MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Hierarchical and Coupled Non-negative Dynamical Systems with Application to Audio Modeling

Simsekli, U.; Le Roux, J.; Hershey, J.R.

TR2013-097 October 2013

Abstract

Many kinds of non-negative data, such as power spectra and count data, have been modeled using non-negative matrix factorization. Even though this modeling paradigm has yielded successful applications, it falls short when the data have certain hierarchical and temporal structure. In this study, we propose a novel dynamical system model that can handle these kinds of complex structures that often arise in non-negative data. We show that our model can be extended to handle heterogeneous data for data-driven regularization. We present convergence-guaranteed update rules for each latent factor. In order to assess the performance, we evaluate our model on the transcription of classical piano pieces, and show that it outperforms related models. We also illustrate that the performance can be further improved by making use of symbolic data.

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

HIERARCHICAL AND COUPLED NON-NEGATIVE DYNAMICAL SYSTEMS WITH APPLICATION TO AUDIO MODELING

Umut Şimşekli,[1] *Jonathan Le Roux,*[2] *John R. Hershey*[2]

[1]Boğaziçi University, Dept. of Computer Engineering, 34342, Bebek, İstanbul, Turkey
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
umut.simsekli@boun.edu.tr, {leroux, hershey}@merl.com

ABSTRACT

Many kinds of non-negative data, such as power spectra and count data, have been modeled using non-negative matrix factorization. Even though this modeling paradigm has yielded successful applications, it falls short when the data have certain hierarchical and temporal structure. In this study, we propose a novel dynamical system model that can handle these kinds of complex structures that often arise in non-negative data. We show that our model can be extended to handle heterogeneous data for data-driven regularization. We present convergence-guaranteed update rules for each latent factor. In order to assess the performance, we evaluate our model on the transcription of classical piano pieces, and show that it outperforms related models. We also illustrate that the performance can be further improved by making use of symbolic data.

Index Terms— Non-negative matrix factorization, Linear dynamical systems, Coupled factorization, Audio modeling

1. INTRODUCTION

Non-negative data arise in a variety of important signal processing domains, such as power spectra of signals, pixels in images, and count data. Non-negative matrix factorization (NMF) and its variants have seen a wide variety of applications to non-negative data analysis in recent years. Given the data matrix $\mathbf{V}$, the goal is to compute a decomposition of the form $\mathbf{V} \approx \mathbf{W}\mathbf{U}$ where $\mathbf{V} \equiv \{v_{fn}\}$, $\mathbf{W} \equiv \{w_{fk}\}$, and $\mathbf{U} \equiv \{u_{kn}\}$ are non-negative matrices of size $F \times N$, $F \times K$, and $K \times N$, respectively [1]. One of the popular approaches for estimating the factors $\mathbf{W}$ and $\mathbf{U}$ is minimizing a cost function between $\mathbf{V}$ and $\mathbf{W}\mathbf{U}$:

$$(\mathbf{W},\mathbf{U})^{\star} = \arg\min_{\mathbf{W},\mathbf{U}} d(\mathbf{V}||\mathbf{W}\mathbf{U}), \qquad (1)$$

where $d(\cdot)$ is a suitable cost function that is usually selected as Euclidean, Kullback-Leibler (KL), or Itakura-Saito (IS) divergences.

NMF-based modeling has been shown to be useful in various domains, including signal processing, finance, bioinformatics, and natural language processing. Even though this modeling paradigm has yielded successful applications, it does not model the hierarchical and dynamical structures that often arise in non-negative data. In this study, we present a novel dynamical system that aims to handle these complex structures. The proposed model is applicable to any kind of non-negative data with temporal and hierarchical properties.

There have been various studies on such structured data revolving around NMF, mostly applied to audio since audio signals have prominent temporal and hierarchical structure. When applied to audio signals, NMF decomposes the magnitude or power spectra $\mathbf{V}$ into a dictionary of spectral templates $\mathbf{W}$ and their corresponding activations $\mathbf{U}$. One property of audio signals is that certain spectral templates tend to be active simultaneously. This property requires the columns of $\mathbf{U}$ to have certain structure where the NMF model falls short. In order to enforce this structure to the NMF model, Lefèvre et al. [2] introduced a penalty term on the activations that favors sparsity at the group level. Grais and Erdoğan [3] proposed a method that regularizes the columns of $\mathbf{U}$ by using Gaussian mixture models. Şimşekli and Cemgil [4] presented a coupled factorization model where they hierarchically decomposed the activation matrix $\mathbf{U}$ into basis and weight matrices where the basis matrix aims to capture the hierarchical structure of the spectral dictionary.

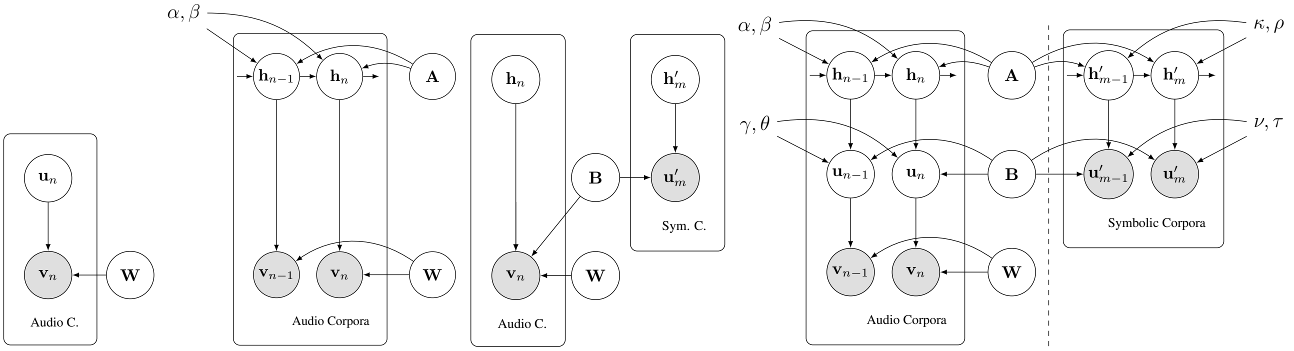
Another drawback of the basic NMF model is that it assumes that the activations at a given time (a column of $\mathbf{U}$) are independent of the activations at any other time, which conflicts with many time-series data including audio signals. To overcome this problem, Smaragdis proposed a convolutive factorization model [7] where the spectral templates encapsulate temporal information in addition to the spectral information. Mysore et al. [8] modeled the temporal structure by using a hidden Markov model (HMM) where each state of the Markov chain has a corresponding spectral dictionary. Ozerov et al.'s factorial scaled hidden Markov model [9] and Nakano et al.'s NMF with Markov-chained bases [10] intend to represent time-varying spectra as state transitions through a limited and fixed number of spectral patterns.

Rather than grafting NMF onto disparate models such as discrete HMMs, a number of approaches have introduced dynamics in a way that is more consistent with the NMF approach. Dikmen and Cemgil [11] related the columns of the activation matrix $\mathbf{U}$ by using gamma Markov chains. Févotte proposed a similar model in [12].

This basic approach was generalized in Févotte et al. [6] to a new non-negative dynamical system model (NDS) that employs a full transition matrix to represent statistical dependencies over time. This model shares the benefit of an HMM in that it can represent complex dynamics, but also has the advantage of having a continuous hidden state that can gradually evolve and adapt to changes in gain. One limitation of this model is that it is difficult to independently control the sparsity of the representation and the sparsity of the dynamics.

In order to address both temporal dependencies and dependencies among activations at a given time, we introduce a new generative model based on NDS. The new model introduces an additional layer of hidden variables between the model of temporal dynam-

The research presented in this paper was conducted while Umut Şimşekli was an intern at MERL. The authors thank Dr. Cédric Févotte for fruitful discussions.



(a) NMF (b) NDS (c) CMF (d) HNDS and CHNDS

Figure 1: The graphical models for (a) the original NMF model [5], (b) non-negative dynamical system (NDS) [6], (c) coupled matrix factorization (CMF) model [4], and (d) hierarchical NDS (HNDS) and coupled hierarchical NDS (CHNDS). The nodes represent the random variables, the arrows determine the conditional independence structure, and the shaded the nodes represent the observed variables. The left part of (d) corresponds to HNDS, defined in Eq. 2-4.

ics and the observations. This is analogous to the use of GMMs in a discrete HMM with continuous observations, where the mixture components can be seen as a layer of additional latent variables. Such a hierarchical model can represent the statistical dependencies of the activations (different rows of $\mathbf{U}$) for a given state of the dynamics. Moreover, in the proposed model, each layer has its own parameters to control sparsity, so that the sparsity of the dynamics can be controlled independently of the sparsity of the rest of the model.

We also introduce an extension of the model that enables the use of heterogeneous data in the manner of [4]. This extension allows the model to learn from data sources with information pertaining to different levels of the model. This type of transfer learning has been shown to improve the performance of multi-level models [4].

The model was evaluated on a polyphonic transcription task consisting of acoustic power spectra of classical piano pieces. The proposed model outperforms previous models in terms of detection of the piano-roll note values, as quantified by the F-measure. In addition, we introduce transfer learning using the symbolic piano-roll representation of different piano pieces, and show that performance is further improved. The rest of the paper makes use of the following notation: bold capital letters denote matrices (e.g., $\mathbf{A}$), $\mathbf{a}_j$ denotes the j^{th} column of $\mathbf{A}$, and a_{ij} denotes a single entry of $\mathbf{A}$.

2. THE MODEL

We define a novel probabilistic model referred to as hierarchical non-negative dynamical system (HNDS):

$$h_{in} = \Big(\sum_j a_{ij} h_{j(n-1)}\Big)\epsilon^h_{in}, \qquad \epsilon^h_{in} \sim \mathcal{G}(\epsilon^h_{in}; \alpha_i, \beta_i) \qquad (2)$$

$$u_{kn} = \Big(\sum_i b_{ki} h_{in}\Big)\epsilon^u_{kn}, \qquad \epsilon^u_{kn} \sim \mathcal{G}(\epsilon^u_{kn}; \gamma_k, \theta_k) \qquad (3)$$

$$v_{fn} = \Big(\sum_k w_{fk} u_{kn}\Big)\epsilon^v_{fn}, \qquad \epsilon^v_{fn} \sim \mathcal{E}(\epsilon^v_{fn}; 1) \qquad (4)$$

where $\mathcal{G}$ and $\mathcal{E}$ denote the gamma and exponential distributions, respectively, and $\mathcal{G}$ is defined using shape and inverse scale parameters. The columns of $\mathbf{H}$ form a Markov chain which represents all temporal dependency via the transition matrix $\mathbf{A} \in \mathbb{R}^{I \times I}$ and the random innovations ϵ^h_{in}. In NDS, the variables $\mathbf{H}$ are directly used as activations of the basis functions $\mathbf{W}$. Here, the intermediate variables $\mathbf{U}$ serve as the activations and are expressed as a random linear transformation of $\mathbf{H}$, the distribution of which is governed by the innovations ϵ^u_{kn}. We assume independent scale-invariant Jeffreys prior on $\mathbf{h}_1$: $p(h_{i1}) \propto \frac{1}{h_{i1}}$, and for further regularization and numerical stability, we assume exponential priors on $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{W}$: $a_{ij} \sim \mathcal{E}(a_{ij}; \lambda_A),\ b_{ki} \sim \mathcal{E}(b_{ki}; \lambda_B),\ w_{fk} \sim \mathcal{E}(w_{fk}; \lambda_W)$. For simplicity, we constrain the innovations ϵ to have mean 1 by taking $\alpha_i = \beta_i$, $\gamma_k = \theta_k$. The graphical models for the proposed HNDS model and related models are given in Fig. 1.

Note that the innovations ϵ^u_{kn} control the strength of the dynamics: as their variance increases the model gradually reduces to NMF. If we neglect the innovations ϵ^h_{in} and ϵ^u_{kn} (set them to 1), whereas NDS reduces to the NMF-like factorization $\mathbf{V} \approx \mathbf{WH}$, HNDS reduces to the factorization $\mathbf{V} \approx \mathbf{WBH}$, where in both cases $\mathbf{H}$ is constrained by the Markovian dynamics.

In audio terms, v_{fn} denotes the observed power spectra, w_{fk} denotes the spectral dictionary, and u_{kn} denotes the activations. The indices f, n, and k denote the frequency bins, the time frames, and the spectral templates, respectively. For audio, we use multiplicative exponential noise on the observations, which is equivalent to choosing the IS divergence as the cost function in Eq. 1 [5].

For the particular case of polyphonic music modeling, we can suppose that each column of $\mathbf{W}$ represents the spectral information of a single note, $\mathbf{B}$ captures the chord structure of a given piece as its columns model the different combinations of the notes, $\mathbf{H}$ determines which chords are active at a given time frame, and $\mathbf{A}$ captures the chord progressions. The innovation term ϵ^u_{kn} imparts the HNDS model with the ability to handle variations in the relative strength of the notes within each chord, such as in the case of different voicings of the same chord.

In addition to the basic model, we propose an extension to do transfer learning using heterogeneous data. Recent studies suggest that providing additional sources of information to audio models can increase the performance on different tasks [4, 13]. Such a transfer-learning paradigm is compelling for music signals because large amounts of symbolic music data are available. Symbolic music data in the form of $\mathbf{U}' \equiv \{u'_{km}\}$ encodes whether the note k is active at time frame m or not. If $\mathbf{V}$ and $\mathbf{U}'$ have similar harmonic properties like tonality, it is reasonable to assume that $\mathbf{U}'$

Table 1: Update rules for $\mathbf{U}$ and $\mathbf{H}$. The factors can be updated at each iteration to the value $\frac{\sqrt{b^2-4ac}-b}{2a}$ where each factor has different a, b, and c values. The update rules for $\mathbf{H}'$ are identical to those of $\mathbf{H}$ up to replacing the parameters α_i, β_i, γ_k, and θ_k with κ_i, ρ_i, ν_k, and τ_k.

	a	b	c
u_{kn}	$\sum_f \frac{w_{fk}}{\hat{v}_{fn}} + \frac{\theta_k}{\hat{u}_{kn}}$	$1-\gamma_k$	$-u_{kn}^2 \sum_f \frac{v_{fn}}{\hat{v}_{fn}^2} w_{fk}$
h_{in} $(n=1)$	$\sum_j \frac{\alpha_j}{\hat{h}_{j(n+1)}} a_{ji} + \sum_k \frac{\gamma_k}{\hat{u}_{kn}} b_{ki}$	1	$-\hat{h}_{in}^2 \Big[\sum_k \frac{u_{kn}}{\hat{u}_{kn}^2} \theta_k b_{ki} + \sum_j \frac{h_{j(n+1)}\beta_j}{\hat{h}_{j(n+1)}^2} a_{ji}\Big]$
h_{in} $(1<n<N)$	$\sum_k \frac{\gamma_k}{\hat{u}_{kn}} b_{ki} + \frac{\beta_i}{\hat{h}_{in}} + \sum_j \frac{\alpha_j}{\hat{h}_{j(n+1)}} a_{ji}$	$1-\alpha_i$	$-\hat{h}_{in}^2 \Big[\sum_k \frac{u_{kn}}{\hat{u}_{kn}^2} \theta_k b_{ki} + \sum_j \frac{h_{j(n+1)}\beta_j}{\hat{h}_{j(n+1)}^2} a_{ji}\Big]$
h_{in} $(n=N)$	$\sum_k \frac{\gamma_k}{\hat{u}_{kn}} b_{ki} + \frac{\beta_i}{\hat{h}_{in}}$	$1-\alpha_i$	$-\hat{h}_{in}^2 \Big[\sum_k \frac{u_{kn}}{\hat{u}_{kn}^2} \theta_k b_{ki}\Big]$

has a similar underlying probabilistic model where it shares the so-called 'chord dictionary' $\mathbf{B}$ and 'chord transition' matrix $\mathbf{A}$ with $\mathbf{U}$. Hence we extend HNDS to the coupled hierarchical non-negative dynamical system (CHNDS) by introducing the following terms:

$$h'_{im} = \Big(\sum_j a_{ij} h'_{j(m-1)}\Big)\epsilon^{h'}_{im}, \qquad \epsilon^{h'}_{im} \sim \mathcal{G}(\epsilon^{h'}_{im}; \kappa_i, \rho_i) \quad (5)$$

$$u'_{km} = \Big(\sum_i b_{ki} h'_{im}\Big)\epsilon^{u'}_{km}, \qquad \epsilon^{u'}_{km} \sim \mathcal{G}(\epsilon^{u'}_{km}; \nu_k, \tau_k) \quad (6)$$

where $\mathbf{H}' \in \mathbb{R}^{I\times M}$ denotes the chord activations for $\mathbf{U}'$. Note that the symbolic music $\mathbf{U}'$ does not necessarily belong to the same piece nor does it have the same number of time frames as $\mathbf{V}$. It is used as side information that can contribute to the estimation of the latent factors. As before, we set $\kappa_i = \rho_i$, and $\nu_k = \tau_k$. CHNDS reduces to the model in [4] if we discard the Markovian prior on $\mathbf{H}$ and the noise terms ϵ^h_{in}, $\epsilon^{h'}_{im}$, ϵ^u_{kn}, and $\epsilon^{u'}_{km}$.

3. INFERENCE

In this section, we present convergence-guaranteed update rules for maximum a-posteriori (MAP) estimation in the proposed model. In particular, we use the majorization-minimization (MM) algorithm [12] which monotonically decreases the intractable MAP objective function by minimizing a tractable upper-bound constructed at each iteration. This algorithm is a block-coordinate descent algorithm which alternatively updates each latent factor given its current value and the other factors. For more details, the reader is referred to [12].

The MM algorithm yields multiplicative update rules for $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{B}$. The update rule for $\mathbf{W}$ is given as follows:

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_n \frac{v_{fn}}{\hat{v}_{fn}^2} u_{kn}}{\sum_n \frac{u_{kn}}{\hat{v}_{fn}} + \lambda_W}}, \quad (7)$$

where $\hat{v}_{fn} = \sum_k w_{fk} u_{kn}$. The update rule for $\mathbf{A}$ contains terms that come from both the audio and symbolic data models:

$$a_{ij} \leftarrow a_{ij} \sqrt{\frac{\beta_i \sum_{n=2}^{N} \frac{h_{in} h_{j(n-1)}}{\hat{h}_{in}^2} + \rho_i \sum_{m=2}^{M} \frac{h'_{im} h'_{j(m-1)}}{\hat{h}'^2_{im}}}{\alpha_i \sum_{n=2}^{N} \frac{h_{j(n-1)}}{\hat{h}_{in}} + \kappa_i \sum_{m=2}^{M} \frac{h'_{j(m-1)}}{\hat{h}'_{im}} + \lambda_A}} \quad (8)$$

where $\hat{h}_{in} = \sum_j a_{ij} h_{j(n-1)}$ and $\hat{h}'_{im} = \sum_j a_{ij} h'_{j(m-1)}$. The update rule for $\mathbf{B}$ is as follows:

$$b_{ki} \leftarrow b_{ki} \sqrt{\frac{\theta_k \sum_n \frac{u_{kn}}{\hat{u}_{kn}^2} h_{in} + \tau_k \sum_m \frac{u'_{km}}{\hat{u}'^2_{km}} h'_{im}}{\gamma_k \sum_n \frac{h_{in}}{\hat{u}_{kn}} + \nu_k \sum_m \frac{h'_{im}}{\hat{u}'_{km}} + \lambda_B}} \quad (9)$$

where $\hat{u}_{kn} = \sum_i b_{ki} h_{in}$ and $\hat{u}'_{km} = \sum_i b_{ki} h'_{im}$. Note that the update equations for the HNDS model can be obtained by setting $\kappa_i = \rho_i = \nu_k = \tau_k = 0$, $\forall i, k$, which is equivalent to having infinite noise variance on the symbolic data.

The update equations of $\mathbf{U}$, $\mathbf{H}$, and $\mathbf{H}'$ involve finding roots of second order polynomials. The corresponding update rules are given in Table 1.

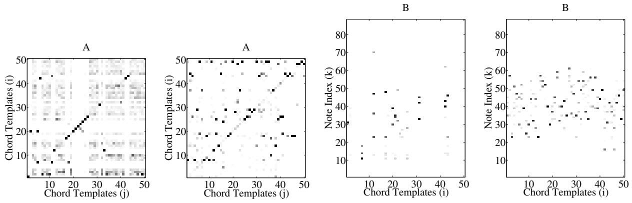
4. EXPERIMENTS

In order to illustrate the performance of our model, we conducted several experiments on transcription of polyphonic piano pieces, where the aim is to recognize the notes and the time intervals in which they are played. Although HNDS and CHNDS can be applied to many other applications, we choose to evaluate it on this transcription problem as the activation matrix $\mathbf{U}$ is then likely to have a physical interpretation which will enable us to introduce symbolic data as side information.

In our experiments, we use the MIDI Aligned Piano Sounds (MAPS) database [14]. In order to train the spectral dictionary $\mathbf{W}$, we use $K = 88$ monophonic piano sounds with 44.1 kHz sampling rate. In all experiments, the audio is subdivided into frames of 93 ms with 50% overlap, and only frequencies in the 0 to 4.3 kHz range are used, for a total of $F = 400$ frequency bins. As the test set, we use 10 excerpts of 10 seconds ($N = 217$ frames) that are gathered from 2 different pieces from Bach.

In order to evaluate the transcription performance, we first threshold the estimated activation matrix $\mathbf{U}$ and compare the resulting binary matrix with the true transcription by computing the precision, recall, and F-measure values defined as follows: precision $=$ true positive$/($true positive $+$ false positive$)$, recall $=$ true positive$/($true positive $+$ false negative$)$, and F-measure $=$ $(2 \times$ precision $\times$ recall$)/($precision $+$ recall$)$. The precision denotes the ratio of the number of correct notes to that of retrieved notes. Similarly, the recall denotes the ratio of the number of correct notes to that of true notes.

In all experimental settings, we follow a semi-supervised approach where we initially train the spectral dictionary $\mathbf{W}$ by using the isolated note sounds. Then, during testing, we hold $\mathbf{W}$ fixed and estimate the other variables (including $\mathbf{A}$ and $\mathbf{B}$). In the first part, we conduct the transcription experiments by using the HNDS model. In this model, the free parameters provide a great amount



(a) A (HNDS) (b) A (CHNDS) (c) B (HNDS) (d) B (CHNDS)

Figure 2: Estimated $\mathbf{A}$ and $\mathbf{B}$ by HNDS and CHNDS (for Bach BWV 847). It can be observed that introducing symbolic data results in better-behaved estimation of $\mathbf{A}$ and $\mathbf{B}$.

of flexibility, yet, they make it harder to train the model as the inference algorithm may get stuck in local optima. This provides a motivation to introduce further constraints on the model.

In the second part, we introduce symbolic music data (i.e. MIDI) to the model for data-driven regularization, where we make use of the CHNDS model. Since the symbolic data matrix $\mathbf{U}'$ is expected to have a similar form to that of an activation matrix, we construct $\mathbf{U}'$ using a damping factor as follows: $u'_{km} = \delta u'_{k(m-1)}$ with $0 < \delta < 1$, provided the note k is active at both time frames $m-1$ and m. This representation mimics the structure of $\mathbf{U}$. In these experiments, the symbolic data and the audio data do not belong to the same piece, but they have similar chord structure.

We compare the performance of the proposed model with existing factorization models, namely the Itakura-Saito NMF [5], the coupled matrix factorization (CMF) [4], and the non-negative dynamical system (NDS) [6]. Examples of $\mathbf{A}$ and $\mathbf{B}$ matrices estimated by HNDS and CHNDS are shown in Fig. 2. For NDS, HNDS and CHNDS, we experiment on three different regimes. Other than the default regime where $\mathbf{A}$ is a full matrix, in a second regime we constrain $\mathbf{A}$ to be diagonal, and in a third regime we further constrain $\mathbf{A}$ to be equal to the identity matrix. Note that the NDS model reduces to smooth NMF [12] when $\mathbf{A}$ is chosen as identity. For all models, we investigate various parameter settings and threshold values. The results with best F-measure values are given in Table 2.

5. DISCUSSION

The results show that our proposed model, HNDS, outperforms the other related models in terms of F-measure. It can be observed that constraining the $\mathbf{A}$ matrix yields better performance for all methods. The models attain the best performance when $\mathbf{A}$ is constrained to be the identity. This was to be expected: constraining the model is likely to make the estimation better-behaved, especially as $\mathbf{A}$ and $\mathbf{B}$ are here estimated on the test data, which is limited; this was also particularly likely to occur on this data, which is well-suited to methods like smooth NMF due to its rather simple dynamics.

Introducing symbolic data (CHNDS) further improves the performance by guiding the estimation of $\mathbf{B}$. We can observe that guidance by a reference transcription helps preventing the occurrence of spurious notes that other models introduce, thus improving the precision value. The best results are achieved by CHNDS with $\mathbf{A} = \mathbb{I}$ and the following parameter settings: $\alpha_i = 4$, $\gamma_k = 1.5$, $\kappa_i = 0.1$, $\nu_k = 2$, $\lambda_W = \lambda_A = \lambda_B = 1$, and $I = 50$.

On this particular data, we found that constraining the transition matrix led to the best results. We plan to investigate the performance of our model on data with richer dynamics, where we hope to better exploit the flexibility of a full transition matrix.

Table 2: Precision (%), Recall (%), and F-measure comparison of the proposed model and related models.

Method	Precision	Recall	F-measure
IS-NMF [5]	58.66	65.01	61.67
Coupled MF [4]	66.81	70.08	68.50
NDS [6]	66.78	79.93	72.77
NDS (diag. $\mathbf{A}$)	67.77	85.92	75.77
NDS ($\mathbf{A} = \mathbb{I}$)	74.95	**89.40**	81.54
HNDS [Proposed]	78.27	86.10	82.00
HNDS (diag. $\mathbf{A}$)	79.62	86.19	82.77
HNDS ($\mathbf{A} = \mathbb{I}$)	80.93	86.50	83.62
CHNDS [Proposed]	81.47	82.42	81.94
CHNDS (diag. $\mathbf{A}$)	79.12	85.22	82.06
CHNDS ($\mathbf{A} = \mathbb{I}$)	**84.50**	84.78	**84.64**

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, vol. 13, 2001, pp. 556–562.

[2] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *ICASSP*, 2011.

[3] E. M. Grais and H. Erdoğan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation," *Comput. Speech Lang.*, vol. 27, pp. 746–762, 2013.

[4] U. Şimşekli and A. T. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *EUSIPCO*, 2012.

[5] 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.

[6] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical sytem with application to speech and audio," in *ICASSP*, 2013.

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

[8] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *LVA/ICA*, 2010.

[9] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *WASPAA*, 2009.

[10] M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *LVA/ICA*, Sep. 2010, pp. 149–156.

[11] O. Dikmen and A. T. Cemgil, "Unsupervised single-channel source separation using Bayesian NMF," in *WASPAA*, 2009.

[12] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *ICASSP*, 2011.

[13] U. Şimşekli, Y. K. Yılmaz, and A. T. Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *ICASSP*, 2012.

[14] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE TASLP*, vol. 18, no. 6, pp. 1643–1654, 2010.