Nonnegative matrix factorization and applications in audio signal processing

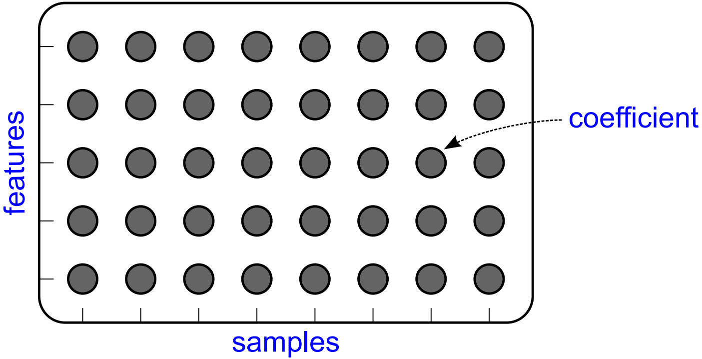
Cédric Févotte
Laboratoire Lagrange, Nice

Machine Learning Crash Course
Genova, June 2015

Outline

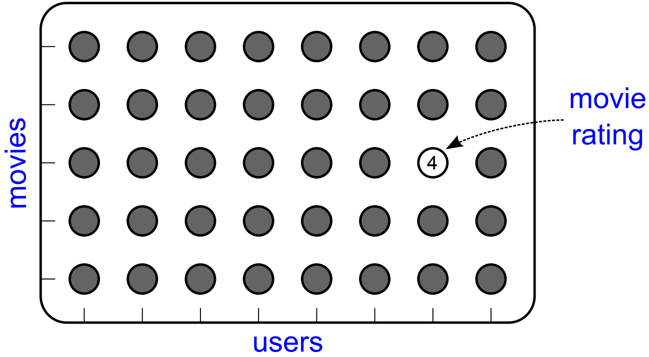
Matrix factorisation models

Data often available in matrix form.



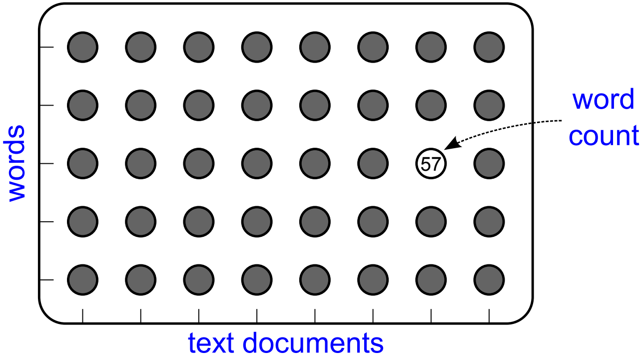
Matrix factorisation models

Data often available in matrix form.



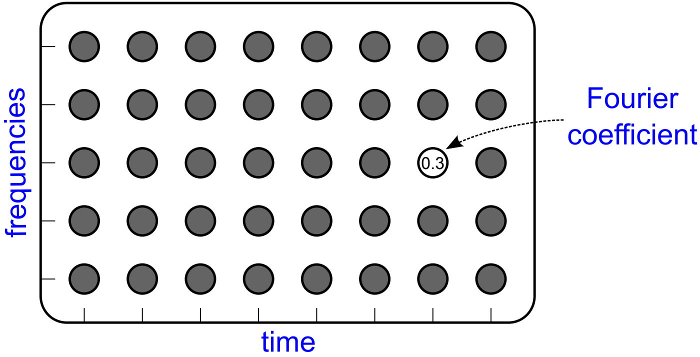
Matrix factorisation models

Data often available in matrix form.



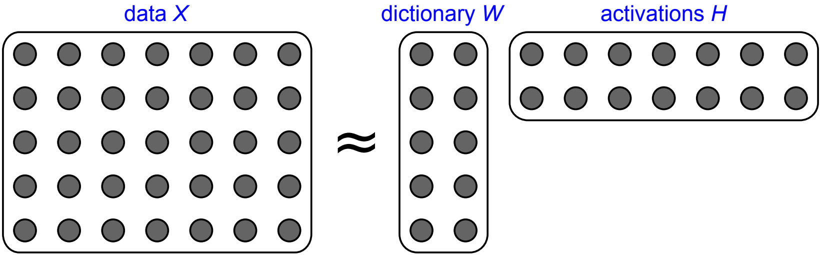
Matrix factorisation models

Data often available in matrix form.



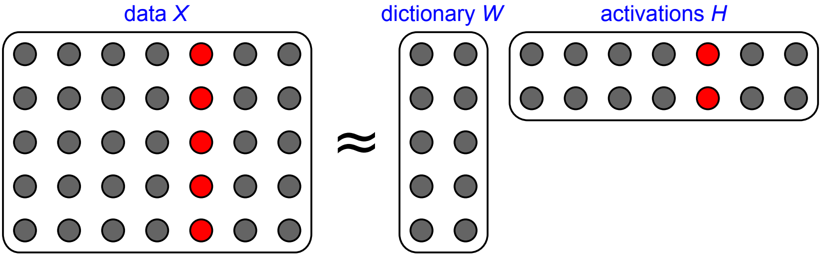
Matrix factorisation models

≈ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



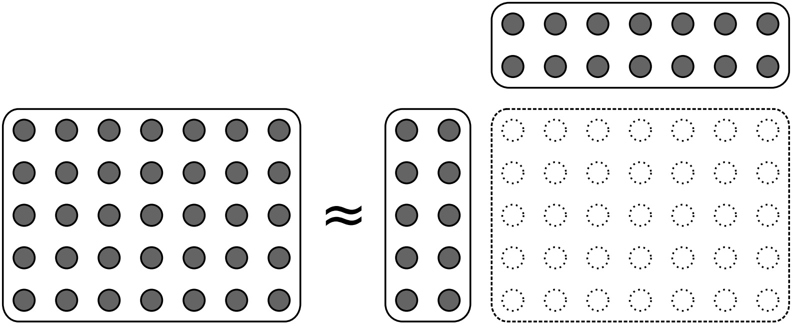
Matrix factorisation models

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



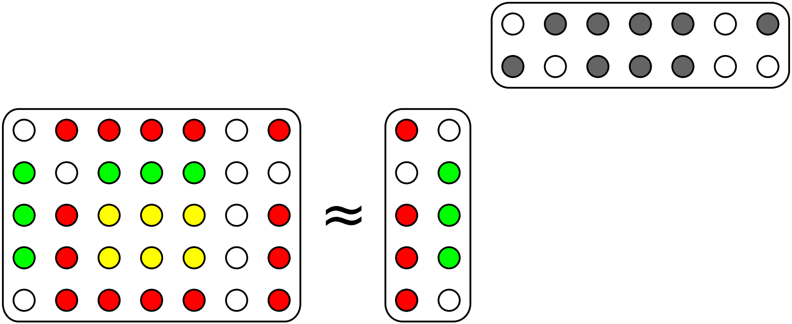
Matrix factorisation models

for dimensionality reduction (coding, low-dimensional embedding)



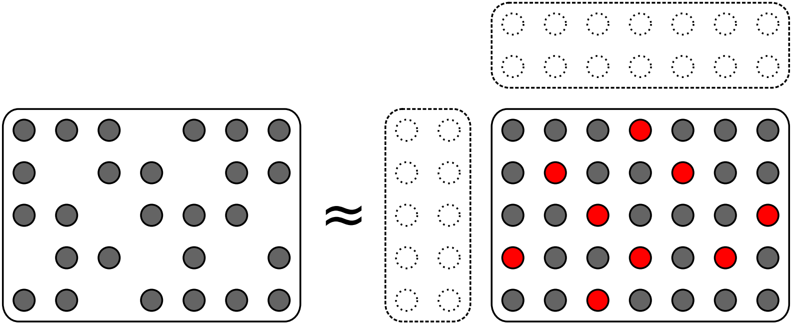
Matrix factorisation models

for unmixing (source separation, latent topic discovery)

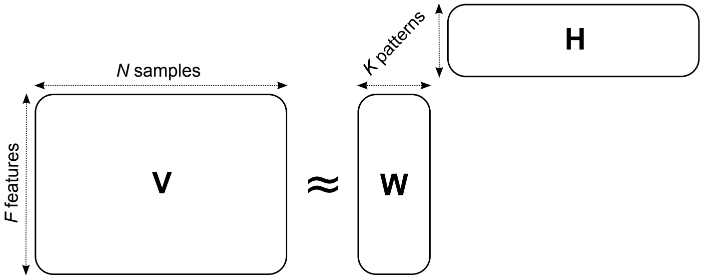


Matrix factorisation models

for interpolation (collaborative filtering, image inpainting)



Nonnegative matrix factorisation

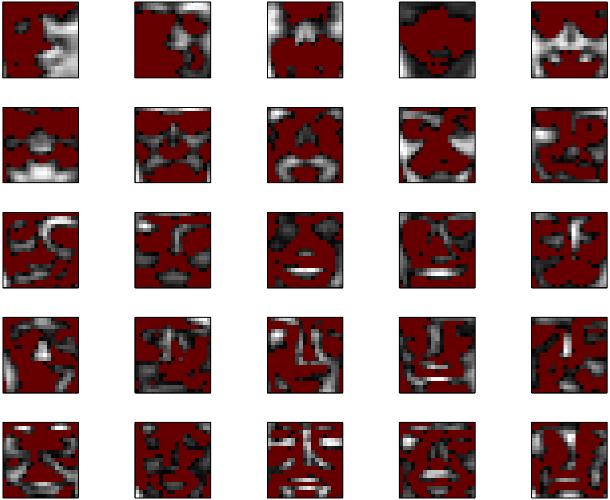


- data **V** and factors **W**, **H** have nonnegative entries.
- nonnegativity of **W** ensures interpretability of the dictionary, because patterns $\mathbf{w}_k$ and samples $\mathbf{v}_n$ belong to the same space.
- nonnegativity of **H** tends to produce part-based representations, because subtractive 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 $K = 25$



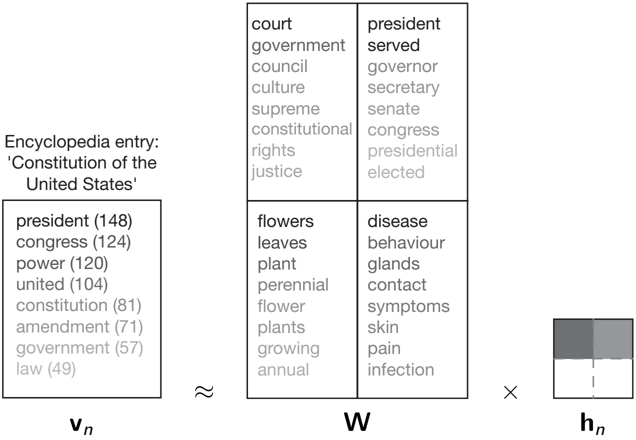
red pixels indicate negative values

NMF dictionary with $K = 25$

experiment reproduced from (Lee and Seung, 1999)

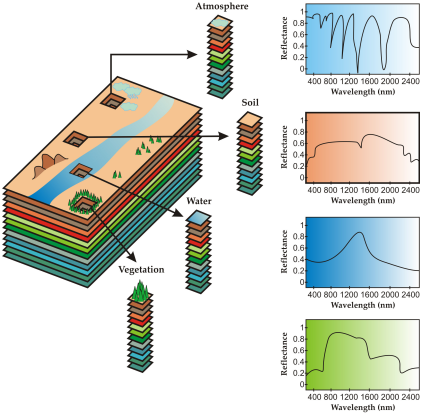
NMF for latent semantic analysis

(Lee and Seung, 1999; Hofmann, 1999)



reproduced from (Lee and Seung, 1999)

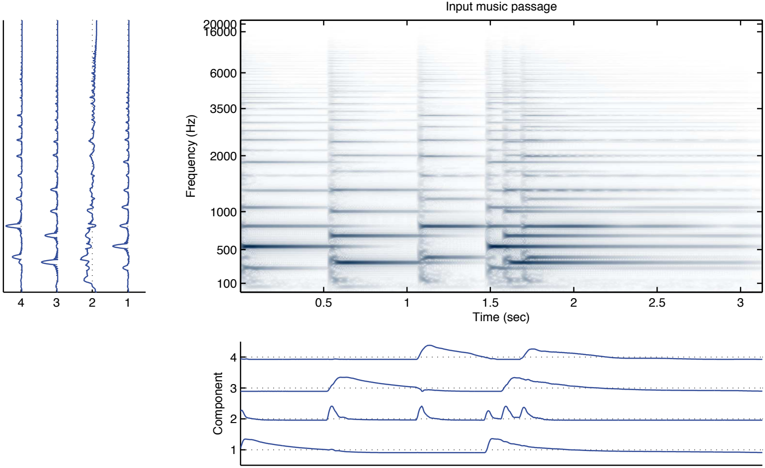
NMF for hyperspectral unmixing
(Berry, Browne, Langville, Pauca, and Plemmons, 2007)



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

NMF for audio spectral unmixing

(Smaragdis and Brown, 2003)



reproduced from (Smaragdis, 2013)

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.

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 $\mathbf{W}$:

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

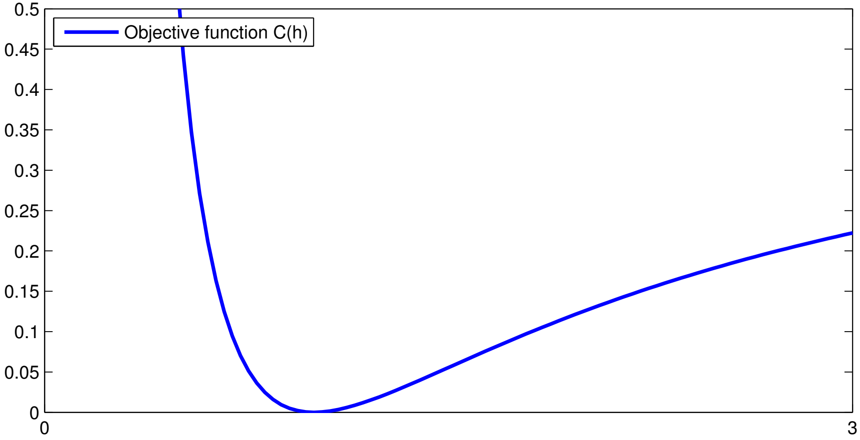
- Essentially left with nonnegative linear regression:

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

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

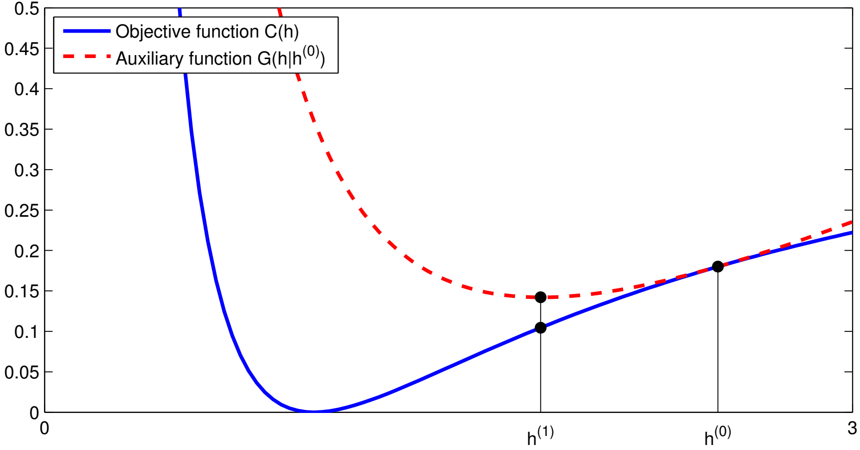
Majorisation-minimisation (MM)

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 $C(\mathbf{h})$.



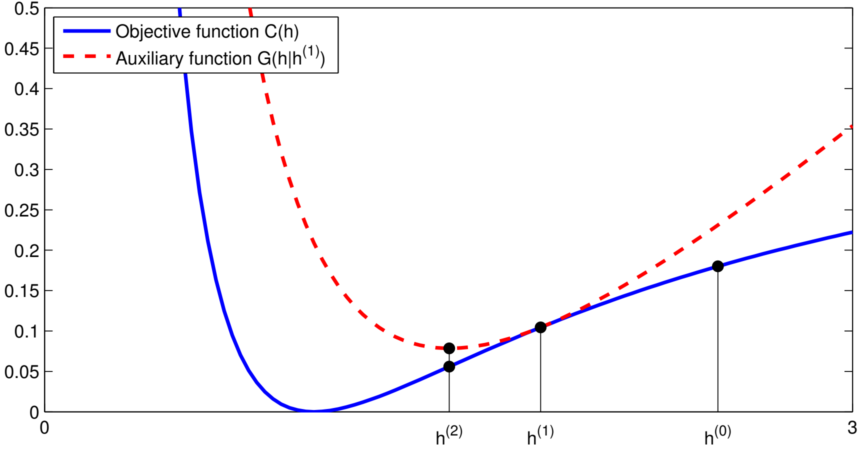
Majorisation-minimisation (MM)

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 $C(\mathbf{h})$.



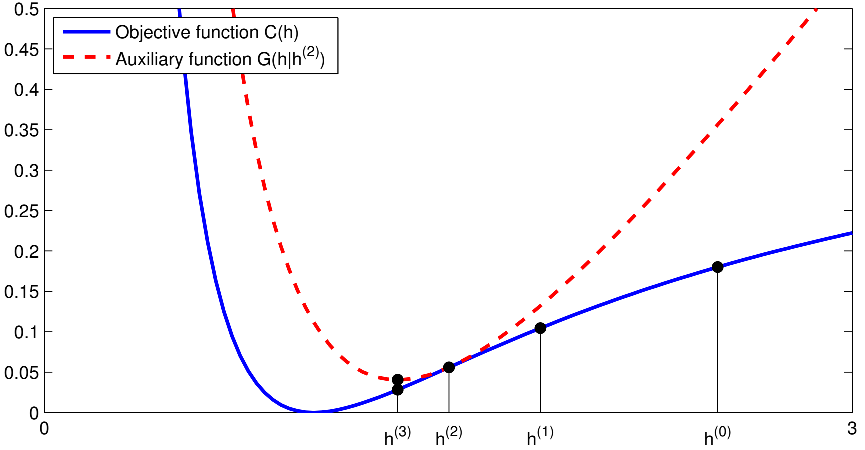
Majorisation-minimisation (MM)

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 $C(\mathbf{h})$.



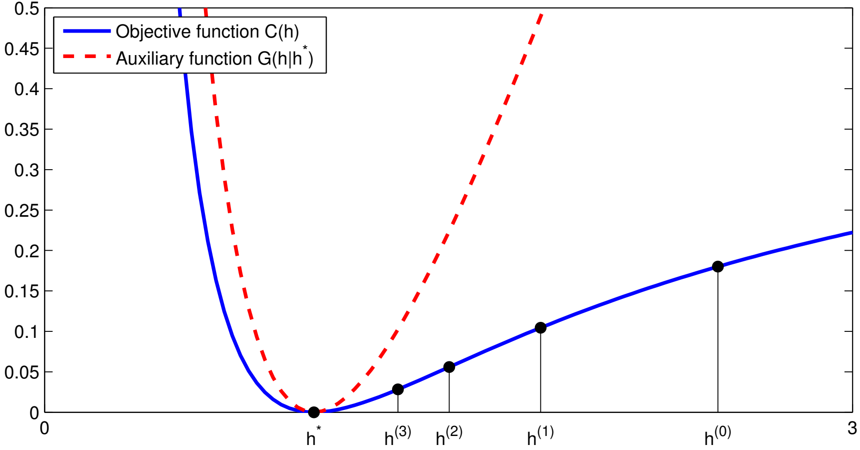
Majorisation-minimisation (MM)

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.
- For most the divergences mentioned, Jensen and tangent inequalities are usually enough.
- In many cases, leads to multiplicative algorithms such that

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

  where
  - $\nabla_{h_k} C(\mathbf{h}) = \nabla^-_{h_k} C(\mathbf{h}) - \nabla^+_{h_k} C(\mathbf{h})$ and the two summands are nonnegative
  - γ is a divergence-specific scalar exponent.
- More details about MM in (Lee and Seung, 2001; Févotte and Idier, 2011; Yang and Oja, 2011).

How to choose a right measure of fit ?

- Squared Euclidean distance is a common default choice.
- Underlies a Gaussian additive noise model such that

$$v_{fn} = [\mathbf{WH}]_{fn} + \epsilon_{fn}.$$

  Can generate negative values – not very natural for nonnegative data.
- Many other options.

Select a right divergence (for a specific problem) by

- comparing performances, given ground-truth data.
- assessing the ability to predict missing/unseen data (interpolation, cross-validation).
- probabilistic modelling:

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

How to choose a right measure of fit ?

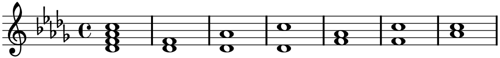
- Let $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$ such that $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$
- then the following correspondences apply with

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

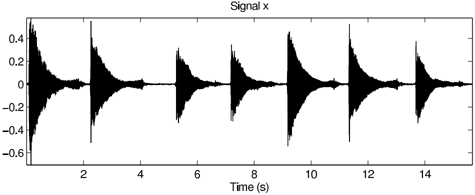
data support	distribution/noise	divergence	examples
real-valued	additive Gaussian	squared Euclidean	many
integer	multinomial	Kullback-Leibler	word counts
integer	Poisson	generalised KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectral data
generally nonnegative	Tweedie	β-divergence	generalises above models

Outline

Piano toy example



(MIDI numbers : 61, 65, 68, 72)



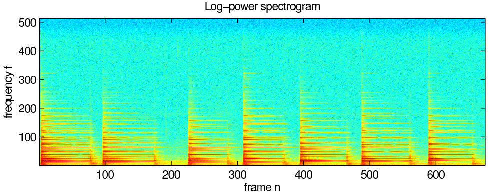
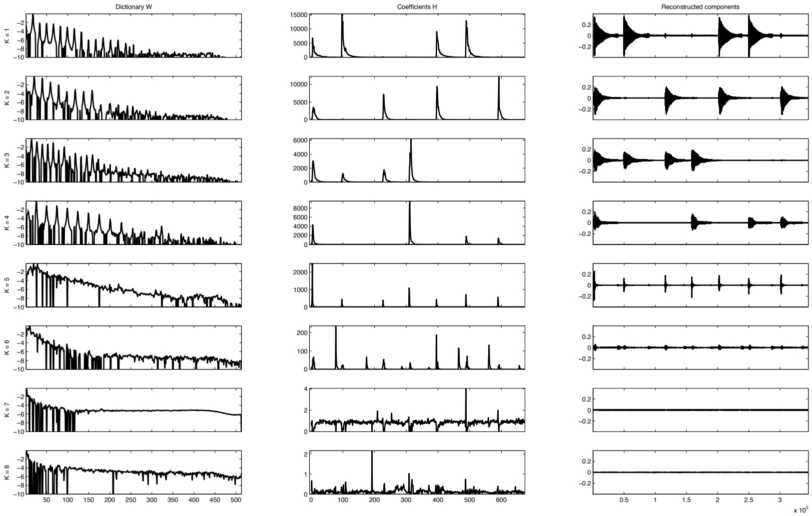


Figure: Three representations of data.

Piano toy example

IS-NMF on power spectrogram with $K = 8$

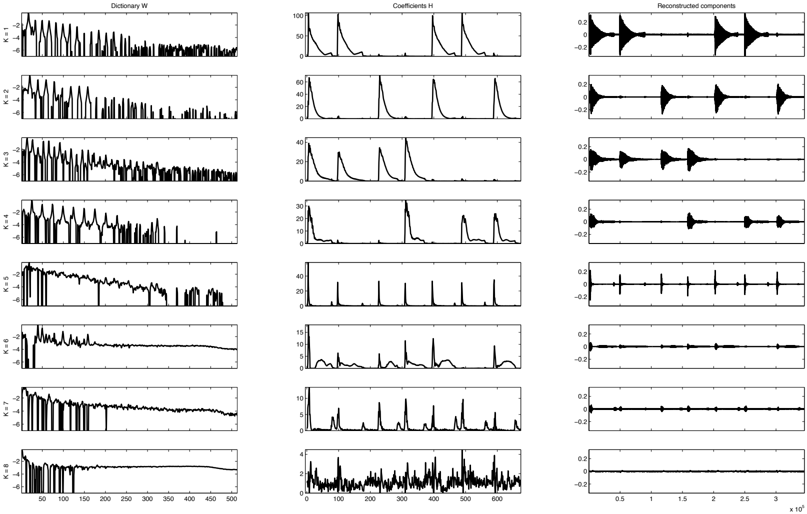


Pitch estimates: 65.0 68.0 61.0 72.0 0 0 0 0

(True values: 61, 65, 68, 72)

Piano toy example

KL-NMF on magnitude spectrogram with $K = 8$

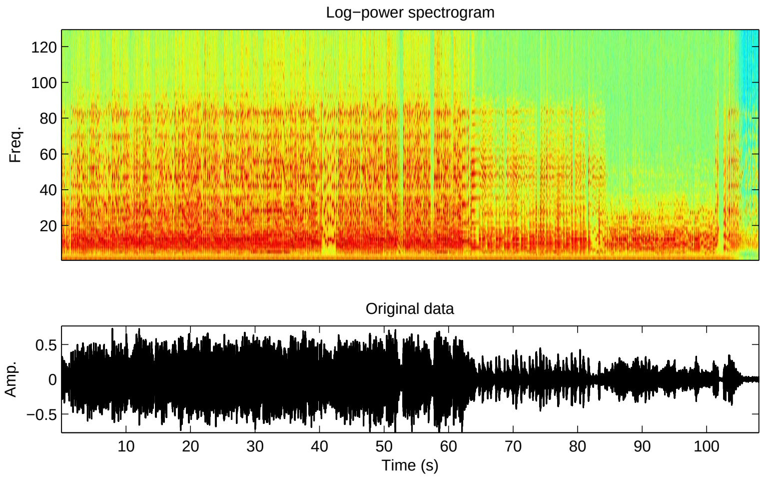


Pitch estimates: 65.2 68.2 61.0 72.2 0 56.2 0 0

(True values: 61, 65, 68, 72)

Audio restoration

Louis Armstrong and His Hot Five



Audio restoration

Louis Armstrong and His Hot Five

$$\text{Original mono} = \underbrace{\textit{Accompaniment}}_{\textit{Comp}.\ 1,9} + \underbrace{\textit{Brass}}_{\textit{Comp}.\ 2,3,5-8} + \underbrace{\textit{Trombone}}_{\textit{Comp}.\ 4} + \underbrace{\textit{Noise}}_{\textit{Comp}.\ 10}$$

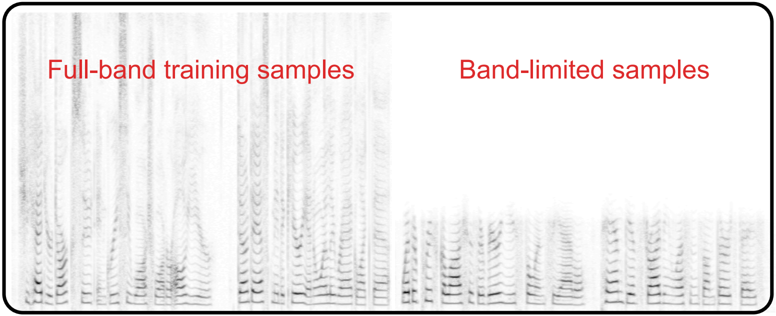
Original mono denoised

Original denoised & upmixed to stereo

Audio bandwidth extension

(Sun and Mazumder, 2013)

$V =$

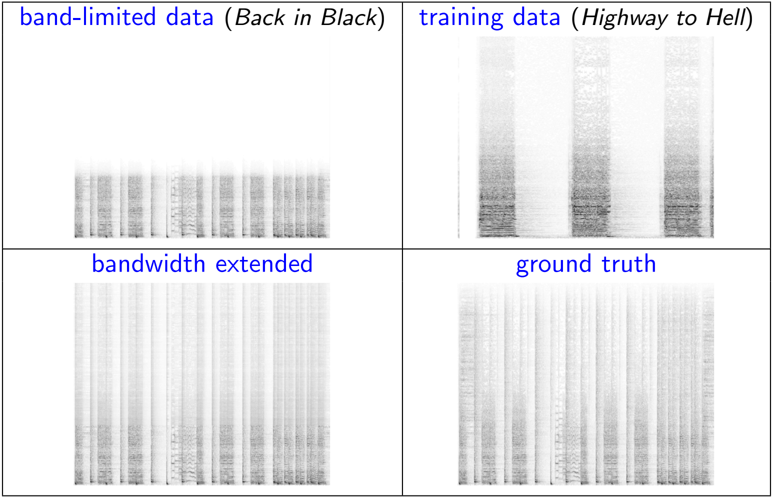


adapted from (Sun and Mazumder, 2013)

Audio bandwidth extension

(Sun and Mazumder, 2013)

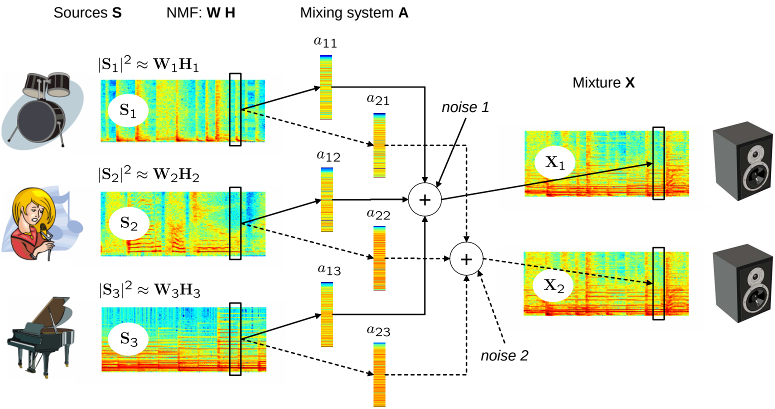
AC/DC example



Examples from `http://statweb.stanford.edu/~dlsun/bandwidth.html`*, used with permission from the author.*

Multichannel IS-NMF

(Ozerov and Févotte, 2010)



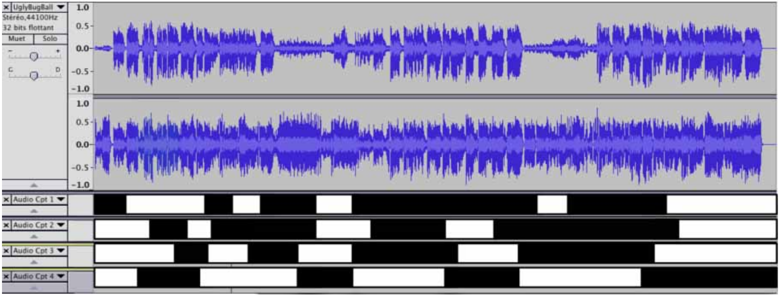
Multichannel NMF problem: Estimate **W**, **H** and **A** from **X**

- Best scores on the *underdetermined speech and music separation* task at the Signal Separation Evaluation Campaign (SiSEC) 2008.
- IEEE Signal Processing Society 2014 Best Paper Award.

User-guided multichannel IS-NMF

(Ozerov, Févotte, Blouet, and Durrieu, 2011)

- the decomposition is guided by the operator: source activation time-codes are input to the separation system.
- set forced zeros in **H** when a source is silent.



References I

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.

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.

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. 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 http://www.unice.fr/cfevotte/publications/journals/neco11.pdf.

References II

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 http://www.unice.fr/cfevotte/publications/journals/neco09_is-nmf.pdf.

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.

A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3): 550–563, Mar. 2010. doi: 10.1109/TASL.2009.2031510. URL http://www.unice.fr/cfevotte/publications/journals/ieee_asl_multinmf.pdf.

A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL http://www.unice.fr/cfevotte/publications/proceedings/icassp11d.pdf.

References III

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'03)*, Oct. 2003.

D. L. Sun and R. Mazumder. Non-negative matrix completion for bandwidth extension: a convex optimization approach. In *Proc. IEEE Workshop on Machine Learning and Signal Processing (MLSP)*, 2013.

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.