Low-rank time-frequency synthesis

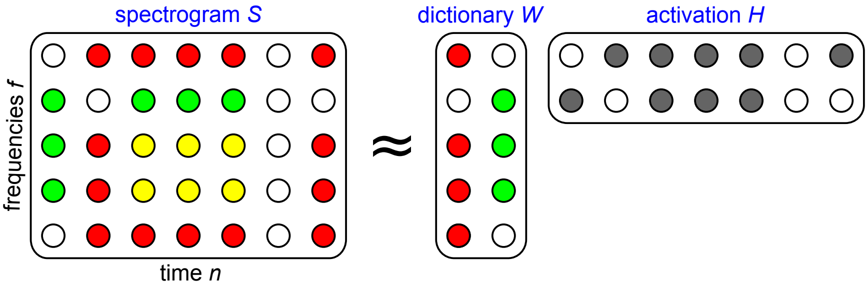
Cédric Févotte[1] & Matthieu Kowalski[2]

[1]Laboratoire Lagrange, Nice
(CNRS, Observatoire de la Côte d'Azur & Université Nice Sophia Antipolis)

[2]Laboratoire des Signaux et Systèmes (LSS), Gif-sur-Yvette
(CNRS, CentraleSupelec & Université Paris-Sud)

SPARS 2015
Cambridge, England

Spectral unmixing by NMF



- $y_{fn} = \sum_t x(t)\phi^\star_{fn}(t)$: short-time Fourier transform (STFT) of temporal signal $x(t)$.
- $s_{fn} = |y_{fn}|^2$: power spectrogram.
- NMF extracts recurring spectral patterns from the data by solving

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \mathrm{D}(\mathbf{S}|\mathbf{WH}).$$

- Successful applications in audio source separation and music transcription.

Itakura-Saito NMF & the Gaussian composite model

(Févotte, Bertin, and Durrieu, 2009)

- Low-rank variance model of analysis coefficients (STFT):

$$y_{fn} \sim N_c(0, [\mathbf{WH}]_{fn}).$$

- Log-likelihood equivalent to Itakura-Saito (IS) divergence:

$$-\log p(\mathbf{Y}|\mathbf{WH}) = D_{\text{IS}}(|\mathbf{Y}|^2|\mathbf{WH}) + \text{cst}.$$

- Underlies a Gaussian composite model (GCM):

$$y_{fn} = \sum_k y_{kfn},$$
$$y_{kfn} \sim N_c(0, w_{fk} h_{kn}).$$

- Given estimates of **W** and **H**, latent STFT components can be estimated by Wiener filter:

$$\hat{y}_{kfn} = \frac{w_{fk} h_{kn}}{[\mathbf{WH}]_{fn}} y_{fn}.$$

- Inverse-STFT of $\{\hat{y}_{kfn}\}_{fn}$ produces temporal components such that $x(t) = \sum_k \hat{c}_k(t)$.

Low-rank time frequency synthesis (LRTFS)

(Févotte and Kowalski, 2014)

- Low-rank variance model of synthesis coefficients:

$$x(t) = \sum_{fn} \alpha_{fn} \, \phi_{fn}(t) + e(t),$$
$$\alpha_{fn} \sim N_c(0, [\mathbf{WH}]_{fn}),$$
$$e(t) \sim N_c(0, \lambda).$$

  - $\phi_{fn}(t)$: time-frequency atom (e.g., from a Gabor frame),
  - α_{fn}: synthesis coefficient,
  - $e(t)$: residual term.
- LRTFS is a generative model of raw data $x(t)$.
- Like in the GCM, the synthesis coefficients have a latent composite structure:

$$\alpha_{fn} = \sum_{k} \alpha_{kfn},$$
$$\alpha_{kfn} \sim N_c(0, w_{fk} h_{kn}).$$

- Given estimates of **W** of **H**, latent coefficients α_{kfn} can be estimated from their posterior mean and temporal components can be reconstructed as $\hat{c}_k(t) = \sum_{fn} \hat{\alpha}_{kfn} \, \phi_{fn}(t)$.

Relation to sparse Bayesian learning (SBL)

- Generative signal model in vector/matrix form:

$$\mathbf{x} = \mathbf{\Phi}\boldsymbol{\alpha} + \mathbf{e}.$$

  - $\mathbf{x}$, $\mathbf{e}$: vectors of signal and residual time samples (size T),
  - $\boldsymbol{\alpha}$: vector of synthesis coefficients α_{fn} (size FN),
  - $\mathbf{\Phi}$: time-frequency dictionary (size $T \times FN$).

- Synthesis coefficients model in vector/matrix form:

$$p(\boldsymbol{\alpha}|\mathbf{v}) = N_c(\boldsymbol{\alpha}|\mathbf{0}, \mathrm{diag}(\mathbf{v})).$$

  - $\mathbf{v}$: vector of variance coefficients $v_{fn} = [\mathbf{WH}]_{fn}$ (size FN).

- Similar to sparse Bayesian learning (Tipping, 2001; Wipf and Rao, 2004) except that the variance parameters are tied together by the low-rank structure $\mathbf{WH}$.

- Optimise

$$
\begin{aligned}
C_{\mathrm{JL}}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{H}, \lambda) &\stackrel{\text{def}}{=} -\log p(\mathbf{x}, \boldsymbol{\alpha} | \mathbf{W}, \mathbf{H}, \lambda) \\
&= \frac{1}{\lambda} \|\mathbf{x} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|_2^2 + D_{\mathrm{IS}}(|\boldsymbol{\alpha}|^2 | \mathbf{v}) + \log(|\boldsymbol{\alpha}|^2) + \mathrm{cst.}
\end{aligned}
$$

- Possible EM algorithm using the procedure of (Figueiredo and Nowak, 2003) based on the hidden variable $\mathbf{z}$ such that

$$
\begin{aligned}
\mathbf{x} &= \boldsymbol{\Phi}\mathbf{z} + \mathbf{e}_1, \\
\mathbf{z} &= \boldsymbol{\alpha} + \sqrt{\beta}\, \mathbf{e}_2,
\end{aligned}
$$

  with $\mathbf{e}_1 \sim N_c(\mathbf{0}, \lambda \mathbf{I} - \beta \boldsymbol{\Phi}\boldsymbol{\Phi}^*)$ and $\mathbf{e}_2 \sim N_c(\mathbf{0}, \mathbf{I})$ (condition applies on β).
- Leads to a form of iterative shrinkage algorithm that scales well with real-world signal dimension.
  Shrinkage operator involves Itakura-Saito NMF of power posterior expectation of $\boldsymbol{\alpha}$ at each iteration.

Estimation in LRTFS

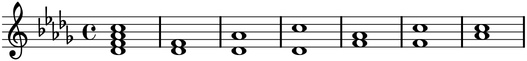
Maximum marginal likelihood (ML)

- Integration of $\boldsymbol{\alpha}$ from the joint likelihood (like in SBL)

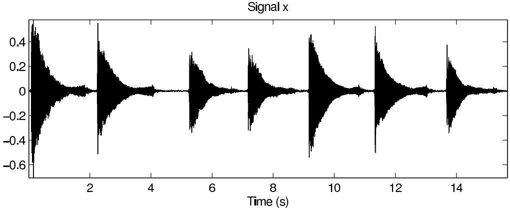
$$
\begin{aligned}
C_{\mathrm{ML}}(\mathbf{W}, \mathbf{H}, \lambda) &\stackrel{\text{def}}{=} -\log p(\mathbf{x}|\mathbf{W}, \mathbf{H}, \lambda) \\
&= -\log \int_{\boldsymbol{\alpha}} p(\mathbf{x}|\boldsymbol{\alpha}, \lambda) p(\boldsymbol{\alpha}|\mathbf{WH}) d\boldsymbol{\alpha}
\end{aligned}
$$

- Possible EM algorithm treating $\boldsymbol{\alpha}$ as the hidden variable.
  Tractable algorithm but does not scale well with dimension.

Toy example



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



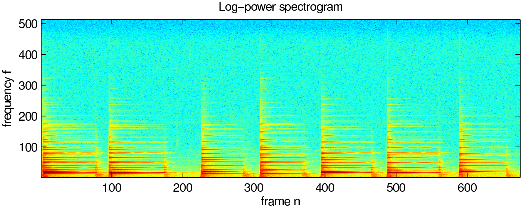
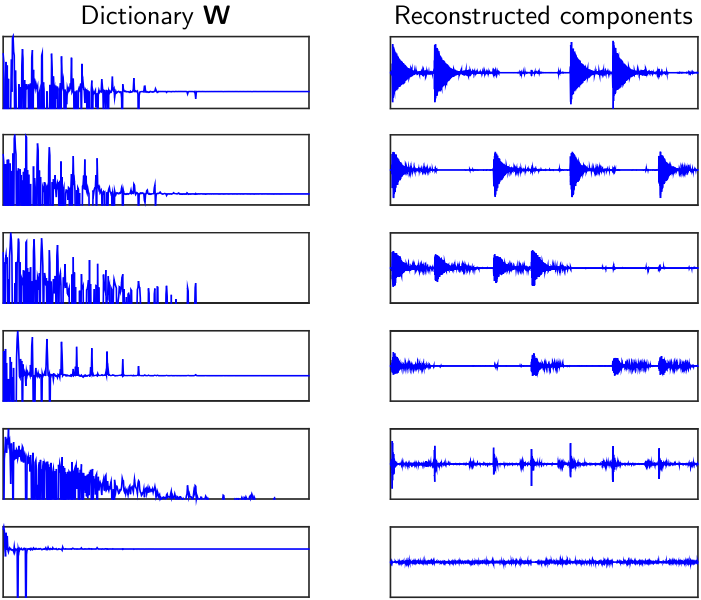


Figure: Three representations of data.

Toy example



Multi-resolution LRTFS

- LRTFS allows for multi-resolution hybrid representations:

$$\mathbf{x} = \mathbf{\Phi}_a \boldsymbol{\alpha}_a + \mathbf{\Phi}_b \boldsymbol{\alpha}_b + \mathbf{e}.$$

  - $\mathbf{\Phi}_a$ and $\mathbf{\Phi}_b$ are time-frequency dictionaries with different resolutions,
  - $\boldsymbol{\alpha}_a$ and $\boldsymbol{\alpha}_b$ have their own latent low-rank structure $\mathbf{W}_a\mathbf{H}_a$ and $\mathbf{W}_b\mathbf{H}_b$.

- Not possible with standard NMF !

- Previous optimisation strategies apply by concatenation:

$$\mathbf{\Phi} = [\mathbf{\Phi}_a \ \mathbf{\Phi}_b]$$

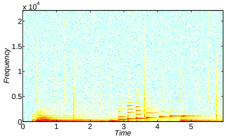
$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_a \\ \boldsymbol{\alpha}_b \end{bmatrix}$$

- Other hybrid decompositions are possible.
  E.g., low-rank layer + sparse layer (forthcoming EUSIPCO paper)

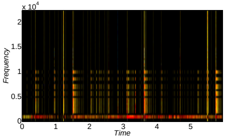
Hybrid decomposition of jazz music

$$\boxed{\mathbf{x} = \mathbf{\Phi}_{\text{tonal}}\,\boldsymbol{\alpha}_{\text{tonal}} + \mathbf{\Phi}_{\text{transient}}\,\boldsymbol{\alpha}_{\text{transient}} + \mathbf{e}}$$

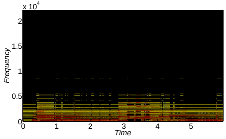
spectrogram of **x**



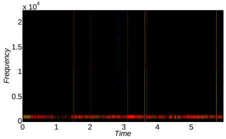
$\log|\hat{\boldsymbol{\alpha}}_{\text{transient}}|$

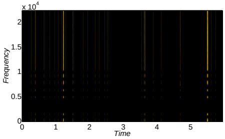


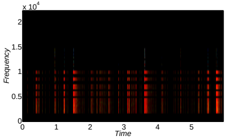
$\log|\hat{\boldsymbol{\alpha}}_{\text{tonal}}|$



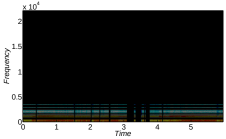
Latent rank-1 components from transient layer

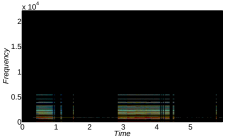


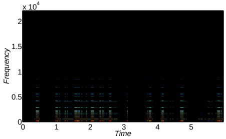




Latent rank-1 components from tonal layer



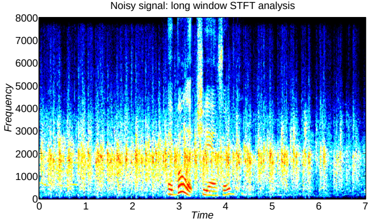


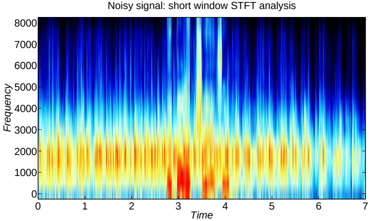


Speech enhancement in applause noise

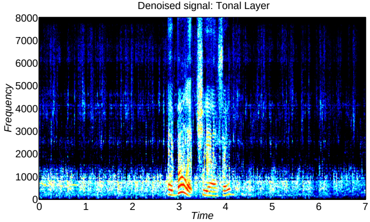
$$\mathbf{x} = \mathbf{\Phi}_{\text{tonal}} \left(\alpha^{\text{speech}}_{\text{tonal}} + \alpha^{\text{noise}}_{\text{tonal}} \right) + \mathbf{\Phi}_{\text{transient}} \left(\alpha^{\text{speech}}_{\text{transient}} + \alpha^{\text{noise}}_{\text{transient}} \right) + \mathbf{e}$$

Spectrograms of x with short and large resolutions

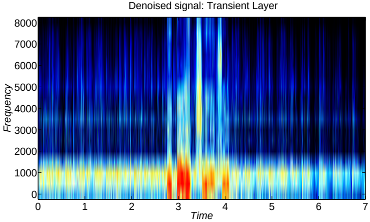




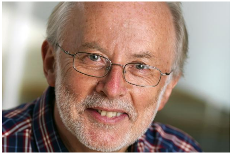
$\log |\hat{\alpha}^{\text{speech}}_{\text{tonal}}|$



$\log |\hat{\alpha}^{\text{speech}}_{\text{transient}}|$



A dedication to Bill Fitzgerald



Prof. William J. Fitzgerald (1948-2014)

An Introduction to
Bayesian Inference Applied to Signal and Data Processing

W J Fitzgerald

Signal Processing Laboratory, Department of Engineering,
University of Cambridge,
Cambridge CB2 1PZ, UK.
wjf@eng.cam.ac.uk

16/5/2001

(my first tutorial on Bayesian inference)

References

C. Févotte and M. Kowalski. Low-rank time-frequency synthesis. *NIPS*, 2014.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 2009.

M. Figueiredo and R. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 2003.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 2001.

D. P. Wipf and B. D. Rao. Sparse Bayesian Learning for Basis Selection. *IEEE Transactions on Signal Processing*, 2004.