Non-Negative Group Sparsity with Subspace Note Modelling for Polyphonic Transcription

Ken O'Hanlon, Hidehisa Nagano,
Nicolas Keriven and Mark D. Plumbley

To cite this paper :
Ken O'Hanlon, Hidehisa Nagano, Nicolas Keriven and Mark D. Plumbley
IEEE/ACM Transactions on Audio, Speech and Language Processing,
Vol. 23 (3), pp.530 - 542, 2016

DOI: 10.1109/TASLP.2016.2515514
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7384716

Non-Negative Group Sparsity with Subspace Note Modelling for Polyphonic Transcription

Ken O'Hanlon, *Member, IEEE,* Hidehisa Nagano, *Senior Member, IEEE,*
Nicolas Keriven, *Student Member, IEEE,* Mark D. Plumbley, *Fellow, IEEE*

***Abstract*—Automatic Music Transcription (AMT) can be performed by deriving a pitch-time representation through decomposition of a spectrogram with a dictionary of pitch-labelled atoms. Typically, Non-negative Matrix Factorisation (NMF) methods are used to decompose magnitude spectrograms. One atom is often used to represent each note. However, the spectrum of a note may change over time. Previous research considered this variability using different atoms to model specific parts of a note, or large dictionaries comprised of datapoints from the spectrograms of full notes. In this paper the use of subspace modelling of note spectra is explored, with group sparsity employed as a means of coupling activations of related atoms into a pitched subspace. Stepwise and gradient-based methods for non-negative group sparse decompositions are proposed. Finally, a group sparse NMF approach is used to tune a generic harmonic subspace dictionary, leading to improved NMF-based AMT results.**

***Index Terms*—Group sparsity, automatic music transcription, non-negative matrix factorisation, stepwise optimal**

I. Introduction

AUTOMATIC Music Transcription (AMT) seeks to derive pitch-time activations from a musical signal. Spectrogram factorisations provide one approach to this problem, and are particularly appropriate when the signal is comprised of instruments with fixed pitch, such as a piano. Often, in audio signal processing a magnitude, or power, spectrogram is used with methods based on Non-negative Matrix Factorisation (NMF) [1]. In this case, NMF seeks to approximate the non-negative spectrogram $\mathbf{S} \in \mathbb{R}_+^{M \times N}$ such that

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

where $\mathbf{D} \in \mathbb{R}_+^{M \times K}$ is a dictionary matrix, with an atom, $\mathbf{d}_k$, in each column and $\mathbf{X} \in \mathbb{R}_+^{K \times N}$ is an activation matrix in which each row, $\mathbf{x}^k$, relates the activations of the corresponding atom, $\mathbf{d}_k$. When the atoms are pitch labelled a pitch-time representation can be derived from the activation matrix.

NMF is an unsupervised learning algorithm which typically uses multiplicative updates to perform the approximation (1). NMF was first proposed for AMT by Smaragdis and Brown [2]. They considered that each note in a signal may need to be played in isolation at least once in order to learn a meaningful atom for that note [2] as overlap of signal elements in the spectrogram, common in musical signals, presents difficulty in separation of factors. While NMF may separate notes that are not played in isolation [3], there is a tendency to learn atoms that are not meaningful, with energy concentrated in few dimensions [3] [4] while the selected learning order, or factorisation rank, K, is seen to effect AMT performance [3].

Supervised NMF, or Non-negative Matrix Decomposition (NMD) [5], using a fixed dictionary, provides a means to perform AMT that avoids problems associated with NMF. However, NMD performance degrades if a dictionary is not suited to the signal [4]. Typically NMD is performed with one atom used to model each note [5]. Improved AMT using multi-atom note modelling is reported in [6], where it is suggested that using several atoms better captures variation in spectral shape over the duration of a note. Similarly, different atoms are used to model the attack, sustain and decay states of a piano note with Hidden Markov Models used to determine transitions between states [7] [8]. Alternatively, use of low-rank subspaces, learnt offline using NMF, to model notes is considered in [4]; however degraded performance is reported.

An alternative approach to note modelling for NMD is taken in [9] where a dictionary comprised of the frames of isolated note spectrograms is used in order to capture the variability in the note spectrum. This datapoint dictionary is overcomplete ($K > M$), and Orthogonal Matching Pursuit (OMP) [10] is used to decompose a spectrogram. Difficulty in selecting an appropriate stopping condition for OMP in this context is identified [9], while broken temporal continuity in spectrogram decompositions using greedy pursuits is reported in [11]. However, the potential advantage of stepwise pursuits in the case of multi-instrument signals is noted in [12].

Harmonic variants of NMF [4] [13] [14] constrain the learning in order to learn meaningful atoms and avoid the rank selection problem. These approaches initialise with one atom, estimating an expected note spectrum, specified for each note. Racinski et al [13] place zeroes at all positions of an atom not expected to contain a harmonic partial of the associated note. The zeroes, and harmonic structure, are maintained by multiplicative updates used in NMF. Vincent et al [4] propose a semi-supervised NMF approach using a hierarchical dictionary, in which each high-level harmonic atom is defined by learning a superposition of several low-level fixed narrowband atoms sharing the same fundamental frequency. While this method is considered state-of-the-art for NMF-based AMT, we consider that the harmonic constraint may be over-restrictive, particularly in the case of semi-percussive instruments such as the piano.

Ken O'Hanlon is with the Centre for Digital Music, Queen Mary University of London. Nicolas Keriven is with INRIA Rennes-Bretagne Atlantique. Hidehisa Nagano is with NTT Communications Science Laboratories, NTT Corporation. Mark Plumbley is with the Centre for Vision Speech and Signal Processing, University of Surrey. All authors were at the Centre for Digital Music, Queen Mary when this research was performed.

This research was funded by EPSRC Platform Grant EP/K009559/1, EPSRC Grant EP/L027119/1 and EPSRC Grant EP/J010375/1.

A. Contributions of this paper

In this paper, the use of subspace models of note spectra is considered, whereby a note is represented by a group of atoms, such as seen in Fig. 3 and Fig. 5, that may be co-active and have no explicit temporal dependencies. We consider that negative results reported for this model [4] are due to the lack of a strategy to couple groups of atoms into pitched subspaces. We propose to use group sparsity for this purpose, and develop a suite of non-negative group sparse algorithms. OMP-based methods are first proposed [15], but noted problems with this class of approaches in this context [9] [11] lead us to propose an alternative stepwise method, employing backwards elimination [16]. We previously proposed these approaches in [15] [16], and offer a direct comparison here alongside a further comparison of subspace and datapoint modelling [9].

Group sparsity is then extended to NMF with β-divergence. We propose a novel group sparse penalty that scales in a linear fashion to β-divergence, for which we provide an auxiliary function that affords a monotonic group sparse NMF algorithm. We then employ this approach in a dictionary tuning method, applied to a restructured version of the harmonic dictionary used in [4], whereby the hard harmonic constraint is dropped. Part of this work was described in [17] and is augmented here through monotonic descent algorithm with scale invariant group sparse penalty and further evaluation. We also propose a new onset detector for NMF-based AMT.

In the next section some relevant background information and baseline methods are briefly described. Following this, the proposed group sparse methods are outlined in section III. Section IV introduces the dictionary tuning approach, before evaluation of all proposed approaches is given in section V. Finally the paper concludes with pointers to further work.

II. Background and Baseline Methods

A. Group Sparsity

SPARSE approximation seeks a signal representation that is predominated by zeros. Given a signal, $\mathbf{s} \in \mathbb{R}^M$, and a dictionary, $\mathbf{D} \in \mathbb{R}^{M \times K}$, with unit ℓ_2 norm atoms in each column, $\mathbf{d}_k$, the sparse approximation problem is defined as a penalised least squares problem

$$\mathbf{x} \leftarrow \min_{\mathbf{x}} \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_0 \tag{2}$$

where $\|\mathbf{x}\|_0 = |\mathbf{x} \neq 0|$ is referred to as the ℓ_0 pseudonorm and λ is a parameter controlling the sparsity of the representation. Different approaches may be used to approximate (2). Greedy methods such as OMP [10], outlined in Fig. 1, form a representation by iteratively adding the atom most correlated with the residual signal, $\mathbf{r}$, to the sparse support, Γ. The supported atoms, indexed by $\mathbf{D}_\Gamma$, are projected onto the signal, giving the interim coefficients, $\mathbf{x}_\Gamma$, which are used to recalculate the residual. This iteration is performed until a predefined stopping condition such as residual energy level, or a number of selected atoms, is met. An alternative to matching pursuits is to replace the ℓ_0 penalty in (2), considered a difficult problem, with an ℓ_1 norm, where $\|\mathbf{x}\|_1 = \sum_n |x_n|$. This approach, referred to as ℓ_1 minimisation or Basis Pursuit Denoising [18], allows approximation of (2) with convex optimisation methods.

- *Input* : $\mathbf{D} \in \mathbb{R}^{M \times N}$; $\mathbf{s} \in \mathbb{R}^M$
- *Initialise* : $\mathbf{r} = \mathbf{s}$; $\Gamma = \{\}$
- Repeat
  - *Select atom with index*
  $$\hat{k} = \arg\max_k |\langle \mathbf{d}_k, \mathbf{r} \rangle| \tag{3}$$
  - *Add to support*
  $$\Gamma = \Gamma \cup \hat{k} \tag{4}$$
  - *Backproject support onto signal*
  $$\mathbf{x}_\Gamma \leftarrow \min_{\mathbf{x}} \|\mathbf{s} - \mathbf{D}_\Gamma \mathbf{x}\|_2^2 \tag{5}$$
  - *Calculate new residual*
  $$\mathbf{r} = \mathbf{s} - \mathbf{D}_\Gamma \mathbf{x}_\Gamma \tag{6}$$
- Until stopping condition met

Fig. 1: Orthogonal Matching Pursuit.

Group sparse representations incorporate the assumption that certain atoms tend to be active together, as demonstrated in Fig. 2. Given the set $\mathcal{J} = \{\mathcal{J}^j\}$, where $\mathcal{J}^j$ contains the indices of the jth group, the notation

$$\begin{aligned} \mathbf{D}[j] &= [\mathbf{d}_{\mathcal{J}^j(1)}, ..., \mathbf{d}_{\mathcal{J}^j(|\mathcal{J}^j|)}] \\ \mathbf{x}[j] &= [x_{\mathcal{J}^j(1)}, ..., x_{\mathcal{J}^j(|\mathcal{J}^j|)}]^T \end{aligned}$$

is used for the jth group of the dictionary, $\mathbf{D}[j]$, and of the coefficient vector, $\mathbf{x}[j]$, where $\mathcal{J}^j(i)$ is the ith member of the jth set of indices. The notation $\mathbf{x}[j, i]$ is used to refer to the ith member of the jth group of $\mathbf{x}$.

Group sparse variants of OMP replace (3) with a group selection criteria, and add all atoms in the selected group, indexed by $\hat{j}$, to the support : $\Gamma = \Gamma \cup \mathcal{J}^{\hat{j}}$. The most well-known group sparse greedy method is Block-OMP (B-OMP) [19] which uses the selection criterion:

$$\hat{j} = \arg\max_j \|\phi[j]\|_2 \tag{7}$$

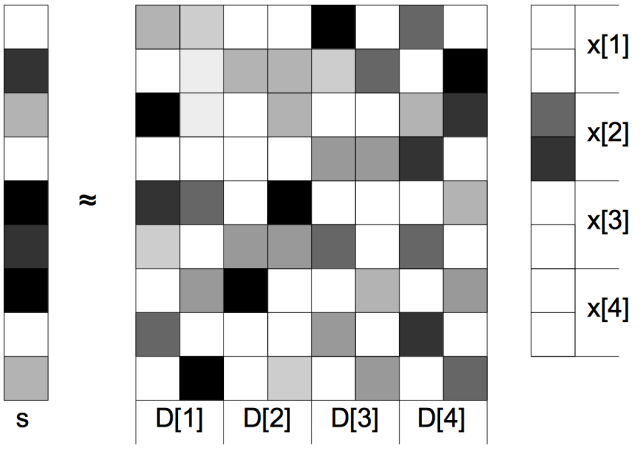

Fig. 2: Graphical description of the group sparse problem showing dictionary with groups notated $\mathbf{D}[j]$ and one active group $\mathbf{x}[2]$. White blocks denote zeroes.

where $\phi[j] = \mathbf{D}[j]^T\mathbf{r}$. Subspace Matching Pursuit [20] is a similar approach which uses the selection criterion:

$$\hat{j} = \arg\min_j \|\mathbf{r} - \pi_j(\mathbf{r})\|_2 \tag{8}$$

where $\pi_j(\mathbf{r})$ is the projection of $\mathbf{r}$ onto the subspace $\mathbf{D}[j]$.

The group sparse problem can also be considered using a penalised least squares approach. In this case a mixed-norm $\ell_{p,q}$ penalty is employed where

$$\|\mathbf{x}\|_{p,q} = \left(\sum_j \left(\sum_{i \in \{1,\ldots,|\mathcal{J}^j|\}} x[j,i]^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}. \tag{9}$$

In the case where it is expected that few groups are active, with many atoms active in each group, an $\ell_{p,0}$ norm is considered ideal, typically with $p = 2$ [19]. Similar to Basis Pursuit Denoising [18], relaxation using an $\ell_{p,1}$ penalty is considered [19]. Other mixed norm penalties, such as the $\ell_{1,2}$ term proposed in [21] which seeks to have many groups active with few atoms in each group supported, are known. A list of some of these penalties is given in [22].

B. Non-negative methods

Spectrogram decompositions are often performed on non-negative spectra with a non-negative constraint applied to the dictionary and activations. Stepwise methods such as OMP require modification to explicitly accommodate this constraint [23]. The least squares backprojection (5) is replaced with Non-Negative Least Squares (NNLS) :

$$\mathbf{x} \longleftarrow \min_{\mathbf{x}} \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0. \tag{10}$$

NNLS is a well studied problem for which many different methods have been proposed [24]. The classic NNLS algorithm [25] is a greedy stepwise algorithm, similar to OMP, that considers a positive only selection criteria

$$\hat{k} = \arg\max_k \mathbf{d}_k^T\mathbf{r} \tag{11}$$

and backprojects using an iterative loop. In each iteration a least squares projection is performed and atoms displaying a negative coefficient are ejected from the active set, Γ. Iterations continue until the non-negative constraint is met. NNLS possesses a natural stopping condition that no inactive atoms have a positive correlation with the residual. Non-negative OMP (NN-OMP) [23], apart from the stipulation of normalised atoms, can be considered a truncated NNLS algorithm terminating upon a predetermined stopping condition.

The ℓ_1 penalised approach can also be used for non-negative sparse approximation using typical ℓ_1 solvers with the non-negative constraint applied [26] [27] or penalised NMD approaches [28]. However, NNLS can be considered a sparse algorithm as the non-negative constraint performs an innate regularisation [29], and is shown empirically to outperform non-negative ℓ_1 minimisation [29]. In AMT experiments we have observed little difference between such non-negative ℓ_1-approximation and NNLS.

Gradient-based methods, often based on NMF, are generally preferred for spectrogram decompositions. While stepwise methods employ the Euclidean distance, other cost functions are considered superior for audio signal processing [2] [4] [30]. In particular, it was shown that the Kullback-Leibler (KL) divergence

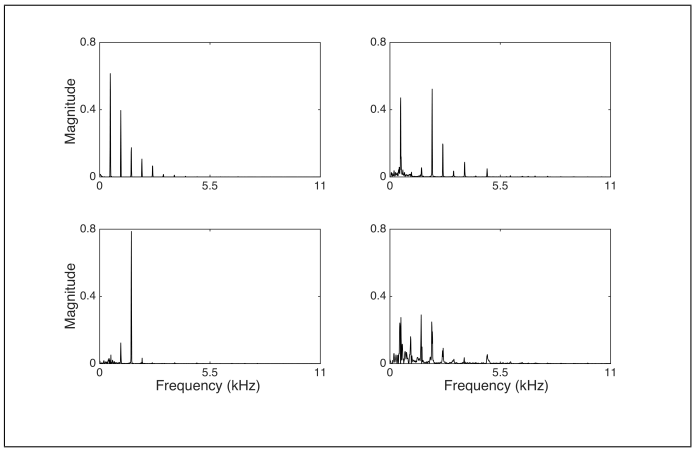

Fig. 3: Group of atoms forming a subspace representing one note.

$$\mathcal{C}_{KL}(\mathbf{s}|\mathbf{z}) = \sum_n s_n \log \frac{s_n}{z_n} - s_n + z_n \tag{12}$$

where $\mathbf{z} = \mathbf{D}\mathbf{x}$ is the current estimate, outperforms Euclidean distance in the original paper considering NMF for AMT [2]. The generalised β-divergence [31]

$$C_\beta(\mathbf{s}|\mathbf{z}) = \frac{1}{\beta(\beta-1)} \sum_m s_m^\beta + (\beta-1)z_m^\beta - \beta(s_m z_m^{\beta-1}) \tag{13}$$

generalises popular cost functions such as Euclidean distance ($\beta = 2$), with KL (12) and Itakuro-Saito (IS) divergences as limit cases as $\beta \to \{1, 0\}$, respectively. NMD experiments described in [4] [5] report superior AMT results for $\beta = 0.5$.

Similar to NNLS, sparsity is a known side effect of NMF due to non-negative regularisation [1]. Nonetheless it is relatively common to enhance this implicit sparsity by using penalty terms, which are easily accommodated in NMF. Typically an ℓ_1 penalty is considered [32] [33] [34]; however this may not always be effective [33] [27]. Concave penalties, such as the log based penalty, $\sum_n \log(1+x_n)$, used with audio signals in [6], may be attractive as they tend to be sparser than the ℓ_1 norm. Penalised NMF approaches with β-divergence generally lead to the multiplicative updates [34]

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left[\frac{\mathbf{D}^T[\mathbf{S} \otimes [\mathbf{D}\mathbf{X}]^{[\beta-2]}]}{[\mathbf{D}^T[\mathbf{D}\mathbf{X}]^{[\beta-1]}] + \lambda\Psi(\mathbf{X})} \right]^{[\varphi(\beta)]} \tag{14}$$

$$\mathbf{D} \leftarrow \mathbf{D} \otimes \left[\frac{[\mathbf{S} \otimes [\mathbf{D}\mathbf{X}]^{[\beta-2]}]\mathbf{X}^T}{[[\mathbf{D}\mathbf{X}]^{[\beta-1]}\mathbf{X}^T] + \lambda\Psi(\mathbf{D})} \right]^{[\varphi(\beta)]} \tag{15}$$

where $\otimes$ denotes elementwise multiplication, $\mathbf{x}^{[.]}$ denotes elementwise exponentiation of a vector or matrix, the matrix division is also elementwise, $\Psi(\mathbf{D})$ and $\Psi(\mathbf{X})$ typically describe the gradient of the penalty term, and $\varphi(\beta)$ is a parameter that varies with β and the penalty used to ensure descent of the cost function. For the range $0 \leq \beta \leq 2$, a value of $\varphi(\beta) = 1$ is given in [35] for the unpenalised case, while $\varphi(\beta) = 1/(3-\beta)$ is given when a ℓ_2^2 penalty is applied [34].

III. Non-negative Group Sparse Methods

In order to apply the subspace model for AMT using magnitude spectrogram, non-negative group sparse algorithms are proposed.

A. Non-negative Group Sparse OMP

Non-negative group sparse OMP methods are simply derived, similar to NN-OMP [23], by using NNLS backprojection and enforcing non-negativity in the selection step. We derive a non-negative B-OMP (NN-BOMP) selection criterion from (7) by considering only positive inner products:

$$\hat{j} = \arg\max_j \|\phi^+[j]\|_2 \tag{16}$$

when $\phi^+ = \mathcal{I}\phi$ where $\mathcal{I}$ is an "is positive" indicator function. We previously proposed the Non-Negative Nearest Subspace OMP (NN-NS-OMP) [15], using the selection criteria:

$$\hat{j} = \arg\min_{\mathbf{x},j} \|\mathbf{r} - \mathbf{D}[j]\mathbf{x}[j]\|_2^2 \quad s.t. \quad \mathbf{x}[j] \geq 0 \tag{17}$$

where $\mathbf{x}[j]$ is the NNLS solution vector for the decomposition of the residual over the subspace $\mathbf{D}[j]$. While (17) can be considered a non-negative variant of Subspace Matching Pursuit selection criteria (8), the non-negative constraint implies that the solution to (17) is not accessible through dictionary-residual multiplication as the active set must be determined for each group, requiring the use of NNLS. This is computationally demanding, as NNLS is calculated for each group at each frame q times, where q is the number of groups to be selected. For a minute of music, sampled at $44.1\,\text{kHz}$, with a hopsize of 1024 samples, or $\sim 23.2\,\text{ms}$, a dictionary containing 88 pitched groups, and an average polyphony of $q = 5$, NN-NS-OMP requires more than 10^6 blockwise NNLS calculations. A fast, exact, variant of NN-NS-OMP upper bounds the norm of the NNLS projection by the lower of the norm of the least squares projection $\pi_j(r)$ (8) and the NN-BOMP coefficient (16), both available through dictionary-residual multiplication, in order to prune the number of NNLS calculations. The details of this approach are left to [36] [37].

B. Backwards Elimination

Problems with corruption of time continuity [11] and difficulty in selecting an apt stopping condition [9] are reported when matching pursuits are employed for AMT. Indeed, it would seem that greedy pursuits may not be appropriate for AMT decompositions. Pursuit algorithms are known to give accurate results when the dictionary elements are uncorrelated, or incoherent [38]. However, non-negative dictionaries are innately coherent [23], a problem accentuated by harmonicity in a dictionary representing pitched notes, where consonant notes are represented by coherent atoms. As a result, it is observed that even initial atom/note selections with greedy methods may be incorrect when two related pitches are present, and a correction mechanism is desirable.

Bi-directional pursuits that alternate between forward selection and backwards elimination have been recently proposed [39] [40] [41]. Some of these approaches [41] [39] [42] are also *stepwise optimal*, in that the sparse cost function (2) is optimised at each elimination or selection. For example, given the current support, Γ, a forward optimal step is given by

$$\hat{k} = \arg\min_{k,\mathbf{x}} \|\mathbf{s} - \mathbf{D}_{\{\Gamma \cup k\}}\mathbf{x}\|_2^2. \tag{18}$$

In comparison, OMP selects the atom that optimises the least squares error relative to the current residual :

$$\hat{k} = \arg\min_{k,x} \|\mathbf{r} - \mathbf{d}_k x\|_2^2. \tag{19}$$

Fast stepwise optimal selection and elimination criteria, derived using block matrix updates, are proposed in the Greedy Sparse Least Squares approach [42].

The non-negative constraint suggests a simple stepwise optimal strategy for the problem at hand. In particular an initial sparse solution can be derived using NNLS, which has a natural stopping condition. Subsequently the necessity to alternate between forwards and backwards steps in order to correct early errors is removed, and an elimination only strategy, referred to as Backwards From NNLS (BF-NNLS) [16] is proposed. The group sparse variant, GBF-NNLS, is outlined in Fig. 4. After the initial NNLS (20) is performed the set of active groups, Γ, is identified before entering the main loop of iterative elimination. At the start of each iteration the index ordered set of active atoms, Π, is denoted and the inverse of the Gram matrix of the active set of the dictionary, $\mathbf{F}$, is calculated. The group elimination cost, Δ_j, equal to the difference in ℓ_2^2 error before and after elimination, is given by (21) where $\mathbf{F}[\bar{j}][\bar{j}]$ denotes the square block of the matrix $\mathbf{F}$ with row/column indices related to the active atoms from the jth group. The calculation of the group elimination cost is derived using block matrix inverse updates, generalising the atomic elimination step used in [42]. The group, indexed by $\hat{j}$, displaying the minimum elimination cost is then eliminated from the support. NNLS is then performed using only the supported groups, before the iteration is re-entered.

- *Input* : $\mathbf{D} \in \mathbb{R}^{M \times N}; \mathbf{s} \in \mathbb{R}^M; \mathcal{J}$
- *Initialise* :

$$\mathbf{x} \leftarrow \arg\min_x \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2 \quad s.t. \quad \mathbf{x} \geq 0 \tag{20}$$

$$\Gamma = \{j | \|\mathbf{x}[j]\| > 0\}$$

- Do While $\bar{\Delta}_j \leq \lambda$ (or $|\Gamma| > q$)

$$\Pi = \{k | \mathbf{x}_k > 0\}; \quad \mathbf{F} = [\mathbf{D}_\Pi^T \mathbf{D}_\Pi]^{-1}$$

  - *Select group*

$$\hat{j} = \arg\min_j \Delta_j = \arg\min \mathbf{x}[\bar{j}]^T \left[\mathbf{F}[\bar{j}][\bar{j}]\right]^{-1} \mathbf{x}[\bar{j}] \tag{21}$$

$$where \quad \bar{j} = \{i \mid \Pi(i) \in \mathcal{J}^j\}$$

  - *Eliminate* : $\Gamma \leftarrow \Gamma \backslash \hat{j}; \quad \mathbf{x}[j] = 0$
  - *Reproject*

$$\mathbf{x}_\Gamma \leftarrow \arg\min_x \|\mathbf{s} - \mathbf{D}_\Gamma \mathbf{x}\|_2 \quad s.t. \quad \mathbf{x} \geq 0$$

Fig. 4: Group Backwards From NNLS algorithm.

In practice NNLS is performed on the full spectrogram, $\mathbf{S}$, before elimination, in order to determine the stopping condition, λ, which is calculated using a parameter, δ:

$$\lambda = \delta \times \max_{j,n}[H]_{j,n} \tag{22}$$

where $\mathbf{H}$ is a group coefficient matrix

$$[H]_{j,n} = \|\mathbf{D}[j]\mathbf{x}_n[j]\|_2. \tag{23}$$

This approach is similar to that used in [4] and [5] for thresholding approaches. The optimal value of δ is seen to be consistent across spectrograms of different transforms [37] in thresholding approaches. As this consistency is desirable, a modified group sparse cost function is used [16]

$$\mathcal{C}_{mod} = \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_{\perp,0} \tag{24}$$

where $\|\mathbf{x}[j]\|_\perp = \|\mathbf{D}[j]\mathbf{x}[j]\|_2$. The modified cost (24) simply replaces the typically used ℓ_2^2 error (2), with the ℓ_2 error. Experiments in [16] verify that use of this cost function maintains the scaling property seen in thresholding. The modified elimination cost, $\bar{\Delta}_{\hat{j}}$, is only explicitly considered in the stopping condition, as the ordering of elimination costs is the same as that of the standard elimination cost $\Delta_{\hat{j}}$, from which it is calculated

$$\bar{\Delta}_{\hat{j}} = \sqrt{\|\mathbf{r}\|_2^2 + \Delta_{\hat{j}}} - \|\mathbf{r}\|_2 \tag{25}$$

where $\mathbf{r}$ is the current residual.

C. GS-β-NMF with $\ell_{2,\beta}^\beta$ penalty

A variant of group sparse NMF, using IS ($\beta = 0$), was proposed for source separation in [43], with a log-based penalty applied to ℓ_1 norm group coefficients:

$$\lambda \sum_l \log(a + \|\mathbf{x}[l]\|_1) \tag{26}$$

leading to updates generalised by (14) (15). A strategy for estimation of the parameters λ, a, in (26) is given in [43], however it is unclear whether this strategy extends to cost functions other than IS. The penalty (26) is also employed with KL divergence in [22], in which case optimisation is performed by using a convex-concave projection algorithm. As an alternative, we now propose a group penalty using an ℓ_2 norm group coefficient that is scale invariant to β-divergence.

An alternative to the log-based sparse penalty is the ℓ_p^p quasinorm measure [44] which is also concave when $p < 1$

$$\|\mathbf{x}\|_p^p = \|\mathbf{x}^{[p]}\|_1 \tag{27}$$

where $\mathbf{x}^{[p]}$ denotes elementwise exponentiation, and can form a tighter approximation to the ℓ_0 pseudonorm as $p \to 0$ than either ℓ_1 or the log-based penalty. We considered a small value of p in [17], but now propose to use a ℓ_β^β penalty with the β-divergence, noting the scaling relationship when $\beta > 0$

$$\frac{\mathcal{C}_\beta(\mathbf{s}|\mathbf{z})}{\|\mathbf{x}\|_\beta^\beta} = \frac{\mathcal{C}_\beta(a\mathbf{s}|a\mathbf{z})}{\|a\mathbf{x}\|_\beta^\beta}. \tag{28}$$

This relationship implies consistent sparse penalisation, relative to a given λ, regardless of scale. For the KL-divergence ($\beta = 1$), the ℓ_1 penalty gives constant penalisation, which is explained in terms of dispersion factors of exponential distributions in [34] [45]. This scale invariance is desirable, hence we use an $\ell_{2,\beta}^\beta$ penalty for the GS-β-NMF problem

$$\mathbf{X}, \mathbf{D} \leftarrow \arg\min_{\mathbf{X},\mathbf{D}} \mathcal{C}_\beta(\mathbf{S}|\mathbf{D}\mathbf{X}) + \frac{\lambda}{\beta}\sum_{n=1}^N \|\mathbf{y}_n\|_{2,\beta}^\beta \tag{29}$$

for $\beta \in]0, 2]$ where $\mathcal{C}_\beta$ is given by (13) and

$$[Y]_{k,n} = [X]_{k,n} \times \|\mathbf{d}_k\|_2 \tag{30}$$

is considered in order to accommodate the ℓ_2 norm constraint on each atom. For KL ($\beta = 1$), the $\ell_{2,1}$ penalty is employed in (29), giving a convex cost function and linear scaling, unlike the approach in [22]. KL with $\ell_{2,1}$ penalty was previously used for group sparse NMD [46], however a monotonic algorithm was not developed in [46], and is offered here.

Majorisation-Minimisation (MM) methods are used to derive monotonic descent algorithms for β-NMF [47] [35] and penalised β-NMF [34]. MM approaches consider an auxiliary function $\mathcal{G}(\mathbf{x}, \hat{\mathbf{x}})$ defined by the properties

$$\mathcal{C}(\hat{\mathbf{x}}) = \mathcal{G}(\hat{\mathbf{x}}, \hat{\mathbf{x}}); \qquad \mathcal{C}(\mathbf{x}) \leq \mathcal{G}(\mathbf{x}, \hat{\mathbf{x}}) \tag{31}$$

where $\mathcal{C}(\mathbf{x})$ denotes $\mathcal{C}(\mathbf{s}|\mathbf{D}\mathbf{x})$, and $\hat{\mathbf{x}}$ is referred to as an auxiliary variable. In practical terms, the auxiliary vector $\hat{\mathbf{x}}$ is set to the current estimate of the coefficient vector, which is considered a constant. Optimisation of the auxiliary function

$$\mathbf{x} \leftarrow \arg\min_{\mathbf{x}} \mathcal{G}(\mathbf{x}, \hat{\mathbf{x}}) \tag{32}$$

then results in optimisation of the function $\mathcal{C}(\mathbf{x})$. Separability of auxiliary functions in terms of individual variables such that

$$\mathcal{G}(\mathbf{x}|\hat{\mathbf{x}}) = \sum_k \mathcal{G}(x_k|\hat{\mathbf{x}}) + C \tag{33}$$

where C is a constant, is desirable allowing decoupling of the optimisation [35]. Summed variables are separable, and a MM approach is used for the ℓ_2^2 penalty, or $\sum_k \mathbf{x}_k^2$, in [34].

The $\ell_{2,\beta}^\beta$ penalty is separable groupwise as $\|\mathbf{y}[j]\|_{2,\beta}^\beta = \sum_j \|\mathbf{y}[j]\|_2^\beta$. Development of a MM approach for (29) requires an auxiliary function for $\|\mathbf{y}[j]\|_2^\beta$. This can be achieved using the weighted arithmetic-geometric inequality

$$(a^v b^w)^{\frac{1}{v+w}} \leq \frac{va + wb}{v + w} \tag{34}$$

and setting $a = \|\mathbf{y}[j]\|_2^2$, $b = \|\hat{\mathbf{y}}[j]\|_2^2$, $v = \beta$, $w = 2 - \beta$:

$$\|\mathbf{y}[j]\|_2^\beta \leq \frac{\beta}{2}\frac{\|\mathbf{y}[j]\|_2^2}{\|\hat{\mathbf{y}}[j]\|_2^{2-\beta}} + \left(1 - \frac{\beta}{2}\right)\|\hat{\mathbf{y}}[j]\|_2^\beta. \tag{35}$$

with equality only when $\|\hat{\mathbf{y}}[j]\|_2 = \|\mathbf{y}[j]\|_2$. The inequality (35) has previously been derived, less the group notation, using Young's inequality [48], and through calculating a quadratic function that is tangent to the concave left hand side at the current estimate [49] [44]. The second term of the right hand side in (35) is a constant in terms of $\|\mathbf{y}[j]\|$ as it is given in terms of the auxiliary variable, allowing the auxiliary function for the ℓ_2^β penalty to be given as

$$\mathcal{G}_{\ell_2^\beta}(y_k|\hat{\mathbf{y}}[j]) = \frac{\beta y_k^2}{2\|\hat{\mathbf{y}}[j]\|_2^{2-\beta}} + C. \tag{36}$$

where $k \in \mathcal{J}^j$, and C denotes a constant in terms of $\|\mathbf{y}[j]\|$. Considering unit norm atoms ($\mathbf{Y} = \mathbf{X}$), the gradient of (36) relative to x_k is

$$\nabla_{x_k} \mathcal{G}_{\ell_2^\beta}(x_k|\hat{\mathbf{x}}[j]) = \frac{\beta x_k}{\|\hat{\mathbf{x}}[j]\|_2^{2-\beta}}. \tag{37}$$

Elementwise multiplication of (37) by $\hat{x}_k/\beta x_k$, consideration that (36) is separable in each column of $\mathbf{X}$, and setting $\hat{x} \to x$ in the convention of (14) leads to

$$[\Psi(\mathbf{X})]_{k,n} = \frac{x_{kn}}{\|\mathbf{x}_n[j]\|_2^{2-\beta}} \tag{38}$$

which is inserted into (14), to optimise (29) relative to $\mathbf{X}$.

For the dictionary update the auxiliary function (36) is not separable in the columns of $\mathbf{X}$ and needs to be summed over all activations $\mathcal{G}_{\ell_2^\beta}(d_{mk}|\hat{\mathbf{Y}}[j]) = \sum_n \mathcal{G}_{\ell_2^\beta}(d_{mk}|\hat{\mathbf{y}}_n[j])$. Otherwise, a similar process to that used to derive (38) gives

$$[\Psi(\mathbf{D})]_{m,k} = [D]_{m,k} \sum_n \frac{x_{kn}^2}{\|\mathbf{x}_n[j]\|_2^{2-\beta}} \tag{39}$$

where $k \in \mathcal{J}^j$, which is inserted into (15), to optimise GS-β-NMF (29) relative to $\mathbf{D}$. Subsequent normalisation such that $\mathbf{x}_{kn} = \mathbf{x}_{kn} \times \mathbf{d}_k$; $\mathbf{d}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|}$, does not affect the cost (29) due to (30). For both (14) with (38) and (15) with (39) a value of $\varphi(\beta) = (3-\beta)^{-1}$ guarantees monotonicity through a similar MM strategy as used for ℓ_2^2 penalised β-divergence in [34].

IV. Dictionary Tuning with GS-β-NMF

Use of GS-β-NMF requires knowledge of the partitioning of the dictionary, unless a group clustering strategy is applied. An application of GS-β-NMF for dictionary tuning in the former case is now proposed. Unlike dictionary learning, which seeks to discover a dictionary in a purely data-driven manner, dictionary tuning considers initialising a dictionary that is fit for purpose and preferably generic, and allowing it to morph into a better version of itself in its immediate context, while maintaining its labelling. For this purpose, we restructure the adaptive harmonic dictionary (AHD) proposed by Vincent et al [50] [4] [14], which models a pitched atom by superposition of narrowband atoms of the same pitch, such as seen in Fig. 5. A similar model was earlier proposed by Virtanen and Klapuri [51] using a linear frequency scale. The AHD [4] uses a logarithmic Equivalent Rectangular Bandwidth Transform (ERBT) scale, which may be more robust to inharmonicity due to larger spacing between higher frequency bins.

In [4] it is considered that an atom, $\mathbf{e}_j$, representing the full spectrum of the jth note is formed from a superposition of several narrowband harmonic atoms, $\mathbf{D}[j]$, of similar pitch

$$\mathbf{e}_j = \mathbf{D}[j]\mathbf{u}[j] \tag{40}$$

where $\mathbf{D}$ is the AHD and $\mathbf{u} \in \mathbb{R}^K$. The spectrogram is then approximated by

$$\mathbf{S} \approx \mathbf{E}\mathbf{X} \tag{41}$$

where $\mathbf{X} \in \mathbb{R}^{88 \times N}$. In this way $\mathbf{E}$ can be considered the top-level of a hierarchical dictionary in which each atom, $\mathbf{e}_j$ is formed as a linear mixture of several columns of $\mathbf{D}$, the fixed AHD. Semi-supervised NMF algorithms were proposed

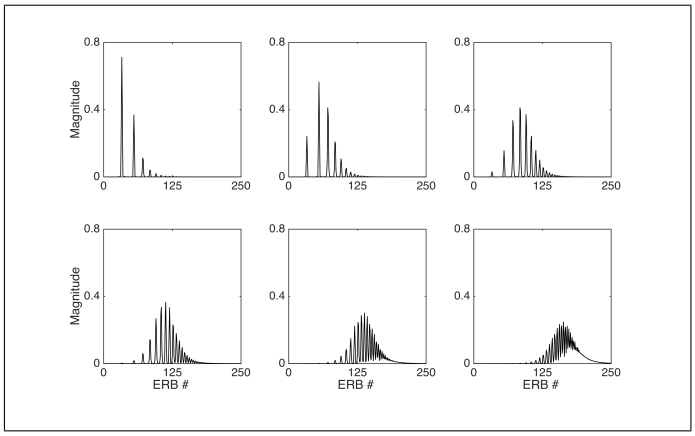

Fig. 5: Group of atoms used to represent one note in adaptive harmonic dictionary.

for the approximation (41) using a perceptually weighted Euclidean distance [50] and β-divergence [4], which we refer to here as Harmonic NMF (H-NMF). H-NMF used alternating multiplicative updates to estimate the pitch activation matrix $\mathbf{X}$, and then the spectral shape of each individual atom in $\mathbf{E}$ by updating $\mathbf{u}[j]$. The low-level dictionary, $\mathbf{D}$, is not updated.

An alternative perspective on the AHD is taken here, in which the top-level dictionary, $\mathbf{E}$, is excluded, and the dictionary $\mathbf{D}$ is group structured, that is, the narrowband atoms used to represent one note form a subspace. In this case, GS-β-NMD (14) (38) can be used as a decomposition algorithm, leading to note representations that can vary at each individual time frame, as the signal can now be decomposed with 88 pitched subspaces rather than with 88 atoms, as in H-NMF. In particular, the coupling between narrowband atoms is effected through data in H-NMF, while it is effected simply by the group sparse penalty in GS-β-NMD. In this way, H-NMF may present different results for a given piece of data when learning is performed on a subset, or superset, of that data while GS-β-NMD will present the same result as each decomposition is independent of other time frames.

A potential weakness of H-NMF, and GS-β-NMD, in this context, is the strict harmonic model of the AHD due to the narrowband atoms being fixed. Non-harmonic elements present in a signal, such as the resonances of a piano body, may then lead to false detections of atoms that best capture their energy. It is considered that relaxation of the harmonic constraint in the AHD may be beneficial. For this purpose the dictionary tuning approach, outlined in Fig. 6, that uses GS-β-NMF to update the AHD, is proposed here. In order for the harmonic constraint to be dropped, a small value, $\epsilon_{mk} = \sum_m d_{mk}/(4 \times M)$, is added to all elements of the dictionary allowing them to be updated. While it is expected that the coupling of atoms within a group will maintain the pitch identity of each atom, some steps are made to explicitly encourage this behaviour through slowing the dictionary updates. An initial decomposition is performed using NMD in order to form a reasonable approximation of the signal, in which case the dictionary can be expected to change less than it might from a random coefficient initialisation. Then when the dictionary tuning starts, two coefficient matrix updates are performed for each dictionary update. Finally, the dictionary

- **Input** $\mathrm{S} \in \mathbb{R}^{M\times N}, \mathbf{D} \in \mathbb{R}^{M\times K}$, $\lambda, \mathcal{J}, \beta, a, b$
- **Initialise**
  - $x_{kn} = 0.01 \; \forall \; \{k, n\}$
  - for i=1:a
    - Perform NMD using (14) with $\Psi(\mathbf{X}) = 0$
- **Dictionary Tuning**
  - for i=1:b
    - Update dictionary using (15) with (39)
    - Normalise: $\mathbf{x}^k \leftarrow \mathbf{x}^k \times \|\mathbf{d}_k\|_2$; $\mathbf{d}_k \leftarrow \mathbf{d}_k / \|\mathbf{d}_k\|_2$
    - Update activation matrix (14) with (38) ($\times 2$)
- **Output** $\mathbf{X}, \mathbf{D}$

Fig. 6: Dictionary tuning with GS-β-NMF.

update itself is further stabilised through addition of an extra term, $\mu = 1$, to the numerator and denominator in the dictionary update (15). This reduces the step size taken for each dictionary element, particularly in the case where the numerator and denominator are small, and more likely to result in very large steps that may introduce instability to the pitch labelling of the dictionary.

V. Evaluation

EXPERIMENTS were performed with the group sparse algorithms to evaluate their use for AMT. Further objectives include evaluating subspace modelling relative to the datapoint approach. A dataset was formed from the EnStDkCl subset of the MAPS database [52], containing live recordings of 30 pieces of classical piano played by a Disklavier piano. The Disklavier is an upright piano, that is capable of robotic acoustic playback with piano strings struck by electromechanically actuated hammers. This robotic setup leads to acoustic signals with a reliable ground truth that affords a more rigorous experimental setup that is not available for other instruments. The acoustic nature of these signals has previously led to a large divergence in results, particularly for onset detection, relative to MIDI playback files in the MAPS dataset [14] [53] [54]. In particular, the difference reported in [54] is over 20%

The first $30\,$s of each piece in the dataset were downsampled to $22.05\,$kHz. For each piece a Short-Time Fourier Transform (STFT) spectrogram [50] with window size 2048 and 75% overlap, leading to a hopsize of $\sim 23.2\,$ms, was formed. A further set of spectrograms using an ERBT of dimension $M = 250$, similar to the AHD, and with similar temporal resolution, was also formed. Dictionaries were learnt offline from a set of signals containing isolated notes, also from the EnStDkCl subset of the MAPS database. For each of the 88 notes on the piano scale, an STFT spectrogram of the corresponding isolated note was computed, with similar parameters as the spectrograms of the dataset. Subspaces were learnt from each spectrogram using Euclidean distance NMF for a range of values of rank $P \in \{1, ..., 7\}$ in order to compare the effects of subspace size. An example subspace of rank $P = 4$ is shown in Figure 3. The dictionary was formed by concatenating the individual pitched subspaces with appropriate labelling, with each atom normalised to unit ℓ_2 norm. A datapoint dictionary was formed from the same spectrograms. In order to omit silent segments at the start, onset detection was performed on each isolated note spectrogram and 50 spectra, representing $\sim 1.16\,$s of audio including, and subsequent to, the onset were extracted. These datapoints were normalised and formed a subdictionary representing the given note. The datapoint dictionary was formed by concatenation of these subdictionaries. Equivalent dictionaries were also formed using the ERBT.

A. Experiment A

Spectrogram decompositions were performed to compare the stepwise methods and NNLS for subspace and datapoint dictionaries. As the selection of a stopping condition for OMP is known to be problematic [9], the sparsity, or polyphony, at each frame is given in order to allow fair comparison of the different approaches. NN-BOMP and NN-NS-OMP were both run for all values of $P \in \{1, ..., 7\}$, noting that both algorithms revert to NN-OMP when $P = 1$, and were stopped when q_n, the number of notes active at the nth frame, groups were selected. OMP was used with the datapoint dictionaries, in which case selection of different atoms of the same pitch was allowed, and a stopping condition specifying that q_n notes are selected at each time frame was used.

NNLS was run using the datapoint dictionary and with the subspace dictionaries, in which case it is referred to as GT-NNLS. An early stopping strategy, after 100 iterations, was used for NNLS with the datapoint dictionaries, as convergence may not occur due to the dimensions of the dictionary. To allow comparison with the OMP based approaches, a q-thresholding was performed whereby the q_n notes displaying the largest pitch-grouped coefficients (23) at each frame were selected and all other pitches set to zero. Similarly, GBF-NNLS was performed until only q_n groups were active. GBF-NNLS was also employed with the datapoint dictionaries, with grouping of active atoms of a similar pitch in each column of the initial NNLS activation matrix. For each decomposition the sets of true positive, tp, and false positive, fp, detections are denoted for all pieces, and the results are described in terms of $\mathcal{F}$-measure which, when the sparsity level is known, is given simply by $\mathcal{F} = \frac{|tp|}{|tp|+|fp|}$. The STFT spectrograms and corresponding dictionaries were used.

Results for this set of experiments are shown in Figure 7. The subspace methods are seen to improve on the case of $P = 1$, with the increase in $\mathcal{F}$-measure being of the order of $3 \sim 5\%$ at the optimal value of $P = 5$. NN-NS-OMP is seen to be more consistent than NN-BOMP, and with optimal P performs similar to OMP using the datapoint dictionaries. NN-BOMP, for some values of P performs worse than NN-OMP. The thresholded NNLS approaches outperform the OMP methods while GBF-NNLS adds further improvements in all cases. For the subspace dictionaries, GBF-NNLS is seen to improve on GT-NNLS by around 3% except when $P = 1$, for which similar $\mathcal{F}$-measure is given. For the datapoint dictionaries, the improvement is smaller. GBF-NNLS, at optimal P, is also seen to perform similar to NNLS and GBF-NNLS using the large datapoint dictionary.

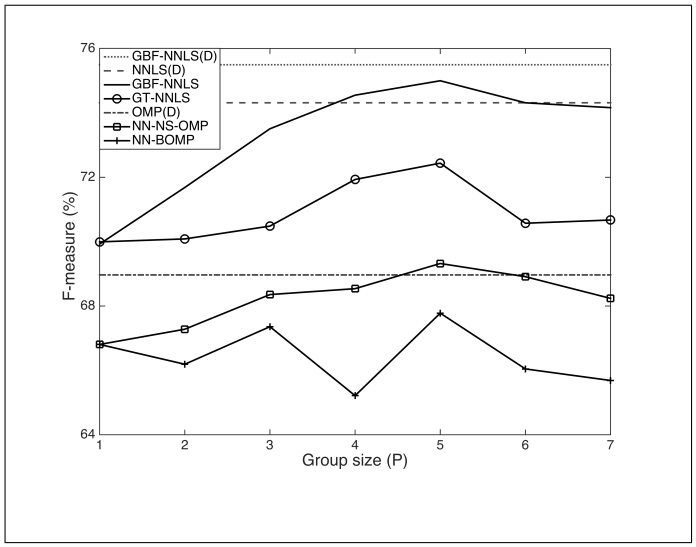

Fig. 7: AMT results for OMP and NNLS approaches with subspace and datapoint dictionaries and known polyphony. Datapoint dictionary methods denoted by (D)

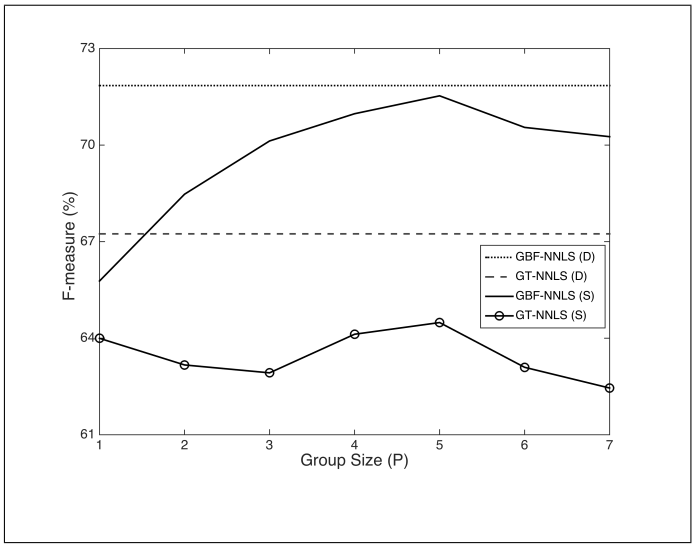

Fig. 8: $\mathcal{F}$-measure for AMT, relative to groupsize P for (G)BF-NNLS and (G)T-NNLS with subspace dictionaries (S), and with datapoint dictionary (D).

B. Experiment B

Experiments giving a more realistic comparison of NNLS and backwards elimination approach, without known polyphony, were performed. Thresholding, similar to that described in [50] [4] [5] was performed for the NNLS approaches. In the case of GT-NNLS, and NNLS with the datapoint dictionaries, thresholding of the group coefficient matrix, $\mathbf{H}$, (23) using the δ parameter (22) was performed for a variety of values of $\delta \in \{15, ..., 50\}$ dB in steps of 1 dB. At each value of δ, for all pieces, the ground truth and binarised thresholded group matrix are compared, with true positives, false positives and false negatives, *fn*, denoted from which the common Precision, $\mathcal{P}$, Recall, $\mathcal{R}$ and $\mathcal{F}$-metrics

$$\begin{aligned} \mathcal{P} &= |tp| \,/\, (|tp| + |fp|) \\ \mathcal{R} &= |tp| \,/\, (|tp| + |fn|) \\ \mathcal{F} &= 2 \times \mathcal{P} \times \mathcal{R}/(\mathcal{P} + \mathcal{R}) \end{aligned}$$

were derived. The optimal results in terms of $\mathcal{F}$-measure, found at δ_{opt} applied across all pieces, were recorded. For GBF-NNLS, with both dictionaries, the stopping condition is calculated from the coefficient matrix of the NNLS decomposition (22), with experiments similarly run for various values of δ. Again the STFT spectrograms were employed.

Results are shown in Fig 8, where the difference between GBF-NNLS and GT-NNLS is more marked than in *Exp. A* with differences of $\sim 7\%$ seen for the subspace dictionaries, and $\sim 5\%$ for the datapoint dictionaries. GT-NNLS varies little relative to P. GBF-NNLS(S) improves on NNLS with the datapoint dictionaries by $\sim 4\%$, and again approaches the performance of the GBF-NNLS with datapoint dictionaries.

We also compared to two other methods that are designed for use with overcomplete dictionaries. ASNA [55] is a stepwise method that uses the KL cost function (12), a selection criterion based on the KL gradient, and Newton steps to perform the signal estimation. The ASNA approach was run for 100 iterations, similar to NNLS. Another KL-based method, proposed in [56], seeks to minimise $\mathcal{C}_{KL}(\mathbf{s}|\mathbf{D}\mathbf{x}) - \lambda\|\mathbf{x}\|_2$. We tested for several values of $\lambda = 2^{-a}; a \in \{0, 1, ...6\}$ and found $\lambda = 1/4$ to perform best, concurring with the optimal range expressed in [56]. We ran until a convergence criterion as using the few iterations suggested by the authors [56] resulted in poor performance. The results are given in Table I, where the two unpenalised stepwise methods, ASNA and T-NNLS are seen to perform similarly. In terms of the penalised methods the GBF-NNLS with the datapoint dictionary performs slightly better than the KL-ℓ_2 approach [56]. GBF-NNLS with the subspace dictionary is comparable to these methods.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
T-NNLS	66.9	67.6	67.2
ASNA [55]	66.9	66.4	66.6
KL-ℓ_2 [56]	71.1	69.1	70.1
GBF-NNLS	75.9	68.2	71.9
GBF-NNLS (S)	76.4	67.2	71.5

TABLE I
COMPARISON OF METHODS USING DATAPOINT DICTIONARY TO GBF-NNLS WITH A SUBSPACE DICTIONARY WITH (P = 5). (S) DENOTES SUBSPACE DICTIONARY

C. Experiment C

Experiments were run to test the effectiveness of the NMF-based approaches, comparing the use of the proposed GS-β-NMF dictionary tuning and GS-β-NMD using AHD with H-NMF [4]. The ERBT spectrograms were employed.

The H-NMF algorithm was run using code supplied by the authors, and the AHD was produced using the default settings provided. For the group sparse approaches, normalisation of the atoms to unit ℓ_2 norm is performed. H-NMF was run with $\beta = 0.5$ for which superior performance is reported [4]. GS-β-NMD / NMF was run with $\beta = 0.5$ and also with the KL-divergence ($\beta = 1$). For KL and $\beta^{(0.5)}$, a value of $\lambda = 1$ was seen in early experiments to be apt and was used for all experiments. Dictionary tuning using GS-β-NMF was run for $b = 30$ iterations, after an initial $a = 30$ iterations of NMD.

	$\mathcal{P}$	$\mathcal{R}$	δ_{opt}	$\mathcal{F}$
KL-NMD	72.3	69.6	32	70.9
β-NMD	74.5	69.4	33	71.9
H-NMF [4]	70.3	65.3	29	67.7
GS-KL-NMD	67.9	67.1	30	67.5
GS-β-NMD	70.5	65.3	29	67.8
GS-KL-NMF	75.4	68.5	31	71.8
GS-β-NMF	75.5	70.5	34	72.9

TABLE II

FRAMEWISE RESULTS FOR SUPERVISED KL AND β-NMD USING OPTIMAL ONE ATOM PER PITCH DICTIONARY, H-NMF [4], PROPOSED GS-β-NMD APPROACHES AND GS-β-NMF DICTIONARY TUNING APPROACHES.

	$\mathcal{P}$	$\mathcal{R}$	δ_{opt}	$\mathcal{F}$
KL-NMD	84.8	76.3	29	80.3 (74.2)
β-NMD	88.7	76.7	30	82.3 (76.0)
H-NMF [4]	77.2	74.4	28	75.8 (71.7)
GS-KL-NMD	78.0	69.4	27	73.4 (67.6)
GS-β-NMD	75.5	72.4	27	73.9 (69.2)
GS-KL-NMF	82.7	75.7	28	79.1 (72.7)
GS-β-NMF	83.1	77.7	29	80.3 (75.4)

TABLE III

ONSET ANALYSIS RESULTS FOR PROPOSED GS-β-NMD / NMF APPROACHES, COMPARED AGAINST H-NMF [4], AND SUPERVISED KL AND β-NMD USING OPTIMAL ONE ATOM PER PITCH DICTIONARY. NUMBERS IN BRACKETS FOR IN THE $\mathcal{F}$ COLUMN SHOW RESULTS FOR THRESHOLDING ON ACTIVATIONS RATHER THAN THE PROPOSED ACTIVATION DIFFERENCES.

GS-β-NMD used a flat initialisation on the activations, setting $[X]_{k,n} = 0.01 \forall \{k, n\}$, and were run until the cost function was seen to decrease by less than 0.5% over 5 iterations. For comparison, $\beta^{(0.5)}$-NMD and KL-NMD were used to perform AMT using a dictionary with one atom per pitch, and were run with similar initialisation and stopping conditions as GS-β-NMD. Framewise analysis is performed in a similar manner to the previous experiments.

1) Onset Analysis: An onset analysis is also performed. Typically this is performed using a simple threshold-based onset detector, which is triggered when a threshold is surpassed and sustained for a minimum of 3 frames [4] [17] [3]. A true positive is denoted when a detected onset falls within a tolerance of 50 ms from a ground truth onset, with other detections denoted as false positives, and undetected ground truth onsets denoted as false negatives. Analysis of onset detection is performed in a similar manner to the framewise case, using thresholding relative to a range of values of the δ parameter (22), and with results given for $\mathcal{P}, \mathcal{R}, \mathcal{F}$.

We have previously observed systematic problems with the described onset detector [57], including not capturing retriggered notes and detecting spurious false onsets when the activation level is near the threshold. We propose some modifications. In order to avoid spurious triggering a small median filter is applied to each row of the activation matrix, $\mathbf{h}^k$, resulting in a smoothed coefficient matrix $\hat{\mathbf{H}}$. In order to capture retriggered notes, difference matrices such that $[A]_{k,n} = [H]_{k,n} - [H]_{k,n-1}$ are used. The differences are calculated for both $\mathbf{H}$ and $\hat{\mathbf{H}}$. A search for candidate onsets considers only points where the elements of both difference matrices are above a threshold, and, similar to above, the subsequent two activations, $[H]_{k,n+1}, [H]_{k,n+2}$ are above the threshold. Often two or more of these points are found adjacent to each other. Pruning is performed by eliminating any candidate point for which either of the two earlier time frames is also a candidate. Remaining candidates are deemed onsets, and linear interpolation between the activations of the candidate candidate $[H]_{k,n}$ and the earlier point in the activation vector, $[H]_{k,n-1}$ is used to estimate the onset time. As above, the threshold is estimated for a range of values of δ, the thresholding parameter.

2) Results: Table II displays the results for the framewise analysis. GS-β-NMD, for both KL and $\beta^{(0.5)}$, is seen to perform similar to H-NMF in framewise analysis. Improvements of $\sim 5\%$ for $\beta^{(0.5)}$ and $\sim 4\%$ for KL are observed using dictionary tuning. In the case of $\beta^{(0.5)}$, all metrics increase by 1% relative to β-NMD using the optimal dictionary, and also by $\sim 2\%$ relative to our previous results [17], using the monotonic descent algorithm.

The onset-based analysis results are given in Table III. We first note that the proposed onset detector performs between $4 \sim 7\%$ better than the original, with the largest increases for the NMD results using the dictionaries learnt offline and the smallest for GS-NMDs with the fixed AHD. H-NMF performs better than the GS-β-NMD approaches, by $\sim 2.5\%$. Using GS-β-NMF dictionary tuning, performance for both cost functions exceeds that of H-NMF. For $\beta^{(0.5)}$ the $\mathcal{F}$-measure is almost 5% higher than for H-NMF, and 2% lower than β-NMD while being similar to KL-NMD.

As GS-β-NMF was run for a fixed number of iterations, rather than to a convergence condition, a subsequent GS-β-NMD is run using the new tuned dictionary, for each piece, with the purpose of confirming that a better dictionary has been learnt, rather than a favourable local minima found. The results for these post-NMD approaches are seen in Table IV. Here it is seen that the results acheived using dictionary tuning with $\beta^{(0.5)}$ are almost maintained by post-NMD using either KL or $\beta^{(0.5)}$. However, the framewise results achieved by dictionary tuning using KL are seen to degrade with post-NMD.

D. Discussion

The evaluation validates the proposed approaches; the subspace model is seen to be similar to the larger datapoint models, and the backwards elimination strategy improves over NNLS and OMP. More significantly, the dictionary tuning approach improves over H-NMF [4], and performs almost as well as β-NMD with a dictionary learnt offline. Furthermore, the modified onset detector leads to improvements.

Further AMT experiments were performed to directly compare the stepwise and NMF-based methods. First, GBF-NNLS and GS-β-NMD were run with ERBT subspace dictionaries. GBF-NNLS, with a similar datapoint dictionary, is also compared. GBF-NNLS and GS-β-NMD were then used for post

	Frames			Onsets		
	DT	$\beta^{(0.5)}$	KL	DT	$\beta^{(0.5)}$	KL
$\beta^{(0.5)}$	72.9	72.7	73	80.3	79.6	79.4
KL	71.8	69.5	70.3	79.1	78.0	78.5

TABLE IV

$\mathcal{F}$-MEASURE FOR FRAMEWISE AND ONSET ANALYSIS WITH POST-TUNING NMD. VALUES ON LEFT GIVE COST FUNCTION USED FOR DICTIONARY TUNING. DT INDICATES THE RESULTS FROM THE DICTIONARY TUNING. OTHER COLUMN HEADS DESCRIBE COST USED FOR POST-TUNING NMD.

	Frames	Onsets
GBF-NNLS (D)	73.3	76.9
GBF-NNLS (S)	73.2	76.1
GS-β-NMD (S)	73.8	82.0
GS-KL-NMD (S)	74.1	83.2
GS-β-NMD (T)	72.7	79.6
GBF-NNLS (T)	69.9	71.8
NN-NS-OMP (T)	69.7	74.4

TABLE V
$\mathcal{F}$-MEASURE FOR FRAMEWISE AND ONSET ANALYSIS COMPARING GBF-NNLS WITH GS-β-NMD. (D) DENOTES DATAPOINT; (S) DENOTES SUBSPACE (S), AND (T) DENOTES DICTIONARIES TUNED USING GS-β-NMF

dictionary tuning NMD, using AHDs tuned for each piece using GS-β-NMF. NNLS was observed not to converge with the tuned AHDs, which was due to narrowband atoms from different pitches of the AHD being highly correlated to each other. NN-NS-OMP was instead used for the initial decomposition, selecting a maximum of 22 groups, with results given. A similar experimental setup to *Exp. C*, with framewise and onset analysis, was employed, For the subspace dictionaries, results given are for the optimal P for each algorithm.

The results are given in Table V. GBF-NNLS performs similar to GS-β-NMD algorithms in terms of framewise measures with the subspace dictionaries. However, in terms of onset measures, and also framewise measures with the AHDs, the performance of GBF-NNLS is lesser relative to GS-β-NMD. In the case of the tuned AHDs, some error may be effected by the correlated elements that caused convergence problems for NNLS. Alternatively, the results may be interpreted in terms of the quality of model. The subspace dictionary provides a more descriptive model than the AHD by including implicit temporal information, while the approximate additivity assumed by NMF-based AMT may be less effective in the presence of transients and onsets [57]. From this perspective, GBF-NNLS performs well when the spectra can be well-approximated, as also seen in Table I, but is not so robust as the NMF-based methods which outperfom GBF-NNLS in other cases.

Framewise detection is improved by group sparse NMD with the subspace dictionaries relative to standard NMD for both KL and $\beta^{0.5}$, as seen in Table II, and are also above that seen with dictionary tuning. These results suggest that there is some room for improvement in dictionary tuning, or learning. The improvement is larger for GS-KL-NMD, for which onset-based metrics also increase. We suggest that KL is better than $\beta^{(0.5)}$ for GS-NMD, with $\beta^{(0.5)}$ superior for dictionary tuning, which we also observe to a greater extent in experiments not reported here. The concave penalty used with $\beta^{(0.5)}$ may improve the dictionary updating as low-energy elements are penalised more while convexity of the $\ell_{2,1}$-penalised KL-divergence may be preferable for performing GS-β-NMD.

A common feature of many AMT methods is the use of temporal information. Temporal information is an obvious prior for audio signals, and some NMF approaches perform frame to frame penalisation to encourage smoothness [33] [14] [54] [58]. While smoothing is considered useful for source separation [33], its validity in the context of AMT for piano is questioned in [14] [54], where little or no improvement is reported. A large improvement in AMT is reported when temporal constraints are employed to transition between different states of piano notes [7], which may be due to the weaker shift-invariant note model used. Such methods may be improved by constrained training of a subspace dictionary, with states represented by active subsets of atoms, with overlapping subsets for adjacent states. Such staggered activation patterns are often implicitly present in the subspace dictionaries, and may be observed in spectrogram decompositions, even in a polyphonic setting. In terms of dictionary tuning, we consider that it may be possible to learn this implicit temporal information by using one, or more, constrained group-wise subspace learning steps, in a manner akin to the Block K-SVD [59], rather than gradual updating, early in the dictionary tuning approach.

	50 ms	100 ms	100 ms (Best δ)
GS-β-NMF	80.3	84.3	86.4
GS-KL-NMD	83.2	87.3	89.5

TABLE VI
$\mathcal{F}$-MEASURE FOR ONSET DETECTION WITH DIFFERENT WINDOW SIZES.

The use of long-term temporal dependencies for AMT is considered necessary in [60]. Structured sparse decomposition methods may provide further possibilities. It is easy to observe in NMF-based AMT the problem of low-energy elements of sustained notes being overpowered by false positives related to higher energy active notes [57]. We previously observed improved AMT, in both analyses, simply by using a low offset threshold, clustering adjacent active atoms into molecules, and subsequently determining activation over a whole molecule rather than for individual atoms [15] [37], using stepwise methods. An enhanced clustering approach, possibly using an excitation-decay model may be considered. However, this may require reliable onset detection, itself one of the primary goals of AMT.

In order to probe the potential for improved onset detection, we further inspect the proposed onset detector by re-running GS-β-NMF with AHD and GS-KL-NMD with the subspace dictionary, with 50 ms and 100 ms onset tolerances compared. Results are given in Table VI, where a jump of $\sim 4\%$ is seen for the larger tolerance. A further increase of $\sim 2\%$ is observed when the optimum δ per song is used. We note a tendency towards higher precision than recall, as also seen in Table III. Careful consideration, through alignment to the spectrogram, or leveraging onset classification methods [53] [61] may possibly bridge this performance jump.

VI. CONCLUSIONS

In this paper the use of group sparsity with subspace modelling for piano transcription was explored. Non-negative group sparse algorithms, a dictionary tuning approach, and an onset detector for NMF-based AMT were proposed and experimentally validated. The subspace model performed similar to a datapoint model, and an elimination based stepwise method was seen to counter noted problems of OMP in this context. A proposed monotonic group sparse β-NMF with a scale invariant penalty term was used for tuning a harmonic dictionary previously proposed in [4], leading to improved NMF-based AMT. In particular, dictionary tuning counters the problems of NMD, as the dictionary is adapted to the signal,

and unsupervised NMF as rank selection and separability problems are avoided. Some possibilities for incorporating temporal information into the group sparse methods for further improving piano transcription were then discussed.

Further work will include extending the range of grouping strategies, allowing different problems to be considered. One potential application is for multi-instrument signals [12] [62], where grouping could be performed on instrument and pitch-labels. NMF-based decompositions tend towards co-activity of instruments in this case and stepwise methods incorporating penalty terms that encourage e.g. temporal continuity of a given instrument-pitch combination may provide an alternative perspective.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for useful comments, and the authors of [50] [4] for making the code for H-NMF freely available, in particular to Roland Badeau for discussions during his visit to the Centre for Digital Music. The code to reproduce the experiments in the paper is available at *https://code.soundsoftware.ac.uk/projects/gs_bnmf*.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS 14)*, Denver, 2000, pp. 556–562.

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

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, 2007, pp. 65–68.

[4] 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, March 2010.

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

[6] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, 2004, pp. 318–325.

[7] E. Benetos and S. Dixon, "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, Winter 2012.

[8] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011.

[9] S. K. Tjoa and K. J. Ray Liu, "Factorization of overlapping harmonic sounds using approximate matching pursuit," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, 2011, pp. 257–262.

[10] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 1993, vol. 1, pp. 40–44.

[11] J. J. Carabias-Orti, P. Vera-Candeas, F. J. Canadas-Quesada, and N. Ruiz-Reyes, "Music scene adaptive harmonic dictionary for unsupervised note-event detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 473–486, March 2010.

[12] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 116–128, January 2008.

[13] S. Raczynski, N. Ono, and S. Sagayama, "Extending non-negative matrix factorisation - a discussion in the context of multiple frequency estimation of musical signals," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, 2009, pp. 934–938.

[14] 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, March 2010.

[15] K. O'Hanlon, H. Nagano, and M. D. Plumbley, "Structured sparsity for automatic music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, 2012, pp. 441–444.

[16] N. Keriven, K. O'Hanlon, and M. D. Plumbley, "Structured sparsity using backwards elimination for automatic music transcription," in *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, 2013, pp. 1–6.

[17] 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)*, Florence, 2014, pp. 3112 – 3116.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, December 1998.

[19] Y. C. Eldar, P. Kuppinger, and H. Bolsckei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[20] A. Ganesh, Z. Zhou, and Y. Ma, "Separation of a subspace sparse signal: Algorithms and conditions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, 2009, pp. 3141–3144.

[21] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, vol. 68, no. 1, pp. 49–67, February 2006.

[22] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[23] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of non-negative sparse solutions to underdetermined systems of equations," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4813–4820, November 2008.

[24] D. Chen and R. J. Plemmons, "Nonnegativity constraints in numerical analysis," in *A. Bultheel and R. Cools (Eds.), Symposium on the Birth of Numerical Analysis*. 2009, pp. 109–140, World Scientific Press.

[25] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice Hall, 1974.

[26] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

[27] J. Rapin, J. Bobin, A. Larue, and J. Starck, "Robust non-negative matrix factorization for multispectral data with sparse prior," in *Proceedings of the 7th Conference on Astronomical Data Analysis (ADA 7)*, Cargese, 2012.

[28] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Proceedings of the SPIE conference (Wavelets XI)*, Baltimore, 2005, pp. 327–339.

[29] M. Slawski and M. Hein, "Sparse recovery by thresholded non-negative least squares," in *Advances in Neural Information Processing Systems (NIPS 24)*, Granada, 2011, pp. 1926–1934.

[30] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, March 2009.

[31] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended smart algorithms for non-negative matrix factorization," *Lecture notes in Artificial Intelligence, 8th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, vol. 4029, pp. 548–562, 2006.

[32] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, November 2004.

[33] T. Virtanen, "Monaural sound source separation by non-negative matrix factorisation with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[34] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592 – 1605, July 2013.

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

[36] K. O'Hanlon and M. D. Plumbley, "Non-negative group sparsity," in *Proceedings of the IMA Conference on Numerical Linear Algebra and Optimisation*, Birmingham, 2012.

[37] K. O'Hanlon, *Automatic Music Transcription using structure and sparsity*, Ph.D. thesis, Queen Mary University of London, 2013.

[38] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions in Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[39] B. L. Sturm and M. G. Christensen, "Cyclic matching pursuits with multiscale time-frequency dictionaries," in *Conference Record of the 44th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2010, pp. 581–585.

[40] H. Huang and A. Makur, "Backtracking-based matching pursuit method for sparse signal reconstruction," *IEEE Signal Processing Letters*, vol. 18, no. 7, pp. 391–394, July 2011.

[41] B. Varadarajan, S. Khudanpur, and T. D. Tran, "Stepwise optimal subspace pursuit for improving sparse recovery," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 27–30, January 2011.

[42] B. Moghaddam, A. Gruber, Y. Weiss, and S. Avidan, "Sparse regression as a sparse eigenvalue problem," in *Information Theory and Applications Workshop (ITA)*, San Diego, 2008, pp. 219 –225.

[43] A. Lefevre, F. Bach, and C. Fevotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[44] M. A. T. Figueirido, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2980–2991, November 2007.

[45] U. Simsekli, A. T. Cemgil, and Y. K. Yilmaz, "Learning the beta-divergence in Tweedie compound Poisson matrix factorization models," in *Proceedings of The 30th International Conference on Machine Learning (ICML)*, Atlanta, 2013.

[46] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition.," in *INTERSPEECH*, 2012.

[47] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative update algorithms for nonnegative matrix factorization with the β-divergence," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Kittila, 2010, pp. 283–288.

[48] J. De Leeuw and G. Michailidis, "Drawing data graphs by pushing and pulling," http://gifi.stat.ucla.edu/janspubs/1999/notes/deleeuw_michailidis_U_99c.pdf, 1999.

[49] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3437 – 3440.

[50] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorisation for polyphonic music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, 2008, pp. 109–112.

[51] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, 2002, pp. 1757 – 1760.

[52] 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, August 2010.

[53] S. Bock 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)*, Kyoto, 2012, pp. 121–124.

[54] N. Bertin, R. Badeau, and E. Vincent, "Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription.," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 29–32.

[55] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2277–2289, November 2013.

[56] P. Smaragdis, "Polyphonic pitch tracking by example," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011.

[57] K O'Hanlon, H. Nagano, and M. D. Plumbley, "Oracle analysis for automatic music transcription," in *Proceedings of 9th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, London, 2012.

[58] M. Nakano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Infinite-state spectrum model for music signal analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2011, pp. 1972–1975.

[59] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar, "Dictionary optimization for block-sparse representations," *Signal Processing, IEEE Transactions on*, vol. 60, no. 5, pp. 2386–2395, 2012.

[60] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchoff, and Anssi Klapuri, "Automatic music transcription: Breaking the glass ceiling," in *Proceedings of the 13th Conference of the International Society for Music Information Retrieval (ISMIR)*, Porto, 2012, pp. 379–384.

[61] F. Weninger, C. Kirst, Bjorn B. Schuller, and H. J. Bungartz, "A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6–10.

[62] E. Benetos, S. Ewert, and T. Weyde, "Automatic transcription of pitched and unpitched sounds from polyphonic music," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.