Robust nonnegative matrix factorisation with the β-divergence and applications in imaging

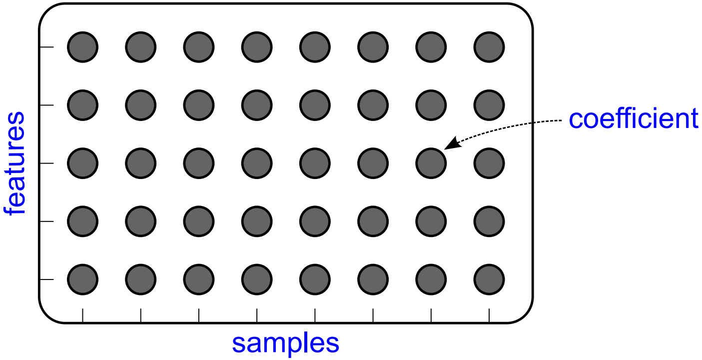
Cédric Févotte
Institut de Recherche en Informatique de Toulouse

Imaging & Machine Learning
Institut Henri Poincaré
April 2019

Outline

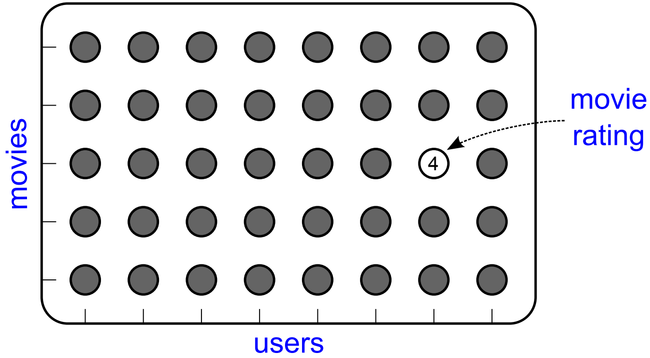
Matrix factorisation models

Data often available in matrix form.



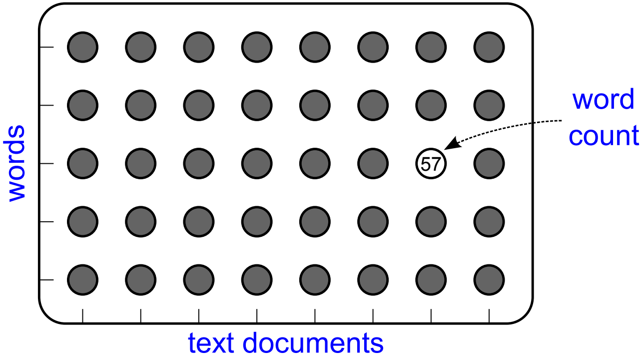
Matrix factorisation models

Data often available in matrix form.



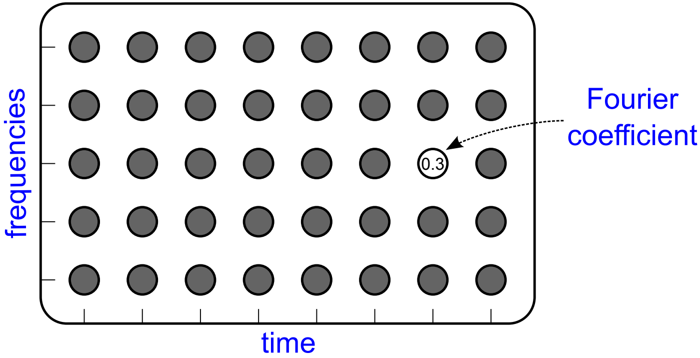
Matrix factorisation models

Data often available in matrix form.



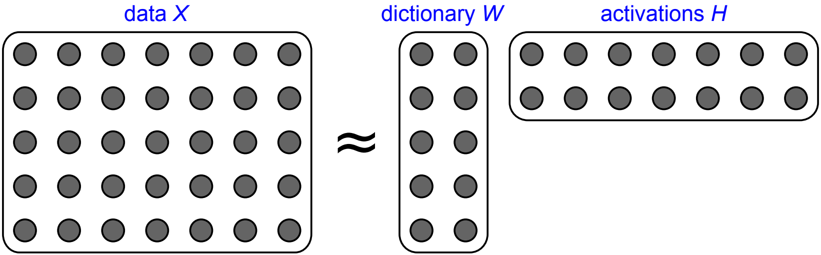
Matrix factorisation models

Data often available in matrix form.



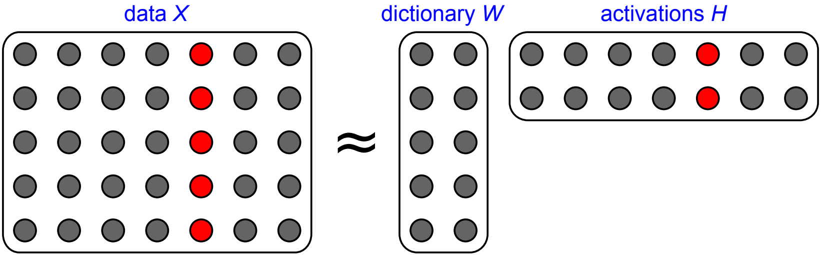
Matrix factorisation models

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



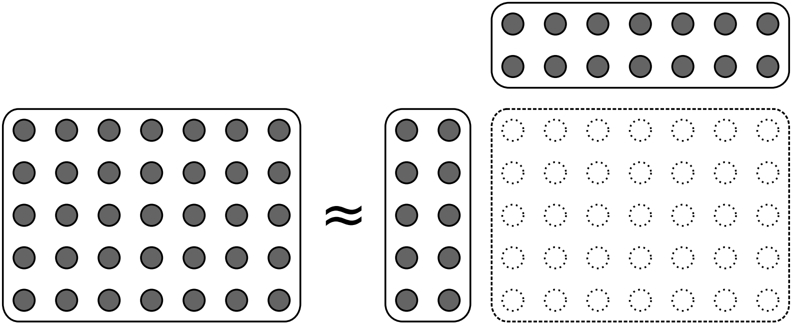
Matrix factorisation models

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



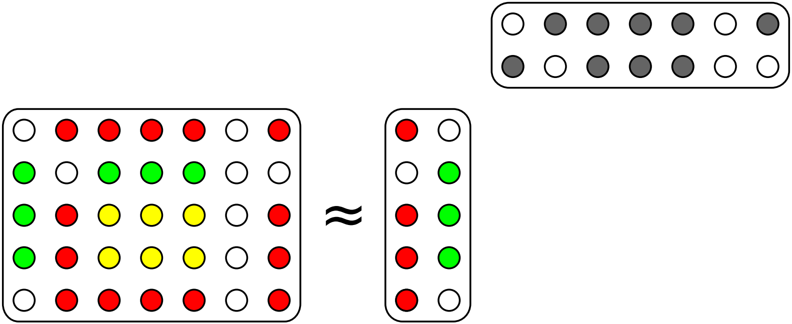
Matrix factorisation models

for dimensionality reduction (coding, low-dimensional embedding)



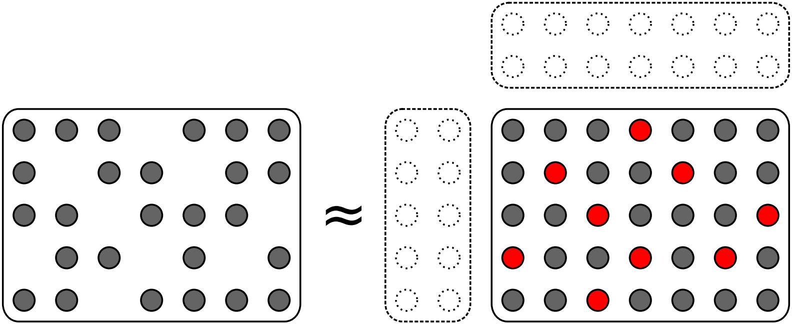
Matrix factorisation models

for unmixing (source separation, latent topic discovery)

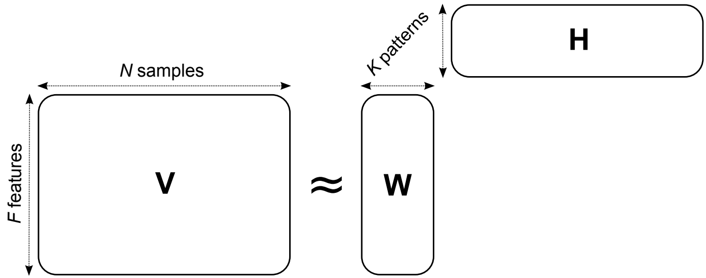


Matrix factorisation models

for interpolation (collaborative filtering, image inpainting)



Nonnegative matrix factorisation

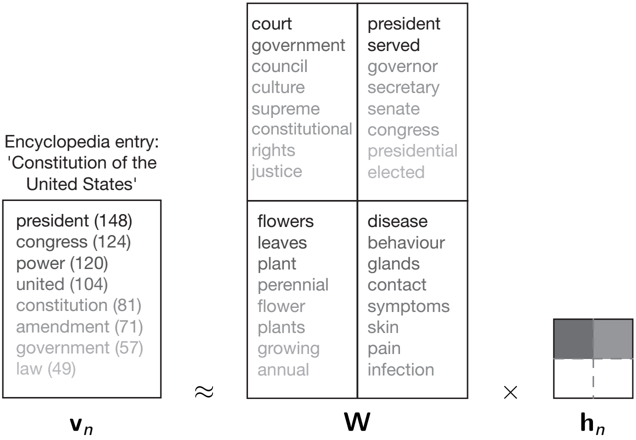


- data $\mathbf{V}$ and factors $\mathbf{W}$, $\mathbf{H}$ have nonnegative entries.
- nonnegativity of $\mathbf{W}$ ensures interpretability of the dictionary, because patterns $\mathbf{w}_k$ and samples $\mathbf{v}_n$ belong to the same space.
- nonnegativity of $\mathbf{H}$ tends to produce part-based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

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

NMF for latent semantic analysis

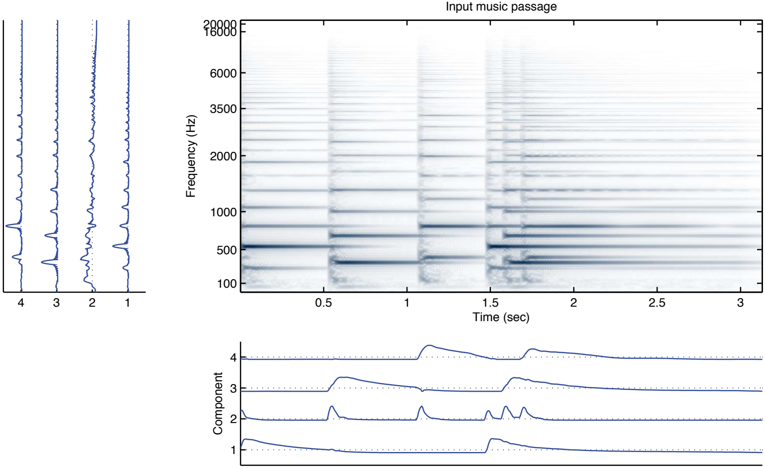
(Lee and Seung, 1999; Hofmann, 1999)



reproduced from (Lee and Seung, 1999)

NMF for audio spectral unmixing

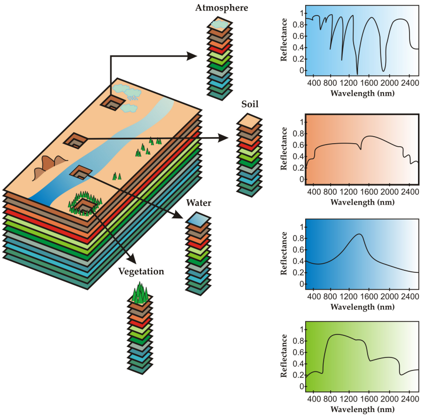
(Smaragdis and Brown, 2003)



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 minimisation problem

Minimise a measure of fit between $\mathbf{V}$ and $\mathbf{WH}$, 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)

Regularisation terms often added to $D(\mathbf{V}|\mathbf{WH})$ for sparsity, smoothness, dynamics, 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	squared Euclidean	many
integer	multinomial*	weighted KL	word counts
integer	Poisson	generalised KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectrogram
generally nonnegative	Tweedie	β-divergence	generalises 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)

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

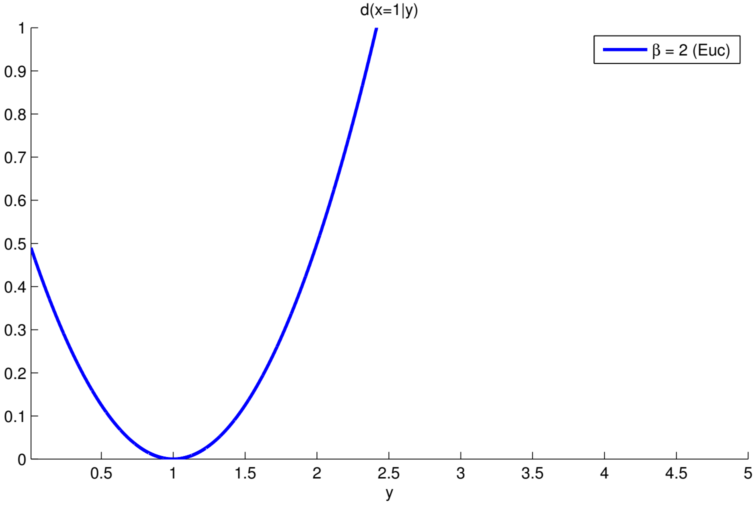
Special cases:

- squared Euclidean distance ($\beta = 2$)
- generalised Kullback-Leibler (KL) divergence ($\beta = 1$)
- Itakura-Saito (IS) divergence ($\beta = 0$)

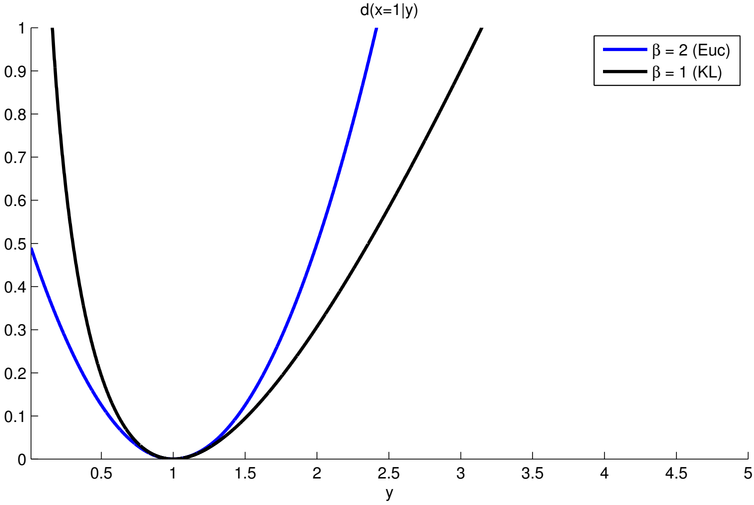
Properties:

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

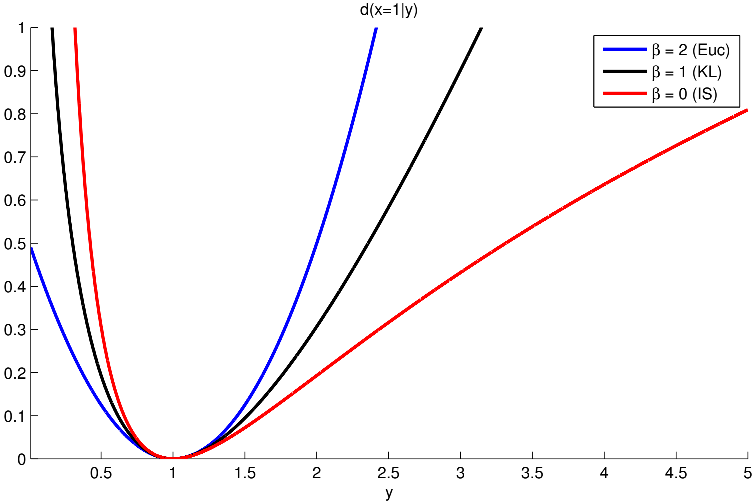
The β-divergence



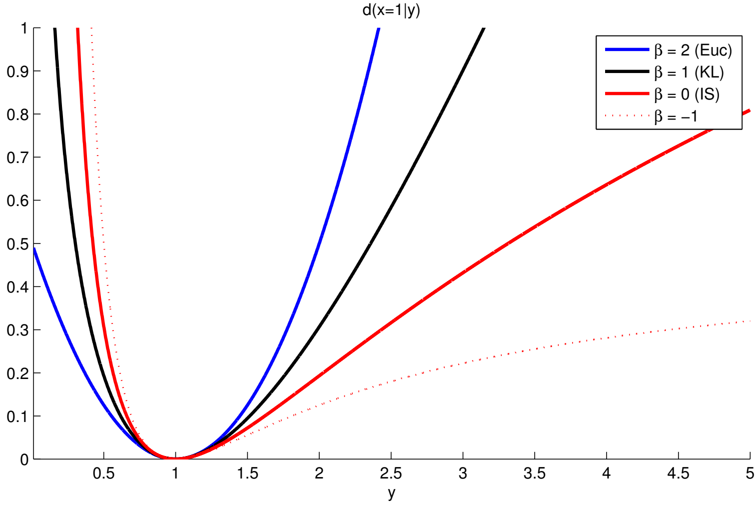
The β-divergence



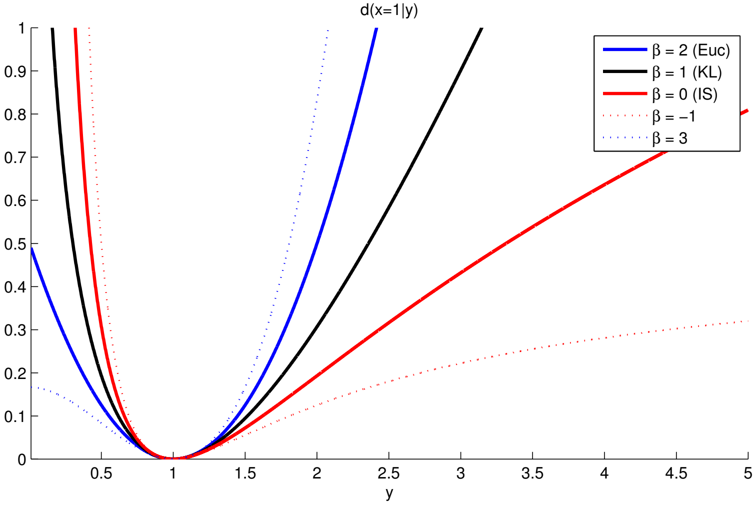
The β-divergence



The β-divergence



The β-divergence



Common NMF algorithm design

- 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:

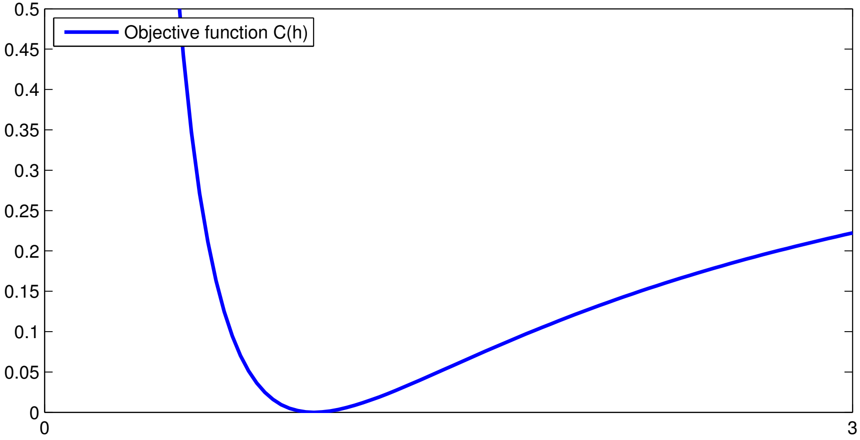
$$\min_{\mathbf{h}\geq\mathbf{0}} C(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h})$$

  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, initialisation is an issue.

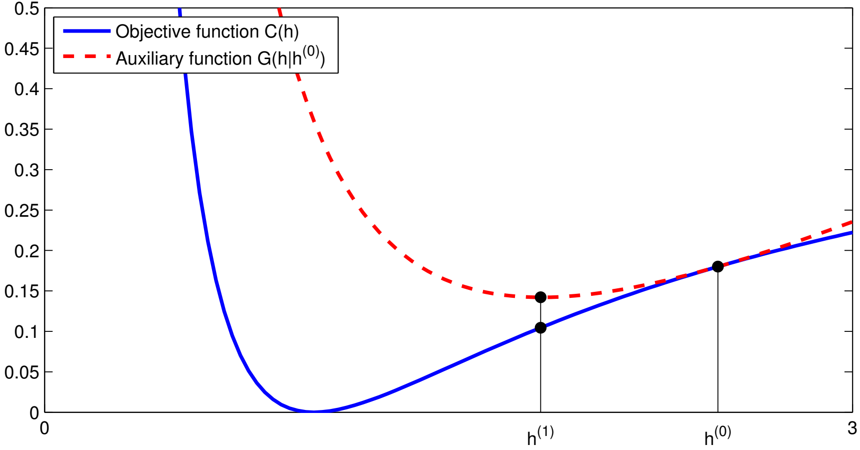
Majorisation-minimisation (MM)

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



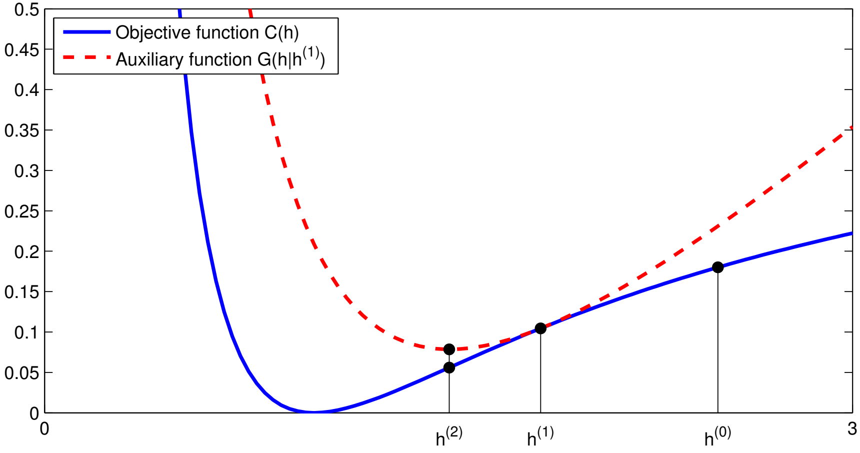
Majorisation-minimisation (MM)

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



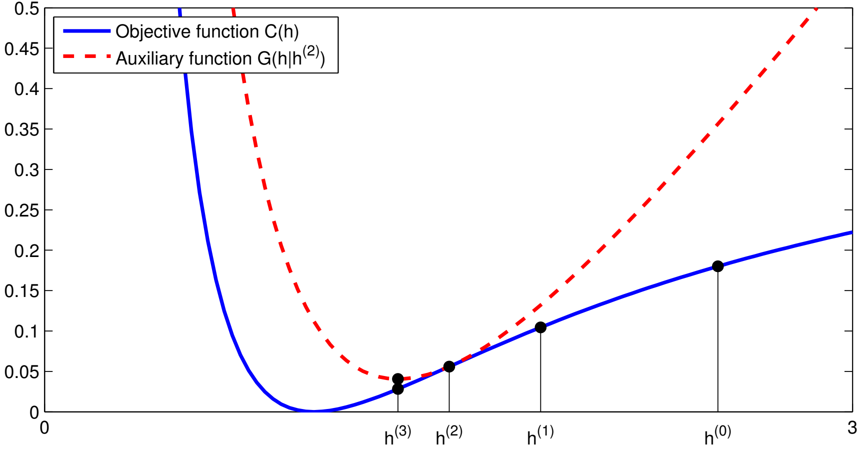
Majorisation-minimisation (MM)

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



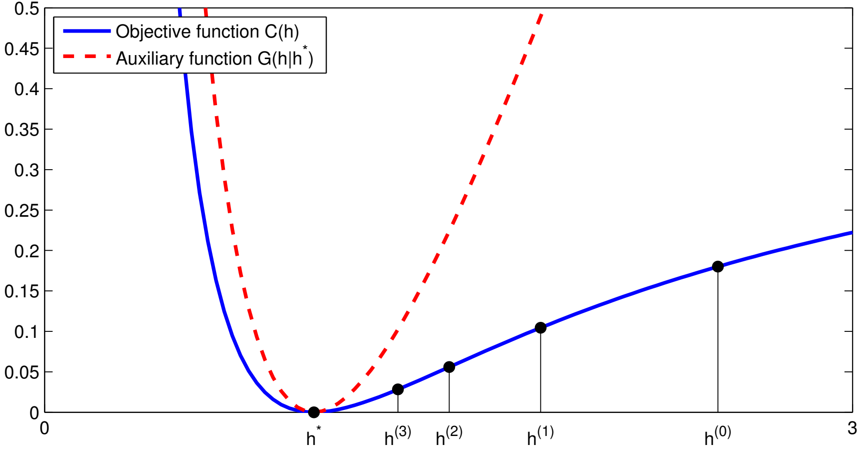
Majorisation-minimisation (MM)

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



Majorisation-minimisation (MM)

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



Majorisation-minimisation (MM)

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

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

Majorisation-minimisation (MM)

- Finding a good & workable local majorisation 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 goes left.
  - γ is a divergence-specific scalar exponent.
- Details in (Févotte and Idier, 2011; Yang and Oja, 2011; Zhao and Tan, 2018)

Outline

Generalities

- Matrix factorisation models
- Nonnegative matrix factorisation (NMF)

Optimisation for NMF

- Measures of fit
- Majorisation-minimisation

Applications in imaging

- Hyperspectral unmixing in remote sensing
- Factor analysis in dynamic PET

Hyperspectral model selection by matrix completion

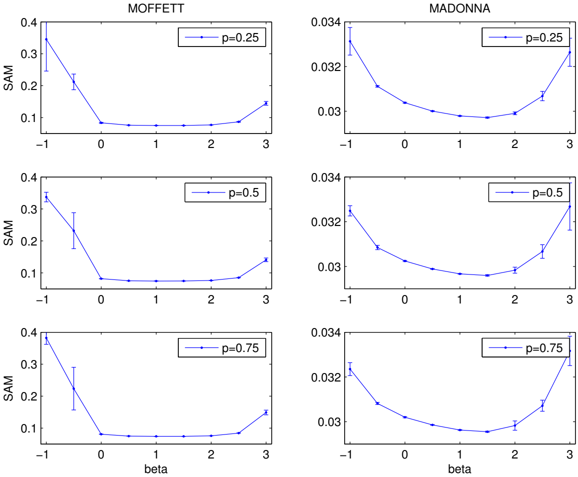
(Févotte and Dobigeon, 2015)

- **Data**: 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.
- a percentage of the pixels is randomly removed.
- $\mathbf{W}$ and $\mathbf{H}$ estimated with $K = 3$ ($\sim$ ground truth) and various values of β.
- missing pixels are reconstructed from $\hat{\mathbf{V}} = \mathbf{WH}$.
- evaluation using the average spectral angle mapper (aSAM):

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

Hyperspectral model selection by matrix completion

(Févotte and Dobigeon, 2015)



Recommended value $\beta \approx 1.5$
(compromise between Poisson and additive Gaussian noise).

Nonlinear hyperspectral unmixing

(Févotte and Dobigeon, 2015)

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

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

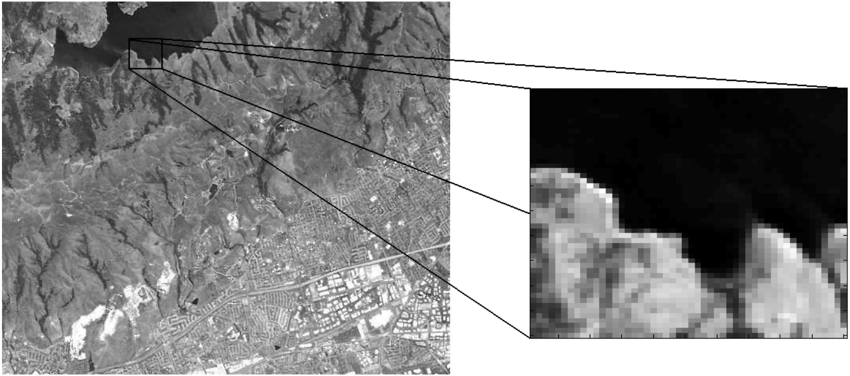
- Often, $\mathbf{r}_n$ has a parametric form, e.g., linear combination of quadratic components $\{\mathbf{w}_k \odot \mathbf{w}_j\}_{kj}$ (Nascimento and Bioucas-Dias, 2009; Fan et al., 2009; Altmann et al., 2012)
- Nonlinear effects usually affect few pixels only.
- We treat them as non-parametric 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.
- Optimised with majorisation-minimisation.

Nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Moffett Field data



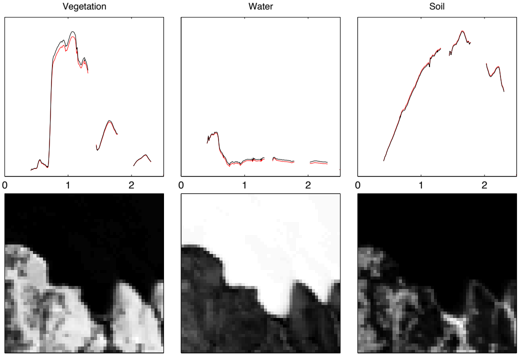
reproduced from (Dobigeon, 2007)

Nonlinear hyperspectral unmixing

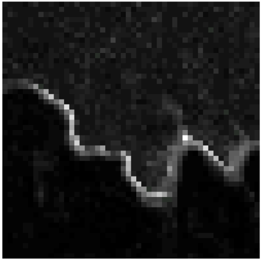
(Févotte and Dobigeon, 2015)

Unmixing results

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



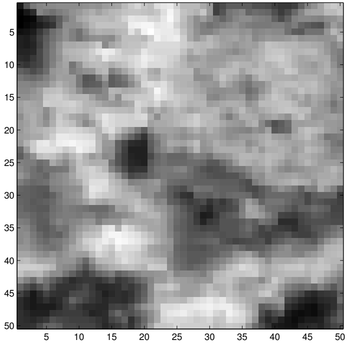
outlier energy $\{\|\mathbf{r}_n\|\}_n$
($\beta = 1$)



Outlier term captures specific water/soil interactions.

Nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

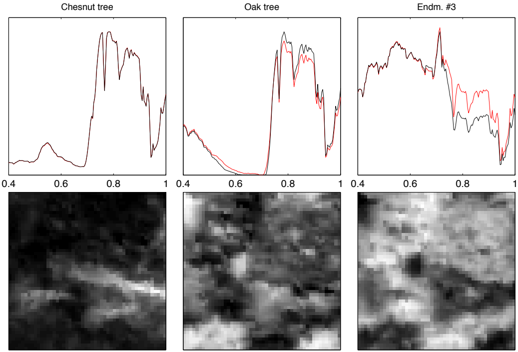
Villelongue/Madonna data (forested area)



Unmixing results

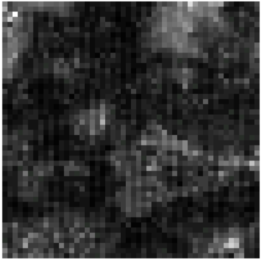
spectral endmembers & activation maps

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



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

$(\beta = 1)$



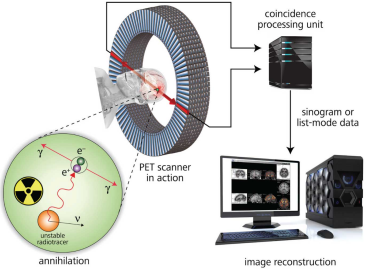
Outlier term seems to capture patterns due to sensor miscalibration.

Outline

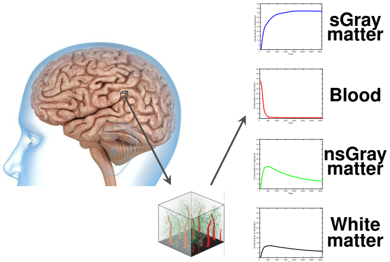
Factor analysis in dynamical PET

(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

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

Optimised with majorisation-minimisation.

Unmixing results

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

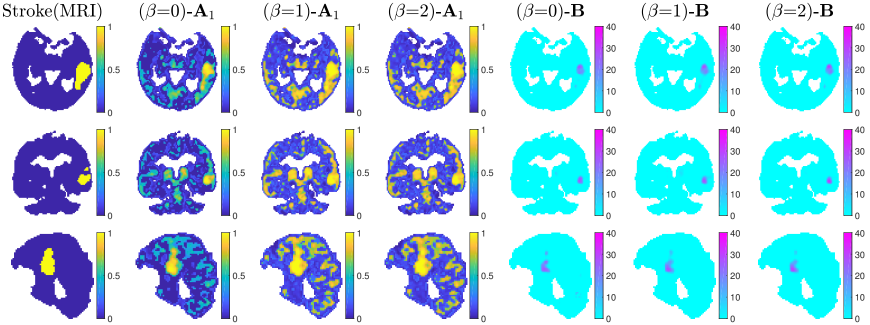


Fig.: Specific-binding activation (h_{1n}) and variability maps ($\|\mathbf{b}_n\|_{2,1}$) in three different planes and for three values of β

Conclusions

- NMF can efficiently unmix composite data in imaging problems.
- Application-specific variants have been proposed.
- The β-divergence can be adjusted to the statistics of the noise.
- Majorisation-minimisation works well in this setting.

ERC-funded **postdoc positions** in machine learning & signal processing:

- Multimodal data processing for multimedia artistic creation (with Tim van Cruys)
- Learning with low-rank models (with Emmanuel Soubies)
- Bayesian deep learning (with Nicolas Dobigeon)

`http://projectfactory.irit.fr/`

Plenary speakers

Yuejie Chi, CMU
Emilie Chouzenoux, Univ. Paris-Est
Mark Davenport, Georgia Tech
Monika Dörfler, Univ. Vienna
Pier Luigi Dragotti, ICL
Bhaskar Rao, UC San Diego
Simon Thorpe, CNRS
Lenka Zdeborova, CNRS

Special talk by **Michael I. Jordan**, UC Berkeley

`http://spars-workshop.org/`

References I

Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret. Supervised nonlinear spectral unmixing using a post-nonlinear mixing model for hyperspectral imagery. *IEEE Transactions on Image Processing*, 21 (6):3017–3025, June 2012.

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.

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

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*, -:–, 2019. ISSN 0278-0062. doi: 10.1109/TMI.2019.2906828.

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

References II

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.

N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero. Nonlinear unmixing of hyperspectral images: Models and algorithms. *IEEE Signal Proccessing Magazine*, 31(1):89–94, Jan. 2014.

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 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.

References III

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

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.

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.

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

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.

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.

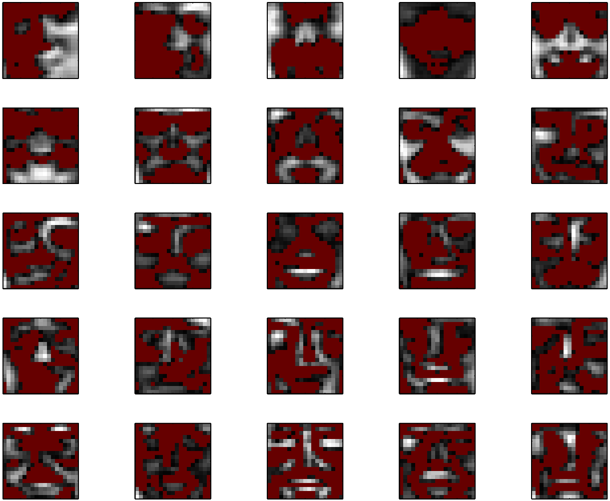
References IV

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.

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.

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)