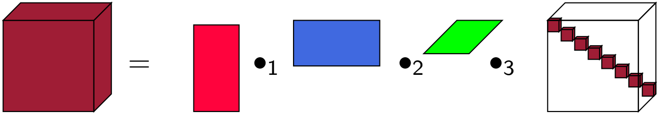
$$\vdots$$

$$\mathbf{M}_i = \mathbf{A} \quad \boldsymbol{\Sigma}_i \quad \mathbf{B}_i^T$$

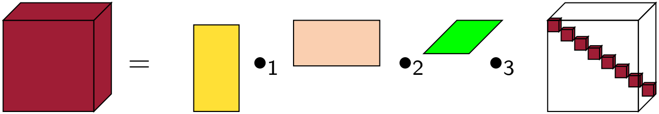


Example: LC-MS data.

Data Fusion with tensors



$$\mathcal{T}_1 = (\mathbf{A}_1 \otimes \mathbf{B}_1 \otimes \mathbf{C}_1)\, \mathcal{I}_R$$



$$\mathcal{T}_2 = (\mathbf{A}_2 \otimes \mathbf{B}_2 \otimes \mathbf{C}_2)\, \mathcal{I}_R$$

Example: Fluorescence and NMR data. Often $\mathbf{C}_1 := \mathbf{C}_2$. But:

- the sampling rates can be different?
- the relation may not be trivial? Can it be learned?
- how does coupling affect the cost function?

General Framework using a Bayesian approach [Cabral Farias, Cohen,2015]

- Parameters $\boldsymbol{\theta}_i = \left[\ \text{vec}(\mathbf{A}_i); \text{vec}(\mathbf{B}_i); \text{vec}(\mathbf{C}_i) \ \right]$ are random
- Known prior distribution $p\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\right)$ and likelihoods $p(\mathcal{Y}_i|\boldsymbol{\theta}_i)$
- Deterministic point of vue: $\boldsymbol{\theta}_i = \phi_i(\boldsymbol{\theta}^*)$ for some fixed function ϕ_i.

MAP estimation under conditionnal independance

$$\underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}{\arg\max}\ p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N|\mathcal{Y}_1, \ldots, \mathcal{Y}_N) = \underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}{\arg\min}\ \Upsilon(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$$

$$\begin{aligned} \Upsilon(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) &= -\sum_{i=1}^{N} \log p(\mathcal{Y}_{\mathbf{i}}|\boldsymbol{\theta}_i) - \log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \\ &= \text{ data fitting terms } + \text{ coupling} \end{aligned}$$

Some flexible coupled LRA models

Noisy exact coupling on $\mathbf{C}_i$

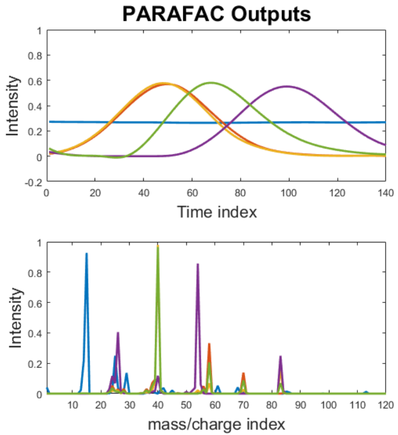
$$\forall\ i \in [1, N], \quad \begin{cases} \mathcal{T}_i & = (\mathbf{A}_i \otimes \mathbf{B}_i \otimes \mathbf{C}_i)\mathcal{I}_R + \mathcal{E}_i \\ \mathbf{C}_i & = \mathbf{C}^* + \mathbf{\Gamma}_i \\ \mathbf{\Gamma}_i & \sim \mathcal{AN}\left(\mathbf{0}, \frac{1}{\sigma_{c,i}^2}\mathbf{I} \otimes \mathbf{I}\right) \end{cases}$$

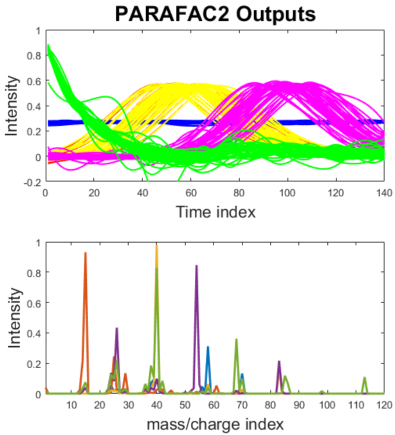
$$\Upsilon(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \mathbf{C}^*) = -\sum_{i=1}^{N} \frac{1}{\sigma_1^2} \|\mathcal{T}_i - (\mathbf{A}_i \otimes \mathbf{B}_i \otimes \mathbf{C}_i)\mathcal{I}_R\|_F^2 - \sum_{i=1}^{N} \frac{1}{\sigma_{ci}^2} \|\mathbf{C}_i - \mathbf{C}^*\|_F^2$$

PARAFAC2 [Harshman,1972][Bro,1999][Cohen,2018]

$$\forall\ i \in [1, N], \quad \begin{cases} \mathbf{M}_i & = \mathbf{A}_i \mathbf{\Sigma}_i \mathbf{B}_i^T + \mathbf{E}_i \\ \mathbf{A}_i & = \mathbf{A}^* \\ \mathbf{B}_i & = \mathbf{P}_i \mathbf{B}^* \\ \mathbf{P}_i^T \mathbf{P}_i & = \mathbf{I} \end{cases}$$

PARAFAC2 vs PARAFAC on LC-MS data





Many other solutions can be thought of to tackle subject variability!

Perspectives

Variability along time / sensors

- Characterize variations along time/sensors within a multiway model? Or in a statistical manner, *i.e.* with priors on the evolution of coupled parameters?
- Applications: automatic stereo transcriptions, temporal spectral unmixing and super resolution. . .

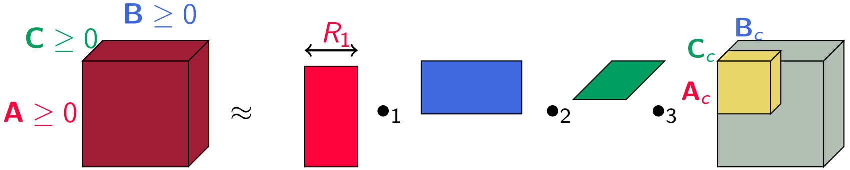
Interactions between machine learning and multimodality

- Learning the coupling relationship
- Tensor dictionary learning

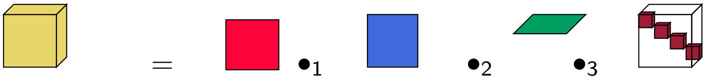
Challenge III: Constrained large decompositions.

I S B N 2-205-00921-4

Unconstrained compression . . .



$$\mathcal{T} \approx (\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W})\, \mathcal{G}$$



$$\mathcal{G} = (\mathbf{A_c} \otimes \mathbf{B_c} \otimes \mathbf{C_c})\, \mathcal{I}_R$$

$$\mathcal{T} \approx (\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W})\, \mathcal{G} = (\mathbf{UA_c} \otimes \mathbf{VB_c} \otimes \mathbf{WC_c})\, \mathcal{I}_R$$

...but constrained decomposition!

Compressed domain NN CP:

$$\min_{\mathbf{A}_c,\mathbf{B}_c,\mathbf{C}_c} \quad \|\mathcal{G} - (\mathbf{A}_c \otimes \mathbf{B}_c \otimes \mathbf{C}_c)\mathcal{I}\|$$
$$\text{sub. to} \quad \widehat{\mathbf{U}}\mathbf{A}_c, \widehat{\mathbf{V}}\mathbf{B}_c, \widehat{\mathbf{W}}\mathbf{C}_c \geqslant 0$$

Issue	Solution
Easy unconstrained/difficult constrained	Unconstrained solution $\rightarrow$ projection
Difficult exact projection $\widehat{\mathbf{U}}\mathbf{A}_c$	Approximate projection

Approximate projection and PROCO-ALS

Approximate projection $\mathbf{\Pi}$:

Given Least Squares update $\widehat{\mathbf{A}}_c$

1. Decompression: $\widehat{\mathbf{A}} := \widehat{\mathbf{U}}\widehat{\mathbf{A}}_c$
2. Projection: $\widehat{\mathbf{A}} := \left[\widehat{\mathbf{A}}\right]^+$
3. Compression: $\widehat{\mathbf{A}}_c := \widehat{\mathbf{U}}^\mathsf{T}\widehat{\mathbf{A}}$

$$\mathbf{\Pi}\left[\widehat{\mathbf{A}}\right] = \mathbf{U}^\mathsf{T}[\mathbf{U}\mathbf{A}_c]^+$$

Projected and compressed framework (PROCO) [Cohen,2014]

Other possible algorithms and related problems

- PROCO-ALS [Cohen,2014], Compressed-AOADMM [Cohen,2016]

$$\begin{aligned}&\text{minimize } \|\widehat{\mathcal{G}} - (\mathbf{A}_c \otimes \mathbf{B}_c \otimes \mathbf{C}_c)\mathcal{I}_R\|_F^2 \\ &\text{w.r.t. } \mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c \\ &\text{s.t. } \widehat{\mathbf{U}}\mathbf{A}_c \succeq 0\end{aligned}$$

- Tensorlab 3.0 [Vervliet,2016]

$$\begin{aligned}&\text{minimize } \| \left(\widehat{\mathbf{U}} \otimes \widehat{\mathbf{V}} \otimes \widehat{\mathbf{W}}\right) \widehat{\mathcal{G}} - (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\mathcal{I}_R\|_F^2 \\ &\text{w.r.t. } \mathbf{A}, \mathbf{B}, \mathbf{C} \\ &\text{s.t. } \mathbf{A} \succeq 0\end{aligned}$$

- AOADMM [Huang,2015], FastNNLS [Bro,1997], ANLS

$$\begin{aligned}&\text{minimize } \|\mathcal{T} - (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\mathcal{I}_R\|_F^2 \\ &\text{w.r.t. } \mathbf{A}, \mathbf{B}, \mathbf{C} \\ &\text{s.t. } \mathbf{A} \succeq 0\end{aligned}$$

Application in Fluorescence Spectroscopy

Fluorescence spectroscopy data:

excitation spectra
emission spectra
mixtures

multimodal chemometrics data set from Acar *et al*[1]

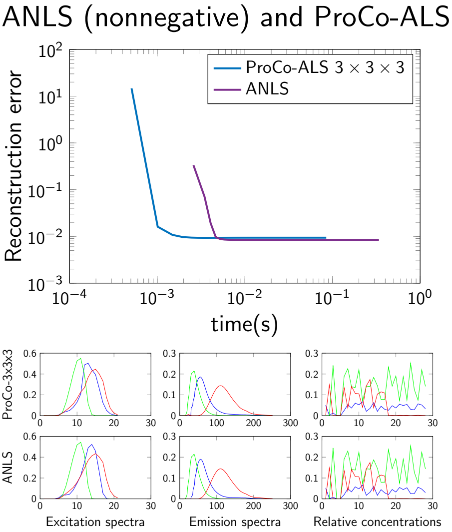
Description

5 compounds: Valine-Tyrosine-Valine (Val), Tryptophan- Glycine (Gly), Phenylalanine (Phe), Maltoheptaose (Mal) and Propanol (Pro)

Nb. of excitation wave lengths	21 (**A**)
Nb. of emission wave lengths	251 (**B**)
Nb. of Mixtures	28 (**C**)
Missing values	30% (replaced by zeros)

[1] E. Acar, A.J. Lawaetz, M.A. Rasmussen, and R. Bro. Structure-revealing data fusion model with applications in metabolomics. In Conf. Proc. IEEE Eng. Med. Biol. Soc., pages 6023– 6026. IEEE, 2013

Application to Fluorescence Spectroscopy



Conclusions and Perspectives

Studied topics

- Identification through known dictionaries.
- Multimodality and Subject Variability in matrix/tensor low rank factorization models for chemometrics/neuroimaging.
- Constrained tensor compression and decomposition, especially in the context of nonnegativity.

Things I am (or would like to be) working on

- Flexible dictionary constraints, tensor dictionary learning.
- Data fusion for temporal series of hyperspectral images.
- Multispectral/Hyperspectral fusion for spectral unmixing.
- Audio source separation with tensor models, which calls for new tensor decomposition models and non-euclidean error metrics.

Thank you for your attention!

State-of-the-art (non-exhaustive)

- **Continuous approaches**
  Lasso, GLUP [Ammanouil 2014], FGNSR [Gillis 2016]
  + Robust, optimization criterion. − Slow.
- **Greedy/Non-iterative method**
  - Geometric algorithms (pure pixel hypothesis)
    N-FINDR [Winter 1999], VCA [Nascimento 2005], SPA [Gillis 2014, Businger Golub 1965]
  - Matching pursuit approaches
    SDSOMP[X.Fu 2013, Tropp 2006]

  + Fast − Not robust, No explicit criterion
- **Pixel-wise brute force algorithms**
  MESMA [Roberts 1998], MESLUM, AUTOMCU, AMUSES [Degerickx 2017]
  + Flexible − No low rank property, Slow.
- **Statistical methods**

Experiment on the URBAN HSI

	$r = 6$		$r = 8$	
	Time (s.)	Rel. err.	Time (s.)	Rel. err.
RAND-wo	0.00	7.87	0.00	11.66
d-RAND-wo	22.46 (13)	5.09	34.87 (18)	5.35
RAND-av	0.02	11.51	0.02	9.60
d-RAND-av	23.91 (13)	4.65	30.77 (15)	4.65
RAND-be	0.00	13.77	0.00	5.54
d-RAND-be	22.01 (11)	4.36	36.18 (19)	4.16
VCA	2.01	18.38	1.86	20.11
d-VCA	26.89 (15)	5.83	29.06 (14)	5.05
SPA	0.30	9.58	0.30	9.45
d-SPA	24.37 (13)	4.67	28.61 (14)	4.62
SNPA	24.34	9.63	36.72	5.64
d-SNPA	23.04 (13)	4.94	27.94 (13)	**3.97**
H2NMF	19.02	5.81	22.35	5.47
d-H2NMF	26.66 (15)	**4.05**	28.92 (14)	4.24
FGNSR-100	2.73	5.58	2.55	4.62
d-FGNSR-100	26.72 (14)	4.36	20.81 (8)	4.04

Table: Numerical results for the Urban data set.