About this Non-Negative Business

Paris Smaragdis
paris@illinois.edu
10 years ago to the day ...

2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

October 19-22, 2003, New Paltz, NY

DINNER
6:00PM-7:20PM

SESSION N: RESYNTHESIS AND CROSS-SYNTHESIS
7:20PM-8:20PM

7:20pm  Rejection Phenomena in Inter-Signal Voice Transplantations
        Werner Verhelst and Henk Brouckxon, Vrije Universiteit Brussel, Brussels, Belgium

7:40pm  Discrimination of Sustained Musical Instrument Sounds Resynthesized With Randomly Altered
        Harmonic Amplitudes
        Andrew B. Horner, Hong Kong University of Science and Technology, Kowloon, Hong Kong
        James W. Beauchamp, University of Illinois at Urbana-Champaign, Urbana, IL, USA

8:00pm  Time-Scale Modification of Music Using a Subband Approach Based on the Bark Scale
        David Dorr, Dublin Institute of Technology, Dublin, Ireland
        Robert Lawlor, National University of Ireland, Maynooth, Ireland

SESSION O: MUSIC SIGNAL PROCESSING - MUSIC TRANSCRIPTION
8:20PM-9:20PM

8:20pm  Non-Negative Matrix Factorization for Polyphonic Music Transcription
        Paris Smaragdis, Mitsubishi Electric Research Lab, Cambridge, MA, USA
        Judith C. Brown, Wellesley College, Wellesley, MA, USA

8:40pm  Generative Model Based Polyphonic Music Transcription
        Ali Taylan Cemgil and Bert Kappen, University of Nijmegen, The Netherlands
        David Barber, Edinburgh University, UK
What is this talk about?

- What are all these “non-negative” papers?
- What is special about this approach?
- What can we do with it?
  - And why should we bother?
Traditional signal processing

- Axiom 1: “Thou shall love the Gaussian”
  - Why? It makes the math easy

- Gave rise to least squares models:

\[ y(t) = x(t) + n(t) \]

What we get  \( y(t) \)  
What we want  \( x(t) \)  
Gaussian noise  \( n(t) \)
A misunderstood model

- Abusing the noise model

\[ y(t) = x(t) + n(t) \]

- Other sounds are not Gaussian noise!
  - In fact neither is your target sound

"Other sounds"
And the impending revolution

- mid-90’s: The ICA community
  - Sources are not really Gaussian

- mid-2000’s: Compressive Sensing
  - Data is sparse in the right domain

- mid-2000’s: Non-Negative Models
  - We only care about positive-valued quantities
Picking a meaningful domain

- Waveforms are not that intuitive, we instead use spectrograms to examine audio signals.
Decomposing spectrograms

- What are the building blocks of spectrograms?
  - Standard question in machine learning

- The low-rank matrix factorization:

\[ X \approx WH \]
The usual suspect

- Principal Component Analysis: \( X = W \cdot H \)

orthonormal, decorrelated
Why is this result meaningless?

- This least-squares/Gaussian model is counter-intuitive for sound
  - Makes use of cross-cancellation

- We perceive scenes additively
  - We need an additive decomposition!
Non-Negative Matrix Factorization

- All factors are positive-valued: \( X \approx W \cdot H \)
- Resulting reconstruction is additive

\[ X = W \cdot H \]

Input music passage

Component

About this non-negative business
Why is this a better model?

1) It allows us to intuitively model sounds
   - All quantities mean something

2) The model parameters are additive
   - This also means we are invariant to mixtures

We can easily redefine previous work
   - And reap the benefits!
Wiener filtering / Spectral subtraction

- Learn “noise” spectrum, and filter/subtract
- And it doesn’t work with complex noises ...
- Extra complications due to negative values
The non-negative version

- Learning a sound model
  - An additive dictionary instead of a spectrum

\[ \mathbf{X} \approx \mathbf{W} \cdot \mathbf{H} \]

Linear combinations of these, explain these
Denoising

- Explain a mixture with the existing model
- Add new elements to explain the rest of the signal

Still the same model

\[ \mathbf{X} \approx [ \mathbf{W}_u \quad \mathbf{W}_k ] \cdot [ \mathbf{H}_u \quad \mathbf{H}_k ] \]
Reconstruction

- Parts-wise reconstruction:

\[ X = X_u + X_k \approx W_u \cdot H_u + W_k \cdot H_k \]

Spectrogram of unknown target

Spectrogram of known “noise”

Extracted target

Extracted "noise"
Why bother?

- Better statistical fit for the data
  - Results in better sounding outputs

- Flexible learning of “noise” model
  - No need to simply temporally segment
    - Spatial guidance, user guidance, TF guidance, ...

- Demo time!
Layer editing options

- Original drum loop
- Extracted layers
  - No tambourine
  - No congas
  - Congas!
- Remix

- Piano + Soprano
  - Soprano layer
  - Piano layer
- Remixed layers

- Music layer
- Voice layer

Selective pitch shifting
So what?

- We can resolve mixtures well
  - But what’s the use of that?
  - My mantra: “Separation is useless”

- What matters is the additivity of the model
  - Allows us to not care about mixing

\[
H_{x(t)+y(t)} \approx H_{x(t)} + H_{y(t)}
\]
Sound classification/detection

- Machine learning approaches are a poor fit
  - Can’t use winner-takes-all classification

- The real question: How active is each class?
  - Not whether it exists
A challenging example
The non-negative treatment

- Decompose as:

\[ x_t = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \cdot \begin{bmatrix} h_{1,t} \\ h_{2,t} \\ h_{3,t} \\ h_{4,t} \end{bmatrix} \]

- Energies in \( h \) express presence of each sound

\( h_{1,t} \), \( h_{2,t} \), \( h_{3,t} \), \( h_{4,t} \)
“Additive” sound recognition

- We can now find simultaneous sound classes
Adding the temporal dimension

- To be serious we should use Markov models.
- The non-negative HMM:
Advantages over GMM HMMs

- No need for factorial models
- Sum of models = model of sum of sounds
Speaker separation challenge

- WER doesn’t drop drastically with maskers

![Graph showing the performance of different scenarios]

- Clean
- -9dB
- -6dB
- -3dB
- 0dB

Legend:
- Baseline
- Speaker 2
- Speaker 1
Parameter estimation in mixtures

- Estimate parameters of only one sound in mix
  - Usually hard due to mixing

- Associate components with parameter
  - Learn on tagged data

- Explain new input with model
  - Use component / parameter association
Example: Pitch tracking

- Works fine on clean sounds
- Fails miserably on dense mixtures ...

![Pitch tracking example](image)
The non-negative pitch tracker

- Learn model from tagged data:
  \[ x_t \rightarrow p_t \]
  \[ x_t \approx W \cdot h_t \]

- Associate components & pitch:
  \[ P(W_i \rightarrow p_t) \propto h_{i,t} / \sum_i h_{i,t} \]

- Associate pitch to new inputs:
  \[ y_t \approx W \cdot h_t \]
  \[ P(y_t \rightarrow p_t) \propto \sum_i h_{i,t} p_i / \sum_i h_{i,t} \]

### Input music passage

<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Component Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>1.5</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>2000</td>
</tr>
<tr>
<td>2.5</td>
<td>3000</td>
</tr>
<tr>
<td>3</td>
<td>4000</td>
</tr>
<tr>
<td>3.5</td>
<td>5000</td>
</tr>
<tr>
<td>4</td>
<td>6000</td>
</tr>
</tbody>
</table>

### About this non-negative business
Result

- Sharp pitch probabilities on mixture

And also works for phonemes, sound class, loudness, and other parameters
And I could go on and on ...

- Echo-cancellation, dereverberation, multi-modal processing, missing data, convolutive models, tensor versions, ...

- Rich literature on non-negative models
  - Lots of WASPAA/ICASSP papers
So what is coming up next?

- **Theory:**
  - Problem definition, parameter estimation, convergence properties, variations and generative models, dynamical systems, ...

- **Practical directions:**
  - Multi-channel data formulations
  - Alternative TF front-ends
  - Efficient formulations for big data
Rethinking the array

- We can re-conceptualize beamforming
  - Example case: Lots of cell phones in concert
    - All recordings will be bad and non-synced

![Image of people holding cell phones at a concert](image-url)
A non-negative take

- Joint component analysis
  - Common components are of interest
  - Non-common components are noise
  - Optional priors from reference recordings
Example case

Original input

Lowpass & interference

Highpass & interference

Bandpass & clipping
Recovered signal

- Recovery of full bandwidth
- Suppression of uncommon elements
- Not sensitive to non-linearities/synchronization

Original input

Recovered signal
Alternative TF front ends

- The STFT has poor frequency resolution
  - We can do better with other transforms
    - Constant-Q, reassigned spectra, sinusoidal models, ...

- But that data is not in a matrix format!!
  - Reformulate NMF as a function approximation
  - Allows us to use arbitrary TF representations
Sinusoidal model example

Sinusoidal Modeling

Irregular Input

1st NMF Component from STFT

2nd NMF Component from STFT

1st Non-Regular Component

2nd Non-Regular Component
Reassigned spectra example

Long FFT window

Short FFT window

Heassigned spectogram

Clustered reassigned spectrum
NMF for big data

- How do we analyze huge recordings?
  - Operate on landmark space instead
To conclude

- The wild west is in non-negative models
  - Can they be the new Gaussian?

- A more perceptual take on analysis
  - Still on unclear math ground though

- Thanks!
  - And many thanks to Nick Bryan, Minje Kim, Gautham Mysore, Madhu Shashanka