Mixtures of Gamma Priors for Non-negative Matrix Factorization Based Speech Separation

Tuomas Virtanen[1] and Ali Taylan Cemgil[2]

[1] Tampere Univ. of Technology, Korkeakoulunkatu 1, FI-33720 Tampere, Finland
[2] Boğaziçi University, Dept. of Computer Eng., TR-34342 Istanbul, Turkey

Abstract. This paper deals with audio source separation using supervised non-negative matrix factorization (NMF). We propose a prior model based on mixtures of Gamma distributions for each sound class, which hyperparameters are trained given a training corpus. This formulation allows adapting the spectral basis vectors of the sound sources during actual operation, when the exact characteristics of the sources are not known in advance. Simulations were conducted using a random mixture of two speakers. Even without adaptation the mixture model outperformed the basic NMF, and adaptation furher improved slightly the separation quality. Audio demonstrations are available at `www.cs.tut.fi/~tuomasv`.

1 Introduction

Separation of mixtures of sound sources has many applications in the computational analysis of audio, speech enhancement, and noise-robust speech recognition. Particularly, non-negative matrix factorization (NMF) and its extensions have produced good results [1,2,3].

The signal model in non-negative spectrogram factorization approximates the spectrum vector $\mathbf{x}_t$ in frame t as a weighted sum of N basis vectors $\mathbf{b}_n$:

$$\mathbf{x}_t \approx \sum_{n=1}^{N} \mathbf{b}_n g_{n,t}, \tag{1}$$

where $g_{n,t}$ is the gain of the nth component in frame $t = 1, \ldots, T$.

The basis vectors and gains can be estimated by minimizing the error of the approximation (1). In audio signal processing, the divergence

$$\sum_{t=1}^{T} \sum_{f=1}^{F} d(x_{t,f}, \sum_{n} b_{n,f} g_{n,t}) \tag{2}$$

where $d(p,q) = p \log(p/q) - p + q$, has turned out to produce good results [1]. Here $b_{n,f}$ denotes the fth entry of $\mathbf{b}_n$, and f is the frequency index. The same procedure can be derived from a maximum likelihood perspective

$$p(\mathbf{x}_{1:T}|\mathbf{b}_{1:N}, g_{1:T,1:N}) = \sum_{t=1}^{T} \sum_{f=1}^{F} \delta(x_{t,f} - \sum_{n=1}^{N} s_{t,f}^{n}) \prod_{n=1}^{N} p(s_{t,f}^{n}|b_{n,f} g_{n,t}) \tag{3}$$

where each spectrogram entry $x_{t,f}$ equals the sum $x_{t,f} = \sum_{n=1}^{N} s_{t,f}^n$ of component spectrograms $s_{t,f}^n$ having Poisson distribution $p(s_{t,f}^n|b_{n,f}g_{n,t}) = Po(s_{t,f}^n; b_{n,f}g_{n,t})$ [5,6]. The divergence (2) can be efficiently minimized using the multiplicative update rules proposed in [7].

When training material of each source in isolation is available, NMF can be used in "supervised mode", i.e., to train class conditional basis spectra of each source in advance [2,3,4]. In the training phase, all the training material of a specific sound class is first concatenated into a single signal, and the spectrogram of the resulting signal is then decomposed into a sum of components using NMF. This results to a class specific set of basis vectors for each source. For the actual separation, the trained basis vector sets of all the source classes are combined and a mixture signal can then be processed using the learned spectra. The previous studies have kept the basis vectors fixed and re-estimated the gains only.

In the real world scenarios, it is either not possible to have training material of a particular target source, or the acoustic conditions in the training and actual operation stages vary. In these situations, adaptive models may be advantageous. One obvious possibility is to train prior distributions $p(\mathbf{b}_n|\Theta)$ instead of fixed parameters $\mathbf{b}_n^*$. Rennie et al. [4] obtained better results in the separation of two speakers by using prior distributions instead of fixed spectra. The computational burden caused by prior distributions can be alleviated if appropriate conjugate priors are chosen, so that one can retain the efficiency of the original NMF algorithm in maximum a posterior (MAP) estimation [5] as explained in Section 2, or in a full Bayesian treatment [6].

This paper discusses the supervised use of NMF where the basis vectors are trained in advance using material where each sound class is present in isolation. We propose here a practical procedure to estimate a Gamma mixture prior model for basis vectors. Section 4 shows simulations using mixtures of two speakers, where the proposed method is shown to outperform the existing ones.

2   Supervised Non-negative Spectrogram Factorization

The characteristics of acoustic sources in real environments are highly variable, hence it is advantageous to have adaptive models that can capture these characteristics. In a probabilistic framework, this can be accomplished by using prior distributions for the basis vectors $\mathbf{b}_n$ instead of fixing them. Formally, in the training phase of supervised non-negative spectrogram factorisation, we ideally wish to estimate class-conditional hyperparameters Θ_c for each source class $c = 1 \ldots C$ by maximising the marginal log-likelihood:

$$\mathcal{L}(\Theta_c) = \int p(s_c|\mathbf{b}_c, g_c)p(\mathbf{b}_c|\Theta_c)p(g_c|\Theta_c)d\mathbf{b}_c dg_c \qquad (4)$$

where s_c is a known spectrogram from a source class c, and $\mathbf{b}_c$ and g_c denote all the basis vector and gains of class c, respectively. Then, the actual separation of mixture spectrogram $\mathbf{x}$ is achieved via computation of

$$p(s_{1:C}|\mathbf{x}, \Theta_{1:C}) = \int \delta(\mathbf{x}_{1:T} - \sum_{c=1}^{C} s_c) \left[\prod_{c=1}^{C} p(s_c|\mathbf{b}_c, g_c)p(\mathbf{b}_c|\Theta_c)p(g_c|\Theta_c) \right] d\mathbf{b}_{1:C} dg_{1:C}$$

However, these integrals can be hard to evaluate and more practical approaches are taken in practice, such as computing MAP estimates for $\mathbf{b}_c$ and g_c.

As a prior for basis vectors $p(\mathbf{b}_c|\Theta_c)$, one can use a Gamma distribution $\mathcal{G}(b_{n,f}; k_{n,f}, \theta_{n,f})$ for each element $b_{n,f}$ of each basis vector $\mathbf{b}_n$. In the MAP framework with the Poisson observation model it results to minimizing the sum of the divergence (2) and the penalty term

$$\sum_{n=1}^{N} \sum_{f=1}^{F} (k_{n,f} - 1) \log(b_{n,f}) - b_{n,f}\theta_{n,f} \qquad (5)$$

which is the logarithm of the Gamma distribution [5], up to additive terms which are independent of the basis vector entries.

A typical gain prior $p(g|\Theta_c)$ is an exponential distribution with rate parameter λ which translates to the penalty term $\lambda \sum_n \sum_t g_{n,t}$. Sparse prior for the gains has been found to improve the separation quality [2].

During separation, when the basis vectors are fixed, the MAP estimation of gains can be obtained by applying iteratively updates

$$\mathbf{r}_t = \mathbf{x}_t./\sum_n \mathbf{b}_n g_{n,t} \qquad\qquad g_{n,t} \leftarrow g_{n,t}\frac{\mathbf{r}^\mathsf{T}\mathbf{b}_n}{\mathbf{1}^\mathsf{T}\mathbf{b}_n + \lambda}, \quad n = 1, \ldots, N, \qquad (6)$$

where ./ denotes element-wise division and $\mathbf{1}$ is a all-one column vector. Similarly, when the gains are fixed, the basis vectors can be updated via

$$b_{n,f} \leftarrow \frac{k_{n,f} - 1 + b_{n,f}\sum_t (g_{n,t}x_{t,f}/\sum_{n'} g_{n',t}b_{n',f})}{1/\theta_{n,f} + \sum_t g_{n,t}} \qquad (7)$$

which is guaranteed to increase the posterior probability of the basis vectors when $k_{n,f} \geq 1$ [5]. It is important to note that under this formalism, the basis is adapted during the actual separation.

3   Training Mixture of Gamma Priors

In single channel source separation, when two or more sources overlap in time and frequency, we have to use redundancy of the sources to achieve good sound source separation. Redundancy in frequency can be used efficiently when basis vectors correspond to entire spectra of sound events instead of just parts of their spectra. Basis vectors corresponding to entire spectra can be trained by restricting only one basis vector to be active in each frame.

We make two assumptions which allow us to train efficiently distributions of basis vectors corresponding to entire spectra: 1) only one basis vector is active in each frame, and 2) the training data can be normalized so that the normalized observations correspond to observed basis vectors. The first assumption can be viewed as an extreme case of sparseness whereas the second cancels out the effect of gains in training basis vector priors. While this is omitting the variation in the Poisson model, we find this procedure virtually identical to the more principled approach where $\mathbf{b}_c$ are considered as latent.

A prior for the basis vectors based on the above assumptions can be trained using a mixture model. In the sequel, we omit the class label c as each class is learned separately. The model for a source class is

$$p(\mathbf{b}|\Theta) = \sum_{n=1}^{N} w_n \prod_{f=1}^{F} \mathcal{G}(b_{n,f}; k_{n,f}, \theta_{n,f}),\qquad(8)$$

where k, θ are the shape and scale parameters of individual Gamma distributions and w_n are the prior weights. All the hyperparameters are denoted as $\Theta = (k, \theta, w)$. We do not model the dependencies between frequency bins so that the distribution of a mixture component is the product of its frequency marginals.

3.1   Training Algorithm

The observations are first preprocessed by normalizing each observation vector $\mathbf{b}_t$ so that the norm of $\log(\mathbf{b}_t + \epsilon)$, where ϵ is a small fixed scalar, is unity. The EM algorithm is initialized by running the k-means algorithm with random initial clusters for 10 iterations using the normalized log-spectrum observations to get cluster centroid vectors $\boldsymbol{\mu}_n$. Centroids of linear observations are then calculated as $\mu_{n,f} = e^{\mu_{n,f}} - \epsilon$. From the linear cluster centroids we estimate the initial Gamma distribution parameters as $k_{n,f} = \mu_{n,f}^2$ and $\theta_{n,f} = 1/\mu_{n,f}$. This generates a Gamma distribution having mean $\mu_{n,f}$ and variance 1. Cluster weights are set to $w_n = 1/N$. The iterative estimation procedure is as follows:

1. Evaluate the posterior distribution $z_{n,t}$ that the nth cluster has generated the tth observation as

$$z_{n,t} = \frac{w_n \prod_f \mathcal{G}(b_{t,f}; k_{n,f}, \theta_{n,f})}{\sum_{n'=1}^{N} w_{n'} \prod_f \mathcal{G}(b_{t,f}; k_{n',f}, \theta_{n',f})}\qquad(9)$$

2. Re-estimate the mixture weights as

$$w_n = \frac{\sum_{t=1}^{T} z_{n,t}}{\sum_{n'=1}^{N} \sum_{t=1}^{T} z_{n',t}}.\qquad(10)$$

3. Re-estimate the shape parameters by solving

$$\log(k_{n,f}) - \psi(k_{n,f}) = \log\left(\frac{\sum_t z_{n,t} b_{t,f}}{\sum_t z_{n,t}}\right) - \sum_t \log(b_{t,f}) z_{n,t}\qquad(11)$$

using the Newton-Raphson method, where $\psi(k_{n,f}) = \Gamma'(k_{n,f})/\Gamma(k_{n,f})$ is the digamma function. We used 10 iterations, and the previous estimates of $k_{n,f}$ as initial values.

4. Re-estimate the scale parameters as

$$\theta_{n,f} = \frac{\sum_{t=1}^{T} z_{n,t} b_{t,f}}{k_{n,f} \sum_t z_{n,t}}.\qquad(12)$$

The steps 1-4 are repeated for 100 iterations, or until the algorithm converges. In order to prevent too narrow clusters, we found it advantageous to restrict the variance of each cluster above a fixed minimum m after each iteration as follows. The variance of each Gamma distribution is $\mu_{n,f}\theta_{n,f}^2$. For clusters for which the variance is smaller than the minimum limit m, we calculate the ratio $y_{n,f} = m/(\mu_{n,f}\theta_{n,f}^2)$, and then modify the distribution parameters as $\theta_{n,f} \leftarrow \theta_{n,f}y_{n,f}$ and $k_{n,f} \leftarrow k_{n,f}/y_{n,f}$. The above procedure sets the variance of the cluster to m without changing its mean. We also found it advantageous to keep the mixture weights fixed for the first 90 iterations.

3.2 Alternative Gamma Prior Estimation Methods

In addition to the Gamma mixture model, we tried out alternative methods for generating the priors. In general, one can generate a Gamma distribution from fixed basis vectors $b_{n,f}$ obtained with NMF (or by any other algorithm) by selecting arbitrary shape k, and then calculating the scale as $\theta_{n,f} = b_{n,f}/k$. The mean of the resulting distribution equals $b_{n,f}$ and its variance $b_{n,f}^2/k$ scales quadratically with the mean.

In addition to direct training of the Gamma mixture model parameters, we obtained good results by applying a Gaussian mixture model for the log-spectrum observations and then deriving the corresponding Gamma mixture model by matching the moments of each cluster. We calculated the log-spectrum as $\log(\mathbf{b}_t + \epsilon)$ and then trained a Gaussian mixture model, which mean and variance are denoted as $\mu_{n,f}$ and $\sigma_{n,f}^2$, respectively. The mean and variance of the linear observations are $\tilde{\mu}_{n,f} = e^{\mu_{n,f}+\sigma_{n,f}^2/2}$ and $\tilde{\sigma}_{n,f}^2 = (e^{\sigma_{n,f}^2} - 1)e^{2\mu_{n,f}+\sigma_{n,f}}$, respectively. Gamma distributions of linear observations can be generated by matching the mean and variance as $k_{n,f} = \tilde{\mu}_{n,f}^2/\tilde{\sigma}_{n,f}$ and $\theta_{n,f} = \tilde{\sigma}_{n,f}/\tilde{\mu}_{n,f}$.

4   Simulations

We evaluated the performance of the proposed methods in separating signals consisting of two speakers of different genders. We used the Grid corpus [8], which consists of short sentences spoken by 34 speakers. We generated 300 random test signals where three sentences spoken by a male speaker and a female speaker were mixed. Each test signal was generated by concatenating random three sentences of a random male speaker, concatenating random three sentences of a random female speaker, and mixing the signals at equal power level.

The data representation is similar to the one used in [2]: the signals were first filtered with a high-frequency emphasis filter, then windowed into 32 ms frames using a Hamming window with 50 % overlap between adjacent frames. DFT was used to calculate the spectrum of each frame, and the spectra were decimated to 80-band Mel frequency scale by weighting and summing the DFT bins.

In the training phase we learned a model for both genders. We used leave-one-out training where the model of a gender was trained by excluding each test speaker at time from the training data, resulting in 18 male and 16 female models in total. We used only every 10th sentence (in the alphabetical order) of

the training data to keep the computation time reasonable. The purpose of the leave-one-out training was to simulate a situation where the exact characteristics of the target and interfering sources were not known in advance.

4.1   Training Algorithms

Four different algorithms were tested in training the priors:

- NMF estimates fixed priors using the sparse NMF algorithm [1] which uses the divergence criterion (2). Sparseness factor which produced approximately the best results was used (the optimal value was different in testing). In order to test basis vector adaptation, Gamma distributions were generated using the procedure explained in Section 3.2 with parameter $k = 0.01$.
- Gamma mixture model was trained using the algorithm in Section 3.1.
- Gaussian mixture model was trained using log-spectrum observations, and the Gamma mixture model was generated as explained in Section 3.2.
- Gaussian mixture model was trained using linear-spectrum observations and the Gamma mixture model was generated by matching the moments.

The above algorithms are denoted as NMF, Gamma, Gaussian-log, and Gaussian-lin, respectively. All the algorithms were tested with 30 and 70 components per speaker. We found that it is advantageous to control the variance of the trained distributions by scaling their parameters as $k_{n,f} = k_{n,f}/q$ and $\theta_{n,f} = \theta_{n,f}q$, which retains the mean of the distribution but scales its variance by q. Value $q = 0.1$ produced approximately the best results. Normalizing each basis vector to unity norm and scaling the distributions accordingly by multiplying the scale parameter was also found to improve the results slightly.

4.2   Testing

In the test phase, the bases of male and female speakers obtained by a particular training algorithm were concatenated. Each of the 300 test signals was processed using sparse NMF by applying the update rules (6) and (7). Sparseness factor λ which produced approximately best result was used.

All the algorithms were tested with fixed and adaptive bases: adaptive bases used the distributions obtained from the training, whereas fixed based were set equal to the mean of each prior distribution. The basis vectors in all the algorithms were initialized with the mean of each prior distribution, and random positive values were used to initialize the gains.

The basis vectors and gains were estimated using each test signal at time. The weighted sum of male basis vectors in frame t is calculated as $\mathbf{m}_t = \sum_{n \in \mathcal{M}} \mathbf{b}_n g_{n,t}$, where $\mathcal{M}$ is the set of male basis vectors. Similarly, the weighted sum $\mathbf{f}_t$ of female basis vectors is calculated using the set of female speaker basis vectors. The male speaker spectrum $\hat{\mathbf{m}}$ in each frame is then reconstructed as

$$\hat{\mathbf{m}}_t = \mathbf{x}_t . * \mathbf{m}_t ./ (\mathbf{m}_t + \mathbf{f}_t), \tag{13}$$

where $.*$ and denotes element-wise multiplication. Female spectra are obtained as $\mathbf{x}_t - \hat{\mathbf{m}}_t$.

The quality of separation was measured by the signal-to-noise ratio of the separated spectrograms. The SNRs were averaged over both the speakers in all the test signals.

4.3   Results

The average signal-to-noise ratios of each of the tested algorithm are illustrated in Table 1. The Gamma and Gaussian-log methods produce clearly better results than NMF and Gaussian-lin. Adaptive bases increase the performance of Gamma method, but for other methods the effect is small. A larger number of components improves significantly the performance of all the methods except NMF.

Sparseness in testing was found to improve the quality of the separation slightly. Figure 1 illustrates the performance of the Gamma method with different sparseness factors λ. Sparseness improved more clearly the performance of the NMF training, but the results are omitted because of space limitation restrictions. Figure 1 also illustrates the effect of scaling the variances of the distributions. Value $q = 0$ corresponds to fixed priors, and larger values (adaptation) improve the quality slightly up to certain value of q.

All the parameters of the training algorithms were not completely optimized for this application, so final judgment about the relative performance of Gamma and Gauss-log methods cannot be made. However, the results show that these models perform clearly better than NMF in training the basis vectors.

Table 1. Average signal-to-noise ratios of the tested methods in dB, obtained with fixed and adaptive basis vectors and with either 30 or 70 components per source. The best algorithm in each column is highlighted with bold face font.

method	30 components		70 components	
	fixed	adaptive	fixed	adaptive
NMF	5.68	5.71	5.57	5.55
Gamma	**6.55**	**6.73**	6.95	**7.04**
Gaussian-log	6.54	6.56	**7.02**	7.03
Gaussian-lin	3.34	3.27	3.75	3.71
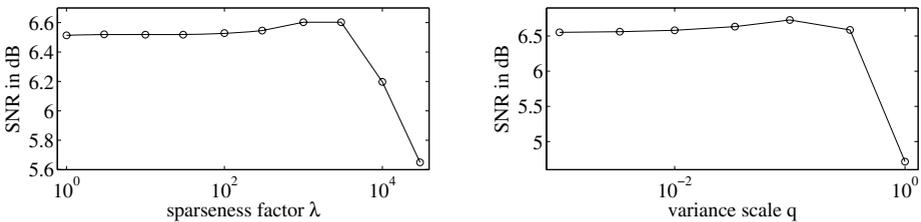
Fig. 1. The effect of the sparseness factor λ (left panel) and the distribution variance scale q (right panel) on the average signal-to-noise ratio

5   Conclusions

We have proposed the use of a Gamma mixture model in representing the basis vector distributions in supervised non-negative matrix factorization based sound source separation. The proposed method is shown to produce better results than previous sparse NMF training. In addition to better separation quality, the method also simplifies the training since there is no need to tune the sparseness factor in NMF. Mixture model training also opens up new possibilities of incorporating hidden state variables in the model, which allow modeling the temporal dependency in the signals.

References

1. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing 15(3) (2007)
2. Schmidt, M.N., Olsson, R.K.: Single-channel speech separation using sparse non-negative matrix factorization. In: Proceedings of the International Conference on Spoken Language Processing, Pittsburgh, USA (2006)
3. Wilson, K.W., Raj, B., Smaragdis, P.: Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In: 9th Annual Conf. of the Int. Speech Communication Association (Interspeech 2008), Brisbane, Australia (2008)
4. Rennie, S.J., Hershey, J.R., Olsen, P.A.: Efficient model-based speech separation and denoising using non-negative subspace analysis. In: Proceedings of IEEE Int. Conf. on Audio, Speech and Signal Processing, Las Vegas, USA (2008)
5. Virtanen, T., Cemgil, A.T., Godsill, S.: Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In: Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Las Vegas, USA (2008)
6. Cemgil, A.T.: Bayesian inference in non-negative matrix factorisation models. Technical Report CUED/F-INFENG/TR.609, University of Cambridge (July 2008)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of Neural Information Processing Systems, Denver, USA, pp. 556–562 (2000)
8. Cooke, M.P., Barker, J., Cunningham, S.P., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Soc. of America 120(5) (2006)