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

Generalities about NMF
- Concept of NMF
- Majorization-minimization algorithms

Itakura-Saito NMF
- A statistical model of the STFT
- Piano decomposition example

Multichannel IS-NMF
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.
Nonnegative matrix factorization (NMF)

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

$$V \approx WH$$

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

Dimensions:

- If $W$ tall ($K < F$), NMF produces a low-rank approximation.
- If $W$ 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

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

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

and \(\mathbf{W}\) is learnt from the set of data vectors \(\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_N]\).
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 \( V = [v_1 \ldots v_N] \).

- **nonneg. of** \( W \) ensures *interpretability* of the dictionary
  (features \( w_k \) and data \( v_n \) belong to same space).

- **nonneg. of** \( H \) tends to produce *part-based representations*
  because subtractive combinations are forbidden.

NMF as a constrained minimization problem

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 certain properties of $W$ or $H$ (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)
- $\alpha$-divergence (Cichocki et al., 2008)
- $\beta$-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)
- $\alpha$-divergence (Cichocki et al., 2008)
- $\beta$-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 $H$ given $W^{(i-1)}$ and $W$ given $H^{(i)}$.  

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

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

$$V \approx WH \iff V^T \approx H^T W^T$$
Common NMF algorithm design

- Block-coordinate update of $H$ given $W^{(i-1)}$ and $W$ given $H^{(i)}$.
- The updates of $W$ and $H$ are equivalent by transposition:
  \[ V \approx WH \iff V^T \approx H^T W^T \]
- The objective function is separable in the columns of $H$ or the rows of $W$:
  \[ D(V|WH) = \sum_n D(v_n|Wh_n) \]
Common NMF algorithm design

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

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

- 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 nonnegative linear regression

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

Numerous references in the image restoration literature (Richardson, 1972; Lucy, 1974; Daube-Witherspoon and Muehllehner, 1986)
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 the objective function $C(h)$](image)
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 the objective function and auxiliary function with points $h^{(0)}$ and $h^{(1)}$]
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 and auxiliary function]
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 $C(h)$. 
Local convergence

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

Cédric Févotte (CNRS) 
Itakura-Saito NMF
Application to music signal processing
(Smaragdis and Brown, 2003)

$$V \approx w \times h$$

spectrogram patterns
activation coefficients
Outline

Generalities about NMF
  Concept of NMF
  Majorization-minimization algorithms

Itakura-Saito NMF
  A statistical model of the STFT
  Piano decomposition example

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
(Féotte, 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 $H$. 

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

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

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

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 Gamma multiplicative noise with mean value 1.

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,

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

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

Itakura-Saito NMF
  A statistical model of the STFT
  Piano decomposition example

Multichannel IS-NMF
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 $H$ when a source is silent.
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 $W$ and/or $H$ (Féotte and Cemgil, 2009; Hoffman et al., 2010; Féotte et al., 2011; Dikmen and Féotte, 2011)


References V


