PROBABILISTIC LATENT TENSOR FACTORIZATION FRAMEWORK FOR AUDIO MODELING

Ali Taylan Cemgil[†]*, *Umut Şimşekli*[†], *Yusuf Cem Sübakan*[⋆]

Dept. of Computer Engineering[†] ,
Dept. of Electrical and Electronics Engineering[⋆],
Boğaziçi University,
34342 Bebek, İstanbul, Turkey
{taylan.cemgil,umut.simsekli,cem.subakan}@boun.edu.tr

ABSTRACT

This paper introduces probabilistic latent tensor factorization (PLTF) as a general framework for hierarchical modeling of audio. This framework combines practical aspects of graphical modeling of machine learning with tensor factorization models. Once a model is constructed in the PLTF framework, the estimation algorithm is immediately available. We illustrate our approach using several popular models such as NMF or NMF2D and provide extensions with simulation results on real data for key audio processing tasks such as restoration and source separation.

Index Terms— Audio Modeling, Probabilistic Latent Tensor Factorization, Factor graphs, Statistical Inference, Message Passing

1. INTRODUCTION

The last decade has witnessed a rapid development of statistical modeling techniques for various audio applications related to music information retrieval and content analysis, such as transcription or source separation.

A particularly useful modeling paradigm, leading to practical and useful algorithms has been based on matrix factorization. As a particular example, given an observed audio spectrogram X as a matrix of frequency and time indices f and t, one searches for a decomposition of form

$$X(f,t) \approx \hat{X}(f,t) = \sum_i D(f,i)E(i,t) \tag{1}$$

Typically, the goal is to find optimal matrices D^* and E^* such that

$$(D^*, E^*) = \arg\min_{D,E} d(X, \hat{X}) \tag{2}$$

where d is a divergence (a quasi-squared-distance) typically taken as Euclidian, Kullback-Leibler (KL) or Itakura-Saito (IS). The β-*divergence* generalizes all this divergences and enables a unified treatment [1, 2, 3]

$$d_\beta(x,y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right) & \beta \notin \{0,1\} \\ x(\log x - \log y) + (y - x) & \beta = 1 \\ x/y - \log(x/y) - 1 & \beta = 0 \end{cases} \tag{3}$$

Pioneering work on Nonnegative Matrix Factorization (NMF) for audio processing [4] has demonstrated that, provided that model order is properly chosen, the computed factors D and E tend to be semantically meaningful as they correlate well with the intuitive notion of spectral templates and a musical score. Following the various extensions and improvements have been proposed for transcription or source separation [2]. NMF and related extensions have also a natural interpretation as probabilistic generative models [5, 3].

This paper introduces probabilistic latent tensor factorization (PLTF) as a general framework for hierarchical modeling of audio. PLTF derives inspiration from two apparently independently developed tools, namely probabilistic graphical models of statistical machine learning [6] and tensor decompositions of multiway analysis [7]. The key motivation behind PLTF is that many useful models scattered in the audio and music processing literature can be expressed compactly using a tensor factorization and contraction (summing over a set of indices) formalism; we will give several examples later in the paper. In statistical machine learning literature, it is standard to represent a multivariate probability distribution as a product of local potential functions that describe interactions between random variables. A popular graphical representation for such objects is a *factor graph*; this is a bipartite graph of factor nodes (typically shown as black squares) and variable nodes (shown as white circles). Each factor node corresponds to a local function and each variable node corresponds to a random variable. The inference algorithm (e.g. for computing marginal distributions and moments) can be implemented as a message passing algorithm on the factor graph [6].

In PLTF, we represent a tensor model by a factor graph, where now factor tensors correspond to factor nodes and indices correspond to variable nodes. An index i is connected to a tensor node Z if it appears as an index of Z. One novel contribution of PLTF is that, once a model is represented in this form, the inference algorithm to estimate the tensor factorization can also be derived automatically from the factor graph specification. Note that, unlike in probabilistic graphical models, in PLTF, the factor graph does not represent a probability measure; only the algebraic representation is analogous. Yet, this analogy enables us to derive novel message passing algorithm. Perhaps more importantly, this gives a flexibility for building increasingly more complex hierarchical models easily without much extra effort; we believe that this is both of practical and theoretical interest to the audio processing community.

*Funded by the scientific and technological research council of Turkey (TÜBİTAK) grant number 110E292, project “Bayesian matrix and tensor factorizations (BAYTEN)”. U.Ş. is also supported by a Ph.D. scholarship from TÜBİTAK.

1.1. Probabilistic Latent Tensor Factorization

The latent tensor factorization model [8] is given as a natural extension of the matrix factorization model of (1)

$$X(v_0) \approx \hat{X}(v_0) = \sum_{\bar{v_0}} \prod_{\alpha} Z_\alpha(v_\alpha), \tag{4}$$

where our goal is computing an approximate factorization of a given a multiway array X in terms of a product of individual factors Z_α, some of which are possibly fixed. The product $\prod_\alpha Z_\alpha(v_\alpha)$ is collapsed over a set of indices, hence the factorization is latent. The optimization problem is again minimization of $d(X, \hat{X})$. Here, we define

V	set of all indices in a model
V_0	set of visible indices
V_α	set of indices in Z_α
$\bar{V_\alpha} = V - V_\alpha$	set of all indices not in Z_α

We use small letters as v_α to refer to a particular setting of indices in V_α. For example, in this framework, the NMF model of [9], introduced in (1) would be represented via the dictionary matrix $Z_1 \equiv D$, the excitations $Z_2 \equiv E$, and the index sets $V = \{i, j, k\}$, $V_0 = \{f, t\}$, $V_1 = \{f, i\}$, and $V_2 = \{i, t\}$. The factor graph corresponding to the NMF model is shown in Table 2.

1.2. Inference

The inference, i.e., estimation of the latent factors Z_α can be achieved via iterative optimization (see [8]). One can obtain the following compact fixed point equation where each Z_α is updated in an alternating fashion fixing the other factors $Z_{\alpha'}$ for $\alpha' \neq \alpha$

$$Z_\alpha \leftarrow Z_\alpha \circ \frac{\Delta_\alpha(M \circ X \circ \hat{X}^{\beta-2})}{\Delta_\alpha\left(M \circ \hat{X}^{\beta-1}\right)}, \tag{5}$$

where $\circ$ is the Hadamard product (element-wise product) and M is a $0-1$ mask array where $M(v_0) = 1$ ($M(v_0) = 0$) if $X(v_0)$ is observed (missing). In this iteration, the key quantity is the Δ_α function that is defined as

$$\Delta_\alpha(A)(v_\alpha) \equiv \sum_{\bar{v}_\alpha} \left(A(v_0) \prod_{\alpha' \neq \alpha} Z_{\alpha'}(v_{\alpha'}) \right). \tag{6}$$

For updating Z_α, we need to compute this function twice for arguments $A = M \circ X \circ \hat{X}^{\beta-2}$ and $A = M \circ \hat{X}^{\beta-1}$. As an example, it is easy to verify that the update equations for the KL-NMF problem (for $\beta = 1$) are obtained as a special case of (5). Further cases are summarized in Table 1. A key observation is that the Δ_α function is computing a product of tensors and collapses this product over indices not appearing in Z_α. Algebraically, this is equivalent to computing a marginal sum; a task for which several graph based algorithms exist.

It is also easy to regularize the model or incorporate prior knowledge (such as sparsity). For example, in the case of the KL divergence, we can choose a gamma prior model

$$Z_\alpha(v_\alpha) \sim \mathcal{G}(Z_\alpha(v_\alpha); A_\alpha(v_\alpha), B_\alpha(v_\alpha)/A_\alpha(v_\alpha))$$

Table 1: Update rules for different β values

β	Cost Function	Multiplicative Update Rule
0	Itakura-Saito	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\Delta_\alpha(M \circ X/\hat{X}^2)}{\Delta_\alpha\left(M/\hat{X}\right)}$
1	Kullback-Leibler	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\Delta_\alpha(M \circ X/\hat{X})}{\Delta_\alpha(M)}$
2	Euclidean	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\Delta_\alpha(M \circ X)}{\Delta_\alpha\left(M \circ \hat{X}\right)}$

where $\mathcal{G}$ denotes the gamma distribution $\mathcal{G}(x; a, b) = b^a x^{a-1} \exp(-bx)/\Gamma(a)$. In this case, the update equation is slightly altered and becomes

$$Z_\alpha \leftarrow \frac{(A_\alpha - 1) + Z_\alpha \circ \Delta_\alpha(M \circ X/\hat{X})}{A_\alpha/B_\alpha + \Delta_\alpha\,(M)} \tag{7}$$

For the general case of the β divergence, the choice of priors are more delicate [10], which we omit from this publication.

2. HIERARCHICAL FACTORIZATIONS FOR AUDIO

The NMF model has obvious limitations due to unrealistic modelling assumptions; spectral template components at each frequency bin are weighted with the same coefficient. To capture richer temporal variations observed in real audio signals, in [11], Smaragdis introduced the *non-negative matrix factor deconvolution* (NMFD) that is defined by

$$\begin{aligned} \hat{X}(f,t) &= \sum_{\tau,i} D(f,\tau,i) E(i, \overbrace{t-\tau}^{d}). \\ &= \sum_{\tau,i,d} D(f,\tau,i) E(i,d) Z(d,t,\tau) \end{aligned} \tag{8}$$

Here, we have introduced a new dummy index d and define a new factor $Z(d,t,\tau) = \delta(d - t + \tau)$ to express this model in our framework. Here, Z is a constant factor not to be updated during the iterations. Again, the update equations are immediately available from (5). For example, for KL cost, after straightforward simplifications, one obtains the Δ functions required for the updates

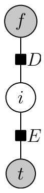
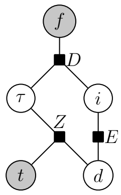
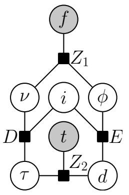
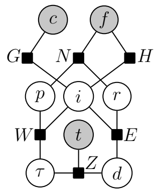
$$\Delta_D(A)(f,\tau,i) = \sum_t A(f,t) E(i, t-\tau) \tag{9}$$

$$\Delta_E(A)(i,d) = \sum_{f,t} A(f,t) D(f, t-d, i) \tag{10}$$

where each function need to be computed for $A = M \circ X/\hat{X}$ and $A = M$. These are convolutions, hence computation can be further accelerated via FFT.

The convolutive model has been further extended by Schmidt and Mørup [12] as the Non-negative Matrix Factor 2D Deconvolution (NMF2D) to factorize a log-frequency spectrogram (constant-Q) using a model that can represent both temporal structure and the pitch changes when an instrument plays different notes. The key idea of this elegant model is that on log-frequency index, modulations correspond to shifts. We can reformulate the model in the

Table 2: Models, index sets and factor graphs. For NMF, NMFD, NMF2D, D, E denote the dictionary and the excitations; for SF-SSNTF G are gains of sources, H is a filter, N is a harmonic dictionary and W are harmonic weights. Gray shaded nodes are visible indices. In all models f, t correspond to frequency and time frame, In NMF* models, i is the template index and ν, τ are the 'local' frequency and time indices of spectral templates. In SF-SSNTF, i, p, r, c correspond to instrument, harmonic, note label and channel indices.

	Symbol	NMF	NMFD	NMF2D	SF-SSNTF
Model	V	$\{f,t,i\}$	$\{f,t,\tau,i,d\}$	$\{f,t,\nu,\tau,i,\phi,d\}$	$\{c,t,f,i,p,r,\tau,d\}$
Observed	V_0	$\{f,t\}$	$\{f,t\}$	$\{f,t\}$	$\{c,t,f\}$
Latent	$\bar{V}_0$	$\{i\}$	$\{\tau,i,d\}$	$\{\nu,\tau,i,\phi,d\}$	$\{i,p,r,\tau,d\}$
Factors		$\{f,i\}$ $\{i,t\}$	$\{f,\tau,i\}$ $\{d,i\}$ $\{d,t,\tau\}$	$\{\nu,\tau,i\}$ $\{\phi,d,i\}$ $\{\nu,f,\phi\}$ $\{d,t,\tau\}$	$\{c,i\}$ $\{f,i\}$ $\{f,p,r\}$ $\{p,i,\tau\}$ $\{r,i,d\}$ $\{d,t,\tau\}$
					 c f G N H p i r W t E Z τ d

PLTF framework as

$$\hat{X}(f,t) = \sum_{i,\phi,\tau} D(\overbrace{f-\phi}^{\nu},\tau,i)E(\phi,\overbrace{t-\tau}^{d},i) \qquad (11)$$

$$= \sum_{i,\phi,\tau,\nu,d} D(\nu,\tau,i)E(\phi,d,i)Z_1(\nu,f,\phi)Z_2(d,t,\tau)$$

here $Z_1 = \delta(\nu - f + \phi)$ and $Z_2 = \delta(d - t + \tau)$ are fixed. We don't derive explicitly the update equations here; these follow again directly from (5). Both models are shown in Table 2.

A related model, proposed by [13], FitzGerald et al. is the *Source-Filter Sinusoidal Shifted Nonnegative Tensor Factorization Model* (SF-SSNTF). A model in the same spirit is also proposed in [14] by Klapuri et al. The model mimics physically inspired source-filter models of audio production in the spectral domain, such as a harmonic excitation multiplied by spectral envelope of a body response filter and is defined by

$$\hat{X}(c,t,f) = \sum_{i,p,r,\tau} G(c,i)H(f,i)N(f,p,r)W(p,i,\tau)E(r,i,\overbrace{t-\tau}^{d}) \qquad (12)$$

where G is the gain of each channel, H is the formant filter, N is the harmonic dictionary, W is the harmonic weight tensor, and E is the excitation tensor. This model is fairly complex to describe as it contains both convolutive and hierarchical elements; a derivation and implementation from scratch is also not straightforward. Again, by defining a dummy index d and setting $Z = \delta(d - t + \tau)$ we obtain the rightmost model given in Table 2, for which the update equations are directly available in our framework.

2.1. Extensions

In this section, based on the PLTF framework, we will propose extensions to the models introduced in the previous section. These concentrate mainly on modeling spectral templates hence focus on the dictionary but don't exploit temporal continuity or sparsity. For example, in two dimensional non-negative factor deconvolution model, we wish to interpret the excitation tensor $E(\phi, d, i)$ as a piano-roll like representation, where a large value indicates the presence of note ϕ at time d, played by the i'th source. Hence, it seems more natural to model the elements of this tensor to reflect statistical properties of piano rolls. For NMF models, temporal continuity and smoothness can be enforced via Markovian priors such as Gamma chains [15] or changepoint models [16]. Such hierarchical models also capture sparsity and continuity, but the inference schemata can become fairly complicated to describe; here we develop two related approaches that fit directly to the PLTF framework. The decompositions are:

$$E(\phi,d,i) = \sum_{k,l} B(k,l)C(k,d-l,\phi,i) \qquad \text{(I)} \qquad (13)$$

$$E(\phi,d,i) = \sum_{k,l} B(d,k)C(k,\phi,i) \qquad \text{(II)} \qquad (14)$$

The first approach (I) is in the spirit of convolutive models, where we decompose the excitations as shifted and scaled versions of vectors from a predetermined excitation dictionary B where $B(k,l)$ denotes the l'th element of k'th basis vector. Here, $C(k,u,\phi,i)$ is a tensor which dictates where the continuous basis functions will be replicated in time. Note that for each note ϕ and source i, we convolve two sequences to have the corresponding excitation vector in time but the catalog is shared, reducing significantly the number of free parameters. The second decomposition (II) is a simpler and is based on a basis spline approach. Here we use a dictionary B where for each k, the basis vector $B(k,:)$ has the shape of a locally concentrated triangle: by superposition of these basis vectors we can model piecewise linear functions with knot points located at triangle centers. All the extended models and the corresponding factor graphs are shown in Figure1. In the next section, we will illustrate the performance of our models on two audio processing tasks, namely restoration and source separation.

3. RESULTS AND CONCLUSIONS

In our restoration experiments, we used a database of 50 short mono audio examples sampled at 44.1kHz used in [13] (available online).

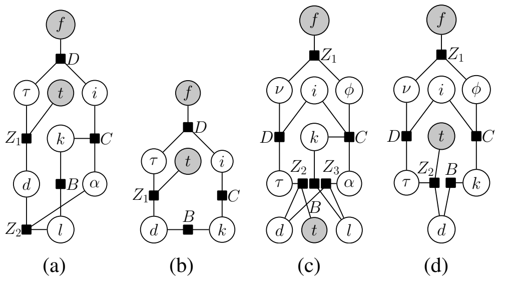


Figure 1: Graphical representations of the extended models. a) NMFD+I b) NMFD+II c) NMF2D+I d) NMF2D+II

Table 3: Evaluation of the models on missing audio restoration

	SNR			MSE		
	IS	KL	EUC	IS	KL	EUC
NMFD	2.99	4.74	5.05	4.43	2.91	2.68
SF-SSNTF	−0.28	5.09	5.06	15.00	2.57	2.59
NMFD + I	3.01	6.00	6.91	5.89	2.23	1.68
NMFD + II	5.00	5.79	5.80	2.74	2.20	2.17

For each example, we compute a spectrogram with framelength of 1024 samples with no overlap. Then, we remove randomly blocks of 10 consecutive time frames, corresponding to approx. 250ms gaps. In total, 20 per cent of each audio file was removed but the gaps are quite long. We compute the signal-to-noise ratio (SNR) and the mean squared error (MSE) using the true and predicted magnitude spectrogram coefficients. The performances of the models on restoration are given in Table 3, we see that our extensions are effective. For source separation experiment we use the same database, where we simply sum pairs of examples. For each mixture, we compute a constant-Q-transform and iteratively estimate the sources. For performance evaluation, we compute the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR). The results are shown in Table 4. In this case, the extensions seem to be somewhat less effective. The detailed derivations and the evaluation results are available on `http://www.cmpe.boun.edu.tr/~umut/pltf_audio`

3.1. Conclusion and Future Work

We have introduced PLTF as a general framework for hierarchical modeling of audio. PLTF combines practical aspects of graphical modelling such as ease of model construction and systematic development of an inference algorithm. The approach is particularly handy for the treatment of complicated tensor factorisation models. We haven't investigated Bayesian techniques for incorporating conjugate priors for regularization, as well as model selection and comparison issues, i.e., questions regarding the cardinality of latent indicies (such as choosing the number of spectral templates, the size of the catalog etc.) or comparing between two alternative TF models. As the models get increasingly more complicated, model selection and/or regularisation issues become central and we perceive the need for a full Bayesian treatment. Fortunately, these computations can also be carried out in a mechanical fashion and this is our current active research. Other technical issues are automatic inference code generation from a model specification and parallelization.

Table 4: Evaluation of models on blind source separation

Model	SDR	SIR	SAR
NMF2D	6.10	19.00	7.50
SF-SSNTF	≈ 8.00	≈ 24.00	≈ 8.00
NMF2D + I	6.19	19.84	6.84

4. REFERENCES

[1] A. Cichoki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorization*. Wiley, 2009.

[2] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[3] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *EUSIPCO*, 2009.

[4] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA*, 2003, pp. 177–180.

[5] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, 2009.

[6] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm." *IEEE Transactions on Information Theory*, vol. 47, pp. 498–519, 2001.

[7] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, pp. 455–500, 2009.

[8] Y. K. Yilmaz and A. T. Cemgil, "Probabilistic latent tensor factorization," in *LVA/ICA*, 2010, pp. 346–353.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–791, 1999.

[10] Y. K. Yilmaz and A. T. Cemgil, "Algorithms for probabilistic latent tensor factorisation with beta divergence," *Submitted to Signal Processing*, 2011.

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

[12] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *ICA*, 2006.

[13] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended non-negative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[14] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and EM algorithm," in *ICASSP*, 2010.

[15] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *ICASSP*, 2008, pp. 1825–1828.

[16] U. Şimşekli and A. T. Cemgil, "Probabilistic models for real-time acoustic event detection with application to pitch tracking," *JNMR*, 2011 (to appear).