OPTIMAL WEIGHT LEARNING FOR COUPLED TENSOR FACTORIZATION WITH MIXED DIVERGENCES

Umut Şimşekli, Beyza Ermiş, A. Taylan Cemgil
Boğaziçi University
Dept. of Computer Engineering
34342, Bebek, İstanbul, Turkey

Evrim Acar
University of Copenhagen
Faculty of Science
DK-1958 Frederiksberg C, Denmark

ABSTRACT

Incorporating domain-specific side information via coupled factorization models is useful in source separation applications. Coupled models can easily incorporate information from source modalities with different statistical properties by estimating shared factors via divergence minimization. Here, it is useful to use mixed divergences, a specific divergence for each modality. However, this extra freedom requires choosing the correct divergence as well as an optimal weighting mechanism to select the relative ‘importance’. In this paper, we present an approach for determining the relative weights, framed as dispersion parameter estimation, based on an inference framework introduced previously as Generalized Coupled Tensor Factorization (GCTF). The dispersion parameters play a key role on inference as they form a balance between the information obtained from multimodal observations. We tackle the problem of optimal weighting by maximum likelihood exploiting the relation between $\beta$-divergences and the family of Tweedie distributions. We demonstrate the usefulness of our approach on a drum source separation application.

Index Terms— Informed source separation, Coupled Matrix/Tensor Factorization, Tweedie models, $\beta$-divergence

1. INTRODUCTION

Blind source separation methods are popular tools and have been shown to be useful in many domains [1]. However, when it comes to separating complex acoustic signals using only a single channel of observations, these methods fall short as they don’t exploit the highly structured patterns in such signals. Recent studies show that providing domain-specific information to source separation methods increases the separation performance [2, 3, 4].

One way to incorporate domain-specific knowledge is to jointly analyze side information with the observed audio mixtures using coupled factorization models. For instance, Kim et al. present a coupled matrix factorization model for drum source separation in polyphonic music signals by incorporating side information obtained from a collection of drum recordings [5]. Similarly, coupled factorizations are used for audio source separation in [3], where the audio mixture is coupled with isolated sounds and symbolic music data.

Coupled factorization models jointly factorize data from multiple information sources by extracting common factors. We first consider the following simple coupled matrix factorization model and, in the next section, we construct a drum separation model based on a more complicated coupled factorization problem. We define the toy coupled model as follows:

$$X_1(f, t) \approx \hat{X}_1(f, t) = \sum_i Z_1(f, i)Z_2(i, t)$$

$$X_2(f, n) \approx \hat{X}_2(f, n) = \sum_i Z_1(f, i)Z_3(i, n)$$

where $X_1$ and $X_2$ are two observed matrices decomposed by a Nonnegative Matrix Factorization (NMF) model. In an audio processing application, $X_1$ and $X_2$ may denote audio spectra (where $f$ denotes frequency and $t$ and $n$ denote time frames), $Z_1$ may correspond to a spectral dictionary while $Z_2$ and $Z_3$ may denote the corresponding excitations. Here, the factor $Z_1$ exists in both models, which makes these models coupled. If $X_1$ consists of a mixture of sound sources to be separated, we can incorporate further information to the source separation model using such coupled factorization models.

Given the observed matrices $X_1$ and $X_2$, the coupled factorization problem estimates the factors, i.e., $Z_{1:3} \equiv \{Z_1, Z_2, Z_3\}$; in other words, solves the following optimization problem:

$$\left(Z_{1:3}\right)^* = \arg\min_{Z_{1:3}} \frac{1}{\phi_1} d_1(X_1 || \hat{X}_1) + \frac{1}{\phi_2} d_2(X_2 || \hat{X}_2)$$

(1)

where $d_1(\cdot)$ and $d_2(\cdot)$ are (possibly different) divergences, usually chosen as the Euclidean, Kullback-Leibler (KL) or the Itakura-Saito (IS) divergences. In our context, the (inverse)
weights $\phi_1$ and $\phi_2$ will be called dispersion parameters and balance the information obtained from different modalities. We have previously illustrated the effect of correctly choosing a divergence on coupled models [6]. Similarly, dispersion parameters are also expected to play a significant role on the performance of coupled models.

While coupled factorization models have been widely studied in many fields [6, 7, 8], determining the right weighting scheme remains to be a major challenge in data fusion studies [9]. Wilderjans et al. [10] proposed a maximum likelihood approach for estimating the weights under Gaussian observation models. In this study, we tackle this problem by using a probabilistic approach, which makes use of the relation between the $\beta$-divergence and the family of Tweedie distributions and enables us to learn the dispersion parameters by maximizing the likelihood. We demonstrate that estimating the dispersion is both straightforward and useful in audio source separation.

2. DRUM SEPARATION MODEL

In this section, we present a coupled matrix factorization model for drum separation for professionally recorded audio. This model combines the information that is gathered from the audio mixture, isolated drum sounds and an approximate transcription of drum track of the audio mixture.

Suppose we observe the magnitude spectrum of an audio mixture $X_1(f, t)$, where $f$ and $t$ denote the frequency and time frame indices, respectively. Here, we assume that matrix $X_1$ is decomposed by using an NMF model:

$$X_1(f, t) \approx \hat{X}_1(f, t) = \sum_i D(f, i)G(i, t)$$

where $D$ is the spectral dictionary and $G$ is the corresponding excitation matrix. Since the aim of this model is drum track separation, we will assume that some of the spectral templates, say the first $I_b$ columns of $D$, denoted as $D(:, 1 : I_b)$ model the background sources and the remaining model the drum track. Suppose we observe another magnitude spectrum $X_2(f, n)$, which is obtained from a database of drum sounds. Here $f$ is again the frequency index and $n$ is the time frame index. We can also decompose this matrix using a similar approach:

$$X_2(f, n) \approx \hat{X}_2(f, n) = \sum_i D(f, i)T_1(f, i)E(i, n)$$

where $D$ is the same spectral dictionary in Equation 2 and $E$ is the excitation matrix for the example drum sounds. Here, $T_1$ is a pre-determined binary matrix that makes sure that the drum sounds use only the related spectral templates: $T_1$ takes values of 0 for the background part of the dictionary and 1 otherwise: $T_1(:, 1 : I_b) = 0$ and $T_1(:, I_b + 1 : I) = 1$, where $I$ is the number of spectral templates. So far, we have the coupled factorization model of [5], which incorporates the spectral information that is obtained from the drum sounds to the drum separation model.

As we are modeling musical signals, we may also assume the excitation matrix $G$ is composed of the superposition of some certain patterns that repeat over time. With this assumption, we can also factorize the matrix $G$ using another NMF model as follows: $G(i, t) = \sum_k B(i, k)F_2(k, t)$, where $B$ is the dictionary for the excitations and $F$ denotes the excitations that correspond to this dictionary. By making use of the relation between an excitation matrix and a musical score, we can also couple the matrices $B$ and $F$ with an approximate transcription of the drum track as follows:

$$X_3(i, t) \approx \hat{X}_3(i, t) = \sum_k B(i, k)T_2(i, k)F(k, t).$$

Here, $X_3(i, t)$ takes a constant value if a drum event $i$ (e.g. snare hit, hi-hat hit, etc.) is present at time frame $t$, and becomes 0 otherwise. Furthermore, $T_2$ is another pre-determined binary matrix similar to $T_1$, where $T_2(1 : I_b, :) = 0$ and $T_2(I_b + 1 : I, :) = 1$. Note that, decomposition of the excitation matrix by this approach is shown to be useful in musical source separation [3].

Finally, we can define the combined model as follows:

$$\hat{X}_1(f, t) = \sum_{i,k} D(f, i)B(i, k)F(k, t)$$

Mixture (5)

$$\hat{X}_2(f, n) = \sum_i D(f, i)T_1(f, i)E(i, n)$$

Drum (6)

$$\hat{X}_3(i, t) = \sum_k B(i, k)T_2(i, k)F(k, t)$$

MIDI (7)

Figure 1 illustrates this model. The goal is to estimate the latent factors $D$, $B$, $F$, and $E$ as the sources can be separated by Wiener filtering after the factors $D$, $B$, and $F$ are obtained. This problem can be formulated as an optimization problem with the following objective function with mixed divergence:

$$d(X_{1;3}||\hat{X}_{1;3}) = \sum_{\nu=1}^3 \frac{1}{\phi_\nu} d_\nu(X_\nu||\hat{X}_\nu),$$

where $d_\nu(\cdot)$ are divergence functions and $\phi_\nu$ are dispersion parameters as defined in Section 1. Since we are dealing with audio signals, selecting $d_1(\cdot)$ and $d_3(\cdot)$ as Itakura-Saito divergence would be appropriate as suggested in [11]. Besides, we may want to select $d_3(\cdot)$ as the KL divergence. However, we may wish to give different weights to each observed matrix. In particular, we know that the transcription matrix $X_3$ is not very accurate, we don’t need to fit this matrix precisely. The dispersion parameters play a key role here, as they determine the noise variance of each observed matrix. For this particular example, selecting a large $\phi_3$ seems to be an accurate modeling strategy.
3. GENERALIZED COUPLED TENSOR FACTORIZATION

The Generalized Coupled Tensor Factorization (GCTF) framework [12] is a generalization of matrix and tensor factorization models to jointly factorize more than one multiway array (tensor or matrix). The formal definition of the GCTF framework is as follows:

\[ X_\nu(u_\nu) \approx \hat{X}_\nu(u_\nu) = \sum_{u_\alpha} \prod_{\alpha} Z_\alpha(v_\alpha)^{R_{\nu,\alpha}} \]  

where \( \nu = 1, \ldots, |\nu| \) is the observed tensor index and \( \alpha = 1, \ldots, |\alpha| \) is the factor index. In this framework, the goal is computing an approximate factorization of given observed tensors \( X_\nu \) in terms of a product of individual factors \( Z_\alpha \), some of which are possibly shared. Here, we define \( V \) as the set of all indices in a model, \( U_\nu \) as the set of visible indices of the tensor \( X_\nu \), \( V_\alpha \) as the set of indices in \( Z_\alpha \), and \( U_\nu = V - U_\nu \) as the set of invisible indices that is not present in \( X_\nu \). We use small letters as \( v_\alpha \) to refer to a particular setting of indices in \( V_\alpha \). Furthermore, \( R \) is a coupling matrix that is defined as follows: \( R_{\nu,\alpha} = 1 \) (0) if \( X_\nu \) and \( Z_\alpha \) connected (otherwise). In other words, the coupling matrix \( R_{\nu,\alpha} \) specifies the factors \( Z_\alpha \) that effect the observed tensor \( X_\nu \).

As the product \( \prod_{\alpha} Z_\alpha(v_\alpha) \) is collapsed over a set of indices, the factorization is latent. The optimization problem is to minimize the total discrepancy between the observations \( X_\nu \) and the model output \( \hat{X}_\nu \), as given by a divergence function \( d_\nu(X_\nu||\hat{X}_\nu) \). This divergence is a quasi-squared distance typically taken as Euclidean, KL or IS. The objective function to be minimized has a similar form as in Equation 8:

\[ Z_{1:|\alpha|} = \arg \max_{Z_{1:|\alpha|}} \sum_{\nu} \frac{1}{\phi_\nu} d_\nu(X_\nu||\hat{X}_\nu). \]

In order to illustrate our notation, we define the drum separation model in the GCTF notation as follows. We can define \( Z_1:6 \equiv \{D, B, F, T_1, E, T_2\} \), the observed index sets: \( U_1 = \{f, i\} \), \( U_2 = \{f, n\} \), and \( U_3 = \{i, t\} \), the index sets of the factors: \( V_1 = \{f, i\} \), \( V_2 = \{i, k\} \), \( V_3 = \{k, t\} \), \( V_4 = \{f, i\} \), \( V_5 = \{i, n\} \), and \( V_6 = \{i, k\} \).

3.1. Estimation

Estimation of the latent factors \( Z_\alpha \) can be achieved via iterative methods, by fixing all factors \( Z_\alpha \) for \( \alpha' \neq \alpha \) but one \( Z_\alpha \) and updating in an alternating fashion (see [12]). For non-negative data and factors, the update has a simple form:

\[ Z_\alpha \leftarrow Z_\alpha \odot \frac{\sum_{\nu} R^{\nu,\alpha} \phi_\nu^{-1} \Delta_{\alpha,\nu'}(\hat{X}_\nu^{-p_\nu} \circ X_\nu)}{\sum_{\nu} R^{\nu,\alpha} \phi_\nu^{-1} \Delta_{\alpha,\nu'}(X_\nu^{-1})}, \]

where \( \odot \) is the element-wise product (Hadamard product) and the parameter \( p_\nu \) determines the cost function to be used for \( X_\nu \): for \( p_\nu = \{0, 1, 2\} \) correspond to the \( \beta \)-divergence [13] that unifies Euclidean, KL, and IS cost functions, respectively. The key quantity in the above update equation is the \( \Delta_{\alpha,\nu'} \) function that is defined as follows:

\[ \Delta_{\alpha,\nu'}(A) = \left[ \sum_{v_\alpha \cap \tilde{v}_\alpha} A(v_\alpha) \prod_{\alpha' \neq \alpha} Z_{\alpha'}(v_{\alpha'}) R^{\nu,\alpha'} \right] \]

For updating \( Z_\alpha \), we need to compute this function twice for arguments \( A = X_\nu^{-p_\nu} \circ X_\nu \) and \( A = X_\nu^{1-p_\nu} \).

4. LEARNING THE DISPERSION PARAMETERS

Exponential dispersion models (EDM’s) are a well-studied family of distributions and have found place in various fields. The Tweedie distribution is a special case of the EDM [14], itself a generalization of the more familiar natural exponential family. There is a close connection between \( \beta \)-divergences and the family of Tweedie distributions where the relation was stated by Cichocki et al. [13] and elaborated in [15]. A
Table 1: Tweedie distributions with corresponding normalizing constants, divergence forms and maximum a-posteriori estimation of the dispersion parameters. Here $N_\nu$ denotes the number of elements in $X_\nu$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>Divergence</th>
<th>Distribution</th>
<th>Normalizing Constant</th>
<th>Divergence Form</th>
<th>$\phi^*_\nu$ (MAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 − $\beta$</td>
<td>$\beta$-divergence</td>
<td>Tweedie</td>
<td>$z_{p_\nu}(X_\nu; \phi_\nu)$</td>
<td>$\frac{1}{2} \sum { (X_\nu - \bar{X}_\nu)^2 }$</td>
<td>$d_{p_\nu}(X_\nu; \bar{X}_\nu)$</td>
</tr>
<tr>
<td>0</td>
<td>EUC</td>
<td>Gaussian</td>
<td>$(2\pi)^{-1/2}$</td>
<td>$\frac{1}{2} \sum u_{\nu} (X_\nu - X_\nu)^2$</td>
<td>$d_0(X_\nu; X_\nu) + \beta_\nu$</td>
</tr>
<tr>
<td>1</td>
<td>KL</td>
<td>Poisson</td>
<td>$\frac{X_\nu}{\nu} \frac{\nu^{\nu/2}}{\Gamma(\nu/2)} e^{-X_\nu/\nu}$</td>
<td>$\sum u_{\nu} \log \frac{X_\nu}{\bar{X}<em>\nu} - X</em>\nu + \bar{X}_\nu$</td>
<td>$d_1(X_\nu; X_\nu) + \beta_\nu$</td>
</tr>
<tr>
<td>2</td>
<td>IS</td>
<td>Gamma</td>
<td>$(\Gamma(1/\nu)(\nu^{1/\nu})^{1/\nu}; X_\nu)^{-1}$</td>
<td>$\sum u_{\nu} X_\nu - \log \frac{X_\nu}{\bar{X}_\nu} - 1$</td>
<td>$d_2(X_\nu; X_\nu) + \beta_\nu$</td>
</tr>
<tr>
<td>3</td>
<td>−</td>
<td>Inverse Gaussian</td>
<td>$(2\pi X_\nu^2)^{-1/2}$</td>
<td>$\sum u_{\nu} (X_\nu - X_\nu)^2 \frac{X_\nu}{X_\nu + \bar{X}_\nu}$</td>
<td>$d_3(X_\nu; X_\nu) + \beta_\nu$</td>
</tr>
</tbody>
</table>

The Tweedie distribution is specified by a simple power relation between its mean and variance as: $Var\{x\} = \phi x^p$, where $\bar{x}$ is the mean (also called expectation parameter), $p = 2 - \beta$ is the index parameter of the $\beta$-divergence and $\phi$ is the dispersion parameter. They also fully characterize the dispersion model and can be written in the following moment form

\[\mathbb{P}(x; \bar{x}, \phi, p) = z_p(x, \phi) \exp \left(-\frac{1}{\phi} d_p(x; \bar{x})\right)\]  \hspace{1cm} (13)

where $z_p(\cdot)$ is a suitable normalization constant that depends on the index parameter $p$. Since $z_p$ does not depend on $\bar{x}$, for fixed $p$ and $\phi$, solving a maximum likelihood problem for $\bar{x}$ is indeed equivalent to minimization of the $\beta$-divergence. However, for fixed $\bar{x}$, solving a maximum likelihood problem for $\phi$ depends on $z_p(\cdot)$ that has a closed form characteristic function:

\[\phi^* = \arg \max_{\phi} \log z_p(x, \phi) - \frac{1}{\phi} d_p(x; \bar{x}).\]  \hspace{1cm} (14)

In this study, we focus on learning the dispersion parameters $\phi_\nu$ for integer values of $p_\nu = \{0, 1, 2, 3\}$. Other values of $p_\nu$ are possible but technically more difficult because the normalizing constants $z_p$ are given as power series without an explicit analytical form [16]. From the probabilistic interpretation, maximization of likelihood with respect to the dispersion parameters $\phi_\nu$ when the other parameters are given in a Tweedie model, provides a data driven formulation for choosing relative weights. In addition, it is easy to verify that when mean is fixed, the conjugate prior of the dispersion parameter is the inverse Gamma distribution. Hence, we assume an inverse Gamma prior on $\phi_\nu$: $\phi_\nu \sim IG(\alpha_\nu; \beta_\nu)$. Surprisingly, none of the references we are aware of used this conjugate prior for the dispersion parameter.

Tweedie distribution generalizes the well-known distributions where the choices of $p_\nu = \{0, 1, 2, 3\}$ correspond to the Gaussian, Poisson, Gamma, and inverse Gaussian distributions, respectively. We estimate the optimal dispersion for these distributions by setting the derivative of the log-likelihood to zero and then solving it for $\phi_\nu$. Note that, since the normalizing constants of the Poisson and the Gamma distributions are problematic (see Table 1), we replace the terms involving $\log \Gamma(\cdot)$ functions with the so called Stirling’s approximation: $\log \Gamma(n) \approx -\frac{1}{2} \log n + n \log n - n$. Finally, we obtain the optimal dispersion parameters in closed form as presented in Table 1.

As a summary, at each iteration of the estimation algorithm, we first update all factors $Z_\alpha$ via Equation 11 and compute the mean parameters $\bar{X}_\nu$ via Equation 9. Then, for each observed tensor we compute the maximum a-posteriori (MAP) estimation of the dispersion parameters $\phi^*_\nu$ as in Table 1. This coordinate ascent procedure is iterated until convergence.

Note that, the Poisson distribution in its well-known form, is an exponential dispersion model with $\phi_0 = 1$. When we introduce a dispersion parameter, we re-define domain of the probability distribution on integer multiples of $\phi_\nu$: on a grid $\{0, 2\phi_\nu, 3\phi_\nu, \ldots\}$. Therefore, for the observations with $p_\nu = 1$, we should set $X_\nu \leftarrow X^\text{org} / \phi_\nu$ at the beginning of each iteration, where $X^\text{org}$ denotes the original observation.

5. EXPERIMENTS AND CONCLUSION

In this section, we illustrate our method on the drum source separation model defined in Section 2. Here, we conduct our experiments on a famous pop song ‘Chasing Pavements’ by Adele. Firstly, we estimate the factors without using the dispersion parameters (i.e. setting $\phi_{1:3} = 0$) and then we also estimate the dispersion parameters by using the proposed method while making inference.

We compute the magnitude spectrum of a 20 second excerpt of the piece and obtain $X_1$. For $X_2$, we compute the spectra of the drum sounds that are obtained from the RWC Musical Instrument Sound Database. Finally, we compute $X_3$.
by using an approximate transcription of the drum track of the piece, which can be obtained from online MIDI databases. In our experiments we used $|I| = 813$ (the number of spectral templates); 13 templates for the drum part and 800 templates for the harmonic part, $|I| = 100$, and we set the divergence parameters as $p_1 = 2$, $p_2 = 2$, and $p_3 = 1$. The hyper-parameters are selected as $\alpha_1:3 = 100$, $\beta_1 = 9$, $\beta_2 = \beta_3 = 9900$. The sources are separated via Wiener filtering.

Figure 2 visualizes the drum source separation results for the particular experiment. It can be visually observed that estimating the dispersion parameters while estimating the other factors yields better results. We can see that, the high frequency components of the drum sounds cannot be recovered if the dispersion parameters are fixed to the same number. The resulting audio files and further experimental results on a tensor completion problem can be found in http://www.cmpe.boun.edu.tr/~umut/eusipco2013.

**Conclusion:** The dispersion parameters play a significant role on the performance of coupled matrix/tensor factorization models. In this study, we presented a method for estimating the dispersion parameters in coupled models. Our method follows a systematic approach, where we formulated this problem as a maximum likelihood estimation by making use of the relation between the $\beta$-divergence and the family of Tweedie distributions.

We applied our method on a coupled drum source separation model. We observed a certain improvement on results when the dispersion parameters are estimated while making inference on the coupled factorization model.

### 6. REFERENCES


