Efficient Model Selection for Speech Enhancement Using a Deflation Method for Nonnegative Matrix Factorization

Minje Kim
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
minje@illinois.edu

Paris Smaragdis
University of Illinois at Urbana-Champaign,
Adobe Research
paris@illinois.edu

***Abstract*—We present a deflation method for Nonnegative Matrix Factorization (NMF) that aims to discover latent components one by one in order of importance. To do so we perform a series of individual decompositions, each of which stands for a deflation step. In each deflation we obtain a dominant component and a nonnegative residual, and then the residual is further used as an input to the next deflation in case we want to extract more components. With the help of the proposed additional inequality constraint on the residual during the optimization, the accumulated latent components at any given deflation step can approximate the input to some degree, whereas NMF with an inaccurate rank assumption often fail to do so. The proposed method is beneficial if we need efficiency in deciding the model complexity from unknown data. We derive multiplicative update rules similar to those of regular NMF to perform the optimization. Experiments on online speech enhancement show that the proposed deflation method has advantages over NMF: namely a scalable model structure, reusable parameters across decompositions, and resistance to permutation ambiguity.**

***Index Terms*—Blind source separation, Speech enhancement**

I. Introduction

Nonnegative Matrix Factorization [1], [2] has been widely used for audio analysis, because its parts-based representation matches the additive nature of audio mixtures, e.g. a speech signal with cross-talks, music with simultaneous playing of instruments, etc. A tricky part of NMF learning is that the user has to specify an appropriate number of latent variables. On the contrary to the other linear decomposition models with straightforward deflation methods, such as the power iteration for Principal Component Analysis (PCA), NMF therefore suffers from a series of problems: poor reusability that requires training from scratch whenever we add or remove latent variables and a permutation ambiguity of the estimated components.

In this paper we present a deflation method for NMF that also preserves the original parameter nonnegativity and parts-based representation. Our goal is to achieve an incremental way of adding latent variables on top of the previously learned model, in order to better refine it. This approach is computationally cheaper than starting over the entire NMF run with a different guess about the model complexity. Additionally, it learns components in order of importance as they get discovered. While the desired properties are achieved by employing an additional inequality constraint as in [3], we also introduce an explicit use of a residual matrix to streamline the optimization, which is eventually more similar to that of the standard NMF algorithm. As a result, multiplicative update rules are proposed to learn the optimal basis at each deflation that best describes the dominant component. Meanwhile, a nonnegative residual is also learned, which is then sent as an input to the next decomposition.

We propose to use this deflation NMF method for an online semi-supervised speech enhancement task where the proposed method plays a big role in estimating ranks of the various kinds of unknown noise, the job the ordinary NMF is not suitable for.

II. Nonnegative Matrix Factorization

A. NMF as a Constrained Optimization Problem

NMF takes a nonnegative matrix $V \in \mathbb{R}_+^{M \times N}$ as an input, where $\mathbb{R}_+$ stands for nonnegative real values. Then, it finds factor matrices $W \in \mathbb{R}_+^{M \times K}$ and $H \in \mathbb{R}_+^{K \times N}$, whose product minimizes an error function, $\mathcal{D}(V||WH)$, between the input and the reconstruction. It is common to use β-divergence as a generalized error function,

$$\mathcal{D}_\beta(x||y) = \begin{cases} \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)}, & \beta \in \mathbb{R}\backslash\{0,1\} \\ x(\log x - \log y) + (y - x), & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1, & \beta = 0, \end{cases} \tag{1}$$

since it covers Frobenius norm, unnormalized KL-divergence, and Itakura-Saito divergence when $\beta = 2$, 1, and 0, respectively, as special cases [4]. NMF with β-divergence is defined as follows:

$$\arg\min_{W,H} \mathcal{D}_\beta(V||WH), \text{s.t. } W \geq 0, H \geq 0, \tag{2}$$

where the inequality holds element-wise. The NMF algorithm solves this constrained optimization by representing the gradient descent method via a set of multiplicative update rules. For the nonnegative initial parameters W and H, the update rules are:

$$W \leftarrow W \odot \frac{\left\{(WH)^{\cdot(\beta-2)} \odot V\right\} H^\top}{(WH)^{\cdot(\beta-1)} H^\top}, \; H \leftarrow H \odot \frac{W^\top \left\{(WH)^{\cdot(\beta-2)} \odot V\right\}}{W^\top (WH)^{\cdot(\beta-1)}}, \tag{3}$$

where $\odot$ represents the Hadamard product and division and exponentiation are carried in the element-wise manner, too.

B. Properties of NMF

NMF and its extension have been popular for music transcription [5]. In Figure 1 we see results from NMF runs on a signal with three notes (C5-G5-C6) when $\beta = 2$. The input matrix (a) is a magnitude spectrogram of the signal after applying Short-Time Fourier Transform (STFT). After learning three NMF components we see in (b) that the spectral patterns of the three notes are represented as columns of the matrix W, whereas H tells us the temporal activations. However, these results are permuted, i.e. the most significant note (the third one) is captured by the second component. Furthermore, we see in (c) and (d) that NMF runs with a suboptimal number of components fail to analyze the notes accurately.

III. The Deflation Method for NMF

With an additional inequality constraint on the original NMF objective function, we can achieve the desired deflation ability. We use the superscript $^{(i)}$ to indicate variables involved in the i-th deflation. The goal is to reconstruct the i-th input, $V^{(i)}$, with i-th parameter vectors $w^{(i)} \in \mathbb{R}_+^{M \times 1}$ and $h^{(i)} \in \mathbb{R}_+^{1 \times N}$. For instance,

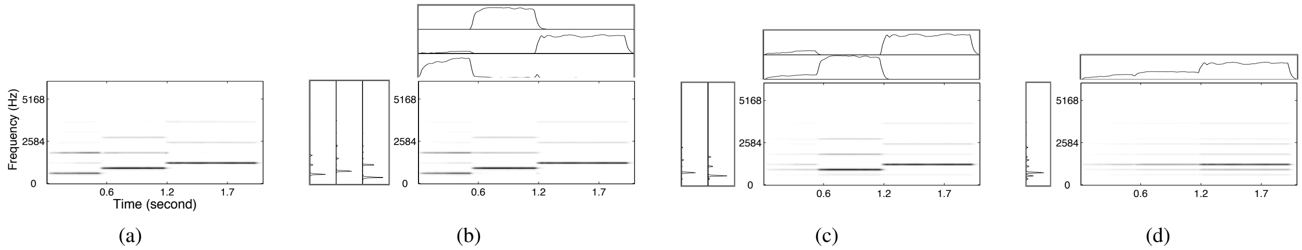


Fig. 1. NMF runs on 3 musical notes. (a) The input spectrogram (b) NMF results with three assumed components (c) two components (d) only one component

$V^{(1)}$ is simply the input matrix whereas $V^{(2)} = V^{(1)} - w^{(1)}h^{(1)}$, and so on. In i-th deflation, we solve the following opimization problem:

$$\underset{w^{(i)},h^{(i)}}{\arg\min} \quad |V^{(i)} - w^{(i)}h^{(i)}|_F^2$$
$$\text{s.t. } w^{(i)} \geq 0,\ h^{(i)} \geq 0,\ V^{(i)} \geq w^{(i)}h^{(i)}, \tag{4}$$

where the additional inequality constraint $V^{(i)} \geq w^{(i)}h^{(i)}$ prevents the reconstruction $w^{(i)}h^{(i)}$ from exceeding the input. Therefore, it is different from running NMF with only one component as in Figure 1 (d), because the NMF reconstruction breaks this constraint with spurious harmonics. Nonnegative Matrix Underapproximation (NMU), which is defined with the Frobenius norm as its error function, solves the optimization problem by using Lagrange multipliers for the new inequality constraint [3]. Its update rules are divided into two parts: updates for original parameters, $w^{(i)}$ and $h^{(i)}$, using Hierarchical Alternating Least Squares (HALS) [6], and a gradient descent for the Lagrange multipliers. In the following section we generalize this problem with β-divergence and propose a more compact optimization by employing a residual matrix as an unknown parameter, which in turn works like the Lagrange multipliers in the HALS updates, but with its own nonnegativity constraint.

A. The Deflation Update Rules with Nonnegative Residuals

With another parameter $R^{(i)}$ for the residuals, $R^{(i)} = V^{(i)} - w^{(i)}h^{(i)}$, now the optimization is defined as follows:

$$\underset{w^{(i)},h^{(i)},R^{(i)}}{\arg\min} \quad \mathcal{D}_\beta(V^{(i)}||w^{(i)}h^{(i)} + R^{(i)}) + \frac{1}{2}\lambda\ |R^{(i)}|_F^2$$
$$\text{s. t.} \quad w^{(i)} \geq 0, \quad h^{(i)} \geq 0, \quad R^{(i)} \geq 0, \tag{5}$$

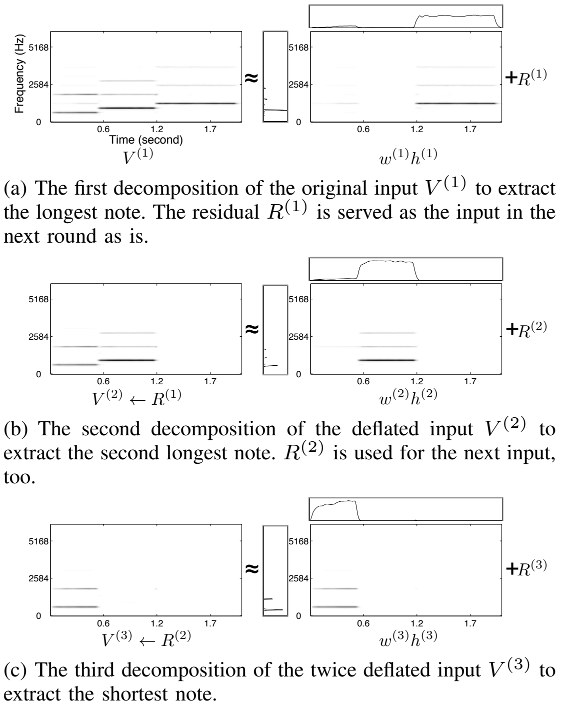
where the additional inequality constraint is absolved into the residual $R^{(i)}$. To avoid trivial solutions, where $R^{(i)} = V^{(i)}$, we also need a regularization term $\frac{1}{2}\lambda|R^{(i)}|_F^2$ with λ as a control parameter. The new constraint with the residual matrix is easier to be incorporated into the multiplicative update rules with β-divergence if we gather the negative and the positive terms of the derivatives into the numerator and the denominator of the multiplier, respectively:

$$w^{(i)} \leftarrow w^{(i)} \odot \frac{\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-2)} \odot V^{(i)}\right\} h^{(i)\top}}{\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-1)}\right\} h^{(i)\top}}, \tag{6}$$

$$h^{(i)} \leftarrow h^{(i)} \odot \frac{w^{(i)\top}\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-2)} \odot V^{(i)}\right\}}{w^{(i)\top}\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-1)}\right\}}, \tag{7}$$

$$R^{(i)} \leftarrow R^{(i)} \odot \frac{\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-2)} \odot V^{(i)}\right\}}{\left\{(w^{(i)}h^{(i)}+R^{(i)})^{\cdot(\beta-1)}\right\} + \lambda R^{(i)}}. \tag{8}$$

The nonnegative residual $R^{(i)}$ plays a big role in the deflation method. First, its nonnegativity constraint prevents the rank-1 reconstruction from exceeding the input, and then it serves as the input to the next deflation. Nothing stops the multiplicative update rules from evolving into more sophisticated optimization techniques, once the residual remains nonnegative throughout the process. However, the approximation error at every deflation tends to be propagated to the next deflation. As a result, the entire approximation gets relatively worse than the oracle NMF result with the correct rank assumption when the true rank of the input becomes larger. However, we believe that the advantages of the proposed method, i.e. ordered components, reusability of parameters, and the ability of estimating model complexity, can compensate for the downside in some applications.



(a) The first decomposition of the original input $V^{(1)}$ to extract the longest note. The residual $R^{(1)}$ is served as the input in the next round as is.

(b) The second decomposition of the deflated input $V^{(2)}$ to extract the second longest note. $R^{(2)}$ is used for the next input, too.

(c) The third decomposition of the twice deflated input $V^{(3)}$ to extract the shortest note.

Fig. 2. The deflation NMF results on the three notes.

B. Properties of the deflation NMF

Figure 2 demonstrates the advantage of the deflation NMF on the same signal used in Figure 1. In (a) it first learns the longest note (C6) as the most important component in terms of the signal energy. In the second run in (b), we learn the second longest note, G5. In (c) the shortest note C5 is finally extracted.

C. A Comparison of the Convergence Behaviors

Figure 3 is the proportion of the absolute error to the sum of the input at every iteration. It can be calculated with $\frac{\sum_{f,t}|V-WH|_{f,t}}{\sum_{f,t}V_{f,t}}$ in the NMF case and $\frac{\sum_{f,t}|V-\sum_i w^{(i)}h^{(i)}|_{f,t}}{\sum_{f,t}V_{f,t}}$ in the deflation case, repsectively. At every run (or at every deflation) with a fixed number

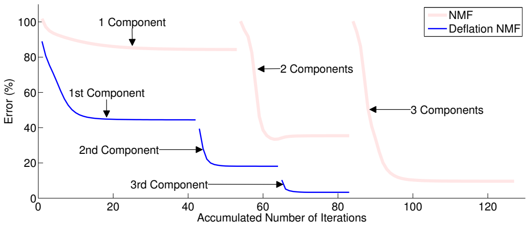


Fig. 3. Error curves of NMF and deflation NMF when adding components: thick pink lines for NMF and thin blue lines for the deflation method. The discontinuities represent either the addition of a new component (deflation) or starting over with a new number of components (NMF).

of components the algorithms stop if five error values in a row are in the range of 0.01%. The bigger loop of the systems, which decides the number of components, stops when the error reaches below 10%. In the three notes case, both NMF and the deflation NMF could achieve this amount of error with three components.

As the figure shows, we can see that the first deflated component of the proposed method explains more than 50% of the input, whereas NMF with only one component explains about 15%. This correlates with the NMF reconstruction of only one component in Figure 1 (d) being so poor. By trials with more components, namely two and three, NMF can provide smaller errors. However, all the three trials start from high error values as NMF randomly re-initializes its parameters without any hint from the previous runs. On the other hand, the deflation method starts from the residual of the previous run, so that it can continue to reduce the error by adding more components. It is true that NMF with correct number of components, three in our case, converges after only 44 iterations, whereas the deflation method consumes 84 iterations. However, if we take the previous wrong guesses of NMF into account the total number of NMF iterations increases up to 127. Likewise, the proposed deflation method can converge faster when NMF has to find the unknown model complexity by trial and error.

D. Computational Complexity

The standard and deflation update rules of NMF have complexity of $\mathcal{O}(MKN)$ and $\mathcal{O}(MN)$ per an iteration, respectively. If we assume that they converge at the same number of iterations I and we know the optimal rank K, both algorithms have complexity of $\mathcal{O}(MKNI)$, since we do the deflation K times. However, if we have to try different ranks from 1 to K for the NMF runs, the complexity of NMF grows to $\mathcal{O}(K^2MNI)$, which is worse than the unaffected complexity of the deflation method.

E. A Compromised Version: Group Deflation

We also present a compromised version of the deflation method to mitigate propagation of reconstruction errors down to the next deflation. Suppose the input matrix can be ideally decomposed into K latent components, such as $V = \sum_{i=1}^{K} Y^{(i)} = \sum_{i=1}^{K} w^{(i)} h^{(i)}$, where $Y^{(i)}$ stands for i-th ideal latent component with rank-1. It is possible that there exists an reconstruction error between $Y^{(i)}$ and $w^{(i)} h^{(i)}$, which is not guaranteed to be addressed in the succeeding $(i+1)$-th decomposition. Therefore, it is usual that the deflation error is propagated as we keep adding more components.

To resolve this issue we relax the algorithm so that it allows a group of components to be added simultaneously rather than only one at a time. Although we lost ordering among the components in the same group, we can reduce the number of deflation from K to G, where G is the number of groups deflated. This can be useful when a roughly good guess of the number of components is still important.

F. Relationships with the Other Methods

Among the series of Nonnegative Matrix Partial Co-Factorization (NMPCF) algorithms [7]–[9], the unsupervised one in [8] is relevant to the deflation method, where NMPCF extracts common basis vectors from column-blocks of a matrix. It assumes that repeating patterns can be captured by the common basis vectors whereas non-shared bases hold less-repeating ones during its simultaneous decomposition on the multiple column-blocks. The model reduces to the proposed deflation method if we shrink the size of the column-blocks into only one column. Then, a single basis vector $w^{(i)}$ captures the common component across all the input vectors, whereas another basis per a column vector holds the residual component, whose concatenation builds the residual matrix $R^{(i)}$. This is an analogous intuition about the proposed deflation method: if something keeps repeating in most of the input vectors, it tends to be captured by the most significant basis vector.

The proposed model can be also seen as a non-probabilistic alternative to nonparametric Bayesian methods, such as the hierarchical Dirichlet process [10], which attempts to provide the same benefits. For those who are familiar with NMF however, the proposed method is easier to utilize when it comes to different choices of β rather than 1, which only corresponds to the multinomial topic models.

IV. Numerical Experiments

Online speech enhancement is a task where a few frames are available for processing as a buffer. This problem usually assumes that either the type of noise or the identity of the speaker is unknown. Furthermore, since the optimal rank of the noise can differ by their types and at different time points, a proper model complexity estimation is necessary. Also, if this system should run in real-time, we are not supposed to spend a lot of time to investigate the optimal noise model.

An online speech enhancement system with a semi-supervised separation technique was proposed in [11]. It uses an ordinary topic model that corresponds to NMF with KL-divergence, assuming that an exact noise dictionary is always available for an unseen noisy signal. Since it does not know about the speaker identity, the model allocates some basis vectors for the speech part and learns them from the test signal whereas the noise dictionary is fixed. We start from the same idea, but assume that we only know about the speech source, not noise. Although acquiring a clean training signal for a particular speaker is also difficult or expensive, we have some ways to get around this limitation, such as the universal speech model [12]. In this paper, we learned the speaker-specific dictionary W_s (40 bases) in advance from a female speaker in the TIMIT corpus, leaving out a sentence for testing. As for the noise signals, we use 10 different noise signals proposed in [11], whose labels are presented in Fig. 4. Each noise signal is added to the test speech with 0dB Signal-to-Noise Ratio (SNR).

Algorithm 1 summarizes the online separation procedure using the deflation NMF method. We set $N_B = 60, \beta = 1, G_{max} = 8, \lambda = 2$, and $\eta = 0.01$, each of which is for the buffer size, KL-divergence, the maximum number of deflations, the regularization parameter, and the stopping criterion for deflations, respectively. For a given time frame t, a buffer B is defined with the most recent N_B mixture spectra (line 4). After initializing the parameters, e.g. with random numbers (line 5), we perform a semi-supervised separation from line 6 to 12, coupled with the deflation method. In other words, we update the

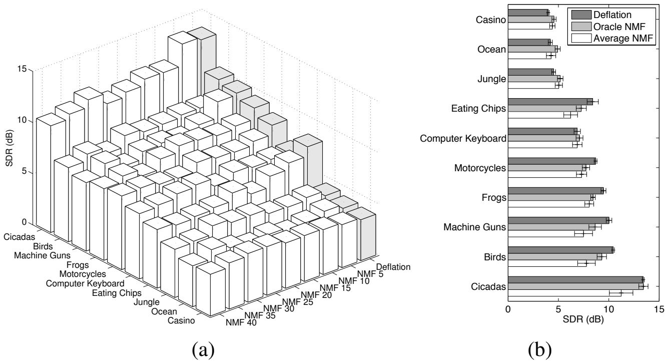
Algorithm 1 Online speech enhancement by using deflation NMF

```
1: Input: Mixture spectra $X_t$ and a speech dictionary $W_s$
2: Define paramters: $N_B, \beta, G_{max}, \lambda, \eta$
3: for $t = N_B$ to $T$ do
4:     $B \leftarrow X_{t-N_B+1:t}, \quad g \leftarrow 1$
5:     Initialize parameters: $H_s, H_n^{(g)}, W_n^{(g)}, R^{(g)}$
6:     repeat
7:         $W \leftarrow [W_s, W_n^{(g)}], \quad H \leftarrow [H_s; H_n^{(g)}]$
8:         $H_s \leftarrow H_s \odot \frac{W_s^\top \{(WH)^{\cdot(\beta-2)} \odot B\}}{W_s^\top \{(WH)^{\cdot(\beta-1)}\}}$
9:         $H_n^{(g)} \leftarrow H_n^{(g)} \odot \frac{W_n^{(g)\top} \{(WH+R^{(g)})^{\cdot(\beta-2)} \odot B\}}{W_n^{(g)\top} \{(WH+R^{(g)})^{\cdot(\beta-1)}\}}$
10:        $W_n^{(g)} \leftarrow W_n^{(g)} \odot \frac{\{(WH+R^{(g)})^{\cdot(\beta-2)} \odot B\} H_n^{(g)\top}}{\{(WH+R^{(g)})^{\cdot(\beta-1)}\} H_n^{(g)\top}}$
11:        $R^{(g)} \leftarrow R^{(g)} \odot \frac{\{(WH+R^{(g)})^{\cdot(\beta-2)} \odot B\}}{\{(WH+R^{(g)})^{\cdot(\beta-1)}\} + \lambda R^{(g)}}$
12:    until Convergence
13:    repeat
14:        $B \leftarrow R^{(g)}, g \leftarrow g + 1$
15:        Initialize parameters: $H_n^{(g)}, W_n^{(g)}, R^{(g)}$
16:        repeat
17:            $H_n^{(g)} \leftarrow H_n^{(g)} \odot \frac{W_n^{(g)\top} \{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-2)} \odot B\}}{W_n^{(g)\top} \{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-1)}\}}$
18:            $W_n^{(g)} \leftarrow W_n^{(g)} \odot \frac{\{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-2)} \odot B\} H_n^{(g)\top}}{\{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-1)}\} H_n^{(g)\top}}$
19:            $R^{(g)} \leftarrow R^{(g)} \odot \frac{\{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-2)} \odot B\}}{\{(W_n^{(g)} H_n^{(g)} + R^{(g)})^{\cdot(\beta-1)}\} + \lambda R^{(g)}}$
20:        until Convergence
21:    until $|R^{(g)}|_F < \eta |X_{t-N_B+1:t}|_F$ or $g \geq G_{max}$
22: end for
```

coupled parameters W and H as if they are the deflation parameters in (6)-(8), except the fixed bases W_s for speech. Once this first decomposition is done we decide whether to deflate more by checking on the norm of the residual matrix $R^{(1)}$ is bigger than the threshold, $\eta = 0.01$, times the norm of the original data. If we decide to proceed, we learn the new parameters for another group of 5 components (line 14 to 20) by using the residual $R^{(1)}$ as input. Likewise, at each new deflation g, we add five additional noise components on top of the current one. We repeat the deflation until we reach the maximal number of deflations or the residual becomes smaller than the threshold (line 13 to 21). For each deflation 10 iterations were enough to converge. Moreover, if we initialize the parameters with the one learned for $(t-1)$-th frame instead of random numbers (line 5 and 15), we can start from good initializations. The system met the stopping criterion before the maximun number of deflation (8 groups) most of the time. After separation, we recover the speech part by multiplying the input with the proportion of the speech reconstruction out of the mixture reconstruction, $X_{t-N_B+1:t} \odot \frac{W_s H_s}{W_s H_s + \sum_g W_n^{(g)} H_n^{(g)}}$.

For a comparison, we conduct a similar online semi-supervised separation as in [11], but with noise components learned from the test signal by the usual NMF updates. In Algorithm 1, the loop for the deflation (line 13 to 21) is discarded in this case. Also, the deflation updates (line 9 to 11) are replaced with ordinary NMF updates as in (3) with a fixed number of components. We tried 8 different numbers of components, $\{5, 10, \cdots, 40\}$, same as in the deflation case. Since we have to fix the number of noise components with one of those 8 numbers throughout the entire separation, the model cannot adapt to the temporal dynamics of the noise.

One can think of an adaptive system for the NMF case as well, where several different ranks are tried for a given frame and the optimal model complexity at every given time is chosen. However, it is not efficient, because we do not have a criterion which noise model is the best without calculating the Signal-to-Distortion Ratio (SDR); running a lot of of different NMF models is a computationally complex procedure with $\mathcal{O}(K^2LMN)$; we cannot reuse the learned model from the previous time frame if the model complexity changes in the next frame.



(a) (b)

Fig. 4. (a) SDR of the enhancement results for each choice of systems (number of components in NMF systems or the adaptive deflation technique) and for each noise. Every bar is an average of 10 repeated experiments with same settings. (b) Comparison of the deflation method versus the oracle NMF result (with the optimal number of components), and the average of all NMF choices. Standard deviation is provided as the error bars.

Figure 4 (a) shows the separation results. First, NMF runs with different choices of noise components produce indeed diverse results. Also, each different type of noise has its own rank that eventually introduces performance variation to the system. Note that the NMF results with highest bars for the given noise types are not actually the optimal models, since there could be a lot of different optima over time. The proposed deflation NMF model, however, can cope with the dynamics and could provide better performances than any NMF settings most of the time.

Figure 4 (b) further emphasizes the merit of the deflation method. The white bars are the average results of all the choices for NMF model complexities from 5 to 40. The deflation method is significantly better than this ordinary setting except some comparable cases. If we compare our results with the oracle NMF results with the globally optimal number of noise components (the maximum of the NMF results per each noise in Figure 4 (a)), the proposed method is still better. It is also noticeable that the deflation method enjoys less performance variation with narrower error bars than the oracle NMF results.

V. Conclusion

This paper presented a deflation strategy for NMF to incrementally learn the latent components. To this end we proposed an explicit use of a nonnegative residual, which is also estimated in the multiplicative manner similarly to the other NMF parameters. Experimental results on the speech source separation task show that the method works with expected properties, such as the flexible model structure, reuse of estimated parameters, and the control over the permutation. As a future work, we plan to explore the generative models of the proposed method and their relationships to existing nonparametric Bayesian models.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1319708.

References

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[2] ——, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[3] N. Gillis and F. Glineur, “Using underapproximations for sparse non-negative matrix factorization,” *Pattern Recognition*, vol. 43, no. 4, pp. 1676–1687, Apr. 2010.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[5] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[6] A. Cichocki, R. Zdunek, and S. Amari, “Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization,” in *Lecture Notes in Computer Science*. Springer, 2007, vol. 4666, pp. 169–176.

[7] J. Yoo, M. Kim, K. Kang, and S. Choi, “Nonnegative matrix partial co-factorization for drum source separation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[8] M. Kim, J. Yoo, K. Kang, and S. Choi, “Blind rhythmic source separation: Nonnegativity and repeatability,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[9] ——, “Nonnegative matrix partial co-factorization for spectral and temporal drum source separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[11] Z. Duan, G. J. Mysore, and P. Smaragdis, “Online plca for real-time semi-supervised source separation,” in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 34–41.

[12] D. L. Sun and G. J. Mysore, “Universal speech models for speaker independent single channel source separation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.