LEARNING COMPLEX-VALUED LATENT FILTERS WITH ABSOLUTE COSINE SIMILARITY

Anh H. T. Nguyen, V.G. Reju, Andy W. H. Khong, and Ing Yann Soon

School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
Email: nguyenha001@e.ntu.edu.sg, {reju, andykhong, eiysoon}@ntu.edu.sg.

ABSTRACT

We propose a new sparse coding technique based on the power mean of phase-invariant cosine distances. Our approach is a generalization of sparse filtering and K-hyperlines clustering. It offers a better sparsity enforcer than the L_1/L_2 norm ratio that is typically used in sparse filtering. At the same time, the proposed approach scales better than the clustering counterparts for high-dimensional input. Our algorithm fully exploits the prior information obtained by preprocessing the observed data with whitening via an efficient row-wise decoupling scheme. In our simulating experiments, the algorithm produces better estimates than previous approaches do. It yields better separation of live recorded speech mixtures as well.

***Index Terms*—** Row-wise decoupling, cosine similarity, blind source separation, sparse component analysis.

1. INTRODUCTION

In this paper, we focus on the problem of estimating the complex-valued latent filters from observed data when there are more filters than the data dimension (i.e., *under-determined* mixing process or over-complete representation). The data is assumed to follow complex-valued linear model

$$\mathbf{x}[k] = \mathbf{A}\mathbf{s}[k],\ k = 1, 2, \ldots, K, \tag{1}$$

where $\mathbf{x}[k] = [x_1[k], \ldots, x_M[k]]^T \in \mathbb{C}^{M\times 1}$ is the collection of observed signals , $\mathbf{s}[k] = [s_1[k], \ldots, s_N[k]]^T \in \mathbb{C}^{N\times 1}$ is the collection of latent source signals, k is the sample index, M is number of observed signals, N is number of sources, K is the number of samples, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in \mathbb{C}^{M\times N}$ is the mixing matrix where its jth column $\mathbf{a}_j = [a_{1j}, \ldots, a_{Mj}]^T \in \mathbb{C}^{M\times 1}$ denotes a latent filter.

When $M = N$, provided the mixing matrix is invertible, this learning problem has been successfully addressed by independent component analysis (ICA) in which a set of directions are found so that the projections of data onto these directions are maximally non-Gaussian [1]. However, ICA is not applicable in under-determined cases where $M < N$ due to the non-invertible mixing matrix.

Fortunately, natural signals such as speech in the time-frequency domain or image in the wavelet domain are sparse in the sense that they have only a few non-zero elements [2, 3]. With sufficient degree of sparseness, the sources may be approximately *disjoint-orthogonal*, i.e., it is likely that there is only one dominant source at a particular sample index [4]. For such sources, majority of the observed data will concentrate in directions specified by columns of $\mathbf{A}$. As an example, assuming the first source being dominant at the kth index, we have $\mathbf{x}[k] \approx \mathbf{a}_1 s_1[k]$ and therefore $\mathbf{x}[k]$ and $\mathbf{a}_1$ are approximately collinear in complex vector space. Equivalently, the Hermitian angle between them [5],

$$\theta_H(\mathbf{x}[k], \mathbf{a}_1) = \arccos(\left|\mathbf{a}_1^H \mathbf{x}[k]\right| / (\|\mathbf{a}_1\|_2 \|\mathbf{x}[k]\|_2)) \tag{2}$$

is close to zero. Since scaling $\mathbf{x}[k]$ or $\mathbf{a}_1$ by an arbitrary complex scalar number does not change the Hermitian angle [6], one can recover the complex-valued latent filters by clustering the data based on the following phase-invariant cosine distance

$$D^2(\mathbf{x}[k], \widehat{\mathbf{a}}_j) = 1 - \cos^2\theta_H(\mathbf{x}[k], \widehat{\mathbf{a}}_j), \tag{3}$$

where $\widehat{\mathbf{a}}_j$ is the jth centroid. Here, the cosine squared of Hermitian angle is used because of its simple derivative.

By replacing the Euclidean distance with the above distance in K-means clustering, the K-hyperlines (KHL) algorithm was proposed [7, 8]. In K-hyperlines, the centroids, which form an estimation of the mixing matrix, are the (local) minimizers of the objective function

$$J^{(\mathrm{KHL})}(\widehat{\mathbf{A}}) = \tfrac{1}{K}\sum_{k=1}^{K} \min_{j=1,\ldots,N} D^2(\mathbf{x}[k], \widehat{\mathbf{a}}_j). \tag{4}$$

Similar to K-means, K-hyperlines is a hard clustering algorithm due to the inclusion of the minimum function inside its cost function. This minimum function effectively partitions the data into disjoint groups and therefore the K-hyperlines algorithm is only optimal when the sources are perfectly disjoint, which is rarely the case. On the other hand, while the soft extensions of K-hyperlines exist in literature (e.g., Gaussian mixture model of line orientations [9]), solving for the centroids via the expectation-maximization (EM) is computationally intensive due to phase-invariant property of the distance metric. In fact, the Gaussian mixture model (GMM)

of line orientations has a cost per iteration in the order of $O(NKM^2)$ which corresponds to the cost of calculating N weighted covariance matrices. More importantly, since clustering objectives are well-known to have local minima due to the non-convexity of their objective functions, minimizing them with EM scheme prevents us from using advanced optimization algorithms such as the accelerated gradient methods which are able to deal with non-convexity [10]. The main objective of this work is to develop a smooth approximation of K-hyperlines which produces a more accurate estimate as well as one that can be solved efficiently by any gradient-based algorithm.

2. PROPOSED ALGORITHM

We propose the power mean of phase-invariant cosine distance as a suitable objective function for recovering the mixing matrix from mixtures of approximately disjoint sources. We further improve the robustness of our algorithm by exploiting the prior information that one obtains from prewhitening the observed data. This is achieved via an inline row-wise decoupling scheme which results in a smooth unconstrained optimization problem. As a consequence, our algorithm can be solved by any gradient-based method.

2.1. The power mean of phase-invariant cosine distance

The power mean of order r defined by

$$\mu(y_1, y_2, \ldots, y_N;\, r) = \left[\tfrac{1}{N}\textstyle\sum_{j=1}^{N} y_j^r\right]^{1/r} \tag{5}$$

is a generalization of the arithmetic mean that quantifies the central tendency of a group of positive numbers. Its functional value skews toward the minimum value for $r < 1$. In fact, $\mu(y_1, y_2, \ldots, y_N;\, r) \to \min(y_1, y_2, \ldots, y_N)$ when $r \to -\infty$ [11]. As such, it can be considered a smooth approximation of the minimum function. Substituting the power mean into (4) yields the following cost function

$$J(\widehat{\mathbf{A}};\, r) = \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\left[\tfrac{1}{N}\sum_{j=1}^{N}\left(D^2(\mathbf{x}[k], \widehat{\mathbf{a}}_j)\right)^r\right]^{1/r}. \tag{6}$$

Since the power mean is increasing w.r.t. r, we have $J(\widehat{\mathbf{A}};\, r) \geq J^{(\mathrm{KHL})}(\widehat{\mathbf{A}}) \geq 0$. In other words, the proposed objective function is bounded below by a constant and therefore any optimization algorithm, which guarantees to decrease the cost function, will eventually converge to a local minimum.

In practice, it is common to decorrelate the observed data with a whitening filter to facilitate the learning process [1]. Given that the sources are zero-mean and uncorrelated with unit variance, the mixing matrix, which models the relationship between the pre-whitened observed data and the sources, is unitary when $M = N$ and is semi-unitary when $M < N$, that is $\mathbf{A}\mathbf{A}^H = \mathbf{I}_M$. Incorporating this prior information into our objective function yields the following constrained optimization problem

$$\min_{\widehat{\mathbf{A}}}\ J(\widehat{\mathbf{A}};\, r),\ \text{s.t.}\ \widehat{\mathbf{A}}\widehat{\mathbf{A}}^H = \mathbf{I}_M. \tag{7}$$

Now, for each sample, we define the magnitude-squared cosine-similarity vector $\mathbf{f}[k]$ where its jth component is given by $\cos^2\theta_H(\mathbf{x}[k], \widehat{\mathbf{a}}_j)$. When the sources are approximately disjoint, $\mathbf{x}[k]$ will have only N distinct directions which are collinear to the columns of $\mathbf{A}$. As a result, the columns of $\widehat{\mathbf{A}}$ include all directions specified by columns of the true mixing matrix if one component of $\mathbf{f}[k]$ is sufficiently close to 1 for every k. At the same time, since $\left\|\widehat{\mathbf{A}}^H\mathbf{x}[k]\right\|_2 = \|\mathbf{x}[k]\|_2$ for any $\widehat{\mathbf{A}}$ which is semi-unitary, one can show that

$$\textstyle\sum_{j=1}^{N} \|\widehat{\mathbf{a}}_j\|_2^2 \cos^2\theta_H(\mathbf{x}[k], \widehat{\mathbf{a}}_j) = 1,\ \forall k. \tag{8}$$

Intuitively, the higher the largest component of $\boldsymbol{f}[k]$, the lower the weighted sum of the remaining components. This implies one can reconstruct the mixing matrix up to some permutation and scaling ambiguity by minimizing the sparsity-promoting penalty of $\mathbf{f}[k]$, assuming each column of the mixing matrix contributes a similar amount of information so that all $\|\widehat{\mathbf{a}}_j\|_2^2$ are approximately equal when we are close to the optimal solution. Indeed, the power mean of order r where $r < 1$ is suitable for this purpose because it belongs to a class of sparsity enforcing functions, namely, strictly Schur-concave functions [12, pp. 138-139][13]. However, (8) implies that the "soft" minimum of $1 - \boldsymbol{f}[k]$ is a better sparsity enforcer than the "soft" minimum of $\boldsymbol{f}[k]$. In the determined cases where $M = N$, since $\|\widehat{\mathbf{a}}_j\|_2^2 = 1,\ \forall j$, the effect is apparent because forcing a component of $\boldsymbol{f}[k]$ to 1 will cause all other components to vanish. In short, the power mean of phase-invariant cosine distance under semi-unitary constraint is suitable for learning the mixing matrix from mixtures of disjoint sparse sources.

The optimization problem in (7) can be solved with projected gradient descent or projected quasi-Newton in which the following constraint projection

$$\widehat{\mathbf{A}} \leftarrow (\widehat{\mathbf{A}}\widehat{\mathbf{A}}^H)^{-1/2}\widehat{\mathbf{A}} \tag{9}$$

is applied after every parametric update. This projection will greatly reduce the convergence rate. As a result, the idea to solve a semi-unitary constrained problem without actually performing the orthonormalization after each update has been explored. In [14], this is accomplished by regularizing the cost function with the reconstruction cost so that $\widehat{\mathbf{A}}$ is approximately semi-unitary. On the other hand, the authors of [15] decouple each row of $\widehat{\mathbf{A}}$ successively via the Schur complement of a matrix formed from all other rows. Both approaches have their disadvantages, the former is inexact while the later is order-dependent. On the contrary, by simply substituting the constrained projection into the cost function, we introduce an equivalent unconstrained cost

$$\min_{\mathbf{B}}\ J((\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B};\, r) \tag{10}$$

which performs the exact row-wise decoupling simultaneously for all rows in place. In essence, we minimize the orig-

inal cost function w.r.t. an auxiliary matrix and then compute the final solution via a single projection given in (9). Although it is seemingly difficult to compute the gradient of nested matrix functions, for our particular cost, we will show that one can find its gradient efficiently from the original gradient in $O(MN^2)$.

2.2. Efficient inline row-wise decoupling

Let us consider the composite cost function in the form of $J(\widehat{\mathbf{A}};\ r)$ where $\widehat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$. Here, according to Wirtinger calculus [16, 17], one should treat the argument and its conjugate as independent variables (i.e., we consider $\mathbf{B}^*$ as a constant when deriving the derivative of J w.r.t. $\mathbf{B}$ and vice versa). Moreover, the direction which yields maximal change in the value of J w.r.t. $\mathbf{B}$ is $[\nabla_{\mathbf{B}^*}J]_{ij} = 2\partial J/\partial b_{ij}^*$. Applying the chain rule of matrix functions given in [17], one obtains

$$\begin{aligned}&\mathrm{vec}(\nabla_{\mathbf{B}^*}J) = \\ &\quad(\mathcal{D}_{\mathbf{B}^*}\widehat{\mathbf{A}})^T\mathrm{vec}(\nabla_{\widehat{\mathbf{A}}^*}J)^* + (\mathcal{D}_{\mathbf{B}}\widehat{\mathbf{A}})^H\mathrm{vec}(\nabla_{\widehat{\mathbf{A}}^*}J),\end{aligned} \quad (11)$$

where $\mathcal{D}_{\mathbf{B}^*}\widehat{\mathbf{A}} = \partial\mathrm{vec}(\widehat{\mathbf{A}})/\partial\mathrm{vec}(\mathbf{B}^*)^T$ and $\mathcal{D}_{\mathbf{B}}\widehat{\mathbf{A}} = \partial\mathrm{vec}(\widehat{\mathbf{A}})/\partial\mathrm{vec}(\mathbf{B})^T$ are the Jacobian matrices of $\widehat{\mathbf{A}}$, and $\mathrm{vec}(\cdot)$ denotes the vectorization operator. Letting $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ be the economy-sized singular value decomposition (eSVD) of $\mathbf{B}$, where $\mathbf{U} \in \mathbb{C}^{M\times M}$ is unitary, $\mathbf{\Sigma} \in \mathbb{R}_{>0}^{M\times M}$ is diagonal, $\mathbf{V} \in \mathbb{C}^{N\times M}$ is semi-unitary, we define several intermediate variables given by $\mathbf{B}_1 = \mathbf{B}\mathbf{B}^H = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^H$, $\mathbf{B}_2 = \mathbf{B}_1^{-1} = \mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^H$, $\mathbf{B}_3 = \mathbf{B}_2^{1/2} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^H$ and $\widehat{\mathbf{A}} = \mathbf{B}_3\mathbf{B} = \mathbf{U}\mathbf{V}^H$. Applying the identity $d(\mathbf{X}\mathbf{Y}) = d\mathbf{X}.\mathbf{Y} + \mathbf{X}.d\mathbf{Y}$ and the vectorization operator on previously defined variables, reordering with $\mathrm{vec}(\mathbf{X}\mathbf{Y}\mathbf{Z}) = (\mathbf{Z}^T \otimes \mathbf{X})\mathrm{vec}\mathbf{Y}$, and simplifying, we obtain

$$\mathbf{W} = (\mathbf{\Sigma}\otimes\mathbf{I}_M + \mathbf{I}_M\otimes\mathbf{\Sigma})^{-1}(\mathbf{V}^H\otimes\mathbf{\Sigma}^{-1}\mathbf{U}^T), \quad (12)$$

$$(\mathcal{D}_{\mathbf{B}}\widehat{\mathbf{A}})^H = (\mathbf{I}_N\otimes\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^H) - (\mathbf{V}^*\mathbf{\Sigma}\otimes\mathbf{U})\mathbf{W}^*, \quad (13)$$

$$(\mathcal{D}_{\mathbf{B}^*}\widehat{\mathbf{A}})^T = -\mathbf{K}_{NM}(\mathbf{U}\otimes\mathbf{V}^*\mathbf{\Sigma})\mathbf{W}, \quad (14)$$

where $\otimes$ denotes the Kronecker product and $\mathbf{K}_{NM}$ is the commutation matrix defined by $\mathbf{K}_{NM}\mathrm{vec}\mathbf{X} = \mathrm{vec}(\mathbf{X}^T)$ for any $\mathbf{X} \in \mathbb{C}^{N\times M}$. Here, when finding the matrix differentiation of $\widehat{\mathbf{A}}$, we use spectral decomposition to reduce the inversion of a generic MN-by-MN matrix to the inversion of a diagonal matrix $(\mathbf{\Sigma}\otimes\mathbf{I}_M + \mathbf{I}_M\otimes\mathbf{\Sigma})$. We further exploit the unique structure of this diagonal matrix to reduce the computational complexity and memory storage. In fact, we observe for an arbitrary matrix $\mathbf{M}$ that $(\mathbf{\Sigma}\otimes\mathbf{I}_M + \mathbf{I}_M\otimes\mathbf{\Sigma})^{-1}\mathrm{vec}(\mathbf{M}) = \mathrm{vec}\left(\mathbf{M}\oslash(\mathbf{1}_M\boldsymbol{\sigma}^T + \boldsymbol{\sigma}\mathbf{1}_M^T)\right)$ where $\boldsymbol{\sigma} = \mathrm{diag}(\mathbf{\Sigma}) \in \mathbb{R}_{>0}^{\mathrm{M}\times 1}$, $\mathbf{1}_M = [1,\ldots,1]^T \in \mathbb{R}^{M\times 1}$ and $\oslash$ denotes the element-wise division. Consequently, substituting (13) and (14) into (11) and then unvectorizing, we obtain

$$\mathbf{C} = -(\mathbf{\Sigma}^{-1}\mathbf{U}^H(\nabla_{\widehat{\mathbf{A}}^*}J)\mathbf{V})\oslash(\mathbf{1}_M\boldsymbol{\sigma}^T + \boldsymbol{\sigma}\mathbf{1}_M^T), \quad (15)$$

$$\nabla_{\mathbf{B}^*}J = \mathbf{U}(\mathbf{C}^H + \mathbf{C})\mathbf{\Sigma}\mathbf{V}^H + \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^H(\nabla_{\widehat{\mathbf{A}}^*}J). \quad (16)$$

This implies that one can replace any $\nabla_{\widehat{\mathbf{A}}^*}J$ and obtain the gradient of decoupled objective function $\nabla_{\mathbf{B}^*}J$ in $O(MN^2)$ which is the cost of calculating eSVD of $\mathbf{B}$. Since the time complexity of evaluating $\nabla_{\widehat{\mathbf{A}}^*}J$ from (6) is $O(KMN)$, the computational cost of finding $\nabla_{\mathbf{B}^*}J$ from $\nabla_{\widehat{\mathbf{A}}^*}J$ is negligible. Consequently, the total cost-per-iteration of our approach is $O(KMN)$, which is asymptotically M times faster than the soft-clustering approach. Interestingly, sparse filtering (SF), a feature-learning algorithm, has a same asymptotic cost to our proposed algorithm [18]. As we will explain, our algorithm outperforms the sparse filtering method in estimating $\mathbf{A}$ under disjoint orthogonal assumption.

2.3. Connection to sparse filtering

Sparse filtering minimizes the L_1/L_2 norm ratio as follows

$$\min_{\widehat{\mathbf{A}}}\ J^{(SF)}(\widehat{\mathbf{A}}) = E\Big\{\ \left\|\widehat{\mathbf{A}}^H\mathbf{x}\right\|_1 / \left\|\widehat{\mathbf{A}}^H\mathbf{x}\right\|_2\Big\}. \quad (17)$$

Suppose we constrained $\widehat{\mathbf{A}}$ to be semi-unitary, the objective function of sparse filtering is equivalent to $E\Big\{\sum_{j=1}^N \|\widehat{\mathbf{a}}_j\|_2\cos\theta_H(\mathbf{x}[k],\widehat{\mathbf{a}}_j)\Big\}$. As oppose to our proposed objective function and clustering methods, the objective of sparse filtering depends on the norm of the latent filters inside the sparsity penalty. This may lead to the degeneration of some columns of $\widehat{\mathbf{A}}$. Nevertheless, if we further assume $\|\widehat{\mathbf{a}}_j\|_2 \approx 1$, the sparsity penalty of sparse filtering can be understood as the squared root of the power mean of $\boldsymbol{f}[k]$ for $r = 0.5$. This sparsity enforcer is less flexible than our method as well as not suitable for mixtures of approximately disjoint sources because it is the "soft" minimum of $\boldsymbol{f}[k]$ rather than $1 - \boldsymbol{f}[k]$.

3. PERFROMANCE EVALUATION

We compare the performance of our proposed method (PM) to related approaches such as KHL [7], GMM [9], and SF [18]. Our algorithm is implemented using Nesterov accelerated gradient [19]. While we set $r = -0.5$ and use a fixed learning rate of 1 for all experiments in this paper, these hyper-parameters should be found via cross validation. For GMM, its parameters are set as suggested by its authors. For SF, we alter its original code to work with complex-valued signals. All the algorithms share a similar preprocessing step, initial mixing matrix, and stopping criteria. We employ the average mixing error ratio (MER) as our performance criterion [20]. Given $\widehat{\mathbf{a}}_j$ be the estimation of jth column $\mathbf{a}_j$, the average mixing error ratio in decibels (dB) is then defined as

$$\mathrm{MER} = (20/N)\sum_j \log\Big(\ \left\|\mathbf{a}_j^{\mathrm{coll}}\right\| / \left\|\mathbf{a}_j^{\mathrm{orth}}\right\|\Big), \quad (18)$$

where $\mathbf{a}_j^{\mathrm{coll}}$ and $\mathbf{a}_j^{\mathrm{orth}}$ are, respectively, collinear component and orthogonal component of $\mathbf{a}_j$ in $\widehat{\mathbf{a}}_j$. MER is computed using the BSS_EVAL toolbox[1]. A higher MER implies better performance.

[1] http://bass-db.gforge.inria.fr/bss_eval/

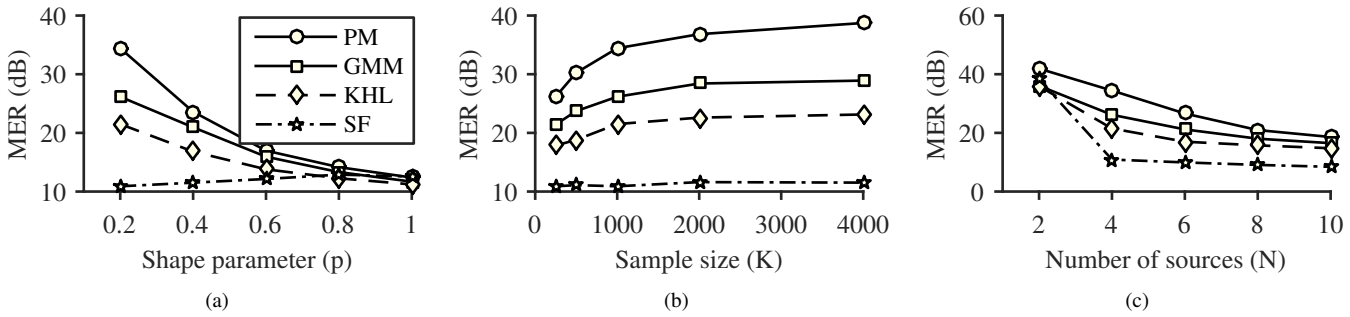

Fig. 1: Average MER in estimation of 2×4 mixing matrix w.r.t. : a) Sparseness. b) Sample size. c) Number of sources

Table 1: Output SDR and SIR in dB for 2mic_4src_5cm subset of SiSEC dev1 dataset

RT60	130ms				250ms			
Source	4 males		4 females		4 males		4 females	
Perf. metric	SDR	SIR	SDR	SIR	SDR	SIR	SDR	SIR
PM	**4.55**	**8.27**	3.80	6.38	**3.67**	**6.06**	**3.57**	**5.36**
[22]	4.1	6.38	**4.47**	**6.48**	3.55	5.07	3.5	4.85
[9]	3.31	-	3.92	-	2.62	-	3.49	-
Input	-4.81	-4.60	-4.76	-4.68	-4.79	-4.64	-4.83	-4.71

In each simulation, we report the average performance of 100 trials. New sources and a mixing matrix are created randomly for each trial. The real part and the imaginary part of the mixing matrix are generated independently with standard normal distribution. The sources are drawn according to complex generalized Gaussian distribution $f(s) \propto \exp(-|s|^p)$ where p is the shape parameter [21]. The smaller the shape parameter is, the sparser the source. The default values of the sample size and the shape parameter are respectively $K = 10^3$ and $p = 0.2$. Figs.˜1(a) and (b) illustrate the performance of estimating 2×4 mixing matrix w.r.t. the sparseness of the sources and sample size. The performance w.r.t. to the number of sources is shown in Fig.˜1(c), given the number of observed signals is 2. One can see that our proposed algorithm outperforms all its peers in most cases.

To evaluate our algorithm w.r.t. real-world data, we consider the problem of separating convolutively mixed speeches, i.e., $x_i(t) = \sum_{j=1}^{N}(a_{ij} * s_j)(t)$ where $a_{ij}(t)$ denotes the unknown impulse response between the ith microphone and the jth source. Using short-time Fourier transform (STFT) with an appropriate choice of window length [23], one can convert the convolutive mixing process in time domain to complex-valued mixing process in time-frequency domain as

$$\mathbf{x}(\tau, f) \approx \mathbf{A}(f)\mathbf{s}(\tau, f). \quad (19)$$

Following the current state-of-the-art systems given in [9, 22], we estimate $\widehat{\mathbf{A}}(f)$ at each bin separately with our method then compute the following soft-mask

$$M(j, \tau, f) = \frac{\exp(\beta \cos^2 \theta_H(\mathbf{x}(\tau, f), \widehat{\mathbf{a}}_j(f)))}{\exp(\beta \sum_l \cos^2 \theta_H(\mathbf{x}(\tau, f), \widehat{\mathbf{a}}_l(f)))}, \quad (20)$$

where $\beta > 0$ is a predefined constant which controls the softness of the mask. We introduce the above soft-mask since it yields less distortion in the reconstructed speeches compared to traditional binary masking. Next, the source index of the mask is properly reordered across the frequency-bin using multi-band permutation alignment [24]. Finally, we extract the sources via time-frequency masking and inverse STFT.

We compare our method with [9] and [22] on *dev1* dataset of the Signal Separation Evaluation Campaign (SiSEC) [20]. This dataset contains 16 mixing scenarios of speech signals. In each scenarios, two recordings of three or four people speaking are recorded in an actual room environment using two microphones 5 cm or 1 m apart. The room reverberation time is either 130 ms or 250 ms. We measure the performance of signal separation using signal distortion ratio (SDR) and signal interference ratio (SIR) [25]. We set $\beta = 12.5$ and use the same STFT parameters with [9]. Table 1 contains the separation results on 4 most challenging scenarios of SiSEC *dev1* dataset[2]. In this subset, while our method and [22] have the same average SDR of 3.9 dB, the average SIR of our method is 6.5 dB compared to 5.7 dB of the algorithm proposed in [22]. For all 16 mixing scenarios, in comparison to [22], our algorithm yields 0.18 dB improvement in average SDR (5.55 dB vs. 5.37 dB) and 1.67 dB improvement in SIR (9.65 dB vs. 7.98 dB). Interestingly, our proposed method requires at most one minute on a typical Intel i7 (3.4 Ghz) machine for any multi-channel mixture in SiSEC datasets while the method in [22] requires up to one hour on an Intel i5 (3.4 Ghz) machine.

4. CONCLUSIONS

We presented an inline decoupling sparse coding technique for mixing matrix estimation. The algorithm combines the benefits of clustering approach and sparse filtering approach by exchanging the L_1/L_2 norm ratio in sparse filtering with the power mean of phase-invariant cosine distance. Experiments on simulated data and real data show promising results.

[2]The SIR values of [22] are retrieved from https://sisec.wiki.irisa.fr/

5. REFERENCES

[1] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[2] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Process.*, vol. 81, no. 11, pp. 2353–2362, 2001.

[3] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE, 44th Annu. Meeting*, vol. 3813, 1999, pp. 188–195.

[4] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2002, pp. I–529–I–532.

[5] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Math.*, vol. 69, no. 1, pp. 95–103, 2001.

[6] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.

[7] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sanei, "K-hyperline clustering learning for sparse component analysis," *Signal Process.*, vol. 89, no. 6, pp. 1011–1022, 2009.

[8] P. D. O'Grady and B. A. Pearlmutter, "Hard-LOST: Modified K-Means for oriented lines," in *Proc. Irish Signals Syst. Conf.*, 2004, pp. 247–252.

[9] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[10] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[11] P. S. Bullen, *Handbook of means and their inequalities*, 2nd ed. Dordrecht, The Netherlands: Kluwer, 2003, vol. 560.

[12] A. W. Marshall, I. Olkin, and B. Arnold, *Inequalities: Theory of majorization and its applications*, 2nd ed. New York: Springer-Verlag, 2011.

[13] K. Kreutz-Delgado, B. D. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Convex/Schur-convex (csc) log-priors and sparse coding," in *Proc. 6th Joint Symp. Neural Computation*, 1999, pp. 65–71.

[14] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Neural Inform. Process. Syst.*, 2011, pp. 1017–1025.

[15] P. A. Rodriguez, M. Anderson, X.-L. Li, and T. Adali, "General non-orthogonal constrained ICA," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2778–2786, Jun. 2014.

[16] T. Adali and P. J. Schreier, "Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation," *IEEE Signal Process. Mag.*, vol. 5, no. 31, pp. 112–128, Sep. 2014.

[17] A. Hjørungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[18] J. Ngiam, Z. Chen, S. A. Bhaskar, P. W. Koh, and A. Y. Ng, "Sparse filtering," in *Proc. Neural Inform. Process. Syst.*, 2011, pp. 1125–1133.

[19] Y. Nesterov, *Introductory lectures on convex optimization*. New York: Kluwer, 2004, vol. 87.

[20] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Independent Component Anal. Signal Separation (ICA'09)*, 2009, pp. 734–741.

[21] M. Novey, T. Adali, and A. Roy, "A complex generalized Gaussian distribution - Characterization, generation, and estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1427–1433, Mar. 2010.

[22] J. Cho and C. D. Yoo, "Underdetermined convolutive BSS: Bayes risk minimization mased on a mixture of super-Gaussian posterior approximation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 828–839, Mar. 2015.

[23] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[24] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digital Signal Process.*, vol. 31, pp. 79–92, 2014.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.