

# INDEPENDENT VECTOR ANALYSIS WITH MORE MICROPHONES THAN SOURCES

Robin Scheibler and Nobutaka Ono

Tokyo Metropolitan University, Tokyo, Japan

## ABSTRACT

We extend frequency-domain blind source separation based on independent vector analysis to the case where there are more microphones than sources. The signal is modelled as non-Gaussian sources in a Gaussian background. The proposed algorithm is based on a parametrization of the demixing matrix decreasing the number of parameters to estimate. Furthermore, orthogonal constraints between the signal and background subspaces are imposed to regularize the separation. The problem can then be posed as a constrained likelihood maximization. We propose efficient alternating updates guaranteed to converge to a stationary point of the cost function. The performance of the algorithm is assessed on simulated signals. We find that the separation performance is on par with that of the conventional determined algorithm at a fraction of the computational cost.

**Index Terms**— Blind source separation, independent vector analysis, overdetermined, optimization, array signal processing

## 1. INTRODUCTION

We address the problem of blindly separating  $K$  sound sources recorded with  $M$  microphones when  $K < M$ . By far the most popular technique for blind source separation (BSS) is independent component analysis (ICA) which only requires statistical independence of the sources [1]. A convolutive sound mixture is written

$$\hat{x}_m[t] = \sum_{k=1}^K (\hat{a}_{mk} \star \hat{s}_k)[t], \quad (1)$$

where  $\hat{x}_m[t]$  is the  $m$ -th microphone signal,  $\hat{s}_k[t]$  is the  $k$ -th source signal, and  $\hat{a}_{mk}[t]$  is the impulse response between the two. The operator  $\star$  denotes convolution. In the time-frequency domain, convolution becomes frequency-wise multiplication and we have

$$x_{mf n} = \sum_{k=1}^K a_{mk f} s_{k f n}, \quad (2)$$

where  $x_{mf n}$  and  $s_{k f n}$  are the short-time Fourier transforms (STFT) [2] of  $\hat{x}_m[t]$  and  $\hat{s}_k[t]$ , respectively, and  $a_{mk}[f]$  is the discrete Fourier transform of  $\hat{a}_{mk}[t]$ . Finally,  $f = 1, \dots, F$  and  $n = 1, \dots, N$  are the discrete frequency bin and frame indices, respectively. This is an approximation valid when the Fourier transform is sufficiently longer than the impulse response. In this form, the separation problem can be solved by applying ICA to every frequency sub-band independently [3]. Unfortunately, the assignment of output signals to sources in each of the sub-bands is unknown

This work was supported by a JSPS post-doctoral fellowship and grant-in-aid (no. 17F17049), and the SECOM Science and Technology Foundation. The research presented in this paper is reproducible. Code and data are available at <https://github.com/onolab-tmu/overiva>.

and the correct permutation must be recovered. Clustering is a popular solution for permutation alignment [4]. Nevertheless, this extra step is notoriously hard to get right and avoiding it is desirable. Independent vector analysis (IVA) does just that by considering the problem as joint separation over frequencies [5, 6]. The computationally efficient, hyperparameter-free, method for ICA and IVA known as iterative projection [7, 8, 9] forms the basis of our work.

Both for ICA and IVA, the determined case, i.e.,  $K = M$ , is the most straightforward. It allows to do a change of variables and directly maximize the likelihood of the separated signals. In practice, however, using extra microphones adds robustness and increases performance. This is the so-called *overdetermined* case with  $K < M$ . Unfortunately, the aforementioned change of variables cannot be done anymore. A straightforward solution to this problem is to run the algorithm for  $M$  sources, and retain the  $K$  outputs with the largest power. Alternatives to power-based selection exist, for example [10]. Due to the large number of parameters,  $\mathcal{O}(M^2)$ , to estimate, such approaches come with a high computational cost. Ideally, we want to estimate no more than  $\mathcal{O}(KM)$  parameters.

Several methods with better complexities have been proposed. These methods fall broadly in two categories. First, some methods not based on the aforementioned change of variables can directly tackle the overdetermined case [11, 12], but some require regularization [13]. Second, methods that first reduce the number of channels to  $K$  and then apply a determined separation algorithm. This is done for example by selecting the best  $K$  channels [14, 15], or by principal component analysis (PCA) [15, 16, 17]. Nevertheless, these methods inherently risk removing some target signal upfront, irremediably degrading performance. Anecdotaly, a few methods have been proposed for instantaneous mixtures [18, 19], and in the time-domain [20]. All the above methods are single mixture methods that require permutation alignment. Few techniques have been proposed for overdetermined IVA. The single source case, i.e.,  $K = 1$ , known as independent vector extraction (IVE), has been tackled with a gradient ascent method [21].

We propose *OverIVA*, an algorithm to perform IVA with  $K < M$ . The proposed algorithm is hyperparameter-free, guaranteed to converge, and only requires the estimation of  $\mathcal{O}(KM)$  parameters. We derive two variants based on the Laplace and time-varying Gaussian source distributions. The resulting algorithms can be seen as extensions of IVE [21] to more than one source, and with the fast converging updates of AuxIVA [8]. Numerical experiments reveal its separation performance to be comparable to that of full  $M$ -channels IVA at a fraction  $K/M$  of the computational cost. We also find that adding extra microphones fails to improve the performance when using PCA as a pre-processing in diffuse noise.

The rest of this paper is organized as follows. Section 2 describes the hypotheses and signal model. In Section 3, we derive the proposed algorithm. The numerical experiments are discussed in Section 4. Section 5 concludes.

## 2. MODEL

The microphone signals  $\mathbf{x}_{fn} = [x_{1fn}, \dots, x_{Mfn}]^\top \in \mathbb{C}^M$  at frequency  $f$  and time  $n$  is modelled as

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{\Psi}_f \mathbf{z}_{fn}, \quad (3)$$

where  $\mathbf{s}_{fn} = [s_{1fn}, \dots, s_{Kfn}]^\top \in \mathbb{C}^K$  contains the source signals,  $\mathbf{z}_{fn} \in \mathbb{C}^{M-K}$  is a vector of noise, and  $\mathbf{A}_f \in \mathbb{C}^{M \times K}$  and  $\mathbf{\Psi}_f \in \mathbb{C}^{M \times M-K}$  are the respective mixing matrices. Our objective is to estimate the *demixing matrix*  $\widehat{\mathbf{W}}_f \in \mathbb{C}^{M \times M}$  such that the source vector  $\mathbf{s}_{fn}$  is recovered from the measurements

$$\begin{bmatrix} \mathbf{s}_{fn} \\ \mathbf{\Phi}_f \mathbf{z}_{fn} \end{bmatrix} = \widehat{\mathbf{W}}_f \mathbf{x}_{fn}. \quad (4)$$

The matrix  $\mathbf{\Phi}_f$  is an arbitrary invertible linear transformation reflecting that we do not aim at separating the noise components. Indeed, we may even choose  $\mathbf{\Phi}_f$  to simplify the task at hand. Namely, we choose it so that

$$\widehat{\mathbf{W}}_f = \begin{bmatrix} \mathbf{W}_f \\ \mathbf{U}_f \end{bmatrix} \quad \text{with} \quad \begin{aligned} \mathbf{W}_f &= [\mathbf{w}_{1f} \ \cdots \ \mathbf{w}_{Kf}]^H \in \mathbb{R}^{K \times M}, \\ \mathbf{U}_f &= [\mathbf{J}_f \ \ -\mathbf{I}_{M-K}] \in \mathbb{R}^{M-K \times M}, \end{aligned} \quad (5)$$

with  $\mathbf{J}_f \in \mathbb{C}^{M-K \times K}$ . With a slight abuse of notation, we let  $\mathbf{z}_{fn} = \mathbf{U}_f \mathbf{x}_{fn}$ . Following blind source separation principles, we will assume that the target sources have some non-Gaussian distribution. On the other hand, because we do not want to separate the noise components, they are likely to stay mixed and thus their distribution can be assumed close to Gaussian. However, the Gaussianity of the background by itself will turn out to be ineffective at separating the foreground components. We thus rely on orthogonal constraints to further help separation [22, 21]. We formalize this intuition with the following hypothesis.

1. The separated sources are statistically independent

$$\mathbf{s}_{kn} \perp \mathbf{s}_{k'n'}, \quad \forall k \neq k', n, n' \quad (6)$$

where we use the notation  $\mathbf{s}_{kn} \in \mathbb{C}^F$  to mean the vector of frequency components of the  $k$ -th source vector at frame  $n$ . In addition, the separated sources have a *time-varying* circular Gaussian distribution (or Laplace, see Section 3.1)

$$p_{\mathbf{s}}(\mathbf{s}_{kn}) = \frac{1}{\pi^F r_{kn}^F} e^{-\frac{\|\mathbf{s}_{kn}\|^2}{r_{kn}}}, \quad (7)$$

where  $r_{kn}$  is the variance of source  $k$  at time  $n$ .

2. The separated background noise vectors have a *time-invariant* complex Gaussian distribution across microphones

$$p_{\mathbf{z}_f}(\mathbf{z}_{fn}) = \frac{1}{\pi^{M-K} |\det(\mathbf{R}_f)|} e^{-\mathbf{z}_{fn}^H (\mathbf{R}_f)^{-1} \mathbf{z}_{fn}} \quad (8)$$

where  $\mathbf{R}_f$  is the (unknown) spatial covariance matrix of the noise (after separation). Moreover, the separated background noise is statistically independent across frequencies.

3. The sources and background span orthogonal subspaces after separation, namely,

$$\mathbf{0} = \frac{1}{N} \mathbf{Y}_f \mathbf{Z}_f^H = \mathbf{W}_f \mathbf{C}_f \mathbf{U}_f^H, \quad \text{with} \quad \mathbf{C}_f = \frac{1}{N} \mathbf{X}_f \mathbf{X}_f^H, \quad (9)$$

where  $\mathbf{X}_f = [\mathbf{x}_{f,1}, \dots, \mathbf{x}_{f,N}]$ ,  $\mathbf{Y}_f = \mathbf{W}_f \mathbf{X}_f$ , and  $\mathbf{Z}_f = \mathbf{U}_f \mathbf{X}_f$ . The matrix  $\mathbf{C}_f$  is the covariance of the input signal.

Based on these hypothesis, we can write explicitly the likelihood function of the data and find the demixing matrices maximizing it. A few points are in order. We assume the covariance matrix of the noise is rank  $M - K$ . In practice, this means that we will not be able to remove noise that has the same steering vector as one of the sources. Independence of noise across frequencies is a simplifying assumption and is typically not fulfilled. We confirm in the experiment of Section 4 that this does not seem to be a problem. One can also wonder how the algorithm can tell apart sources from noise. While we do not offer a precise analysis, we conjecture that the  $K$  strongest sources have a very non-Gaussian distribution. On the contrary, the mix of the noise and remaining weaker sources will have a distribution closer to Gaussian. As such, we expect the maximum likelihood to choose the strongest sources automatically.

## 3. ALGORITHM

By using (7) and (8), and omitting all constants, we can write the negative log-likelihood of the observed data

$$\begin{aligned} \mathcal{J} = & -2N \sum_f \log |\det(\widehat{\mathbf{W}}_f)| + \sum_{kn} \left( F \log r_{kn} + \frac{\|\mathbf{s}_{kn}\|^2}{r_{kn}} \right) \\ & + \sum_{fn} \left( \log |\det(\mathbf{R}_f)| + \mathbf{z}_{fn}^H (\mathbf{R}_f)^{-1} \mathbf{z}_{fn} \right). \end{aligned} \quad (10)$$

where  $\|\mathbf{s}_{kn}\|^2 = \sum_f |\mathbf{w}_{kf}^H \mathbf{x}_{fn}|^2$ . The first term is due to the change of variables. First, one can show that the gradient of (10) with respect to  $\mathbf{R}_f$  is zero when  $\mathbf{R}_f = \mathbf{U}_f \mathbf{C}_f \mathbf{U}_f^H$ . Furthermore, for this choice of  $\mathbf{R}_f$ , regardless of the choice of  $\mathbf{U}_f$ , we have

$$\sum_n \mathbf{z}_{fn}^H \mathbf{R}_f^{-1} \mathbf{z}_{fn} = \text{tr} \left( \mathbf{R}_f^{-1} \mathbf{Z}_f \mathbf{Z}_f^H \right) = N(M - K). \quad (11)$$

As a consequence, once  $\mathbf{R}_f$  has been fixed, the background part of the cost function can be ignored for the estimation of  $\widehat{\mathbf{W}}_f$ .

The minimization of (10) with respect to  $\mathbf{W}_f$  can be carried out as in AuxIVA [8] via the iterative projection method. Because direct minimization for  $\mathbf{W}_f$  is difficult, this method minimizes (10) alternatively with respect to  $\mathbf{w}_{kf}$ ,  $k = 1, \dots, K$ .

$$\begin{aligned} r_{kn} & \leftarrow \frac{1}{F} \sum_f |\mathbf{w}_{kf}^H \mathbf{x}_{fn}|^2, & \mathbf{V}_{kf} & \leftarrow \frac{1}{N} \sum_n \frac{1}{r_{kn}} \mathbf{x}_{fn} \mathbf{x}_{fn}^H, \\ \mathbf{w}_{kf} & \leftarrow \left( \widehat{\mathbf{W}}_f \mathbf{V}_{kf} \right)^{-1} \mathbf{e}_k, & \mathbf{w}_{kf} & \leftarrow \frac{\mathbf{w}_{kf}}{\left( \mathbf{w}_{kf}^H \mathbf{V}_{kf} \mathbf{w}_{kf} \right)^{-\frac{1}{2}}}. \end{aligned} \quad (12)$$

Once these updates have been applied, we must modify the lower part of the demixing matrix, i.e.  $\mathbf{J}_f$ , so that the noise subspace stays orthogonal. For fixed  $\mathbf{W}_f$ , we can solve (9) for  $\mathbf{J}_f$  and obtain

$$\mathbf{J}_f = \left( \mathbf{E}_2 \mathbf{C}_f \mathbf{W}_f^H \right) \left( \mathbf{E}_1 \mathbf{C}_f \mathbf{W}_f^H \right)^{-1}, \quad (13)$$

where  $\mathbf{E}_1 = [\mathbf{I}_K \ \mathbf{0}_{K \times M-K}]$  and  $\mathbf{E}_2 = [\mathbf{0}_{M-K \times K} \ \mathbf{I}_{M-K}]$ .

The final algorithm applying updates to  $\mathbf{W}_f$  and  $\mathbf{J}_f$  alternatively is detailed in Algorithm 1. Each of the updates from (12) and (13) set the gradient of the cost function to zero with respect to the parameter optimized. Thus, the value of the cost function is non-increasing under these updates. While convergence to a global minimum is not guaranteed, convergence to a stationary point is. Concerning the initial value of  $\mathbf{W}_f$ , we find that a rectangular identity matrix is satisfactory.

**Input** : Microphones signals  $\{\mathbf{x}_{fn}\}$ , # sources  $K$   
**Output**: Separated signals  $\{\mathbf{s}_{fn}\}$   
 $\mathbf{s}_{fn} \leftarrow \mathbf{x}_{fn}, \forall f, n$   
 $\mathbf{W}_f \leftarrow [\mathbf{I}_M \mathbf{0}_{K \times M - K}], \forall f$   
 $\mathbf{J}_f \leftarrow \mathbf{0}_{M-K \times K}, \forall f$   
**for** loop  $\leftarrow 1$  **to** max. iterations **do**  
  **for**  $k \leftarrow 1$  **to**  $K$  **do**  
     $r_{kn} \leftarrow \frac{1}{F} \sum_f |s_{kfn}|^2, \forall n$   
    **for**  $f \leftarrow 1$  **to**  $F$  **do**  
       $\mathbf{V}_{kf} \leftarrow \frac{1}{N} \sum_n \frac{1}{r_{kn}} \mathbf{x}_{fn} \mathbf{x}_{fn}^H$   
       $\mathbf{w}_{kf} \leftarrow (\widehat{\mathbf{W}}_f \mathbf{V}_{kf})^{-1} \mathbf{e}_k$   
       $\mathbf{w}_{kf} \leftarrow \mathbf{w}_{kf} (\mathbf{w}_{kf}^H \mathbf{V}_{kf} \mathbf{w}_{kf})^{-\frac{1}{2}}$   
       $s_{kfn} \leftarrow \mathbf{w}_{kf}^H \mathbf{x}_{fn}, \forall n$   
       $\mathbf{J}_f \leftarrow (\mathbf{E}_2 \mathbf{C}_f \mathbf{W}_f^H) (\mathbf{E}_1 \mathbf{C}_f \mathbf{W}_f^H)^{-1}$   
    **end**  
  **end**  
**end**

Algorithm 1: OverIVA

### 3.1. Laplace overdetermined IVA

The algorithm presented so far assumes a time-varying Gaussian distribution of source vectors. It is possible to change the model to a *time-invariant* circular Laplace distribution as in AuxIVA [8]. Under this new source model, the cost function becomes

$$\mathcal{L} = -2N \sum_f \log |\det(\widehat{\mathbf{W}}_f)| + \sum_{kn} \|\mathbf{s}_{kn}\|_2 + \sum_{fn} \log p_{z_f}(\mathbf{z}_{fn}).$$

Ignoring constants, one can show that this new cost function is majorized by (10) for the specific choice [8]

$$r_{kn} = 2 \sqrt{\sum_f |\mathbf{w}_{kf}^H \mathbf{x}_{fn}|^2}. \quad (14)$$

In this case, OverIVA becomes an auxiliary function based optimization procedure that is still guaranteed to converge to a stationary point.

### 3.2. Computational Complexity

When the number of time frames  $N$  is larger than the number of microphones  $M$ , the runtime is dominated by the computation of the weighted covariance matrix  $\mathbf{V}_{kf}$ . The computational complexity in that case is  $\mathcal{O}(KFM^2N)$ . When the number of microphones is larger, the bottleneck is the matrix inversion with complexity  $\mathcal{O}(KFM^3)$ . The total complexity of the algorithm is thus

$$\mathcal{C}_{\text{OverIVA}} = \mathcal{O}(KFM^2 \max\{M, N\}). \quad (15)$$

The leading  $K$  comes from the number of demixing filters (one per source), and  $F$  is the number of frequency bins. In contrast, conventional AuxIVA needs to update all  $M$  demixing filters, which leads to complexity

$$\mathcal{C}_{\text{AuxIVA}} = \mathcal{O}(FM^3 \max\{M, N\}). \quad (16)$$

The overall complexity is thus reduced by a factor  $K/M$ . This is significant in many practical cases as the number of target sources is rarely larger than four, and the number of microphones can easily be over ten for larger arrays.

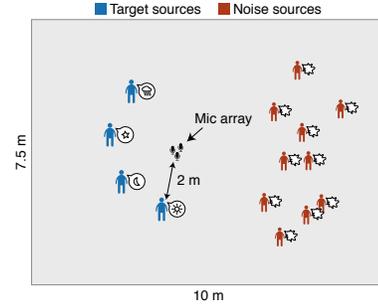


Figure 1: Setup of the simulated experiment.

## 4. PERFORMANCE EVALUATION

In this section, the separation and runtime performances of the proposed and conventional algorithms are compared.

### 4.1. Setup

We simulate a  $10\text{m} \times 7.5\text{m} \times 3\text{m}$  room with reverberation time of 300 ms using the image source method [23] implemented in the `pyroomacoustics` Python package [24]. We place a half-circular microphone array of radius 4 cm at  $[4.1, 3.76, 1.2]$ . The number of microphones is varied from 2 to 8. Between 2 and 4 target sources are placed equispaced on an arc of  $120^\circ$  of radius 2 m centered at the microphone array and at a height of 1.2 m. Diffuse noise is created by 10 additional sources on the opposite side of room. This setup, illustrated in Fig. 1, is that of a few speakers holding a meeting in a noisy open office.

After simulating propagation, the variances of target sources are fixed to  $\sigma_k^2 = 1$  (at an arbitrary reference microphone). The signal-to-noise and signal-to-interference-and-noise ratios are defined as

$$\text{SNR} = \frac{\frac{1}{K} \sum_{k=1}^K \sigma_k^2}{\sigma_n^2}, \quad \text{SINR} = \frac{\sum_{k=1}^K \sigma_k^2}{Q\sigma_i^2 + \sigma_n^2}, \quad (17)$$

where  $\sigma_i^2$  and  $\sigma_n^2$  are the variances of the  $Q$  interfering sources and uncorrelated white noise, respectively. We set them so that  $\text{SNR} = 60$  dB and  $\text{SINR} = 10$  dB. Speech samples of approximately 20 s are created by concatenating utterances from the CMU Sphinx database [25]. The experiment is repeated 50 times for different attributions of speakers and speech samples to source locations. The simulation is conducted at a sampling frequency of 16 kHz. The STFT frame size is 4096 samples with half-overlap and uses a Hann window for analysis and matching synthesis window. We compare OverIVA to three methods.

1. **AuxIVA**: Full IVA with  $M$  channels, followed by picking the  $K$  strongest outputs.
2. **PCA+AuxIVA**: Reduce the number of channels to  $K$  via PCA, followed by IVA. This is only done when  $K \geq 2$ .
3. **OGIVEw**: For  $K = 1$ , orthogonally constrained independent vector extraction (OGIVEw) [21].

We further compare the time-varying Gauss and Laplace versions of all these algorithms. AuxIVA-based algorithms are run for 100 iterations. OGIVEw is run for 4000 iterations with step size of 0.01. The scale of the separated signals is restored by projecting back on the first microphone [26].

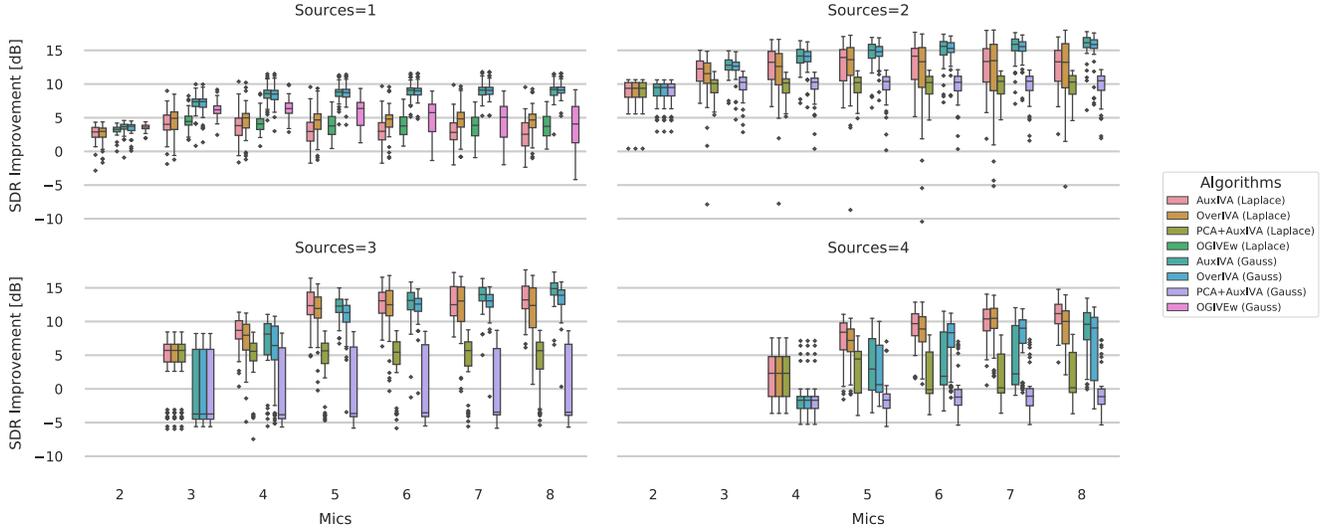


Figure 2: Box-plots of signal-to-distortion ratio (SDR, top row) improvements between mixture and separated signals. Dots represent outliers. The number of sources increases from 1 to 4 left to right and top to bottom. The number of microphones increases from 2 to 8 on the horizontal axis.

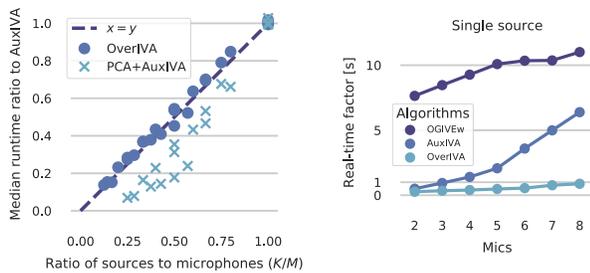


Figure 3: Left, ratio of median runtimes of OverIVA/PCA+AuxIVA to full AuxIVA. Right, runtime per second of audio (i.e., real-time factor) for single source extraction.

#### 4.2. Runtime Performance

To verify the claim of Section 3.2, we measured the runtime of 100 runs of each algorithm and compute the median. As shown in Fig. 3, on the left, the ratio of the runtime of OverIVA to that of AuxIVA follows closely the predicted  $K/M$ . Unsurprisingly, PCA+AuxIVA is much more computationally efficient since it only performs IVA on  $K$  channels. However, its separating performance falls short as discussed in the next section.

For a single source, as shown in Fig. 3, right, OverIVA is very fast and has real-time factor (RTF) less than one for up to 8 microphones (using 100 iterations). Comparatively, AuxIVA has RTF less than one only up to 3 microphones. We also find that our straightforward Python implementation of OGIVEw is not competitive. Let us note that OGIVEw requires many gradient ascent iterations that might run faster in a compiled language such as C or C++.

#### 4.3. Separation Performance

The separation performance of the algorithms is assessed in terms of signal-to-distortion ratio (SDR) as defined in [27]. These metrics

are computed using the `mir_eval` toolbox [28]. Fig. 2 shows box-plots of SDR improvements (with respect to the mixture signal).

We find that of all algorithms, OverIVA and AuxIVA perform best and similarly over all cases investigated. It is interesting to notice a large gap between the determined case (where both algorithms are identical) and using one extra microphone. Just the one extra input signal boosts SDR by 3 to 4 dB. Adding further microphones consistently improves SDR, albeit at a slower pace. In the single source extraction scenario (i.e.,  $K = 1$ ), OverIVA turns out to be perfectly suitable and largely outperforms the state-of-the-art method OGIVEw. When  $K \geq 2$ , the PCA+AuxIVA method falls short in terms of separation, with virtually no improvement when using more microphones. This is likely due to the diffuse noise, since PCA is only optimal when the noise is uncorrelated across channels. Finally, the difference between using Gauss or Laplace models seems consistent across algorithms. For 1 and 2 sources, Gauss IVA performs better than Laplace IVA. However, the trend reverses for 3 and 4 sources. We conjecture that the Laplace AuxIVA might be more robust to mismatched initialization. Using more microphones seems to make the gap in performance disappear.

## 5. CONCLUSION

We introduced OverIVA, a hyperparameter-free algorithm for blind source separation with more microphones than sources. The algorithm applies the efficient updates from auxiliary function-based IVA while maintaining orthogonality between the signal and noise subspaces. A parametrization of the demixing matrix that reduces the number of parameters to estimate is introduced to reduce complexity. We show that using more microphones indeed increases, sometimes dramatically, performance, and that OverIVA solves the problem at a fraction of the cost of full IVA. We also verify that the algorithm performs largely over the state-of-the-art in the so-called blind source extraction (single source) case. Future work will focus on applying the algorithm to recorded data and assessing its performance for real-time implementation.

## 6. REFERENCES

- [1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [2] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, Jun. 1977.
- [3] P. Smaragdakis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, Nov. 1998.
- [4] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE ISCAS*, New Orleans, LA, USA.
- [5] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Advances in Cryptology – ASIACRYPT 2016*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 601–608.
- [6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Dec. 2006.
- [7] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," *LVA/ICA*, vol. 6365, no. 6, pp. 165–172, 2010.
- [8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, Oct. 2011, pp. 189–192.
- [9] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proc. IEEE ICASSP*, Kyoto, JP, Mar. 2012, pp. 2417–2420.
- [10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," *Proc. EUSIPCO*, 2015.
- [11] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," in *Proc. NOLTA98*, 1998, pp. 923–926.
- [12] S. Amari, "Natural gradient learning for over- and under-complete bases in ICA," *Neural computation*, vol. 11, no. 8, pp. 1875–1883, 1999.
- [13] T. Nishikawa, H. Abe, H. Saruwatari, and K. Shikano, "Overdetermined blind separation of acoustic signals based on miso-constrained frequency-domain ICA," in *Proc. ICA*, Kyoto, JP, Apr. 2004, pp. IV–3143 – 3146.
- [14] —, "Overdetermined blind separation for convolutive mixtures of speech based on multistage ICA using subarray processing," in *Proc. IEEE ICASSP*, Montreal, CA, May 2004, pp. I–225.
- [15] C. Osterwise and S. L. Grant, "On over-determined frequency domain bss," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 956–966, May 2014.
- [16] M. Joho and R. L. P. Mathis, H, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proc. ICA*, Helsinki, Finland, Jun. 2000, pp. 81–86.
- [17] I. Lee, T. Kim, and T.-W. Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind Speech Separation*. Dordrecht: Springer, Dordrecht, 2007, pp. 169–192.
- [18] X. Fu and W.-K. Ma, "A simple closed-form solution for overdetermined blind separation of locally sparse quasi-stationary sources," in *Proc. IEEE ICASSP ICASSP 2012 - 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, JP, Mar. 2012, pp. 2409–2412.
- [19] M. Souden, S. Affes, and J. Benesty, "A new approach to blind separation of two sources with three sensors," *Vehicular Technology Conference*, pp. 1–5, 2006.
- [20] K. I. Diamantaras and T. Papadimitriou, "Subspace-based channel shortening for the blind separation of convolutive mixtures," *Signal Processing, IEEE Transactions on*, vol. 54, no. 10, pp. 3669–3677.
- [21] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Dec. 2018.
- [22] J. F. Cardoso, "On the performance of orthogonal source separation algorithms," in *EUSIPCO*, Edinburgh, UK, Sep. 1994, pp. 776–779.
- [23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, p. 943, 1979.
- [24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proc. IEEE ICASSP*, Calgary, CA, Apr. 2018, pp. 351–355.
- [25] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.
- [26] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, Oct. 2001.
- [27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.
- [28] C. Raffel, B. McFee, E. J. Humphrey, J. Salomon, O. Nieto, D. Liang, D. P. W. Ellis, C. C. Raffel, B. Mcfee, and E. J. Humphrey, "mir\_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014.