Itakura-Saito NMF: un modèle probabiliste à facteurs latents pour la transformée de Fourier court-terme

Cédric Févotte

Laboratoire Lagrange, Nice

Peyresq, juin 2013

Outline

Nonnegative matrix factorization (NMF)

Given a *nonnegative* matrix $\mathbf{V}$ of dimensions $F \times N$, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH}$$

where $\mathbf{W}$ and $\mathbf{H}$ are *nonnegative* matrices of dimensions $F \times K$ and $K \times N$, respectively.

Nonnegative matrix factorization (NMF)

Given a *nonnegative* matrix $\mathbf{V}$ of dimensions $F \times N$, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH}$$

where $\mathbf{W}$ and $\mathbf{H}$ are *nonnegative* matrices of dimensions $F \times K$ and $K \times N$, respectively.

Dimensions:

- If $\mathbf{W}$ tall ($K < F$), NMF produces a low-rank approximation.
- If $\mathbf{W}$ fat ($K > F$), NMF produces an overcomplete representation (e.g., sparse coding).

An unsupervised part-based representation

Along VQ, PCA or ICA, NMF provides an **unsupervised linear representation** of data

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
data vector		"explanatory variables"	"regressors"
		"basis", "dictionary"	"expansion coefficients"
		"patterns"	"activation coefficients"

and $\mathbf{W}$ is learnt from the set of data vectors $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$.

An unsupervised part-based representation

Along VQ, PCA or ICA, NMF provides an **unsupervised linear representation** of data

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
data vector		"explanatory variables"	"regressors"
		"basis", "dictionary"	"expansion coefficients"
		"patterns"	"activation coefficients"

and $\mathbf{W}$ is learnt from the set of data vectors $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$.

- **nonneg. of W** ensures *interpretability* of the dictionary (features $\mathbf{w}_k$ and data $\mathbf{v}_n$ belong to same space).
- **nonneg. of H** tends to produce *part-based* representations because subtractive combinations are forbidden.

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

NMF as a constrained minimization problem

Minimize a measure of fit between data $\mathbf{V}$ and model $\mathbf{WH}$, subject to nonnegativity of $\mathbf{W}$ and $\mathbf{H}$:

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

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

Regularization terms are often added to $D(\mathbf{V}|\mathbf{WH})$ to favor certain properties of $\mathbf{W}$ or $\mathbf{H}$ (sparsity, smoothness).

Divergences used in NMF

(selected references)

- Euclidean distance (Paatero and Tapper, 1994; Lee and Seung, 2001)
- Kullback-Leibler divergence (Lee and Seung, 1999; Finesso and Spreij, 2006)
- α-divergence (Cichocki et al., 2008)
- β-divergence (Cichocki et al., 2006; Févotte and Idier, 2011)
- Bregman divergences (Dhillon and Sra, 2005)

Divergences used in NMF

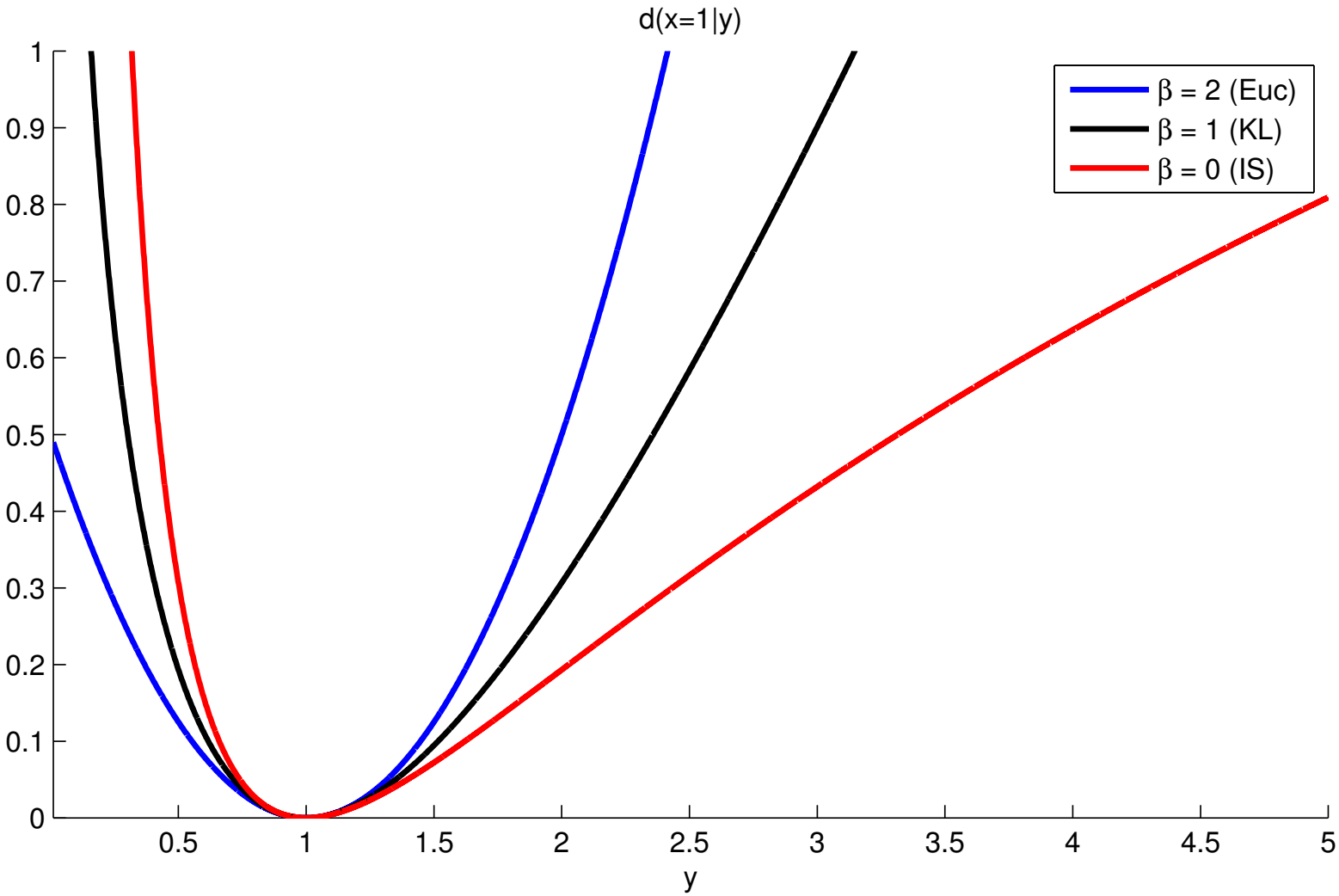
(selected references)

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

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

The Itakura-Saito divergence

(Itakura and Saito, 1968)



Common NMF algorithm design

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

Common NMF algorithm design

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

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

Common NMF algorithm design

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

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

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

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

Common NMF algorithm design

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

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

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

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

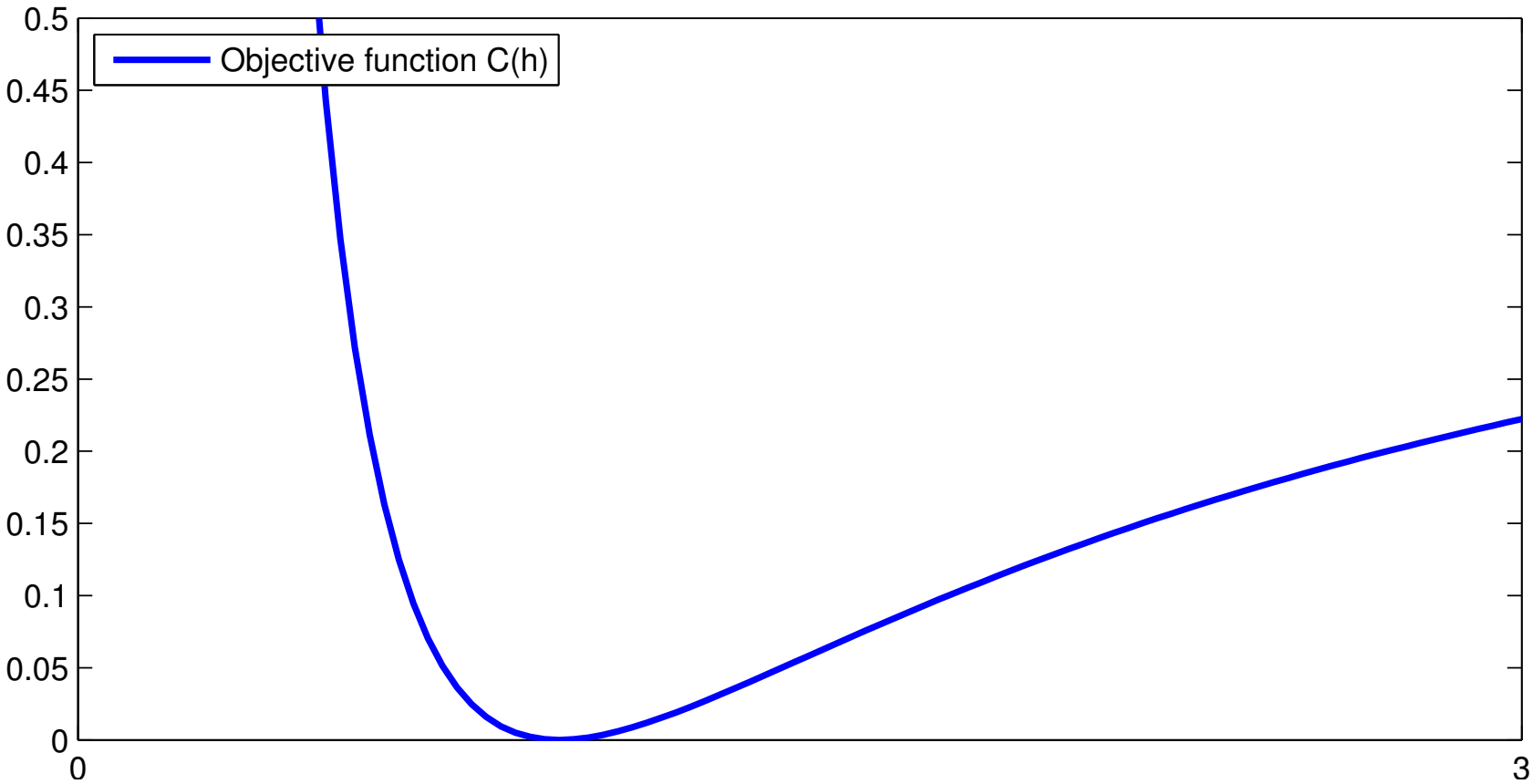
- In the end we are left with *nonnegative linear regression*

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

Numerous references in the image restoration literature (Richardson, 1972; Lucy, 1974; Daube-Witherspoon and Muehllehner, 1986)

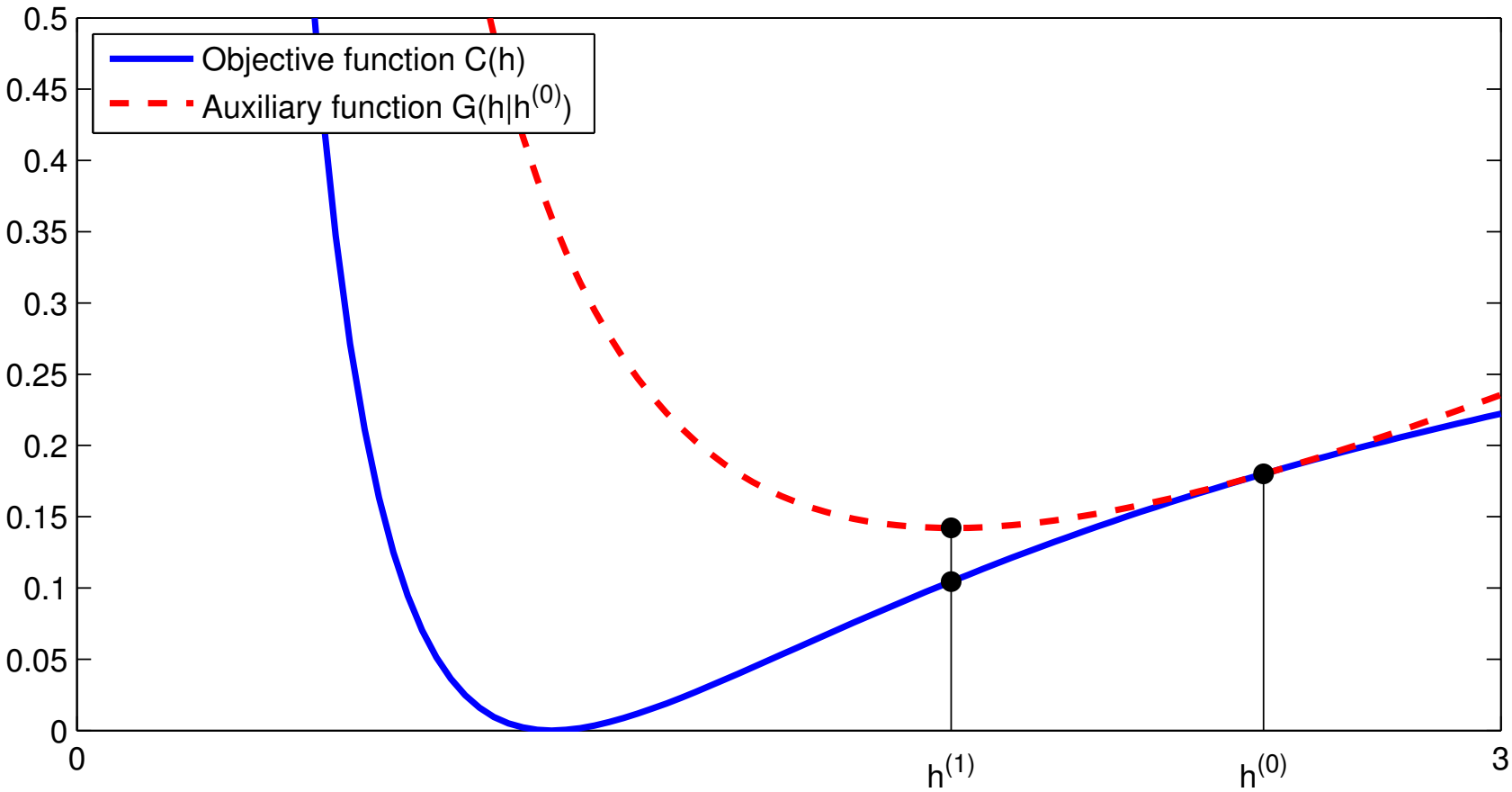
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



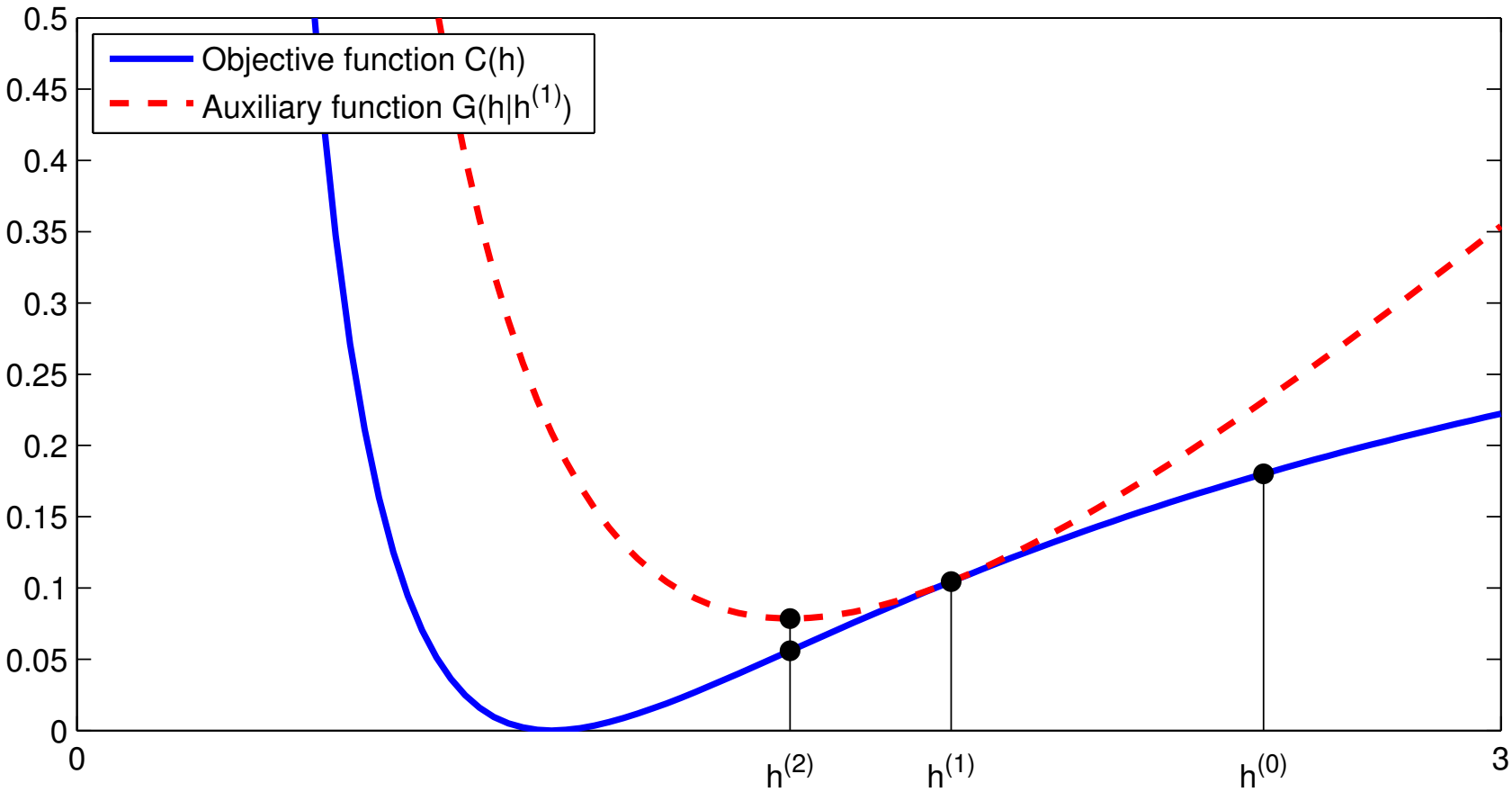
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



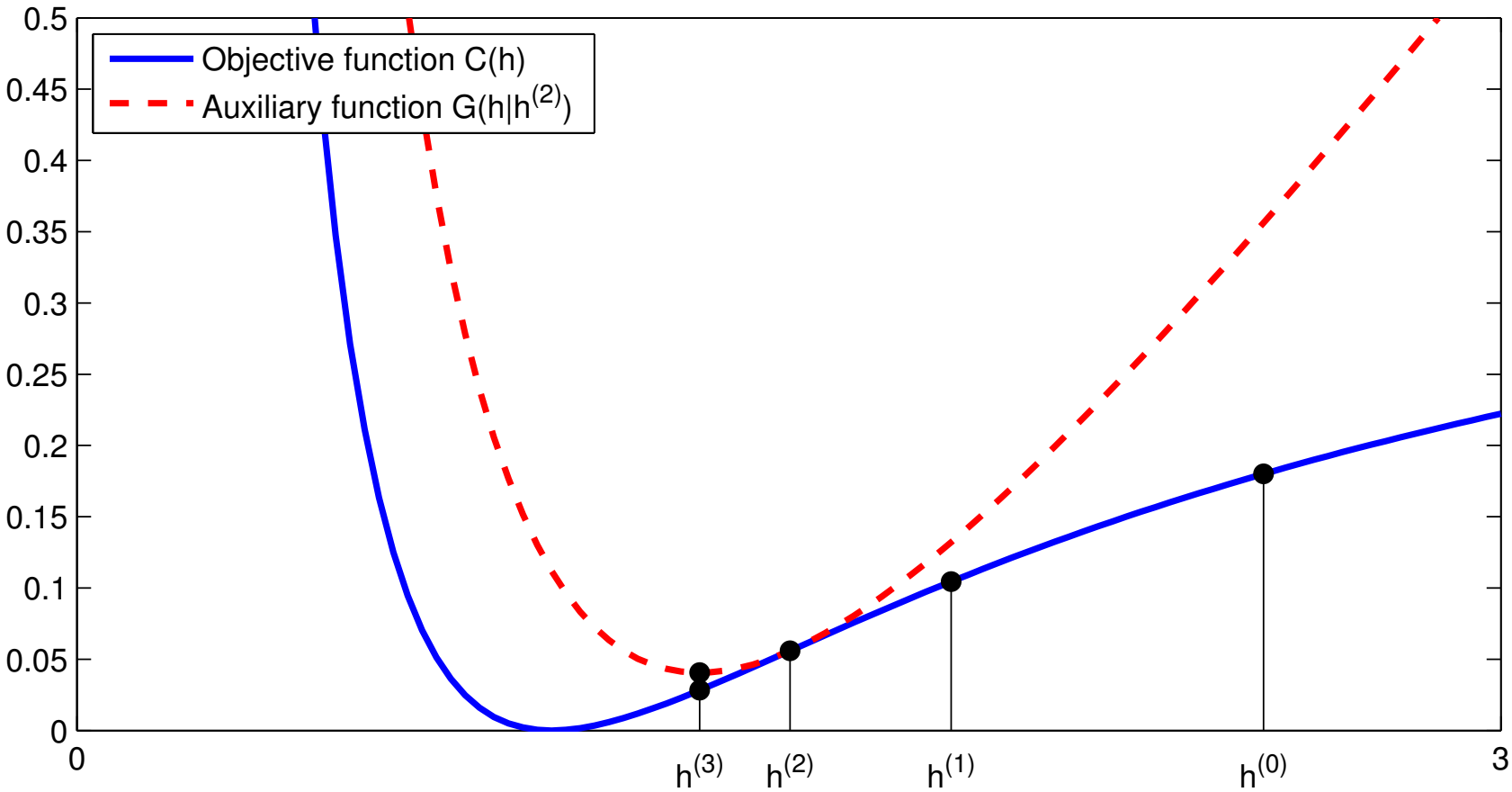
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



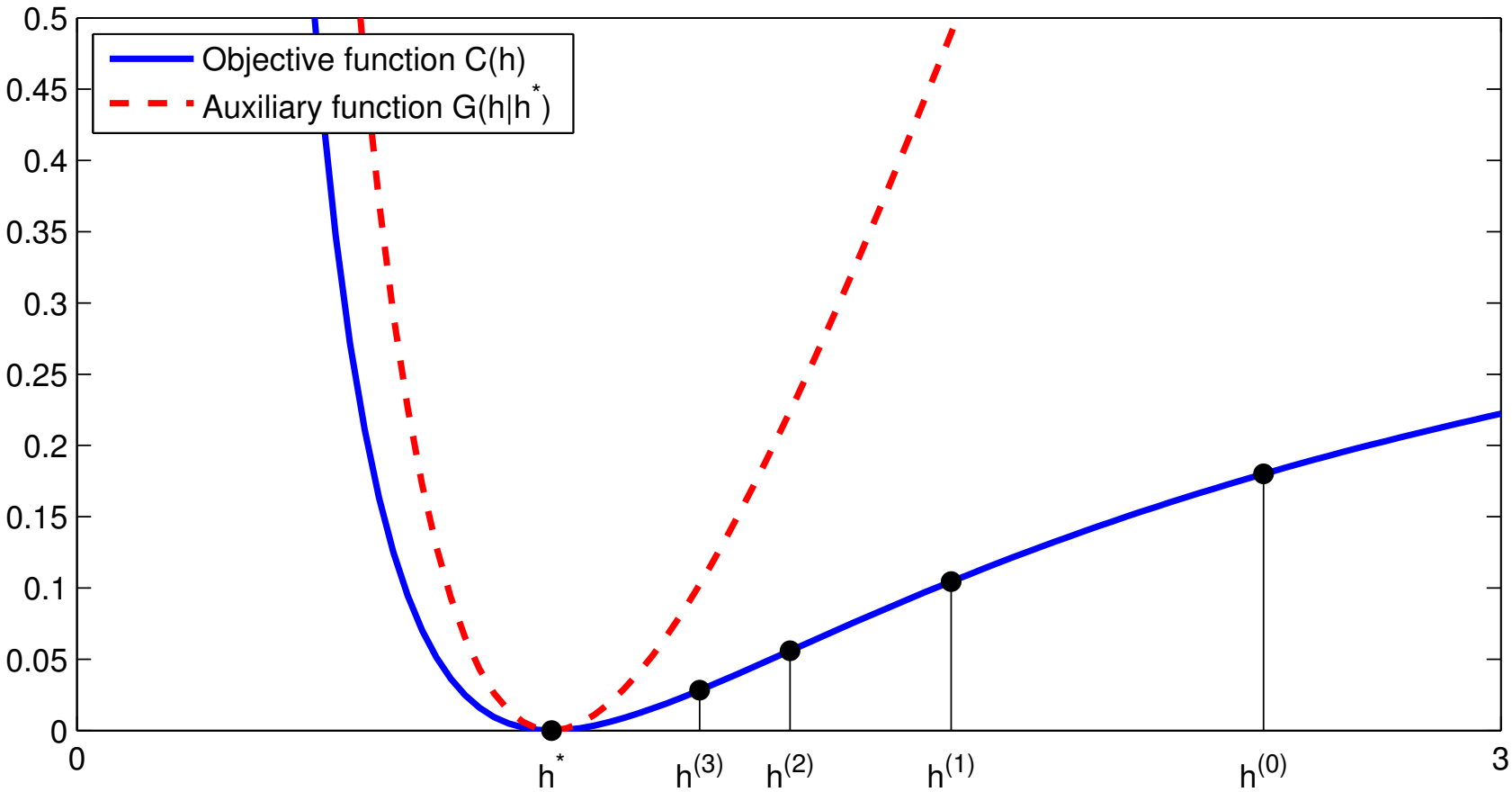
Majorization-minimization (MM)

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 $C(\mathbf{h})$.

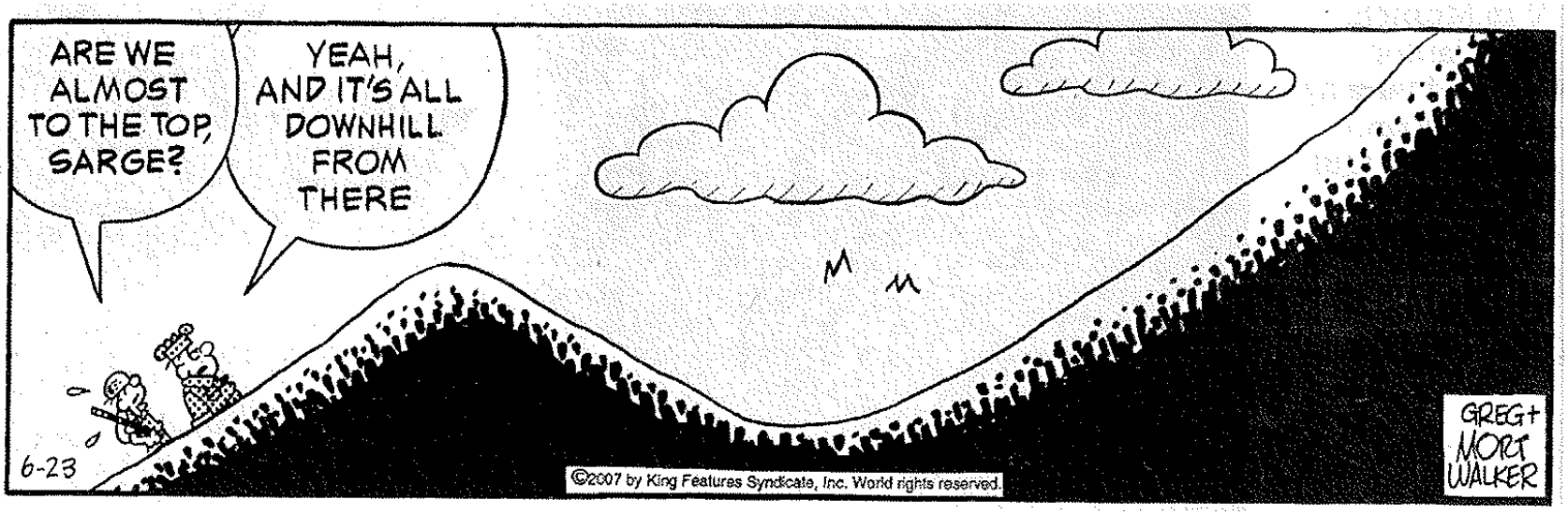


Majorization-minimization (MM)

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 $C(\mathbf{h})$.



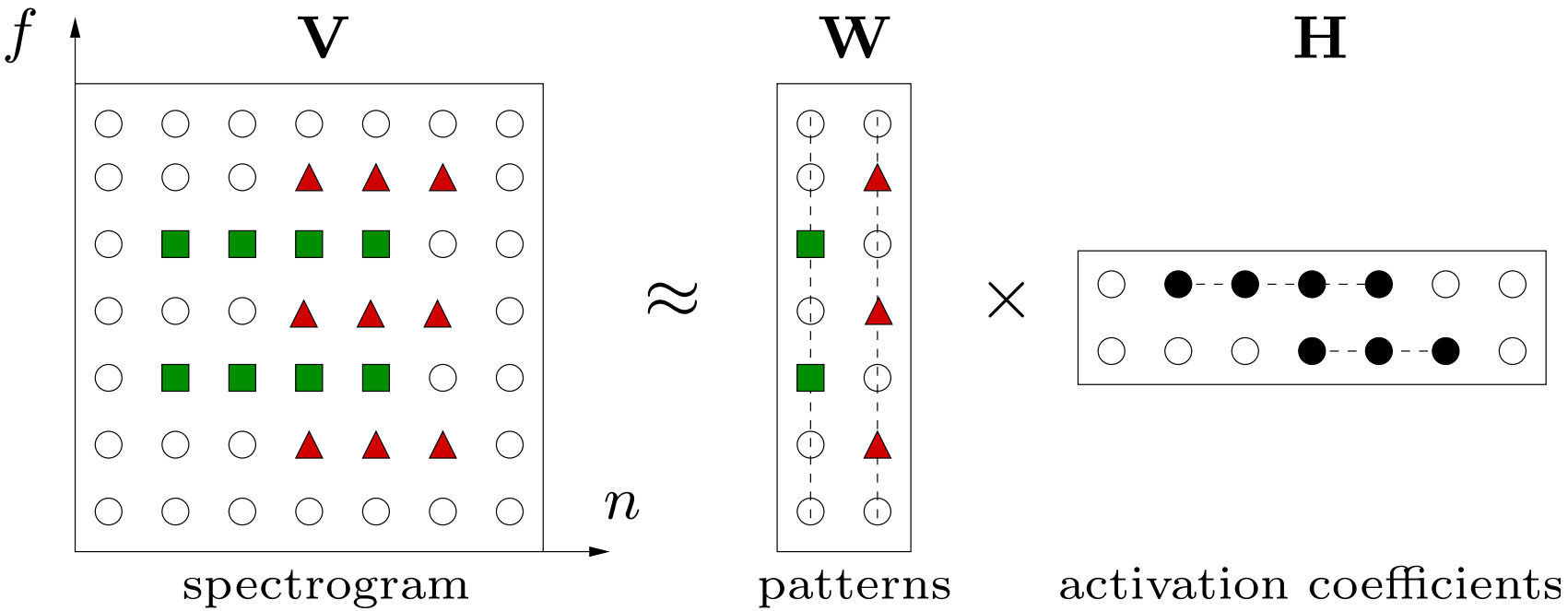
Local convergence



- If $d(x|y)$ is convex w.r.t to y, $D(\mathbf{V}|\mathbf{WH})$ convex w.r.t either $\mathbf{W}$ or $\mathbf{H}$ but not both.
- Not even true if $d(x|y)$ not convex w.r.t y.

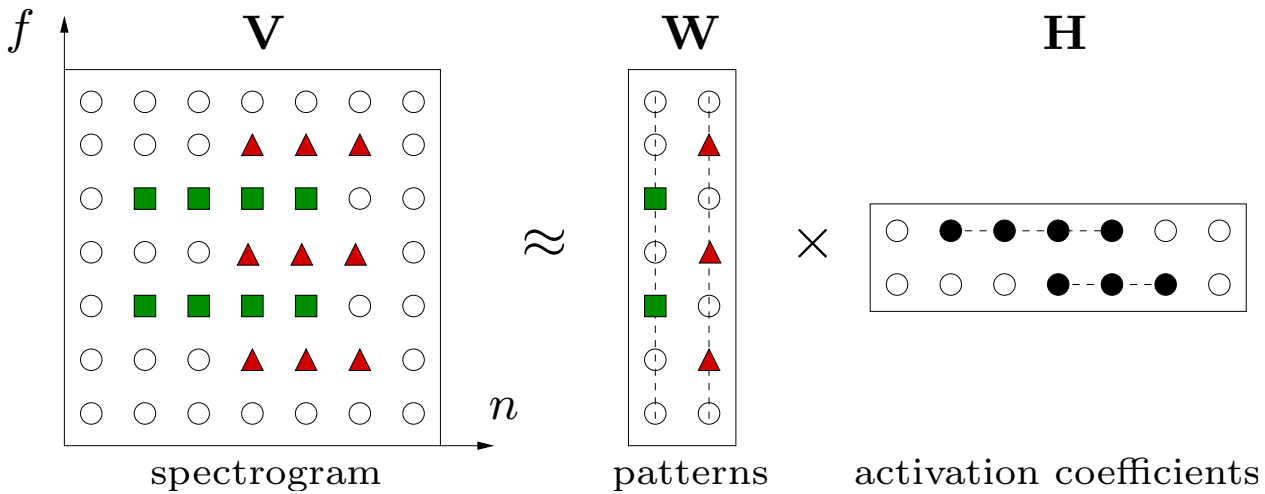
Application to music signal processing

(Smaragdis and Brown, 2003)



Outline

Model choices



- Magnitude or power spectrogram ?
- Which measure of fit should be used for the factorization ?
- NMF approximates the spectrogram by a sum of rank-one spectrograms. How can we invert these ? What about phase ?

Itakura-Saito NMF: a generative approach

(Févotte, Bertin, and Durrieu, 2009)

Let $\mathbf{X} = \{x_{fn}\}$ be the (complex-valued) STFT of the signal. Assume

$$x_{fn} = \sum_{k=1}^{K} c_{k,fn}$$

$$c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn})$$

and the components $c_{1,fn}, \ldots, c_{K,fn}$ are independent given $\mathbf{W}$ and $\mathbf{H}$.

Itakura-Saito NMF: a generative approach

(Févotte, Bertin, and Durrieu, 2009)

Let $\mathbf{X} = \{x_{fn}\}$ be the (complex-valued) STFT of the signal. Assume

$$x_{fn} = \sum_{k=1}^{K} c_{k,fn}$$
$$c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn})$$

and the components $c_{1,fn}, \ldots, c_{K,fn}$ are independent given $\mathbf{W}$ and $\mathbf{H}$. Then

$$-\log p(\mathbf{X}|\mathbf{W},\mathbf{H}) = D_{IS}(|\mathbf{X}|^2|\mathbf{WH}) \quad + cst.$$

Additivity assumed in the STFT domain. Phase is preserved in the model, though in a noninformative way (uniform distribution).

Related work by Benaroya et al. (2003); Parry and Essa (2007)

Itakura-Saito NMF: a generative approach

(Févotte, Bertin, and Durrieu, 2009)

Main message: Itakura-Saito NMF of the power spectrogram corresponds to maximum likelihood estimation in a well-defined generative composite model of the STFT.

This in particular gives a statistically grounded way of reconstructing the components:

$$\hat{c}_{k,fn} = \mathrm{E}\{c_{k,fn}|\mathbf{X},\mathbf{W},\mathbf{H}\} = \underbrace{\frac{w_{fk}h_{kn}}{\sum_j w_{fj}h_{jn}}}_{\text{time-freq. mask}} x_{fn}$$

Lossless decomposition: $x_{fn} = \sum_k \hat{c}_{k,fn}$

Itakura-Saito NMF: a generative approach

(Févotte, Bertin, and Durrieu, 2009)

Alternatively, IS-NMF can be interpreted as maximum likelihood in multiplicative noise:

$$v_{fn} = |x_{fn}|^2 = [\mathbf{WH}]_{fn} \cdot \epsilon_{fn}$$

where ϵ_{fn} is Gamma multiplicative noise with mean value 1.

Related work by Abdallah and Plumbley (2004).

Noteworthy properties of the IS divergence

- The IS divergence is scale-invariant:

$$d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y)$$

Implies higher accuracy in the representation of data with large dynamic range, such as audio spectra. In contrast,

$$\begin{aligned} d_{EUC}(\lambda x|\lambda y) &= \lambda^2 d_{EUC}(x|y) \\ d_{KL}(\lambda x|\lambda y) &= \lambda d_{KL}(x|y) \end{aligned}$$

- The IS divergence in nonconvex (inflexion at $y = 2x$); was found to lead to more local minima in practice.

Small-scale example

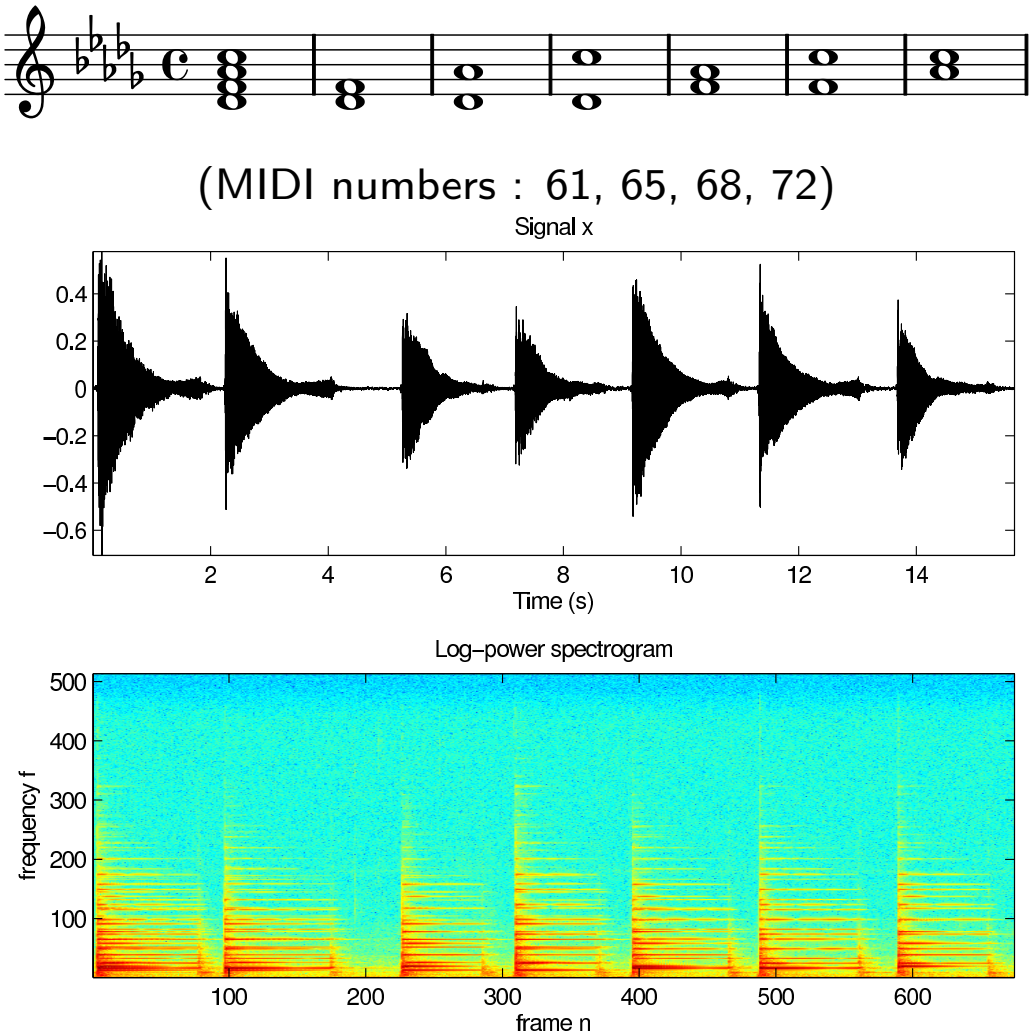
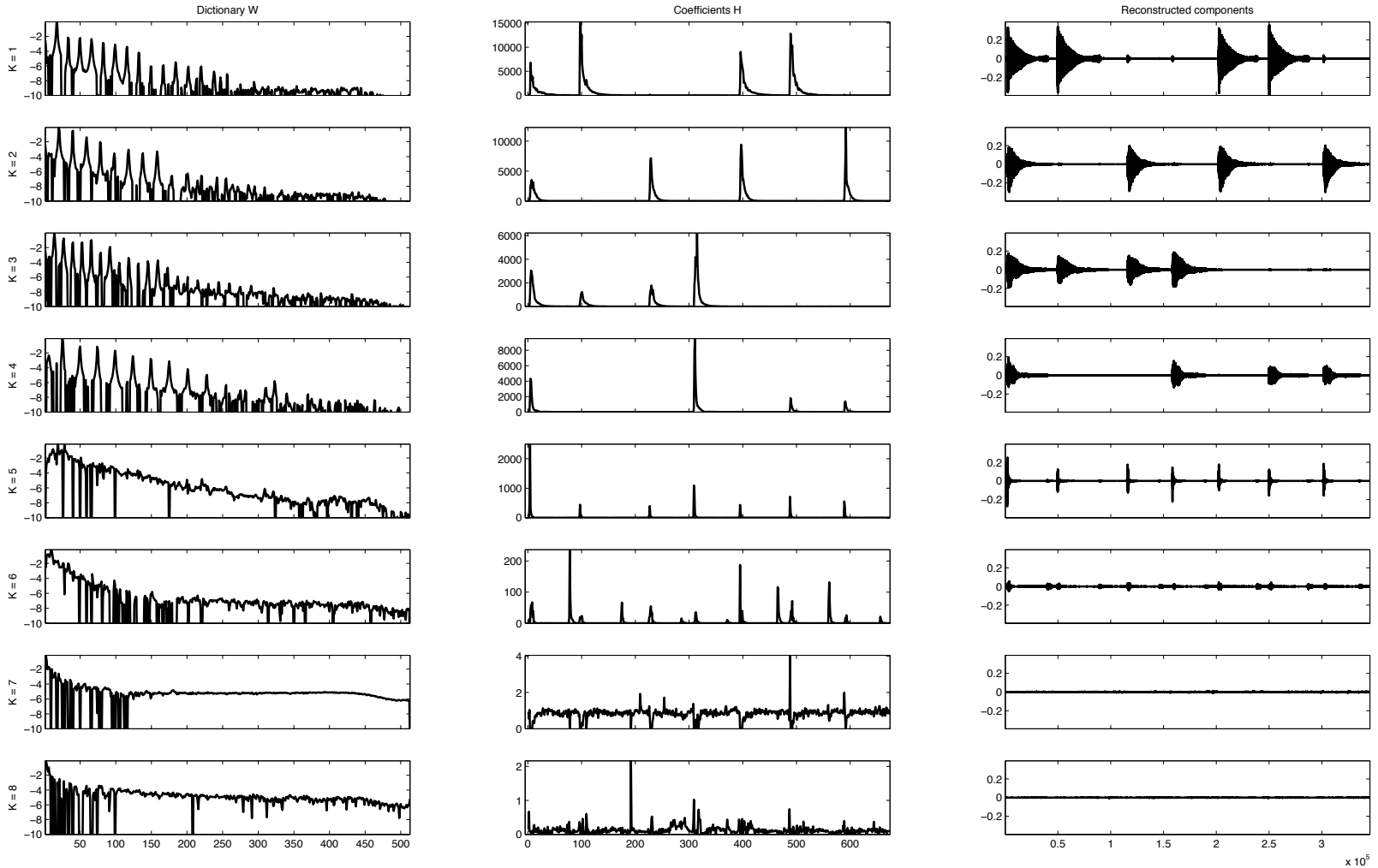


Figure: Three representations of data.

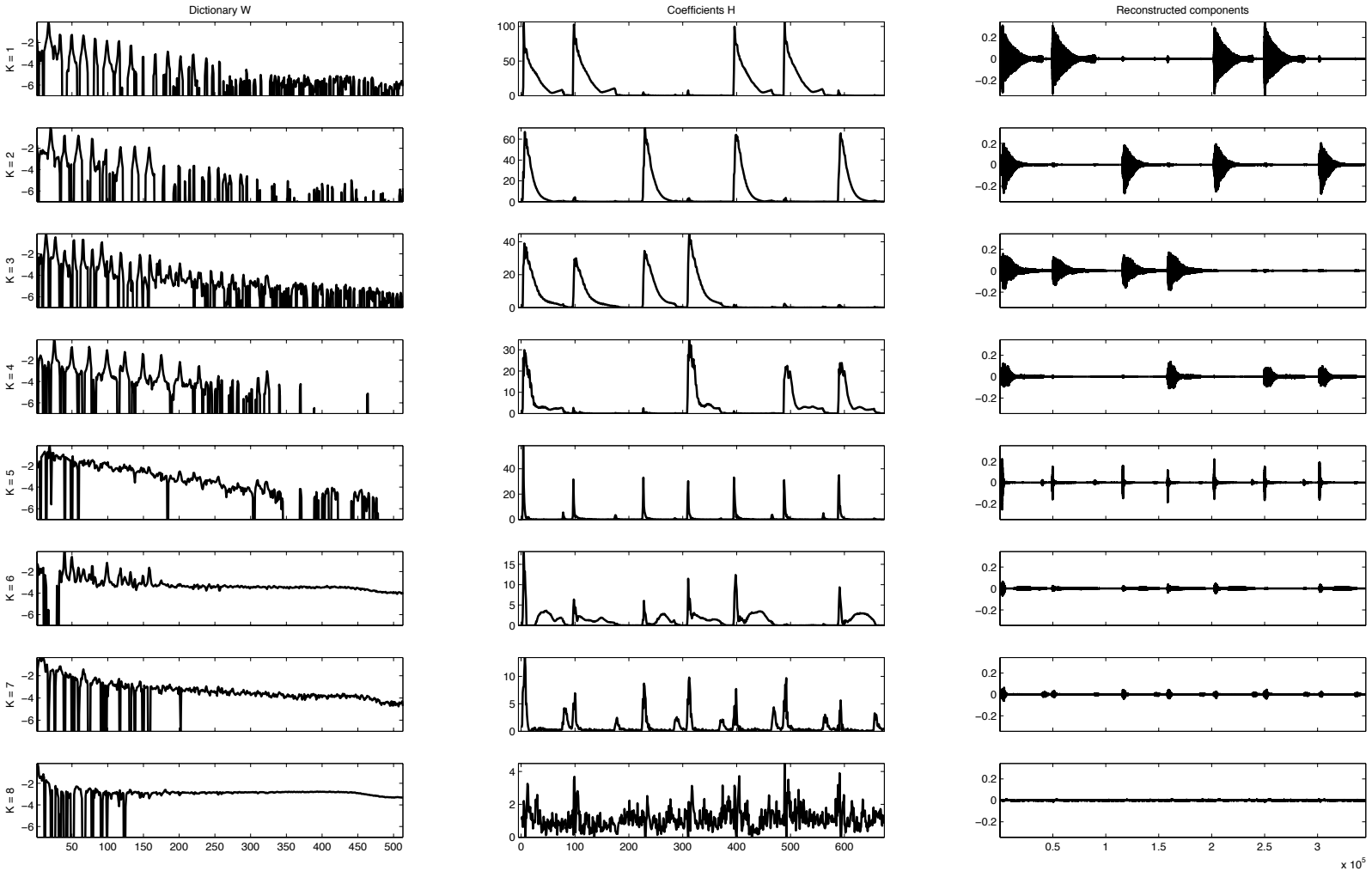
IS-NMF on power spectrogram with $K = 8$



Pitch estimates: 65.0 68.0 61.0 72.0 0 0 0 0

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

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



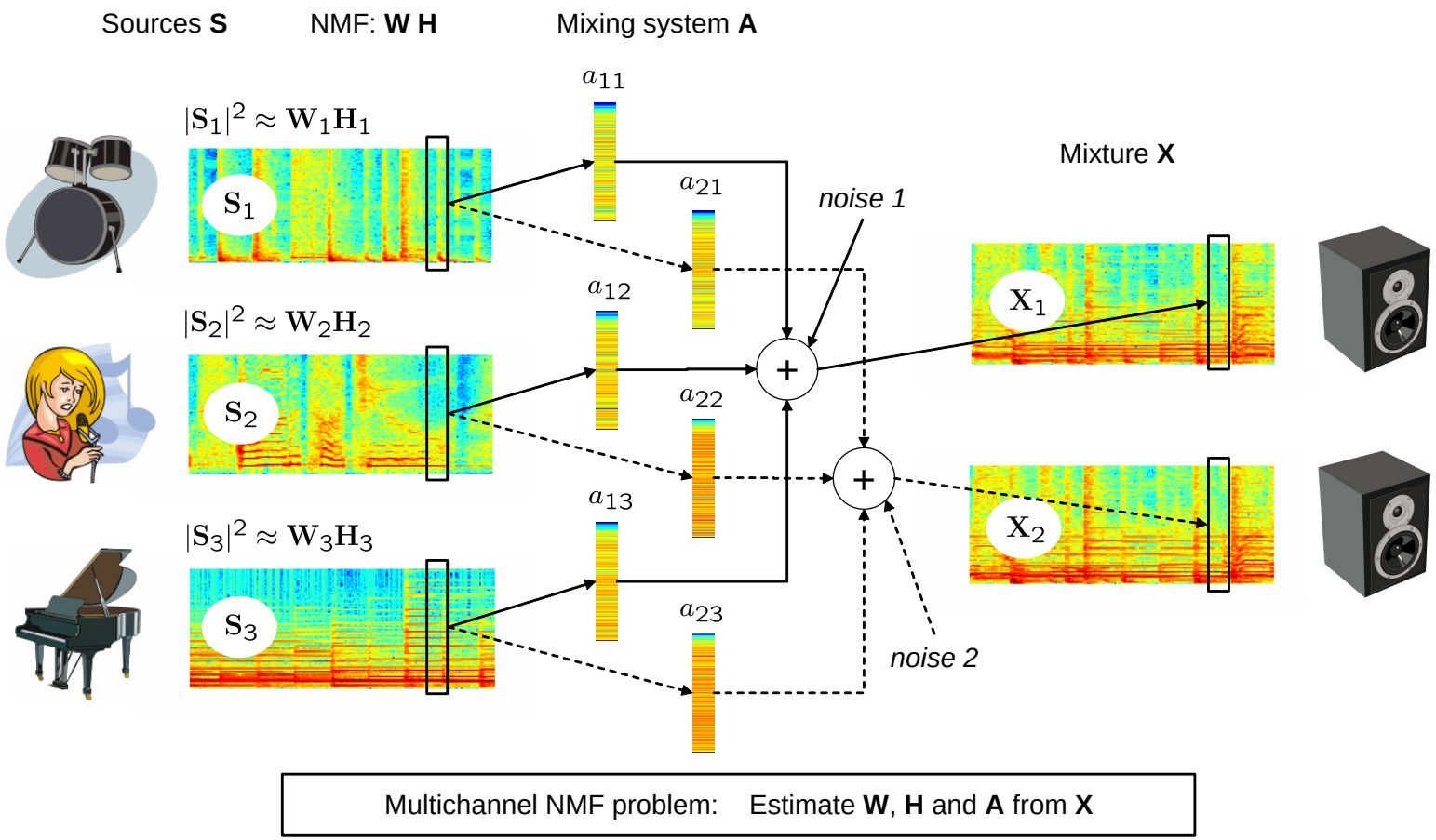
Pitch estimates: 65.2 68.2 61.0 72.2 0 56.2 0 0

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

Outline

Multichannel IS-NMF

(Ozerov and Févotte, 2010)

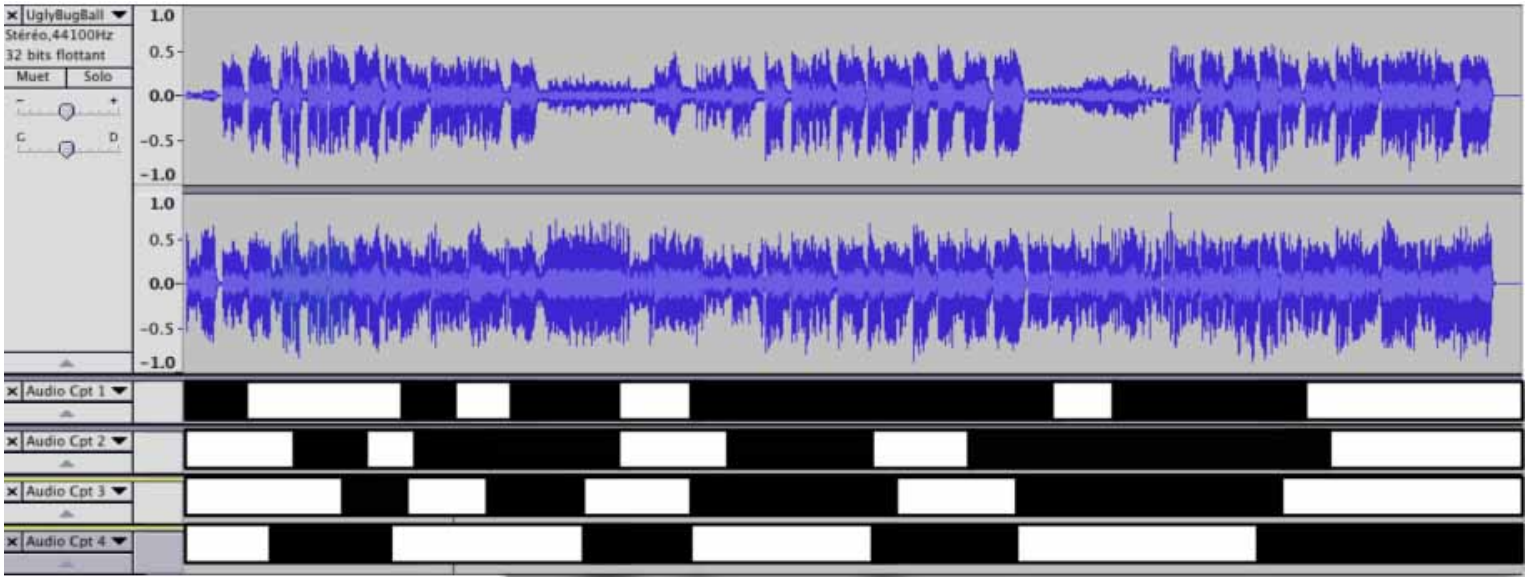


Best scores on the *underdetermined speech and music separation* task at the Signal Separation Evaluation Campaign (SiSEC) 2008.

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.
- The temporal segmentation is reflected in the form of zeros in $\mathbf{H}$ when a source is silent.



Conclusions

- Itakura-Saito NMF of the power spectrogram is underlain by a statistical model of superimposed Gaussian components.
- This model is relevant to the representation of audio signals.
- Algorithms can be designed in a principled way in the majorization-minimization setting.
- Possible extension to multichannel data for audio source separation.

- The latent statistical model opens doors to fully Bayesian approaches that integrates over $\mathbf{W}$ and/or $\mathbf{H}$ (Févotte and Cemgil, 2009; Hoffman et al., 2010; Févotte et al., 2011; Dikmen and Févotte, 2011)

References I

S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by nonnegative sparse coding of power spectra. In *Proc. 5th International Symposium Music Information Retrieval (ISMIR)*, pages 318–325, Barcelona, Spain, Oct. 2004.

L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for Wiener based source separation with a single sensor. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 613–616, Hong Kong, 2003.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

W. L. Buntine and A. Jakulin. Discrete component analysis. In *Lecture Notes in Computer Science*, volume 3940, pages 1–33. Springer, 2006. URL `http://www.springerlink.com/content/d53027666542q3v7/`.

J. F. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th ACM international Conference on Research and Development of Information Retrieval (SIGIR)*, pages 122–129, 2004.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009(Article ID 785152):17 pages, 2009. doi:10.1155/2009/785152.

References II

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th 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.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2005.

O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2267–2275, Granada, Spain, Dec. 2011. MIT Press. URL `http://www.unice.fr/cfevotte/publications/proceedings/nips11.pdf`.

References III

C. Févotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009. URL `http://www.unice.fr/cfevotte/publications/proceedings/eusipco09a.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 `http://www.unice.fr/cfevotte/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 `http://www.unice.fr/cfevotte/publications/journals/neco09_is-nmf.pdf`.

C. Févotte, O. Cappé, and A. T. Cemgil. Efficient Markov chain Monte Carlo inference in composite models with space alternating data augmentation. In *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, pages 221 – 224, Nice, France, June 2011. URL `http://www.unice.fr/cfevotte/publications/proceedings/ssp11.pdf`.

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

References IV

M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proc. 27th International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010.

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`.

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc 6th International Congress on Acoustics*, pages C–17 – C–20, Tokyo, Japan, Aug. 1968.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1135–1145, May 2007.

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.

References V

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`.

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.

R. M. Parry and I. Essa. Phase-aware non-negative spectrogram factorization. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 536–543, London, UK, Sep. 2007.

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

References VI

M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008 (Article ID 947438):8 pages, 2008. doi:10.1155/2008/947438.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003.

P. Smaragdis, M. Shashanka, and B. Raj. A sparse non-parametric approach for single channel separation of known sounds. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1705–1713. MIT Press, 2009.