Learning the β-Divergence in Tweedie Compound Poisson Matrix Factorization Models

Umut Şimşekli UMUT.SIMSEKLI@BOUN.EDU.TR
Dept. of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey

Ali Taylan Cemgil TAYLAN.CEMGIL@BOUN.EDU.TR
Dept. of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey

Yusuf Kenan Yılmaz KENAN@SIBNET.COM.TR
Sibnet Computers Ltd, 34742 Kadıköy, Istanbul, Turkey

Abstract

In this study, we derive algorithms for estimating mixed β-divergences. Such cost functions are useful for Nonnegative Matrix and Tensor Factorization models with a compound Poisson observation model. Compound Poisson is a particular Tweedie model, an important special case of exponential dispersion models characterized by the fact that the variance is proportional to a power function of the mean. There are several well known matrix and tensor factorization algorithms that minimize the β-divergence; these estimate the mean parameter. The probabilistic interpretation gives us more flexibility and robustness by providing us additional tunable parameters such as power and dispersion. Estimation of the power parameter is useful for choosing a suitable divergence and estimation of dispersion is useful for data driven regularization and weighting in collective/coupled factorization of heterogeneous datasets. We present three inference algorithms for both estimating the factors and the additional parameters of the compound Poisson distribution. The methods are evaluated on two applications: modeling symbolic representations for polyphonic music and lyric prediction from audio features. Our conclusion is that the compound poisson based factorization models can be useful for sparse positive data.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Non-negative Matrix Factorization (NMF) is a widely used algorithm for data analysis. The goal is calculation a factorization of the form:

$$X(i,j) \approx \hat{X}(i,j) = \sum_k Z_1(i,k) Z_2(k,j) \qquad (1)$$

where X is the given data matrix, $\hat{X}$ is an approximation to X, and Z_1, and Z_2 are non-negative factor matrices. This model has been applied to various fields including signal processing, finance, bioinformatics, and natural language processing (Cichocki et al., 2009). One of the most popular approaches for computing factorizations is based on minimization of a divergence D:

$$(Z_1^*, Z_2^*) = \underset{Z_1, Z_2 \geq 0}{\arg\min}\, D(X||\hat{X}). \qquad (2)$$

In practice, a separable divergence $D(X||\hat{X}) = \sum_{i,j} d(X(i,j)||\hat{X}(i,j))$ is used. Some popular divergence (i.e., cost) functions are special cases of the β-divergence, defined as $p = 2 - \beta$:

$$d_p(x;\hat{x}) = \frac{x^{2-p}}{(1-p)(2-p)} - \frac{x\hat{x}^{1-p}}{1-p} + \frac{\hat{x}^{2-p}}{2-p} \qquad (3)$$

where p is an index parameter. By taking appropriate limits it is easy to verify that d_p is the Euclidean distance square, information divergence or Itakura-Saito divergence (Févotte et al., 2009) for $p = 0, 1$ and 2 respectively.

The key idea of the current paper is to exploit the close connection between β-divergences and a particular exponential family, the so-called Tweedie models (Yılmaz & Cemgil, 2012). It turns out that Tweedie

densities, to be described in more detail in the following section, can be written in the following moment form

$$\mathbb{P}(x;\hat{x},\phi,p) = \frac{1}{Z(x,\phi,p)}\exp\left(-\frac{1}{\phi}d_p(x;\hat{x})\right) \quad (4)$$

where $\hat{x}$ is the mean, ϕ is the dispersion and p is the index parameter of the β-divergence defined in (3). An important property is that the normalization constant Z does not depend on $\hat{x}$; hence it is easy to see that for fixed p and ϕ, solving a maximum likelihood problem for $\hat{x}$ is indeed equivalent to minimization of the β-divergence.

Note that for the familiar Gaussian case, we have $d_0(x;\hat{x}) = (x-\hat{x})^2/2$

$$\mathbb{P}(x;\hat{x},\phi,p=0) = \frac{1}{\sqrt{2\pi\phi}}\exp\left(-\frac{1}{\phi}(x-\hat{x})^2/2\right) \quad (5)$$

the dispersion is simply the variance. As for all admissible p we have a similar form, Tweedie models generalize the established theory of least squares linear regression to more general noise models (restricted to identity link functions).

Matrix factorization (MF) is often viewed as a divergence minimization problem, and various algorithms for solving the optimization problem in (2) have been proposed. Often, multiplicative updates are used in practice for their simplicity, yet many extensions and variations have been proposed (Yılmaz et al., 2011). However, the divergence minimization perspective does not provide a complete picture of MF models. One key question is the choice of the divergence. In practice, several divergence functions are tried on the problem and models are evaluated according to an application specific success criterion. Another problem arises in collective factorization, for example when we wish to decompose several matrices collectively as in the following block matrix model

$$[X_1, X_2] \approx [\hat{X}_1, \hat{X}_2] = Z_1[Z_2, Z_3]. \quad (6)$$

This can be viewed as a coupled factorization of X_1 and X_2 where the factor Z_1 is being shared. If the data matrices are representing different modalities, it is natural that we might want to choose a cost function that puts more emphasis on one matrix using weights as

$$\text{Cost}(Z_{1:3}) = \phi_1^{-1}D_{p_1}(X_1||Z_1Z_2) + \phi_2^{-1}D_{p_2}(X_2||Z_1Z_3).$$

We will refer to such cost functions as mixed β-divergences. The probabilistic perspective provides here a natural, data driven formulation in choosing the relative weights by maximization of a joint likelihood with respect to the dispersion parameters ϕ_ν and possibly the individual divergences D_{p_ν} via determination of p_ν for $\nu = 1, 2$.

2. Exponential Dispersion Models and the Tweedie Family

The Tweedie family is a particular exponential dispersion model (EDM) (Jørgensen, 1997). EDM's are a well-studied family of distributions and have found place in various fields. It has an important role at statistical data analysis as the response distribution of the generalized linear models (McCulloch & Nelder, 1989).

An exponential dispersion model (in canonical form) can be defined by a two parameter density as follows (Jørgensen, 1997):

$$\mathbb{P}(x;\theta,\phi) = h(x,\phi)\exp\left\{\frac{1}{\phi}\left(\theta x - \kappa(\theta)\right)\right\} \quad (7)$$

where θ is the canonical (natural) parameter, ϕ is the dispersion parameter and κ is the cumulant (log-partition) function ensuring normalization. Here, $h(x,\phi)$ is the base measure and is independent of the canonical parameter. For EDM, it is easy to verify that the mean $\hat{x}$ (also called expectation parameter) and the variance $\text{Var}\{x\}$ are obtained directly from the first and second derivatives of $\kappa(\cdot)$ with respect to the canonical parameter

$$\kappa'(\theta) = \langle x\rangle_{p(x;\theta,\phi)} \equiv \hat{x} \quad (8)$$

$$\kappa''(\theta) = \frac{1}{\phi}\text{Var}\{x\} \equiv v(\hat{x}). \quad (9)$$

Here $v(\hat{x})$, the second derivative, is also known as the variance function (Tweedie, 1984; Bar-Lev & Enis, 1986; Jørgensen, 1997).

As a special case of EDMs, Tweedie distributions $\mathcal{TW}(x;\hat{x},\phi,p)$ specify the variance function as

$$v(\hat{x}) = \hat{x}^p \quad (10)$$

The variance function is related to the p'th power of the mean, therefore it is called a power variance function. Note that this choice directly dictates the form of $\kappa(\theta)$ that can be solved as

$$\kappa(\theta) = \begin{cases} \frac{1}{2-p}\left((1-p)\theta\right)^{\frac{2-p}{1-p}} & p \neq 1,2 \\ -1-\log(-\theta) & p = 2 \\ \exp(\theta) & p = 1 \end{cases}. \quad (11)$$

Here, different choices for p yield well-known important distributions such as the Gaussian ($p = 0$), Poisson ($p = 1$), compound Poisson ($1 < p < 2$), Gamma

($p = 2$) and inverse Gaussian ($p = 3$) distributions. Excluding the interval $0 < p < 1$ for which no EDM exists, for all other values of p not mentioned above, one obtains stable distributions (Jørgensen, 1997).

In this study, we focus on the inference in the matrix/tensor factorization models with $p \in (1, 2)$ and p is unknown. Tweedie distribution with $p \in (1, 2)$ is equivalent to the compound Poisson distribution and has a support for continuous positive data and a discrete probability mass at zero. The presence of the discrete mass at zero makes this distribution suitable for many applications where observations are often zero but sometimes are positive. Handling this using a single family has been illustrated to be useful in many applications, including actuarial science (no claim/claim amount), rainfall modeling (no rain/rain amount), fishery prediction (no catch/some catch) (Dunn & Smyth, 2005).

Maximum likelihood estimation of the compound Poisson distribution is relatively simple only if the index parameter p is known beforehand. If p is not known, it is a quite challenging task to make inference on the compound Poisson models. Related to this problem, in (Zhang, 2012) the authors present likelihood-based inferential methods and a Monte Carlo EM algorithm for making inference in compound Poisson models. In another recent study (Lu et al., 2012), the authors present a score matching method for finding the best p for the simpler case where they assumed unitary dispersion. In this study, we present three methods for making inference in matrix/tensor factorization models with compound Poisson observation models. In the first and the second methods, we follow a variational approach, where in the third method we integrate out the dispersion parameter. We evaluate the proposed methods on two applications. Firstly, we evaluate our methods on modeling symbolic representations for polyphonic music. Secondly, we define a novel coupled tensor factorization model and evaluate our methods on prediction of the lyrics of a song from its audio features.

3. The Compound Poisson Distribution

The goal in this section is to give a compact characterization of the compound Poisson distribution as a Tweedie model (Jørgensen, 1997). We will show that the Tweedie density with $p \in (1, 2)$ coincides with the compound Poisson density. A random variable x that is the sum of n independent and identically distributed Gamma random variables is compound Poisson distributed, when n is Poisson distributed. The generative model is (Jørgensen, 1997):

$$x = \sum_{i=1}^{n} g_i \tag{12}$$

where n and g_i are

$$n \sim \mathcal{PO}(n; \lambda) \qquad g_i \sim_{\text{iid}} \mathcal{G}(g_i; a, b) \tag{13}$$

Here, $\mathcal{PO}$ and $\mathcal{G}$ denote the Poisson and Gamma densities, respectively. The marginal density $\mathbb{P}(x)$ is compound Poisson. More compactly, we can also write $x|n \sim \mathcal{G}(x; an, b)$.

To show the equivalence to the Tweedie, we first note that the cumulant generating function (CGF) $K_u(s)$ of a random variable u with density $\mathbb{P}(u)$ is defined as $K_u(s) = \log G_u(e^s)$ where $G_u(z) = \langle z^u \rangle_{\mathbb{P}(u)}$ is a generating function. From basic probability theory, we know that the generating function of the sum of a random number of iid variables is obtained by nesting as $G_x(z) = G_n(G_g(z))$, where

$$G_n(z) = \exp(\lambda(z - 1)) \quad G_g(z) = (1 - \log(z)/b)^{-a}$$

are generating functions for the Poisson and Gamma densities. By substitution we obtain the CGF of x as

$$K_x(s) = \lambda((1 - s/b)^{-a} - 1). \tag{14}$$

Now, we will show that we obtain the same CGF starting from the power variance assumption. We can easily verify that CGF for EDM in (7) is given by (Jørgensen, 1997; Dunn & Smyth, 2005)

$$K_x(s; \theta, \phi) = \frac{1}{\phi} \left(\kappa(s\phi + \theta) - \kappa(\theta) \right). \tag{15}$$

If we substitute the expression for $\kappa(\theta)$ in (11) and then express the result as a function of the expectation parameter $\hat{x}$ by noting that

$$\theta = \frac{\hat{x}^{1-p}}{1 - p} \tag{16}$$

(as $d\theta/d\hat{x} = v(\hat{x})^{-1} = \hat{x}^{-p}$), we obtain

$$K_x(s; \theta, \phi) = \frac{\hat{x}^{2-p}}{(2 - p)\phi} \left(\left(1 - s\phi(p - 1)\hat{x}^{p-1}\right)^{\frac{2-p}{1-p}} - 1 \right)$$

that has the same form as (14). By matching term by term, we see that the Tweedie distribution for $1 < p < 2$ is the compound Poisson distribution with the following parameter mapping:

$$\lambda = \frac{\hat{x}^{2-p}}{\phi(2 - p)}, \quad a = \frac{2 - p}{p - 1}, \quad b = \frac{\hat{x}^{1-p}}{\phi(p - 1)}. \tag{17}$$

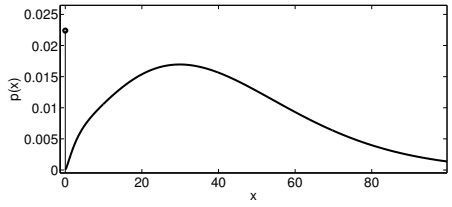


Figure 1. The compound Poisson distribution with $p = 1.3$, $\phi = 5$, and $\hat{x} = 40$. Note that the probability mass at zero makes this distribution suitable for sparse positive data.

By using this mapping, the joint distribution can be written as follows:

$$\begin{aligned}\mathbb{P}(x, n|\hat{x}, \phi, p) =& \mathbb{P}(x|n, \hat{x}, \phi, p)\mathbb{P}(n|\hat{x}, \phi, p) \\ =& \Big[\exp(-\frac{\hat{x}^{2-p}}{(2-p)\phi})\Big]^{[n=0]} \\ & \Big[\exp(-\frac{n}{p-1}\log(\phi) + n\frac{2-p}{p-1}\log\frac{x}{p-1} \\ & - n\log(2-p) - \log\Gamma(n+1) \\ & - \log\Gamma(\frac{2-p}{p-1}n) - \log(x) \\ & - \frac{1}{\phi}\left(\frac{\hat{x}^{1-p}x}{(p-1)} + \frac{\hat{x}^{2-p}}{(2-p)}\right))\Big]^{[n>0]}.\end{aligned} \tag{18}$$

It turns out that $\sum_n p(x, n|\cdot)$ does not have a closed form. Here, Dunn and Smyth provide numerical methods for approximate computation (Dunn & Smyth, 2005), but we propose here two simpler algorithms. An example pdf of a compound Poisson distribution is given in Figure 1.

4. Parameter Estimation

An interesting property of the joint distribution in (18) is that $\hat{x}$ and n are conditionally independent given the index parameter p and the dispersion ϕ, as the joint factorizes such that there are no cross terms that contain both $\hat{x}$ and n. Besides, the terms that depend on $\hat{x}$ are specified by the β-divergence. Therefore, any standard algorithm that minimizes the beta divergence can be used here.

When dealing with factorization models (i.e $\hat{x}$ is decomposed into some latent factors), we seek the best factorization whose form can vary depending on the application. If we consider the model that is defined in (6), maximum likelihood estimation of the factors under mixed cost functions can be achieved by iteratively applying the multiplicative update rules given in (Yılmaz et al., 2011). The update rule for the factor Z_1 can be written as follows:

$$Z_1 \leftarrow Z_1 \circ \frac{\sum_{\nu=1}^{2} \phi_\nu^{-1}\Delta_\nu(M_\nu \circ X_\nu \circ \hat{X}_\nu^{-p_\nu})}{\sum_{\nu=1}^{2} \phi_\nu^{-1}\Delta_\nu(M_\nu \circ \hat{X}_\nu^{1-p_\nu})} \tag{19}$$

where p_ν are the index parameters, ϕ_ν are the dispersion parameters, $A \circ B$ and $\frac{A}{B}$ denotes element-wise product and division of two matrices A and B, respectively. Here, $\Delta_\nu(\cdot)$ are functions that are defined as follows:

$$\Delta_1(A) = AZ_2^\top \tag{20}$$

$$\Delta_2(A) = AZ_3^\top \tag{21}$$

where $\top$ denotes the matrix transpose. Besides, M_ν is a binary matrix of size X_ν that have values of 1 (0) where X_ν is observed (missing).

When p_ν and ϕ_ν are not known beforehand, the inference problem gets complicated. In this study, we focus on estimating p_ν and ϕ_ν when $p_\nu \in (1, 2)$. Since p_ν and ϕ_ν are conditionally independent from the factors, given the mean parameter, our methods can be used in any matrix and tensor factorization model. Therefore, we stick to our vector notation where we define $x \equiv \mathbf{vec}(X_\nu)$, $\hat{x} \equiv \mathbf{vec}(\hat{X}_\nu)$, $m \equiv \mathbf{vec}(M_\nu)$, and ν denotes the observed matrix/tensor index for the case when we have multiple (most likely multimodal) observed matrices/tensors. Here, $\mathbf{vec}(\cdot)$ is the vectorization operator (i.e. the colon operator in Matlab).

In the next subsections, we present three novel inference methods for estimating the index parameter in Tweedie compound Poisson models. In the first and the second methods we follow a variational approach, where in the third method we integrate out the dispersion parameter and make inference on the marginal distribution.

4.1. Variational Approach

In this section, we present two variational methods, namely the Iterative Conditional Modes (ICM) and the Expectation-Maximization (EM) algorithms.

The ICM algorithm iteratively maximizes over the parameters n, ϕ, and p given x and $\hat{x}$. Even though the maximization over n is intractable, we can find the mode n^* by approximating the $\log\Gamma(\cdot)$ functions in (18) by using Stirling's approximation, as proposed in (Dunn & Smyth, 2005). The mode has the following analytical form:

$$n^*(i) = \frac{x(i)^{2-p}}{(2-p)\phi}. \tag{22}$$

Maximizing the dispersion parameter ϕ is straightforward, however, since the index parameter p and ϕ are closely related to the variance and may affect each other, it can be necessary to regularize ϕ in order to have a better estimate of p. It is easy to verify that the conjugate prior of the dispersion parameter is the inverse Gamma distribution. Therefore, here we assume an inverse Gamma prior on ϕ: $\phi \sim \mathcal{IG}(\phi; \alpha_\phi, \beta_\phi)$. The optimal dispersion, given the other parameters is as follows:

$$\phi^* = \frac{\left(\sum_i \frac{m(i)\hat{x}(i)^{1-p}x(i)}{(p-1)} + \frac{m(i)\hat{x}(i)^{2-p}}{(2-p)}\right) + \beta_\phi}{\frac{\sum_i m(i)n^*(i)}{p-1} + \alpha_\phi + 1}. \quad (23)$$

Surprisingly, none of the references we are aware of used this conjugate prior. In the next section we will use this property to analytically integrate out the dispersion parameter.

The last step of the ICM algorithm is to compute the maximization over p. Since the optimal p does not have an analytical solution, we consult numerical methods. As the domain of p is limited to $(1, 2)$, we run a simple line search procedure in order to estimate the index parameter p.

To sum up, at each iteration of the estimation algorithm, we first estimate the factors and compute the mean parameter $\hat{x}$. Then, we compute the parameters n^* and ϕ^* that are described above, and finally we compute the optimal index parameter p. This procedure is run until convergence.

The EM algorithm is quite similar to the ICM algorithm in algorithmic sense, where we merely replace n^* with the expectation $\langle n \rangle$ in (23). Unfortunately, computing this expectation is also intractable. Therefore, we use a numerical method that is similar to the one proposed in (Dunn & Smyth, 2005). By using the fact that the conditional distribution of n is unimodal, we approximate the expectation by numerically computing it around the mode which is defined in (22). The rest of the EM algorithm is the same as the ICM algorithm.

4.2. Integrating out the Dispersion Parameter

The dispersion parameter plays a key role when there are more than one observed tensor (see (19)). However, when we have only one observed tensor, the dispersion parameter does not contribute to the estimation of the factors in a factorization model as it cancels out in the multiplicative update rules.

In this section we integrate out the dispersion parameter ϕ and n and make inference on the marginal distribution. When assumed an inverse Gamma prior on ϕ, we obtain the following marginal distribution:

$$\begin{aligned}\mathbb{P}(x, n) = &\left[\exp\Big(\alpha_\phi(\log\beta_\phi - \log(\frac{\hat{x}^{2-p}}{2-p} + \beta_\phi)\Big)\right]^{[n=0]} \\ &\left[\exp\Big(n\frac{2-p}{p-1}\log\frac{x}{p-1} - n\log(2-p)\right. \\ &- \log\Gamma(n+1) - \log\Gamma(\frac{2-p}{p-1}n) - \log(x) \\ &- (\alpha_\phi + \frac{n}{p-1})\log(\beta_\phi + \frac{\hat{x}^{1-p}x}{(p-1)} + \frac{\hat{x}^{2-p}}{(2-p)}) \\ &+ \alpha_\phi\log\beta_\phi + \log\Gamma(\alpha_\phi + \frac{n}{p-1}) \\ &\left.- \log\Gamma(\alpha_\phi)\Big)\right]^{[n>0]}. \quad (24)\end{aligned}$$

In order to estimate the index parameter p, we also marginalize out n by using numerical methods. Finally, the optimal p is found by a line search algorithm, similar to ICM and EM.

5. Experiments

In order to evaluate our methods, we conduct experiments on both synthetic and real data. Due to space limitations, in this paper we only present the experiments that we conduct on real data. The other experiments can be found in `http://www.cmpe.boun.edu.tr/~umut/icml2013`.

5.1. Polyphonic Music Modeling

Along with the rapid development of computational power and statistical modeling techniques, factorization-based music modeling has become popular. This paradigm has been shown to be successful in many applications including polyphonic pitch transcription, source separation and audio restoration.

Recent studies suggest that, when designed properly, polyphonic pitch transcription methods with higher level musical models yield better transcription performance (Boulanger-Lewandowski et al., 2012). In this section, we present a tensor factorization model for symbolic musical data modeling. This model can be used as a side model for factorization-based audio models.

Symbolic music representation is similar to the sheet representation of music, where symbolic data contain high level musical information, such as note onset times, note durations, and the pitch of the notes that occur in a musical piece. Musical Instrument Digital Interface (MIDI) is one of the standards of symbolic

music representation.

One disadvantage of the symbolic representation is that it does not reflect the temporally varying characteristics of the musical notes. We have the information of the velocities at the note onsets, however we cannot obtain the damping structure that the notes naturally have. Therefore, in order to have a better representation, we quantize the time into time-frames and encode the musical information into a matrix $X \equiv \{X(n,t)\}$ where n is the note index and t is the time frame index. Here $X(n,t)$ simulates the time-varying velocity (volume) of note n during time frame t. For instance, if the note n is active at both the time-frame t and $t+1$, then the velocities have the following relation: $X(n,t+1) = \alpha X(n,t)$ where $0 < \alpha < 1$. This representation mimics the structure of an excitation matrix of the Nonnegative Matrix Factorization model for audio signals (Smaragdis & Brown, 2003).

By construction, only a couple of notes will be active at a given time frame t, therefore X will consist of mostly zeros and some positive values. We can observe that assuming a compound Poisson observation model is quite reasonable as the compound Poisson distribution has a nonnegative probability mass at 0 and a continuous density on positive values.

In this study, we use Nonnegative Matrix Factor Deconvolution (NMFD) model (Smaragdis, 2004) in order to model the modified symbolic musical data. Apart from using the benefits of the NMF model, this model is also capable of modeling the temporal information of the music. We can define the model as follows:

$$X(n,t) \approx \hat{X}(n,t) = \sum_{\tau,k} D(n,\tau,k) E(k,t-\tau) \quad (25)$$

where D is the dictionary tensor and E encapsulates the corresponding excitations.

Since we have only one observed tensor in this model, we can use all three of the inference methods that have been described. In order to evaluate our methods on modeling the symbolic data, we firstly erase some columns (time frames) of the data, then reconstruct the missing parts by using the NMFD model. This reconstruction problem is not trivial as entire time frames (columns of X) can be missing.

In our experiments we use the MIDI Aligned Piano Sounds (MAPS) database (Emiya et al., 2010). We use 10 excerpts from 5 different classical music pieces. After generating the X matrices from the symbolic data, we randomly erase some columns of the data which are going to be reconstructed later on. In order to obtain the reconstructed symbolic data, we simply combine

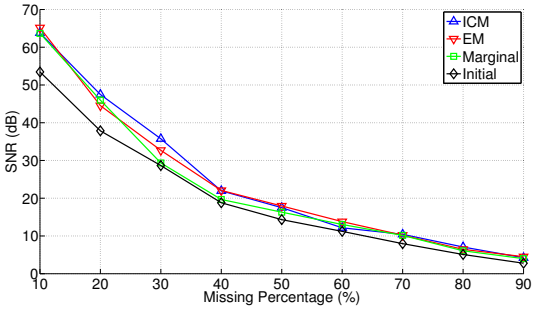
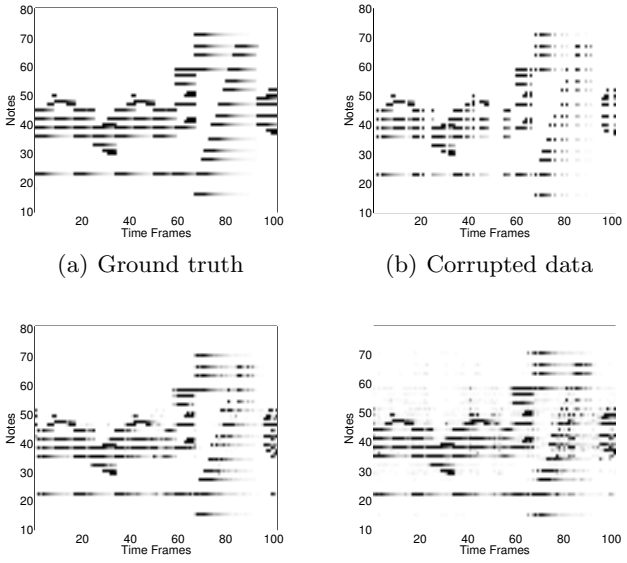


Figure 2. Results of the experiments. Initial SNR is computed by substituting 0 as missing values.



(a) Ground truth (b) Corrupted data

(c) Reconst. (compound Poisson) (d) Reconst. (Gaussian)

Figure 3. Visualization of the symbolic music reconstruction.

the observed parts of X and the estimated parts of $\hat{X}$: $M \circ X + (1-M) \circ \hat{X}$, where M is the binary mask that is introduced in Section 4. We evaluate and compare the performances of our methods by measuring the signal-to-noise ratio (SNR) between the corrupted and the reconstructed symbolic musical data.

In our experiment settings, the duration of the excerpts is 10 seconds, where we use time frames of 93 milliseconds. We select $\alpha_\phi = 5$ and $\beta_\phi = 3$, $|k| = 50$, and $|\tau| = 5$ for all methods, where $|\cdot|$ denotes cardinality. The results are shown in Figure 2.

The results suggest that, the methods always improve the quality of the corrupted symbolic data. The ICM and the EM algorithm give similar results, where the Bayesian method seems to be more sensitive to the missing data than the variational methods. The estimated index parameter p differs for each piece that is reconstructed. Besides, each algorithm finds different p values: the average values for the index parameter are 1.01 (ICM), 1.19 (EM), and 1.26 (Bayesian). For all methods, we get about 4 dB SNR improvement where 50% of the data is missing; gracefully degrading from 10% to 90% missing data. Figure 3 visualizes an example reconstruction. It can be observed that the compound Poisson model yields a better reconstruction, where the Gaussian model introduces spurious notes.

As the results are encouraging even when quite long portions of the data are missing, we can say that modeling the polyphonic music with this approach seems reasonable and might produce good results when used in more complicated models.

5.2. Coupled Audio and Lyrics Modeling

In this section, we illustrate how our approaches can be used with multimodal data. Coupled factorization models have been shown to be useful at fusing information from multimodal data (Şimşekli et al., 2012). Here, we illustrate how the index parameter p and the corresponding dispersion ϕ will be estimated under coupled models with mixed observation models where at least one of the observation model is the compound Poisson model.

We present a novel coupled matrix factorization model which combines audio features and the lyrics of songs. The aim of this application is to predict the bag-of-words representation of the lyrics of a song given its audio features. This is an interesting application which tries to estimate the keywords that should exist in the lyrics of a song by making use of its audio features and the information from other songs.

Suppose we observe the matrices $X_1 \equiv \{X_1(f,s)\}$ and $X_2 \equiv \{X_2(w,s)\}$, where X_1 contains the song-level audio features and X_2 contains the bag-of-words representation of the lyrics of the songs in their columns. Here, f denotes the audio feature index, s is the song index, w is the word index. We decompose these matrices by using the NMF model as follows:

$$X_1(f,s) \approx \hat{X}_1(f,s) = \sum_k D_1(f,k)E_1(k,s) \quad (26)$$

$$X_2(w,s) \approx \hat{X}_2(w,s) = \sum_n D_2(w,n)E_2(n,s) \quad (27)$$

where D_1 and D_2 are the dictionary matrices and E_1 and E_2 are the corresponding excitation matrices. By also assuming a low rank model over the excitation matrices, we hierarchically factorize the excitations by using another NMF model as follows:

$$E_1(k,s) = \sum_r B_1(k,r)C(r,s) \quad (28)$$

$$E_2(n,s) = \sum_r B_2(n,r)C(r,s), \quad (29)$$

where B_1 and B_2 are the dictionaries for the excitations. With a final assumption that a particular song would use the same columns of the dictionaries B_1 and B_2, we can say that it would have the same excitations. By this approach, we can relate the audio features to the lyrics. We define the ultimate coupled model as follows:

$$\hat{X}_1(f,s) = \sum_{k,r} D_1(f,k)B_1(k,r)C(r,s) \quad (30)$$

$$\hat{X}_2(w,s) = \sum_{n,r} D_2(w,n)B_2(n,r)C(r,s). \quad (31)$$

Figure 4 visualizes this model. Note that, an NMF-based approach is proposed for modeling lyrics in (Dikmen & Févotte, 2012) and the authors report successful results.

One can come up with many different applications by using this model; in this study, we focus on the prediction of the lyrics of a song in a bag-of-words representation. It is fairly easy to predict the lyrics of a particular song by using this model: we mark the related parts of the binary mask M_2 (see Section 4) as unobserved, then make predictions by using $\hat{X}_2$.

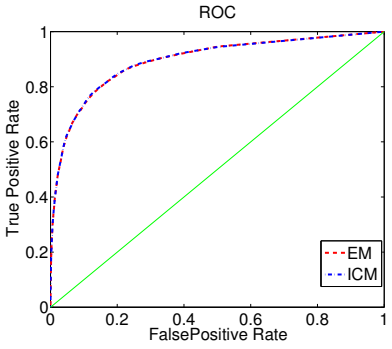


Figure 5. The ROC curve belonging to the word detection performance.

In our experiments we use the Million Song Dataset (MSD) and the MusiXmatch dataset (Bertin-Mahieux et al., 2011). The MSD is a free collection of audio

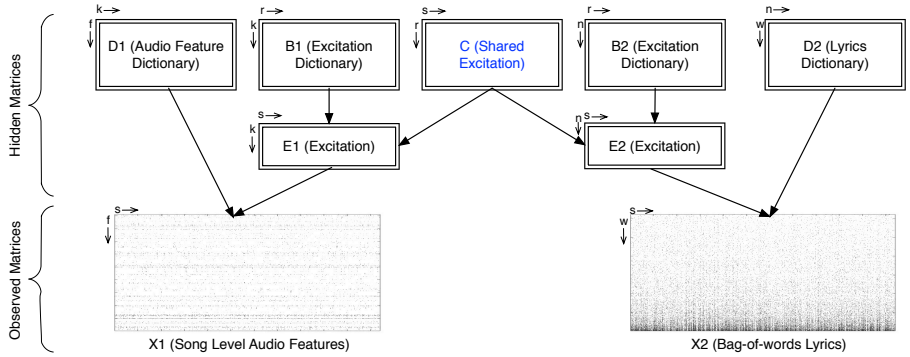


Figure 4. Visualization of the coupled factorization model. The blocks visualize the matrices and the relation between them. The lower-case letters and arrows near the blocks represent the indices of a particular matrix.

features and metadata that are gathered from a large number of music tracks. These features include the key, tempo, time signature, duration, genre tags, year, loudness, and the chroma features of the songs. We use the song level features of random 500 pop songs where we use 2827 features for each song, yielding an audio feature matrix X_1 of size 2827×500.

The MusiXmatch dataset contains the lyrics of the songs in a bag-of-words representation. This dataset contains more than 230 thousand songs, all being matched with the ones of MSD. Here, we use the number of occurrences of the most common 5000 words of each song, where these 5000 words cover over 92% of all the words in the dataset. We use the same songs that are selected while constructing X_1. Therefore, we have the lyrics matrix X_2 of size 5000×500, where each column of X_2 holds a bag-of-words lyrics of a song.

In our experiment settings, we select $p_1 = 1$ with unitary dispersion, which corresponds to the Poisson observation model. Note that, we could also optimize the dispersion ϕ_1, but this is out of the scope of this study. We set $|k| = |n| = 25$ and $|r| = 10$. In order to estimate the factors, we use the method that is presented in (Yılmaz et al., 2011). At each run, we estimate the factors, the index parameter p_2, and the dispersion ϕ_2. We predict the lyrics of random 10 songs at once and we repeat this process 5 times.

In order to assess the quality of the predictions, we measure the word detection performance. We estimate the predictions $\hat{X}_2$ and then consider the words as detected if the corresponding entries in $\hat{X}_2$ are above some threshold. We compute the true positive and the false positive rates as the performance metrics.

Figure 5 visualizes the results. It can be observed that both algorithms yield very similar results. We get more than 80% of true positive rate while keeping the false positive rate less than 20%. Besides, the ICM algorithm seems more advantageous since its computational requirements are much lower than the EM algorithm. These results are encouraging since the lyrics are predicted by solely using the song level audio features.

6. Conclusion

The compound Poisson distribution is a useful distribution for sparse data as it has a discrete probability mass at zero and a support for continuous positive data. In this study, we presented inference methods for estimating the index and the dispersion parameter of the Tweedie compound Poisson models. In the first two methods, we followed a variational approach, where in the third method we estimated the index parameter by using its marginal distribution. One of the contributions of this study is to make use the conjugate prior on the dispersion parameter, which has not been investigated in the literature yet.

We evaluated and compared our methods on real data. Firstly, we evaluated our methods on modeling symbolic representations for polyphonic music. Secondly, we defined a novel coupled tensor factorization model and evaluated our methods on prediction of the lyrics of a song from its audio features. Our conclusion is that the compound poisson based factorization models can be useful for sparse positive data.

Acknowledgments

Funded by TÜBİTAK grant number 110E292, project Bayesian matrix and tensor factorizations (BAYTEN). U. Ş. is also supported by a Ph.D. scholarship from TÜBİTAK.

References

Bar-Lev, S. K. and Enis, P. Reproducibility and natural exponential families with power variance functions. *Annals of Stat.*, 14, 1986.

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Boulanger-Lewandowski, Nicolas, Bengio, Yoshua, and Vincent, Pascal. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *International Conference on Machine Learning (ICML)*, 2012.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S. *Nonnegative Matrix and Tensor Factorization.* Wiley, 2009.

Şimşekli, U., Yılmaz, Y. K., and Cemgil, A. T. Score guided audio restoration via generalised coupled tensor factorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

Dikmen, Onur and Févotte, Cédric. Maximum marginal likelihood estimation for nonnegative dictionarylearning in the gamma-poisson models. *IEEE Transactions on Signal Processing*, 60(10):5163–5175, 2012.

Dunn, P. K. and Smyth, G. S. Series evaluation of tweedie exponential dispersion model densities. *Stats. & Comp.*, 15:267–280, 2005.

Emiya, V., Badeau, R, and David, B. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6): 1643–1654, 2010.

Févotte, C., Bertin, N., and Durrieu, J. L. Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Computation*, 21:793–830, 2009.

Jørgensen, B. *The Theory of Dispersion Models.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1997.

Lu, Zhiyun, Yang, Zhirong, and Oja, Erkki. Selecting β-divergence for nonnegative matrix factorization by score matching. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, volume 7553 of *Lecture Notes in Computer Science*, pp. 419–426, Lausanne, Switzerland, 2012. Springer.

McCulloch, C. E. and Nelder, J. A. *Generalized Linear Models.* Chapman and Hall, 2nd edition, 1989.

Smaragdis, P. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *ICA*, pp. 494–499, 2004.

Smaragdis, P. and Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

Tweedie, M. C. An index which distinguishes between some important exponential families. *Statistics: applications and new directions, Indian Statist. Inst., Calcutta*, pp. 579–604, 1984.

Yılmaz, Y. K. and Cemgil, A. T. Alpha/beta divergences and tweedie models. *arXiv:1209.4280 v1*, 2012.

Yılmaz, Y. K., Cemgil, A. T., and Şimşekli, U. Generalised coupled tensor factorisation. In *NIPS*, 2011.

Zhang, Yanwei. Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing*, accepted, 2012.