Structured sparsity for automatic music transcription

O'Hanlon, K; Nagano, H; Plumbley, MD; International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)

STRUCTURED SPARSITY FOR AUTOMATIC MUSIC TRANSCRIPTION

Ken O'Hanlon ⋆ *Hidehisa Nagano*⋆† *Mark D. Plumbley* ⋆

⋆Queen Mary University of London
Centre for Digital Music
† NTT Communication Science Laboratories
NTT Corporation

ABSTRACT

Sparse representations have previously been applied to the automatic music transcription (AMT) problem. Structured sparsity, such as group and molecular sparsity allows the introduction of prior knowledge to sparse representations. Molecular sparsity has previously been proposed for AMT, however the use of greedy group sparsity has not previously been proposed for this problem. We propose a greedy sparse pursuit based on nearest subspace classification for groups with coherent blocks, based in a non-negative framework, and apply this to AMT. Further to this, we propose an enhanced molecular variant of this group sparse algorithm and demonstrate the effectiveness of this approach.

Index Terms— Transcription, non-negative, structured sparsity

1. INTRODUCTION

Sparse coding is the attempt to derive a representation vector, $\mathbf{t}$, predominated by zero coefficients, or sparse, of a signal $\mathbf{s}$ from a given dictionary $\mathbf{D}$, where $\mathbf{s} \approx \mathbf{Dt}$. In this work, we consider sparse approximation in a non-negative framework. Formally stated, for the noisy case, the non-negative sparse representation problem seeks the minimisation:

$$\min_t \frac{1}{2}\|\mathbf{s} - \mathbf{Dt}\|_2^2 + \lambda\|\mathbf{t}\|_0 \; s.t. \; \mathbf{t} \geq \mathbf{0}. \tag{1}$$

Many algorithms have been proposed for performing sparse approximation, the most popular families of which are greedy methods, which typically add the atom most correlated with the residual signal at each iteration, and global optimisation methods, which relax the ℓ_0 norm sparse penalty for a ℓ_1 norm sparse penalty, allowing the problem to be solved with convex optimisation methods.

Structured sparsity allows the introduction of prior knowledge to sparse approximation algorithms. This knowledge can be general as in block, or group sparsity [1][2], where the assumption is made that groups of atoms tend to be supported coincidentally, or application specific, such as molecular pursuits [3] which have been applied to audio signals selecting time-related atoms at each iteration.

Musical signals can be seen as inherently sparse in that only a small subset of notes are seen to be active at a given time. Automatic music transcription is the attempt to produce a mid-level pitch-time representation, often called a piano roll, which relates the note events in a musical piece. Resarch in AMT is often delineated into methods which use either offline training or online training. For instance, a state-of-the art online method using Bayesian NMF with time-smoothness and harmonic constraints is proposed in [4]. There exists several works in the literature which aim to use sparse representation methods with dictionaries learnt offline to tackle the AMT problem. For example, Abdallah and Plumbley [5] used a non-negative sparse coding algorithm based on a Bayesian formulation of sparsity to produce a piano roll using a dictionary of full-spectrum basis atoms. Leveau et al. [6] proposed using Matching Pursuit (MP) with instrument and note specific dictionaries of harmonic atoms for instrument recognition, and also used a molecular representation for polyphonic transcription.

We are currently developing a non-negative sparse dictionary learning based AMT system, and the work presented here considers the sparse approximation step of the proposed system. It was observed in [5] that transcription is improved using several atoms per note, better capturing the spectral dynamics of a note. To this end we propose using group sparsity, with groups of pitch-related atoms. A variant of Non-Negative Basis Pursuit (NN-BP) which outputs group coefficients is considered. We propose a non-negative group greedy pursuit tailored for the coherent groups we use, which we term Non-negative Nearest Subspace OMP (NN-NS-OMP). We also propose a molecular version of NN-NS-OMP, promoting time persistence in the sparse representation. We demonstrate improved transcription performance using these structured approaches on a dataset of polyphonic piano pieces.

2. BACKGROUND

2.1. Non-negative sparse representations

In [7], Hoyer proposed a non-negative sparse coding (NNSC) algorithm, based on a multiplicative update,

$$t_{k,n} \longleftarrow t_{k,n} \frac{[\mathbf{D}^{\mathrm{T}}\mathbf{S}]_{k,n}}{[\mathbf{D}^{\mathrm{T}}\mathbf{DT}]_{k,n} + \lambda}$$

which seeks to minimise the cost function

$$\min_T \frac{1}{2}\|\mathbf{S} - \mathbf{DT}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{T}\|_{1,1} \; s.t. \; \mathbf{T} \geq 0 \tag{2}$$

which is seen to be similar to the ℓ_1 relaxation of the sparse representation problem with an added non-negativity con-

This research is supported by ESPRC Leadership Fellowship EP/G007144/1 and EU FET-Open Project FP7-ICT-225913 "SMALL"

straint. In [8], NN-BP was proposed, using several iterations of NNSC, followed by a hard thresholding.

A non-negative variant of OMP (NN-OMP), outlined in *Algorithm 1* is proposed in [9] which differs from OMP by constraining the atom selection to positive coefficients, and by using non-negative least squares(NNLS) to calculate the coefficients of the supported atoms at each iteration.

Algorithm 1 NN-OMP [9]

Input
$\mathbf{D} \in \Re_+^{M\times K}$**;** $\mathbf{s} \in \Re_+^{\mathbf{M}}$
Initialise
$i = 0$**;** $\mathbf{r}^0 = \mathbf{S}$**;** $\mathbf{t}^0 = \mathbf{0}$**;** $\Gamma^0 = \{\}$**;**
repeat
$i = i + 1$
$\hat{k} = \arg\max_k \langle \mathbf{d}_k, \mathbf{r}^{\mathbf{i-1}} \rangle$
$\Gamma^i = \Gamma^{i-1} \cup \hat{k}$
$\mathbf{t}_{\Gamma^i} = \min_z \|\mathbf{D}_{\Gamma^i}\mathbf{z} - \mathbf{s}\|_2^2;\ \mathbf{z} \geq 0$
$\mathbf{r}^{\mathbf{i}} = \mathbf{s} - \mathbf{D}\mathbf{t}^{\mathbf{i}}$
until stopping condition met

2.2. Structured Sparsity

Group sparsity assumes that groups of atoms tend to be active at the same time. Usually it is assumed that the groups are of the same size and adjacent in a given dictionary, $\mathbf{D} \in \Re^{M\times K}$ with $K = L \times P$ where L is the number of groups and P is the amount of atoms per group, allowing us to define the set of group indices:

$$G = \{G_l | G_l = \{P \times (l-1) + 1, ..., P \times l\} \forall l \in \{1, ..., L\}\}.$$

Several variants of OMP, such as Block-OMP (B-OMP) [1] and Subspace Matching Pursuit (SMP) [2] have been proposed which incorporate the group assumption. These algorithms differ from OMP by using group selection criteria and by adding all members of a elected group to the sparse support at each iteration. In B-OMP the group selection criterion is given by

$$\hat{l} = \arg\max_l \|\mathbf{D}_{G_l}^T \mathbf{r}^{\mathbf{i-1}}\|_2 \tag{3}$$

while the SMP selection criterion is given by

$$\hat{l} = \arg\min_l \|\mathbf{r}^{\mathbf{i-1}} - \pi_l(\mathbf{r}^{\mathbf{i-1}})\|_2 \tag{4}$$

where $\pi_l(\mathbf{y})$ is the projection operator of the signal y onto the subspace $\mathbf{D}_{G_l}$.

Daudet [3] proposed Molecular Matching Pursuit (MMP) for separation of transient, tonal and noise elements of an audio signal. A modified matching pursuit selects transient molecules consisting of wavelet trees, or tonal molecules consisting of time-persisting fourier atoms. To isolate spurious noise in the Fourier domain, tonal atoms were selected from a time-smoothed spectrogram with coefficents based on the values of several time related atoms. When a tonal atom is selected, a molecule is "grown" by tracking in both time directions until the coefficient values in the original spectrogram disappeared below a threshold. Subsequently all atoms in the molecule are added to the sparse support.

3. METHOD

3.1. Dictionary Training

We use a simple dictionary training method, learning subdictionaries each of size P atoms using NMF[10] on STFTs of isolated notes from the MAPS [11] database. We learn a subdictionary for each of $L = 88$ notes corresponding to MIDI notes 21 to 108, composing our block-based dictionaries from the note-specific subdictionaries. An alternative viewpoint is to regard each subdictionary as a subspace and the dictionary as a union of subspaces.

3.2. Non-negative structured sparse approximation

We modify slightly the NN-BP algorithm, imposing group structure post-hoc by deriving a group coefficient matrix, $\mathbf{GC} \in \Re^{\mathrm{M}\times\mathrm{L}}$ by summing the group coefficients. A hard threshold is then determined by a parameter δ and the largest entry in $\mathbf{GC}$, to which it is then applied elementwise.

Algorithm 2 NN-BP(GC)

Input
$\mathbf{D} \in \Re_+^{M\times K}$ **,** $\mathbf{S} \in \Re_+^{M\times N}$ **,** δ**,** $\mathbf{T}^0 = \mathbf{D}^{\mathrm{T}}\mathbf{S}$
repeat
$t_{k,n} \longleftarrow t_{k,n} \dfrac{[\mathbf{D}^{\mathrm{T}}\mathbf{S}]_{k,n}}{[\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{T}]_{k,n} + \lambda}$
until stopping condition met
$\mathbf{GC}_{l,n} = \sum \mathbf{T}_{G_l,n}\ \forall(l,n)$
$\mathbf{GC}_{l',n'} = 0\ \forall\{l',n'\}\ s.t.\ \mathbf{GC}_{l',n'} < \delta\|\mathbf{GC}\|_{\infty,\infty}$

Our subdictionaries tend to display high inner-group coherence, or sub-coherence [1] due to both non-negativity and harmonic relationship. For this reason, the B-OMP [1] selection criteria (3) is unsuitable for this problem. Due to the lack of a explicit projection operator, such as $\pi(\mathbf{y})$ in (4), in the non-negative framework used in these experiments the SMP [2] selection criteria (4) is also unsuitable. We propose a similar greedy group sparse algorithm, NN-NS-OMP, outlined in *Algorithm 3* which seeks the group containing the nearest subspace with non-negative coefficients at each iteration.

Algorithm 3 NN-NS-OMP

Input
$\mathbf{D} \in \Re_+^{M\times K}$ **,** $\mathbf{s} \in \Re_+^M$
Initialise
$\Gamma = \{\}$ $\quad \mathbf{r}^0 = \mathbf{s}$ $\quad i = 0$
repeat
$i = i + 1$
$\hat{l} = \arg\min_l \|\mathbf{r}^i - \mathbf{D}_{G_l}\Theta\|_2^2\ s.t.\ \Theta \geq 0\ \forall l = \{1, ..., L\}$
$\Gamma_i = \Gamma_{i-1} \cup \mathbf{D}_{G_{\hat{l}}}$
$\mathbf{t}_{\Gamma_i} = \arg\min_z \|\mathbf{s} - \mathbf{D}_{\Gamma_i}\mathbf{z}\|_2^2$
$\mathbf{r}^i = \mathbf{s} - \mathbf{D}\mathbf{t}^i$
until stopping condition met

We also propose a molecular version of this algorithm, M-NN-NS-OMP, outlined in *Algorithm 4*, which accepts an input binary atom support, Γ derived from $\mathbf{GC}$ output from NN-BP(GC). The reasons for inputting Γ are twofold; affording cheaper computation, as we need only calculate NNLS for atoms supported in Γ at each iteration, instead of all atoms

as in NN-NS-OMP. Also we no longer rely on a threshold, as in MMP, to stop growing the molecule, instead using discontinuities in Γ to delineate the endpoints of a molecule.

This algorithm perfoms sparse approximation on the matrix as a whole, in which it differs from NN-NS-OMP which operates on vectors. In this case we do not select a nearest subspace per se, but derive a group coefficient Θ from the sum of the NNLS coefficient vector $\mathbf{x}$ for the group at each time bin supported in Γ. Similar to tonal elements in MMP[3], a smoothed coeffcient matrix $\bar{\Theta}$ is derived with a rectangular window, of size α. We select the largest atom coefficient from $\bar{\Theta}$, from which we grow a molecule using Γ to define the endpoints, as mentioned above.

Algorithm 4 M-NN-NS-OMP

Input
$\mathbf{D} \in \Re_+^{M\times K}$, $\mathbf{S} \in \Re_+^{M\times N}$, $\Gamma \in \{0,1\}^{L\times N}$, G, α
Initialise
$i = 0$; $\Phi = 0^{L\times N}$; $B = \{\beta_n | \beta_n = \{\} \forall n \in \{1,..,N\}\}$
repeat
$i = i + 1$
Get group coefs Θ and smoothed coefs $\bar{\Theta}$
$\mathbf{x}_{G_l,n} = \arg\min_x \|\mathbf{r}_n^i - \mathbf{D}_{G_l}\mathbf{x}\|_2^2 \ s.t.\ \mathbf{x} \geq 0 \ \forall l \in \Gamma_n$
$\Theta_{l,n} = \|\mathbf{x}_{G_l,n}\|$; $\bar{\Theta}_{l,n} = \sum_{n'=n}^{n+\alpha-1} \Theta_{l,n'}/\alpha$
Select initial atom and grow molecule
$\{\hat{l}, \hat{n}\} = \arg\max_{l,n} \bar{\Theta}_{l,n}$
$n_{min} = \min \bar{n} \ s.t.\ \Gamma_{\hat{l},\Xi} = 1, \ \Xi = \{\bar{n}, ..., \hat{n}\}$
$n_{max} = \max \bar{n} \ s.t.\ \Gamma_{\hat{l},\Xi} = 1, \ \Xi = \{\hat{n}, ..., \bar{n}\}$
$\beta_n = \beta_n \cup \hat{l} \ \forall n \in \Xi = \{n_{min}, ..., n_{max}\}$
Calculate current coefficients and residual
$\mathbf{t}_{G_{\beta_n},n} = \min_t \|\mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}\|_2^2 \quad \forall n \in \Xi$
$\mathbf{r}_n^{i+1} = \mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}_{G_{\beta_n}} \ \forall n \in \Xi$
until stopping condition met

4. EXPERIMENTS

We perform transcription on a subset of MAPS [11], a database of midi-aligned piano pieces, also used in [4]. The database contains many sets of piano pieces, most recorded using high quality samples, and all pieces come with a standardised ground truth. The subset we used was recorded live on a Disklavier, and, similar to [4] we use the first thirty seconds of each piece. We downsampled each piece in the subset to $22.05kHz$, and used the STFT with a window size of $90ms$ with a 50% overlap to produce a magnitude spectrogram.

MAPS also contains samples of isolated chords and notes, from which we trained our subdictionaries. We trained dictionaries for various sizes of $P = 1$ to 5 in order to investigate the possible effect on performance of group size. We note that when $P = 1$, the algorithms revert to non-group sparse implementations.

We performed sparse approximation on the spectrograms using NN-BP(GC), NN-NS-OMP and M-NN-NS-OMP. We set NN-BP(GC) to run for 100 iterations. For NN-NS-OMP, we set three stopping conditions; at each time bin we allowed no more than 7 atoms to be selected; once the residual error norm dropped below 5% of the original signal norm; we also stopped when adding a new atom would reduce the residual norm by less than 2% of the original signal norm. For the NN-BP and NN-NS-OMP, we perform some post-processing on **GC**. We assume a minimum note length of 3 time bins and threshold out atoms which are continuously supported in the time domain for a lesser duration. We stopped the M-NN-NS-OMP, when $\|\bar{\Theta}\|_{\infty,\infty}$ dropped below a threshold of 2, and a persistence factor $\alpha = 5$ was used.

	Onset based			Frame based		
	Acc	Rec	F	Acc	Rec	F
NN-BP	79.9	71.8	75.6	61.2	76.0	67.8
NN-BP (GC)	76.8	73.1	74.9	50.3	82.6	62.5
NN-OMP	75.8	72.9	74.3	73.8	60.7	66.6
NN-NS-OMP	73.2	73.6	73.4	77.9	61.4	68.7
M-OMP	78.3	74.3	76.3	69.1	73.6	71.3
M-NN-NS-OMP	78.8	77.3	78.1	71.8	79.3	75.3
Marolt[13]	63.7	53.6	58.0	-	-	-
B-NMF[4]	46.6	45.3	45.0	-	-	-

Table 1. Transcription results comparing group sparse methods ($P = 3$) with non-group sparse methods. Results for Marolt method and B-NMF from [4] shown for comparision

We used both onset-based and a frame based measures to compare the performance of the algorithms. We detect onsets using a simple threshold-based method similar to [4]. An onset is detected when a threshold is exceeded and subsequently sustained for a minimum duration of three time bins. We set the threshold to an amplitude of 5.5, noting that better results can be obtained with other thresholds with respect to P. We register a true positive, tp when we find an onset for a note within 1 time bin of the ground truth, as ascertained from the aligned onset files supplied in MAPS. A false positive, fp is registered when the system detects an onset which is not present in the ground truth. False negatives, fn are recorded for each note in the ground truth which is not detected by the system.

We also record the frame-based performance, in which we compare the midi ground truth with the output from the sparse algorithms at each frame. Here we register a tp when a point in the time-frequency domain is supported by the ground truth and the transcription results. Similarly, points supported only in the ground truth, and transcription results register tn and fn respectively.

The following metrics are then used to measure performance; accuracy $Acc = \frac{tp\times 100}{tp+fp}\%$; recall $Rec = \frac{tp\times 100}{tp+fn}\%$ and the F-measure $F = \frac{2\times(Acc\times Rec)}{Acc+Rec}$.

5. RESULTS

In *Table 1*, we compare results for the group methods against their corresponding non-group methods. We note that the molecluar methods show the best results for F-value for both types of transcription, and that M-NN-NS-OMP improves upon the M-NN-OMP.

NN-BP(GC) performs worse with group coefficients for all measures except recall, with very poor accuracy for frame based metrics. This can be explained by the fact that the algorithm does not exploit the group structure in the multiplicative update, but uses a post hoc summation of group

P		BP(GC)	NN-NS-OMP	M-NN-NS-OMP
2	*Acc*	78.9	75.7	78.8
	Rec	72.2	73.1	76.2
	F	75.4	74.1	77.5
5	*Acc*	76.3	73.3	78.6
	Rec	73.2	74.2	77.8
	F	74.7	73.8	78.2

Table 2. Onset based metrics for group methods (abbreviated names) relative to group size P.

P		BP(GC)	NN-NS-OMP	M-NN-NS-OMP
2	*Acc*	53.5	75.0	69.0
	Rec	79.4	60.9	76.4
	F	63.9	67.2	72.5
5	*Acc*	48.8	78.7	72.9
	Rec	83.5	61.8	80.0
	F	61.6	69.2	76.3

Table 3. Frame Onset based metrics for group methods (abbreviated names) relative to group size P.

coefficients. However the recall is high, which prompts good results from the molecular approaches which input Γ from NN-BP(GC). In this way the M-NN-NS-OMP can be seen as a denoising step on the NN-BP(GC), capturing the salient parts of the signal.

The NN-NS-OMP shows slightly poorer performance in the onset detection task, but an improvement in the frame based detection. Informal experiments suggest that results are superior to other group versions of OMP we have tried, but that is not described here.

Finally a comparision is made with the results described for onset detection in [4]. Our results compare favourably with the neural network-based method proposed by Marolt in [13]. The results for the Bayesian NMF method are state-of-the-art for a signal decomposition based method, and direct comparision with offline-learning based methods is unfair, except to show the relative difficulty of the problems.

In *Tables 2 & 3*, we show results for different values of P, which indicate the effects of group size on performance. In particular we note the increase in almost all metrics for the molecular approach with group size.

6. CONCLUSIONS AND FURTHER WORK

We have a presented a structured non-negative sparse music transcription system, with promising results, demonstrating that group and molecular sparsity may enhance transcription performance. We intend to test the proposed method with datasets and metrics used by other researchers in AMT to allow a full comparision to be made.

7. REFERENCES

[1] Y.C. Eldar et al., "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. SP*, vol. 58, pp. 3042–3054, 2010.

[2] A. Ganesh et al., "Separation of a subspace=sparse signal: Algorithms and conditions," in *Proc. ICASSP 2009*.

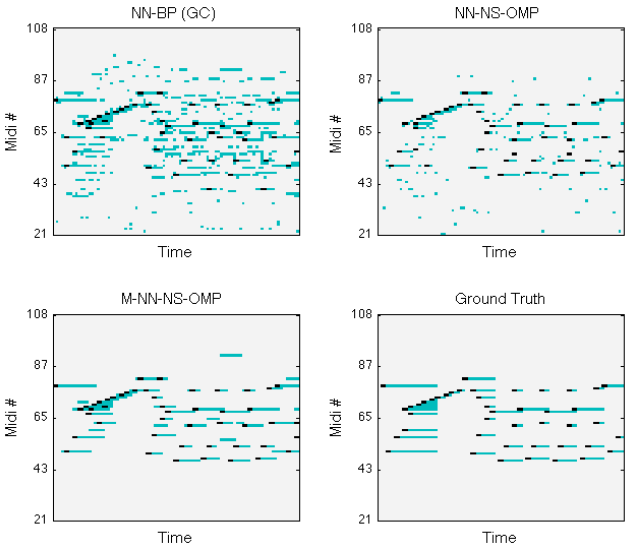

Fig. 1. Transcriptions results for with denoted algorithms and ground truth from 5 second excerpt, with $P = 3$. Detected onsets coloured in black (extended forward in time).

[3] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. ASLP*, pp. 1808–1816, 2006.

[4] N. Bertin et al., "Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription," *IEEE Trans. ASLP*, vol. 18, pp. 538–549, 2010.

[5] S.A. Abdallah and M.D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. ISMIR 2004*, 2004, pp. 318–325.

[6] P. Leveau et al., "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. ASLP*, pp. 116–128, 2008.

[7] P.O. Hoyer, "Non-negative sparse coding," *Proc. NNSP 2002*, pp. 557–565, 2002.

[8] M. Aharon et al., "K-svd and its non-negative variant for dictionary design," in *Proc. of the SPIE conference wavelets*, 2005, pp. 327–339.

[9] A.M. Bruckstein et al., "On the uniqueness of non-negative sparse solutions to underdetermined systems of equations," *IEEE Trans. IT*, vol. 54, pp. 4813–4820, 2008.

[10] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Proc. NIPS 2000*.

[11] V. Emiya et al., "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. ASLP*, vol. 18, pp. 1643–1654, Aug. 2010.

[12] E. Oja, *Subspace Methods of Pattern Recognition*, Res. Studies Press, 1983.

[13] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans on Multimedia*, pp. 439 – 449, 2004.