Itakura-Saito nonnegative matrix factorization and friends for music signal decomposition

Cédric Févotte

CNRS LTCI; Télécom ParisTech

ESI, Vienna
December 2012

Outline

Nonnegative matrix factorization (NMF)

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

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}$$

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

K is usually chosen such that $F\,K + K\,N << F\,N$, hence reducing the data dimension, but not always.

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"

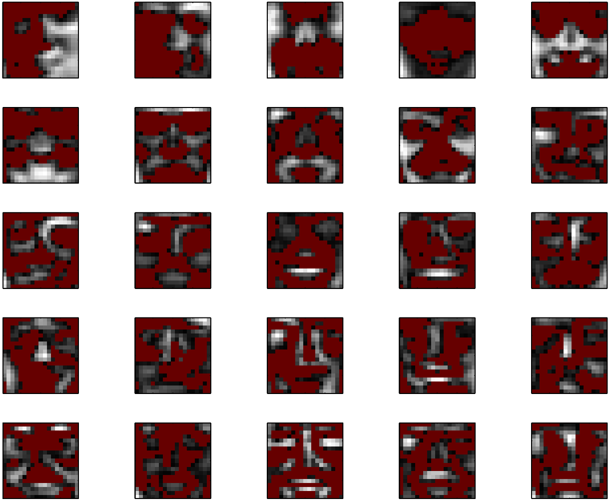
and $\mathbf{W}$ is learnt from the set of data vectors $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$.

- **nonneg. of $\mathbf{W}$** ensures *interpretability* of the dictionary (features $\mathbf{w}_k$ and data $\mathbf{v}_n$ belong to same space).
- **nonneg. of $\mathbf{H}$** 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).

49 images among 2429 from MIT's CBCL face dataset

PCA dictionary with $K = 25$



red pixels indicate negative values

NMF dictionary with $K = 25$

as shown in (Lee and Seung, 1999)

NMF as a constrained minimization problem

We seek to 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 sparsity or smoothness of $\mathbf{W}$ or $\mathbf{H}$.

Identifiability

Trivial scale and order ambiguities between the columns of $\mathbf{W}$ and the rows of $\mathbf{H}$.

There may be more complex geometrical ambiguities if the data does not fill the positive orthant "sufficiently well".

- In the exact/noiseless case $\mathbf{V} = \mathbf{W}^{\star}\mathbf{H}^{\star}$, the elements of the true dictionary $\mathbf{W}^{\star}$ need to belong to facets of the positive orthant (Donoho and Stodden, 2004).
- In the approximate/noisy case, the situation is less clear, see (Laurberg, Christensen, Plumbley, Hansen, and Jensen, 2008). Adding regularization terms certainly help.

β-divergence

A very popular cost function in NMF is the β-divergence (Basu et al., 1998; Eguchi and Kano, 2001; Cichocki and Amari, 2010), given by

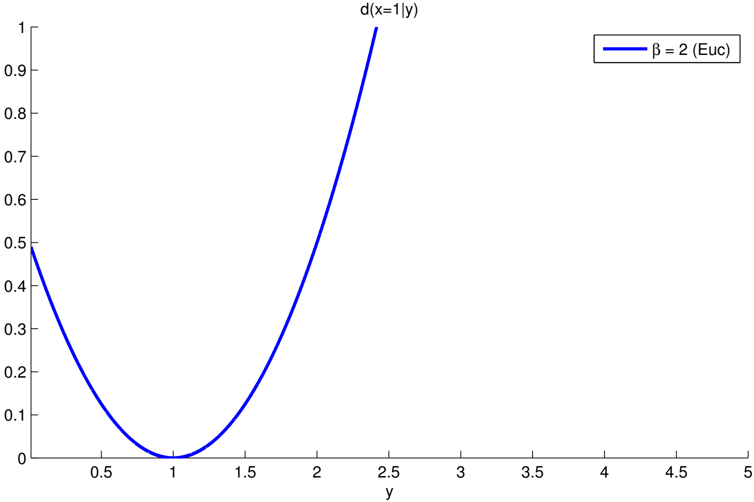
$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log\frac{x}{y} + (y-x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases}$$

which takes the

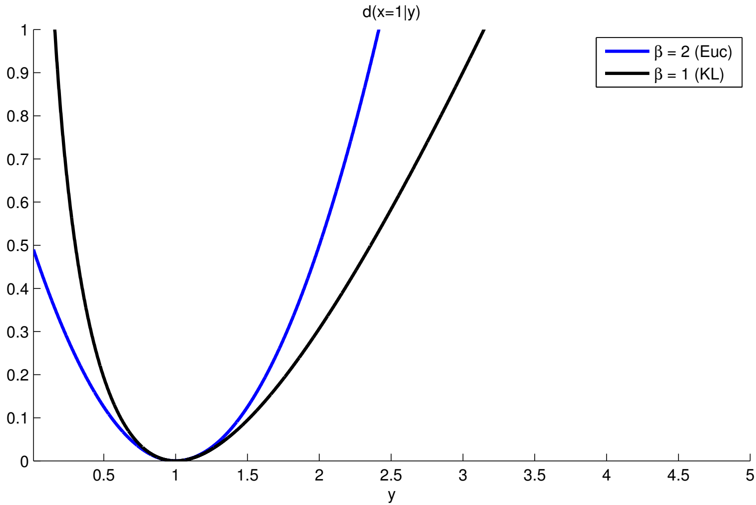
- Euclidean distance ($\beta = 2$)
- Kullback-Leibler (KL) divergence ($\beta = 1$)
- Itakura-Saito (IS) divergence ($\beta = 0$)

as special cases.

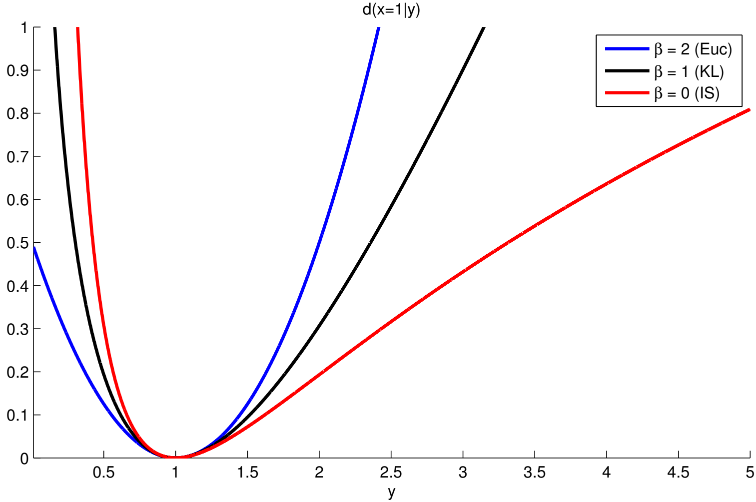
β-divergence



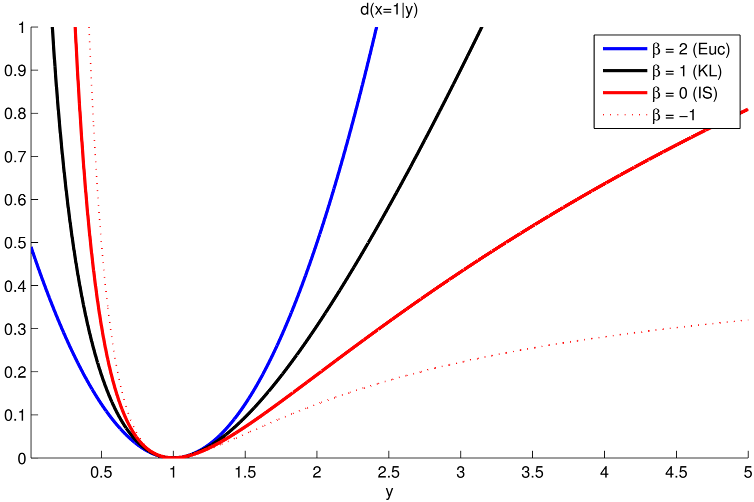
β-divergence



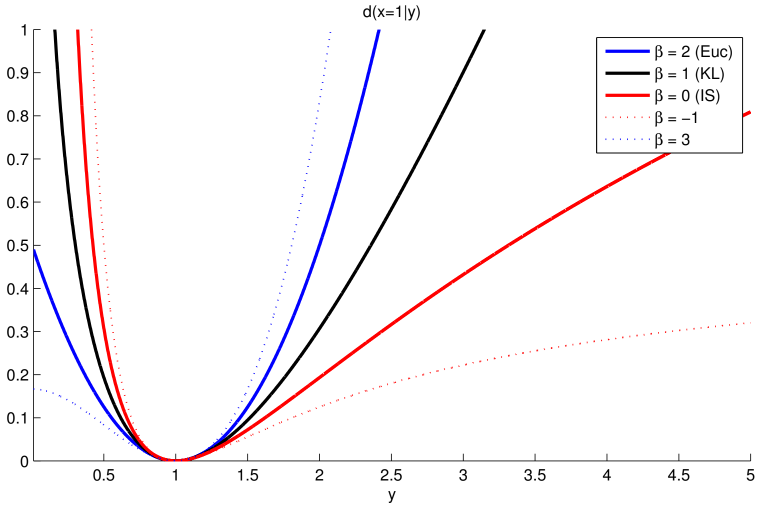
β-divergence



β-divergence



β-divergence



Common NMF algorithm design

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

$$\min_{\mathbf{H}\geq\mathbf{0}} D(\mathbf{V}|\mathbf{W}^{(i-1)}\mathbf{H}), \quad \min_{\mathbf{W}\geq\mathbf{0}} D(\mathbf{V}|\mathbf{W}\mathbf{H}^{(i)})$$

- The updates of $\mathbf{W}$ and $\mathbf{H}$ are equivalent by symmetry:

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

- 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

In the end we are left with

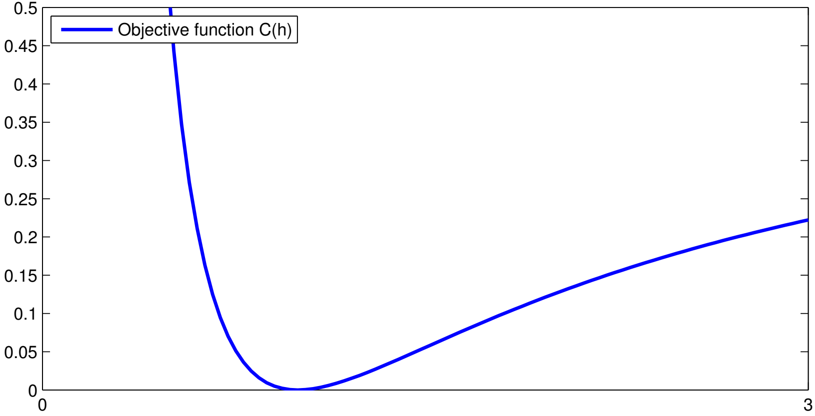
$$\min_{\mathbf{h} \geq \mathbf{0}} C(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h})$$

which is a *nonnegative linear regression* problem that has received considerable attention in image restoration, e.g.,

- (Richardson, 1972; Lucy, 1974) with KL divergence
- (Daube-Witherspoon and Muehllehner, 1986; De Pierro, 1993) for Euclidean distance
- (Cao, Eggermont, and Terebey, 1999) for Itakura-Saito divergence (aka Burg entropy)

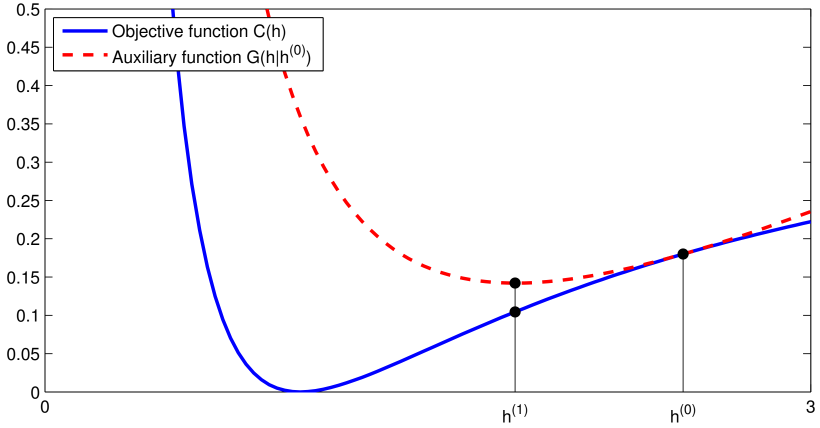
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



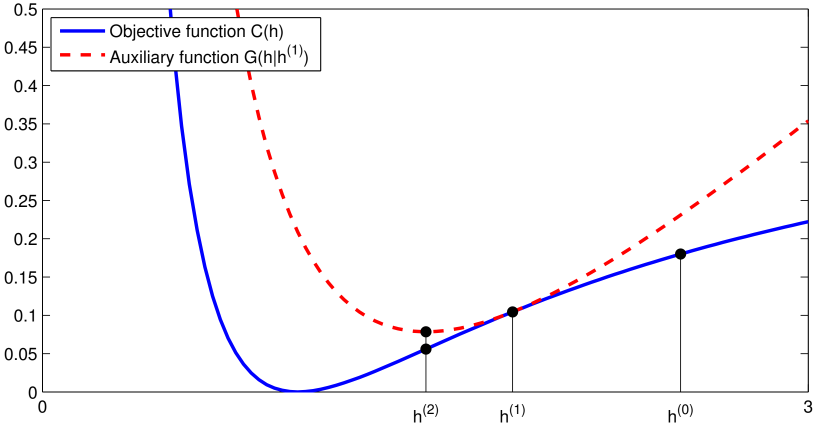
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



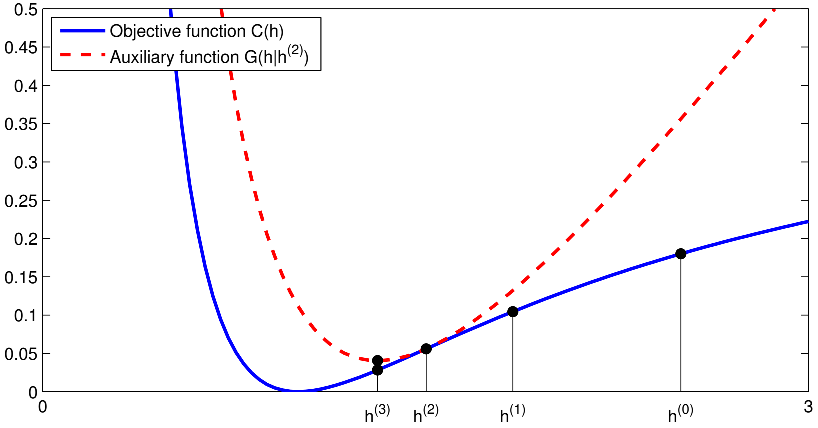
Majorization-minimization (MM)

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 $C(\mathbf{h})$.



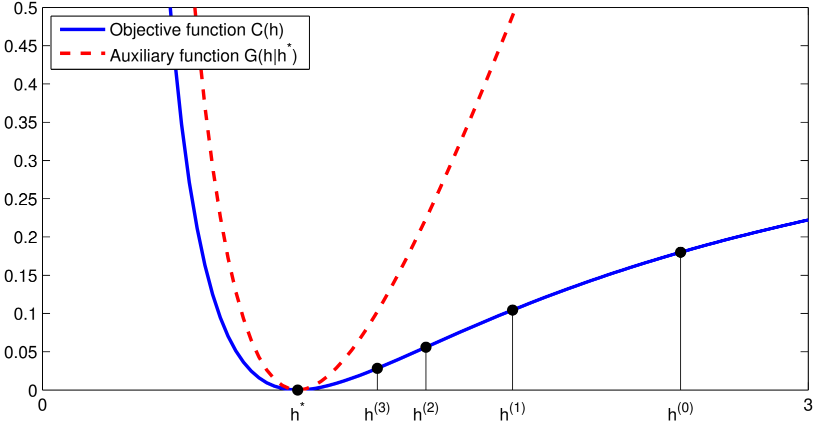
Majorization-minimization (MM)

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)

Majorize convex and concave parts separately

$$C(\mathbf{h}) = \underbrace{\breve{C}(\mathbf{h})}_{\text{Maj. by Jensen's ineq.}} + \underbrace{\frown{C}(\mathbf{h})}_{\text{Maj. by tangent}}$$

Majorization-minimization (MM)

Majorize convex and concave parts separately

$$C(\mathbf{h}) = \underbrace{\breve{C}(\mathbf{h})}_{\text{Maj. by Jensen's ineq.}} + \underbrace{\widehat{C}(\mathbf{h})}_{\text{Maj. by tangent}}$$

In the end, MM for β-NMF leads to

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

where $\gamma(\beta)$ is a scalar β-dependent exponent, see (Nakano et al., 2010; Févotte and Idier, 2011). Similar update for $\mathbf{W}$.

Multiplicative form preserves nonnegativity given nonnegative initialization.
Very easy to implement, of complexity $\mathcal{O}(FKN)$ per iteration.

Other algorithms

Specific to the Euclidean distance in most cases.

- Alternating nonnegative least squares (Paatero and Tapper, 1994; Berry et al., 2007)
- Projected gradient descent (Lin, 2007)
- Interior-point gradient descent (Merritt and Zhang, 2005)
- Quasi-Newton methods (Zdunek and Cichocki, 2007; Kim et al., 2008)
- Active set methods (Kim and Park, 2008)
- Fast coordinate descent (Hsieh and Dhillon, 2011; Cichocki and Anh-Huy, 2009)

(selected references)

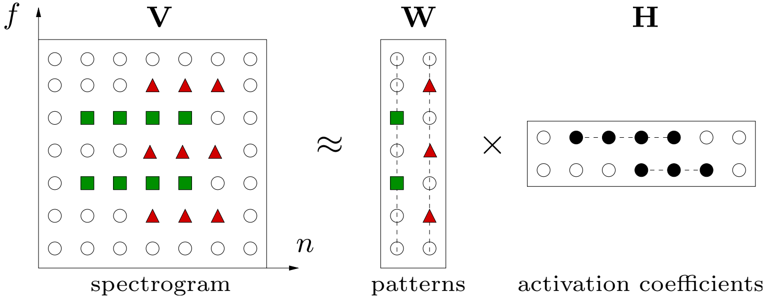
Some applications of NMF...

- environmetrics (Paatero and Tapper, 1994)
- video summarization (Cooper and Foote, 2002)
- text mining (Lee and Seung, 1999; Xu et al., 2003)
- gene expression analysis (Brunet et al., 2004)
- Scotch whiskies clustering (Young et al., 2006) (!)
- hyperspectral imaging (Berry et al., 2007)
- portfolio diversification (Drakakis et al., 2007)
- clustering of protein interactions (Greene et al., 2008)
- food consumption analysis (Zetlaoui et al., 2010)
- image denoising and inpainting (Mairal et al., 2010)

(selected references)

...and in particular

- music signal processing (Smaragdis and Brown, 2003)



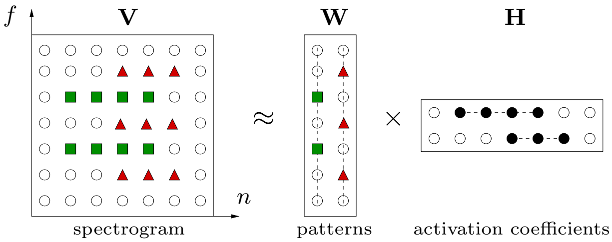
Outline

Generalities about NMF
- Concept of NMF
- The β-divergence: a unifying measure of fit
- Majorization-minimization algorithms

Itakura-Saito NMF
- Statistical model
- Piano decomposition example

Variants of IS-NMF
- Regularized IS-NMF
- Multichannel IS-NMF

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:

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

where ϵ_{fn} is exponential noise.

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.

Other statistical factor models of the spectrogram

Latent factor models for count data inspired from text analysis:

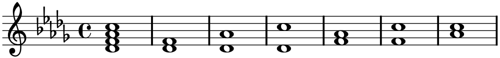
- Poisson models (Le Roux et al., 2007; Cemgil, 2009), similar to *GaP* (Canny, 2004)
- Multinomial models (Shashanka et al., 2008; Smaragdis et al., 2009), similar to *PLSI* (Hofmann, 1999) or *LDA* (Blei et al., 2003; Buntine and Jakulin, 2006)

Not generative models:

- Data $|x_{fn}|$ is modeled as integer.
- Additivity is assumed at the magnitude level

$$|x_{fn}| = \sum_k |c_{k,fn}|.$$

Small-scale example



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

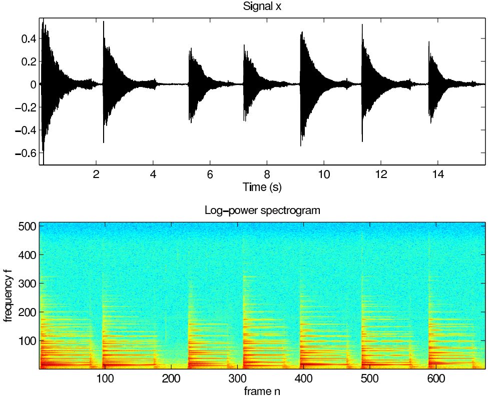
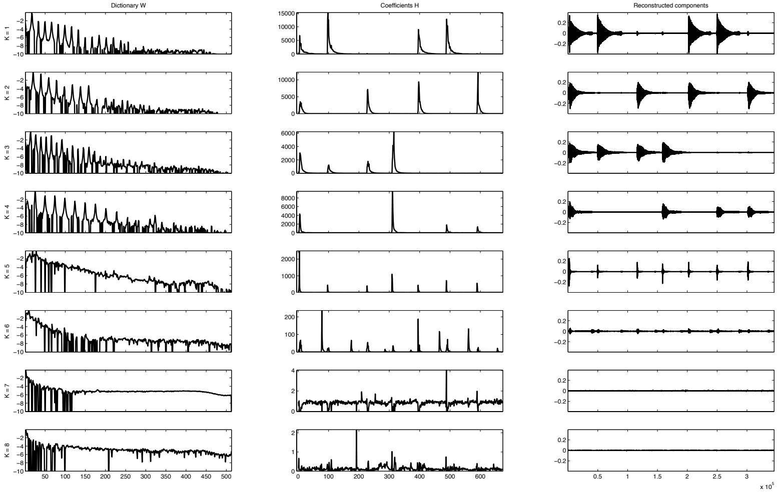


Figure: Three representations of data.

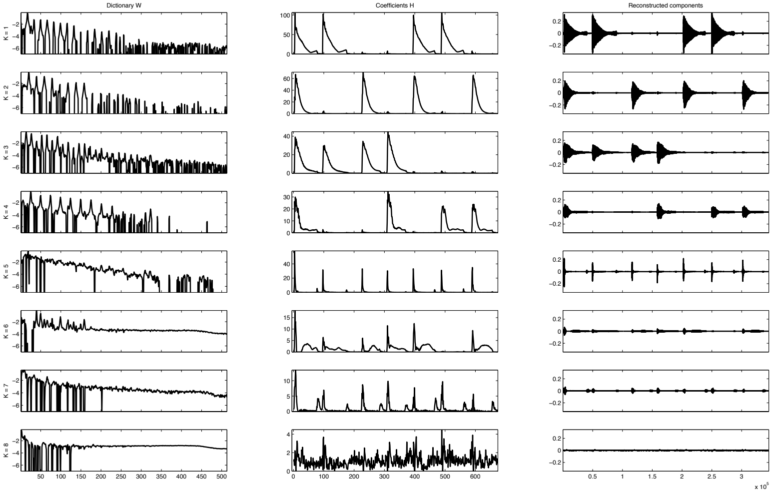
IS-NMF on power spectrogram with $K = 8$



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

Regularized NMF

One usually wants to reflect prior information/expected characteristics about $\mathbf{H}$ or $\mathbf{W}$ in the form of penalty terms.

MM algorithms are easily adapted to penalized NMF, where a term $R(\mathbf{W})$ or $R(\mathbf{H})$ is added to the objective function.

Ex.: regularizer on $\mathbf{H}$

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

Only the minimization step is changed. (Which may become non-tractable, in which case $R(\mathbf{H})$ can be majorized itself.)

Regularized NMF

One usually wants to reflect prior information/expected characteristics about $\mathbf{H}$ or $\mathbf{W}$ in the form of penalty terms.

MM algorithms are easily adapted to penalized NMF, where a term $R(\mathbf{W})$ or $R(\mathbf{H})$ is added to the objective function.

Ex.: regularizer on $\mathbf{H}$

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

Only the minimization step is changed. (Which may become non-tractable, in which case $R(\mathbf{H})$ can be majorized itself.)

Smooth Itakura-Saito NMF

Audio exhibit time persistence/redundancy. Should be taken into account in the factorization for

- more accurate estimation of **H**, and **W**,
- reduced identifiability ambiguities,
- perceptually more pleasant component reconstructions.

$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} C(\mathbf{W},\mathbf{H}) = D_{IS}(\mathbf{V}|\mathbf{W}\mathbf{H}) + \lambda \sum_{n=2}^{N} D(\mathbf{h}_n|\mathbf{h}_{n-1})$$

Smooth Itakura-Saito NMF

Audio exhibit time persistence/redundancy. Should be taken into account in the factorization for

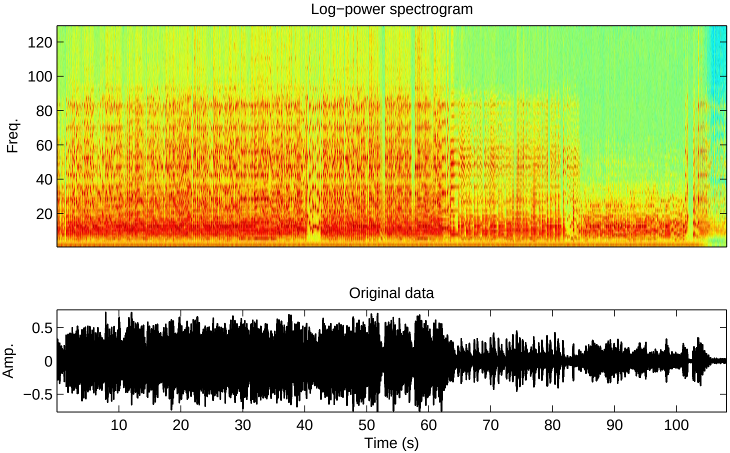
- more accurate estimation of **H**, and **W**,
- reduced identifiability ambiguities,
- perceptually more pleasant component reconstructions.

$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} C(\mathbf{W},\mathbf{H}) = D_{IS}(\mathbf{V}|\mathbf{W}\mathbf{H}) + \lambda \sum_{n=2}^{N} D(\mathbf{h}_n|\mathbf{h}_{n-1})$$

Choosing $D(\mathbf{h}_n|\mathbf{h}_{n-1}) = \sum_k d_{IS}(h_{kn}|h_{k,n-1})$ approximately corresponds to MAP estimation with an inverse-Gamma Markov chain prior on $\{h_{kn}\}_n$. See (Févotte, 2011).

Smooth Itakura-Saito NMF

Louis Armstrong and His Hot Five



Smooth Itakura-Saito NMF

Effect of regularization

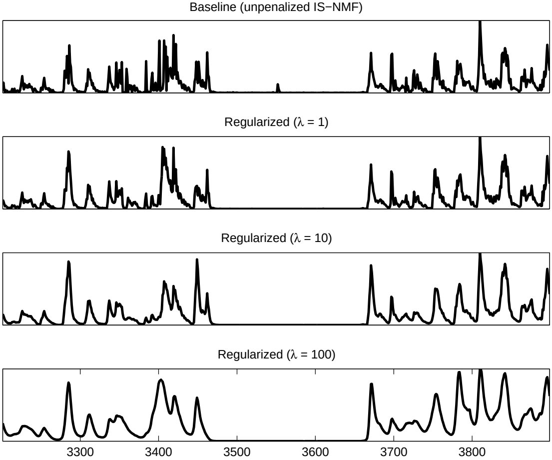
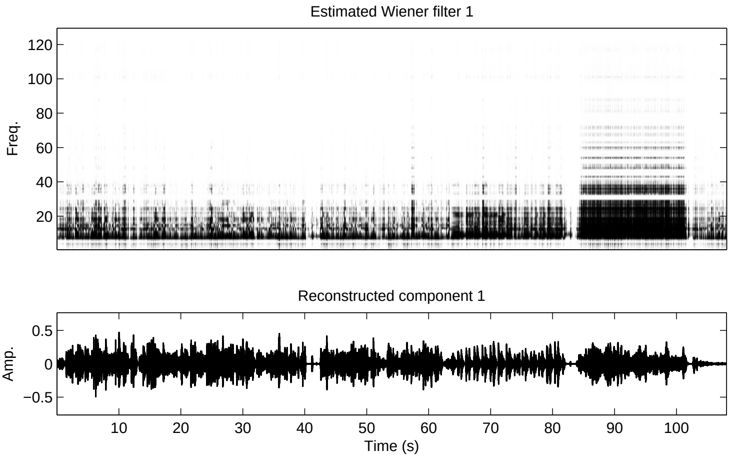


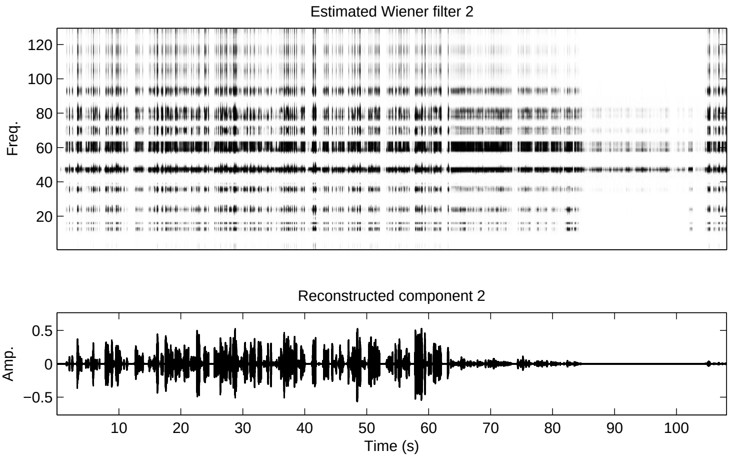
Figure: Segment of one of the rows of **H** for different values of λ.

Smooth Itakura-Saito NMF
Component 1



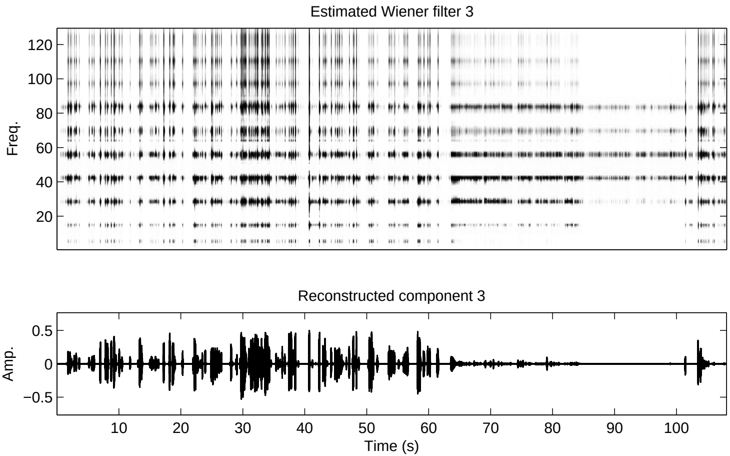
Smooth Itakura-Saito NMF

Component 2



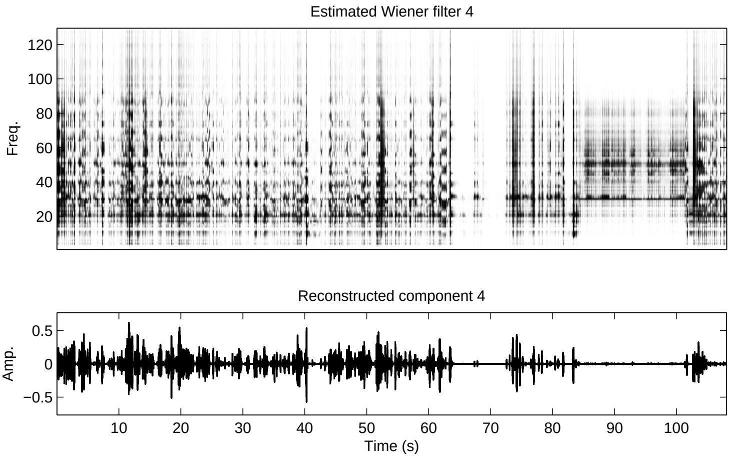
Smooth Itakura-Saito NMF

Component 3



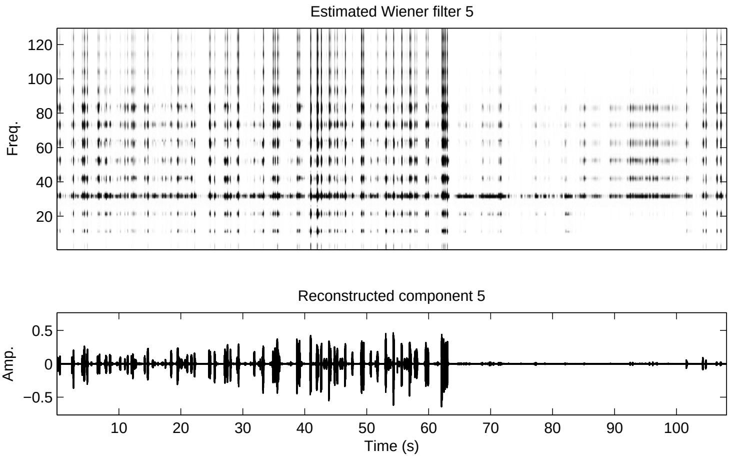
Smooth Itakura-Saito NMF

Component 4



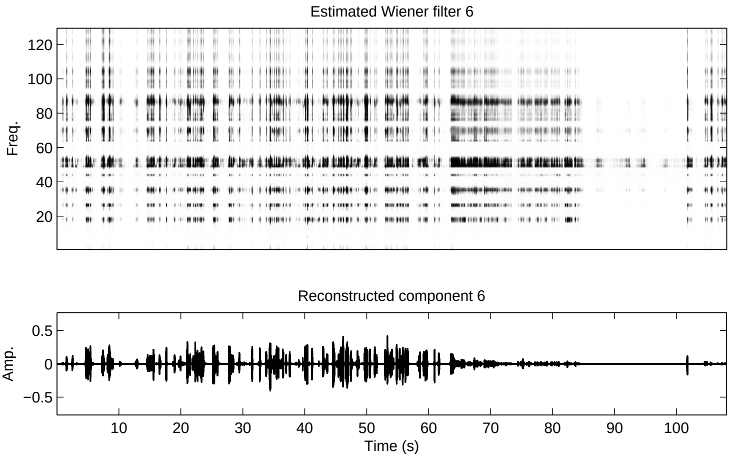
Smooth Itakura-Saito NMF

Component 5



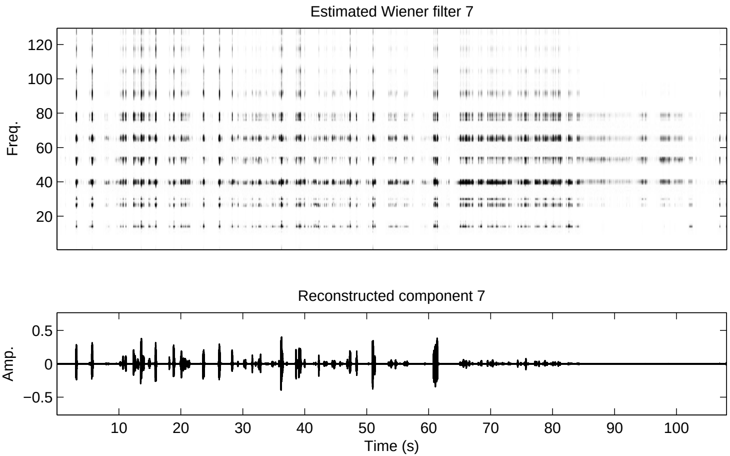
Smooth Itakura-Saito NMF

Component 6



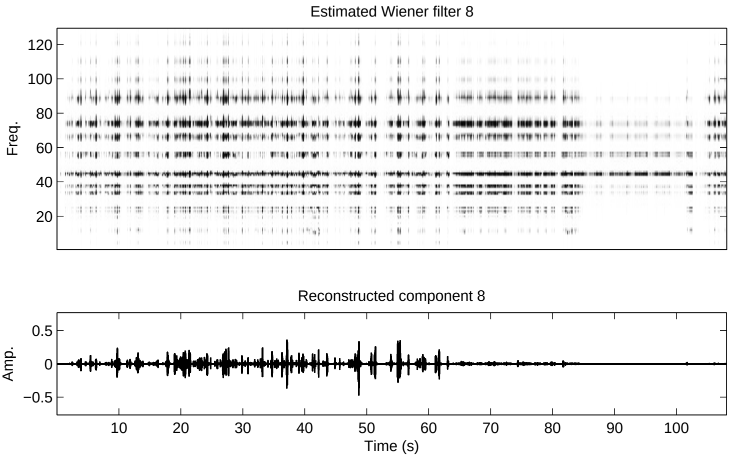
Smooth Itakura-Saito NMF

Component 7



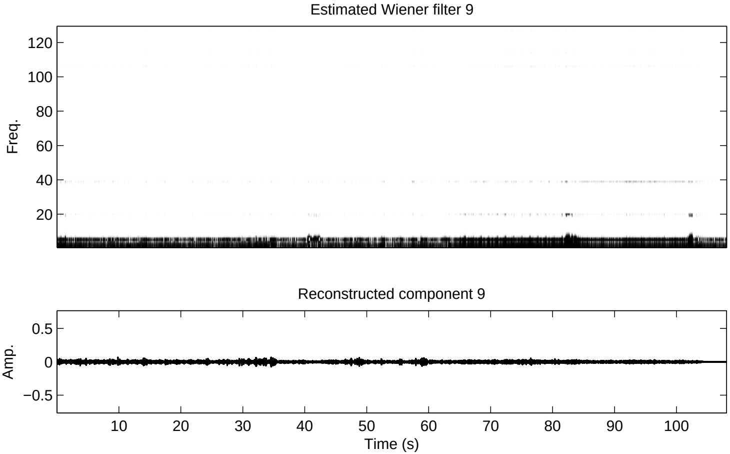
Smooth Itakura-Saito NMF

Component 8



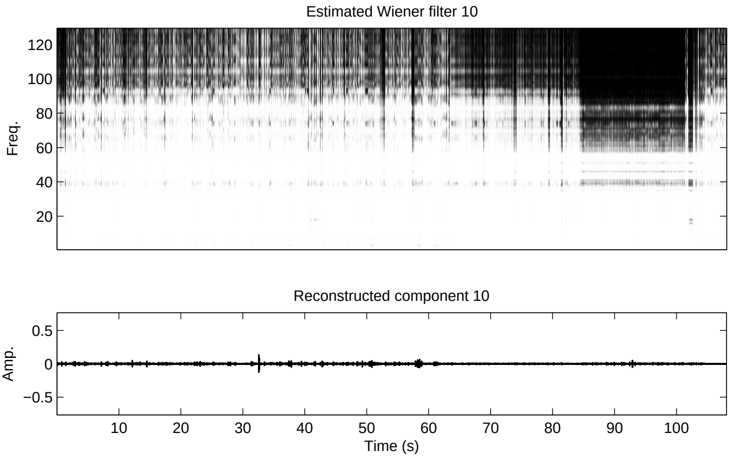
Smooth Itakura-Saito NMF

Component 9



Smooth Itakura-Saito NMF

Component 10



Smooth Itakura-Saito NMF

Full audio restoration example

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

IS-NMF with group sparsity

(Lefèvre, Bach, and Févotte, 2011)

Expected activation structure in music:

$$\mathbf{H} = \begin{bmatrix} \times & \times & \times & \times & 0 & 0 & \times & \times & \times & \times \\ \times & \times & \times & \times & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & \times & \times & 0 & 0 \\ \times & 0 & \times & 0 & \times & 0 & \times & 0 & \times & 0 \\ \times & 0 & \times & 0 & \times & 0 & \times & 0 & \times & 0 \end{bmatrix}$$

Exploit group structure to automate component grouping.

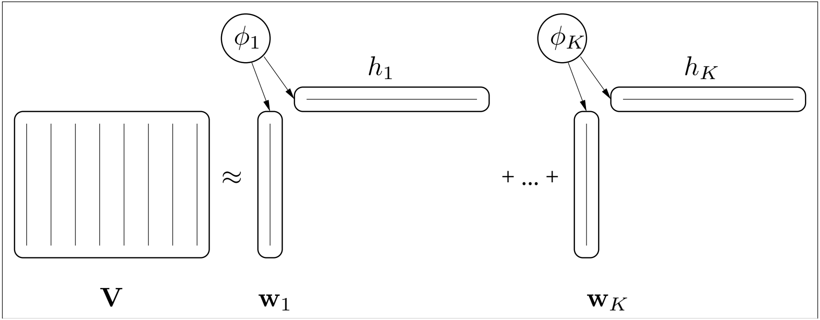
$$\mathbf{h}_n = [\ \underbrace{\times\ \times}_{\text{Source 1}}\ \ \underbrace{\times\ \times}_{\text{Source 2}}\ \ \underbrace{\times\ \times}_{\text{Source 3}}\ ]^T$$

E.g., group sparsity: $R(\mathbf{h}_n) = \|\mathbf{h}_{1,n}\|_2 + \|\mathbf{h}_{2,n}\|_2 + \|\mathbf{h}_{3,n}\|_2$

Automatic relevance determination in NMF

(Tan and Févotte, 2009, in press)

Tie each column $\mathbf{w}_k$ and row h_k through their prior, via a common "relevance" (scale) parameter: $p(\mathbf{w}_k|\Phi_k)$, $p(h_k|\Phi_k)$.



Some relevance parameters converge naturally towards a small constant and the corresponding components are pruned.

Automatic relevance determination in NMF

(Tan and Févotte, 2009, in press)

Under half-normal or exponential priors for $\mathbf{w}_k$, h_k and inverse-Gamma prior for ϕ_k, MAP boils down to minimizing

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

where

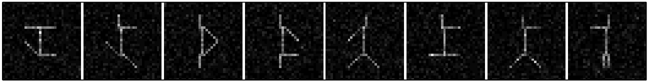
- $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ or $\|\mathbf{x}\|_1$
- b is a sparsity shape parameter

Concave term $\log(x + b)$ induces group-sparsity at the column & row level.
Optimization can be handled in the MM framework.

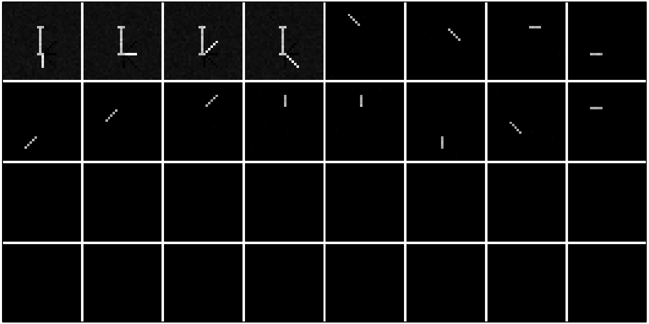
Automatic relevance determination in NMF

Swimmer data decomposition

(a) Noisy data



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



Automatic relevance determination in NMF

Piano data decomposition

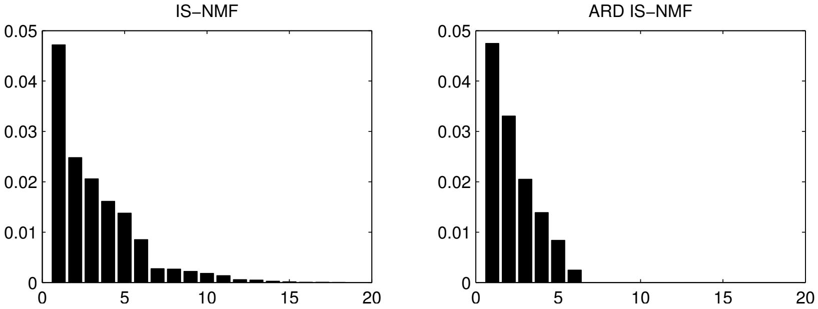
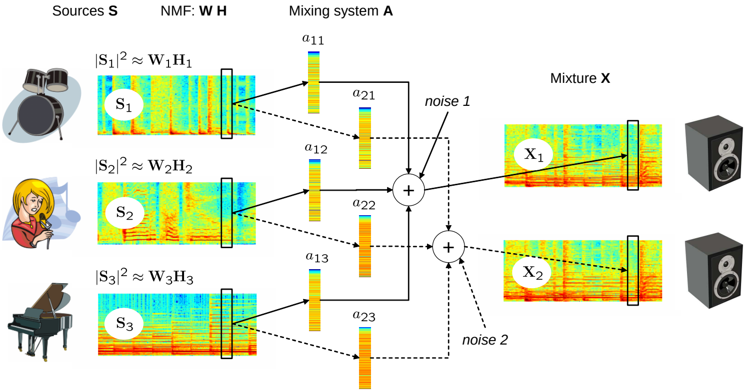


Figure: Standard deviation of reconstructed components with IS-NMF and ARD IS-NMF applied to previously used piano data with $K = 18$. ARD IS-NMF retains the six "ground truth" components only.

Multichannel IS-NMF

(Ozerov and Févotte, 2010)



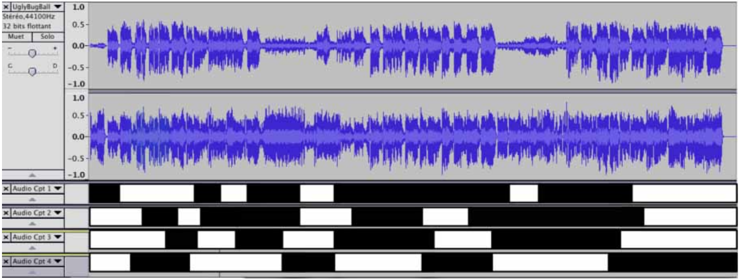
Multichannel NMF problem: Estimate **W**, **H** and **A** from **X**

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

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 **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. Regularized variants.
- Possible extension to multichannel data for audio source separation.
- Multiplicative updates make user-guided separation easy.

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. Regularized variants.
- Possible extension to multichannel data for audio source separation.
- Multiplicative updates make user-guided separation easy.

- 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.

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.

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.

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.

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

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of the National Academy of Sciences*, pages 4164–4169, Mar. 2004.

References II

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.

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.

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.

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 and P. Anh-Huy. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *Proc. IEEE Workshop on Multimedia Signal Processing*, 2002.

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.

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://perso.telecom-paristech.fr/~fevotte/Proceedings/nips11.pdf.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Journal of Mathematical Sciences*, 6(2), June 2007.

References IV

S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical report, Institute of Statistical Mathematics, June 2001. URL http://www.ism.ac.jp/$\sim$eguchi/pdf/Robustify_MLE.pdf. Research Memo. 802.

C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL http://perso.telecom-paristech.fr/~fevotte/Proceedings/icassp11a.pdf.

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.tsi.enst.fr/~fevotte/Proceedings/eusipco09_1.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://arxiv.org/pdf/1010.1763v3.

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.tsi.enst.fr/~fevotte/Journals/neco09_is-nmf.pdf.

References V

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://perso.telecom-paristech.fr/~fevotte/Proceedings/ssp11.pdf.

D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, 24 (15):1722–1728, 2008.

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.

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

D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, 1 (1):38–51, 2008.

H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

H. Laurberg, M. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, Article ID 764206, 2008.

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.

A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL `http://perso.telecom-paristech.fr/~fevotte/Proceedings/icassp11c.pdf`.

References VII

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, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:10–60, 2010.

M. Merritt and Y. Zhang. Interior-point gradient method for large-scale totally nonnegative least squares problems. *Journal of optimization theory and applications*, 126(1):191–202, 2005.

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.

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.tsi.enst.fr/~fevotte/Journals/ieee_asl_multinmf.pdf`.

References VIII

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. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL `http://perso.telecom-paristech.fr/~fevotte/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.

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.

References IX

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.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, St-Malo, France, Apr. 2009. URL `http://www.tsi.enst.fr/~fevotte/Proceedings/spars09.pdf`.

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*, in press. URL `http://arxiv.org/abs/1111.6085`.

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. 26th International ACM SIGIR conference on research and development in information retrieval (SIGIR'03)*, pages 267–273, Toronto, Canada, 2003.

S. S. Young, P. Fogel, and D. Hawkins. Clustering Scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14(1):11–13, June 2006.

References X

R. Zdunek and A. Cichocki. Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8):1904–1916, Aug. 2007.

M. Zetlaoui, M. Feinberg, P. Verger, and S. Clémençon. Extraction of food consumption systems by non-negative matrix factorization (nmf) for the assessment of food choices. Technical report, Arxiv, 2010. URL http://hal.archives-ouvertes.fr/docs/00/48/47/94/PDF/NMF_food.pdf.