Automatic Music Transcription Using Low Rank Non-Negative Matrix Decomposition

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*—Automatic Music Transcription (AMT) is concerned with the problem of producing the pitch content of a piece of music given a recorded signal. Many methods rely on sparse or low rank models, where the observed magnitude spectra are represented as a linear combination of dictionary atoms corresponding to individual pitches. Some of the most successful approaches use Non-negative Matrix Decomposition (NMD) or Factorization (NMF), which can be used to learn a dictionary and pitch activation matrix from a given signal. Here we introduce a further refinement of NMD in which we assume the transcription *itself* is approximately low rank. The intuition behind this approach is that the total number of distinct activation patterns should be relatively small since the pitch content between adjacent frames should be similar. A rank penalty is introduced into the NMD objective function and solved using an iterative algorithm based on Singular Value thresholding. We find that the low rank assumption leads to a significant increase in performance compared to NMD using β-divergence on a standard AMT dataset.**

I. Introduction

Sparse and low rank models [1][2] have seen a lot of interest in signal processing. A challenging problem for such methods is that of Automatic Music Transcription (AMT), which is the task of isolating and enumerating the pitches present in a recorded music signal. The nature of music presents several difficulties which are in general not well addressed by standard matrix factorization approaches. For example, exemplar dictionaries built using isolated pitches have been shown to be highly correlated which can make them difficult to separate [3]. Another issue is that nearly all common techniques which rely on factorizing a given signal implicitly treat each frame as independent. This assumption is clearly violated by musical signals in which the pitch content is (locally) smooth.

In this work, we show how AMT performance can be improved by placing an additional low rank assumption on the activation matrix, as part of an Non-negative Matrix Factorization (NMF) or Non-negative Matrix Decomposition (NMD) model. We argue that this constraint is sensible, since although any activation matrix may have many thousands of columns, the ground truth activation tends to have low rank structure. Classical matrix factorizations consider each audio frame independently, potentially leading to rapidly changing supports which is not present in the groundtruth.

Here we focus on the problem of estimating the active pitches given a *fixed* pitch dictionary, noting that having established the merits of the proposed approach it can be easily extended to the case where the dictionary is learned at the same time. The structure of this paper is a follows: in Section II we describe the use of NMD for AMT and its potential shortcomings. In Section III we introduce the proposed low-rank approach based on the standard NMD non-negative matrix updates and singular value thresholding. Section IV outlines an experiment using a popular AMT dataset, where we show that the rank constrained approach is competitive with related published work.

II. Automatic Music Transcription using Non-negative Matrix Decomposition

Given a matrix $\mathbf{S}$ whose columns correspond to magnitude spectra of short-time audio frames and a dictionary $\mathbf{D}$ of isolated pitches, Non-negative Matrix Decomposition (NMD) seeks a factorization of the form

$$\mathbf{S} \approx \mathbf{DC} \tag{1}$$

which minimizes some distortion or distance measure $d(\mathbf{S}, \mathbf{D}\,\mathbf{C})$ between the observed signals $\mathbf{S}$ and the reconstruction in the dictionary given by $\mathbf{DC}$ with non-negativity constraints on all the matrix elements. For the purposes of this work, we distinguish NMD – which seeks a factorization over a *given* dictionary – from the standard Non-negative Matrix Factorization (NMF) – which seeks both the activations *and* a (non-negative) dictionary $\mathbf{D}$. While the approach outlined in this paper can be easily applied to NMF, we chose to deal solely with the NMD case in order to isolate the effect of the proposed modified $\mathbf{C}$ updates. In both cases, the non-negativity constraints are natural for AMT since we are modelling the observed magnitude spectrum as a linear combination of individual pitches.

As a similarity measure, we consider the family of β-divergences which have been shown to perform well in AMT. For matrices $\{\mathbf{X}, \mathbf{Y}\} \in \mathbb{R}^{m \times n}$, define $d_\beta(\mathbf{X}, \mathbf{Y}) =$

ISBN 978-0-9928626-7-1 © EURASIP 2017

$\sum_{i=1}^{m}\sum_{j=1}^{n} d_\beta(x_{ij}, y_{ij})$ where

$$d_\beta(x,y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) & \beta \in \mathcal{C} \\ x\log\frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{2}$$

and $\mathcal{C}$ is the set $\mathbb{R}_+ \setminus \{0,1\}$. In fact, the generalized β-divergence includes several well-known similarity measures as special-cases, such as Euclidean distance ($\beta = 2$), Kullback-Leibler divergence ($\beta = 1$) and the Itakura-Saito divergence ($\beta = 0$). For signals $\mathbf{S} \in \mathbb{R}_+^{d\times n}$ and a fixed dictionary $\mathbf{D} \in \mathbb{R}_+^{d\times k}$, the β-NMD problem seeks the solution to

$$\min_{\mathbf{C}} \quad d_\beta(\mathbf{S}, \mathbf{DC}) \quad \text{such that} \quad \mathbf{C} \in \mathbb{R}_+^{k\times n}.$$

A suitable $\mathbf{C}$ satisfying the constraint can be found using the multiplicative update rule [1] [4]

$$\mathbf{C} \leftarrow \mathbf{C} \odot \left(\frac{\mathbf{D}^T(\mathbf{S} \odot (\mathbf{DC})^{\beta-2})}{\mathbf{D}^T(\mathbf{DC}^{\beta-1})} \right) \tag{3}$$

where $\odot$ is the Hadamard product and the division/exponentiation are applied entry-wise. This update rule is basically a gradient descent scheme with a judiciously chosen step size – by writing the update step in terms of a multiplication instead of addition (as in standard gradient descent), all the terms will remain non-negative throughout. It has been shown that this rule results in a consistent decrease in the objective function for a range of β values.

A weakness of current NMD/NMF methods is that individual frames are treated independently. In the case of AMT this is a poor assumption since in general adjacent frames will share similar pitch content. To combat this, some works have attempted to apply smoothness constraints on the activations post-hoc using additional processing steps such as Hidden Markov Models, in which the obtained activations are treated as observables of the true hidden (pitch) state [5]. Still others have explored probabilistic generalizations of NMF which include assumptions about the smoothness of the activations (i.e. the rows of $\mathbf{C}$ should be smooth) [6] or have investigated the use of Recurrent Neural Networks which take into account the previous states during the inference [7][8].

III. Low Rank Non-Negative Matrix Decomposition

The β-NMD method can be used to produce a fixed-rank estimate of the initial signal matrix $\mathbf{S}$ based on the dictionary $\mathbf{D}$ and coefficients $\mathbf{C}$ (i.e. the rank is of the reconstructed matrix is equal to the number of columns of $\mathbf{D}$). For AMT however, we note that the resulting coefficient matrix $\mathbf{C}$ has additional structure not exploited by traditional NMF/NMD algorithms; namely time smoothness. In this context we expect that (i) once activated, pitches remain activated for a significant amount of time and (ii) adjacent frames should be similar. The goal of this work is to use this intuition as part of the algorithm in order to produce 'smoother' estimates $\mathbf{C}$. As a motivating example, Fig. 1 displays the activation matrix for a short excerpt of a piece of classical piano music using β-NMD (top), the activation with spurious elements removed (middle) and just the isolated spurious elements (bottom). While much of the true pitch content is successfully captured, we observe many spurious activations where harmonically related atoms have entered the model. Our goal is to remove the noisy activations while keeping the true activations by assuming that the activation matrix $\mathbf{C}$ is *also* low-rank. Using this assumption, we can model the observed activations as consisting of a low-rank part (corresponding to the true activation patterns) together with a "noisy" component (corresponding to short-time spurious activations) and therefore we can improve performance by removing the short-time components in $\mathbf{C}$. We will refer to this approach as *Low Rank Non-negative Matrix Decomposition* (LR-NMD).

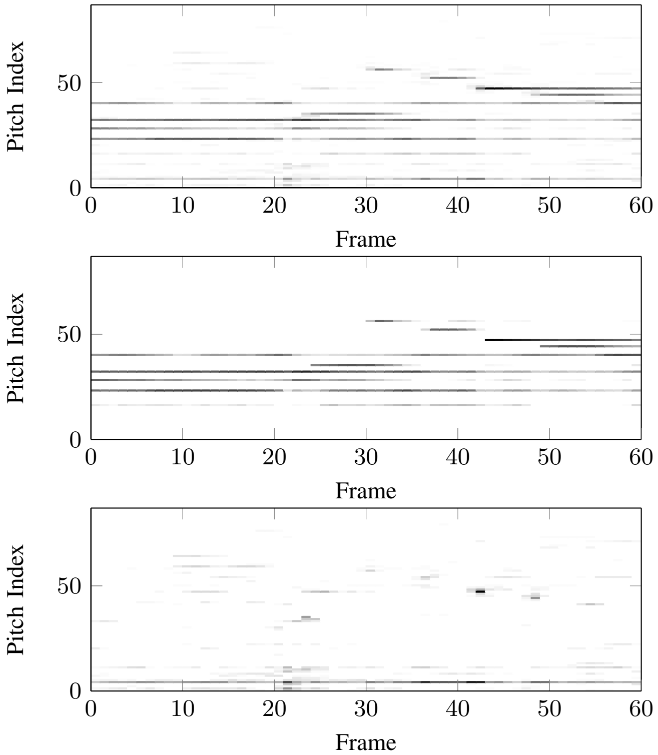

Fig. 1. Transcription matrix $\mathbf{C}$ using β-NMD (top), only the ground-truth pitches (middle) and spurious activations (bottom).

Practically, given a signal matrix $\mathbf{S}$ and a dictionary $\mathbf{D}$, the problem aims to solve

$$\min_{\mathbf{C}\in\mathbb{R}_+^{k\times n}} \quad d_\beta(\mathbf{S}, \mathbf{DC}) \quad \text{such that} \quad \operatorname{rank}(\mathbf{C}) \le k \tag{4}$$

where $d_\beta(\mathbf{S}, \mathbf{DC})$ is the measure-of-fit between the signals and their reconstructions given by the β-divergence outlined above.

To make the optimization easier, following [9] we can relax the rank constraint on $\mathbf{C}$ using the nuclear-norm defined by

$$\|\mathbf{C}\|_* = \text{trace}\left(\sqrt{\mathbf{C}^T\mathbf{C}}\right) = \sum_{i=1}^{m} \sigma_i(\mathbf{C}) \tag{5}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

where $m = \min\{k, n\}$ and $\{\sigma_i(\mathbf{C})\}_{i=1}^{m}$ are the singular values of $\mathbf{C}$. This gives the final proposed LR-NMD optimization problem

$$\min_{\mathbf{C}\in\mathbb{R}_+^{k\times n}} \quad d_\beta(\mathbf{S}, \mathbf{DC}) + \lambda\|\mathbf{C}\|_* \tag{6}$$

where λ is a regularization constant which must be chosen. In effect, this corresponds to an ℓ_1-norm penalty on the singular values of $\mathbf{C}$ which encourages sparsity (i.e. most of the singular values will be 0) and hence $\mathbf{C}$ will be low-rank.

We solve this optimization using an iterative scheme; at each step we update $\mathbf{C}$ using the β-NMD rule given in equation (3) followed by a Singular Value Thresholding operation [10]. Using its Singular Value Decomposition, $\mathbf{C}$ can then be written as

$$\mathbf{C} = \mathbf{U\Sigma V}^T \tag{7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values. Writing $\boldsymbol{\sigma} = \text{diag}(\mathbf{\Sigma})$ we apply an element-wise shrinkage and thresholding operation $\sigma_i \leftarrow \max(\sigma_i - \lambda, 0)$ to $\boldsymbol{\sigma}$ and let $\mathbf{C} = \mathbf{U}\ \text{diag}(\boldsymbol{\sigma})\ \mathbf{V}^T$. This is essentially a projected (or proximal) gradient descent scheme [11] similar to Iterative Shrinkage and Thresholding (ISTA) [12] and other proximal gradient techniques seen in signal processing and compressed sensing, but applied in the context of non-negative matrix factorization (in fact, the singular value thresholding operator is the proximity operator for the nuclear-norm [10]). One problem is that after thresholding, small negative values can appear in $\mathbf{C}$ and therefore after each iteration we simply set these values to 0.

We also found a small performance increase by using a slightly modified update rule for $\mathbf{C}$: while the nuclear norm is non-diffentiable, it has a well-defined subgradient [13] with $\mathbf{UV}^T \in \partial\|\mathbf{C}\|_*$ [14]. Incorporating this into the NMD update gives

$$\mathbf{C} \leftarrow \mathbf{C} \odot \left(\frac{\mathbf{D}^T\big(\mathbf{S} \odot (\mathbf{DC})^{\beta-2}\big) - \lambda\mathbf{M}^-}{\mathbf{D}^T(\mathbf{DC}^{\beta-1}) + \lambda\mathbf{M}^+} \right) \tag{8}$$

where $\mathbf{M}^+$ and $\mathbf{M}^-$ are respectively the matrices formed using just the positive and negative entries of $\mathbf{UV}^T$. This update rule minimizes (6) using sub-gradient descent, but resulted in only a minor improvement without the singular value thresholding strategy. The full LR-NMD algorithm is outlined in Fig. 2.

```
Input: Signals $\mathbf{S}$, dictionary $\mathbf{D}$, regularizer $\lambda$,
    $\beta$-parameter.
Output: Activation matrix $\mathbf{C}$.
    Initialization: Initialize $\mathbf{C}$ using small non-negative
    values.
while not converged do
    Update $\mathbf{C}$ using (8).
    Compute the SVD: $\mathbf{U\Sigma V}^T = \mathbf{C}$.
    Update $\mathbf{\Sigma}$ by singular value thresholding:
        $\mathbf{\Sigma} \leftarrow \text{thresh}_\lambda(\mathbf{\Sigma})$.
    Update $\mathbf{C} \leftarrow \mathbf{U\Sigma V}^T$.
    Set any negative elements in $\mathbf{C}$ to 0.
end while
return $\mathbf{C}$.
```

Fig. 2. Proposed algorithm.

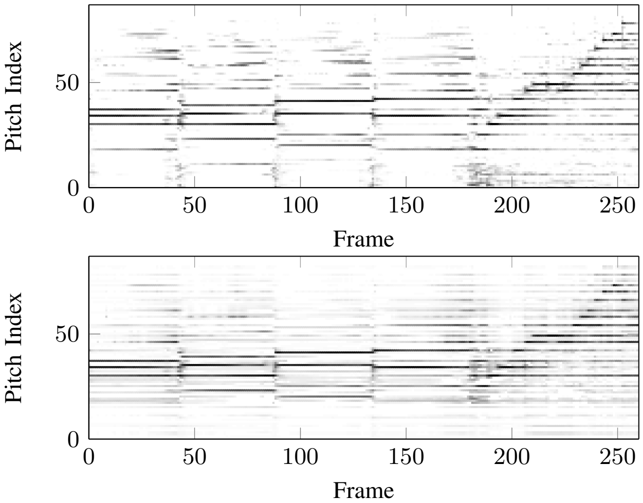

Fig. 3. Transcription results using β-NMD (top) and propsed LR-NMD (bottom) with a fixed pitch dictionary.

IV. Experiment

The algorithm was evaluated using the MIDI Aligned Piano Sounds (MAPS) dataset [15], which consists of recording of western classical piano music together with ground-truth MIDI transcriptions. This set contains several subsets corresponding to different piano types and recording environments and for testing purposes we chose the "EnStDkcl" subset. For each recording we took the first 30-seconds and processed it using an Equivalent Rectangular Bandwidth (ERB) transform on 25-windows and a frequency resolution of 250 bins. The ERB is perceptually motivated and has been shown to perform well for the AMT problem [3]. MAPS also contains isolated recordings of individual piano keys which we use to build the pitch dictionary by learning one atom per-pitch. This results in a dictionary of size 250 × 88, with each column corresponding to a single pitch.

The transcription quality was evaluated using frame-wise $\mathcal{F}$-measure which is a common metric for AMT (see for example [5] [16] [17]). For each recording, we decomposed it over the learned dictionary using both β-NMD and LR-NMD with $\beta = 0.5$ (this value has been found to work best for AMT). The target transcription consists of a binary time-pitch matrix, with a 1 indicating the presence of a pitch and a 0 indicating absence. To binarize the output transcription matrix $\mathbf{C}$, we threshold it by setting to zero any element with value less than $m + c\sigma$ (and 1 otherwise), where m and σ are the mean and standard deviation of the activations for that column and c is a constant. Other thresholding strategies have been used in the literature, for example based on the maximum activation in $\mathbf{C}$, but we didn't find a meaningful difference in the performance in terms of $\mathcal{F}$-measure of either approach.

We count the number of true-positive (tp), false-positive (fp) and false-negative (fn), from which we calculate the precision ($\mathcal{P}$), recall ($\mathcal{R}$) and $\mathcal{F}$-measure:

$$\mathcal{P} = \frac{\text{tp}}{\text{tp}+\text{fp}}, \quad \mathcal{R} = \frac{\text{tp}}{\text{tp}+\text{fn}}, \quad \mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}}. \tag{9}$$

A. Results and Discussion

The results for each method are summarized in Fig. 4, where we see find that the low-rank penalty considerably improves performance. In Fig. 2 we compare the resulting outputs of standard β-NMD against LR-NMD. As expected, we find that the low-rank approach shows less short-time components. In particular, the short time components have been considerably 'smoothed' in the low-rank case. We also note that the iterative scheme is important, giving much better results than simply applying the thresholding as a post-processing step.

To test the effect of the hyperparameters c and λ we created a smaller dataset of 5 tracks and varied the values of λ and c. The results are presented in Figures 5 and 6. While some care must be taken to set these values for best performance, in general LR-NMD consistently outperforms β-NMD across a range of reasonable values.

Overall we find that the low rank assumption results in a robust improvement. An absolute $\mathcal{F}$-measure improvement of 2.25 over β-NMD is non-trivial for AMT and is similar to the gains using other techniques, for example the 1.8 improvement reported in [3] using a per-frame dictionary conditioning step with an ERB transform of the same dimension and the same dataset. In Fig. 4 we also include results (as reported in [18]) on "EnStDkcl" obtained using a Probabalistic Latent Component Analysis (PLCA) method by Benetos et al [19] and the Harmonic Non-negative Matrix Decomposition (H-NMD) of Vincent et al [20]. The approach of Cheng et al [18] represents the state-of-the-art for this task using NMD at 79.01 reported $\mathcal{F}$-measure; while also relying on β-divergence, they consider a more complicated model with onsets and decays treated separately, together with dynamic-programming to infer the activations.

Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
H-NMD [20]	-	-	58.84
PLCA [19]	-	-	67.79
Wβ-NMF [3]	-	-	73.70
β-NMD [21]	73.31	69.31	71.25
LR-NMD (proposed)	73.83	73.17	73.50

Fig. 4. Framewise $\mathcal{F}$-measure AMT results for the MAPS "EnStDkcl" subset. Italics indicate results taken directly from the relevant publication. Dashes indicate that the given metric was not reported.

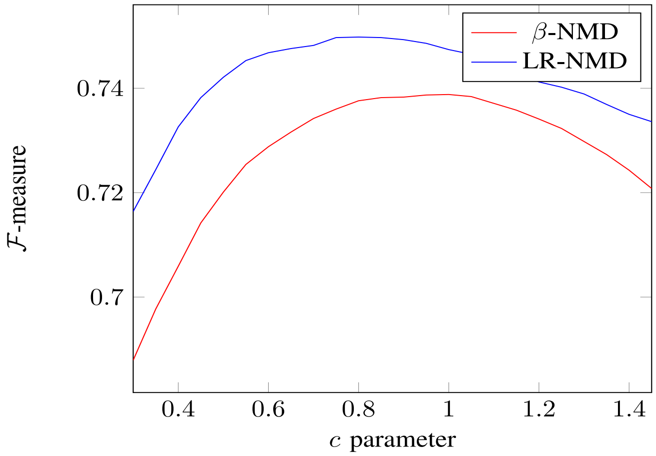

Fig. 5. AMT results for different threshold parameters c.

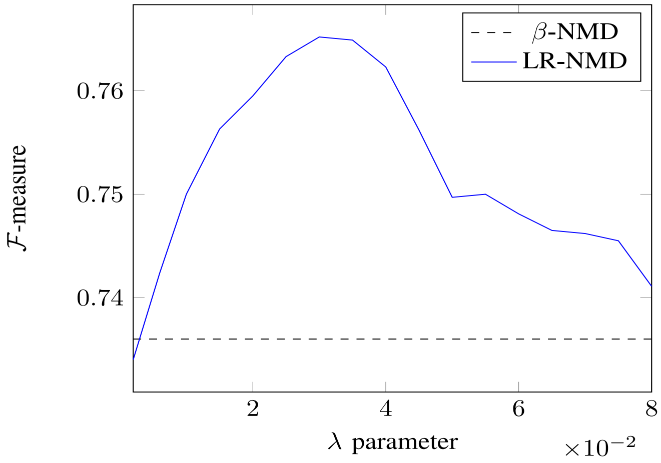

Fig. 6. AMT results for different λ.

V. Conclusion

We proposed a low-rank model for Automatic Music Transcription, based on the idea that a good transcription should be free from short-time spurious elements. To achieve this we introduced an augmented update rule for β-NMD, which iterates between non-negative multiplicative updates and singular value thresholding. The proposed approach gives improved performance for the AMT task and is simple to implement, which makes it easy to add to existing NMF-based systems.

For future work, similar ideas could be applied to more complicated NMF AMT algorithms involving group sparsity [22] or non-negative dictionary learning. So far, we have used a setting of 0.5 for β which has previously found to work well for AMT, however it is possible that the performance could be further improved by fine-tuning the β parameter for the low-rank approach. Another avenue is incorporating low-rank updates into an NMF model, for example using the Wβ-NMF dictionary updates.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under Grant Agreement n$^{\text{O}}$ 607290 SpaRTaN.

REFERENCES

[1] 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.

[2] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[3] 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.

[4] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[5] T. B. Yakar, P. Sprechmann, R. Litman, A. M. Bronstein, and G. Sapiro, "Bilevel sparse models for polyphonic music transcription." in *14th International Society for Music Information Retrieval Conference*, 2013, pp. 65–70.

[6] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[7] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, speech and signal processing (ICASSP), 2012*. IEEE, 2012, pp. 121–124.

[8] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 927–939, 2016.

[9] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[10] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[11] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011, pp. 185–212.

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[13] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *Lecture Notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, 2003.

[14] G. A. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, vol. 170, pp. 33–45, 1992.

[15] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[16] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1727–1741, 2013.

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

[18] T. Cheng, M. Mauch, E. Benetos, S. Dixon *et al.*, "An attack/decay model for piano transcription," in *17th International Conference on Music Information Retrieval (ISMIR)*, 2016.

[19] E. Benetos, T. Weyde *et al.*, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," pp. 701–707, 2015.

[20] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[21] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *ISMIR-11th International Society for Music Information Retrieval Conference*, 2010, pp. 489–494.

[22] K. O'Hanlon and M. D. Plumbley, "Polyphonic piano transcription using non-negative matrix factorisation with group sparsity," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 3112–3116.

ISBN 978-0-9928626-7-1 © EURASIP 2017