Efficient Learning of Harmonic Priors for Pitch Detection in Polyphonic Music

Pablo A. Alvarado, Dan Stowell
Published in: arXiv.org

Centre for Digital Music
Queen Mary University of London
Introduction
  Automatic music transcription

Method
  Pitch detection model
    Activation process
    Component process
  Leave one out model
  Variational inference

Results
  Transcription of polyphonic music

Conclusions
Automatic music transcription (AMT)

AMT consists in updating our beliefs about the symbolic description (piano-roll) of a piece of music, after observing a corresponding audio recording [1, 2].

\[
p(\text{piano-roll}|\text{signal}) = \frac{p(\text{signal}|\text{piano-roll})p(\text{piano-roll})}{p(\text{signal})}.
\]

AMT [1].
Content

Introduction
  Automatic music transcription

Method
  Pitch detection model
    Activation process
    Component process
  Leave one out model
  Variational inference

Results
  Transcription of polyphonic music

Conclusions
We address the transcription problem from a time-domain source separation perspective, as in [3].

Given an acoustic signal $D = \{t_n, y_n\}_{n=1}^{N}$, we use the regression model

$$y(t) = \sum_{m=1}^{M} f_m(t) + \epsilon,$$

$$= \sum_{m=1}^{M} \phi_m(t)w_m(t) + \epsilon,$$

where the set of functions $\{\phi_m(t)\}_{m=1}^{M}$ and $\{w_m(t)\}_{m=1}^{M}$ are called activation processes, and quasi-periodic component processes respectively.
Introduction to Gaussian Processes

- Gaussian processes (GPs) are probability distributions over functions.

\[ g(t) \sim \mathcal{GP}(\mu(t), k(t, t')) . \]

- The covariance function or kernel \( k(t, t') \) defines the properties of the random function \( g(t) \), such as smoothness, frequency content.

- Any finite number of function evaluations \( \mathbf{g} = [g(t_1), \ldots, g(t_N)]^\top \) follows a multivariate normal distribution.
Activations are defined as non-linearly transformed GPs.

\[ y = \sum_{m=1}^{M} \phi_m(t)w_m(t) + \epsilon. \]
Activation process

**Sigmoid model:**
Activations correspond to

\[ \phi_m(t) = \sigma(g_m(t)), \]

where

\[ \sigma(x) = [1 + \exp(-x)]^{-1}, \]

and

\[ g_m(t) \sim \mathcal{GP}(0, \hat{k}_m(r)), \quad r = |t - t'|. \]
**Activation process**

**Softmax model:**
To introduce dependences between all activations we use the softmax function [4, 5]

\[ \phi_m(t) = \frac{\exp(g_m(t))}{\sum_{\forall j} \exp(g_j(t))}, \]

where

\[ \sum_{\forall m} \phi_m(t) = 1, \quad \text{for all } t, \]

and

\[ g_m(t) \sim \mathcal{GP}(0, \hat{k}_m(r)), \quad r = |t - t'|. \]
Component process

Component processes exhibit the frequency content of music notes.

\[ y = \sum_{m=1}^{M} \phi_m(t) w_m(t) + \epsilon. \]
The quasi-periodic processes follow

\[ w_m(t) \sim \mathcal{GP}(0, k_m(r)), \]

where

\[ k_m(r) = \sum_{\forall j} \sigma^2_{j,m} e^{-\lambda_j m r} \cos(\omega_{0j,m} r). \]

We seek to make

\[ \mathcal{F}\{k_m(r)\} \approx |\mathcal{F}\{y_m(t)\}|, \]

where \( y_m(t) \) corresponds to the audio recording of an isolated sound event with pitch \( m \).
Leave one out model (LOO)

Standard model:

\[ y = \sum_{m=1}^{M} \phi_m(t)w_m(t) + \epsilon. \]

LOO model:

\[ y(t) = \phi_i(t)w_i(t) + \phi_j(t)w_j(t) + \epsilon. \]

- Activations follow \( \phi_i(t) = \sigma(g_i(t)) \).
- Components follow

\[ w_i(t) \sim \mathcal{GP}(0, k_i(r)), \]

\[ w_j(t) \sim \mathcal{GP} \left( 0, \sum_{\forall m \neq i} k_m(r) \right). \]
Variational inference

\[ \text{posterior} = p(\text{components, activations}|\text{data}) \]

- GPs posteriors are in general computationally expensive to compute.
- In this case the posterior also does not have closed form.
Variational inference

\[ \text{approximation} = q(\text{components, activations} | \text{data}) \]

- GPs posteriors are in general computationally expensive to compute.
- In this case the posterior also does not have closed form.
- The key idea is to approximate to posterior with optimization [2].
- We first choose a family of probability distributions, then we try to find the member of that family closest to the exact posterior.
Toy example

audio signal = source_1 + source_2

Mixture signal.
Toy example

\[
\text{audio signal} = \text{source}_1 + \text{source}_2 \\
\text{source}_m = (\text{component process})_m \times (\text{activation process})_m
\]
Toy example

audio signal = source\(_1\) + source\(_2\)

source\(_m\) = (component process)\(_m\) \times (activation process)\(_m\)
Toy example

\[
\text{audio signal} = \text{source}_1 + \text{source}_2
\]

\[
\text{source}_m = (\text{component process})_m \times (\text{activation process})_m
\]
Toy example

audio signal = source_1 + source_2

source_m = (component process)_m \times (activation process)_m
Toy example

audio signal = source_1 + source_2

source_m = (component process)_m × (activation process)_m

Quasi-periodic process 1 (A4)

Envelope process 1 (A4)

Quasi-periodic process 2 (E5)

Envelope process 2 (E5)

Posterior after 50 iterations.
Toy example

audio signal = source_1 + source_2

source_m = (component process)_m \times (activation process)_m

Posterior after 500 iterations.
Toy example

audio signal = source₁ + source₂

sourceₘ = \((\text{component process})ₘ \times (\text{activation process})ₘ\)
Content

Introduction
   Automatic music transcription

Method
   Pitch detection model
      Activation process
      Component process
   Leave one out model
   Variational inference

Results
   Transcription of polyphonic music

Conclusions
Results: transcription of polyphonic music

Test data

- Synthetic electric guitar audio signal [3].
- Mixture sounds: C4, E4, G4, C4+E4, C4+G4, E4+G4, and C4+E4+G4.
- Duration 14 seconds, sampled at 16KHz.

![Test mixture signal.](image)

![Ground truth piano-roll.](image)
Results: transcription of polyphonic music

Training data
- Three isolated sound events with pitches C4 (261.63Hz), E4 (329.63Hz), G4 (392.00Hz), respectively.

Experiments
- Detection of pitches C4, E4 (using standard model & sigmoid (SIG) or softmaxv(SOF)).
- Detection of all three pitches C4, E4, G4 (using SIG-LOO).

Learning hyperparameters
- Tuned manually (TM).
- Maximising the marginal likelihood (ML).
- Reducing the MSE between $\mathcal{F} \{ k_m(r) \}$ and $|\mathcal{F} \{ y_m(t) \}|$ (FL, proposed method).
Results: transcription of polyphonic music

Pitch detection using SIG-LOO model.

Ground truth piano-roll.

Pitch detection using FL.

Pitch detection using ML.

Pitch detection using TM.
## Results: transcription of polyphonic music

| Model       | TM     | ML     | FL     |
|-------------|--------|--------|--------|
| SIG         | 89.54% | 59.23% | 98.68% |
| SOF         | 86.28% | 55.28% | 97.15% |
| SIG-LOO     | 76.21% | 84.86% | 98.19% |

**Table:** F-measure for SIG, SOF models detecting two pitches (first two rows), and F-measure for SIG-LOO model detecting three pitches (bottom row), using three different learning approaches: TM, ML, and FL.
Content

Introduction
   Automatic music transcription

Method
   Pitch detection model
      Activation process
      Component process
   Leave one out model
   Variational inference

Results
   Transcription of polyphonic music

Conclusions
Conclusions

▶ We proposed a GP regression approach for pitch detection in polyphonic music.
▶ We introduced Matérn mixture kernel able to reflect the complex frequency content of sounds of single notes.
▶ The proposed approach allows to introduce prior beliefs about smoothness, positive-values constrains, and correlation between activations.
▶ Pitch detection results suggest that a set of proper frequency content priors over of the sound events to be detected are more relevant than encouraging dependency between activations.
▶ The linear scalability of the LOO model regarding the number of pitches makes it appropriate to detect more than just three pitches.
[1] A. T. Cemgil, S. J. Godsill, P. H. Peeling, and N. Whiteley, “Bayesian statistical methods for audio and music processing,” *The Oxford Handbook of Applied Bayesian Analysis*, 2010.

[2] D. Blei, A. Kucukelbir, and J. McAuliffe, “Variational inference: A review for statisticians,” *arXiv preprint arXiv:1601.00670*, 2016.

[3] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, “Beyond nmf: Time-domain audio source separation without phase reconstruction.” in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
Thank you for your attention.
Questions?