Dictionary learning methods and single-channel source separation

Augustin Lefèvre

November 29, 2012

UCL
Université catholique de Louvain

Acknowledgements

Ministere de la Recherche

European Research Council

Willow team

Sierra team
TSI Telecom ParisTech

Acknowledgements II

Host institutions :

Interuniversity Attraction Pole

Pole INMA

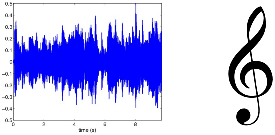
Current collaborators :

Pierre-Antoine Absil

François Glineur

Nicolas Gillis

From raw signals to intelligible information

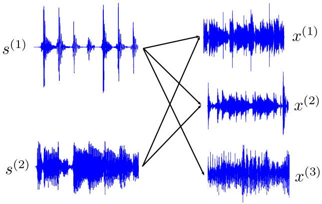


(a) Transcription of polyphonic signals

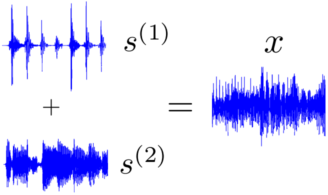
```
Susie kchrr
I'm in the
subway
pffrrrrt
Meet me at
?x%r square
at 9 in front
of pfffrrt
```

(b) Speech recognition in complex environments

What is source separation ?



(c) Overdetermined



(d) Underdetermined

How do we define a source ?

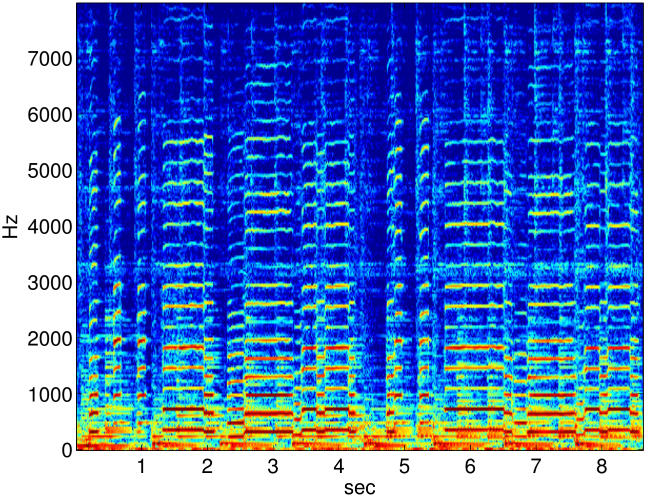
Different sources may sound similar.

How do sources interact ?

Outline

Time-frequency representations

$$
\begin{array}{ccccc}
x \in \mathbb{R}^T & \rightarrow & X \in \mathbb{C}^{F \times N} & \rightarrow & V_{fn} = |X_{fn}|^2 \,, \\
s^{(g)} & \rightarrow & S^{(g)} & \rightarrow & V_{fn}^{(g)} = |S_{fn}^{(g)}|^2 \,.
\end{array}
$$



Nonnegative Matrix Factorization

Reduce the number of unknowns to explain redundancy in the data :

$$V = \underbrace{(W^{(1)} H^{(1)})}_{\hat{V}^{(2)}} + \underbrace{(W^{(2)} H^{(2)})}_{\hat{V}^{(1)}}\,.$$

$W \in \mathbb{R}_+^{F \times K}$ is a dictionary with K basis elements ($K < F$).
$H \in \mathbb{R}_+^{K \times N}$ is a matrix of activation coefficients.
Enforce (pointwise) nonnegativity of the input :

$$W^{(g)} \geq 0, H^{(g)} \geq 0 \Rightarrow \hat{V}^{(g)} \geq 0\,.$$

1) W fixed, H unknown : nonnegative linear model.
2) (W, H) unknown : nonnegative matrix factorization.

(Paatero & Tapper, 1994; Smaragdis & Brown, 2003)

Itakura-Saito NMF

$$\begin{array}{cc} \min_{W,H} & \sum_{fn} d_{IS}(V_{fn}, (WH)_{fn}) \\ \text{s.t.} & W \geq 0, H \geq 0 \end{array}$$

$d_{IS}(x, y) = \frac{x}{y} - \log(\frac{x}{y}) - 1.$
$d_{IS}(x, y) \geq 0.$
$d_{IS}(x, y) = 0 \Rightarrow x = y.$
$d_{IS}(\lambda x, \lambda y) = d_{IS}(x, y)$

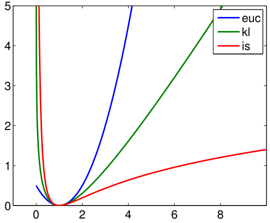


Figure: Plot of $d_{IS}(1, x)$ alongside Kullback-Leibler and Euclidean distance.

Probabilistic interpretation of Itakura-Saito NMF

$V_{\cdot n} \in \mathbb{R}_+^F$ observed power spectrum at time n.

$$V_{fn} = \left| \sum_g S_{fn}^{(g)} \right|^2 \qquad S_{fn}^{(g)} \sim \mathcal{N}_c(0, \mathrm{diag}(\sum_k W_{fk}^{(g)} H_{kn}^{(g)}))\,.$$

(Févotte et al., 2009)

- Phase of spectrograms is assumed uninformative.
- Reconstruct $S^{(g)}$ from $\hat{V}^{(g)}$ and X in a principled way.

$$S_{fn}^{(1)} = \frac{\hat{V}_{fn}^{(1)}}{\hat{V}_{fn}^{(1)} + \hat{V}_{fn}^{(2)}} X_{fn} \qquad \text{keep the same phase as the mixture}$$

- Select the number of components, cheaper than cross-validation.

(Tan & Févotte, 2009; Hoffmann et al., 2010; Lefèvre et al., 2011)

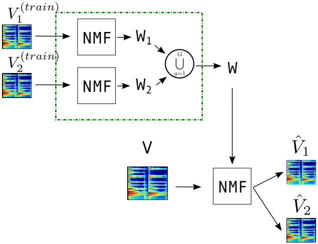
Finding a dictionary

What dictionary should we use ?

1) Ask a physicist to design the dictionary for you.
2) Use a large collection of samples from source 1 and source 2.

Storing all samples from source 1 and source 2 into memory is inconvenient and violates the assumption $K < F$.

Supervised dictionary learning



Having at hand a collection of true source signals decouples learning in two separate problems.

$$
\begin{array}{ll}
\text{Find} & (W, H) \\
s.t. & V^{(g)} = W^{(g)} H^{(g)} \\
 & W \geq 0, H \geq 0
\end{array}
$$

- Combine dictionaries at test time to compute activation coefficients.

Structure

$$\min_H \quad \sum_{fn} \| V_{fn} - (WH)_{fn} \|^2 + \lambda \Psi(H) \, .$$

Few fewer basis elements are used at the same time :
$\Psi(H) = \{$number of nonzero coordinates of $H\}$.

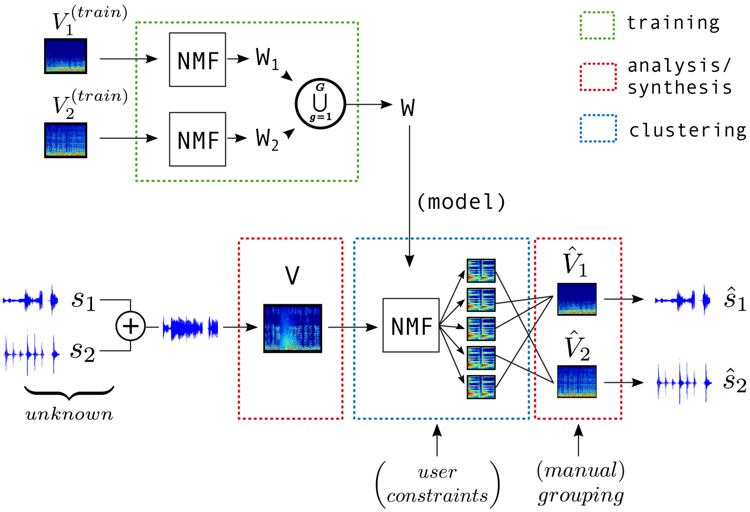
Choice of Ψ reflects assumed structure : temporal continuity at 200ms scale, phonems in speech, etc.

This thesis : Ψ models independence between sources as a group of basis elements.

Assuming simple interactions, we can make weaker assumptions on the dictionary.

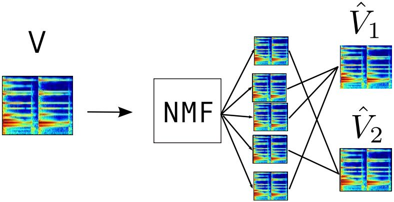
(Hoyer, 2004; Virtanen, 2007; Mysore et al., 2010)

Overview



$V_1^{(train)}$
$V_2^{(train)}$
NMF
NMF
W_1
W_2
$\bigcup_{g=1}^{G}$
W
(model)
training
analysis/
synthesis
clustering
s_1
s_2
unknown
V
NMF
$\hat{V}_1$
$\hat{V}_2$
$\hat{s}_1$
$\hat{s}_2$
(user constraints)
(manual) grouping

Unsupervised learning



If no training data is available to learn $W^{(g)}$ separately, then

$$\begin{array}{ll} \text{Find} & (W, H) \\ s.t. & W^{(1)}H^{(1)} + W^{(2)}H^{(2)} = WH = V\,. \end{array}$$

Not ill-posed any more, but there are still several global optima (nonconvex problem).

Trial and error : find a dictionary that reconstructs the input while enforcing specified structure.

NMF with time-frequency annotations

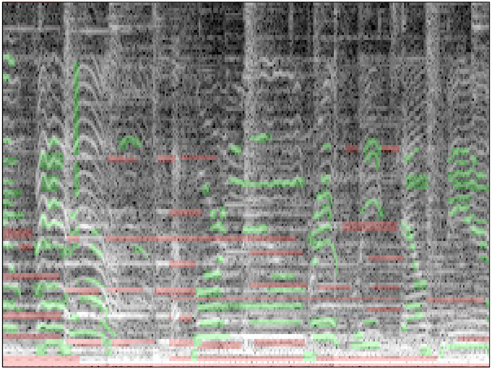


Figure: Example of user annotations in a ten seconds' audio track: green voice. red accompaniment.

NMF with time-frequency annotations

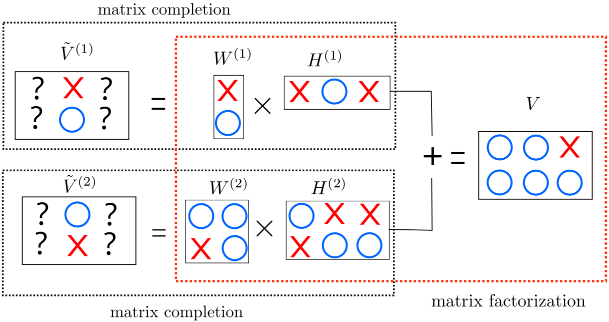


Figure: Semi-supervised NMF consists in solving G matrix completion problems, coupled by a matrix factorization problem.

Robustness to error via relaxation of the constraints (tuning parameter)
Allow "soft" annotations : $M_{fn}^{(g)} \in [0, 1]$.
Discard $M_{fn}^{(g)} = 0.5$.

Towards automatic annotations

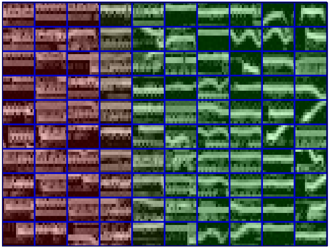


Figure: Time-frequency patches
(green) voice (red) accompaniment

Nearest neighbour.

Quantized nearest-neighbour.

Random Forest.

Experimental results

	% annotated	% correct
track 1	0.23	0.91
track 2	0.10	0.89
track 3	0.29	0.91
track 4	0.17	0.81
track 5	0.22	0.95

Table: Evaluation of user annotations on the SISEC database.

Experimental results

	Track1	
true	accomp	voice
ideal(20%)	15.65	10.34
user(20%)	8.74	3.18
auto	2.44	2.35
baseline	8.20	0.86
lazy	5.07	-5.11

Table: Time-frequency annotations : listening tests

ideal : annotations computed from ground truth (upper-bound).
baseline : NMF with optimally permuted components[1].
auto : automatic annotations.
user : user annotations.
lazy : use $\frac{1}{2}x$ as estimate of each source.

[1]Supposing expert correctly finds best permutation among 10^{18} possibilites ...

New : a convex re-formulation

$$\begin{array}{ll} \min & \|\bar{V} - \sum_{g=1}^{G} V^{(g)}\|_F^2 + \lambda \sum_{g=1}^{G} \|V^{(g)}\|_\star \\ \text{subject to} & M_{fn}^{(g)} V_{fn}^{(g)} = M_{fn}^{(g)} \bar{V}_{fn}^{(g)} \\ & V^{(1)} \geq 0, V^{(2)} \geq 0\,. \end{array} \tag{1}$$

Convex ! So not dependent on initial guess.

Tune the number of components with a continuous parameter λ.

Low-rank (+ nonnegative) solutions of linear underdetermined system of equations.

Nonnegativity constraints are weaker : only source estimates, not dictionary.

New : a convex re-formulation

	SDR	SIR	SAR
nmf	5.9706	13.1744	7.2764
lownuc	7.5785	13.609	9.1102

Table: Average results on 5 audio tracks using 20% of annotations.

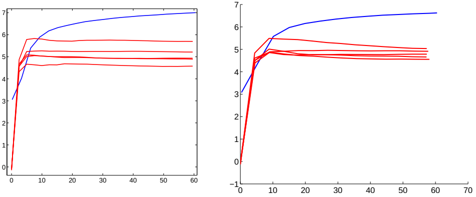


Figure: Source separation quality vs allowed cpu time. (Blue) our method. (Red) NMF.

Convex ! So not dependent on initial guess.

Bottlenecks in NMF

Batch algorithm requires computing and storing matrix-matrix products of the same size as the data set.

Online learning : can't afford to store past data and re-compute activation coefficients.

Large scale learning : $N \to +\infty$, train set is too large to store into memory.

1) Divide-and-conquer strategies (Cao et al., 2007; Mackey et al., 2011).
2) Stochastic updates (Robbins & Monro, 1951).
3) Incremental updates (Neal & Hinton, 1998; Mairal et al., 2010).

On-the-fly updates of the auxiliary function

Batch algorithm works on majorization-minimization

$$\sum_{fn} d_{IS}(V_{fn}, (WH)_{fn}) \leq \sum_{fk} \frac{A_{fk}}{W_{fk}} + B_{fk} W_{fk} \, .$$

H optimized using current estimate $\underline{W}$.

$$\begin{aligned} A_{fk} \leftarrow & \quad \underline{W}_{fk}^2 \textstyle\sum_{n=1}^N V_{fn} (\underline{W} H)_{fn}^{-2} H_{kn} \, , \\ B_{fk} \leftarrow & \qquad \textstyle\sum_{n=1}^N (\underline{W} H)_{fn}^{-1} H_{kn} , \, , \end{aligned}$$

Matrix products in $O(FKN)$ in time and memory.

On-the-fly updates of the auxiliary function

Batch algorithm works on majorization-minimization

$$\sum_{fn} d_{IS}(V_{fn}, (WH)_{fn}) \leq \sum_{fk} \frac{A_{fk}}{W_{fk}} + B_{fk} W_{fk} \, .$$

Draw v at random from V. h optimized using $\underline{W}$.

$$\begin{aligned} A_{fk} &\leftarrow A_{fk} + \underline{W}_{fk}^2 v_f (\underline{W} h)_f^{-2} h_k \, , \\ B_{fk} &\leftarrow B_{fk} + (\underline{W} h)_f^{-1} h_k , \end{aligned}$$

Matrix-vector products in $O(FK)$ in time and memory.

After N draws, same overall number of operations $O(FKN)$.

Memory requirements reduced to $O(FK)$.

How much faster ?

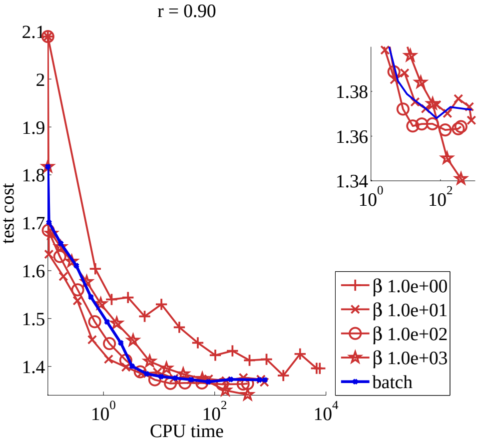


Figure: $N \simeq 10^3$ (30 seconds' excerpt)

How much faster ?

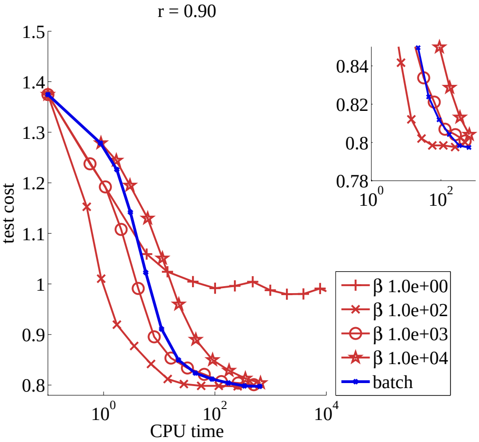


Figure: $N \simeq 10^4$ (4 minutes' audio track)

How much faster ?

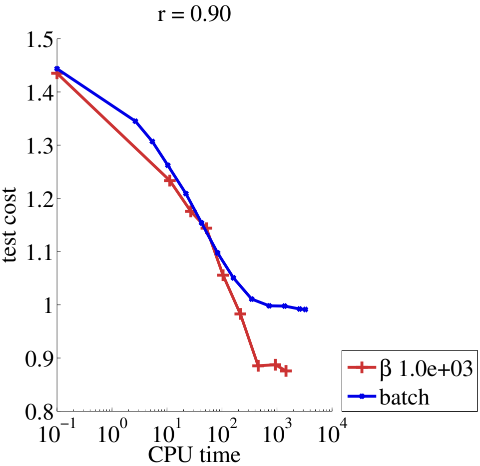


Figure: $N \simeq 10^5$ (1 hour 20 minutes' album)

Conclusion

Machine learning

- "Sensible" solutions to an otherwise underdetermined problem.
- User input gives ideas to design structure.
- Structured decompositions enhance user input.
- Stochastic optimization opens the door to large scale data analysis.

Audio source separation

- Dictionary learning does not replace expert knowledge, it enhances it.
- Audio analysis on larger units : CD, audio collections, and beyond.

Perpectives

Nonnegative decoding in a finite number of iterations.

Automatic annotations using harmonic structure of sound signals (multipitch).

Find other ways to exploit sparsity of time-frequency images.

Audio collections are naturally structured in graphs : we should use that !

Selected Publications I

Adler, Amir, Emiya, Valentin, Jafari, G. Maria, Elad, Michael, Gribonval, Rémi, and Plumbley, Mark D. Audio Inpainting. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

Bach, F. and Jordan, M.I. Blind one-microphone speech separation: A spectral learning approach. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Bengio, S., Pereira, F., Singer, Y., and Strelow, D. Group sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Bertalmío, M., Sapiro, G., Caselles, V., and Ballester, C. Image inpainting. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2000.

Bucak, S. and Gunsel, B. Incremental subspace learning via non-negative matrix factorization. *Pattern Recognition*, 2009.

Cao, B., Shen, D., Sun, J.T., Yang, X., and Chen, Z. Detect and track latent factors with online nonnegative matrix factorization. In *International Joint Conference on Artifical Intelligence (IJCA)*, 2007.

Selected Publications II

Cappé, O., Févotte, C., and Rohde, D. Algorithme em en ligne simulé pour la factorisation non-négative probabiliste. In *Colloque du GRETSI*, 2011.

Daudet, L. Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Transactions on Audio Speech and Language Processing*, 2006.

Duan, Z., Mysore, G.J., and Smaragdis, P. Online PLCA for real-time semi-supervised source separation. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012.

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

Ganseman, J., Scheunders, P., and Dixon, S. Improving plca-based score-informed source separation with invertible constant-q transforms. In *European Signal Processing Conference (EUSIPCO)*, 2012.

Hoffmann, M.D., Blei, D.M., and Cook, P. Bayesian nonparametric matrix factorization for recorded music. In *International Conference on Machine Learning (ICML)*, 2010.

Selected Publications III

Hoyer, P.O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 2004.

Jenatton, R., Audibert, J.-Y., and Bach, F. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 2011.

Lagrange, M., Martins, L.G., Murdoch, J., and Tzanetakis, G. Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

Lefèvre, A., Bach, F., and Févotte, C. Itakura-Saito nonnegative matrix factorization with group sparsity. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2011.

Mackey, L., Talwalkar, A., and Jordan, M.I. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2010.

Selected Publications IV

Mysore, G., Smaragdis, P., and Raj, B. Non-negative hidden markov modeling of audio with application to source separation. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010.

Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, 1998.

Paatero, P. and Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 1994.

Raphael, C. and Han, Y. A classifier-based approach to score-guided music audio source separation. *Computer Music Journal*, 2008.

Recht, B., Fazel, M., and Parrilo, P.A. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review.*, 2010.

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 1951.

Selected Publications V

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

Sprechmann, P., Ramirez, I., Cancela, P., and Sapiro, G. Collaborative sources identification in mixed signals via hierarchical sparse modeling. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2011.

Srebro, N. and Jaakkola, T. Weighted low-rank approximations. In *International Conference on Machine Learning (ICML)*, 2003.

Tan, V.Y.F and Févotte, C. Automatic relevance determination in nonnegative matrix factorization. In *Workshop on Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Tibshirani, R. Regression shrinkage and selection via the lasso. j. *Journal of the Royal Statistical Society : series B*, 1996.

Virtanen, T.O. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio Speech and Language Processing*, 2007.

Selected Publications VI

Wang, D., Vipperla, R., and Evans, N. Online pattern learning for convolutive non-negative sparse coding. In *Interspeech*, 2011.

Yu, G. and Slotine, J.J. Audio classification from time-frequency texture. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2009.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : series B*, 2006.

Zhang, Y. and Scordilis, M.S. Effective online unsupervised adaptation of gaussian mixture models and its application to speech classification. *Pattern Recognition*, 2008.