Nonnegative matrix factorizations with the β-divergence

Cédric Févotte

Institut de Recherche en Informatique de Toulouse (IRIT)

École d'Été de Peyresq
Juin 2024

Outline

Generalities
- Matrix factorization models
- Nonnegative matrix factorization (NMF)

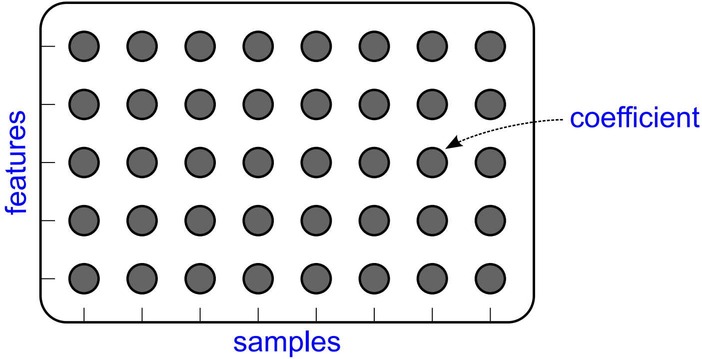
Optimization for NMF
- Measures of fit
- Majorization-minimization
- Other algorithms
- Hyperparameters selection

Regularized NMF
- Common regularizers : sparsity, smoothness
- Automatic relevance determination

Examples in imaging
- Robust NMF for nonlinear hyperspectral unmixing
- Factor analysis in dynamic PET

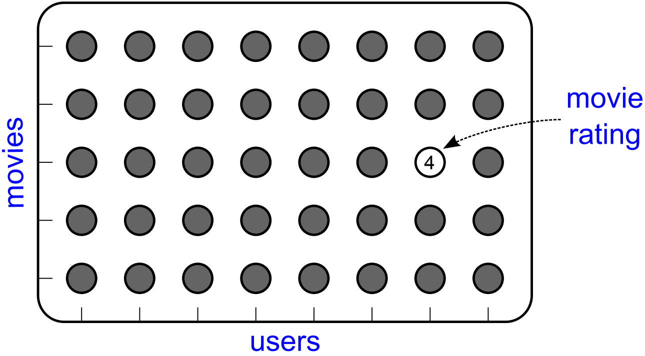
Matrix factorization models

Data often available in matrix form.



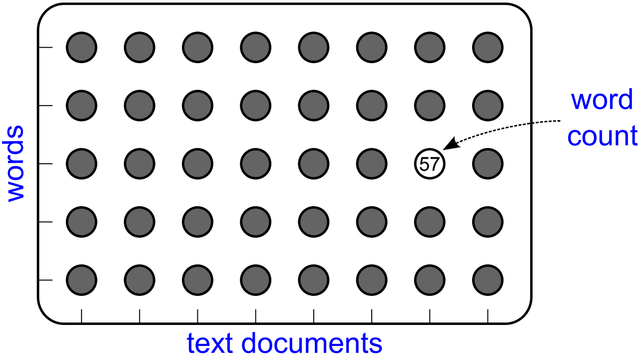
Matrix factorization models

Data often available in matrix form.



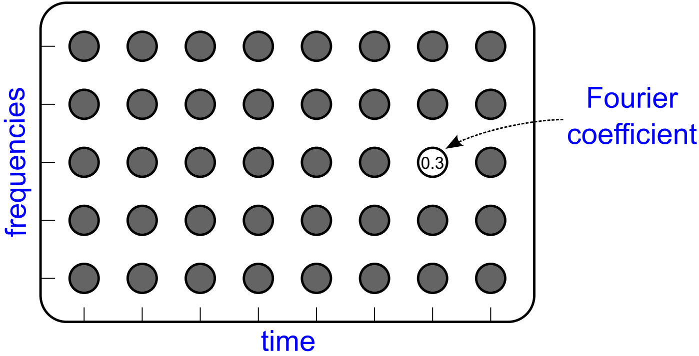
Matrix factorization models

Data often available in matrix form.



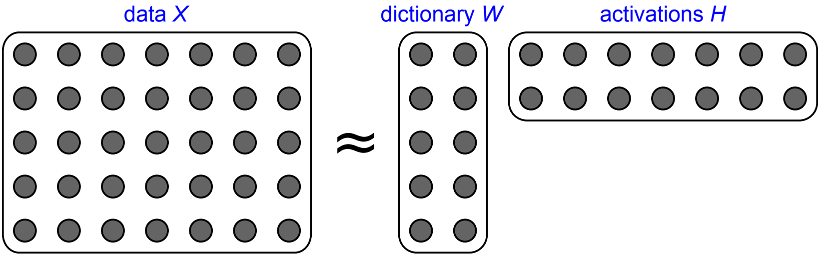
Matrix factorization models

Data often available in matrix form.



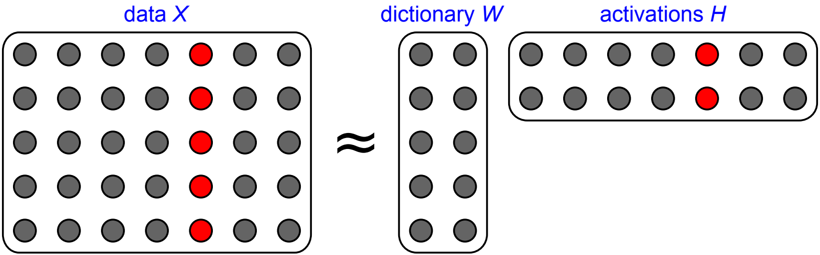
Matrix factorization models

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



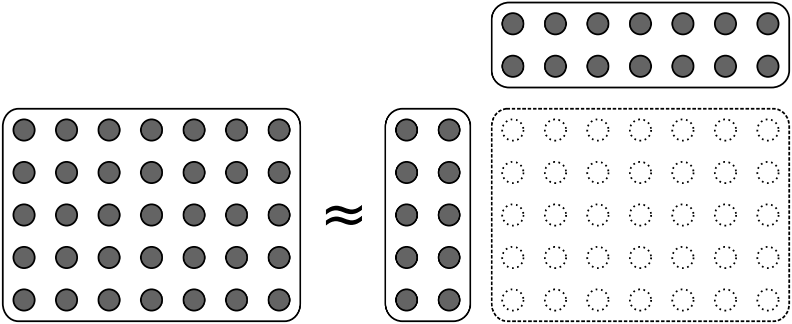
Matrix factorization models

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



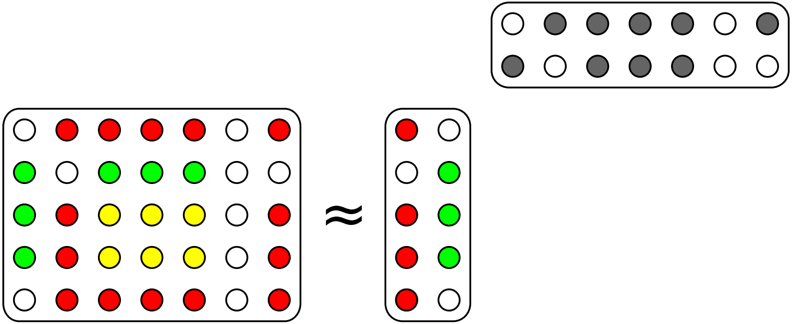
Matrix factorization models

for dimensionality reduction (coding, low-dimensional embedding)



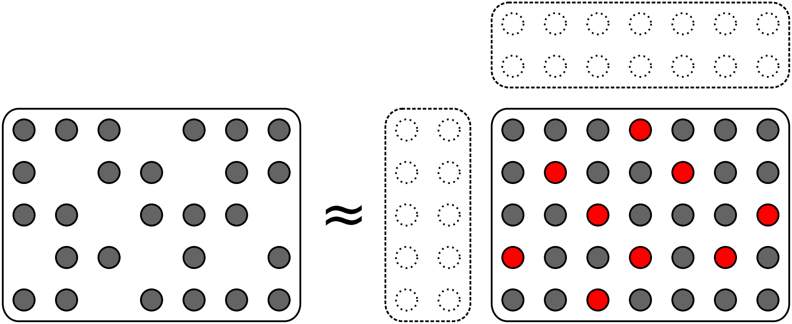
Matrix factorization models

for unmixing (source separation, latent topic discovery)



Matrix factorization models

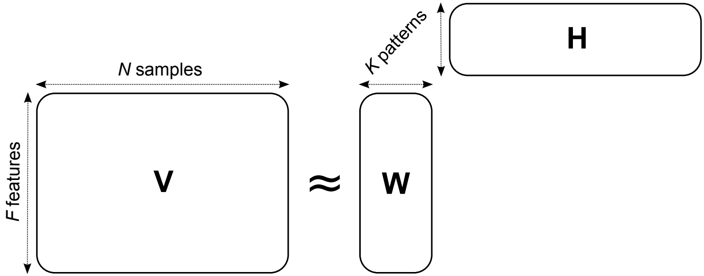
for completion (collaborative filtering, image inpainting)



Matrix factorization models

- Simple generative & interpretable models, popular in unsupervised settings.
- Used in many fields for a long time :
  - Principal component analysis PCA (Pearson, 1901)
  - Factor analysis (Spearman, 1904)
  - Latent semantic analysis LSA (Deerwester et al., 1988)
  - Independent component analysis ICA (Comon, 1994)
  - Nonnegative matrix factorization NMF (Lee & Seung, 1999)
  - Latent Dirichlet allocation LDA (Blei et al., 2003)
  - Sparse dictionary learning, e.g., K-SVD (Aharon et al., 2006)
- Active topics :
  - design of nonconvex optimization algorithms with proven convergence
  - landscape analysis, search for global optima
  - conditions for identifiability
  - rank selection
  - probabilistic models & statistical approaches (e.g., integer-valued or binary data)

Nonnegative matrix factorization

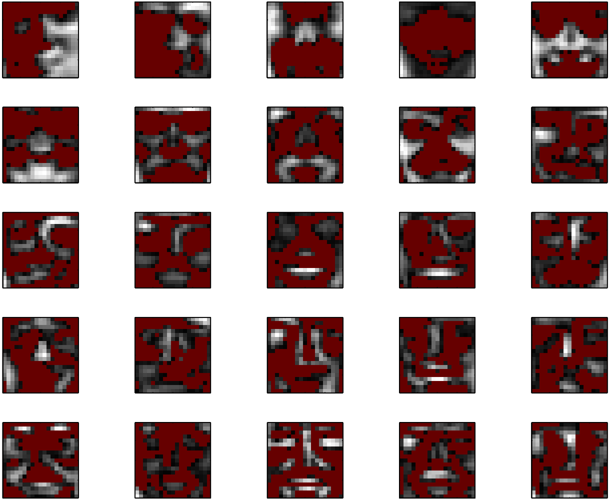


- Data **V** and factors **W**, **H** have nonnegative entries.
- Nonnegativity of **W** ensures interpretability of the dictionary, because patterns $\mathbf{w}_k$ and samples $\mathbf{v}_n$ belong to the same space.
- Nonnegativity of **H** tends to produce part-based representations, because subtractive combinations are forbidden.

Early work by (Paatero and Tapper, 1994), landmark *Nature* paper by (Lee and Seung, 1999)

49 images among 2429 from MIT's CBCL face dataset

PCA dictionary with $K = 25$



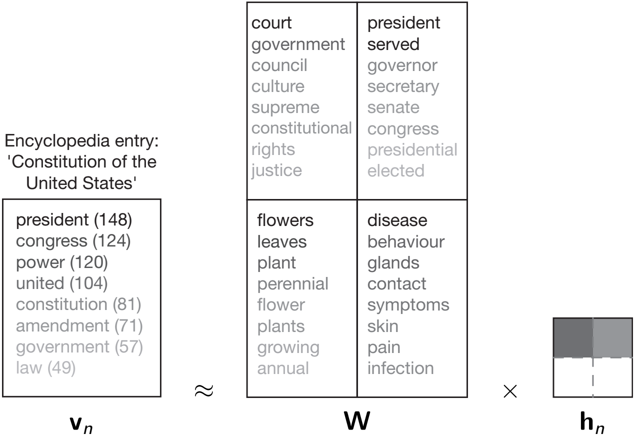
red pixels indicate negative values

NMF dictionary with $K = 25$

experiment reproduced from (Lee and Seung, 1999)

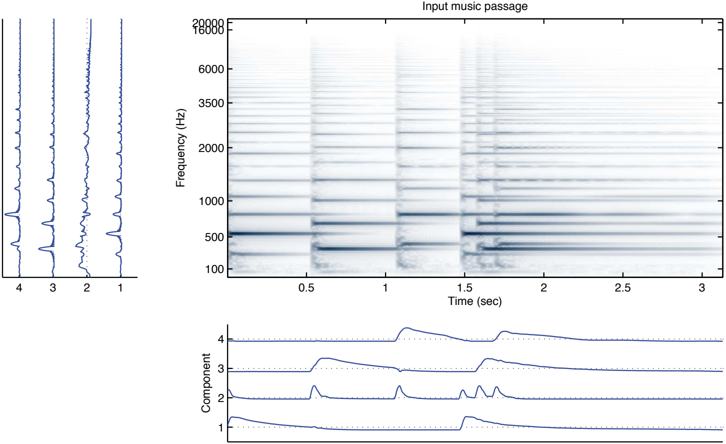
NMF for latent semantic analysis

(Lee and Seung, 1999; Hofmann, 1999)



reproduced from (Lee and Seung, 1999)

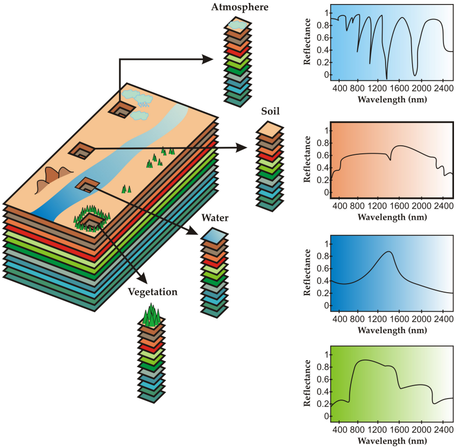
NMF for audio spectral unmixing
(Smaragdis and Brown, 2003)



reproduced from (Smaragdis, 2013)

NMF for hyperspectral unmixing

(Berry, Browne, Langville, Pauca, and Plemmons, 2007)



reproduced from (Bioucas-Dias et al., 2012)

Outline

NMF as a constrained minimization problem

Minimize a measure of fit between **V** and **WH**, subject to nonnegativity :

$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}),$$

where $d(x|y)$ is a scalar cost function, e.g.,

- squared Euclidean distance (Paatero and Tapper, 1994; Lee and Seung, 2001)
- Kullback-Leibler divergence (Lee and Seung, 1999; Finesso and Spreij, 2006)
- Itakura-Saito divergence (Févotte, Bertin, and Durrieu, 2009)
- α-divergence (Cichocki et al., 2008)
- β-divergence (Cichocki et al., 2006; Févotte and Idier, 2011)
- Bregman divergences (Dhillon and Sra, 2005)
- and more in (Yang and Oja, 2011)

Regularization terms often added to $D(\mathbf{V}|\mathbf{WH})$ for sparsity, smoothness, etc.
Nonconvex problem.

Probabilistic models

- Let $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$ such that
  - $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$
  - $p(\mathbf{V}|\mathbf{WH}) = \prod_{fn} p(v_{fn}|[\mathbf{WH}]_{fn})$
- then the following correspondences apply with

$$D(\mathbf{V}|\mathbf{WH}) = -\log p(\mathbf{V}|\mathbf{WH}) + \mathrm{cst}$$

data support	distribution/noise	divergence	examples
real-valued	additive Gaussian	quadratic loss	many
integer	multinomial*	weighted KL	word counts
integer	Poisson	generalized KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectrogram
generally nonnegative	Tweedie	β-divergence	generalizes above models

*conditional independence over f does not apply

The β-divergence

A popular measure of fit in NMF (Basu et al., 1998; Cichocki and Amari, 2010)

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)} \left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \log \frac{x}{y} + (y-x) & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases}$$

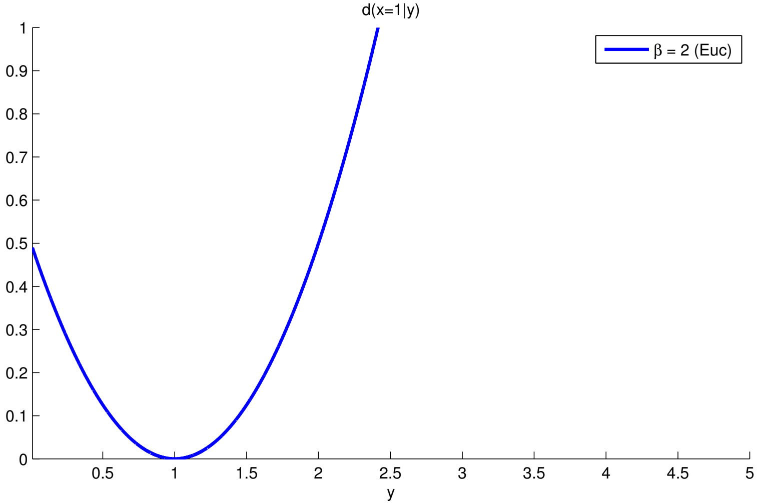
Special cases :

- squared Euclidean distance a.k.a quadratic loss ($\beta = 2$)
- generalized Kullback-Leibler (KL) divergence ($\beta = 1$)
- Itakura-Saito (IS) divergence ($\beta = 0$)

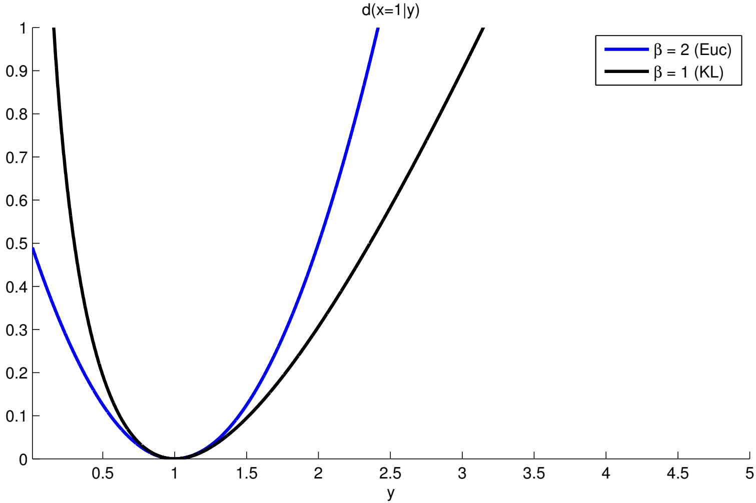
Properties :

- Homogeneity : $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$
- $d_\beta(x|y)$ is a convex function of y for $1 \le \beta \le 2$
- Bregman divergence

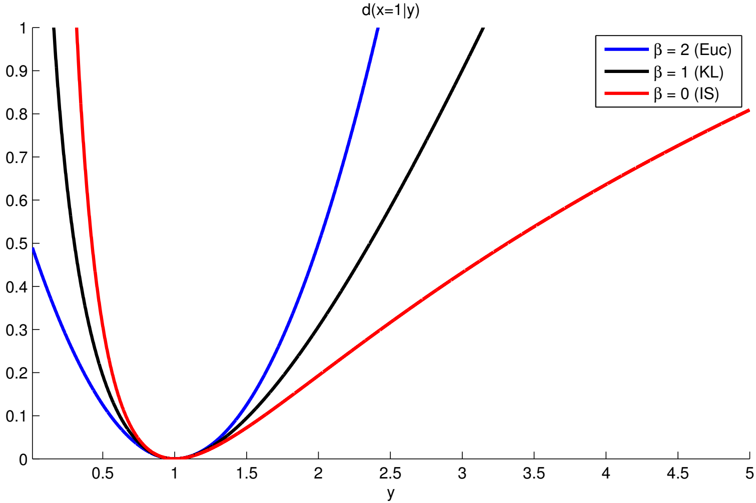
The β-divergence



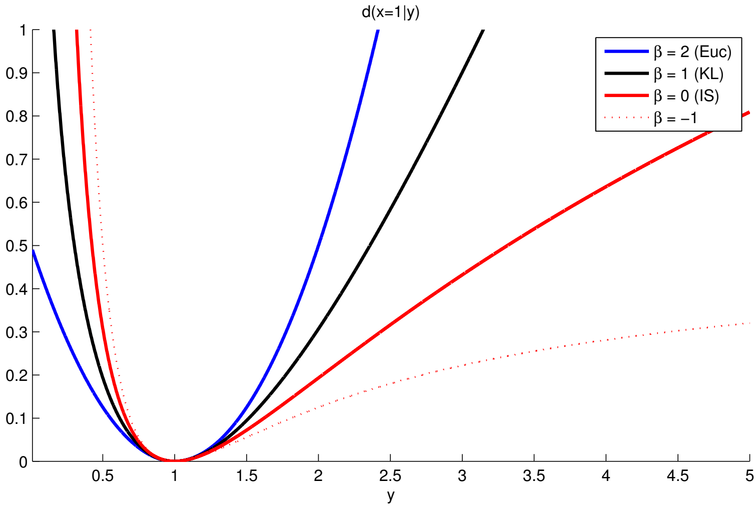
The β-divergence



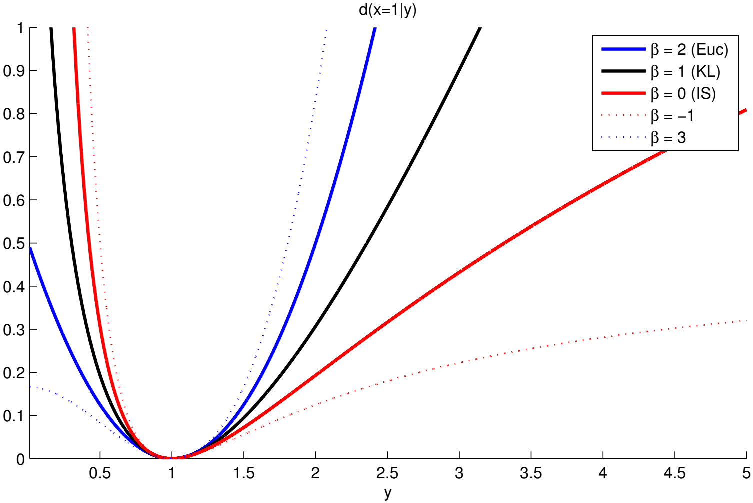
The β-divergence



The β-divergence



The β-divergence



A common NMF algorithm design : alternating methods

- Block-coordinate update of $\mathbf{H}$ given $\mathbf{W}^{(i-1)}$ and $\mathbf{W}$ given $\mathbf{H}^{(i)}$.
- Updates of $\mathbf{W}$ and $\mathbf{H}$ equivalent by transposition :

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \Leftrightarrow \mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$$

- Objective function separable in the columns of $\mathbf{H}$ or the rows of $\mathbf{W}$:

$$D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$$

- Essentially left with nonnegative linear regression :

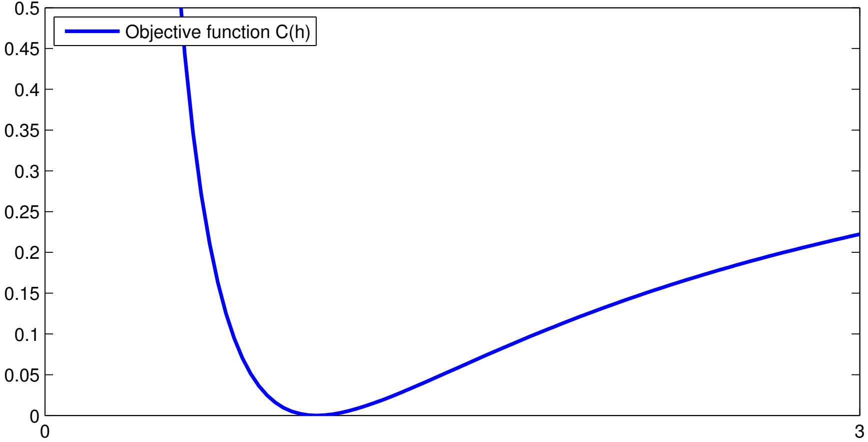
$$\min_{\mathbf{h} \geq \mathbf{0}} C(\mathbf{h}) \overset{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h})$$

  Numerous references in the image restoration literature, e.g., (Richardson, 1972; Lucy, 1974; Daube-Witherspoon and Muehllehner, 1986; De Pierro, 1993)

Block-descent algorithm, nonconvex problem, initialization is an issue.

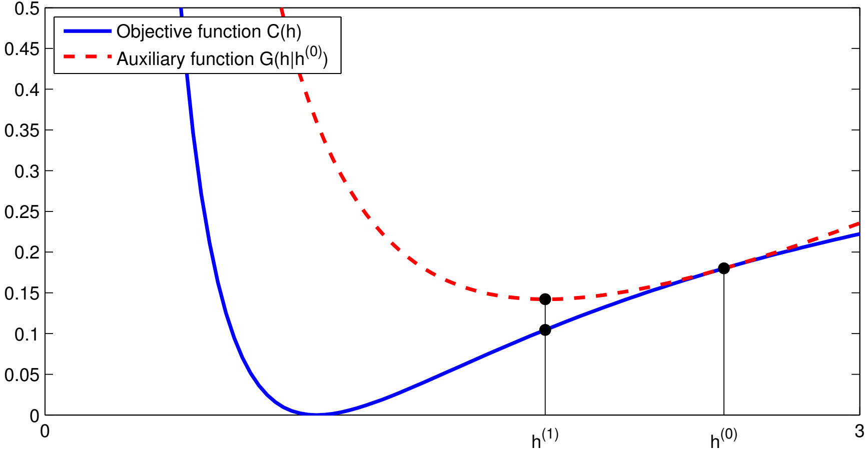
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



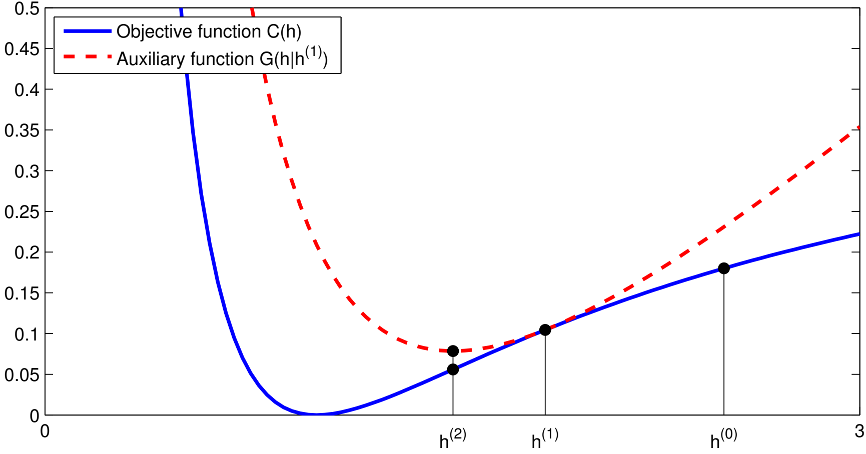
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



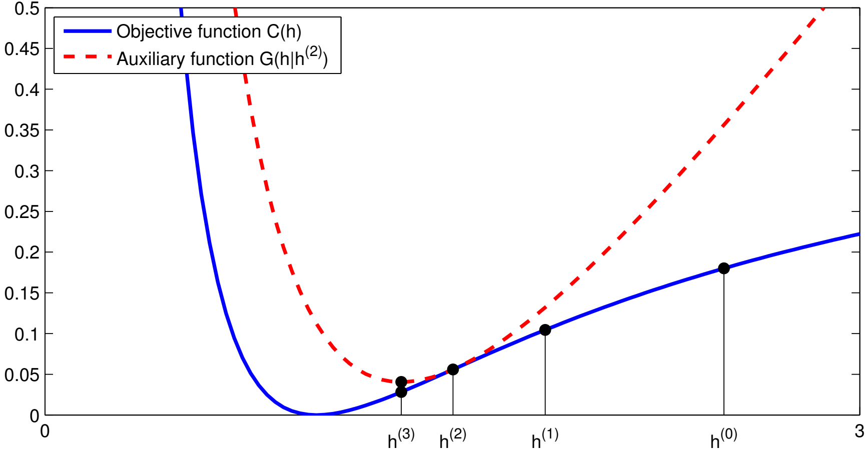
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



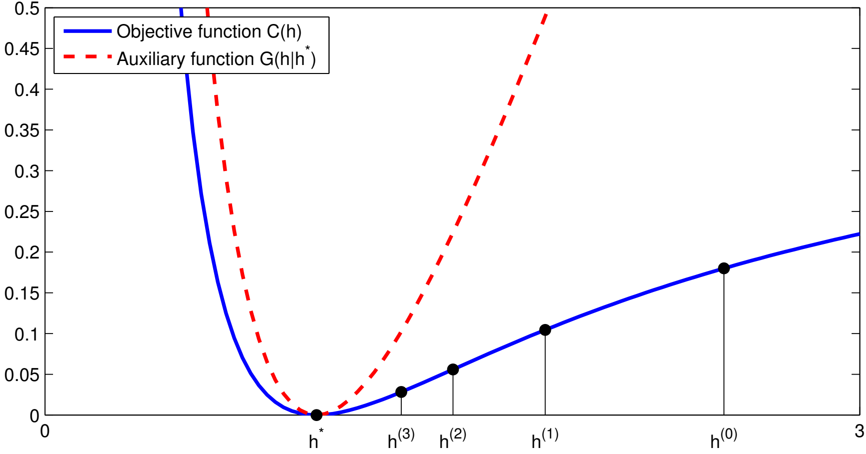
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



Majorization-minimization (MM)

- Finding a good & workable local majorization is the crucial point.
- Treating convex and concave terms separately with Jensen and tangent inequalities usually works. E.g. :

$$C_{\text{IS}}(\mathbf{h}) = \left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right] + \left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right] + cst$$

Majorization-minimization (MM)

- Finding a good & workable local majorization is the crucial point.
- Treating convex and concave terms separately with Jensen and tangent inequalities usually works. E.g. :

$$C_{\mathrm{IS}}(\mathbf{h}) = \left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right] + \left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right] + cst$$

- In most cases, leads to nonnegativity-preserving multiplicative algorithms :

$$h_k = \tilde{h}_k \left(\frac{\nabla^-_{h_k} C(\tilde{\mathbf{h}})}{\nabla^+_{h_k} C(\tilde{\mathbf{h}})}\right)^\gamma$$

  - $\nabla_{h_k} C(\mathbf{h}) = \nabla^+_{h_k} C(\mathbf{h}) - \nabla^-_{h_k} C(\mathbf{h})$ and the two summands are nonnegative.
  - if $\nabla_{h_k} C(\tilde{\mathbf{h}}) > 0$, ratio of summands < 1 and h_k decreases.
  - γ is a divergence-specific scalar exponent.
- Details in (Nakano et al., 2010; Févotte and Idier, 2011; Yang and Oja, 2011)

Example : derivation for the Itakura-Saito divergence

- IS divergence ($\beta = 0$)

$$d_{\text{IS}}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$$

- Nonnegative linear regression with the IS divergence

$$\min_{\mathbf{h}\geq 0} C_{\text{IS}}(\mathbf{h}) = \sum_f d_{\text{IS}}(v_f|[\mathbf{Wh}]_f)$$

$$= \underbrace{\left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right]}_{C_1(\mathbf{h}) \text{ (convex)}} + \underbrace{\left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right]}_{C_2(\mathbf{h}) \text{ (concave)}} + cst$$

Example : derivation for the Itakura-Saito divergence

- Majorization of $C_1(\mathbf{h})$ with Jensen's inequality.
  Let $f(x)$ be a convex function and $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ with $\sum_k \lambda_k = 1$. Then :

$$f\left(\sum_k \lambda_k h_k\right) \leq \sum_k \lambda_k f(h_k).$$

- Let $\tilde{\mathbf{h}} \in \mathbb{R}_+^K$ be the current estimate, $\tilde{\mathbf{v}} = \mathbf{W}\tilde{\mathbf{h}}$ be the current approximation and

$$\lambda_{fk} = \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} = \frac{w_{fk}\tilde{h}_k}{\sum_j w_{fj}\tilde{h}_j} \quad \left(\text{note that } \sum_k \lambda_{fk} = 1\right).$$

- Then, by convexity of $f(x) = x^{-1}$, we may write :

$$C_1(\mathbf{h}) = \sum_f v_f \left(\sum_k w_{fk} h_k\right)^{-1} = \sum_f v_f \left(\sum_k \lambda_{fk} \frac{w_{fk} h_k}{\lambda_{fk}}\right)^{-1}$$

$$\leq \sum_{fk} v_f \frac{\lambda_{fk}^2}{w_{fk} h_k} = \sum_{fk} w_{fk} \frac{v_f}{\tilde{v}_f^2} \frac{\tilde{h}_k^2}{h_k} = G_1(\mathbf{h}|\tilde{\mathbf{h}}).$$

Example : derivation for the Itakura-Saito divergence

- Majorization of $C_2(\mathbf{h})$ with the tangent inequality.

  Let $g(\mathbf{h})$ be a concave function then :

$$g(\mathbf{h}) \leq g(\tilde{\mathbf{h}}) + \nabla g(\tilde{\mathbf{h}})^\top (\mathbf{h} - \tilde{\mathbf{h}}) = \sum_k [\nabla g(\tilde{\mathbf{h}})]_k h_k + cst.$$

- Given $C_2(\mathbf{h}) = \sum_f \log\left(\sum_k w_{fk} h_k\right)$, we have :

$$[\nabla C_2(\tilde{\mathbf{h}})]_k = \nabla_{h_k} C_2(\tilde{\mathbf{h}}) = \sum_f \frac{w_{fk}}{\tilde{v}_f}.$$

- Finally, we may majorize $C_2(\mathbf{h})$ with :

$$G_2(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_{fk} \frac{w_{fk}}{\tilde{v}_f} h_k + cst.$$

Example : derivation for the Itakura-Saito divergence

- In the end, we may majorize $C_{\text{IS}}(\mathbf{h})$ with :

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = G_1(\mathbf{h}|\tilde{\mathbf{h}}) + G_2(\mathbf{h}|\tilde{\mathbf{h}}) + cst$$
$$= \sum_{fk} w_{fk} \left[\frac{v_f}{\tilde{v}_f^2} \frac{\tilde{h}_k^2}{h_k} + \frac{1}{\tilde{v}_f} h_k \right] + cst.$$

- Smooth, convex and separable majorizer. Easily minimized by cancelling its gradient, leading to the MM-based multiplicative update

$$h_k = \tilde{h}_k \left(\frac{\sum_f w_{fk} v_f [\mathbf{W}\tilde{\mathbf{h}}]_f^{-2}}{\sum_f w_{fk} [\mathbf{W}\tilde{\mathbf{h}}]_f^{-1}} \right)^{\frac{1}{2}}.$$

- Algorithm known from (Cao et al., 1999). The $\frac{1}{2}$ exponent can be dropped using majorization-equalization (Févotte and Idier, 2011).

The multiplicative updates (MU) for NMF with β-divergence

- Alternating updates of $\mathbf{W}$ and $\mathbf{H}$.
- In standard practice, only one MM update applied to $\mathbf{W}$ and $\mathbf{H}$, rather than fully solving subproblems $\min_{\mathbf{W}\geq 0} D(\mathbf{V}|\mathbf{WH})$ and $\min_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})$.
- Leads to a valid descent algorithm with multiplicative updates given by :

$$\mathbf{H} \leftarrow \mathbf{H}. \left(\frac{\mathbf{W}^T \left[(\mathbf{WH})^{\cdot(\beta-2)} . \mathbf{V} \right]}{\mathbf{W}^T \left[\mathbf{WH} \right]^{\cdot(\beta-1)}} \right)^{\gamma(\beta)}$$

$$\mathbf{W} \leftarrow \mathbf{W}. \left(\frac{\left[(\mathbf{WH})^{\cdot(\beta-2)} . \mathbf{V} \right] \mathbf{H}^T}{\left[\mathbf{WH} \right]^{\cdot(\beta-1)} \mathbf{H}^T} \right)^{\gamma(\beta)}$$

- Very straightforward implementation, no hyperparameters !
- Nonnegativity is automatically preserved given positive initializations.
- Linear complexity per iteration.
- In practice, minimizing $D(\mathbf{V}+\epsilon|\mathbf{WH}+\epsilon)$ prevents from numerical issues.

Convergence of the iterates

- By design, we have convergence of the objective values $C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH})$.
- What about the iterates ? Only partial answers so far.
- A theoretical challenge arises from the lack of coercivity of the objective : $\|\mathbf{W}\|$ or $\|\mathbf{H}\| \to \infty \not\Rightarrow C(\mathbf{W},\mathbf{H}) \to \infty$.
- Due to the scale indeterminacy : $C(\mathbf{W\Lambda}^{-1},\mathbf{\Lambda H}) = C(\mathbf{W},\mathbf{H})$, with $\mathbf{\Lambda} \to 0$.

Possible remedies (modified problems)

1) Impose $\mathbf{W} \geq \epsilon$, $\mathbf{H} \geq \epsilon$ (Takahashi et al., 2018; Hien and Gillis, 2021).
2) Slightly change the objective function to ensure coercivity (Zhao and Tan, 2018) :

$$C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \epsilon\|\mathbf{W}\|_1 + \epsilon\|\mathbf{H}\|_1$$

MM results in adding ϵ at the denominator of the multiplicative updates.

Other alternating optimization methods

- MM-based multiplicative updates are a simple and competitive choice for many divergences (beyond β-divergences).
- More efficient options have been proposed for specific measures of fit, see books by Cichocki et al. (2009); Gillis (2020)

Quadratic loss (selection)

- Active-set methods (Kim and Park, 2011)
- Hierarchical alternating LS (Cichocki et al., 2007; Gillis and Glineur, 2012)
- Proximal gradient descent (Lin, 2007; Guan et al., 2012; Bolte et al., 2014)
- ADMM (Sun and Févotte, 2014; Huang et al., 2016)

Kullback-Leibler divergence (selection)

- Second-order coordinate descent methods (Hsieh and Dhillon, 2011)
- Hybrid Newton-type algorithms with line search and MU (Hien and Gillis, 2021)

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant (∼Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant (∼Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).
2) Joint MM algorithm for the β-divergence (Marmin, Goulart, and Févotte, 2023a) :
   - Global majorizer constructed using Jensen and tangent inequalities :

$$C(\mathbf{W}, \mathbf{H}) \leq G(\mathbf{W}, \mathbf{H}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$
$$C(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) = G(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$

   - Global minimizer of G not available in closed form. G non-convex.
   - Alternate minimization of G leads to closed-form updates and new multiplicative rules. Important computational savings for some values of β (see paper).

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant (∼Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).
2) Joint MM algorithm for the β-divergence (Marmin, Goulart, and Févotte, 2023a) :
   - Global majorizer constructed using Jensen and tangent inequalities :

$$C(\mathbf{W}, \mathbf{H}) \leq G(\mathbf{W}, \mathbf{H}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$
$$C(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) = G(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$

   - Global minimizer of G not available in closed form. G non-convex.
   - Alternate minimization of G leads to closed-form updates and new multiplicative rules. Important computational savings for some values of β (see paper).
3) Second-order method for β-NMF based on efficient Hessian approximations and tricks to maintain semidefinite positivity (Vandecappelle et al., 2020).

Large-scale NMF

Online NMF

- Large number of samples $N >> F$.
- Update $\mathbf{W}$ as samples $\mathbf{v}_n$ become available.
- Vectors $\mathbf{h}_n$ act as latent variables, minimize :

$$C(\mathbf{W}) = \sum_{n=1}^{N} \min_{\mathbf{h}_n \geq 0} D(\mathbf{v}_n | \mathbf{W}\mathbf{h}_n)$$

- Solved with online MM (Lefèvre et al., 2011; Mairal, 2015; Zhao et al., 2017)

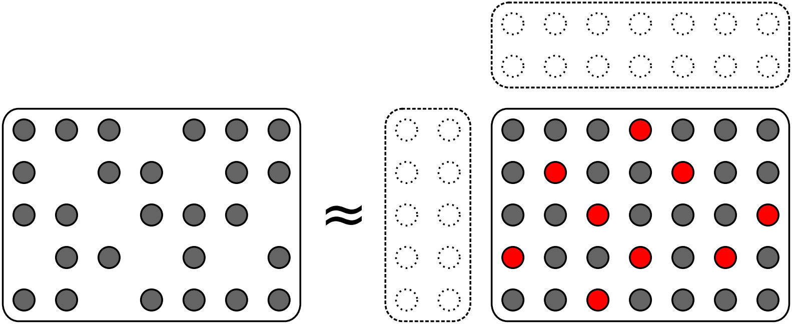
Stochastic NMF

- Large F and N.
- Online NMF with stochastic subsampling :

$$\min_{\mathbf{h}_n \geq 0} D(\mathbf{v}_n[\mathcal{I}] | \mathbf{W}[\mathcal{I}, :]\mathbf{h}_n)$$

where $\mathcal{I}$ is a random set of indices (Mensch et al., 2018).

Selecting hyperparameters K and β with matrix completion



- Matrix completion of held out data using a range of values of β (or K).
- Select β (or K) that best reconstructs held out coefficients v_{fn} with $[\mathbf{WH}]_{fn}$.

Selecting β with matrix completion

- Remove some coefficients of $\mathbf{V}$ randomly.
- Pick a candidate value of β and solve :

$$\min_{\mathbf{W},\mathbf{H}\geq \mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{(f,n)\in\mathcal{O}} d_\beta([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn})$$

  where $\mathcal{O}$ is the set of remaining ("observed") coefficients.
- Optimization can be handled using a mask $\gamma_{fn} \in \{0,1\}$:

$$\sum_{(f,n)\in\mathcal{O}} d_\beta([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}) = \sum_{fn} \gamma_{fn}\, d_\beta([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}) = \sum_{fn} d_\beta(\gamma_{fn}[\mathbf{V}]_{fn}|\gamma_{fn}[\mathbf{WH}]_{fn})$$

- Assess β using a given reconstruction error on held out data :

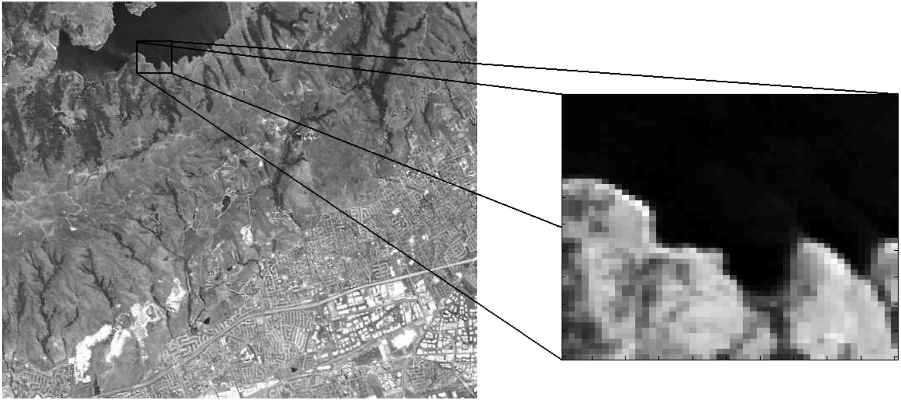
$$L(\beta) = \sum_{(f,n)\in\overline{\mathcal{O}}} \ell(v_{fn}|[\mathbf{WH}]_{fn})$$

- Repeat for other values of β and pick $\hat{\beta}$ with minimum $L(\beta)$.

Selecting β with matrix completion

(Févotte and Dobigeon, 2015)

Moffett Field hyperspectral data



reproduced from (Dobigeon, 2007)

Selecting β with matrix completion

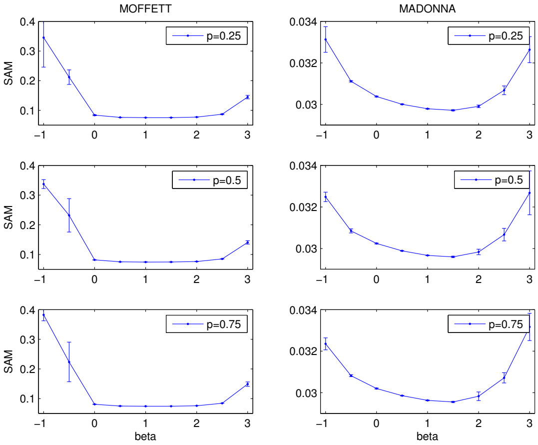
(Févotte and Dobigeon, 2015)

Experimental setting :

- Two unfolded hyperspectral cubes, $F \sim 150$, $N = 50 \times 50$
  - Aviris instrument over Moffett Field (CA), lake, soil & vegetation.
  - Hyspex/Madonna instrument over Villelongue (FR), forested area.
- $K = 3$ ($\sim$ ground truth)
- $\beta \in [-1, 3]$
- Evaluation using the average spectral angle mapper (aSAM) :

$$L(\beta) = \mathrm{aSAM}(\mathbf{V}, \hat{\mathbf{V}}) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{acos}\left(\frac{\langle \mathbf{v}_n, \hat{\mathbf{v}}_n \rangle}{\|\mathbf{v}_n\| \|\hat{\mathbf{v}}_n\|} \right)$$

Selecting β with matrix completion
(Févotte and Dobigeon, 2015)



Estimated value $\hat{\beta} \approx 1.5$ for these datasets (compromise between Poisson and additive Gaussian noise).

Outline

Regularized NMF

- Induce prior information or desired structure on **H** (or **W**) using penalty terms :

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \alpha S(\mathbf{H})$$

- MM algorithms are easily adapted to that setting :

$$D(\mathbf{V}|\mathbf{WH}) \qquad \leq G(\mathbf{H}|\tilde{\mathbf{H}}, \mathbf{W})$$

- Only the minimization step is changed.
- May however become intractable ; sometimes $S(\mathbf{H})$ needs to be majorized itself.
- Similar to adjusting the proximal operator in proximal gradient descent.

Regularized NMF

- Induce prior information or desired structure on $\mathbf{H}$ (or $\mathbf{W}$) using penalty terms :

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{H}) + \alpha S(\mathbf{H})$$

- MM algorithms are easily adapted to that setting :

$$D(\mathbf{V}|\mathbf{W}\mathbf{H}) + \alpha S(\mathbf{H}) \leq G(\mathbf{H}|\tilde{\mathbf{H}}, \mathbf{W}) + \alpha S(\mathbf{H})$$

- Only the minimization step is changed.
- May however become intractable ; sometimes $S(\mathbf{H})$ needs to be majorized itself.
- Similar to adjusting the proximal operator in proximal gradient descent.

Sparse NMF

Goal : promote zeros in $\mathbf{H}$ (or $\mathbf{W}$)

$$\min_{\mathbf{W},\mathbf{H}\geq 0} C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{H}) + \alpha S(\mathbf{H})$$

- Exemple : ℓ_1 norm

$$S(\mathbf{H}) = \|\mathbf{H}\|_1 = \sum_{kn} h_{kn}$$

- Exemple : log-sparsity

$$S(\mathbf{H}) = \sum_{kn} \log(h_{kn} + \epsilon)$$

- Or terms that induce a group structure, e.g., cancel some rows of $\mathbf{H}$.
- Vast literature ! Seminal paper by Hoyer (2004).

Ill-posed problem

- $S(\cdot)$ can be made arbitrary small :

$$C(\mathbf{W}\mathbf{\Lambda}^{-1}, \mathbf{\Lambda}\mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{H}) + S(\mathbf{\Lambda}\mathbf{H})$$

- Need to control $\|\mathbf{W}\|$ to avoid degenerate solutions $\|\mathbf{W}\| \to \infty$, $\|\mathbf{H}\| \to 0$.

Sparse NMF

Remedy 1 : penalized optimization

$$\min_{\mathbf{W},\mathbf{H}\geq 0} C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \alpha S(\mathbf{H}) + \delta\|\mathbf{W}\|$$

- Gentle optimization problem.
- Need to tune an extra parameter δ.

Remedy 2 : constrained optimization

$$\min_{\mathbf{W},\mathbf{H}\geq 0} C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \alpha S(\mathbf{H}) \quad \text{subject to} \quad \forall k, \|\mathbf{w}_k\| = 1$$

- Harder optimization problem.
- More natural in a dictionary learning perspective.

Sparse NMF with unit ℓ_1-norm dictionary constraint

Optimization problem

$$\min_{\mathbf{W},\mathbf{H}\geq 0} C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \alpha S(\mathbf{H}) \quad \text{subject to} \quad \forall k, \|\mathbf{w}_k\|_1 = 1$$

1) Lagragian method (Leplat, Gillis, and Idier, 2021)

Search for saddle points of

$$L(\mathbf{W},\mathbf{H},\boldsymbol{\nu}) = D(\mathbf{V}|\mathbf{WH}) + \alpha S(\mathbf{H}) + \sum_k \nu_k(\|\mathbf{w}_k\|_1 - 1)$$

- $\boldsymbol{\nu} \in \mathbb{R}^K$ is the vector of Lagrangian multipliers. $S(\mathbf{H}) = \|\mathbf{H}\|_1$.
- MM-based block-coordinate algorithm that updates $\mathbf{W}, \mathbf{H}$ given $\boldsymbol{\nu}$.
- Only applies to $\beta \leq 1$ or $\beta \in \{\frac{5}{4}, \frac{4}{3}, \frac{3}{2}, 2\}$.
- Update of $\boldsymbol{\nu}$ given $\mathbf{W}$, $\mathbf{H}$ requires a Newton-Raphson procedure.
- Conceptually well-grounded but limited scope.

Sparse NMF with fixed-norm dictionary constraint

2) Heuristic method (Eggert and Körner, 2004; Le Roux et al., 2015)

Unconstrained optimization using reparametrization :

$$\mathbf{W} \leftarrow \mathbf{W}\mathbf{\Lambda}^{-1} \quad \text{with} \quad \lambda_k = \|\mathbf{w}_k\|_1$$

- Minimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{\Lambda}^{-1}\mathbf{H}) + \alpha S(\mathbf{H})$.
- Heuristic multiplicative algorithm using gradient splitting.
- No convergence guarantees (not even monotonicity of the objective function).

3) Block-descent MM method (Marmin, Goulart, and Févotte, 2023b)

Unconstrained optimization of

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{H}) + \alpha S(\mathbf{\Lambda}\mathbf{H})$$

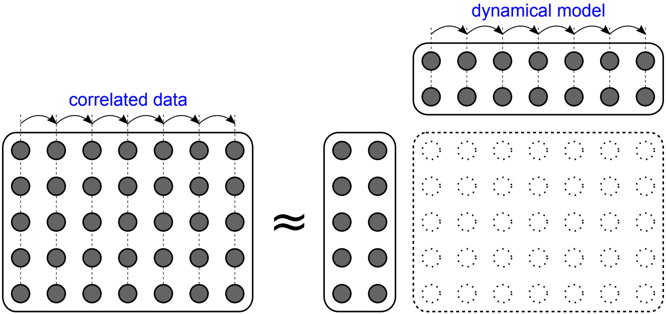
- Shown equivalent to the original problem (after renormalization of the solution).
- Convergent multiplicative MM algorithm for all $\beta \in \mathbb{R}$ ☺
- $S(\mathbf{H}) = \ell_1$ or log-sparsity ☺

Smooth NMF

Impose temporal or spatial regularization, e.g.,

$$S(\mathbf{H}) = \sum_{kn} d(h_{kn}|h_{k(n-1)})$$

- Least squares penalization (Virtanen, 2007; Essid and Févotte, 2013)
- Gamma Markov chains (Smaragdis et al., 2014; Filstroff et al., 2021)

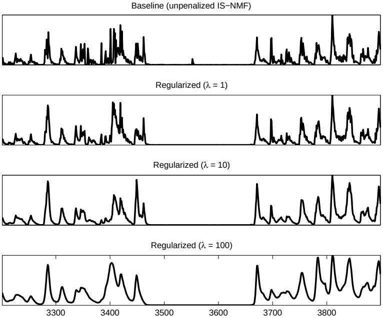


Smooth NMF

Impose temporal or spatial regularization, e.g.,

$$S(\mathbf{H}) = \sum_{kn} d(h_{kn} | h_{k(n-1)})$$

- Least squares penalization (Virtanen, 2007; Essid and Févotte, 2013)
- Gamma Markov chains (Smaragdis et al., 2014; Filstroff et al., 2021)



One row of **H** with increasing smoothness (Févotte, 2011)

Other common regularizers

- Orthogonal NMF : $\mathbf{H}\mathbf{H}^T = \mathbf{I}$.
  Essentially nonnegative clustering (Ding et al., 2006).
- Projective NMF : $\mathbf{H} = \mathbf{W}^T\mathbf{V}$.
  Essentially nonnegative PCA (Yang and Oja, 2010).
- Symmetric NMF : $\mathbf{H} = \mathbf{W}^T$.
  Popular in graph clustering (Kuang et al., 2012; Huang et al., 2013).
- Separable NMF : $\mathbf{W}$ is a subset of columns of $\mathbf{V}$.
  Very active research topic ! (Donoho and Stodden, 2004; Gillis and Vavasis, 2014; Arora et al., 2016).
- Archetypal NMF : $\mathbf{W}$ belongs to the column-range of $\mathbf{V}$.
  A relaxation of separable NMF (Ding et al., 2010; Chen et al., 2014).
- Minimum-volume NMF : penalize the aperture of $\mathbf{W}$.
  Very active research topic ! (Miao and Qi, 2007; Chan et al., 2009) (Leplat, Gillis, and Ang, 2020)

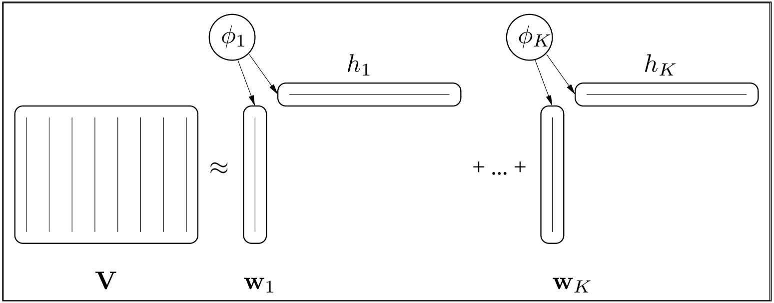
An excellent reference is Nicolas Gillis' book on NMF.

Outline

Automatic relevance determination in NMF

(Tan and Févotte, 2013)

- Another way to select K, inspired by Bayesian PCA (Bishop, 1999).
- Tie each column $\mathbf{w}_k$ and row $\underline{\mathbf{h}}_k$ with a common scale parameter ϕ_k.
- Probabilistic setting with priors $p(\mathbf{w}_k|\phi_k)$ and $p(\underline{\mathbf{h}}_k|\phi_k)$.



- Estimate $\mathbf{W}$ and $\mathbf{H}$ together with the scale parameters ϕ.
- Some scale parameters converge to 0 and the components are pruned.

Automatic relevance determination in NMF

(Tan and Févotte, 2013)

Statistical model

- Observation model : $\mathbf{V} \sim \prod_{fn} \text{Tweedie}(v_{fn}|[\mathbf{WH}]_{fn}, \sigma^2, \beta)$
- half-normal or exponential priors : $\mathbf{w}_k \sim p(\mathbf{w}_k|\phi_k)$ and $\underline{\mathbf{h}}_k \sim p(\underline{\mathbf{h}}_k|\phi_k)$
- inverse-Gamma prior : $\phi_k \sim \text{IG}(\phi_k|a, b)$

Maximum a posteriori estimation

- Boils down to minimizing (using closed-form solution of ϕ_k)

$$C(\mathbf{W}, \mathbf{H}) = D_\beta(\mathbf{V}|\mathbf{WH}) + \lambda \sum_{k=1}^{K} \log\left(\|\mathbf{w}_k\| + \|h_k\| + b\right)$$

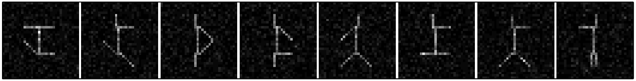
  - $\|\mathbf{x}\| = \frac{1}{2}\|\mathbf{x}\|_2^2$ or $\|\mathbf{x}\|_1$
  - λ is a weight parameter that depends on a and σ^2
  - b acts as a sparsity shape parameter

- Concave term $\log(x + b)$ induces group-sparsity at the column & row level.
- Block-descent multiplicative MM algorithm.
- Follow-up study with more general regularizations by (Cohen and Leplat, 2024).

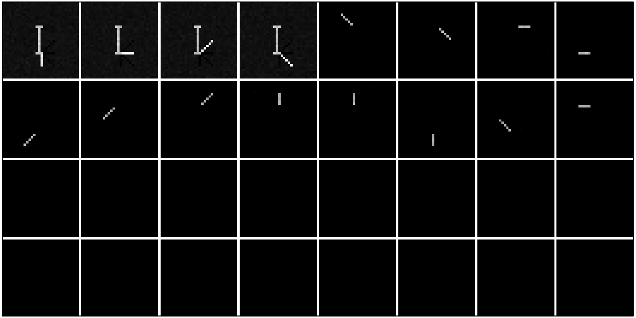
Automatic relevance determination in NMF

Swimmer data decomposition

(a) Noisy data



(b) ℓ_1-ARD decomposition wih $K = 32$



Outline

Robust NMF for nonlinear hyperspectral unmixing

(Févotte and Dobigeon, 2015)

- Variants of the linear mixing model account for "non-linear" effects :

$$\mathbf{v}_n \approx \mathbf{W}\mathbf{h}_n + \mathbf{r}_n$$

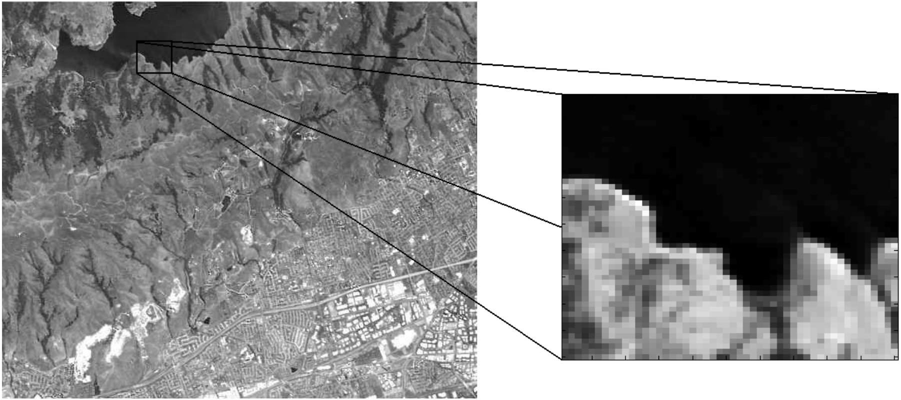
- Often, $\mathbf{r}_n$ has a parametric form such as linear combination of quadratic components $\{\mathbf{w}_k \odot \mathbf{w}_j\}_{kj}$ (Nascimento and Bioucas-Dias, 2009; Fan et al., 2009)
- Nonlinear effects usually affect few pixels only.
- We treat them as non-parametric sparse outliers.

$$\min_{\mathbf{W},\mathbf{H},\mathbf{R}\geq 0} D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H} + \mathbf{R}) + \lambda\|\mathbf{R}\|_{2,1}$$

  where $\|\mathbf{R}\|_{2,1} = \sum_{n=1}^{N} \|\mathbf{r}_n\|_2$ induces sparsity at group level.
- A form of robust NMF (Candès et al., 2009)
- Block descent MM-based algorithm.

Robust NMF for nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Moffett Field data



reproduced from (Dobigeon, 2007)

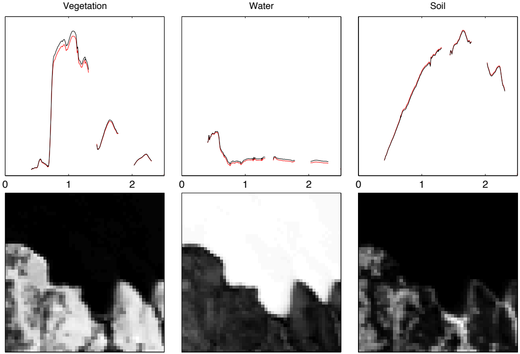
Robust NMF for nonlinear hyperspectral unmixing

(Févotte and Dobigeon, 2015)

Unmixing results

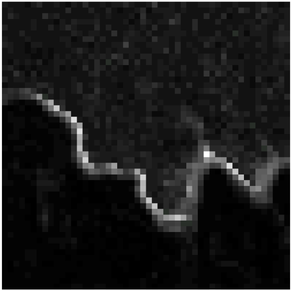
spectral endmembers & activation maps

(**red** : $\beta = 1$, **black** : $\beta = 2$)



outlier energy $\{\|\mathbf{r}_n\|\}_n$

($\beta = 1$)

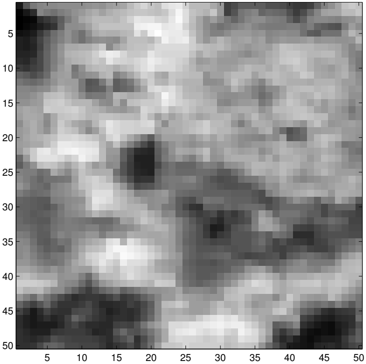


Outlier term captures specific water/soil interactions.

Robust NMF for nonlinear hyperspectral unmixing

(Févotte and Dobigeon, 2015)

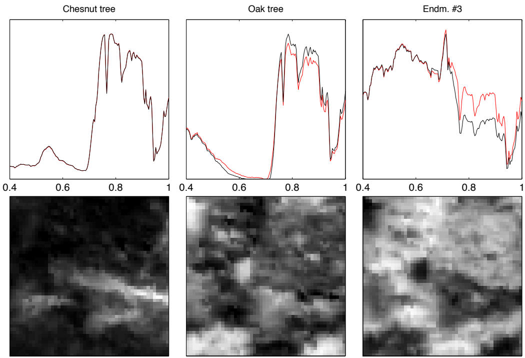
Villelongue/Madonna data (forested area)



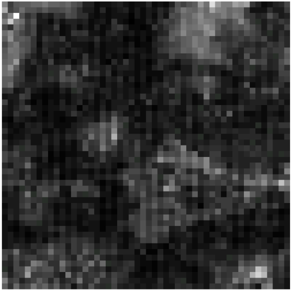
Robust NMF for nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Unmixing results

spectral endmembers & activation maps
(**red** : $\beta = 1$, **black** : $\beta = 2$)



outlier energy $\{\|\mathbf{r}_n\|\}_n$
($\beta = 1$)

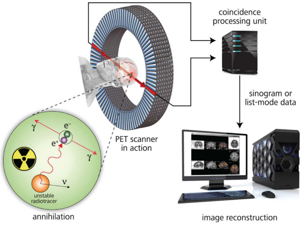


Outlier term seems to capture patterns due to sensor miscalibration.

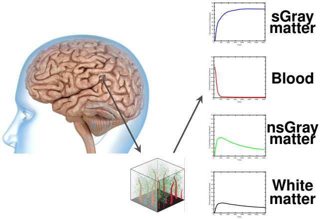
Factor analysis in dynamical PET

(Cavalcanti, Oberlin, Dobigeon, Févotte, Stute, Ribeiro, and Tauber, 2019)

- 3D functional imaging
- Observe the temporal evolution of the brain activity after injecting a radiotracer (biomarker of a specific compound).
- $\mathbf{v}_n$ is the time-activity curve (TAC) in voxel n.
- Neuroimaging : mixed contributions of 4 TAC signatures in each voxel.



Dynamic positron emission tomography



PET voxel decomposition

reproduced from (Cavalcanti, 2018)

Factor analysis in dynamical PET

(Cavalcanti, Oberlin, Dobigeon, Févotte, Stute, Ribeiro, and Tauber, 2019)

Mixing model

- the specific-binding TAC signature varies in space :

$$
\begin{aligned}
\mathbf{v}_n &\approx [\mathbf{w}_1 + \boldsymbol{\delta}_n] h_{1n} + \sum_{k=2}^{K} \mathbf{w}_k h_{kn} \\
&\approx [\mathbf{w}_1 + \mathbf{D}\mathbf{b}_n] h_{1n} + \sum_{k=2}^{K} \mathbf{w}_k h_{kn} \\
&\approx \mathbf{W}\mathbf{h}_n + h_{1n}\,\mathbf{D}\mathbf{b}_n
\end{aligned}
$$

- $\mathbf{D}$ is fixed and pre-trained using labeled or simulated data.

Estimation

$$
\min_{\mathbf{W},\mathbf{H},\mathbf{B}\geq 0} D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H} + \mathbf{1}\,\underline{\mathbf{h}}_1 \odot \mathbf{D}\mathbf{B}) + \lambda \|\mathbf{B}\|_{2,1}
$$

- Optimized with majorization-minimization.

Unmixing results

- real dynamic PET image of a stroke subject injected with a tracer for neuroinflammation.
- MRI ground-truth region of the stroke.

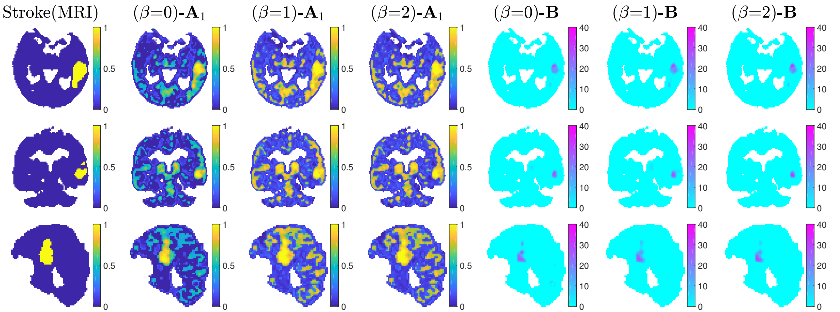


Fig. : Specific-binding activation (h_{1n}) and variability maps ($\|\mathbf{b}_n\|_{2,1}$) in three different planes and for three values of β

Conclusions

- NMF has become a popular data processing tool over the last 25 years.
- Well suited to unmixing problems in unsupervised settings.
- Exciting non-convex optimization problem with non-Euclidean measures of fit.
- MM is a versatile algorithmic framework for NMF :
  - Simple multiplicative algorithms for the β-divergence and beyond.
  - Can be adapted to regularized NMF and variants.
  - More efficient algorithms exist for the quadratic loss.

Funding acknowledgement : *European Research Council, Agence Nationale de la Recherche France, National Research Foundation Singapore*.

References I

S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45(4) :1582–1611, 2016.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3) :549–559, Sep. 1998.

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1) :155–173, Sep. 2007.

J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2) :354–379, 2012.

C. M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1) :459–494, 2014.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis ? *Journal of ACM*, 58(1) : 1–37, 2009.

Y. Cao, P. P. B. Eggermont, and S. Terebey. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2) :286–292, Feb. 1999. doi : 10.1109/83.743861.

Y. C. Cavalcanti. *Factor analysis of dynamic PET images*. PhD thesis, Toulouse INP, 2018.

References II

Y. C. Cavalcanti, T. Oberlin, N. Dobigeon, C. Févotte, S. Stute, M. Ribeiro, and C. Tauber. Factor analysis of dynamic PET images : Beyond Gaussian noise. *IEEE Transactions on Medical Imaging*, 38(9) :2231–2241, Sep. 2019. ISSN 0278-0062. doi : 10.1109/TMI.2019.2906828. URL `https://arxiv.org/pdf/1807.11455`.

T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11) :4418–4432, 2009.

Y. Chen, J. Mairal, and Z. Harchaoui. Fast and robust archetypal analysis for representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

A. Cichocki and S. Amari. Families of Alpha- Beta- and Gamma- divergences : Flexible and robust measures of similarities. *Entropy*, 12(6) :1532–1568, June 2010.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. In *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Charleston SC, USA, Mar. 2006.

A. Cichocki, R. Zdunek, and S.-i. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3d tensor factorization. In *International Conference on Independent Component Analysis and Signal Separation*, 2007.

A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9) :1433–1440, July 2008.

A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorizations : Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

J. E. Cohen and V. Leplat. Efficient algorithms for regularized nonnegative scale-invariant low-rank approximation models. *arXiv :2403.18517*, 2024.

References III

M. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, 5(5) :61 – 66, 1986. doi : 10.1109/TMI.1986.4307748.

A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2) :328–333, 1993. doi : 10.1109/42.232263.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proc. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 126–135. ACM, 2006.

C. H. Q. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1) :45 – 55, 2010. doi : 10.1109/TPAMI.2008.277.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts ? In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

J. Eggert and E. Körner. Sparse coding and NMF. In *Proc. IEEE International Joint Conference on Neural Networks*, pages 2529–2533, 2004.

S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2) :415–425, Feb. 2013. doi : 10.1109/TMM.2012.2228474. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/ieee_multimedia_smoothnmf.pdf.

References IV

W. Fan, B. Hu, J. Miller, and M. Li. Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data. *International Journal of Remote Sensing*, 30(11) :2951–2962, June 2009.

C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp11a.pdf.

C. Févotte and N. Dobigeon. Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 24(12) :4810–4819, Dec. 2015. doi : 10.1109/TIP.2015.2468177. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/tip2015.pdf.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9) :2421–2456, Sep. 2011. doi : 10.1162/NECO_a_00168. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco11.pdf.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3) :793–830, Mar. 2009. doi : 10.1162/neco.2008.04-08-771. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco09_is-nmf.pdf.

L. Filstroff, O. Gouvert, C. Févotte, and O. Cappé. A comparative study of Gamma Markov chains for temporal non-negative factorization. *IEEE Transactions on Signal Processing*, 69 :1614–1626, 2021. doi : 10.1109/TSP.2021.3060000. URL https://arxiv.org/pdf/2006.12843.

References V

L. Finesso and P. Spreij. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications*, 416 :270–287, 2006.

N. Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020.

N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4) :1085–1105, 04 2012. ISSN 0899-7667. doi : 10.1162/NECO_a_00256. URL https://doi.org/10.1162/NECO_a_00256.

N. Gillis and S. A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4) :698–714, 2014. doi : 10.1109/TPAMI.2013.226.

N. Guan, D. Tao, Z. Luo, and B. Yuan. NeNMF : An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6) :2882–2898, 2012.

L. T. K. Hien and N. Gillis. Algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. *Journal of Scientific Computing*, 87(3) :93, 2021.

T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999. URL http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf.

P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5 :1457–1469, 2004.

C. J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1064 – 1072, Aug. 2011.

References VI

K. Huang, N. D. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited : Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1) :211–224, 2013.

K. Huang, N. D. Sidiropoulos, and A. P. Liavas. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing*, 64(19) :5052–5065, 2016. doi : 10.1109/TSP.2016.2576427.

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

D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Prco. SIAM International Conference on Data Mining*, pages 106–117, 2012.

J. Le Roux, F. J. Weninger, and J. R. Hershey. Sparse NMF–half-baked or well done ? Technical report, Mitsubishi Electric Research Labs (MERL), 2015.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401 : 788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, Oct. 2011. URL `https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/waspaa11.pdf`.

V. Leplat, N. Gillis, and A. M. Ang. Blind audio source separation with minimum-volume beta-divergence NMF. *IEEE Transactions on Signal Processing*, 68 :3400–3410, 2020.

V. Leplat, N. Gillis, and J. Idier. Multiplicative updates for nmf with β-divergences under disjoint equality constraints. *SIAM Journal on Matrix Analysis and Applications*, 42(2) :730–752, 2021.

C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19 : 2756–2779, 2007.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79 : 745–754, 1974. doi : 10.1086/111605.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2) :829–855, 2015.

A. Marmin, J. H. de M.. Goulart, and C. Févotte. Joint majorization-minimization for nonnegative matrix factorization with the beta-divergence. *Signal Processing*, 209 :109048, 2023a. URL https://arxiv.org/pdf/2106.15214.

A. Marmin, J. H. de M.. Goulart, and C. Févotte. Majorization-minimization for sparse nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Signal Processing*, 71 :1435–1447, 2023b. doi : 10.1109/TSP.2023.3266939. URL https://arxiv.org/pdf/2207.06316.

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1) :113–128, 2018. doi : 10.1109/TSP.2017.2752697.

L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3) :765–777, 2007. ISSN 0196-2892. doi : 10.1109/TGRS.2006.888466.

M. C. Mukkamala and P. Ochs. Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

References VIII

M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2010)*, Sep. 2010.

J. M. P. Nascimento and J. M. Bioucas-Dias. Nonlinear mixture model for hyperspectral unmixing. In *Proc. SPIE Image and Signal Processing for Remote Sensing XV*, 2009.

P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 :111–126, 1994.

A. Rakotomamonjy. Direct optimization of the dictionary learning problem. *IEEE Transactions on Signal Processing*, 61(12) :5495–5506, 2013.

W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62 :55–59, 1972.

P. Smaragdis. About this non-negative business. WASPAA keynote slides, 2013. URL `http://web.engr.illinois.edu/~paris/pubs/smaragdis-waspaa2013keynote.pdf`.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations : A unified view. *IEEE Signal Processing Magazine*, 31(3) : 66–75, May 2014. doi : 10.1109/MSP.2013.2297715. URL `https://www.irit.fr/~Cedric.Fevotte/publications/journals/spm2014.pdf`.

D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014. URL https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp14a.pdf.

N. Takahashi, J. Katayama, M. Seki, and J. Takeuchi. A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization. *Computational Optimization and Applications*, 71(1) : 221–250, 2018.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7) :1592 – 1605, July 2013. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/pami13_ardnmf.pdf.

M. Vandecappelle, N. Vervliet, and L. De Lathauwer. A second-order method for fitting the canonical polyadic decomposition with non-least-squares cost. *IEEE Transactions on Signal Processing*, 68 :4454–4465, 2020.

T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3) :1066–1074, Mar. 2007.

Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5) :734–749, 2010.

Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22 :1878 – 1891, Dec. 2011. doi : http://dx.doi.org/10.1109/TNN.2011.2170094.

References X

R. Zhao and V. Y. F. Tan. A unified convergence analysis of the multiplicative update algorithm for regularized nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 66(1) :129–138, Jan 2018. ISSN 1053-587X. doi : 10.1109/TSP.2017.2757914.

R. Zhao, V. Y. Tan, and H. Xu. Online nonnegative matrix factorization with general divergences. In *Proc. AISTATS*, 2017.