DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION FOR MULTIPLE PITCH ESTIMATION

Nicolas Boulanger-Lewandowski, Yoshua Bengio and Pascal Vincent
Dept. IRO, Université de Montréal
Montréal, Québec, Canada H3C 3J7
{boulanni, bengioy, vincentp}@iro.umontreal.ca

ABSTRACT

In this paper, we present a supervised method to improve the multiple pitch estimation accuracy of the non-negative matrix factorization (NMF) algorithm. The idea is to extend the sparse NMF framework by incorporating pitch information present in time-aligned musical scores in order to extract features that enforce the separability between pitch labels. We introduce two discriminative criteria that maximize inter-class scatter and quantify the predictive potential of a given decomposition using logistic regressors. Those criteria are applied to both the latent variable and the deterministic autoencoder views of NMF, and we devise efficient update rules for each. We evaluate our method on three polyphonic datasets of piano recordings and orchestral instrument mixes. Both models greatly enhance the quality of the basis spectra learned by NMF and the accuracy of multiple pitch estimation.

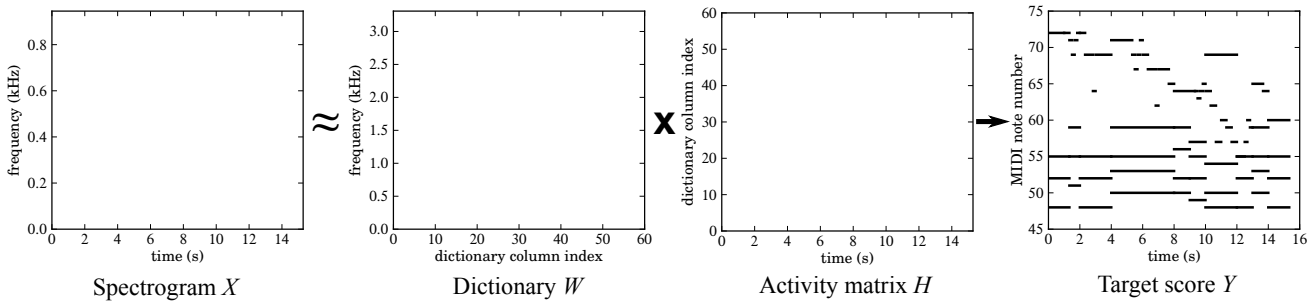
1. INTRODUCTION

Non-negative matrix factorization (NMF) is an unsupervised technique to discover parts-based representations underlying non-negative data [12], i.e. a set of characteristic components that can be combined additively to reconstitute the observations. When applied to the magnitude spectrogram of a polyphonic audio signal, NMF can discover a basis of interpretable recurring note events and their associated time-varying encodings, or *activities*, that together optimally reconstruct the original spectrogram.

In general, the extracted representation will converge to individual note spectra provided the following conditions are met [5]. First, each observed spectrogram frame must be representable as a non-negative linear combination of the isolated note spectra, an approximation that depends on the interference between overlapping harmonic partials in a polyphonic mix but that is nevertheless reasonable [22]. The second condition requires that basis spectra be linearly independent, and the third condition requires that all combinations of individual notes be present in the database. This last assumption is of course difficult to achieve completely but partial combinations seem sufficient in practice. Consequently, the activities extracted by NMF have proven useful as features to detect individual note pitches played simultaneously at a given instant in a polyphonic audio signal, a task known as multiple pitch estimation, and for the related task of transcribing audio excerpts into musical notation [1, 3, 4, 19]. Sparsity, temporal and spectral priors have proven useful to enhance the accuracy of multiple pitch estimation [3, 7, 20].

Since NMF is an unsupervised technique, it can be applied in principle to an unlimited number of musical recordings without the need for ground-truth *pitch labels*. However, such information is often readily available as recorded expressive performances, symbolic sequences (e.g. a MIDI file) or time-aligned musical scores. In those cases, we would like to exploit the pitch information to steer the NMF decomposition in a supervised way to obtain discriminative features more useful for multiple pitch estimation. A few attempts have been made in this direction, notably by adding a linear discriminant analysis (LDA) stage to the activities extracted by NMF [23], or by embedding Fisher-like discriminant constraints inside the decomposition [9, 21, 23]. Discriminative dictionaries have also been developed for sparse coding [15]. Those methods however are designed for classification, which means choosing a single label, whereas multiple pitch estimation is a multi-label task, i.e. multiple pitch labels can be associated with a single spectrogram frame. In this context, we propose two discriminative criteria that maximize inter-class scatter for each label separately and estimate the predictive power of a given decomposition using logistic regressors. Those ideas are applied in the conventional latent variables framework of NMF and in a deterministic autoencoder model to directly maximize test-time discriminative performance. Efficient update rules are devised for each, and we show that our method greatly improves the quality of the basis spectra learned by NMF and the accuracy of multiple pitch estimation on three polyphonic datasets of piano recordings and orchestral instrument mixes.

The remainder of this paper is organized as follows. In Sections 2 and 3, we review the NMF algorithm and its application to multiple pitch estimation. In Sections 4 and 5 we introduce the latent variables and autoencoder discriminative models. We describe our experiments and evaluate our method in Sections 6 and 7.



Spectrogram X | Dictionary W | Activity matrix H | Target score Y

Figure 1. Illustration of the sparse NMF decomposition ($\lambda = 0.01$, $\mu = 10^{-5}$) of an excerpt of Drigo's *Serenade*. Using a dictionary W pretrained on a polyphonic piano dataset, the spectrogram X is transformed into an activity matrix H approximating the piano-roll transcription Y. The columns of W were sorted by increasing estimated pitch for visualization.

2. NON-NEGATIVE MATRIX FACTORIZATION

The NMF method aims to discover an approximate factorization of an input matrix X:

$$\overset{n_f \times n_t}{X} \simeq \overset{n_f \times n_t}{\Lambda} \equiv \overset{n_f \times m}{W} \cdot \overset{m \times n_t}{H} \tag{1}$$

where X is the observed magnitude spectrogram with time and frequency dimensions n_t and n_f respectively, Λ is the reconstructed spectrogram, W is a dictionary matrix of m basis spectra and H is the activity matrix. Non-negativity constraints $W_{i,j} \geq 0, H_{i,j} \geq 0$ apply on both matrices. NMF seeks to minimize the *reconstruction error*, a distortion measure between the observed spectrogram X and the reconstruction Λ. A popular choice is the Euclidean distance:

$$C_{LS} \equiv ||X - \Lambda||^2 \tag{2}$$

with which we will demonstrate our method although it can be easily generalized to other distortion measures in the β-divergence family [11]. Minimizing C_{LS} can be achieved by alternating multiplicative updates to H and W [13]:

$$H \leftarrow H \circ \frac{W^T X}{W^T \Lambda} \tag{3}$$

$$W \leftarrow W \circ \frac{X H^T}{\Lambda H^T} \tag{4}$$

where the $\circ$ operator denotes element-wise multiplication, and division is also element-wise. These updates are guaranteed to decrease the reconstruction error assuming a local minimum is not already reached. While the objective is convex in either W or H separately, it is non-convex in W and H together and thus finding the global minimum is intractable in general.

2.1 Sparsity constraints

In a polyphonic signal with relatively few notes played at any given instant, it is reasonable to assume that active elements H_{ij} should be limited to a small subset of the available basis spectra. To encourage this behavior, a sparsity penalty C_S can be added to the total SNMF objective [10]:

$$C_S = \lambda |H| \tag{5}$$

where $|\cdot|$ denotes the L_1 norm and λ specifies the relative importance of sparsity. In order to eliminate underdetermination associated with the invariance of WH under the transformation $W \rightarrow WD$, $H \rightarrow D^{-1}H$, where D is a diagonal matrix, we impose the constraint that the basis spectra have unit norm. Equation (3) becomes:

$$H \leftarrow H \circ \frac{W^T X}{W^T \Lambda + \lambda} \tag{6}$$

and the multiplicative update to W (equation 4) is replaced by projected gradient descent [14]:

$$W \leftarrow W - \mu(\Lambda - X)H^T \tag{7}$$

$$W_{:i} \leftarrow \frac{W_{:i}}{||W_{:i}||} \tag{8}$$

where $W_{:i}$ is the i-th column of W, μ is the learning rate and $1 \leq i \leq m$.

3. NMF FOR MULTIPLE PITCH ESTIMATION

The ability of NMF to extract fundamental note events from a polyphonic mixture makes it an obvious stepping stone for multiple pitch estimation. In the ideal scenario, the dictionary W contains the spectrum profiles of individual notes composing the mix and the activity matrix H approximately corresponds to the ground-truth score. An example of the sparse NMF decomposition of an excerpt of Drigo's *Serenade* using a dictionary pretrained on a simple polyphonic piano dataset is illustrated in Figure 1. The dictionary contains mostly monophonic basis spectra that were sorted by increasing estimated pitch for visualization. We also observe a clear similarity between the activity matrix and the target score in a piano-roll representation Y.

There are many options to exploit the NMF decomposition to perform actual multiple pitch estimation. The *dictionary inspection* approach [1, 18, 19] consists in estimating the pitch (or lack thereof) of each column of W, which can be done automatically using harmonic combs [20], and to transcribe all pitches for which the associated H_{ij} activities exceed a threshold η:

$$Y_{kj} = 1 \Leftrightarrow \sum_{i|L(i)=k} H_{ij} \geq \eta \tag{9}$$

where $L(i)$ is the estimated pitch label (index) of the i-th basis spectrum. For this method, a new factorization can be performed adaptively for each new piece to analyze, or the dictionary can be pretrained from an extended corpus and kept fixed during testing. Dictionaries can also be constructed from the concatenation of isolated note spectra [3, 4].

Another option is to predict each column of Y from the corresponding column of H using a general-purpose multi-label classifier or a set of binary classifiers, one for each label (note) in the designated range. This obviously requires the use of a fixed dictionary and the availability of annotated pieces to train the classifiers. In this work, we will exclusively employ pretrained dictionaries and we will consider both dictionary inspection and multi-label classification with linear support vector machines (SVM) [17].

4. DISCRIMINATIVE CRITERIA

The simple interpretation of the activity matrix as an approximate transcription usually deteriorates when we increase instrumental diversity, pitch range or polyphony. In this section, we introduce two discriminative criteria exploiting the aligned score information Y to ensure that NMF extracts meaningful features into W and H.

The first criterion is inspired from linear discriminant analysis in that we aim to maximize the inter-class scatter of the H_{ij}, where the classes here refer to the presence or absence of a given pitch label at a given time. We encourage the activities associated with a given basis spectrum to be maximal when its pitch is present in the score and minimal otherwise, such that a unidimensional decision threshold is sufficient to estimate the presence of a note. We first assign a pitch label $L(i)$ to each column i of W, or set $L(i) = -1$ to denote an unpitched basis spectrum. Due to the invariance of WH under the column permutation of W and the equivalent row permutation of H, this assignment can be done arbitrarily as long as the number of basis spectra describing each pitch (q) and the number of unpitched spectra ($\bar{q}$) remain constant. More precisely, this criterion has the form:

$$C_d(H) = \sum_{ij} \begin{cases} -\beta^+ H_{ij} & \text{if } Y_{L(i),j} = 1 \\ \beta^- H_{ij} & \text{if } Y_{L(i),j} = 0 \\ 0 & \text{if } L(i) = -1 \end{cases} \quad (10)$$

where the β^+ and β^- parameters quantify respectively the importance of presence and absence of an H_{ij} element. Note that the limit $\beta^- \to \infty$ corresponds to setting $H_{ij} = 0$ for $Y_{L(i),j} = 0$.

The second proposed criterion does not impose a predetermined structure on the activity matrix, but rather attempts to determine whether H is a good predictor for Y. We introduce a stage of logistic regressors with weight matrix V and bias vector b using H as input:

$$p_{kj} = \sigma((VH)_{kj} + b_k) \quad (11)$$

where $\sigma(x) \equiv (1+e^{-x})^{-1}$ is the element-wise logistic sigmoid function and p is an output matrix of note probabilities, or *probabilistic piano-roll*. We use the cross-entropy as a discriminative criterion for H:

$$C_l(H) = -\alpha \sum_{kj} Y_{kj} \log p_{kj} + (1 - Y_{kj}) \log(1 - p_{kj}) \quad (12)$$

where α is a weighting coefficient. Adding our criteria to the total objective yields the DNMF model:

$$C = C_{LS} + C_S + C_d + C_l. \quad (13)$$

It is easy to show that the Hessian matrices $\nabla^2_H C_d(H)$ and $\nabla^2_H C_l(H)$ are both positive semi-definite and that the DNMF objective remains convex in W or H separately. The multiplicative update rule for H (equation 6) becomes:

$$H \leftarrow H \circ \frac{W^T X}{W^T \Lambda + \lambda + \frac{\partial C_d(H)}{\partial H} + \frac{\partial C_l(H)}{\partial H}} \quad (14)$$

where the gradients are given by:

$$\frac{\partial C_d(H)}{\partial H_{ij}} = \begin{cases} -\beta^+ & \text{if } Y_{L(i),j} = 1 \\ \beta^- & \text{if } Y_{L(i),j} = 0 \\ 0 & \text{if } L(i) = -1 \end{cases} \quad (15)$$

$$\frac{\partial C_l(H)}{\partial H} = \alpha V^T (p - Y). \quad (16)$$

The update rules for W are the same as for sparse NMF and are given by (7) and (8). The V and b parameters are optimized via stochastic gradient descent using the updates:

$$V \leftarrow V - \mu(p - Y)H^T \quad (17)$$

$$b_k \leftarrow b_k - \mu \sum_j (p_{kj} - Y_{kj}). \quad (18)$$

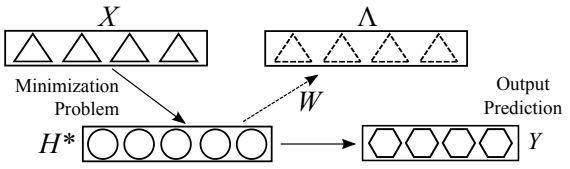


Figure 2. In the DNMF autoencoder model, the input is encoded via a deterministic minimization procedure. The code H^* is trained to reconstruct X and to predict Y.

5. AUTOENCODER MODEL

In the probabilistic latent variables model (LV) underlying NMF, the activities are regarded as hidden variables with joint negative log probability given by (13) and the use of equations (14) and (7-8) during training corresponds to the expectation and maximization phases of an EM algorithm [12]. A subtlety associated with this interpretation arises in testing conditions when the labels Y are unknown. We can resort to equation (6) to infer H, but it is possible to address this issue in a more principled manner with the autoencoder model (AE) presented in this section.

Let us consider the value of H obtained in testing conditions, denoted H^*:

$$H^*(W) \equiv \arg\min_H (C_{LS} + C_S) \quad (19)$$

and let us apply the same discriminative criteria $C_d(H^*)$ and $C_l(H^*)$ on that variable. Since H^* is a purely deterministic function of the input with W the only learned parameter, this model can be assimilated to an autoencoder with the encoding step consisting in a complex minimization problem (equation 19) and the decoding step is the usual linear input reconstruction (equation 1). In addition, the discriminative criteria encourage H^* to be a good predictor of Y. The overall model is depicted in Figure 2. The projected gradient descent update for W becomes:

$$W \leftarrow W - \mu \frac{\partial C(H^*)}{\partial W} \tag{20}$$

$$W_{:i} \leftarrow \frac{W_{:i}}{||W_{:i}||} \tag{21}$$

Since $H^*(W)$ is the result of an optimization process, the gradient of $C(H^*)$ with respect to W is not trivial to compute. We can exploit the convergence guarantee of the multiplicative update (6) to express H^* as an infinite sequence truncated to K iterations:

$$H^* = \lim_{k\to\infty} H^k \simeq H^K \tag{22}$$

where:

$$H^{k+1} = H^k \circ \frac{W^T X}{W^T W H^k + \lambda} \tag{23}$$

from which the gradients are easily computed by back-propagation through iteration k in an efficient $O(K)$ time:

$$\frac{\partial C}{\partial H^k} = \frac{\partial C}{\partial H^{k+1}} \circ \frac{H^{k+1}}{H^k} - W^T W B^k \tag{24}$$

for $0 \leq k < K$, where the auxiliary variable B^k is:

$$B^k = \frac{\partial C}{\partial H^{k+1}} \circ \frac{H^{k+1}}{W^T W H^k + \lambda}. \tag{25}$$

The initial conditions are:

$$\frac{\partial C}{\partial H^K} = W^T(WH^K - X) + \lambda + \frac{\partial C_d}{\partial H^K} + \frac{\partial C_l}{\partial H^K} \tag{26}$$

where the two rightmost terms are given by (15) and (16) with $H = H^K$. The gradient with respect to W is then given by:

$$\frac{\partial C}{\partial W} = \sum_{k=0}^{K-1} \Big[X\Big(\frac{\partial C}{\partial H^{k+1}} \circ \frac{H^{k+1}}{W^T X}\Big) - W(B^k H^{kT} + H^k B^{kT})\Big] + (WH^K - X)H^{KT}. \tag{27}$$

When computing $\partial C/\partial W$, the finite-sequence approximation (22) needs only be accurate in the vicinity of the current value of W, denoted W^0. We can increase efficiency without sacrificing precision by initializing $H^0 \equiv H^*(W^0)$ and keeping K small (< 10). Note also that this gradient may become infinite when W is rank deficient, a condition that arises when combinations of basis spectra momentarily align [8]. This optimization issue is alleviated in practice by two facts: the basis spectra are renormalized after each update (equation 21), and the use of a finite sequence to approximate the gradient tends to smooth out singularities.

6. EVALUATION

We use three datasets to evaluate our method:

RAND is a piano dataset of random chords part of the larger MAPS database [6]. Each chord contains from 2 to 7 notes sampled from the whole piano range with heterogeneous loudnesses. We randomly split the data into training, validation and test sets using a 4:1:1 ratio.

ORC is a random polyphonic dataset similar to RAND, but that includes common orchestral instruments such as violin, cello, trumpet, French horn, saxophone, oboe, bassoon, clarinet, flute and piccolo, in addition to piano and organ. Each of the 3000 tracks contains 5 instruments simultaneously playing in their respective range for 16 seconds and was rendered with the FluidR3 SoundFont [1].

MUS is a collection of classical piano pieces also included in MAPS [6], that contains nine sets created by high-quality software synthesizers (7 sets) and a Yamaha Disklavier (2 sets). Five synthesizer sets were selected for training, with the remaining two held out for validation to avoid overfitting the specific piano tones heard during training. We used the first 30 seconds of each piece from the Disklavier sets for test. The average polyphony for this dataset is 2.9.

The magnitude spectrogram was computed for all datasets by the short-term Fourier transform using a 93 ms sliding Blackman window at 10 ms intervals. Each spectrogram frame (column of X) was normalized and square root compressed to reduce the dynamic range. The ground truth Y was directly inferred from the MIDI files [6].

We evaluate multiple pitch estimation performance with the standard metrics of accuracy, precision, recall and F-measure [2]. Either dictionary inspection or linear SVMs using H^* or X as input serve to estimate the pitches. The SVMs can optionally be replaced by multilayer perceptrons (MLP) [16] for comparison. For each NMF model, the parameters are first selected to maximize accuracy on the validation set and we report the final performance on the test set. Parameters are optimized over predetermined search grids on the following intervals:

$q \in [1, 7]$ $\quad \bar{q} \in [0, 12]$ $\quad \eta \in [0, 20]$
$\beta^{\pm} \in [10^{-6}, 10]$ $\quad \alpha \in [10^{-2}, 10^{2}]$
$\lambda \in [10^{-7}, 2]$ $\quad \mu \in [10^{-6}, 10^{-3}]$

7. RESULTS

To illustrate the effectiveness of our approach, we first evaluate qualitatively the learned basis and pitch activities on polyphonic piano data. The dictionary matrices obtained on RAND via unsupervised NMF (Fig. 3(a)) and DNMF (Fig. 3(b)) are presented in Figure 3 after sorting the columns by increasing estimated pitch. From these results, it is clear that DNMF extracted basis spectra – from a purely polyphonic mix – that correspond much closely to the expected spectrum of individual piano notes. It is thus not surprising that applying those dictionaries to extract pitch activities H^* from an excerpt of the MUS test set (Fig. 4(a)) yielded

[1] http://www.hammersound.net

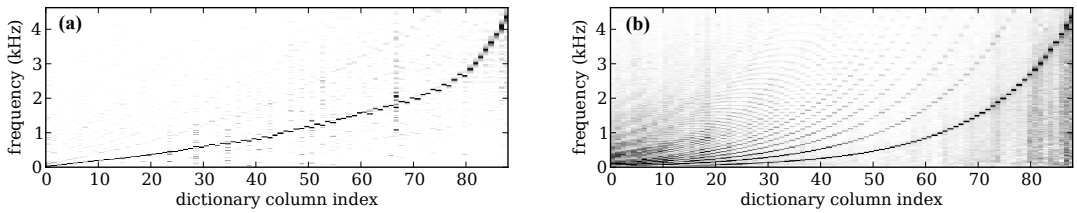


Figure 3. Dictionaries trained ($q = 1$, $\bar{q} = 0$) on the RAND dataset via NMF (a) and DNMF (b). Columns were sorted by increasing estimated pitch for visualization.

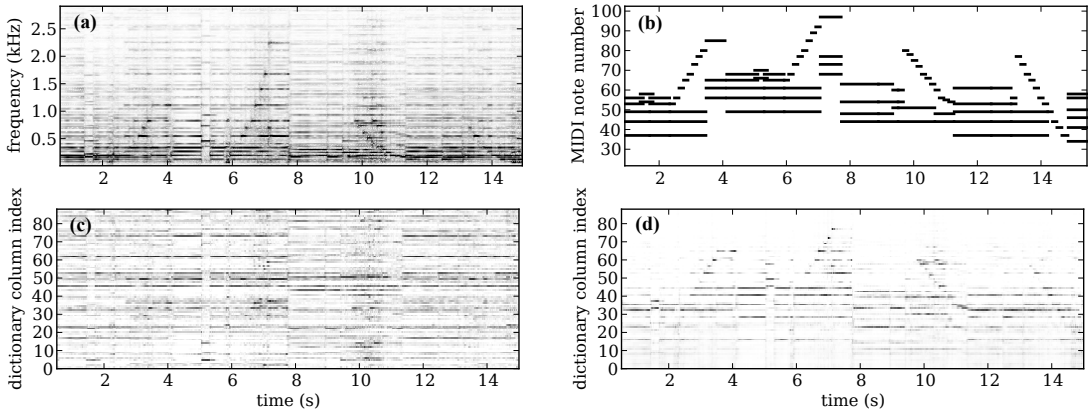


Figure 4. Spectrogram (a) and piano-roll score (b) for the first 15 seconds of an arpeggiated version of *Silent Night, Holy Night* from the MUS test set. Pitch activities H^* (c-d) were estimated for that signal using the pretrained dictionaries in Figure 3(a-b) respectively.

less noisy estimates much closer to the ground-truth score (Fig. 4(b)), as can be observed from Figure 4(c-d).

A more quantitative measure of the discriminative quality of the learned basis is the discriminative ratio r:

$$r(H) = \left(\sum_{i,j|Y_{L(i),j}=1} H_{ij} \Big/ \sum_{i,j|Y_{L(i),j}=0} H_{ij} \right). \quad (28)$$

According to this definition, we obviously favor higher ratios. While r can be made arbitrarily high in training conditions simply by increasing $\beta^{\pm}$, what we really care about is its value in testing conditions $r(H^*)$. Figure 5 shows a significant increase in the test discriminative ratio with our latent variable algorithm compared to the sparse NMF baseline, which indicates a much better pitch label separability. The additional improvement provided by the autoencoder model demonstrates that directly optimizing H^* is useful to increase discriminative performance.

In the next experiments we verify if the discriminative features learned by our models translate in good pitch estimation performance. Frame-level accuracies on the RAND and ORC datasets are presented in Table 1 using dictionary inspection and in Table 2 for multi-label classification. The proposed models outperform the baselines in all cases, especially DNMF-AE used in conjunction with SVMs. Table 3 shows frame-level precision, recall and F-measure results on the MUS test set for common existing NMF variants. Our approach surpasses adaptive unconstrained NMF and is competitive with NMF trained on isolated piano notes and NMF with spectral constraints [20].

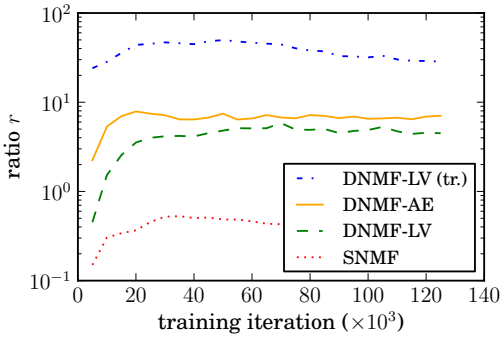


Figure 5. Evolution of the ratio $r(H^*)$ during training on the RAND dataset. "tr" stands for training conditions.

Method	RAND	ORC
NMF	27.6%	30.0%
SNMF	32.3%	43.8%
DNMF-LV	53.2%	**58.8%**
DNMF-AE	**53.4%**	58.6%

Table 1. Multiple pitch estimation accuracy obtained by dictionary inspection on the RAND and ORC datasets.

8. CONCLUSION

We have shown that by exploiting pitch information present in time-aligned musical scores to encourage the extracted features to discriminate against the pitch labels, we can improve the multiple pitch estimation performance of NMF on three datasets of polyphonic music. Interestingly, the

Features	RAND	ORC
Spectrogram	50.9%	55.9%
NMF	56.2%	59.4%
SNMF	55.5%	59.5%
DNMF-LV	60.4%	63.3%
DNMF-AE	**61.6%**	**65.5%**
Spectrogram (MLP)	52.7%	62.0%

Table 2. Multiple pitch estimation accuracy obtained on the RAND and ORC datasets via linear SVMs using the specified feature extraction technique.

NMF variant	Prec.	Rec.	F-meas.
No training			
Unconstrained †	58.9%	60.0%	57.8%
Spectral constraints [20]	71.6%	65.5%	67.0%
Pretrained dictionary			
Isolated note spectra †	68.6%	66.7%	66.0%
Proposed (DNMF-LV)	68.1%	65.9%	66.9%
Proposed (DNMF-AE)	66.8%	**68.7%**	**67.8%**
Other methods			
SONIC [16]	**74.5%**	57.6%	63.6%

Table 3. Average multiple pitch estimation performance of common NMF variants on the MUS (MAPS) piano dataset. †These results are from Vincent [20].

resulting basis spectra closely resemble the spectrum of individual piano notes, even though they were trained on purely polyphonic data without explicit harmonicity constraints. Once that discriminative basis is learned, relevant pitch activity features can be efficiently computed using only standard multiplicative updates.

9. ACKNOWLEDGMENTS

The authors would like to thank NSERC, CIFAR and the Canada Research Chairs for funding.

10. REFERENCES

[1] S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Trans. on Neural Networks*, 17(1):179–196, 2006.

[2] M. Bay, A.F. Ehmann, and J.S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *ISMIR*, 2009.

[3] A. Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *ISMIR*, 2006.

[4] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *ISMIR*, 2010.

[5] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts. In *NIPS 16*, 2003.

[6] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2010.

[7] D. Fitzgerald, M. Cranitch, and E. Coyle. Generalised prior subspace analysis for polyphonic pitch transcription. In *DAFX 8*, 2005.

[8] G.H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, pages 413–432, 1973.

[9] N. Guan, D. Tao, Z. Luo, and B. Yuan. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Transactions on Image Processing*, 20(7):2030–2048, 2011.

[10] P.O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing*, pages 557–565, 2002.

[11] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural computation*, 19(3):780–791, 2007.

[12] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[13] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS 13*, 2001.

[14] C.J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, pages 1–8, 2008.

[16] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, 2004.

[17] G.E. Poliner and D.P.W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, 2007(1):154–164, 2007.

[18] P. Smaragdis. Polyphonic pitch tracking by example. In *IEEE WASPAA*, pages 125–128, 2011.

[19] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE WASPAA*, pages 177–180, 2003.

[20] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, 2010.

[21] Y. Wang and Y. Jia. Fisher non-negative matrix factorization for learning local features. In *ACCV*, 2004.

[22] C. Yeh and A. Röbel. The expected amplitude of overlapping partials of harmonic sounds. In *ICASSP*, 2009.

[23] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks*, 17(3):683–695, 2006.