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

Cédric Févotte

CNRS LTCI; Télécom ParisTech

ESI, Vienna
December 2012
Outline

Generalities about NMF
  Concept of NMF
  The $\beta$-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
Nonnegative matrix factorization (NMF)

Given a \textit{nonnegative} matrix \( V \) of dimensions \( F \times N \), NMF is the problem of finding a factorization

\[ V \approx WH \]

where \( W \) and \( H \) are \textit{nonnegative} matrices of dimensions \( F \times K \) and \( K \times N \), respectively.

\( K \) is usually chosen such that \( FK + KN \ll FN \), 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

\[ v_n \approx W h_n \]

- \( v_n \) : data vector
- \( W \) : “explanatory variables”
- \( h_n \) : “regressors”
- “basis”, “dictionary”
- “patterns”
- “expansion coefficients”
- “activation coefficients”

and \( W \) is learnt from the set of data vectors \( V = [v_1 \ldots v_N] \).
An unsupervised part-based representation

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

\[
\mathbf{v}_n \approx \mathbf{W} \mathbf{h}_n
\]

- \(\mathbf{v}_n\) data vector
- \(\mathbf{W}\) “explanatory variables”
- “basis”, “dictionary”
- “patterns”
- \(\mathbf{h}_n\) “regressors”
- “expansion coefficients”
- “activation coefficients”

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

- **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.

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 $V$ and model $WH$, subject to nonnegativity of $W$ and $H$:

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

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

Regularization terms are often added to $D(V|WH)$ to favor sparsity or smoothness of $W$ or $H$. 
Trivial scale and order ambiguities between the columns of $W$ and the rows of $H$.

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

- In the exact/noiseless case $V = W^*H^*$, the elements of the true dictionary $W^*$ 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.
A very popular cost function in NMF is the $\beta$-divergence (Basu et al., 1998; Eguchi and Kano, 2001; Cichocki and Amari, 2010), given by

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

which takes the

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

as special cases.
$\beta$-divergence

d(x=1|y)

$\beta = 2$ (Euc)

Cédric Féotte (CNRS)

Itakura-Saito nonnegative matrix factorization
Generalities about NMF

Itakura-Saito NMF Variants

Concept of NMF Measure of fit Algorithms

\[ \beta \text{-divergence} \]

\[ d(x=1|y) \]

\[ \beta = 2 \text{ (Euc)} \]

\[ \beta = 1 \text{ (KL)} \]
$\beta$-divergence

\[ d(x=1|y) \]

- $\beta = 2$ (Euclidean)
- $\beta = 1$ (Kullback-Leibler)
- $\beta = 0$ (Itakura-Saito)

Cédric Févotte (CNRS)
Generalities about NMF

Itakura-Saito NMF Variants

Concept of NMF Measure of fit Algorithms

$\beta$-divergence

\[ d(x=1|y) \]

- $\beta = 2$ (Euc)
- $\beta = 1$ (KL)
- $\beta = 0$ (IS)
- $\beta = -1$

Cédric Févotte (CNRS) Itakura-Saito nonnegative matrix factorization
β-divergence

\[ d(x=1|y) \]

- \( \beta = 2 \) (Euc)
- \( \beta = 1 \) (KL)
- \( \beta = 0 \) (IS)
- \( \beta = -1 \)
- \( \beta = 3 \)

Cédric Févotte (CNRS)

Itakura-Saito nonnegative matrix factorization
Common NMF algorithm design

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

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

- The updates of $W$ and $H$ are equivalent by symmetry:

$$
V \approx WH \iff V^T \approx H^T W^T
$$

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

$$
D(V|WH) = \sum_n D(v_n|Wh_n)
$$
In the end we are left with

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

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 $G(h|\tilde{h})$ such that $G(h|\tilde{h}) \geq C(h)$ and $G(\tilde{h}|\tilde{h}) = C(\tilde{h})$. Optimize (iteratively) $G(h|\tilde{h})$ instead of $C(h)$.
Majorization-minimization (MM)

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

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

![Graph showing objective function $C(h)$ and auxiliary function $G(h|h_{(1)})$.](image-url)
Majorization-minimization (MM)

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

Build $G(h|\tilde{h})$ such that $G(h|\tilde{h}) \geq C(h)$ and $G(\tilde{h}|\tilde{h}) = C(\tilde{h})$. Optimize (iteratively) $G(h|\tilde{h})$ instead of $C(h)$. 
Majorize convex and concave parts separately

\[
C(h) = \underbrace{\tilde{C}(h)}_{\text{Maj. by Jensen’s ineq.}} + \underbrace{\tilde{C}(h)}_{\text{Maj. by tangent}}
\]
Majorization-minimization (MM)

Majorize convex and concave parts separately

\[
C(h) = \underbrace{\tilde{C}(h)}_{\text{Maj. by Jensen’s ineq.}} + \underbrace{\tilde{C}(h)}_{\text{Maj. by tangent}}
\]

In the end, MM for $\beta$-NMF leads to

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

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

Multiplicative form preserves nonnegativity given nonnegative initialization.

Very easy to implement, of complexity $O(FKN)$ 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 $\beta$-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
(Févote, Bertin, and Durrieu, 2009)

Let $X = \{x_{fn}\}$ 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 $W$ and $H$. 
Itakura-Saito NMF: a generative approach
(Févotte, Bertin, and Durrieu, 2009)

Let $X = \{x_{fn}\}$ 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 $W$ and $H$. Then

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

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)
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:

\[
\hat{c}_{k,fn} = \mathbb{E}\{c_{k,fn}|X, W, H\} = \frac{w_{fk}h_{kn}}{\sum_j w_{fj}h_{jn}} x_{fn}
\]

\[
\text{time-freq. mask}
\]

Lossless decomposition: \(x_{fn} = \sum_k \hat{c}_{k,fn}\)
Alternatively, IS-NMF can be interpreted as maximum likelihood in multiplicative noise:

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

where \( \epsilon_{fn} \) is exponential noise.

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,

\[ d_{EUC}(\lambda x|\lambda y) = \lambda^2 d_{EUC}(x|y) \]
\[ d_{KL}(\lambda x|\lambda y) = \lambda d_{KL}(x|y) \]

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)
- Multinominal 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 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

Generalities about NMF

Concept of NMF
The $\beta$-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
Regularized NMF

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

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

Ex.: regularizer on $H$

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

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

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

Only the minimization step is changed. (Which may become non-tractable, in which case $R(\mathbf{H})$ can be majorized itself.)
Audio exhibit time persistence/redundancy. Should be taken into account in the factorization for

- more accurate estimation of $\mathbf{H}$, and $\mathbf{W}$,
- reduced identifiability ambiguities,
- perceptually more pleasant component reconstructions.

\[ \min_{\mathbf{W}, \mathbf{H} \geq 0} C(\mathbf{W}, \mathbf{H}) = D_{IS}(\mathbf{V} | \mathbf{WH}) + \lambda \sum_{n=2}^{N} D(h_n | h_{n-1}) \]
Smooth Itakura-Saito NMF

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

- more accurate estimation of $\mathbf{H}$, and $\mathbf{W}$,
- reduced identifiability ambiguities,
- perceptually more pleasant component reconstructions.

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

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

Log-power spectrogram

Original data

Cédric Féotte (CNRS)
Itakura-Saito nonnegative matrix factorization
Smooth Itakura-Saito NMF

Effect of regularization

Figure: Segment of one of the rows of $H$ for different values of $\lambda$. 
Smooth Itakura-Saito NMF

Component 1

Estimated Wiener filter 1

Reconstructed component 1
Smooth Itakura-Saito NMF
Component 2

Estimated Wiener filter 2

Reconstructed component 2
Smooth Itakura-Saito NMF
Component 3

Estimated Wiener filter 3

Reconstructed component 3
Smooth Itakura-Saito NMF

Component 4

Estimated Wiener filter 4

Reconstructed component 4
Smooth Itakura-Saito NMF

Component 5

Estimated Wiener filter 5

Reconstructed component 5
Smooth Itakura-Saito NMF

Component 6
Smooth Itakura-Saito NMF
Component 7

Estimated Wiener filter 7

Reconstructed component 7
Smooth Itakura-Saito NMF
Component 8

Estimated Wiener filter 8

Reconstructed component 8
Smooth Itakura-Saito NMF

Component 9

Estimated Wiener filter 9

Reconstructed component 9
Smooth Itakura-Saito NMF

Component 10

Estimated Wiener filter 10

Reconstructed component 10
Smooth Itakura-Saito NMF

Full audio restoration example

Original mono =

$\underbrace{\text{Accompaniment}}_{\text{Comp. 1,9}} + \underbrace{\text{Brass}}_{\text{Comp. 2,3,5–8}} + \underbrace{\text{Trombone}}_{\text{Comp. 4}} + \underbrace{\text{Noise}}_{\text{Comp. 10}}$

Original mono denoised

Original denoised & upmixed to stereo
IS-NMF with group sparsity
(Lefèvre, Bach, and Févotte, 2011)

Expected activation structure in music:

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

Exploit group structure to automate component grouping.

\[ \mathbf{h}_n = \begin{bmatrix}
  \times \times \\
  \times \times \\
  \times \times \\
\end{bmatrix}^T \\
\text{Source 1} \quad \text{Source 2} \quad \text{Source 3} \]

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 Féotte, 2009, in press)

Tie each column $w_k$ and row $h_k$ through their prior, via a common “relevance” (scale) parameter: $p(w_k|\Phi_k)$, $p(h_k|\Phi_k)$.

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 $w_k$, $h_k$ and inverse-Gamma prior for $\phi_k$, MAP boils down to minimizing

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

where

- $f(x) = \frac{1}{2} ||x||^2_2$ or $||x||_1$
- $b$ 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) $\ell_1$-ARD decomposition with $K = 32$
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 Féotte, 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)


