Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation

Dalia El Badawy, Alexey Ozerov, Ngoc Q. K. Duong

RELATIVE GROUP SPARSITY FOR NON-NEGATIVE MATRIX FACTORIZATION WITH APPLICATION TO ON-THE-FLY AUDIO SOURCE SEPARATION

Dalia El Badawy, Alexey Ozerov and Ngoc Q. K. Duong

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{dalia.elbadawy, alexey.ozerov, quang-khanh-ngoc.duong}@technicolor.com

ABSTRACT

We consider dictionary-based signal decompositions with group sparsity, a variant of structured sparsity. We point out that the group sparsity-inducing constraint alone may not be sufficient in some cases when we know that some bigger groups or so-called supergroups cannot vanish completely. To deal with this problem we introduce the notion of relative group sparsity preventing the supergroups from vanishing. In this paper we formulate practical criteria and algorithms for relative group sparsity as applied to non-negative matrix factorization and investigate its potential benefit within the on-the-fly audio source separation framework we recently introduced. Experimental evaluation shows that the proposed relative group sparsity leads to performance improvement over group sparsity in both supervised and semi-supervised on-the-fly audio source separation settings.

Index Terms— group sparsity, non-negative matrix factorization, audio source separation, universal model

1. INTRODUCTION

Dictionary-based signal decompositions usually benefit from sparsity or structured sparsity constraints, especially in the case of overcomplete dictionaries. Group sparsity is a variant of structured sparsity that became quite popular in both signal processing [1] and machine learning [2]. In particular, several approaches using group sparsity for audio source separation were recently proposed [3, 4]. While the idea behind sparsity is to allow only few coefficients of the decomposition to be active (i.e., having non-negligible energy), the idea behind group sparsity is to allow only few pre-defined groups of coefficients to be active. These groups of coefficients are usually defined relying on some prior knowledge about the signal and on some desired properties of the decomposition. For example, in the context of speaker independent single channel speech denoising problem Sun and Mysore [4] propose modeling speech from an unknown speaker by a union of few dictionaries from a large amount of pre-trained speaker-dependent groups of dictionary patterns. Signal decomposition within this dictionary also called *universal speech model* is achieved via group sparsity, where one group consists of all coefficients within the decomposition that correspond to one pre-trained dictionary.

However, the group sparsity constraint alone may not be sufficient in some cases. Let us explain this by one example. Consider the problem of single channel separation of two speech sources: one male speech source and one female speech source. In line with [4], let us assume that the two sources are modeled by a universal male speech model and a universal female speech model, respectively. A straightforward application of a group sparsity-based approach as in [4] would assume that the mixture of sources is decomposed in the union of several male and female pre-trained dictionaries. In that case it can happen, especially if the female source voice is close to a male's one, that all the active groups (the selected dictionaries) will belong to the universal male speech model. As a consequence, the full mixture will be appointed to only one source. Although, we know that there are two sources. To avoid this problem one solution would consist in explicitly preventing the coefficients corresponding to one universal source model from vanishing altogether.

Formulating such a solution in a more general manner, we introduce in this paper the notion of *relative group sparsity*. We assume that the groups are assembled into so-called *supergroups* (i.e., bigger groups corresponding to the universal speech models in the above example) and we characterize a relative group sparsity constraint as

- **inducing the sparsity of the groups** (as in the group sparsity), while
- **inducing the "anti-sparsity" of the supergroups** (i.e., preventing them from vanishing entirely).

In other words, the group sparsity property is now considered relative to the corresponding supergroup and not within the full set of coefficients.

In this paper we formulate practical criteria and algorithms for relative group sparsity as applied to non-negative matrix factorization (NMF) with Itakura-Saito (IS) divergence [5] and investigate its potential benefit within the on-the-fly audio source separation framework [6] we recently introduced. While other user-guided source separation approaches require from the user certain skills and/or knowledge (e.g., humming or speaking source examples [7, 8]; or annotating the mixture spectrograms [9–12]), this framework allows in principle separating mixtures of any sounds with very light guidance that can be performed by almost any user. Briefly, after having listened to the mixture a user is only required to describe the sources to be separated by some keywords (e.g., "dog barking", "wind", etc.), and some external search engines are then used to retrieve the corresponding source examples for training. In line with the on-the-fly image retrieval approach [13], the source models are then learned *on-the-fly* and used for separation. A demo video of the the on-the-fly audio source separation user interface we have created is available at `http://youtu.be/mBmJW7cy710/`. We have found in [6] that among other methods we tested those based on the universal modeling concept with group sparsity [4] performed the best. However, since these methods are based on several universal models, as in the male/female speech separation example above, they may suffer from exactly the same problem of source vanishing and we expect the proposed relative group sparsity to fix it.

In summary, the main contribution of this paper consists in introducing the notion of relative group sparsity and formulating corresponding practical criteria and algorithms for the NMF model. The potential of this new sparsity is demonstrated within the on-the-fly audio source separation framework. In addition, we investigate the on-the-fly audio source separation in the semi-supervised case, i.e., when retrieved source examples are only available for some but not all sources. We show how relative group sparsity is crucial in that case as well.

The remainder of the paper is organized as follows. Section 2 summarizes the on-the-fly audio source separation framework based on NMF with group sparsity as introduced in [6]. Section 3 briefly overviews the modifications we propose within this framework to efficiently handle the semi-supervised case. In section 4 the proposed criteria and algorithms for NMF with relative group sparsity are described. Numerical results are given in section 5 followed by the conclusion in section 6.

2. SUPERVISED ON-THE-FLY AUDIO SOURCE SEPARATION BASED ON NMF WITH GROUP SPARSITY

This section summarizes the on-the-fly audio source separation framework based on the NMF with group sparsity [6].

2.1. Supervised NMF-based source separation

We consider a single-channel source separation problem with J sources. Let $\mathbf{X}$ and $\mathbf{S}_j$ be the $F \times N$ matrices of the short-time Fourier transform (STFT) coefficients of the observed mixture signal and the j-th source signal, respectively, where F is the number of frequency bins and N the number of time frames. The mixing model writes

$$\mathbf{X} = \sum\nolimits_{j=1}^{J} \mathbf{S}_j. \tag{1}$$

Let $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ be the power spectrogram of the mixture where $\mathbf{A}^{\cdot p}$ is the matrix with entries $[\mathbf{A}]_{ij}^{p}$. NMF aims at decomposing the $F \times N$ non-negative matrix $\mathbf{V}$ as a product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ of dimensions $F \times K$ and $K \times N$, respectively, such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$. In audio applications this decomposition is often done by optimizing the following criterion [5]

$$\min_{\mathbf{H}\geq 0, \mathbf{W}\geq 0} D(\mathbf{V}\|\mathbf{WH}), \tag{2}$$

where $D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{f,n=1}^{F,N} d_{IS}(\mathbf{V}_{fn}\|\hat{\mathbf{V}}_{fn})$ and $d_{IS}(x\|y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$ is the IS divergence [5]. The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are initialized with random non-negative values and are iteratively updated via multiplicative update (MU) rules [5].

In the supervised setting, the factorization of $\mathbf{V}$ is guided by a pre-learned spectral model. In other words, the matrix $\mathbf{W}$ is obtained (and fixed within the optimization) by

$$\mathbf{W} = [\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(J)}], \tag{3}$$

where $\mathbf{W}_{(j)}$ is the spectral model for the j-th source learned via the NMF decomposition of some training examples using (2).

2.2. On-the-fly NMF-based source separation

The above supervised strategy could be also applied in the on-the-fly context by training the NMF spectral models (dictionaries) $\mathbf{W}_{(j)}$ from the retrieved examples. However, this straightforward approach does not address the following two challenges:

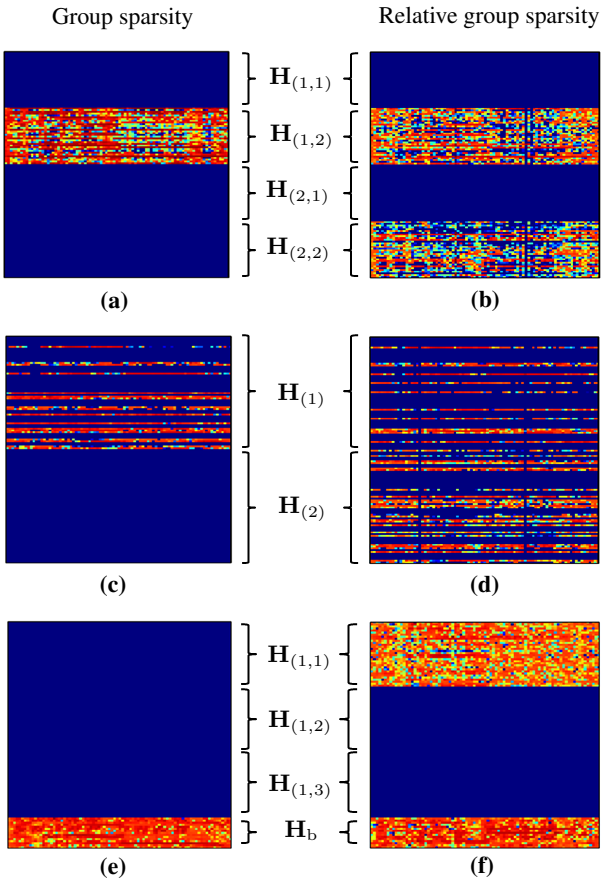

Fig. 1. Examples of estimated activation matrices $\mathbf{H}$. Left column: **(a)** block sparsity in the supervised case, **(c)** component sparsity in the supervised case, and **(e)** block sparsity in the semi-supervised case. Right column: same settings as in left column, but for the proposed relative block/component sparsity.

1. How to deal with irrelevant retrieved examples (i.e., sounds not corresponding to the sources of interest)?
2. How to deal with noisy retrieved examples (i.e., mixtures of relevant and irrelevant sounds)?

To deal with these challenges, it was proposed in [6] to rely on a universal model-based approach with group sparsity [4]. This approach relies mainly on the following steps:

1. Assume L_j examples were retrieved for the j-th source. For the l-th example ($l = 1, \ldots, L_j$) a spectral model $\mathbf{W}_{(j,l)}$ is learned optimizing (2).
2. The mixture spectral model $\mathbf{W}$ is constructed as in (3), where each $\mathbf{W}_{(j)}$ is a universal spectral model obtained by concatenation of the pre-trained example spectral models as $\mathbf{W}_{(j)} = [\mathbf{W}_{(j,1)}, \ldots, \mathbf{W}_{(j,L_j)}]$.
3. The non-negative activation matrix $\mathbf{H}$ is randomly initialized and estimated (while keeping $\mathbf{W}$ fixed) using the following criterion

$$\min_{\mathbf{H}\geq 0} D(\mathbf{V}\|\mathbf{WH}) + \Psi(\mathbf{H}), \tag{4}$$

where

$$\Psi(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log\left(\epsilon + \|\mathbf{H}_{(j,g)}\|_1\right) \tag{5}$$

is a group sparsity-inducing penalty defined as in [4], $\|\cdot\|_1$ denotes the ℓ_1 matrix norm, ϵ and λ_j are some constants, and $\mathbf{H}_{(j,g)}$ $(j = 1, \ldots, G_j)$ are the groups within the activation sub-matrix $\mathbf{H}_{(j)}$ corresponding to the j-th universal source model.

4. Sources are separated using standard Wiener filtering [6].

An iterative algorithm optimizing criterion (4) based on MU rules [14] is summarized by Algorithm 1, where $\eta > 0$ is a constant parameter, $\mathbf{P}_{(j,g)}$ is a matrix with equal entries; its size is the same as $\mathbf{H}_{(j,g)}$, and $\mathbf{P}$ is a matrix concatenating all $\mathbf{P}_{(j,g)}$. This algorithm is almost identical to the one proposed in [3], except that the groups are defined differently and $\mathbf{W}$ is not updated here. It is proven in [3] using a majorization-minimization [15] formulation that these updates with $\eta = 1/2$ are *monotonic*, i.e., they ensure that the cost function in (4) is non-increasing after each iteration.

Algorithm 1 MU rules for NMF with group sparsity

Input: $\mathbf{V}, \mathbf{W}, \boldsymbol{\lambda}$
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
    end for
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  until convergence

In our previous work [6] we introduced two options for defining the groups $\mathbf{H}_{(j,g)}$. First, as in [4], we restrict the groups to be sub-matrices of $\mathbf{H}_{(j)}$ corresponding to the pre-trained example spectral models $\mathbf{W}_{(j,l)}$ (in that case the indices g and l coincide and $G_j = L_j$). This so-called *block sparsity*-inducing strategy allows filtering out irrelevant spectral models $\mathbf{W}_{(j,l)}$, thus dealing with irrelevant retrieved examples (see, e.g., Fig. 1 (a) and (b)). Second, as an alternative solution, we restrict the groups to be lines of $\mathbf{H}_{(j)}$ corresponding to different spectral components (in that case the number of groups G_j simply equals to the number of rows in $\mathbf{H}_{(j)}$). This so-called *component sparsity*-inducing strategy allows filtering out irrelevant spectral components, thus dealing with noisy retrieved examples (see, e.g., Fig. 1 (c) and (d)).

While enforcing group sparsity as in (5) is useful for selecting the appropriate dictionary elements for the decomposition, it does not guarantee that elements from every learned source model are used; in other words, two or more sources in the mixture can be lumped together and expressed using the same dictionary elements making the separation impossible (as in the male/female speech example in the introduction). This "source vanishing" problem was observed in practice quite often and is illustrated in Fig. 1 (a) and (c). It can also be easily seen that increasing the constants λ_j in the penalty (5) (thus decreasing the number of active/emerging groups) increases the chances of source vanishing.

3. PROPOSED SEMI-SUPERVISED ON-THE-FLY AUDIO SOURCE SEPARATION

In this section we briefly describe a novel so-called *semi-supervised* on-the-fly audio source separation setup that was not addressed in [6], but may be very useful in practice. In this setup we assume that for some of sources there are no retrieved examples available. That may happen either in the case when no keywords were provided for some of sources (e.g., because the user has not described all the sources either to save time or because he/she has not recognized some sources) or in the case when for some keywords no examples were retrieved by the search engine (e.g., if there are no sounds matching the corresponding queries).

All these "undescribed" sources are modeled as one background source by a randomly initialized NMF model $\boldsymbol{\theta}_{\mathrm{b}} = \{\mathbf{W}_{\mathrm{b}}, \mathbf{H}_{\mathrm{b}}\}$ with a small number of components (i.e., number of columns in $\mathbf{W}_{\mathrm{b}}$) K_{b}. All the other sources, for which some examples are available, are modeled as in the supervised case by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ (see Fig. 1 (e) and (f)) and the parameters are estimated altogether by optimizing the following criterion

$$\min_{\mathbf{H} \geq 0, \mathbf{W}_{\mathrm{b}} \geq 0, \mathbf{H}_{\mathrm{b}} \geq 0} D(\mathbf{V} \| \mathbf{WH} + \mathbf{W}_{\mathrm{b}}\mathbf{H}_{\mathrm{b}}) + \Psi(\mathbf{H}). \qquad (6)$$

We see that in contrast to criterion (4) $\mathbf{W}_{\mathrm{b}}$ is updated and there is no group sparsity-inducing penalty on $\mathbf{H}_{\mathrm{b}}$. Criterion (6) may be optimized by MU rules that are very similar to those in Algorithm 1, and we omit them here for the sake of conciseness.

Note that the source vanishing problem is even more problematic in this case than in the supervised case. Indeed, we have observed in many cases that the full matrix $\mathbf{H}$ vanishes, which means that the full mixture is modeled by the estimated background model $\boldsymbol{\theta}_{\mathrm{b}}$ (see Fig. 1 (e)). This is very likely due to the fact that $\mathbf{W}_{\mathrm{b}}$ and $\mathbf{H}_{\mathrm{b}}$ are now fully unconstrained in (6), while $\mathbf{W}$ is fixed and $\mathbf{H}$ is constrained by the group sparsity-inducing penalty.

4. NMF WITH RELATIVE GROUP SPARSITY

According to the terminology described in the introduction we here consider $\mathbf{H}_{(j,g)}$ as groups and $\mathbf{H}_{(j)}$, i.e., activation coefficients corresponding to universal models, as supergroups. In order to keep group sparsity, while assuring supergroups anti-sparsity, and thus hopefully fixing the problem of sources vanishing in the case of our application, we propose replacing the group sparsity-inducing penalty (5) by the following relative group sparsity-inducing penalty

$$\Psi_{\mathrm{rel}}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log \left(\frac{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}{\|\mathbf{H}_{(j)}\|_1^{\gamma_j}} \right), \qquad (7)$$

where γ_j are some non-negative constants. The new penalty (7) can be also rewritten as

$$\Psi_{\mathrm{rel}}(\mathbf{H}) = \Psi(\mathbf{H}) - \sum\nolimits_{j=1}^{J} \lambda_j \gamma_j G_j \log \left(\|\mathbf{H}_{(j)}\|_1 \right), \qquad (8)$$

and one can easily understand that, while the new penalty keeps the group sparsity property thanks to $\Psi(\mathbf{H})$, it prevents the ℓ_1-norms of the supergroups from vanishing. Indeed, if $\|\mathbf{H}_{(j)}\|_1$ tends to zero, then $-\log\left(\|\mathbf{H}_{(j)}\|_1\right)$ tends to $+\infty$. Note also that this formulation generalizes the group sparsity approach, since (7) reduces to (5) for $\gamma_j = 0$.

To derive MU rules optimizing criterion (4) (or (6)) with a new penalty $\Psi_{\mathrm{rel}}(\mathbf{H})$ we relied on MU rules derivation heuristics as described in [14, 16]. The resulting MU rules for criterion (4) are summarized in Algorithm 2, where $\mathbf{Q}_{(j,g)}$ is a matrix with equal entries; its size is the same as $\mathbf{H}_{(j,g)}$, and $\mathbf{Q}$ is a matrix concatenating all $\mathbf{Q}_{(j,g)}$ (the same for $\mathbf{P}_{(j,g)}$ and $\mathbf{P}$, as above).

Algorithm 2 MU rules for NMF with relative group sparsity

Input: $\mathbf{V}, \mathbf{W}, \boldsymbol{\lambda}$
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
      $\mathbf{Q}_{(j,g)} \leftarrow \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}$
    end for
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}) + \mathbf{Q}}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  until convergence

During our preliminary studies we also considered the following alternative version of penalty (7)

$$\Psi_{\mathrm{rel2}}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log \left(\epsilon + \frac{\|\mathbf{H}_{(j,g)}\|_1}{\|\mathbf{H}_{(j)}\|_1^{\gamma_j}} \right). \quad (9)$$

The latter penalty cannot be nicely decoupled as penalty (7) in (8). However, it has another potentially attractive property consisting in the fact that with $\gamma_j = 1$ it is fully invariant to the ℓ_1-norm of each supergroup $\mathbf{H}_{(j)}$. Since experimentally we have not found a big difference between penalty (7) and penalty (9) in terms of source separation performance, and since MU updates for penalty (9) are more complex, we decided to use penalty (7) in our experiments.

5. RESULTS

Note first that an informal analysis of activation matrices $\mathbf{H}$ has shown that the source vanishing problem does not happen any more for any value of λ_j and in both supervised and semi-supervised cases. This can be observed in the right column of Fig. 1.

5.1. Dataset and parameter setting

Our experimental setup, that we recall briefly below, is almost as in our previous work on on-the-fly audio source separation [6], except that the number of mixtures is increased from 10 to 15, we here use two sound search engines instead of one in [6], and some parameters are better tuned.

The test dataset we used consists of 15 single-channel mixtures of two sources artificially mixed at 0 dB SNR. The mixtures were sampled at either 16000 Hz or 11025 Hz and vary in duration between 1 and 13 seconds. The sources in the mixtures represent different types of sound ranging from human speech to musical instruments and animal sounds. In our experiments, the example wave files were retrieved from `www.findsounds.com` and `www.freesound.org`. The keywords used included *guitar*, *drum*, *cat*, *dog*, *river*, *chirps*, *rooster*, *bells*, and *traffic*.

For parameter settings, a frame length of 47 ms with 50% overlapping was used for the STFT. The number of iterations for MU updates in all algorithms was 200 for training and 100 for separation. The number of NMF components for each spectral model learned from one example in the universal model was set to 32. Parameters γ_j for relative block sparsity were set to 1. Parameters λ_j were set to $\lambda_0 FNL_j$ where λ_0 was tuned. The number of background components K_{b} in the semi-supervised case was set to 10.

Method	NSDR	NSIR
Block sparsity ($\lambda_0 = 1 \times 10^{-4}$)	5.14	9.80
Component sparsity ($\lambda_0 = 1 \times 10^{-6}$)	5.91	10.67
Relative block sparsity ($\lambda_0 = 1 \times 10^{-4}$)	4.78	9.27
Relative component sparsity ($\lambda_0 = 1 \times 10^{-6}$)	**6.15**	**10.70**

Table 1. Supervised case: Source separation performances averaged over mixtures.

Method	NSDR	NSIR
Block sparsity ($\lambda_0 = 1$)	0.74	4.66
Component sparsity ($\lambda_0 = 2 \times 10^{-8}$)	1.98	6.22
Relative block sparsity ($\lambda_0 = 4 \times 10^{-4}$)	1.68	6.03
Relative component sparsity ($\lambda_0 = 5 \times 10^{-7}$)	**2.31**	**6.64**

Table 2. Semi-supervised case: Source separation performances averaged over mixtures and keywords.

5.2. Simulations

To evaluate source separation performance we used the normalized signal-to-distortion ratio (NSDR) and and the normalized signal-to-interference ratio (NSIR) [17, 18]. We tested four different methods characterized by block vs. component sparsity and by group sparsity vs. relative group sparsity in both supervised and semi-supervised cases. In the semi-supervised case only one of two keywords was retained for each mixture. The group sparsity penalty parameter λ_0 was tuned over a grid for each tested method and λ_0 leading to the highest NSDR was retained in each case. Results are summarized in tables 1 and 2. As can be seen, relative group sparsity improves the results over group sparsity in 3 of 4 cases. This is especially significant in the semi-supervised case where an improvement of 0.94 (0.33) dB NSDR is achieved by relative block (component) sparsity over block (component) sparsity; although the source separation performance of all methods drops significantly compared to the supervised case. Finally, as in [6], we note that component-based sparsity outperforms block-based sparsity owing to its flexibility in choosing the most appropriate spectral components; in particular, the best results in both supervised and semi-supervised cases are obtained by the relative component sparsity method.

6. CONCLUSION

In this paper, we introduced the notion of relative group sparsity for NMF which prevents some bigger groups (supergroups) of coefficients from converging to zero. We formulated practical criteria and algorithms for relative group sparsity and investigated it within the framework of on-the-fly audio source separation, where supergroups correspond to the universal source models. More specifically, we considered two cases: supervised and semi-supervised. Experiments with mixtures containing various sound types showed that the proposed relative group sparsity outperforms conventional group sparsity in both supervised and semi-supervised settings. Future work may be devoted to the investigation of better ways of choosing λ_0 and γ_j as well as for combing relative block sparsity and relative component sparsity-inducing penalties within the same optimization criterion. In addition, while we examined the notion of relative group sparsity in the context of on-the-fly audio source separation and within the NMF-based approaches, we believe that it can be useful for other dictionary-based signal decompositions.

7. REFERENCES

[1] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Interspeech*, 2012, pp. 17–20.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[3] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[4] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[5] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar 2009.

[6] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[7] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[8] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[9] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Int. Conf. on Music Information Retrieval (ISMIR)*, 2012, pp. 115–120.

[10] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 883–887.

[11] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.

[12] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE Int. Conf. on Consumer Electronics - Berlin (ICCE-Berlin)*, 2014.

[13] K. Chatfield and A. Zisserman, "Visor: Towards on-the-fly large-scale object category retrieval," in *Asian Conference on Computer Vision*, vol. 7725 of *Lecture Notes in Computer Science*, pp. 432–446. Springer, 2012.

[14] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

[15] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural and Information Processing Systems 13*, 2001, pp. 556–562.

[17] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 90–93.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.