TELECOM Paris

IP PARIS

Reading note on NMF - TSIA206

Mathis Hardion

June 2023

I. Introduction

This note discusses the methodology and implications of [1], and tentative directions of improvement. It is organized as follows: section II summarizes the main approaches of the paper and how they differ from the methods seen in the course, section III analyses it through a critical lens, and section IV gives possible directions of extension.

II. Summary of the paper

The paper addresses the limits of previous NMF methods for (possibly underdetermined) blind source separation in the multichannel case, namely that they assumed instantaneous mixing, required extra work to rebuild the sources and did not optimally exploit the redundancies between channels.

II.1 Models

The paper considers I convolutive mixtures $\tilde{x}^i(t)$ of J source signals $\tilde{s}^j(t)$ with additive noise $\tilde{b}^i(t)$. The usual narrow-band assumption is made, giving us the mixture model (1) in the STFT domain with F frequency bins and N time frames, where $\mathbf{x}_{f,n} \in \mathbb{C}^I, \mathbf{s}_{f,n} \in \mathbb{C}^J$ and $\mathbf{b}_{f,n} \in \mathbb{C}^I$ contains the STFT coefficients of the corresponding signals at frequency f and window n, and $\mathbf{A}_f \in \mathbb{C}^{I\times J}$ is the mixing matrix at frequency f.

$$\mathbf{x}_{f,n} = \mathbf{A}_f \mathbf{s}_{f,n} + \mathbf{b}_{f,n} \tag{1}$$

The sources are modelled in the frequency domain as a sum of latent components, i.e. the authors assume the existence of $K \geq J$, $(\mathcal{K}_j)_{1\leq j\leq J}$ partitioning $\mathcal{K} \coloneqq \{1,\dots,K\}$, and non-negative $\left(\gamma^k_{f,n}\right)_{k\in\mathcal{K},f,n}$ such that for $j \in \{1,\dots,J\}$,

$$s^j_{f,n} = \sum_{k\in\mathcal{K}_|} c^k_{f,n} \text{ where } c^k_{f,n} \sim \mathcal{N}_c\left(0, \gamma^k_{f,n}\right). \tag{2}$$

This is different from the approach seen in the NMF Lab where the sources to separate were directly the components, whereas here there are J NMF tasks to carry out in order to separate the components of each source. Said differently, the NMF method we have used can be seen as a particular case of this model with $J = 1$ and where we want to retrieve the components c^k. The paper assumes the $c^k_{f,n}$ to be independant over k, f and n which gives

$$s^j_{f,n} \sim \mathcal{N}_c\left(0, \sum_{k\in\mathcal{K}_j} \gamma^k_{f,n}\right). \tag{3}$$

With the STFT matrix of source j $\mathbf{S}_j \coloneqq \left(s^j_{f,n}\right)_{f,n} \in \mathbb{C}^{F\times N}$, the power spectrogram $|\mathbf{S}_j|^2$ verifies under these assumptions

$$\mathbb{E}\left[|\mathbf{S}_j|^2\right] = \left(\sum_{k\in\mathcal{K}_j} \gamma^k_{f,n}\right)_{f,n} \tag{4}$$

which motivates the factorization $\gamma^k_{f,n} = w_{f,k}h_{k,n}$ with $w_{f,k}, h_{k,n} \in \mathbb{R}_+$ allowing us to write the expected power spectrogram as a product of nonnegative matrices

$$\mathbb{E}\left[|\mathbf{S}_j|^2\right] = \mathbf{W}_j\mathbf{H}_j \tag{5}$$

where we denote $\mathbf{W}_j \coloneqq (w_{f,k})_{f,k\in\mathcal{K}_j} \in {\mathbb{R}_+}^{F\times|\mathcal{K}_j|}$ and $\mathbf{H}_j \coloneqq (h_{k,n})_{k\in\mathcal{K}_j,n} \in {\mathbb{R}_+}^{|\mathcal{K}_j|\times N}$. For these parameters, the Maximum Likelihood estimation approach is shown in [2] to be equivalent to the IS-NMF of the spectrogram. The noise is assumed to be stationary, gaussian, spatially and temporally uncorrelated. Under the previous assumptions, the source model is rewritten

$$\mathbf{s}_{f,n} = \mathbf{U}\mathbf{c}_{f,n} \tag{6}$$

with $\mathbf{U} \coloneqq \left(\mathbb{1}_{\mathcal{K}_j}(k)\right)_{k,j} \in \mathbb{R}^{J\times K}$ and $\mathbf{c}_{f,n} \coloneqq \left(c^k_{f,n}\right)_k \in \mathbb{C}^K$. The final model considered is then

$$\mathbf{x}_{f,n} = \mathring{\mathbf{A}}_f\mathbf{c}_{f,n} + \mathbf{b}_{f,n} \tag{7}$$

with the *augmented mixing matrix* $\mathring{\mathbf{A}}_f \coloneqq \mathbf{A}_f\mathbf{U} \in \mathbb{C}^{I\times K}$. The paper highlights a significant advantage of this model in comparison with Gaussian Mixture Models, being that the derived algorithms have complexity linear in the number of components while it grows combinatorially for GMMs.

II.2 Algorithms

The authors present 2 main algorithms, generalizing the usual approaches of Expectation Maximization and Multiplicative Update.

II.2.1 Expectation Maximization

The first approach is to maximize the exact likelihood of the parameters with EM. Writing $p^j_{f,n} \coloneqq \mathbb{E}\left[\left|s^j_{f,n}\right|^2\right]$, $\mathbf{\Sigma}_{\mathbf{s};f,n} = \operatorname{diag}\left(\left(p^j_{f,n}\right)_j\right)$, $\mathbf{\Sigma}_{\mathbf{b};f,n} = \operatorname{diag}\left(\left(\sigma^2_{i,f}\right)_i\right)$ and $\mathbf{\Sigma}_{\mathbf{x};f,n} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s};f,n}\mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b};f,n}$ the covariances of the corresponding signals, the authors derive that ML is equivalent to minimizing the criterion given by (8), where θ contains all the parameters of the model.

$$C_1(\theta) = \sum_{\substack{1\le f\le F\\ 1\le n\le N}} \operatorname{tr}\left(\mathbf{x}_{f,n}\mathbf{x}^H_{f,n}\mathbf{\Sigma}^{-1}_{\mathbf{x};f,n}\right) + \ln\det\mathbf{\Sigma}_{\mathbf{x};f,n} \tag{8}$$

This criterion is invariant under permutation, phase and scaling which creates ambiguities the authors solve by enforcing $\sum_i |a_{i,j;f}|^2 = 1$, $a_{1,j;f} \in \mathbb{R}_+$ and $\sum_f w_{f,k} = 1$. The criterion is then maximized by an EM algorithm adapted for the model. To help with convergence, the authors propose a "simulated annealing" approach consisting in starting with large values of $\sigma^2_{i,f}$ and decreasing them gradually at each iteration, as well as the possibility of adding extra noise at each iteration. The sources are then reconstructed using Wiener filtering followed by inverse STFT.

II.2.2 Multiplicative Updates

The second algorithm presented attempts to maximize the sum of per-channel log-likelihoods, effectively ignoring mutual information between them. This is shown to amount to minimizing criterion (9).

$$C_2(\theta) = \sum_{\substack{1\le f\le F\\ 1\le n\le N\\ 1\le i\le I}} d_{\mathrm{IS}}\left(|x_{i,j}|^2 \middle\| \hat{v}_{i;f,n}\right) \tag{9}$$

With $d_{\mathrm{IS}}(.\|.)$ the IS divergence and

$$\hat{v}_{i;f,n} = \sum_{1\le j\le J} |a_{i,j;f}|^2 \sum_{k\in\mathcal{K}_j} w_{f,k}h_{k,n} \tag{10}$$

Where an additive variance term is ignored due to not being necessary for convergence. $\hat{v}_{i;f,n}$ corresponds to the mixing of the source variances. The authors then derive from (9) the gradients which can be split into a difference between two nonnegative terms, allowing use of the MU scheme. The expressions are altogether rather similar to what we have derived in the course, but involve the vectors $\mathbf{q}_{i,j} = \left(|a_{i,j;f}|^2\right)_f \in \mathbb{R}^F_+$. The sources are then reconstructed with Wiener filtering and ISTFT as usual.

II.3 Experiments

The paper uses multiple criteria to assess the quality of the separation, namely the Signal to Distortion Ratio (SDR) for the source estimates and the Mixing Error Ratio (MER) for the mixing system estimate. These quantitative criteria are not available for all datasets, in which case the criteria were informal through listening to the separations. Two initialization schemes are proposed. First, a "perturbed oracle" method consisting in using a prior method of source separation and adding noise to get the initial parameters. Second, a single-channel NMF decomposition followed by K-means clustering of filters to define the partition $(\mathcal{K}_j)$. The experiments made by the authors underline better SDR performance of the EM algorithm which keeps mutual information between channels. They compare their methods with the (then) state of the art and achieve better SDR with the EM algorithm, but not with the MU scheme. The EM algorithm is however about 4 times slower than the MU one (80min/1000iterations vs 20min/1000iterations). Over the different datasets used, the paper draws the conclusion that the model is more adapted to music than to speech.

II.4 Conclusions

The authors present some possible extensions, such as Bayesian methods for the parameters, clustering methods to automatically infer the hyperparameters (namely the partitioning of $\mathcal{K}$), or Markov chain methods to smoothen the EM estimation. They also give new possible directions to explore which are not addressed by their model, such as nonpoint sources, nonlinear audio effects and more, which are relevant in modern professionally produced music.

III. Critical analysis

III.1 Pros

The paper is thorough on both the theoretical and experimental aspects and presents a more general framework than its predecessors. The presented method of EM manages to improve the results of previous algorithms and obtain a more refined decomposition. At the time of making this note it has been cited 425 times, including recent papers such as [3], which shows there are expansions on the framework of Multichannel NMF (MNMF) to this day. Despite being 13 years old, this paper is still considered state-of-the-art by some researchers [4]. The authors themselves have furthered MNMF following [1], for instance in [5], and especially C. Févotte is still active in the domain of NMF [6]. Overall, this paper appears to have been quite influential in the field of audio processing and particularly source separation and applications.

III.2 Cons

The presented algorithms are still initialization-dependent and strongly rely on previous methods to get the first parameter values. In this sense, it may be unfair to compare the performance of their algorithm with the one used to initialize it, which did not benefit from such an informed headstart. The main algorithm retained, that being the EM algorithm, is also very computationally expensive and while the task is not concerned with real-time, it might not be suited to widely

used audio editing software for instance. Some of the evaluations on the last datasets seemed quite subjective, which can be understood considering the lack of quantitative criteria in this case, but the personal perception of the authors are not that strong of an argument in favor of the implemented algorithm in comparison to others.

IV. Extensions

IV.1 Generalizing the source model

A first possibility of improvement would be to account for a possible correlation of the components $c_{f,n}^k$ over k. If we come back to (2), we could instead write:

$$\mathbb{E}\left[\left|s_{f,n}^j\right|^2\right] = \sum_{k_1\in\mathcal{K}_j}\sum_{k_2\in\mathcal{K}_j}\mathrm{Cov}\left(c_{f,n}^{k_1}, c_{f,n}^{k_2}\right) \tag{11}$$

Then, factorizing $\mathrm{Cov}\left(c_{f,n}^k, c_{f,n}^{k'}\right) = w_{f,k_1}^j \ell_{k_1,k_2}^j h_{k_2,n}^j$, we could write with $\mathbf{W}_j, \mathbf{L}_j, \mathbf{H}_j$ the corresponding matrices:

$$\mathbb{E}\left[|\mathbf{S}_j|^2\right] = \mathbf{W}_j\mathbf{L}_j\mathbf{H}_j. \tag{12}$$

Then, the model presented in [1] is a particular case of (12) with $\mathbf{L}_j = \mathbf{I}_{|\mathcal{K}_j|}$ where $\mathbf{I}_d$ denotes the identity matrix of order d. To remain in the NMF framework, we would need to add the assumption that all the covariances are real and nonnegative. This slightly more general model would lead to 2 NMF tasks per source: the factorization of the spectrogram into $\mathbf{M}_j\mathbf{H}_j$ where $\mathbf{M}_j \in \mathbb{R}^{F\times|\mathcal{K}_j|}, \mathbf{H}_j \in \mathbb{R}^{|\mathcal{K}_j|\times N}$, then the factorization of $\mathbf{M}_j$ into $\mathbf{W}_j\mathbf{L}_j$ where $\mathbf{W}_j \in \mathbb{R}^{|\mathcal{K}_j|}, \mathbf{L}_j \in \mathbb{R}^{|\mathcal{K}_j|\times|\mathcal{K}_j|}$. It would however have the disadvantage to take twice as long as the already computationally expensive methods presented, for the sake of a slight relaxation of the model.

IV.2 Generalizing the optimization criterion

The optimization criteria (8), (9) were derived from a Maximum Likelihood approach leading to the use of the IS divergence, but one could imagine generalizing the algorithms to other β-divergences, considering they are convex for $\beta \in [1, 2]$ as opposed to the IS divergence. We give here the generalization for the criterion of section II.2.2 for simplicity. With the same method and notations as in [1] Appendix B, we obtain

$$\nabla_{\mathbf{q}_{i,j}} D_\beta\left(\mathbf{V}\middle\|\hat{\mathbf{V}}\right) = \left(\hat{\mathbf{V}}_i^{\cdot(\beta-1)}\mathbf{P}_j - \hat{\mathbf{V}}_i^{\cdot(\beta-2)}\cdot\mathbf{V}_i\cdot\mathbf{P}_j\right)\mathbb{1}_{N\times 1}$$

$$\nabla_{\mathbf{W}_j} D_\beta\left(\mathbf{V}\middle\|\hat{\mathbf{V}}\right) = \sum_{i=1}^{I}\mathrm{diag}\left(\mathbf{q}_{i,j}\right)\left(\hat{\mathbf{V}}_i^{\cdot(\beta-1)} - \hat{\mathbf{V}}_i^{\cdot(\beta-2)}\cdot\mathbf{V}_i\right)\mathbf{H}_j^T$$

$$\nabla_{\mathbf{H}_j} D_\beta\left(\mathbf{V}\middle\|\hat{\mathbf{V}}\right) = \sum_{i=1}^{I}\left(\mathrm{diag}\left(\mathbf{q}_{i,j}\right)\mathbf{H}_j\right)^T\left(\hat{\mathbf{V}}_i^{\cdot(\beta-1)} - \hat{\mathbf{V}}_i^{\cdot(\beta-2)}\cdot\mathbf{V}_i\right)$$

which allows generalization of the MU rules.

References

[1] Alexey Ozerov and Cédric Fevotte. "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), pp. 550–563. DOI: 10.1109/TASL.2009.2031510.

[2] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis". In: *Neural Computation* 21.3 (2009), pp. 793–830. DOI: 10.1162/neco.2008.04-08-771.

[3] Kouhei Sekiguchi et al. "Autoregressive Moving Average Jointly-Diagonalizable Spatial Covariance Analysis for Joint Source Separation and Dereverberation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2368–2382. DOI: 10.1109/TASLP.2022.3190734.

[4] Yukoh Wakabayashi, Kouei Yamaoka, and Nobutaka Ono. "Sound Field Interpolation for Rotation-Invariant Multichannel Array Signal Processing". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2286–2298. DOI: 10.1109/TASLP.2023.3282098.

[5] Alexey Ozerov et al. "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 257–260. DOI: 10.1109/ICASSP.2011.5946389.

[6] Arthur Marmin, José Henrique de Morais Goulart, and Cédric Févotte. "Majorization-Minimization for Sparse Nonnegative Matrix Factorization With the β-Divergence". In: *IEEE Transactions on Signal Processing* 71 (2023), pp. 1435–1447. DOI: 10.1109/TSP.2023.3266939.