The Power of Deep without Going Deep?
A Study of HDPGMM Music Representation Learning

tl;dr

- Bayesian nonparametric models can learn music representations as effectively as Deep Learning while being more interpretable.

Motivation

- In the late 2000s - early 2010s, the MIR community explored Bayesian Nonparametric (BN) models.
- After Deep Learning (DL), there are few works exploring BNs.
- BN can offer advantages that DL provides while being more interpretable.

Deep Learning vs. Bayesian Nonparametric

- **High learning capacity**:
  - *Universal approximation theorem vs. Nonparametric nature*
- **Robust to overfitting**:
  - *Dropout/Weight Decay/Augmentation/etc. vs. Bayesian nature*
- **Efficient learning algorithm**:
  - *SGD, ADAM, etc. vs. Online variational inference*
- **Can go "deep"**:
  - *Stacked layers vs. (nested) Hierarchical Dirichlet process prior*
- **Interpretability**:
  - *(almost) black-box vs. can be much better*

Contributions

- Insight into how "good" and transferable the HDPGMM representation is for MIR tasks.
- An implementation of a GPU-accelerated inference algorithm for HDPGMM. [1]

Hierarchical Dirichlet Process Gaussian Mixture Model (HDPGMM)

- Dirichlet Process (DP) can draw distributions of arbitrary dimensionality.
- One of the useful analogies to understand DP is the "stick-breaking" process:

$$\beta'_k \sim \text{Beta}(1, \gamma) \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l) \tag{1}$$

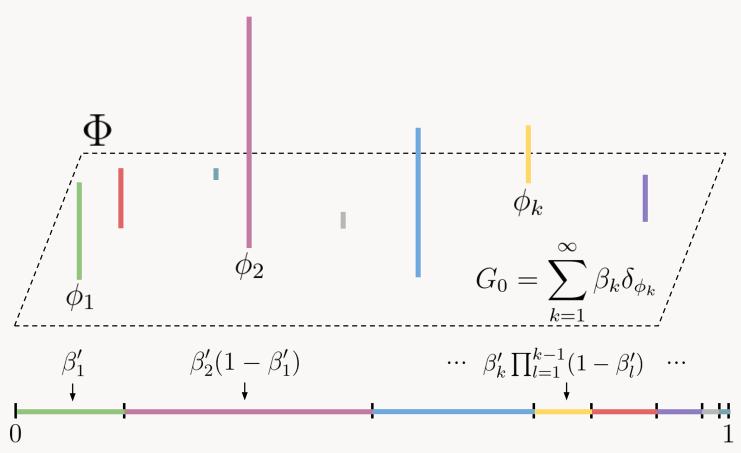

Figure: Illustration of stick-breaking construction

- When β is drawn in this way, we can refer it as $\beta \sim \text{GEM}(\gamma)$
- Employing DP prior as *mixing distribution*, DPMM can find an appropriate number of components for a given dataset.
- It is formally defined as follows:

$$\begin{aligned} \beta|\gamma &\sim \text{GEM}(\gamma) & \phi_k|H &\sim H \\ y_i|\beta &\sim \text{Mult}(\beta) & x_i|y_i, \phi_k &\sim F(\phi_{y_i}) \end{aligned} \tag{2}$$

- DPMM can be extended to the 2-level hierarchy, learning global and group-level components.
- Group naturally arises in many domains, including MIR problems (i.e., lyrics-words, artist-songs, song-time instance features)
- In this work, we set "corpus-level" time instance features as the upper level and "song" as a group of features, being the lower level.

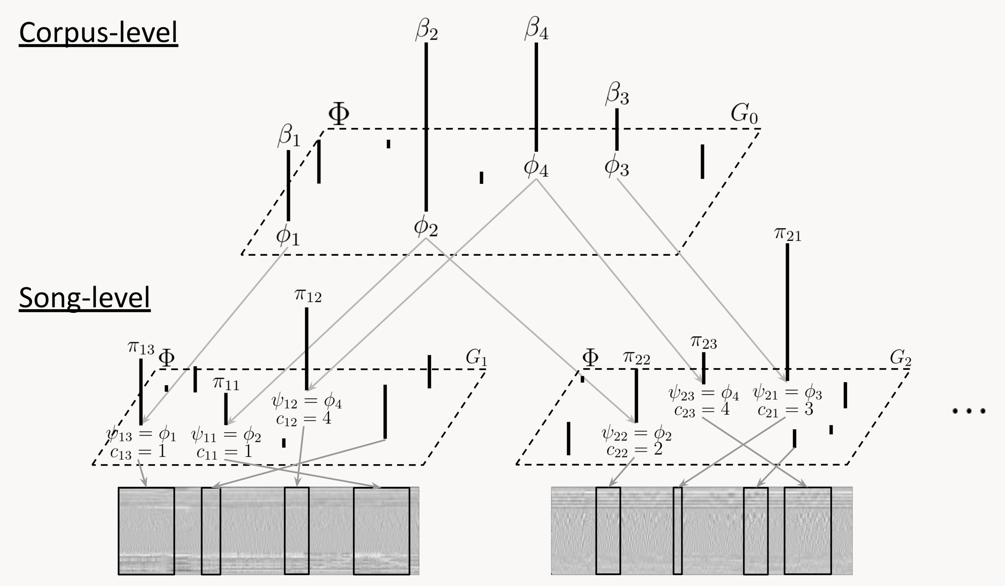

Figure: Illustration of HDP stick-breaking construction

- Song-level components "inherits" the global components with song-specific mixing coefficients π_j.
- Setting F as Gaussian-Inverse Wishart distribution and its parameters θ accordingly, we can model song features

$$\begin{aligned} \pi_j|\alpha_0 &\sim \text{GEM}(\alpha_0) & \theta_{jn} &= \psi_{jz_{jn}} = \phi_{c_{jz_{jn}}} \\ z_{jn}|\pi_j &\sim \text{Mult}(\pi_j) & x_{jn}|z_{jn}, c_{jt}, \phi_k &\sim F(\theta_{jn}) \end{aligned} \tag{3}$$

Inference (Training) / Regularization / Representation / Input Features

- **Online Variational Inference (OVI)** with the mean-field (fully-factorized) approximation.
- Additionally, we "splash" the uniform noise e to the inferred responsibility r_{jn} each time instance to account for the missing data due to the preview clipping.

$$\tilde{r}_{jn} = (1 - \eta_t) r_{jn} + \eta_t e \tag{4}$$

- We employ the (variational) expectation of log-likelihood of samples $\tilde{y}_{jk} = \exp(\mathbb{E}_q[\log p(X_j|c_j, z_j, \phi_k)])$ as the song-level representation.
- Following [2], we employ a set of music audio features as the input features for HDPGMM models.
  - *52 Dimensions*: MFCC (13), ΔMFCC (13), ΔΔMFCC (13), Onset Strength (1), Chroma (2)

Experimental Design

- several models compared
  - **G1**: single multivariate Gaussian parameters (mean-sd) per song
  - **VQCodebook**: approximation of HDPGMM, fitting K-Means globally and employing the post-hoc component frequency per song as the representation.
  - **KIM**: VGG-ish convolutional neural network taking stereo mel-spectrogram as input feature, which is trained with a simple self-supervision objective.
  - **CLMR**: recent DL-based music representations employing advanced self-supervision objective (contrastive learning). It takes time-domain audio samples as input.
- three commonly used MIR downstream tasks are considered:

Dataset	Purpose	no. Samples	no. Classes/no. Users	Acc. Measure
MSD	Repr. Learning	213,354	N/A	N/A
Echonest	Recommendation	40,980	571,355	nDCG
GTZAN	Genre Clf.	1,000	10	F1
MTAT	Autotagging	25,863	50	AUROC

Table: Dataset for training representation (MSD) and downstream tasks evaluation (rest)

Main Results

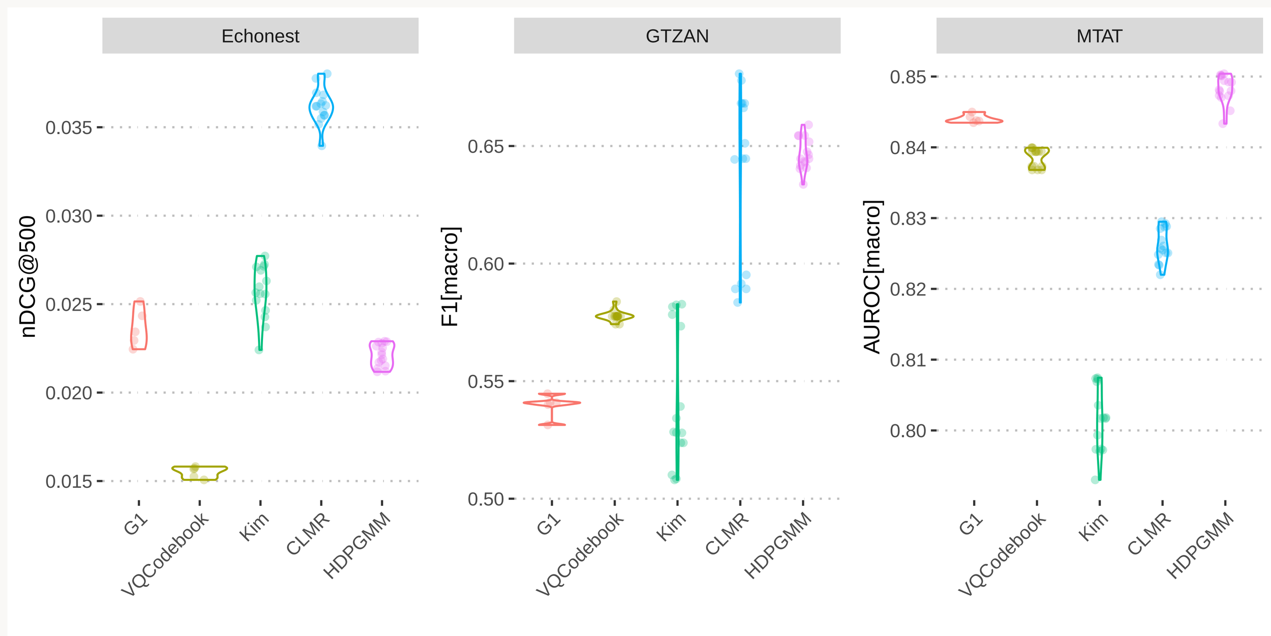

Figure: Main downstream task evaluation results.

- HDPGMM shows the overall comparable "performance" against DL-based representations within our experimental setup.
- HDPGMM representations are competitive to DLs on GTZAN and MTAT, while DL models outperform HDPGMM on Echonest.
- Overall, HDPGMM outperforms simpler non-DL baselines, except on Echonest.

Hyper Parameter Tuning for HDPGMM

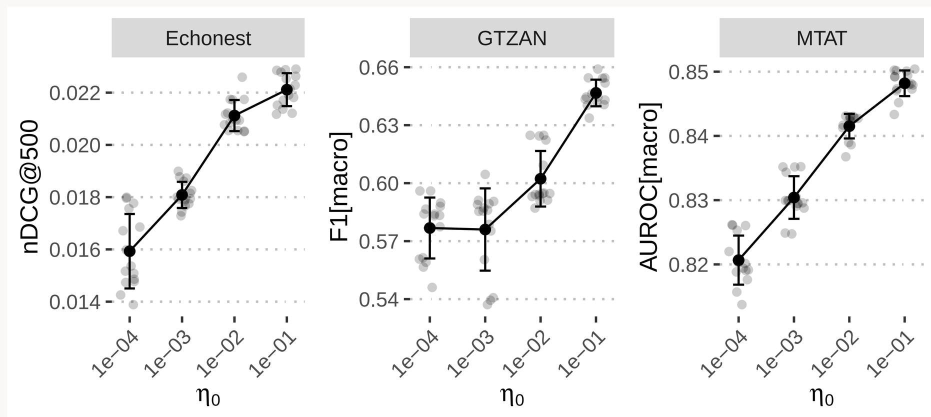

Figure: Effect of regularization factor.

- The additional regularization shows an apparent positive effect up to the range we tested.
- It suggests that employing full-length songs would possibly improve the representation further.

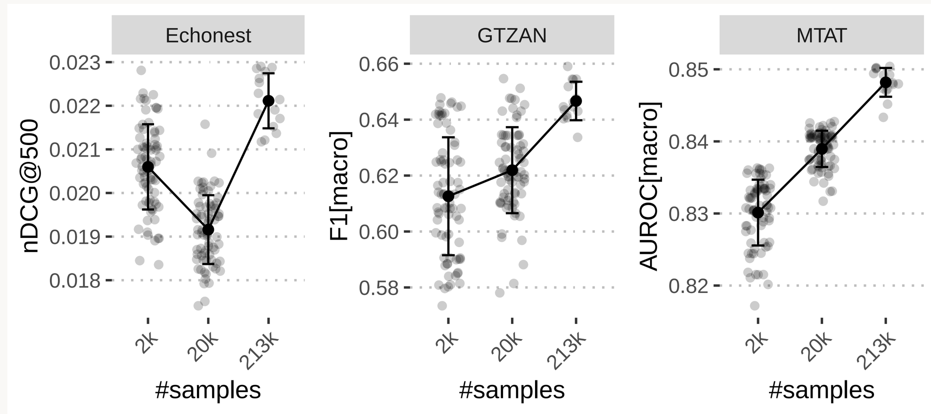

Figure: Effect of the number of training samples.

- The number of training samples also generally indicates a positive effect on the quality of the representation.
- However, it is logarithmic than linear, which suggests:
  - HDPGMM model already generalizes well on the smaller dataset, or
  - It requires exponentially more data to become more competent.

Interpretability

- Knowing what each part of the probabilistic model is supposed to mean and estimating the meaning of components give us a good sense of interpretable representation.
- By intermediating the song-tag assignment matrix from MSD, the semantics of components can be estimated.

Comp1	Comp2	Comp3	Comp4	Comp5
Hip-Hop	country	female vocalists	pop	electronic
pop	rock	singer-songwriter	female vocalists	dance
rnb	pop	pop	female vocalist	electronica
soul	oldies	acoustic	rock	funk
male vocalists	indie	Mellow	Love	electro

Table: Example of tag-based estimation of the per-component semantics.

Conclusion & Future Works

- BN models can learn music representation as effectively as DL while being more interpretable.
- There are several ways to extend BN models
  - semi-supervised learning
  - "deeper" latent structure (nested HDP)
  - sequence-aware models (infinite HMM)

Bibliography

[1] Jaehun Kim. pytorch-hdpgmm, 2022. URL `https://github.com/eldrin/pytorch-hdpgmm`.

[2] Jia-Ching Wang, Yuan-Shan Lee, Yu-Hao Chin, Ying-Ren Chen, and Wen-Chi Hsieh. Hierarchical dirichlet process mixture model for music emotion recognition. *IEEE Trans. Affect. Comput.*, 6(3):261–271, 2015. doi: 10.1109/TAFFC.2015.2415212.

Jaehun Kim and Cynthia C. S. Liem
`j.h.kim@tudelft.nl`
Delft University of Technology

TUDelft