A Block Sparsity Approach to Multiple Dictionary Learning for Audio Modeling

Gautham J. Mysore GMYSORE@ADOBE.COM
Advanced Technology Labs, Adobe Systems Inc., San Francisco, CA 94103, USA

Abstract

Dictionary learning algorithms for audio modeling typically learn a dictionary such that each time frame of the given sound source is approximately equal to a linear combination of the dictionary elements. Since audio is non-stationary data, learning a single dictionary to explain all time frames of the sound source might not be the best modeling strategy. We therefore recently proposed a technique to jointly learn multiple dictionaries such that each time frame of the given sound source is approximately equal to a linear combination of the dictionary elements from one of the many dictionaries. This is equivalent to modeling each time frame with a small subset of all of the dictionary elements in the model, which is analogous to block sparsity on the mixture weights over all dictionary elements. In this paper, we show why there is inherent block sparsity in our model due to its hierarchical nature and why this is useful for audio applications.

1. Introduction

Dictionary learning algorithms have become quite popular for modeling audio (Smaragdis & Brown, 2003; Plumbley et al., 2006). Spectrograms are often the representation of choice as they contain a great deal of structure that can captured by these algorithms. Since spectrograms are non-negative matrices, one of most popular classes of dictionary learning algorithms for modeling audio is non-negative matrix factorization (NMF) (Lee & Seung, 2001) and its probabilistic counterparts such as probabilistic latent component analysis (PLCA) (Smaragdis et al., 2006). These algorithms typically model each time frame of audio (column of the spectrogram) as a linear combination of dictionary elements from a single dictionary (Fig. 1).

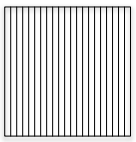

Figure 1. Conceptual model of a single sound source using non-negative matrix factorization. A single dictionary is used to explain the entire sound source.

Since audio is non-stationary data, it could be useful to model a sound source with multiple dictionaries such that each time frame is primarily modeled by a linear combination of the dictionary elements from one of the dictionaries. We therefore recently proposed the non-negative hidden Markov model (N-HMM) (Mysore et al., 2010) (Fig. 2) to do exactly this. It additionally learns a Markov chain to explain the structure of the non-stationarity in the form of temporal dynamics. We found that it is useful for several audio applications (Mysore et al., 2010; Mysore & Smaragdis, 2011; 2012; Han et al., 2012) and yields improved results when compared to using a single dictionary.

Since there is essentially only one dictionary active in each time frame, the mixture weights associated with the other dictionaries are conceptually equal to nearly 0. Therefore, the distribution of mixture weights over all dictionaries tends to be sparse with essentially only a cluster of active mixture weights. This is equivalent to block sparsity (Eldar & Mishali, 2009) on the mixture weights.

2. Probabilistic Models and Block Sparsity

In this section, we first describe the probabilistic model and generative process of the N-HMM. We then ex-

International Conference on Machine Learning Workshop on Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing, Edinburgh, Scotland, UK, 2012.

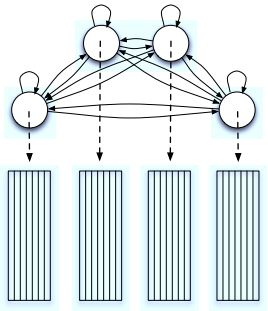

Figure 2. Conceptual model of a single sound source using the non-negative hidden Markov model. Multiple dictionaries account for non-stationarity and the Markov chain accounts for temporal dynamics.

plain why there is inherent block sparsity in this model. Finally, we mention the extension of the N-HMM to model sound mixtures. In these models, each time frame of the spectrogram is viewed as a histogram of "sound quanta" in the same way that a document can be viewed as a histogram of words in topic models (Blei et al., 2003).

2.1. Non-negative Hidden Markov Model

The graphical model is shown in Figure 3. The random variables $D_{1...T}$ form a Markov chain, and the spectra in each time frame are independent given these variables. Each possible value of D_t corresponds to a spectral dictionary. Each dictionary contains a set of dictionary elements (analogous to topics), one of which is selected for each sound quantum by the random variable Z_t. Each dictionary element is a normalized vector over frequencies (analogous to a distribution of words). The frequency associated with a particular quantum is selected by F_t.

The generative process at time frame t is therefore:

1. Choose a dictionary (state):
   $D_t|D_{t-1} \sim \mathsf{Discr}\,[\rho(D_{t-1})]$
2. Repeat for each of v_t quanta:
   - Choose a dictionary element $Z_t \sim \mathsf{Discr}\,[\theta_t(D_t)]$
   - Choose a frequency $F_t \sim \mathsf{Discr}\,[\beta(D_t, Z_t)]$

Here, $\mathsf{Discr}\,[]$ represents the discrete distribution; $\rho(d_t)$ is the column of the Markov transition matrix representing transitions from dictionary d_t; $\theta_t(d_t)$[1] is a vector of normalized mixture weights for dictionary d_t in time frame t; and $\beta(d, z)$ is the normalized dictionary element z of dictionary d. Given the spectrogram of a sound source, these parameters can be estimated using the EM algorithm. The dictionaries and the transition matrix define the model of that sound source whereas the mixture weights are simply the mixing proportions in that specific instance of the sound source. A sample of the dictionaries learned from real speech data is shown in Figure 4.

[1]We denote time–varying distributions and functions with a subscript t.

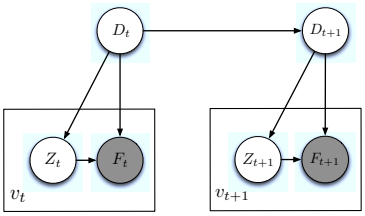

Figure 3. Graphical model of the non-negative hidden Markov model.

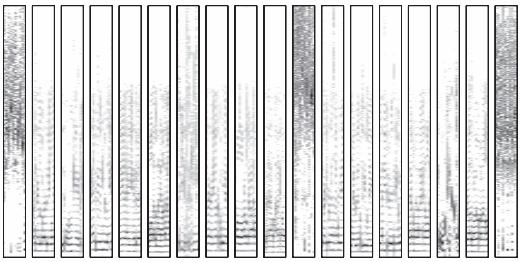

Figure 4. Non-negative hidden Markov model dictionaries that were learned from speech data. Each dictionary contains 10 dictionary elements that are stacked right next to each other. We show a subset of the 40 dictionaries that were learned in this example. We see that these dictionaries roughly correspond to subunits of speech and some are harmonic while others are noise like.

2.2. Block Sparsity

We do not define any explicit notion of block sparsity in our model. However, it has inherent block sparsity due to the hierarchical nature of the model and the iterative nature of the EM algorithm. In this subsection, we provide an intuitive explanation of why this is the case.

As mentioned above, the mixture weights are estimated using the EM algorithm. In the E step, we compute $\gamma_t(d_t)$, which is the marginal posterior probability over the dictionaries, in each time frame. If a given dictionary does not explain the given time frame well, then it will have a low probability in this distribution. Using $\gamma_t(d_t)$, we compute the marginal posterior probability of each dictionary element of each dictio-

nary for each frequency of each time frame as follows:

$$P_t(z_t, d_t|f_t, \bar{\mathbf{f}}) = \gamma_t(d_t) P_t(z_t|f_t, d_t), \quad (1)$$

where $\bar{\mathbf{f}}$ refers to all observations at all time frames. Therefore, if $\gamma_t(d_t)$ has a low probability for dictionary d_t at time frame t, then $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$ will correspondingly be low for all dictionary elements of d_t at all frequencies at that time frame. $P_t(z_t|f_t, d_t)$ is computed using the mixture weights at time t and the dictionary elements of d_t.

In the M step, we use $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$ to estimate the dictionary elements as follows:

$$\beta_f(z, d) = \frac{\sum_t V_{ft} P_t(z, d|f_t, \bar{\mathbf{f}})}{\sum_t \sum_{f_t} V_{ft} P_t(z, d|f_t, \bar{\mathbf{f}})}, \quad (2)$$

where $\beta_f(z, d)$ is the model parameter corresponding to frequency f of dictionary element z of dictionary d. V_{ft} is the magnitude of the input spectrogram data at time t and frequency f. As seen in Eq. 2, $\beta_f(z, d)$ will be estimated such that it primarily explains the time frames in which $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$ is high. It will correspondingly explain the time frames in which $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$ is low, quite poorly.

We can estimate a single distribution over all mixture weights of all dictionaries for time frame t as follows:

$$\theta_t(z_t, d_t) = \frac{\sum_{f_t} V_{ft} P_t(z_t, d_t|f_t, \bar{\mathbf{f}})}{\sum_{z_t} \sum_{d_t} \sum_{f_t} V_{ft} P_t(z_t, d_t|f_t, \bar{\mathbf{f}})}, \quad (3)$$

where $\theta_t(z_t, d_t)$ is the mixture weight of dictionary element z_t of dictionary d_t in time frame t. As with the dictionary elements, the mixture weights are estimated primarily using $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$. If this distribution has a low probability for all z_t of dictionary d_t, then all of the mixture weights that correspond to dictionary d_t will be small. This means that the corresponding block of mixtures weights will have a low value.

In the following EM iteration, $\gamma_t(d_t)$ will correspondingly be low for dictionary d_t in these time frames. Consequently $P_t(z_t, d_t|f_t, \bar{\mathbf{f}})$ will be low for all z_t that corresponds to this dictionary as shown in Eq. 1. The dictionary elements will then be re-estimated with an even lower contribution in these time frames. In practice, over the course of several EM iterations, $\gamma_t(d_t)$ tends to converge to a probability of nearly 0 for this dictionary in these time frames. Similarly, $\gamma_t(d_t)$ tends to converge to a probability of nearly 0 for all but one dictionary, which of course converges to a probability of nearly 1.

Consequently, the mixture weights that correspond to all dictionaries except this dictionary will have a value of nearly 0, which corresponds to block sparsity. The estimated mixture weights for five time frames of speech data are shown in Fig. 5.

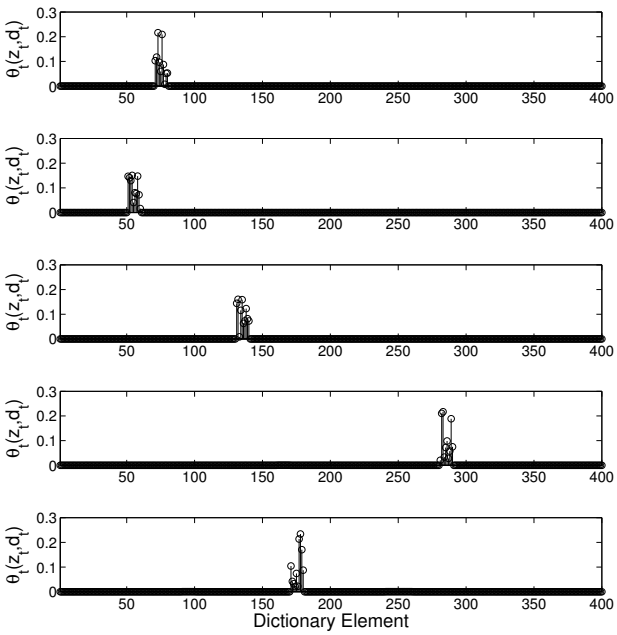

Figure 5. Block sparsity in the mixture weights. This plot shows the mixture weights $\theta_t(z_t, d_t)$ over all dictionary elements of all dictionaries. Each plot corresponds to a different time frame. This particular example has 40 dictionaries with 10 dictionary elements each. Note that we use a single index over all 400 dictionary elements rather than separate indices for d_t and z_t only for plotting purposes. As shown, the mixture weights exhibit a clear block sparsity. In each time frame, all of the mixture weights except for the 10 weights that correspond to a given dictionary, have a probability of nearly 0.

2.3. Modeling Sound Mixtures

In order to model sound mixtures, we extended the N-HMM to the non-negative factorial hidden Markov model (N-FHMM) (Mysore et al., 2010). The N-FHMM models each time frame of the sound mixture as a linear combination of primarily the dictionary elements of one dictionary of each sound source and therefore also has a notion of block sparsity.

3. Applications

Due to the inherent block sparsity in the N-HMM and N-FHMM, each time frame of each sound source is explained primarily by one of its many dictionaries. In this section we explain why this leads to superior performance to using a single dictionary per sound source in various audio applications.

We used these models for supervised source separation (Mysore et al., 2010; Mysore & Smaragdis, 2012) as follows. We first learned an N-HMM for each sound source from isolated training data of the individual sound sources. We then combined these N-HMMs into an N-FHMM by using the learned dictionaries and transition matrices of the individual N-HMMs. The goal was then to estimate the mixture weights of the N-FHMM on mixture test data. Using these mixture weights, we were able to separate the sources [2].

Multiple sound sources often have dictionaries that are quite similar, as is the case when all sources are speech. As shown in Fig. 4, each dictionary corresponds to a fairly specific subunit of speech. In a given time frame of the mixture, each individual source will therefore be explained primarily by one of its many dictionaries. Unless multiple sources happen to correspond to a similar subunit of speech in the same time frame, each source will not be able to explain the other sources well. The estimates of the separated sources therefore tend to not contain much bleed from the other sources. On the other hand, when a single dictionary is used to explain each source (Smaragdis et al., 2007) and the sources are spectrally quite similar, each source will be able to explain the other sources fairly well since all dictionary elements of each source can be used to explain each time frame.

Posing denoising as a source separation problem (Mysore & Smaragdis, 2011) in which the two sources are speech and noise, we obtained superior results to using a single dictionary for speech. This is because because certain parts of speech (e.g. fricatives) can explain noise quite well and the corresponding dictionary elements can be used in every time frame when using a single dictionary. However, when using the N-FHMM, a speech dictionary will only be able to explain noise well in time frames in which speech is noise like.

We also achieved superior results to using a single dictionary in the missing audio data imputation problem (Han et al., 2012)[3] as the structure of the N-HMM helped constrain the problem.

4. Conclusions

In this paper, we showed why the N-HMM has inherent block sparsity and how it is used to jointly learn multiple dictionaries to model a sound source. We also explained why this is benificial for various audio applications.

[2]Sound examples are available at https://ccrma.stanford.edu/~gautham/Site/lva_ica_2010.html and https://ccrma.stanford.edu/~gautham/Site/lva_ica_2012.html

[3]Sound examples are available at http://www.cs.northwestern.edu/~jha222/imputation

References

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *JMLR*, 3:993–1022, January 2003.

Eldar, Yonina C. and Mishali, Moshe. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11): 5302 – 5316, November 2009.

Han, Jinyu, Mysore, Gautham J., and Pardo, Bryan. Audio imputation using the non-negative hidden markov model. In *Proceedings of LVA/ICA*, March 2012.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*, volume 13, pp. 556–562, 2001.

Mysore, G. J., Smaragdis, P., and Raj, B. Non-negative hidden markov modeling of audio with application to source separation. In *Proceedings of LVA/ICA*, Sept. 2010.

Mysore, Gautham J. and Smaragdis, Paris. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In *Proceeding of IEEE ICASSP*, May 2011.

Mysore, Gautham J. and Smaragdis, Paris. A non-negative approach to language informed speech separation. In *Proceedings of LVA/ICA*, March 2012.

Plumbley, Mark D., Abdallah, Samer A., Blumensath, Thomas, Jafari, Maria G., Nesbit, Andrew, Vincent, Emmanuel, and Wang, Beiming. Musical audio analysis using sparse representations. In *Compstat 2006 - Proceedings in Computational Statistics*, pp. 105–117. Physica-Verlag HD, 2006.

Smaragdis, P. and Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of IEEE WASPAA*, October 2003.

Smaragdis, P., Raj, B., and Shashanka, M. Probabilistic latent variable model for acoustic modeling. In *NIPS Workshop on Advances in models for acoustic processing*, December 2006.

Smaragdis, P., Raj, B., and Shashanka, M. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proceedings of ICA*, September 2007.