# BAYESIAN EXTENSIONS TO NON-NEGATIVE MATRIX FACTORISATION FOR AUDIO SIGNAL MODELLING

*Tuomas Virtanen, A. Taylan Cemgil, Simon Godsill*

Signal Processing and Communications Laboratory, University of Cambridge,
Department of Engineering, Trumpington Street, Cambridge, CB2 1PZ, UK

## ABSTRACT

We describe the underlying probabilistic generative signal model of non-negative matrix factorisation (NMF) and propose a realistic conjugate priors on the matrices to be estimated. A conjugate Gamma chain prior enables modelling the spectral smoothness of natural sounds in general, and other prior knowledge about the spectra of the sounds can be used without resorting to too restrictive techniques where some of the parameters are fixed. The resulting algorithm, while retaining the attractive features of standard NMF such as fast convergence and easy implementation, outperforms existing NMF strategies in a single channel audio source separation and detection task.

*Index Terms*— acoustic signal processing, matrix decomposition, MAP estimation, source separation

## 1. INTRODUCTION

Time-frequency energy distributions are of central importance in audio signals analysis; particularly, the magnitude spectrogram representation displays the magnitude of the time-frequency coefficient $x_{\nu,\tau}$ as a function of frequencies $\nu$ and time indices $\tau$. In recent years, one audio modelling approach has focused on non-negativity of the spectrogram matrix $\mathbf{X} = \{x_{\nu,\tau}\}$ and enforcing a factorisation as $\mathbf{X} = \mathbf{T}\mathbf{V}$ where both $\mathbf{T}$ and $\mathbf{V}$ are matrices with positive entries (see [2, 3, 4], and references therein). Here, $\mathbf{T}$ can be interpreted as a codebook of spectra, called basis vectors, and $\mathbf{V}$ is the matrix of their gains in each frame. The success of the model stems from the fact that entities of natural sounds can rather well be approximated as a product of stationary spectrum and time-varying gain. These entities include, for example, individual tones of musical instruments. A basis vector and its gains can represent, for example, the contribution of all the tones of a certain musical instrument having the same pitch, or all the tones of a percussive musical instrument. An advantage of these methods is computational attractiveness due to fast converging iterative matrix factorisation techniques [5].

A problem with the standard NMF objective is that the probabilistic interpretation is not explicit and consequently basis vectors and gains are not well modelled, and as we will show later, are assumed to be independent a-priori for all entries of $\mathbf{T}$ and $\mathbf{V}$. Especially for music signals, due to the physical properties of musical instruments and quasi-periodic structure of music, one could clearly design more informative priors. For example, due to presence of note events that have a constant pitch, gains in adjacent time-frequency atoms tends to be correlated. Similarly, due to harmonicity and constant timbre, basis vectors tend to have typically peaks at harmonically related frequency indicies.

Existing approaches have tried to model prior knowledge about basis vectors by initializing an inference algorithm with a set of basis vectors corresponding to harmonic spectra [3], assuming that this would enable more robust inference. Alternatively, one can learn a set of basis vectors from a training corpus where each source is present in isolation, and then keep the basis vectors fixed and estimate their gains [6]. This latter approach produces good results when the spectral characteristics of the training data are equal to those of the target data. In practice, however, the exact characteristics of the target signal are often not known, and any mismatch between training and target data decreases the quality of the obtained solution. One remedy is adapting all basis vectors but introducing regularisation terms that encode some prior knowledge, such as enforcing temporal continuity. This strategy has been shown to be effective using a heuristic approach where a cost function which penalizes large differences between the gains of adjacent frames [4].

Our goal in this paper is twofold. First we describe in detail the underlying probabilistic generative signal model of the NMF and the nonnegative update equations as a quasi gradient optimisation strategy. Consequently, given the probabilistic model, we can impose various prior structures. Here, we use a Gamma chain prior [1] on the basis vectors $\mathbf{T}$ and gains $\mathbf{V}$. The resulting algorithm outperforms existing NMF strategies and opens up the way for a full Bayesian treatment for model selection.

The paper is organized as follows: Section 2 reviews shortly the objective of non-negative matrix factorisation and the related optimization algorithm. Section 3 presents the probabilistic generative model behind the NMF, and extends it to allow priors for the parameters. Section 4 presents simulation results and Section 5 conclusions.

## 2. NON-NEGATIVE MATRIX FACTORISATION

In NMF, the goal is to find entrywise non-negative matrices $\mathbf{T}$ and $\mathbf{V}$ such that

$$(\mathbf{T}^*, \mathbf{V}^*) = \arg\min_{\mathbf{T}\mathbf{V}} D(\mathbf{X}||\mathbf{T}\mathbf{V}) \tag{1}$$

where $\mathbf{X}$ is an entrywise non-negative matrix and $D$ could be Euclidian distance, or divergence defined as

$$D(\mathbf{X}||\mathbf{Y}) = \sum_{\nu,\tau}[\mathbf{X}]_{\nu,\tau}\log[\mathbf{X}]_{\nu,\tau}/[\mathbf{Y}]_{\nu,\tau} - [\mathbf{X}]_{\nu,\tau} + [\mathbf{Y}]_{\nu,\tau}$$

Here we use the divergence since it has been found to produce better results in audio signal analysis [4]. Since $\mathbf{X}$ is fixed, we can use the divergence function

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

and write the equivalent optimisation problem

$$(\mathbf{T}^*, \mathbf{V}^*) = \arg\min_{\mathbf{TV}} \sum_{\tau,\nu} d([\mathbf{X}]_{\tau,\nu}, [\mathbf{TV}]_{\tau,\nu}) \qquad (3)$$

In general, this optimisation problem is not convex with respect to both $\mathbf{T}$ and $\mathbf{V}$. Therefore finding the global optimum cannot be guaranteed by any optimisation method. However, the problem is convex with respect to $\mathbf{T}$ and $\mathbf{V}$ separately, which allows for finding of locally optimal solutions.

Because of their computational effectiveness and simplicity, the multiplicative updates proposed in [5] have been extensively used to solve the problem (3). The convergence proof in [5], which essentially hinges upon bounding terms $\log \sum_{i=1}^{I} t_{\nu,i} v_{i,\tau}$ of the objective by a variational bound, can be interpreted as follows: for fixed non-negative parameters $x_{\nu,\tau}$ and $t_{\nu,i}$, $\nu = 1, \ldots, F$, $i = 1, \ldots, I$, and variables $v_{i,\tau}$ which are restricted to non-negative values, the function

$$c = \sum_{\nu=1}^{F} d(x_{\nu,\tau}, \sum_{i=1}^{I} t_{\nu,i} v_{i,\tau}) \qquad (4)$$

is non-increasing under simultaneous update of all $v_{i,\tau}$, $i = 1, \ldots, I$ using the rule

$$v_{i,\tau} \leftarrow v_{i,\tau} \frac{\sum_{\nu=1}^{F} t_{\nu,i} x_{\nu,\tau} / (\sum_{i'=1}^{I} t_{\nu,i'} v_{i',\tau})}{\sum_{\nu'=1}^{F} t_{\nu',i}} \qquad (5)$$

The rule has been applied to solve (3) by keeping first $\mathbf{T}$ fixed and applying (5) to update $\mathbf{V}$, then keeping $\mathbf{V}$ fixed and updating $\mathbf{T}$, and repeating the updates until the values converge. In practice, this variational approach has been found to be efficient in estimating $\mathbf{V}$ and $\mathbf{T}$, since it automatically obeys the non-negativity restrictions.

## 3. POISSON OBSERVATION MODEL

In the sequel, we illustrate that the objective (3) can be derived starting from a probabilistic model[1]. Assume that the magnitude at each time-frequency atom $s_{\nu,\tau}^i$ produced by the $i^{\text{th}}$ source is Poisson distributed:

$$s_{\nu,\tau}^i \sim \mathcal{PO}(s_{\nu,\tau}^i; t_{\nu,i} v_{i,\tau}), \qquad (6)$$

where $v_{i,\tau}$ is the gain of the $i^{\text{th}}$ basis vector in frame $\tau$ and $\mathcal{PO}$ is the Poisson distribution defined as

$$\mathcal{PO}(x; \lambda) = e^{-\lambda} \lambda^x / \Gamma(x+1). \qquad (7)$$

Here, $\Gamma(x)$ denotes the gamma (generalised factorial) function. The Poisson distribution is defined only for discrete $x$, but in practise the accuracy of $x$ does need to be limited by having a large integer scale. We assume that the total magnitude of the observed signal $x_{\nu,\tau}$ in each time-frequency point is the sum of the magnitudes of individual sources[2], i.e., $x_{\nu,\tau} = \sum_{i=1}^{I} s_{\nu,\tau}^i$. The sum of independent Poisson-distributed random variables is also a Poisson random variable with intensity parameter equal to the sum of individual intensity parameters. Therefore,

$$p(x_{\nu,\tau}|t_{\nu,1:I}, v_{1:I,\tau}) = \mathcal{PO}(x_{\nu,\tau}; \sum_{i=1}^{I} t_{\nu,i} v_{i,\tau}), \qquad (8)$$

---

[1]Such as an observation has been made before in many studies, but here it is crucial to formulate the mathematical details.

[2]We note that this assumption is physically unrealistic, since in general for two superinposed sources $\xi_1$ and $\xi_2$, the magnitude of the superposition $x = |\xi_1 + \xi_2|$ can not be written as the superposition of the magnitudes of individual sources, i.e. $x \neq |\xi_1| + |\xi_2|$.

where $t_{\nu,1:I}$ denotes the $\nu^{\text{th}}$ column of $\mathbf{T}$ and $v_{1:I,\tau}$ the $\tau^{\text{th}}$ row of $\mathbf{V}$, respectively. Assuming that each time-frequency point is statistically independent conditional on $\mathbf{T}$ and $\mathbf{V}$, the entire model can be denoted using matrix notation by

$$p(\mathbf{X}|\mathbf{T}, \mathbf{V}) = \prod_{\tau,\nu} \frac{e^{-[\mathbf{TV}]_{\nu,\tau}} [\mathbf{TV}]_{\nu,\tau}^{[\mathbf{X}]_{\nu,\tau}}}{\Gamma([\mathbf{X}]_{\nu,\tau} + 1)} \qquad (9)$$

The maximum likelihood solution is given by

$$(\mathbf{T}^*, \mathbf{V}^*) \quad = \quad \arg\max_{\mathbf{T}, \mathbf{V}} \log p(\mathbf{X}|\mathbf{T}, \mathbf{V})$$

where

$$\log p(\mathbf{X}|\mathbf{T}, \mathbf{V})$$
$$= \sum_{\nu,\tau} -[\mathbf{TV}]_{\nu,\tau} + [\mathbf{X}]_{\nu,\tau} \log([\mathbf{TV}]_{\nu,\tau}) - \log(\Gamma([\mathbf{X}]_{\nu,\tau} + 1))$$
$$=^+ - \sum_{\nu,\tau} d([\mathbf{X}]_{\nu,\tau}, [\mathbf{TV}]_{\nu,\tau}) \qquad (10)$$

Here $=^+$ denotes equal up to irrelevant constant terms (i.e. $f \propto g \Leftrightarrow \log f =^+ \log g$). We see that this objective is identical to the objective (3) optimised by NMF.

### 3.1. Gain prior $p(\mathbf{V})$

In the following, we propose that both the basis vectors and gains are unobserved random variables which are to be estimated from the data. We assume that the prior factorises as $p(\mathbf{T}, \mathbf{V}) = p(\mathbf{T})p(\mathbf{V})$. To model continuation across time, we will use a Markov chain on gains. A suitable prior, that guarantees that the gains are strictly non-negative, and positively correlated (i.e., slowly varying in time) can be constructed by a so called Gamma-chain [1]. Here $\mathcal{G}(y; a, b)$ is the gamma distribution defined for $y > 0$ as

$$\mathcal{G}(y; a, b) = y^{a-1} b^{-a} e^{-y/b} / \Gamma(a), \qquad (11)$$

A Gamma chain is constructed by using auxiliary variables $z_{i,\tau}$ as follows.

$$z_{i,1} \quad \sim \quad \mathcal{G}(z_{i,1}; a+1, (ab)^{-1})$$
$$v_{i,\tau}|z_{i,\tau} \quad \sim \quad \mathcal{G}(v_{i,\tau}; a, (z_{i,\tau} a)^{-1})$$
$$z_{\tau+1,i}|v_{i,\tau} \quad \sim \quad \mathcal{G}(z_{\tau+1,i}; a+1, (v_{i,\tau} a)^{-1})$$

Here, $a$ is a coupling parameter that affects the affinity between the gains of adjacent frames. When $a$ is large, adjacent frames are coupled more strongly. The auxiliary variables are needed to ensure positive correlation *and* conjugacy, a technical condition that leads to closed form fixed point equations as in standard NMF. The above Gamma chain is a single parameter version of the model presented in [1] where the value $a_z = a + 1$ is used in the distribution of the auxiliary variables. The resulting model is a collection of independent Gamma chains for the gains of each source $i$.

The relevant terms in the log-prior function are given as

$$\log p(\mathbf{V}, \mathbf{Z}) =^+ \sum_{i=1}^{I} a \log(z_{i,K+1}) - z_{i,1} ab$$
$$+ \sum_{i=1}^{I} \sum_{\tau=1}^{K} 2a[\log(v_{i,\tau}) + \log(z_{i,\tau})] - v_{i,\tau} z_{i,\tau} a - z_{i,\tau+1} v_{i,\tau} a$$
$$=^+ - \sum_{i=1}^{I} \left[ d(a, ab z_{i,1}) + \sum_{\tau=1}^{K} d(a, v_{i,\tau} z_{i,\tau} a) + d(a, v_{i,\tau} z_{i,\tau+1} a) \right]$$

where we define $[\mathbf{Z}]_{i,\tau} \equiv z_{i,\tau}$. $\qquad (12)$

## 3.2. Basis vector prior $p(\mathbf{T})$

In this paper, we assume a prior where each entry of the basis vector matrix is assumed to be independently drawn from a Gamma distribution:

$$p(t_{\nu,i}) = \mathcal{G}(t_{\nu,i}; \alpha_{\nu,i}, \beta_{\nu,i}^{-1}) = t_{\nu,i}^{\alpha_{\nu,i}-1} \beta_{\nu,i}^{\alpha_{\nu,i}} e^{-t_{\nu,i}\beta_{\nu,i}} / \Gamma(\alpha_{\nu,i}) \tag{13}$$

The hyperparameters $\alpha_{\nu,i}$ and $\beta_{\nu,i}$ of the model can be selected individually for each basis vector $t_{1:F,i}$. For example, $\beta_{1:F,i}$ can be selected such that typical basis vectors have peaks at harmonically related frequencies. We assume $p(\mathbf{T}) = \prod_{i=1}^{I} \prod_{\nu=1}^{F} p(t_{\nu,i})$ and consequently, the logarithm of the prior can be written as

$$\log p(\mathbf{T}) =^{+} \sum_{i=1}^{I} \sum_{\nu=1}^{F} (\alpha_{\nu,i} - 1) \log(t_{\nu,i}) - t_{\nu,i}\beta_{\nu,i}$$

$$= -\sum_{i=1}^{I} \sum_{\nu=1}^{F} d(\alpha_{\nu,i} - 1, t_{\nu,i}\beta_{\nu,i}) \tag{14}$$

## 3.3. Inference

Given the model, the joint posterior distribution is given by Bayes' rule $p(\mathbf{Z}, \mathbf{V}, \mathbf{T}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{V}, \mathbf{T}, \mathbf{Z})p(\mathbf{Z}, \mathbf{V}, \mathbf{T})$ which factorises to $p(\mathbf{X}|\mathbf{V}, \mathbf{T})p(\mathbf{Z}, \mathbf{V})p(\mathbf{T})$. The MAP state can be found as

$$\arg \max_{\mathbf{Z}, \mathbf{V}, \mathbf{T}} \{\log p(\mathbf{X}|\mathbf{V}, \mathbf{T}) + \log p(\mathbf{Z}, \mathbf{V}) + \log p(\mathbf{T})\} \tag{15}$$

We substitute the terms in (15) with the expressions in (10), (12), and (14). Since the log-posterior is now written as a sum of the divergence function defined in (2), the MAP estimator can be derived directly by applying the rule (5) on the sum of the terms (10), (12), and (14). To simplify the notation, let us define $m_{\nu,\tau} = x_{\nu,\tau}/(\sum_{i=1}^{I} t_{\nu,i}v_{i,\tau})$ for all $\nu = 1, \ldots, F$ and $\tau = 1, \ldots, K$. The update rule (5) for each of the parameters is given as

$$t_{\nu,i} \leftarrow t_{\nu,i} \frac{(\alpha_{\nu,i} - 1)/t_{\nu,i} + \sum_{\tau=1}^{K} v_{i,\tau}m_{\nu,\tau}}{\beta_{\nu,i} + \sum_{\tau'=1}^{K} v_{i,\tau'}} \tag{16}$$

$$v_{i,\tau} \leftarrow v_{i,\tau} \frac{2a/v_{i,\tau} + \sum_{\nu=1}^{F} t_{\nu,i}m_{\nu,\tau}}{a(z_{i,\tau} + z_{\tau+1,i}) + \sum_{\nu'=1}^{F} t_{\nu',i}} \tag{17}$$

$$z_{i,\tau} \leftarrow \begin{cases} 1/(v_{1,i} + b) & \tau = 1 \\ 2/(v_{i,\tau} + v_{\tau-1,i}) & 1 < \tau < K+1 \\ 1/v_{i,\tau} & \tau = K+1 \end{cases} \tag{18}$$

It can be seen that the update rules differ from the basic NMF updates [5] only by additive terms in the numerator and denominator, which are caused by the priors.

The MAP estimation algorithm works in an iterative fashion, first updating all the basis vectors using (16), then all the gains using (17), and (18), and repeating the updates until the algorithm converges. According to the proof [5], the value of the posterior distribution (15) is guaranteed to be non-decreasing under each of the updates.

By setting $b = 0$, the gain prior becomes independent of the overall level of the gains. Thus, unlike the cost term in [4], the temporal continuity objective implemented by the Gamma chain does not require fixing the scale of the parameters. However, to ensure the numerical stability of the algorithm, in each iteration we scale the variance of the gains of each source to unity, and compensate this by re-scaling the basis vectors and auxiliary variables.

## 4. SIMULATION EXPERIMENTS

The effect of the proposed priors is shown in two studies where the basis vector priors and gain priors are tested separately.
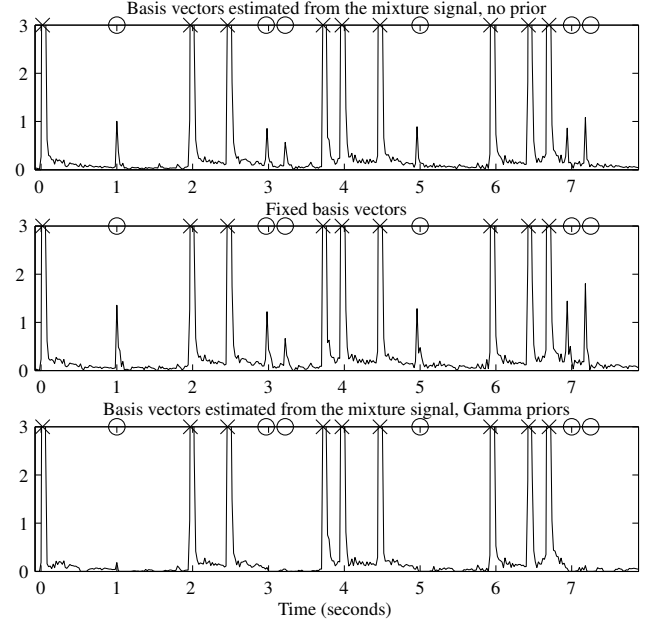


**Fig. 1**. Gains of the bass drum basis vector estimated using three different NMF algorithms. The bass drum onset times are marked with crosses and snare drum onset times with circles, respectively. The proposed method (bottom plot) is able to estimate gains where the interference caused by other sources is smaller than in the other algorithms.

### 4.1. Basis vector priors

In the first study we examine the benefit of the basis vector prior by an example signal of a drum pattern consisting of bass drum, snare drum, and hi-hats. The gain prior is not used here. The magnitude spectrogram of the signal was factorised into three sources ($I = 3$) using three NMF algorithms. The first algorithm, basic NMF, estimates the basis vectors and gains blindly from the mixture signal by minimizing the divergence. The second algorithm uses fixed basis vectors, which were trained for each source using material where the source was present in isolation but the instruments used to produce the sounds were not identical to those in the mixture. The third algorithm is the proposed method which uses Gamma priors for the basis vectors. We set the shape parameter $\alpha_{\nu,i}$ of the prior of all the basis vectors equal to 1, and the scale parameters $\beta_{\nu,i}^{-1}$ equal to the fixed basis vectors trained for the second algorithm.

The gains were estimated using all the three algorithms separately. Figure 1 illustrates the gains corresponding to the bass drum basis vector for three different algorithms. All the algorithms produce large peaks at correct bass drum hits. However, the first and second algorithm also produce smaller erroneous peaks corresponding to the snare drum hits. Because the sounds in the material used to train the basis vectors are not identical to the sounds in the mixture, the second algorithm patches the mismatches by representing parts of snare drum spectra by bass drum basis vectors. On the other hand, because a part of the snare drum spectra can be represented with a bass drum spectra, the blind NMF algorithm is not able to learn the basis vectors accurately enough. The proposed algorithm circumvents these problems and produces gains where snare drum hits do not affect the excitations of the bass drum basis vector.

**Table 1**. Average detection error rates and SDRs of the tested algorithms. The best result in each column is highlighted in bold.

| algorithm | det. error rate (%) | | | SDR (dB) | | |
|---|---|---|---|---|---|---|
| | all | pitched | drums | all | pitched | drums |
| EUC | 28 | 28 | 30 | 6.6 | 7.6 | 4.5 |
| DIV | 26 | 28 | 23 | 7.6 | 9.0 | 4.7 |
| DIV-SQ | **24** | **25** | 22 | 8.5 | 9.8 | **6.0** |
| GAMMA | 25 | 28 | **20** | **10.1** | **12.3** | **6.0** |

## 4.2. Gains

The effect of the Gamma chain prior was evaluated quantitatively in an unsupervised sound source separation task where random acoustic mixtures of tones of musical instruments were separated into sound sources. Basis vector priors were not used in this study. 300 random mixtures were generated by selecting random musical instruments and pitches from an acoustic database described in [4]. Random amount repetitions, timings, durations, etc. were allotted for the tones according to the procedure described in [4].

The baseline algorithms include the basic NMF algorithms based on the minimization of the Euclidean distance between and the divergence between the magnitude spectrograms. These are denoted by EUC and DIV in the following. The NMF algorithm [4] where temporally continuous gains were favored by using a cost term which is the squared difference of the gains of adjacent frames is denoted by DIV-SQ, and the proposed Gamma chain algorithm is denoted by GAMMA. Different values of $a$ were tested and the one which produced approximately the best results was used in the simulations. Basis vector priors were not used in this evaluation.

In the simulations, each mixture was separated into sources by using all the algorithms. At the moment there is no reliable method for the estimation of the number of sources in this framework, and therefore we tested each of the algorithms separately with 5, 10, 15, and 20 basis vectors. Each source was reconstructed as $\hat{s}^i_{\tau,\nu} = x_{\tau,\nu} v_{\tau,i} t_{i,\nu} / (\sum_{i'=1}^{I} v_{\tau,i'} t_{i',\nu})$. The quality of the separation was evaluated by comparing the separated sources to the original sources. Each separated source was assigned to an original source by using the signal-to-distortion ratio (SDR) between them as a distance measure as described in [4]. If an original source was not assigned separated sources, the source is said to be undetected. The detection error rate was calculated as the ratio of the total number of undetected sources and the total number of sources. The quality of the separated sources was measured by calculating the SDR between each separated source and the corresponding original source. Only a single separated source per an original source was used to avoid over-fitting (see [4]). The SDR was averaged over all the sources. The averages were calculated also separately for pitched instrument sources and percussive sources.

## 4.3. Results

The average detection error rate and SDR are shown in Table 1. The results for DIV, EUC, and DIV-SQ are slightly different from those presented in [4], because of the slightly different source reconstruction method. The proposed method allows better average detection accuracy and SDR than the basic NMF algorithms. It produces approximately equivalent average detection error rate to the DIV-SQ method, the performance being slightly worse for pitched instruments and slightly better for percussive instruments. The SDR of pitched instruments obtained with the proposed method is signifi-
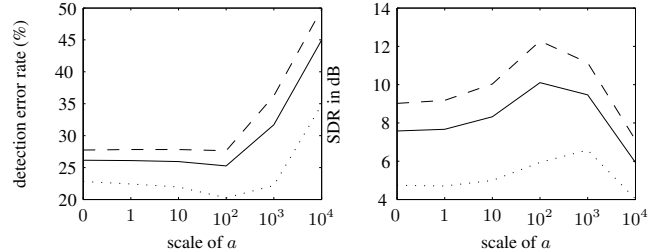


**Fig. 2**. The effect of the coupling parameter $a$ on the average detection error rate and average SDR. The solid line is the average of all sources, the dashed line is the average of pitched instruments, and the dotted line is the average of percussive instruments.

cantly better than the one obtained with the other methods.

The effect of the value of $a$ is illustrated in Figure 2. The case $a = 0$ corresponds to the DIV algorithm. It can be seen that increasing the value of $a$ increases the detection accuracy of percussive instruments and the SDR of all the instruments until a certain point, after which the quality decreases.

## 5. CONCLUSIONS

This paper proposes a Bayesian extension to the NMF where the entries of the unknown matrices are considered as unobserved random variables. We use a Gamma Markov chain prior for the gains and Gamma prior for the basis vectors. These conjugate Gamma priors enable finding the maximum likelihood solution of the parameters by extending the simple and efficient multiplicative updates of the original NMF algorithm, where the likelihood is guaranteed to be non-increasing under each update and therefore the algorithm is guaranteed to converge. The prior structures (both on gains and basis vectors) help to overcome some problems and enable better quality in one-channel source separation than the existing NMF algorithms.

## 6. REFERENCES

[1] A. Taylan Cemgil and Onur Dikmen, "Conjugate Gamma Markov random fields for modelling nonstationary sources," in *Proc. of the 7th Int. Conf. on Independent Component Analysis and Signal Separation*, London, UK, 2007.

[2] Paris Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. of IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics*, New Paltz, USA, 2003.

[3] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. of Int. Conf. on Music Information Retrieval*, Barcelona, Spain, 2004.

[4] Tuomas Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of Neural Information Processing Systems*, Denver, USA, 2000, pp. 556–562.

[6] Bhiksha Raj and Paris Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. of IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics*, New Paltz, USA, 2005.