Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/109579

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
A DYNAMIC LATENT VARIABLE MODEL FOR SOURCE SEPARATION

Anurendra Kumar\textsuperscript{1}, Tanaya Guha\textsuperscript{1}, Prasanta Ghosh\textsuperscript{2}

\textsuperscript{1}Indian Institute of Technology (IIT), Kanpur.
\textsuperscript{2}Indian Institute of Science (IISC), Bangalore.

ABSTRACT

We propose a novel dynamic latent variable model for learning latent bases for time-varying non-negative data. Our model uses a mixture multinomial as the likelihood function and proposes a Dirichlet distribution with dynamic parameters as a prior, which we call the dynamic Dirichlet prior. An expectation maximization (EM) algorithm is developed for estimating the parameters of the proposed model. Furthermore, we connect our proposed dynamic Dirichlet latent variable model (dynamic DLVM) to the two popular latent basis learning methods: probabilistic latent component analysis (PLCA) and non-negative matrix factorization (NMF). We show that (i) PLCA is a special case of the dynamic DLVM, and (ii) dynamic DLVM can be interpreted as a dynamic version of NMF. The effectiveness of the proposed model is demonstrated through extensive experiments on speaker source separation, and speech-noise separation. In both cases, our method performs better than relevant and competitive baselines. For speaker separation, dynamic DLVM shows 1.38 dB improvement in terms of source to interference ratio, and 1 dB improvement in source to artifact ratio.

Index Terms— Latent variable model, Dirichlet distribution, non-negative matrix factorization, source separation

1. INTRODUCTION

Modeling time-varying, non-negative data is critical for many signal processing problems. One such important problem is audio source separation, where time varying, non-negative data arise in the form of magnitude spectra [1]. Source separation is a long standing problem in signal processing, which has widespread applications in speaker recognition, speech enhancement, music editing and audio information retrieval [2, 3].

This paper addresses the problem of modeling time varying non-negative data, looking particularly at the problem of supervised source separation. In the case of supervised source separation, we assume the availability of training data for each source [1, 4]. The training data is used to learn the underlying building blocks i.e., the latent bases for each source. These latent bases are later used to separate the sources from the mixture. Two techniques that have been prominent in this field for learning latent bases are: latent variable model (LVM) [4] and non-negative matrix factorization (NMF) [5, 6]. One of the most popular LVM methods is the probabilistic latent component analysis (PLCA), which has widespread application in acoustic modeling [1]. NMF, on the other hand, is a non-probabilistic approach to learn latent basis with extensive applications to text, image and audio analysis. For certain cost functions, LVM is known to converge to NMF [7], and can be thought of as the probabilistic counterpart of NMF.

The LVM and the NMF in their basic forms do not take into account the temporal correlation in the data. However, many signals, such as music and speech exhibit strong temporal dependence. To address this issue, dynamic variants of LVM and NMF have been developed [8, 9, 10]. Most of these dynamic models capture the temporal dependence in data by imposing temporal constraints on the latent bases and their coefficients [8, 9]. A dynamic variant of PLCA, called the Convolutively PLCA (CPLCA) [8], was proposed to capture the temporal structure in data by assuming the likelihood to be a convolutive mixture. Another dynamic version of PLCA involves dynamic filtering and smoothing, where the coefficient matrix was smoothed by using a vector autoregression (VAR) method [11]. A recent work developed a dynamic NMF (an extension of PLCA) by using an exponential prior [10]. Apart from the LVMs and NMF methods, the hidden Markov model (HMM) has also been extended to model temporal non-negative data [12].

In this paper, we present a novel dynamic LVM for learning latent bases for time varying, non-negative data. Our model uses a mixture multinomial as the likelihood function, and proposes to use a Dirichlet distribution with dynamic parameters as a prior (referred to as the dynamic Dirichlet prior). The mixture multinomial likelihood function is chosen because it is known to yield superior results in source separation and topic modeling [13, 14]. The Dirichlet distribution is the conjugate of multinomial, and has been successfully used (without dynamic properties) as a prior in text modeling [13]. We propose a dynamic variant of the Dirichlet prior in this work, which is particularly suitable for non-negative data. We develop an expectation maximization (EM) algorithm for the proposed model, and derive a maximum a posteriori (MAP) estimate of the parameters. This leads to simple and intuitive update equations due to multinomial-Dirichlet conjugacy. We refer our model as the dynamic Dirichlet latent variable model (dynamic DLVM) in the rest of the paper. Furthermore, we show that (i) the PLCA model is a special case of the dynamic DLVM, and (ii) our model is also a dynamic version of NMF: The effectiveness of the dynamic DLVM is demonstrated through extensive experiments on speaker source separation and speech-noise separation using the SPiB [15] and the TIMIT database [16]. In all cases, our method performs better than several relevant existing methods.

2. PROPOSED DYNAMIC DIRICHLET LATENT VARIABLE MODEL

Let us consider a time varying signal $x(t)$. We represent $x(t)$ spectrographically by taking its short time Fourier transform (STFT), and retaining its scaled magnitude spectrogram $N$

$$N = \gamma |STFT(x(t))| = \gamma X$$

where, $X$ is magnitude spectrogram, $\gamma$ is a large integer which ensures that all the elements in $N$ are integers [1]. The observation spectral data matrix $N$ can be seen as count data, where $N_{ft} \in N$ corresponds to the count of frequency $f$, at a time instant $t$. Each column of the matrix $N$ thus corresponds to the spectral distribution at each time instant. With each frequency count $f \in 1, 2, \ldots, F$, we
We propose to model the temporal dependence between states using a Markovian dependence between states, which uses a dynamic Dirichlet distribution. The dynamic Dirichlet distribution has several appealing properties. We impose a Markovian dependence between states, where \( k \) acts as an indicator for the \( k \)-th latent basis, which is described by a spectral distribution \( P(f|z_k) \).

Latent variable models assume that the underlying cause of an observed variable \( f \) is a set of unobserved latent variables \( z_k \), \( 1 \leq k \leq K \). Marginalizing over the latent bases \( z \), the spectrogram at each time instant \( t \) is a mixture of \( K \) hidden distributions, where \( K \) is a known positive integer

\[
P_t(f) = \sum_{k=1}^{K} P_t(f, z_k) = \sum_{k=1}^{K} P_t(z_k)P(f|z_k) \tag{2}
\]

where, \( P_t(f) \) is the probability of frequency \( f \) at time instant \( t \), \( P(f|z_k) \) is a multinomial distribution similar to PLCA [1] and the coefficients of mixtures are \( P_t(z_k) \). Let us now define a state of the data matrix \( N \) at time \( t \) as \( s_t \); as follows:

\[
s_t = [P_t(z_1), ..., P_t(z_K)]^T = [s_t(1), s_t(2), ..., s_t(K)]^T \tag{3}
\]

We impose a Markovian dependence between states, which uses a Dirichlet distribution. This is described below in detail.

### 2.1. Dynamic DLVM

We propose to model the temporal dependence between states using a Dirichlet distribution with time varying parameters

\[
P(s_t|s_{t-1}, D) = \text{Dir}(\alpha_{t-1} D, s_{t-1} + 1) \tag{4}
\]

where, \( \alpha_t = \sum_f N_{ft} \), \( P(s_1) = \text{Dir}(1) \)

Here, “Dir” denotes the Dirichlet distribution [17]. 1 is an all-one vector, \( \alpha_t \) is the total number of observations at time instant \( t \), and \( D \) is a temporal dependence matrix. For simplicity, we assume \( D \) to be a diagonal matrix with \( D_{kk} = d_k \) for \( 1 \leq k \leq K \), where \( d_k \in \mathbb{R}^+ \) denotes the temporal dependence between two consecutive time instants for the \( k \)-th latent basis. Let us now define pseudo-observation from the previous time instant for each basis \( k \) as \( m_{tk} = \alpha_{t-1} d_k s_{t-1}(k) \). Therefore, Eq. (4) can be rewritten as follows:

\[
P(s_t|s_{t-1}, D) = \frac{\Gamma(\sum_k (m_{tk} + 1))}{\prod_k \Gamma(m_{tk} + 1)} \prod_k s_t(k)^{m_{tk}}
\]

where, \( \Gamma \) is the gamma function. Note that the hyperparameters of the Dirichlet distribution are dynamic, hence, we refer to it as the dynamic Dirichlet distribution in the rest of this paper. The proposed dynamic Dirichlet distribution prior has several appealing properties with intuitive understanding: (i) The generative process (with mixture multinomial as likelihood) allows us to view the spectrogram at time \( t \) as observed count data over \( K \) bases. Static models such as PLCA uses this observation data to estimate the states at each time instant. The dynamic Dirichlet prior allows us to have \( m_{tk} \) extra pseudo-observations for each basis \( k \) (see Eq. (8)), which is the result of multinomial–Dirichlet conjugacy. (ii) The mode of the distribution lies at the normalized pseudo-observations

\[
P_{\text{max}}(s_t(k)|s_{t-1}) = \frac{d_k s_{t-1}(k)}{\sum_k d_k s_{t-1}(k)}
\]

(iii) The variance of each entry of the vector \( s_t \) can be obtained from the properties of Dirichlet distribution [17]

\[
\text{Var}(s_t(k)|s_{t-1}) \propto \frac{1}{(\sum_k m_{tk} + K)^2(\sum_k m_{tk} + K + 1)}
\]

which decreases as total number of observations at previous time instant increases. It is also intuitive, since we expect the distribution to have less variance when we have more prior information from previous time instance. Thus dynamic DLVM assumes the following generative process of the spectrogram \( N \):

- Choose a state, \( s_t \sim \text{Dir}(\alpha_{t-1} D s_{t-1} + 1) \)
- Choose frequency \( f \), \( \alpha_t \) times as follows:
  - Choose a latent basis \( z_t \sim \text{Mult}(s_t) \)
  - Choose a frequency \( f \sim \text{Mult}(P(f|z_t)) \)
- Repeat the above process \( T \) times

where, \( T \) is the total number of time instants. “\( \text{Mult} \)” denotes the multinomial distribution. Fig. 1 presents the graphical model for the generative process.

### 2.2. PLCA as a special case of dynamic DLVM

The relationship between the proposed dynamic DLVM and the well known PLCA [1] model is noteworthy. When the temporal dependence matrix \( D \) is reduced to a zero matrix, the distribution in Eq. (4) becomes a symmetric Dirichlet distribution \( \text{Dir}(1) \). Note that the symmetric Dirichlet distribution \( \text{Dir}(1) \) is nothing but a uniform distribution, and thus the formulation is equivalent to a static PLCA [18]. This can also be intuitively understood as a fact that in the absence of prior information, there exists no preference of any state over others.

### 3. PARAMETER ESTIMATION

In this section, we describe the parameter estimation steps for the proposed dynamic DLVM. We obtain point estimates for \( s_t \) by performing a maximum a posteriori (MAP) estimate.

#### 3.1. Expectation step

The posterior distribution of the latent variable \( z \) is given by

\[
P_t(z_k|f) = \frac{P_t(z_k)P(f|z_k)}{\sum_k P_t(z_k)P(f|z_k)} \tag{5}
\]

#### 3.2. Maximization step

Let us denote the state matrix \( S = [s_1, s_2, ..., s_t] \). Let \( \beta = \{P(f|z), D\} \). We want to maximize the following MAP objective function

\[
\mathcal{L}_{\text{MAP}} = \mathbb{E}_{P_{\beta}(S)} \left( \log P(N, S|\beta) \right) = \mathbb{E}_{P_{\beta}(S)} \left( \log P(N|S, \beta) + \log P(S|\beta) \right)
\]

s.t. \( \sum_f P(f|z_k) = 1, \sum_k s_t(k) = 1, 0 < d_k \)
The objective function $L_{MAP}$ is concave with respect to each parameter $(S, P(f|x), D)$ when others are fixed. Therefore, we update the parameters in an alternating fashion [19].

**Update of $P(f|x)$**

Maximizing the above constrained expected log-likelihood in Eq. (6) with respect to $P(f|z_k)$ yields the following

$$P(f|z_k) = \frac{\sum_f N_{ft}P_t(z_k|f)}{\sum_f \sum_z N_{ft}P_t(z_k|f)}$$

**Update of $S$**

Similarly, maximizing w.r.t. $s_t(k)$, while keeping $D$ fixed yields

$$s_t(k) = \frac{\sum_f N_{ft}P_t(z_k|f) + m_{tk}}{\sum_k (\sum_f N_{ft}P_t(z_k|f) + m_{tk})}$$

**Update of $D$**

Let us define $d = [d_1, ..., d_K]^T$. Maximizing $L_{MAP}$ w.r.t. $d$, keeping $S$ fixed does not have a closed form solution

$$d = \text{argmax}_d \sum_t \log \Gamma(\sum_k (m_{tk} + 1)) - \sum_k \log \Gamma(m_{tk} + 1) + \sum_k m_{tk} \log(s_t(k)) \quad s.t., 0 < d_k.$$ (8)

However, the maximizing function is concave since the Dirichlet distribution is a member of the exponential family. Therefore the local minimum of the function is also a global minimum, which can be obtained via gradient ascent

$$\frac{\partial L_{MAP}}{\partial d_k} = \sum_{t=2}^T \alpha_{t-1}s_{t-1}(k) \left( \psi(\sum_i m_{ti}) + K \right) - \psi(m_{tk} + 1) + \log(s_t(k)).$$

where, $\psi$ is the di-gamma function. It is to be noted that updates of $P(f|x)$ and $S$ are independent of scaling factor $\gamma$. Therefore, we will replace $N$ with $X$ in all the update equations. The complete algorithm is presented in the next section.

**Algorithm 1 Dynamic DLVM as Dynamic NMF**

**Input:** $X$

**Output:** $W, S, d$

**while Not converged do**

$$W_{fk} = W_{fk} \sum_t \frac{X_{ft}}{(WS)_{ft}} S_{kt}$$

$$W_{fk} = W_{fk}/\sum_k W_{fk}$$

**while Not converged do**

$$m_{tk} = \alpha_{t-1}d_{k} s_{t-1}(k)$$

$$S_{kt} = S_{kt} \sum_f W_{fk} \frac{X_{ft}}{(WS)_{ft}} + m_{tk}$$

$S_{kt} = S_{kt}/\sum_t S_{kt}$

Update $d$ using Eq. 9

**end**

**end**

**5. APPLICATION TO SOURCE SEPARATION**

We demonstrate the effectiveness of the proposed dynamic DLVM through its ability to perform supervised source separation. We first learn the basis matrix $W$ for each source, and then separate the sources by employing a standard source separation algorithm following an earlier work [4]. Note that we only use the latent bases (and not the dependency matrix $D$) during the source separation.

We perform two types of source separation experiments: (i) speaker source separation, where mixtures contain speech sources from two speakers, and (ii) noise-source separation, where mixtures contain speech and noise.

**Experimental details**

We follow an experimental setup similar to that described in past works on source separation using PLCA and its variants [20, 21]. The source separation experiments use samples from the TIMIT database [16], and noise data from SPIB [15]. The magnitude spectrograms are obtained by performing STFT on a 64ms window with 16ms overlap. We learn $K = 30$ latent basis in each case, and use the maximum number of iterations (250 for outer loop, 8 for inner loop) as the convergence criterion for Algorithm 1.

To evaluate the performance on source separation, we use four evaluation metrics: signal-to-noise ratio improvement (SNRI) [22], source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) [23, 24, 25]. The later three metrics measure perceptual quality of the separated sources. The SNR improvement (SNRI) of a speaker is calculated by incorporating the phase information and by comparing the improvement in SNR with that of the mixture signal as defined in literature [22, 4].

**Speaker source separation**

Following the experimental setup of the past literature [20, 21], we have used $\sim 25$ seconds of speech (8 to 9 sentences) from 10 speakers (5 male, 5 female) in the database for our experiments. To model each speaker (source), the first $\sim 17$ seconds of the speech is used.
The performance of dynamic DLVM is compared against those of three baseline methods – PLCA [1], PLCA with dynamic filtering [11] and PLCA with dynamic smoothing [11]. Dynamic DLVM performs better than or comparable to the baseline methods in terms of all evaluation metrics. Our model outperforms PLCA by 0.96 dB in SNRI, 0.87 dB in SDR, 1.38 dB in SIR, and 0.46 dB in SAR. The improvement in terms of SAR implies that the artifacts introduced by dynamic DLVM is lesser than the other models. Usually, there is a trade-off between removing noise (measured by SDR and SNRI) and introducing artifacts (measured by SAR). The existing dynamic models [11, 10, 26, 12] while improving SDR often introduce artifacts, which leads to a degraded SAR. However, the proposed model shows simultaneous improvement in SDR and SAR. This indicates an overall better modeling ability of dynamic DLVM, and consequently, a better source separation.

### Speech and noise separation

We consider a speech denoising scenario where prior information about the noise types and the associated training data is available. Both the noise and the speech are first normalized to have zero mean and unit variance. The noisy mixtures were obtained by adding noise (one at a time) to each speaker signal, resulting into a signal to noise ratio of 0 dB. We experiment with five noise types: babble, factory, white, pink and cockpit [15], and speech from 10 speakers used in the speaker source separation experiments. The latent bases are learned for each speaker and noise from their respective training data (same parameter values as before). Fig. 5 presents a sample qualitative result of denoising a source corrupted with babble noise. The performance of dynamic DLVM averaged over 10 mixtures is listed in Table 1, and compared with the baselines. Dynamic DLVM, on average, shows an improvement of 0.9 dB. Note that it performs better than other methods for all noise types, except white noise. This can be explained by the fact that white noise is stationary and has no temporal structure. Nevertheless, for non-stationary noise, the proposed model is able to learn the temporal dependencies in data/noise, which results in better separation. As observed earlier, dynamic DLVM shows 1dB SAR improvement for all noise types as compared to PLCA. This observation supports our earlier claim that our model introduces less artifacts compared to other dynamic models in literature [11].

### 6. CONCLUSION

We proposed a latent variable model, called the dynamic DLVM, for modeling time varying non-negative data. We introduced a new prior (dynamic Dirichlet distribution) and used a multinomial as likelihood function to a dynamic version of NMF. Unlike other dynamic latent variable models, our model does not require any free parameter (except the number of latent bases). Although this work involves modeling magnitude spectra, the proposed model is generic and suitable for modeling other types of non-negative data.

### 7. ACKNOWLEDGEMENT

This work is partially supported by the grant IITK/EE/2015052.
8. REFERENCES


