MULTIPLICATIVE UPDATE RULES FOR NONNEGATIVE MATRIX FACTORIZATION WITH CO-OCCURRENCE CONSTRAINTS

Steven K. Tjoa and K. J. Ray Liu

Signals and Information Group, Dept. of Electrical and Computer Engineering
University of Maryland – College Park, MD 20742 USA
{kiemyang, kjrliu}@umd.edu

ABSTRACT

Nonnegative matrix factorization (NMF) is a widely-used tool for obtaining low-rank approximations of nonnegative data such as digital images, audio signals, textual data, financial data, and more. One disadvantage of the basic NMF formulation is its inability to control the amount of dependence among the learned dictionary atoms. Enforcing dependence within predetermined groups of atoms allows objects to be represented using multiple atoms instead of only one atom. In this paper, we introduce three simple and convenient multiplicative update rules for NMF that enforce dependence among atoms. Using examples in music transcription, we demonstrate the ability of these updates to represent each musical note with multiple atoms and cluster the atoms for source separation purposes.

Index Terms— Dictionary learning, sparse coding, music transcription, source separation.

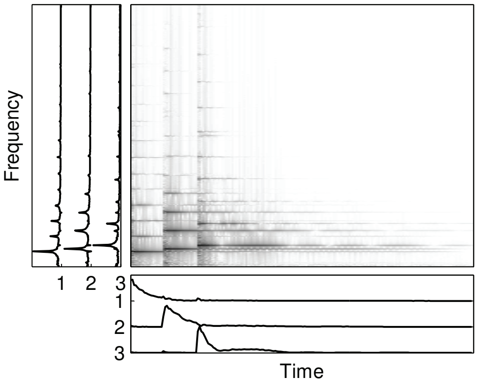

Fig. 1. Nonnegative matrix factorization of the spectrogram $\mathbf{X}$ (top right) into $\mathbf{A}$ (top left) and $\mathbf{S}$ (bottom right) for three piano notes.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) has become a popular tool for discovering structure in a variety of signals. Given a nonnegative matrix $\mathbf{X}$, the objective of NMF is to find two nonnegative matrices, $\mathbf{A}$ and $\mathbf{S}$, that minimizes some divergence between $\mathbf{X}$ and $\mathbf{AS}$. The NMF problem was originally popularized by Paatero and Tapper [1], and Lee and Seung later proposed algorithms for solving the NMF problem using multiplicative update rules [2, 3].

Fig. 1 illustrates the use of NMF when applied to a musical audio signal. The matrix $\mathbf{X}$ represents the magnitude spectrogram of an audio signal containing three notes played by a piano. After decomposition into a rank-three approximation using NMF, the three columns of $\mathbf{A}$ – also referred to as *dictionary atoms* – represent the frequency spectra of the three piano notes, and the three rows of $\mathbf{S}$ represent their corresponding temporal activities. Note how the columns of $\mathbf{A}$ accurately represent the spectra of the three piano notes. To perform source separation, we reconstruct an estimate of the spectrogram for an individual source using only a subset of the learned dictionary atoms.

When used for source separation, the basic formulation of NMF has notable disadvantages. Objects may require more than a single dictionary atom in order to be approximated accurately. For example, Fig. 2 illustrates the decomposition of a spectrogram of an audio signal containing one note played by a violin. Although only one note is played, the vibrato induced by the performer causes the pitch to modulate. As a result, a rank-one approximation is not sufficient to capture this pitch modulation. A user can select multiple dictionary atoms to represent one note. However, in the presence of many other sources, it is unclear which atoms to select. In other words, after learning is complete, there is no straightforward way to cluster multiple dictionary atoms that belong to the same source.

One solution that addresses these problems is to enforce dependence among sets of dictionary atoms by introducing *co-occurrence constraints* – constraints that specify which dictionary atoms are dependent, or co-occur. These co-occurrence constraints have shown to be useful for describing sources with multiple, co-occurring dictionary atoms. Smaragdis et al. [4] proposed the use of cross entropy to enforce the similarity between dictionary atoms belonging to the same source. The atoms are then easily grouped into sets. By decomposing a spectrogram of a drums recording, Smaragdis et al. illustrate that co-occurrence constraints allow each drum sound to be represented more accurately by two dictionary atoms instead of one.

In this paper, we introduce three new update rules to enforce dependence among dictionary atoms by incorporating co-occurrence constraints into NMF. These rules are conceptually simple, easy to implement, and effective for describing sources using multiple dictionary atoms. First, we formulate the NMF problem with co-occurrence constraints. Then, we derive new update rules for minimizing three common divergence metrics. Finally, we illustrate the use of these update rules in the context of music transcription.

2. PROBLEM FORMULATION

The basic NMF problem is formulated as follows. Given a nonnegative matrix $\mathbf{X} \in \mathbb{R}_+^{M \times N}$, we must find nonnegative matrices $\mathbf{A} \in \mathbb{R}_+^{M \times K}$ and $\mathbf{S} \in \mathbb{R}_+^{K \times N}$ that minimize some divergence metric,

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

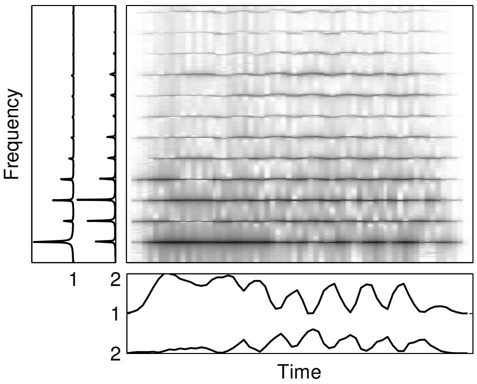

Fig. 2. Nonnegative matrix factorization of the spectrogram $\mathbf{X}$ (top right) into $\mathbf{A}$ (top left) and $\mathbf{S}$ (bottom right) for one violin note. Two atoms are required to capture the pitch modulation due to vibrato.

$\mathrm{d}(\mathbf{X}, \mathbf{AS})$. For any two matrices $\mathbf{X}$ and $\mathbf{Y}$, we define $\mathrm{d}(\mathbf{X}, \mathbf{Y}) = \sum_{m,n} \mathrm{d}(x_{mn}, y_{mn})$. The three divergence metrics we consider are the Euclidean distance,

$$\mathrm{d}_{\mathrm{EUC}}(x, y) = |x - y|^2 \,, \tag{1}$$

the Kullback-Leibler divergence,

$$\mathrm{d}_{\mathrm{KL}}(x, y) = x \log \frac{x}{y} - x + y \,, \tag{2}$$

and the Itakura-Saito divergence,

$$\mathrm{d}_{\mathrm{IS}}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1 \,. \tag{3}$$

These three divergences are special instances of the generalized Bregman divergence [5].

In this paper, for conciseness of notation, we will use $\mathbf{X} \cdot \mathbf{Y}$ to denote element-wise multiplication of matrices $\mathbf{X}$ and $\mathbf{Y}$, $\frac{\mathbf{X}}{\mathbf{Y}}$ to denote element-wise division, and $\mathbf{X}^2$ to denote element-wise exponentiation. Also, we use $\mathbf{1}$ to denote a matrix of ones of appropriate dimension.

Multiplicative update rules for $\mathbf{A}$ and $\mathbf{S}$ have been derived to minimize each of the three divergences [3, 5]. The basic update rules are as follows for the Euclidean distance,

$$\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\mathbf{XS}^T}{\mathbf{ASS}^T} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{A}^T\mathbf{X}}{\mathbf{A}^T\mathbf{AS}} \,, \tag{4}$$

Kullback-Leibler divergence,

$$\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\frac{\mathbf{X}}{\mathbf{AS}}\mathbf{S}^T}{\mathbf{1S}^T} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{A}^T\frac{\mathbf{X}}{\mathbf{AS}}}{\mathbf{A}^T\mathbf{1}} \,, \tag{5}$$

and Itakura-Saito divergence,

$$\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\frac{\mathbf{X}}{(\mathbf{AS})^2}\mathbf{S}^T}{\frac{\mathbf{1}}{\mathbf{AS}}\mathbf{S}^T} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{A}^T\frac{\mathbf{X}}{(\mathbf{AS})^2}}{\mathbf{A}^T\frac{\mathbf{1}}{\mathbf{AS}}} \,. \tag{6}$$

Given a choice of divergence metric, the update rules for $\mathbf{A}$ and $\mathbf{S}$ are usually applied alternately. Because $\mathrm{d}(\mathbf{X}, \mathbf{AS})$ is not jointly convex in $(\mathbf{A}, \mathbf{S})$, the global minimum may not necessarily be achieved. However, it is convex in $\mathbf{A}$ and $\mathbf{S}$ individually, this guaranteeing decrease at each iteration.

To introduce co-occurrence constraints, we must influence the value of the inner product $\mathbf{s}_i^T\mathbf{s}_j$, where $\mathbf{s}_k^T$ is the k^{th} row of $\mathbf{S}$. For instance, using the musical examples in Figs. 1 and 2, a large value for $\mathbf{s}_1^T\mathbf{s}_2$ would indicate that dictionary atoms 1 and 2 co-occur heavily in time, while $\mathbf{s}_1^T\mathbf{s}_2 = 0$ would indicate that the atoms do not co-occur at all. This problem can be formulated as follows:

$$\min_{\mathbf{S}} \mathrm{d}(\mathbf{Q}, \mathbf{SS}^T) \quad \text{s.t. } \mathbf{S} \in \mathbb{R}_+^{K \times N} \,, \tag{7}$$

where $\mathbf{Q} \in \mathbb{R}_+^{K \times K}$ is a pre-defined symmetric matrix such that q_{ij} is low when atoms i and j are not dependent and q_{ij} is high when the atoms are desired to be highly dependent. First, choosing $q_{ii} = 1$ for all i performs normalization upon each row of $\mathbf{S}$. Then we can set $0 \ll q_{ij} \leq 1$ for all pairs of atoms i and j that we desire to be dependent and $0 \leq q_{ij} \ll 1$ for all other pairs of atoms. For d_{KL} and d_{IS}, q_{ij} must be strictly greater than zero for all i and j.

3. PROPOSED UPDATE RULES

Following the derivations by Lee and Seung [3], we derive multiplicative update rules for $\mathbf{S}$ to minimize $\mathrm{d}(\mathbf{Q}, \mathbf{SS}^T)$ for each of the three divergences. Using the Euclidean distance as an example, first we explicitly define $\mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{SS}^T)$:

$$\begin{aligned} \mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{SS}^T) &= ||\mathbf{Q} - \mathbf{SS}^T||_F^2 && (8) \\ &= \mathrm{tr}((\mathbf{Q} - \mathbf{SS}^T)^T(\mathbf{Q} - \mathbf{SS}^T)) \,, && (9) \end{aligned}$$

where $||\mathbf{X}||_F^2$ is the squared Frobenius norm of $\mathbf{X}$, and $\mathrm{tr}(\mathbf{X})$ is the trace of $\mathbf{X}$. Next, we differentiate $\mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{SS}^T)$ with respect to $\mathbf{S}$:

$$\frac{\partial}{\partial \mathbf{S}} \mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{SS}^T) \propto \mathbf{SS}^T\mathbf{S} - \mathbf{QS} \,. \tag{10}$$

Finally, as illustrated by Lee and Seung [3], we construct the multiplicative update term by placing the negative part of $\frac{\partial}{\partial \mathbf{S}}$ in the numerator and the positive part of $\frac{\partial}{\partial \mathbf{S}}$ in the denominator as follows:

$$\mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{QS}}{\mathbf{SS}^T\mathbf{S}} \,. \tag{11}$$

In practice, a small positive number ε is added to the numerator and denominator for three reasons. First, ε prevents division by zero. Second, as long as ε is large enough, this update rule guarantees a decrease in $\mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{SS}^T)$ at each iteration by restricting $\mathbf{S}$ to lie within a local region around the previous instance of $\mathbf{S}$. Third, including ε in this manner does not alter the divergence metric being minimized.

Update rules for $\mathbf{A}$ can be constructed in similar fashion by minimizing $\mathrm{d}(\mathbf{Q}, \mathbf{A}^T\mathbf{A})$. The choice to impose co-occurrence constraints on $\mathbf{A}$ versus $\mathbf{S}$ depends upon the context of the application.

Therefore, we arrive at the finalized update rule for either $\mathbf{A}$ or $\mathbf{S}$ to minimize the Euclidean distance:

$$\boxed{\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\mathbf{AQ} + \varepsilon}{\mathbf{AA}^T\mathbf{A} + \varepsilon} \quad \text{or} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{QS} + \varepsilon}{\mathbf{SS}^T\mathbf{S} + \varepsilon}} \,. \tag{12}$$

Similar derivations using the Kullback-Leibler and Itakura-Saito divergences yield the following two update rules, respectively:

$$\boxed{\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\mathbf{A}\frac{\mathbf{Q}}{\mathbf{A}^T\mathbf{A}} + \varepsilon}{\mathbf{A1} + \varepsilon} \quad \text{or} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\frac{\mathbf{Q}}{\mathbf{SS}^T}\mathbf{S} + \varepsilon}{\mathbf{1S} + \varepsilon}} \tag{13}$$

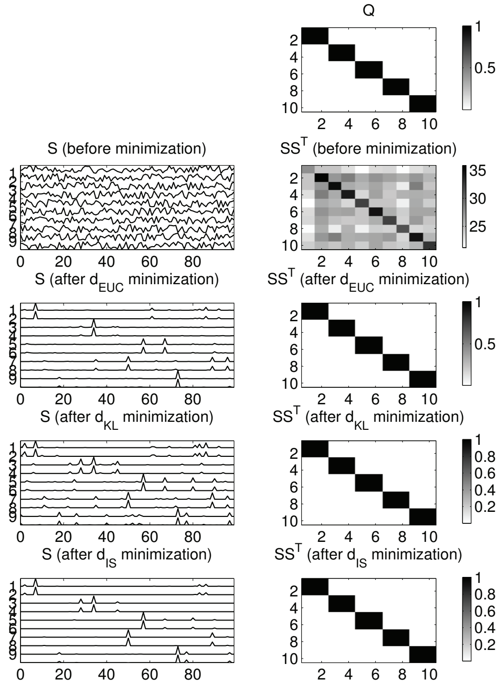

Fig. 3. Minimization of $\mathrm{d}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T)$ for three divergence metrics. Top row: $\mathbf{Q}$. Left column: $\mathbf{S}$ before and after minimization. Right column: $\mathbf{S}\mathbf{S}^T$ before and after minimization.

$$\boxed{\mathbf{A} \leftarrow \mathbf{A} \cdot \frac{\mathbf{A}\frac{\mathbf{Q}}{(\mathbf{A}^T\mathbf{A})^2} + \varepsilon}{\mathbf{A}\frac{\mathbf{1}}{\mathbf{A}\mathbf{A}^T} + \varepsilon} \quad \text{or} \quad \mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\frac{\mathbf{Q}}{(\mathbf{S}\mathbf{S}^T)^2}\mathbf{S} + \varepsilon}{\frac{\mathbf{1}}{\mathbf{S}\mathbf{S}^T}\mathbf{S} + \varepsilon}} \; . \quad (14)$$

To incorporate these co-occurrence constraints into NMF, we first formulate the modified minimization problem, using the Euclidean distance again as an example:

$$\min_{\mathbf{A},\mathbf{S}} \mathrm{d}_{\mathrm{EUC}}(\mathbf{X}, \mathbf{A}\mathbf{S}) + \lambda\, \mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T) \; , \quad (15)$$

where $\lambda > 0$ is a regularization parameter that controls the relative emphasis between $\mathrm{d}_{\mathrm{EUC}}(\mathbf{X}, \mathbf{A}\mathbf{S})$ and $\mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T)$. The proper value for λ depends both upon the data as well as the divergence metric used. Then, using the same method of derivation shown earlier by Lee and Seung [3], we can modify the original update rule for minimizing $\mathrm{d}_{\mathrm{EUC}}$ as follows:

$$\mathbf{S} \leftarrow \mathbf{S} \cdot \frac{\mathbf{A}^T\mathbf{X} + \lambda\mathbf{Q}\mathbf{S} + \varepsilon}{\mathbf{A}^T\mathbf{A}\mathbf{S} + \lambda\mathbf{S}\mathbf{S}^T\mathbf{S} + \varepsilon} \; . \quad (16)$$

Another valid method involves alternating between the updates in Eqs. (4) and (12). Our experiments have shown both options to be roughly equal in accuracy and execution time.

4. EXPERIMENTS

First, we illustrate that the three proposed multiplicative update rules do guarantee decrease in the three divergence metrics at each iteration and that $\mathbf{S}$ does eventually converge. We initialize $\mathbf{S}$ and $\mathbf{Q}$ using the values shown in Fig. 3. We use the update rules in (12), (13), and (14) for 200 iterations to solve the minimization problem in (7) for each of the three corresponding divergence metrics. We see from Fig. 3 that $\mathbf{S}\mathbf{S}^T$ does resemble $\mathbf{Q}$ after minimizing any of the three divergence metrics. By examining $\mathbf{S}$ after convergence, for every pair of rows $(\mathbf{s}_i, \mathbf{s}_j)$ such that $q_{ij} = 1$, we find that $\mathbf{s}_i$ and $\mathbf{s}_j$ are nearly equal.

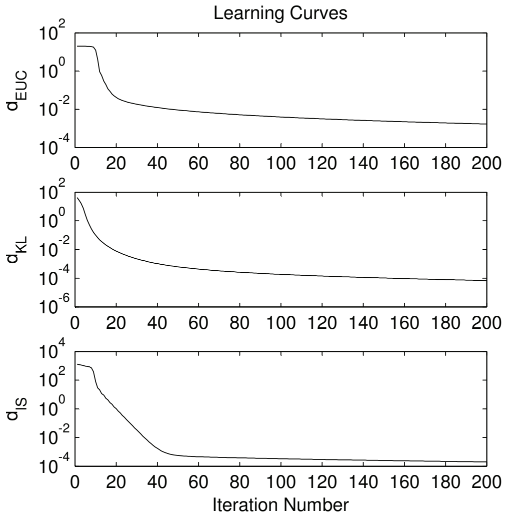

Fig. 4. Learning curves for the examples in Fig. 3. If ε is sufficiently large, then descent of the divergence metrics is guaranteed at each iteration.

Fig. 4 illustrates the learning curves for each of the three minimizations depicted in Fig. 3. The values of $\mathrm{d}_{\mathrm{EUC}}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T)$, $\mathrm{d}_{\mathrm{KL}}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T)$, and $\mathrm{d}_{\mathrm{IS}}(\mathbf{Q}, \mathbf{S}\mathbf{S}^T)$ decrease monotonically as a function of the iteration number, thus confirming that a decrease in the divergence metrics is guaranteed after each iteration of the corresponding update rule as long as ε is sufficiently large. For these experiments, we used $\varepsilon = 0.2$ when minimizing $\mathrm{d}_{\mathrm{EUC}}$, $\varepsilon = 0.2$ when minimizing d_{KL}, and $\varepsilon = 0.6$ when minimizing d_{IS}.

Next, we use the proposed update rules incorporated with NMF to decompose the spectrogram in Fig. 5 containing three notes played by a violin. Because each note is pitch-modulated due to vibrato, multiple atoms are required to accurately represent each note. We initialize a dictionary of six atoms into three groups of two. For the following experiments, we minimize d_{KL} which has qualitatively shown to provide better separation than the other divergence metrics.

Following a co-occurrence model by Wang et al. [6], we define pairs of atoms as "must co-occur", "can co-occur", or "cannot co-occur". We set $q_{ij} = 1$ for atoms that must co-occur, $q_{ij} = 10^{-8}$ for atoms that cannot co-occur, and set $q_{ij} = \mathbf{s}_i^T\mathbf{s}_j$ at each iteration for atoms that can co-occur. In this example, we claim that the atoms representing the second note can co-occur with any of the other atoms, while atoms representing the first note cannot co-occur with atoms of the third note. Atoms within the same group must co-occur by definition. Along with the co-occurrence constraints, to improve the likelihood of co-occurrence within groups, we also impose a smoothness constraint on $\mathbf{S}$ using established NMF modifications [7, 8]. Fig. 5 shows the results of this procedure after minimization. We see that the algorithm does correctly cluster each pair of atoms belonging to the same note. Fig. 6 shows the value of $\mathbf{Q}$ and $\mathbf{S}\mathbf{S}^T$ after minimization.

Finally, we perform the same procedure on the spectrogram in

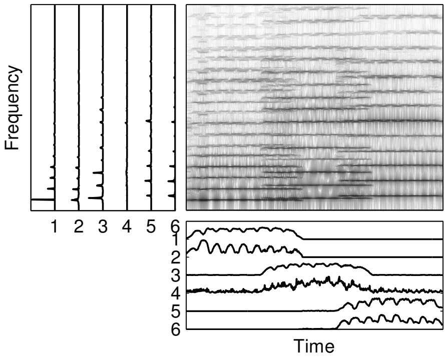

Fig. 5. Factorization of spectrogram with co-occurrence constraints on $\mathbf{S}$ for three violin notes. Six dictionary atoms are grouped into three sets of two atoms.

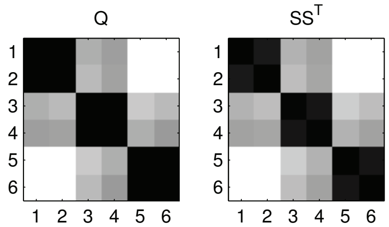

Fig. 6. $\mathbf{SS}^T$ versus $\mathbf{Q}$ for the example in Fig. 5.

Fig. 7 containing five drum beats produced by two drums. However, to enforce dependence in the frequency domain among atoms, we now impose co-occurrence constraints on the columns of $\mathbf{A}$. In a musical context, this constraint is useful whenever there is a percussive sound that emits an initial transient sound followed by a steady-state decaying sound. While the transient and steady-state portions do not co-occur in time, they do overlap considerably in frequency. This behavior is also a property of the piano, xylophone, and similar pitched instruments whose sounds are produced in a percussive manner.

Fig. 7 shows a decomposition using two sets of two atoms each. Each pair of atoms are similar in frequency, as shown in Fig. 8. The transient and steady-state portions of each beat are visible in the matrix $\mathbf{S}$, particularly for the snare drum which occupies a wider frequency bandwidth. Fig. 8 shows the resemblance between $\mathbf{Q}$ and $\mathbf{A}^T\mathbf{A}$ after minimization.

5. CONCLUSIONS

We have proposed novel multiplicative update rules that impose co-occurrence constraints on either of the matrices produced through NMF. These update rules can minimize different divergence metrics, and they integrate easily with the basic NMF multiplicative updates. The constraints are useful when representing objects with multiple atoms, and they provide a natural way to cluster co-occurring atoms. Examples involving music transcription show that these constraints are operate successfully either in the frequency or time domains.

In the future, we believe that these constraints will become useful in many applications addressed by NMF beyond music transcription and source separation. We also plan to investigate the choice of parameters λ and ε which both depend on the magnitude of the data as well as the divergence metric. Intelligent choices for λ and ε can balance the emphasis of the constraints while maximizing the descent per iteration.

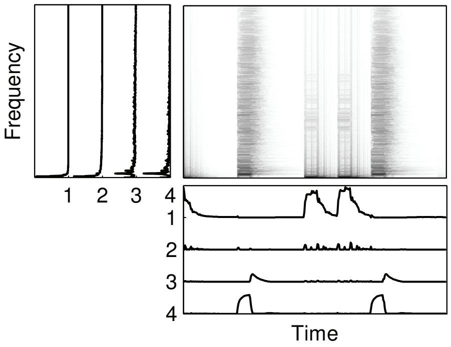

Fig. 7. Factorization of spectrogram with co-occurrence constraints on $\mathbf{A}$ for five drum beats from kick and snare drums. Four dictionary atoms are grouped into two sets of two atoms.

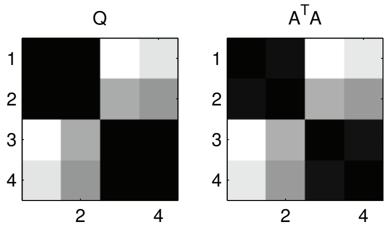

Fig. 8. $\mathbf{A}^T\mathbf{A}$ versus $\mathbf{Q}$ for the example in Fig. 7.

6. REFERENCES

[1] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[2] D.D. Lee and H.S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

[4] P. Smaragdis, M. Shashanka, B. Raj, and G.J. Mysore, “Probabilistic factorization of non-negative data with entropic co-occurrence constraints,” in *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 330–337.

[5] A. Cichocki, R. Zdunek, and S. Amari, “Csiszár’s divergences for non-negative matrix factorization: family of new algorithms,” *Lecture Notes in Computer Science*, vol. 3889, pp. 32, 2006.

[6] F. Wang, T. Li, and C. Zhang, “Semi-supervised clustering via matrix factorization,” in *Proceedings of The 8th SIAM Conference on Data Mining*, 2008.

[7] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[8] A. Cichocki, R. Zdunek, and S. Amari, “New algorithms for non-negative matrix factorization in applications to blind source separation,” in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 2006, vol. 5.