A Hierarchical Latent Mixture Model for Polyphonic Music Analysis

Cian O'Brien and Mark D. Plumbley
Centre for Vision, Speech and Signal Processing
University of Surrey
United Kingdom
Email: {cj.obrien, m.plumbley}@surrey.ac.uk

***Abstract*—Polyphonic music transcription is a challenging problem, requiring the identification of a collection of latent pitches which can explain an observed music signal. Many state-of-the-art methods are based on the Non-negative Matrix Factorization (NMF) framework, which itself can be cast as a latent variable model. However, the basic NMF algorithm fails to consider many important aspects of music signals such as low-rank or hierarchical structure and temporal continuity. In this work we propose a probabilistic model to address some of the shortcomings of NMF. Probabilistic Latent Component Analysis (PLCA) provides a probabilistic interpretation of NMF and has been widely applied to problems in audio signal processing. Based on PLCA, we propose an algorithm which represents signals using a *collection* of low-rank dictionaries built from a base pitch dictionary. This allows each dictionary to specialize to a given chord or interval template which will be used to represent collections of similar frames. Experiments on a standard music transcription data set show that our method can successfully decompose signals into a hierarchical and smooth structure, improving the quality of the transcription.**

I. Introduction

Latent variable techniques represent a diverse collection of algorithms for analyzing signals and data. A common assumption in many signal processing or machine learning applications is that observed signals can be fully or partially explained by some hidden latent factors such as class label. Algorithms for discovering these latent variables under different assumptions include Principal Component Analysis (PCA) [1], Gaussian Mixture Models (GMM) [2], Dictionary Learning (DL) [3], Independent Component Analysis (ICA), Latent Dirichlet Allocation (LDA) [4] and many others. These algorithms have been applied to problems as diverse as document clustering/topic-modelling, image denoising, speech recognition and audio source separation.

We are interested in analyzing polyphonic music signals, which offer many interesting challenges. For example, music signals can be explained by many possible latent factors including music genre, instrumentation, emotion or musical key signature. As an added complication, many of these factors have a hierarchical relationship – for example the key signature of a piece informs which pitches are likely or possible. Furthermore, music is highly temporal so that the state at one instant informs the subsequent states. Our goal with this work is to model these correlated and hierarchical structures. The main application we consider is that of recognizing the pitch content of a recorded piece of music from just the observed signal. In doing so, we will encounter many of the issues discussed above. In the following we state the music transcription problem (Section II), design a general technique for music analysis (Sections III & IV) and evaluate the proposed method on a standard music transcription data set (Section V).

II. Automatic Music Transcription

Automatic Music Transcription (AMT) attempts to reproduce the pitch content of a music signal. That is, it seeks a representation which specifies what musical pitches are present at each time frame. Note that the word “pitch” refers to a musical note in the Western system (i.e., a single piano key) and in terms of a spectrogram, it consists of the fundamental frequency together with a series of higher frequencies and overtones (this series is highly dependent on the instrument, playing style, volume, environment etc). Given a time-frequency matrix $\boldsymbol{X}$ of a recorded music signal, we seek a binary *transcription matrix* which specifies the presence/absence of each musical pitch at every time frame.

The most commonly used approach for AMT is the Non-negative Matrix Factorization (NMF) algorithm [5] [6]. Given a signal matrix with non-negative entries, NMF seeks two non-negative matrices such that their product equals the original signal. In NMF, the observed signals are represented as additive combinations of elements from a learned dictionary. In terms of AMT, the signal matrix $\boldsymbol{X}$ is usually a magnitude-frequency representation of the signal and we seek a factorization of the form

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \tag{1}$$

where $\boldsymbol{W}$ is a matrix whose columns contain individual pitches and the matrix $\boldsymbol{H}$ represents the final transcription. We assume that the desired transcription is a binary matrix indicating the presence or absence of a pitch at each time frame. A commonly used time-frequency representation is the Constant-Q transform which is logarithmic in the frequency axis. In this work we used an Equivalent Rectangular Bandwidth transform which is perceptually motivated and has been shown to work well for AMT.

NMF can be cast as a latent variable model, in which the observed signals are explained by some latent factors – for

AMT these factors are the individual pitches which we would like to infer. We propose a latent variable model which extends the classic NMF algorithm, in order to make it better suited to analyzing music signals. The key points are:

- *Local stationarity*: once active, groups of pitches tend to stay active over many frames. For example, once a piano key is pressed it tends to be held over many time frames and emits a continuous sound with a natural decay. We capture this by building a model which implicitly imposes a *local* low rank constraint, so that related frames will share the same explanatory factors.
- *Hierarchical structure*: related pitches tend to co-occur in the form of intervals, triads and chords[1]. Indeed, we can specify a distinct hierarchy from chord states down to intervallic structure and individual pitches. We achieve this by constructing a series of *local dictionaries* which are composed of combinations of pitches from a *base dictionary*, such that each local model will be built using atoms from the base dictionary.

III. A Hierarchical Latent Mixture Model for Automatic Music Transcription

In this section we propose a latent variable model for the automatic transcription of polyphonic music. We use a probabilistic framework which is closely related to the PLCA model of Smaragdis [7][8], which has in turn been applied to AMT in various forms by Benetos [9][10][11][12]. We model the observed time-frequency signal as a joint distribution $P(f,t)$, where the variable $t \in \mathcal{T}$ is the time location where $P(t)$ supported over the set of frame indices and $f \in \mathcal{F}$ is a frequency bin. This joint distribution induces the following generative model for the observed signals: first sample a frame index t with probability $P(t)$ and then sample an event $P(f \,|\, t)$. As with PLCA, our general strategy will be to factorize the joint distribution $P(f,t)$ using a latent variable model and learn the factors which best fit the signal.

We may factor the joint distribution according to our generative model as follows

$$P(f,t) \;=\; P(t)\, P(f \,|\, t). \tag{2}$$

Further factorizing $P(f \,|\, t)$ gives the asymmetric PLCA model

$$P(f,t) \;=\; P(t) \sum_{p} P(f \,|\, p) P(p \,|\, t). \tag{3}$$

Under this model, $P(f,t)$ is given as a sum over p latent pitches which explain this signal. It has been shown that up to a scaling, this model is equivalent to classical NMF using Kullback-Liebler divergence.

As opposed to standard PLCA, in this work we suppose that each latent model corresponds to a *collection* of pitches from some base dictionary, so that higher level concepts such as intervals and chords can be introduced. Additionally, we aim to represent each frame as a weighted collection of such models. The underlying idea is that similar frames should use the same latent factors in their reconstruction, and that these latent factors will then adapt to specific commonly-occurring pitch combinations. Furthermore by limiting the size of the individual models, the resulting transcription is likely to be locally low rank which has been shown to improve AMT performance over standard NMF/PLCA [13]. For example, a local model with a rank of r can learn r different combinations of pitches in order to account for local changes in active pitches, relative amplitudes and decay profiles.

Formally, suppose that the signal is composed of m latent models, each of which is of rank-r. Each model should be constrained to lie in a so-called *pitched-subspace*, so that they correspond to linear combination of valid pitches. In the language of NMF, the pitched-subspace will correspond to the space defined by the columns of the *base dictionary* $\boldsymbol{D}$ and the signal will be given by a weighted combination of local models derived from $\boldsymbol{D}$. By constraining the rank of each model, the resulting transcription will be locally low-rank in the sense that all frames assigned to a given model will share similar latent factors.

We can introduce this into the model given in (3) by defining

$$P(p \,|\, t) \;=\; \sum_{m,r} P(p \,|\, m,r,t) P(r \,|\, m,t) P(m \,|\, t). \tag{4}$$

Note the similarity between this and the Low Rank Matrix Decomposition approach [13], where here we have explicitly factored the $P(p \,|\, t)$, as opposed to placing a nuclear-norm constraint on the transcription matrix. Combining (4) and (2) gives the proposed *Hierarchical Latent Mixture Model* (HLMM)

$$P(f,t) \;=\; P(t) \sum_{p,m,r} P(f \,|\, p) P(p \,|\, m,r,t) P(r \,|\, m,t) P(m \,|\, t). \tag{5}$$

In practice this is similar to a convolutional template-matching approach such as convolutive NMF [14] or Shift Invariant PLCA [15], but with *groups* of pitch templates. This dynamic has a clustering interpretation, where the factors $P(m \,|\, t)$ can be seen as a soft-assignment of every frame to each model m. It also bears a relationship to group-sparse methods, where collections of pitch templates are encouraged to activate together. Furthermore, correlations between pitches may be discovered by grouping co-occurring pitches together in a single model, which happens naturally during the inference stage.

The proposed model is similar to the Hierarchical Eigeninstruments approach of Grindlay and Ellis [16][17], which models the observed signal as a collection of instrument-specific subspaces. However, their work focussed on multi-instrument transcription and does not exploit any particular structure to improve the transcription quality of a given instrument. Another related idea is the Non-negative Hidden Markov Model (NN-HMM) which was developed for source separation by Mysore et al. [18]. NN-HMM learns a sequence of dictionaries to represent local signals. There are several key difference between the proposed HLMM and these approaches: (i) we

[1] In music theory, an interval describes the musical relationship between two distinct pitches. Triads consist of triplets of pitches with certain intervallic structures and form the building blocks of chords.

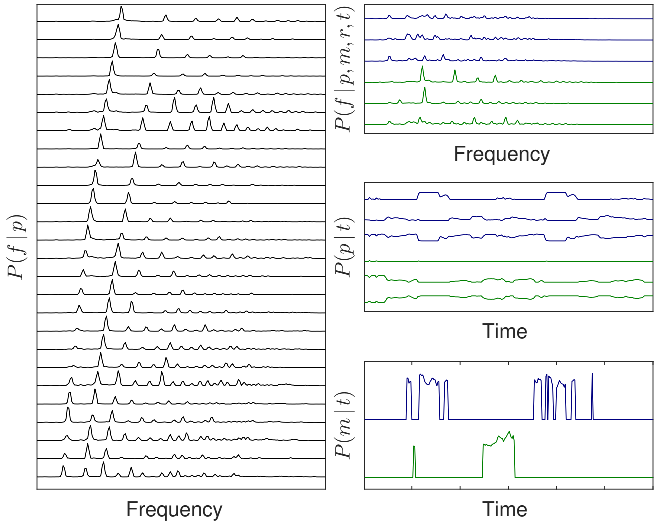

Fig. 1. The proposed approach applied to piano music. In total, a collection of 30 models of rank-3 were learned. **Left**: a collection of templates $P(f\,|\,p)$ from the base pitch dictionary. **Top right**: two examples of the learned local dictionaries $P(f\,|\,p,m,r,t)$. Each colour (blue and green) corresponds to one of the two models, each of which has 3 templates constructed from the base templates $P(f\,|\,p)$. **Middle right**: the learned activations for each model template. **Bottom right**: the model weights $P(m\,|\,t)$ – note the extreme sparsity.

use a *mixture* of models to represent frames so that each model can be highly specialized. (ii) Each model can be used by any frame, unlike the NN-HMM where each frame is constrained to use the currently active dictionary. (iii) We use a *hierarchical* model, where the local dictionaries are themselves constructed from a more fundamental base dictionary. This can be useful in cases where we have domain knowledge about the problem (by constraining the dictionaries to live in the pitched subspace, for example). This is also a natural assumption for many types of signals such as natural images where shapes and patterns are built by combining oriented edges.

IV. Inference

Given an observed time-frequency signal $\pi(f,t)$ we need to fit the best factors in (5). This can be done using the expectation-maximization (EM) algorithm to maximize the following log-likelihood

$$\sum_{f,t} \pi(f,t) \log\big(P(t)P(f\,|\,t)\big) \tag{6}$$

which has been shown to be equivalent to minimizing the KL-divergence between $\pi(f,t)$ and $P(f,t)$ [19]. During the E-step, we compute the posterior distribution of the latent variables p, m and r which by Bayes theorem is given by

$$P(p,m,r\,|\,f,t) \;=\; \frac{P(f\,|\,p)P(p\,|\,m,r,t)P(r\,|\,m,t)P(m\,|\,t)}{P(f\,|\,t)}. \tag{7}$$

During the M-step, the remaining factors are updated using this posterior:

$$P(f\,|\,p) \;=\; \frac{\sum_{m,r} P(p,m,r\,|\,f,t)\,\pi(f,t)}{\sum_{f,m,r} P(p,m,r\,|\,f,t)\,\pi(f,t)} \tag{8}$$

$$P(p\,|\,m,r,t) \;=\; \frac{\sum_{f} P(p,m,r\,|\,f,t)\,\pi(f,t)}{\sum_{f,p} P(p,m,r\,|\,f,t)\,\pi(f,t)} \tag{9}$$

$$P(r\,|\,m,t) \;=\; \frac{\sum_{f,p} P(p,m,r\,|\,f,t)\,\pi(f,t)}{\sum_{f,p,r} P(p,m,r\,|\,f,t)\,\pi(f,t)}. \tag{10}$$

The factor $P(m\,|\,t)$ gives the contribution of model m at time-frame t. For example, one possibility would be a degenerate weighting in which every model contributes equally at each frame by taking $P(m\,|\,t) \approx 1/K$ for all m and t, given K models. In this case the proposed approach would reduce to independently training K rank-r models to fit the full signal and would severely limit the expressive power. Ideally we want each model to specialize in representing a given chord or interval in various decay configurations and therefore we encourage sparsity in the final value by taking

$$P(m\,|\,t) \;=\; \frac{\Big(\sum_{f,p,r} P(p,m,r\,|\,f,t)\,\pi(f,t)\Big)^{\alpha}}{\sum_m \Big(\sum_{f,p,r} P(p,r\,|\,f,t)\,\pi(f,t)\Big)^{\alpha}} \tag{11}$$

where $\alpha \geq 1$. After updating we $P(m\,|\,t)$ we set values below a set threshold to zero before renormalizing. After solving for each latent factor in (3), the joint distribution $P(p,t)$ is given by

$$P(p,t) \;=\; \frac{\sum_{m,r}\; P(p\,|\,m,r,t)P(r\,|\,m,t)P(m\,|\,t)P(t)}{\sum_{p,m,r}\; P(p\,|\,m,r,t)P(r\,|\,m,t)P(m\,|\,t)P(t)}. \tag{12}$$

which is the desired transcription. Finally, we binarize the transcription by setting to zero any values below a threshold and setting the remaining entries to one.

V. Experimental Evaluation

The proposed system was evaluated on 30-second excerpts from the *EnStDkcl* subset of the Midi Aligned Piano Sounds (MAPS) dataset, which consists of recordings of classical piano music. The global dictionary $P(f\,|\,p)$ was initialized by training NMF models on recordings of isolated pitches from the test instrument. To initialize the local dictionaries $P(p\,|\,m,r,t)$, we randomly chose collections of r frames from the signal and decomposed them using NMF with the base dictionary. The activations were initialized by performing NMF over the full signal and the model weights $P(m\,|\,t)$ where set to $1/K$. The full model was then learned by iterating the EM algorithm:

- *E-step*: Compute the posterior using (7).
- *M-step*: Update the latent factors using (8), (9) and (10). Note that these computations can be done in parallel over each model.

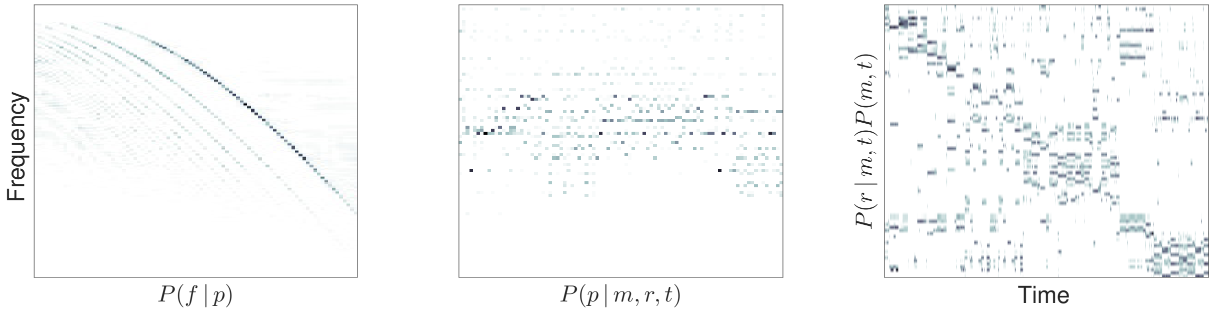

Fig. 2. Learned factors from 30-seconds of piano music. **Left**: the base dictionary. **Middle**: the local dictionaries arranged sequentially, with $K = 30$ and $r = 3$. **Right**: the weighted activations.

After estimating all of the factors, we formed a single large dictionary and activation matrix by concatenating each weighted model. This was used a warm-start for several iterations of NMF to produce the final transcription. The resulting transcription was binarized by setting

$$P(p,t) = \begin{cases} 0 & \text{if} \quad P(p,t) \leq \mu_t + \tau\sigma_t \\ 1 & \text{otherwise} \end{cases}$$

where μ_t is the mean activation value for frame t, σ_t its standard deviation and τ a positive constant. For each track we compute the number of true-postives (N_{tp}), false-positives (N_{fp}) and false-negatives (N_{fn}) are used to calculate the precision ($\mathcal{P}$), recall ($\mathcal{R}$) and F-measure ($\mathcal{F}$)

$$\mathcal{P} = \frac{N_{tp}}{N_{tp} + N_{fp}} \quad \mathcal{R} = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad \mathcal{F} = 2\,\frac{\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}. \quad (13)$$

We compared the results of the proposed Hierarchical Latent Mixture Model (HLMM) to several related approaches: β Non-negative Matrix Decomposition (β-NMD), Weighted β Non-negative Matrix Factorization (Wβ-NMF) [20], low rank Non-negative Matrix Decomposition (LR-β-NMD) [13] and the best performing system reported by O'Hanlon et al. [21], which was Group-Sparse Non-negative Matrix Decomposition using KL-divergence and a subspace pitch-dictionary (GS-KL-NMD). The results are summarized in Table I where we report the per-track average precision, recall and F-measure using the best threshold value τ. The value of the threshold is important and difficult to set a priori – in Figure 3 we show its affect for both HLMM and β-NMF using the first track from the test set.

The proposed method significantly outperforms the standard NMF approach across all metrics. In Figure 2 we present the learned factors. Note the strong diagonal in the weight matrix, which shows that frames tend to favour their local model in reconstruction. Additionally, we see that several models are used by different groups of frames which may indicate repeated structures in the music. The middle figure shows the learned local dictionaries, which have specialized into distinct chord and interval patterns.

Reference	Model	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
	β-NMD	73.13	70.90	71.70
	PLCA	72.26	72.41	72.07
[13]	LR-β-NMD	73.83	73.17	73.50
[20]	Wβ-NMD			73.70
[21]	GS-KL-NMD			74.10
Proposed	HLMM	**75.58**	**76.11**	**75.54**

TABLE I
TRANSCRIPTION RESULTS ON THE *EnStDkcl* DATA SET.

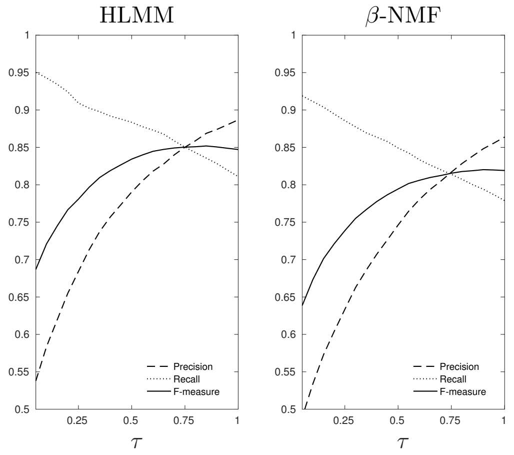

Fig. 3. Precision, recall and F-measure curves for the first track in the *EnStDkcl* data set, for different values of the threshold parameter τ. For HLMM, the best F-measure was 0.852 versus 0.82 for β-NMF.

VI. Conclusion

We have presented an approach to automatically transcribing the pitch content of audio signals. Starting with Probabilistic Latent Component Analysis, we extended the basic factorization algorithm to represent the signal as a weighted collection of models built from a fundamental pitch dictionary. The resulting algorithm can infer common chord and interval combinations which can be used to represent collections of related frames. In the future it would be interesting to investigate adding additional hierarchies to the model; several works have shown the viability of decomposing the pitches into collection of narrow band atoms, thus representing each base pitch using a distinct subspace. This idea could be readily adapted to fit the proposed model, resulting in a hierarchy from narrow-band atoms all the way up to repeated musical structures. Another direction is to investigate the connection between clustering methods, group sparse factorizations and low-rank models suggested by the proposed model.

VII. Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under Grant Agreement no. 607290 SpaRTaN.

References

[1] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[2] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2003, pp. 177–180.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, vol. 148, pp. 8–1, 2006.

[8] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," *Journal of Machine Learning Research*, 2007.

[9] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *8th Sound and Music Computing Conference*, 2011, pp. 19–24.

[10] ——, "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, 2012.

[11] E. Benetos, A. Jansson, and T. Weyde, "Improving automatic music transcription through key detection," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[12] E. Benetos and S. Dixon, "A temporally-constrained convolutive probabilistic model for pitch detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 133–136.

[13] C. O'Brien and M. D. Plumbley, "Automatic music transcription using low rank non-negative matrix decomposition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 1848–1852.

[14] P. D. O'grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*. IEEE, 2006, pp. 427–432.

[15] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2069–2072.

[16] G. Grindlay and D. P. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription." in *ISMIR*, 2010, pp. 21–26.

[17] ——, "Multi-voice polyphonic music transcription using eigeninstruments," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*. IEEE, 2009, pp. 53–56.

[18] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 140–148.

[19] D. Cazau and G. Nuel, "Understanding the probabilistic latent component analysis framework," *arXiv preprint arXiv:1703.05208*, 2017.

[20] K. O'Hanlon and M. D. Plumbley, "Automatic music transcription using row weighted decompositions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 16–20.

[21] K. O'Hanlon, H. Nagano, N. Keriven, and M. D. Plumbley, "Non-negative group sparsity with subspace note modelling for polyphonic transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 530–542, 2016.