Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation

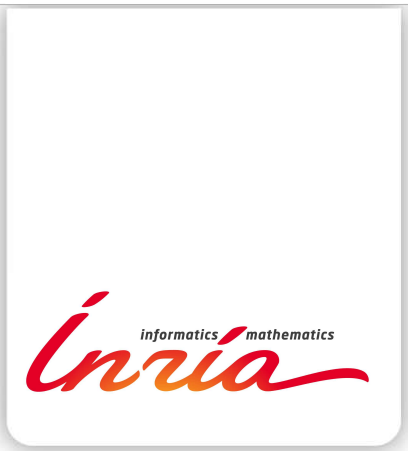
Stanislaw Raczynski, Emmanuel Vincent

Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation

Stanisław A. Raczyński, Emmanuel Vincent

RESEARCH REPORT
N° 434
March 2013
Project-Team PANAMA

ISSN 0249-6399 ISRN INRIA/RR--434--FR+ENG



Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation

Stanisław A. Raczyński, Emmanuel Vincent

Project-Team PANAMA

Research Report n° 434 — March 2013 — 26 pages

Abstract: In this work we present a new Bayesian topic model: latent hierarchical Pitman-Yor process allocation (LHPYPA), which uses hierarchical Pitman-Yor priors for both word and topic posteriors, and generalizes a few of the existing topic models, including the Latent Dirichlet Allocation (LDA), Bigram Topic Model and the Hierarchical Pitman-Yor Topic Model. Using such priors allows for integration of n-grams with a topic model, while smoothing them with the state-of-the-art method.
Our model is evaluated by measuring its cross-entropy on a dataset of musical genre and harmony annotations (3GDB) and by measuring its ability to predict musical genre from chord sequences. Previous work in harmony-based musical genre recognition has used genre-dependent chord n-gram model, but that does not allow for songs to be mixtures of genres. We propose a more flexible approach that models songs as mixtures of latent topics and the genres as mixtures of topics or topic posteriors. In terms of genre prediction accuracy for 9 genres, the proposed approach performs about 35% better than the genre-dependent n-grams, achieving 60.4% of accuracy.

Key-words: topic models, hierarchical Pitman-Yor process, Chinese restaurant process, musical genre recognition, music information retrieval, chord model, genre model

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Allocation de processus hirarchique de Pitman-Yor latente et application la construction des modles de langue musicale bas sur le genre.

Résumé : Dans cet article, nous présentons un nouveau modèle bayésien hiérarchique de topics (thèmes) : LIHPYPA (Latent Hierarchical Pitman-Yor Process Allocation), qui utilise des a-priori hiérarchiques de Pitman-Yor pour la loi de probabilité a-posteriori des mots et des topics, et généralise quelques modèles existants comme le LDA (Latent Dirichlet Allocation) , le modèle bigramme (Bigram Topic Model) et le modèle hiérarchique de Pitman-Yor (Hierarchical Pitman-Yor Topic Model). L'utilisation de tels a-priori nous permet d'intégrer les N-grammes à un modèle de topics tout en les lissant.

Notre modèle est d'abord évalué en mesurant son entropie croisée sur des données d'annotations de genre et d'harmonie (3DGB) puis en mesurant sa capacité à prédire le genre musical à partir d'une séquence d'accords. Les précédents travaux d'estimation du genre musical à partir de l'harmonie utilisent des N-grammes d'accords dépendants du genre, ce qui ne permet pas aux chansons d'être une mélange de plusieurs genres. C'est pourquoi, nous proposons une méthode plus flexible qui modélise les chansons comme des mélanges de topics cachés, et les genres par un mélange de topics ou de lois de probabilité a posteriori. Pour la prédiction sur 9 genres musicaux, l'approche proposée atteint une précision de 60,4% soit 35% de plus que le N-grammes dépendants du genre.

Mots-clés : modéle de sujet, process Pitman-Yor hiérarchique, process de restaurant Chinois, reconaissance de genre musical, extraction d'information musicale, modéle d'accord, modéle de genre musical

1 Introduction

Probabilistic music models are means of incorporating prior musicological knowledge into Music Information Retrieval (MIR) and jointly modelling multiple aspects of musical signals in order to increase the estimation accuracy of all of them [43]. They have already been applied to such MIR problems as composer identification [18, 24], chord transcription from audio signals [25, 42, 15] and from symbolic data [47, 29], automatic harmonization [29, 1, 36], automatic composition [9], music segmentation [32, 19] and polyphonic pitch estimation [47, 30]. Music models are the MIR analogues of the language models that have been successfully used in the fields of natural language processing (NLP) and continuous speech recognition. The world of MIR is slowly adopting language models for its purposes: for example, the n-gram model is nowadays commonly used in MIR [29, 1, 36, 24, 9, 19]. That model, however, does not take into account the fact that the characteristics of music differ between genres and a more flexible, genre-dependent model is therefore desired. Genre-dependent n-gram models have been proposed by Pérez-Sancho *et al.*, who used a collection of n-gram models, one for each genre, for the purpose of musical genre recognition in [27]. The same genre-dependent n-gram model was also used by Lee in [15] for automatic chord transcription from audio. However, this approach does not take into account any overlap between genres and the parameters of the n-grams are not shared between models and as such they are prone to overfitting. In this work we propose to use *topic models*, known from the field of NLP, which allow documents to be mixtures of latent topics and thus effectively dealing with the aforementioned overfitting.

In a typical topic model each observed word (in our case: chord) is assumed to be generated from topic-dependent distribution, and each document (in our case: song) defines a distribution over those topics. Here, we will further assume that the genres are mixtures of topics or their posteriors. The most basic probabilistic topic model is the Latent Dirichlet Allocation (LDA) proposed by Blei *et al.* in [5], which was later generalised as the hierarchical Pitman-Yor topic model (HPYTM) [33] and the Bigram Topic Model (BTM) by Wallach in [44]. In this work we propose a model that generalises all of the above models, that we refer to as the latent hierarchical Pitman-Yor process allocation (LHPYPA). LHPYPA uses hierarchical Pitman-Yor process priors and is therefore capable of using n-grams and advanced smoothing of both the chord and the topic posteriors.

Topic models have already been used in MIR: Hu *et al.* used a simple LDA-based unigram topic model for unsupervised estimation of key profiles from symbolic music data in [13]. Spiliopoulou and Storkey derived a more complex n-gram-based topic model for probabilistic modelling of melodic sequences in [37], however it does not include smoothing.

We evaluate our model by means of cross-entropy calculated on symbolic data, which is a common method for language model evaluation, and by applying it to musical genre recognition (MGR). MGR is one of the fundamental tasks in Music Information Retrieval (MIR) aiming at automatically determining the

genre of a piece of music at hand, represented by an audio recording [31], but may also be a more abstract music representation such as symbolic notes [21], the underlying chord progression [27, 3] or the accompanying lyrics [20]. It is a good testing ground for topic models, because a song will typically exhibit characteristics of more than one genre.

This paper is organised as follows. Section 2 introduces the LHPYPA model, while Section 3 explains the inference procedure for this model. Symbolic evaluation using cross-entropy is presented in Section 4 and MGR results are discussed in Section 5. Finally, conclusion is given in Section 6.

2 Latent hierarchical Pitman-Yor process allocation

2.1 State-of-the-art n-gram and topic models

The currently most popular topic model is the Latent Dirichlet Allocation [5], which is a Bayesian generalization of the probabilistic latent semantic analysis (PLSA) from [12]. In LDA-like models, each song j (consisting of a sequence of observed chords w_t from a dictionary of W, where $t = 1, \ldots, T$ is the position in the sequence) is modelled as a categorical distribution over K hidden topics z_t. Each topic, in turn, defines a different categorical distribution over the observed chords w_t. The chord and topic distributions are unknown and treated as random variables $\boldsymbol{\theta} = \{\theta_{k,j}\}$ and $\boldsymbol{\phi} = \{\phi_{i,k}\}$:

$$\mathrm{P}(z_t = k|j, \boldsymbol{\theta}) = \theta_{k,j}, \quad (1)$$

$$\mathrm{P}(w_t = i|z_t = k, \boldsymbol{\phi}) = \phi_{i,k}. \quad (2)$$

In the basic LDA, the topic and chord posteriors are given Dirichlet priors (see Appendix):

$$\boldsymbol{\theta}_j \sim \mathrm{Dir}(\alpha\mathbf{n}), \quad (3)$$

$$\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta\mathbf{m}), \quad (4)$$

where α and β are the concentration parameters and $\mathbf{n}$ and $\mathbf{m}$ are normalized base measures (expected values of the distribution). Dirichlet priors are used because they are the conjugate priors to the categorical distribution, which simplifies the inference.

LDA makes the so called *bag-of-chords* assumption, *i.e.*, that the order of the observed chords does not matter. We know, however, that the order, called chord progression, is very important for genre discrimination [27]. This prompts us to search for a better prior. The most obvious probabilistic model that takes into account chord order is the n-gram model, which is already commonly used to model chord progressions with $n = 2$ [42, 24, 25, 29, 1, 36, 9, 30] or $n > 2$ [35, 15, 27, 47].

When using n-grams, it is important to use smoothing in order to avoid overfitting due to low amount of training data [35]. Problem with the Dirichlet

priors used in LDA is their limited smoothing capabilities: a simple symmetric Dirichlet prior is equivalent to additive (Laplace) smoothing, while its more flexible asymmetric version to Jelinek-Mercer smoothing [48]. Both methods are not powerful enough to deal with data sparsity of small data sets. The best known n-gram smoothing technique to date is the modified Kneser-Ney, which interpolates between many models of different order and introduces discounting [14, 8]. As it turns out, multi-order interpolation can be achieved by using a hierarchy of Dirichlet distributions, as shown by Teh *et al.* in [41]. Additionally, discounting can be obtained with hierarchical Pitman-Yor processes (HPYP) and in fact Teh has proven that the modified Kneser-Ney smoothing is only an approximation to HPYPs [40].

The idea of using n-grams in topic models was explored before by Wallach in [44]: she replaced the Dirichlet word prior with a hierarchical Dirichlet language model of bigrams, developed earlier by MacKay and Peto in [17], and experimented with what she called the Bigram Topic Model. Later, Sato and Nakagawa proposed another extension to LDA with a order-2 HPYP word prior in [33], in which

$$\begin{aligned} \boldsymbol{\theta}_j &\sim \mathrm{Dir}(\alpha \mathbf{n}), && (5)\\ \boldsymbol{\phi}_k &\sim \mathrm{PY}(d_k, \gamma_k, \boldsymbol{\phi}_0), && (6)\\ \boldsymbol{\phi}_0 &\sim \mathrm{PY}(d_0, \gamma_0, \mathrm{U}), && (7) \end{aligned}$$

where PY() is the Pitman-Yor process distribution, $\mathrm{U}(w_t) = \frac{1}{W}$ is a uniform distribution over words, d_k are the *discount* parameters and γ_k are the *strength* parameters. This formulation, however, does not take into account the chord order and the topic model is given the standard Dirichlet distribution.

2.2 Proposed model

The model proposed in this work extends the work of [44, 33, 39] by placing a HPYP on both the word and the topic models:

$$\begin{aligned} \boldsymbol{\theta}_j &\sim \mathrm{PY}(d_1, \gamma_1, \boldsymbol{\theta}_0), && (8)\\ \boldsymbol{\theta}_0 &\sim \mathrm{PY}(d_0, \gamma_0, \mathrm{V}), && (9)\\ \boldsymbol{\phi}_{k,\mathbf{u}_t(n)} &\sim \mathrm{PY}(d_{k,n}, \gamma_{k,n}, \boldsymbol{\phi}_{k,\mathbf{u}_t(n-1)}), && (10)\\ \boldsymbol{\phi}_{k,\mathbf{u}_t(0)} &\sim \mathrm{PY}(d_{k,0}, \gamma_{k,0}, \mathrm{U}), && (11) \end{aligned}$$

where V is a uniform topic distribution and $\mathbf{u}_t(n) = (w_{t-1}, w_{t-2}, \ldots, w_{t-n+1})$ is the context of the chord w_t of size $n-1$ ($\mathbf{u}_t(1) = \emptyset$), *i.e.*, $n-1$ previously played chords. For $n = 1$, $d_n = 0$, $d_{k,n} = 0$, $\gamma_0 = 0$ and $\gamma_{k,0} = 0$, this model simplifies to LDA with asymmetric priors. For $n = 2$, $d_n = 0$, $d_{k,n} = 0$, $\gamma_0 = 0$ and $\gamma_{k,0} = 0$ it corresponds to the bigram topic model from [44].

It must be noted that a similar, although fully non-parametric, n-gram topic model has also been proposed by Chang and Chien in [7], but there they condition the topics on the context using hierarchical Dirichlet processes and the word n-grams on the topics using hierarchical Pitman-Yor processes, so there

are no song-specific distributions of topics, which makes this model more difficult to use in genre recognition. Also, Blei *et al.* proposed a hierarchical LDA (hLDA) model that allows for an infinite hierarchy of topics using the nested Chinese restaurant process [4].

3 Inference

The key quantity one wants to infer from a topic model is the predictive distribution over a previously unseen test datum $w^{(\text{test})}$ given the training corpus $\mathbf{w}$. To obtain it, we need to integrate out all the latent variables: the latent training topics $\mathbf{z}$, the test topics $\mathbf{z}^{(\text{test})}$, the parameters: $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and the hyperparameters: d_m, γ_m, $d_{k,n}$ and $\gamma_{k,n}$, where $m = \{0, 1\}$. This integration can be performed using an uncollapsed Gibbs sampler [40, 33] akin to the collapsed Gibbs sampling approach used with LDA [11, 44, 45]. A Gibbs sampler samples from the joint distribution of all the variables by sequentially sampling each from the distributions conditional on the values of all the others. The individual topics are drawn from the conditional distribution

$$\mathrm{P}(z_t|\mathbf{w}, \mathbf{z}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{\neg t}) \propto \mathrm{P}(z_t, w_t|\mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{\neg t}), \tag{12}$$

where $\neg t = \{1, 2, \ldots, t-1, t+1, \ldots, T\}$ is a set of time indices excluding the current time index t, $\boldsymbol{\theta}_{\neg t}$ and $\boldsymbol{\phi}_{\neg t}$ are the topic and chord posteriors sampled for all data excluding that at t. The above distribution can be rewritten as a product of these two posteriors:

$$\begin{aligned} &\mathrm{P}(z_t, w_t|\mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{k,\neg t}) = \\ &= \mathrm{P}(w_t|z_t, \mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\phi}_{k,\neg t})\mathrm{P}(z_t|\mathbf{z}_{\neg t}, \boldsymbol{\theta}_{\neg t}) \end{aligned} \tag{13}$$

These can be sampled as seating arrangement in a corresponding Chinese restaurant with the t-th customer removed [40].

3.1 Chinese restaurant analogy

Chord unigram models can be drawn from the Pitman-Yor distribution using a procedure called Chinese restaurant process (CRP) [28]. A Chinese *restaurant* represents the chord posterior $\boldsymbol{\phi}_{k,\mathbf{u}_t(0)}$ for the topic k and consists of the seating arrangement of customers at its *tables* of unbounded number. Each arriving *customer* w_t is seated to a table and served the *dish* i assigned to this table ($w_t = i$) that has been drawn independently from U. The CRP can also be used for sampling in hierarchical Pitman-Yor processes, and therefore sampling for n-gram models, by introducing a hierarchy of restaurants [10, 40, 39]. In this hierarchy, the parent restaurant $\boldsymbol{\phi}_{k,\mathbf{u}_t(n)}$ for the context $\mathbf{u}_t(n)$ and topic k is seating a customer, but also sending him/her to an appropriate child restaurant $\boldsymbol{\phi}_{k,\mathbf{u}_t(n-1)}$. Similarly, topic posteriors can be represented by seating arrangements in restaurants that are serving topics k to customers z_t.

In each restaurant the arriving customer is seated to either an already occupied table with probability proportional to $N_p - d_0$, or to a new table, with probability proportional to $Md_0 + \gamma_0$, where N_p is the number of customers already eating at table p and M is the current number of occupied tables. We can sample the chord and topic posteriors by sequentially *adding* and *removing* customers to the hierarchy of restaurants, *i.e.*, by re-sampling the hierarchical seating arrangements. The chord posterior can be calculated recursively from the current seating arrangement in a hierarchy of chord-serving Chinese restaurants as

$$\begin{aligned} \phi_{i,k,\mathrm{u}_t(n)} &= \frac{N_{i,k,\mathrm{u}_t(n)} - d_{k,n} L_{i,k,\mathrm{u}_t(n)}}{\gamma_{k,n} + N_{i,k}} + \\ &\quad + \frac{\gamma_{k,n} + d_{k,n} M_{k,\mathrm{u}_t(n)}}{\gamma_{k,n} + N_{i,k}} \phi_{i,k,\mathrm{u}_t(n-1)}, \end{aligned} \tag{14}$$

where $L_{i,k,\mathrm{u}_t(n)}$ is the number of tables at which dish i is served in the restaurant for topic k and context $\mathrm{u}_t(n)$.

The hierarchical CRP sampling procedure for the proposed model is a straightforward adaptation of that proposed in [39]. It is outlined in Algorithm 1, while the functions for adding/removing customers and for calculating the posterior are detailed in Algorithm 2.

3.2 Hyper-parameters

Teh described an efficient sampling scheme using auxiliary variables [39], which we adopt in our implementation, but the same results can be achieved using other samplers, *e.g.*, slice sampling [46]. Teh's sampler places a beta hyper-prior on the discount parameters and a gamma hyper-prior on the strengths. The hyper-parameters are tied between all chord restaurant of each topic k.

$$d_{k,n} \sim \mathrm{Beta}(\widetilde{a}_{k,n}, \widetilde{b}_{k,n}), \tag{15}$$

$$\gamma_{k,n} \sim \mathrm{Gamma}(\widetilde{\aleph}_{k,n}, \widetilde{\beth}_{k,n}), \tag{16}$$

where

$$\widetilde{a}_{k,n} = a_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{r=1}^{M_{k,q}-1} (1 - y_{q,r}), \tag{17}$$

$$\widetilde{b}_{k,n} = b_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{i} \sum_{p=1}^{M_{k,q,i}} \sum_{j}^{N_{k,q,p,i}-1} (1 - z_{q,i,p,j}), \tag{18}$$

and

$$\widetilde{\aleph}_{k,n} = \aleph_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{r=1}^{M_{k,q}-1} y_{q,r}, \tag{19}$$

$$\widetilde{\beth}_{k,n} = \beth_{k,n} - \sum_{q=1}^{Q_{k,n}} \log x_q, \tag{20}$$

```
Initialise (z);
Initialise (z^(test));
for s ← 1 to S do
    for t ← 1 to T do
        if s > 1 then
            chordTrainingRestaurants [z_t, u_t(n)].RemoveCustomer (w_t);
            topicTrainingRestaurants [j_t].RemoveCustomer (z_t);
            for k ←1 to K do
                φ_{w_t,k,¬t} = chordTrainingRestaurants [k,
                u_t(n)].DishProbability (w_t);
                θ_{j_t,k,¬t} = topicTrainingRestaurants [j_t].DishProbability
                (z_t);
            end
            z_t = SampleCategorical (φ_{w_t,¬t} θ_{j_t,¬t});
        end
        chordTrainingRestaurants [z_t, u_t(n)].AddCustomer (w_{t,j});
        topicTrainingRestaurants [j_t].AddCustomer (z_t);
    end
    UpdateHyperparamaters (topicTrainingRestaurants [·]);
    for k ← 1 to K do
        UpdateHyperparamaters (chordTrainingRestaurants [k,·]);
    end
    for τ ← 1 to T^(test) do
        if s > 1 then
            topicTestRestaurants [j_τ].RemoveCustomer (z_τ);
            for k ←1 to K do
                φ_{w_τ,k,¬τ} = chordTrainingRestaurants [k,
                u_τ(n)].DishProbability (w_τ);
                θ_{j_τ,k,¬τ} = topicTestRestaurants [j_τ].DishProbability (z_τ);
            end
            z_τ = SampleCategorical (φ_{w_τ,¬τ} θ_{j_τ,¬τ});
        end
        topicTestRestaurants [j_τ].AddCustomer (z_τ);
    end
    UpdateHyperparamaters (topicTestRestaurants [·]);
end
```

Algorithm 1: The Gibbs sampler for LHPYPA. t is used to index the training data $\mathbf{w}$, $\mathbf{z}$ and $\mathbf{j}$ (document numbers) and τ to index the test data $\mathbf{w}^{(\text{test})}$, $\mathbf{z}^{(\text{test})}$ and $\mathbf{j}^{(\text{test})}$. C-style notation is used: square brackets denote accessing an element of an array of restaurants and the full stop stands for invoking a method of an object (a restaurant). The sans-serif font denotes variables, while the `mono-spaced font` functions and methods. The dot notation $[\,\cdot\,]$ stands for using the entire array.

`Restaurant:: AddCustomer(`i`)`
  if $n > 1$ **then**
    $\mathrm{P}(i|\mathbf{u}_t(n-1)) =$ restaurants $[\mathbf{u}_t(n-1)]$.`DishProbability` (i);
    restaurants $[\mathbf{u}_t(n-1)]$.`AddCustomer` (i);
  else
    $\mathrm{P}(i|\mathbf{u}_t(n-1)) = \mathrm{U}(i)$;
  end
  sit customer at table p with probability proportional to $\max(0, N_{i,p} - d_n)$;
  sit customer at new table with probability proportional to $(\gamma_n + d_n L)\mathrm{P}(i|\mathbf{u}_t(n-1))$;

`Restaurant:: RemoveCustomer(`i`)`
  if $n > 1$ **then**
    restaurants $[\mathbf{u}_t(n-1)]$.`RemoveCustomer` (i);
  end
  remove customer from table p with probability proportional to $N_{i,p}$;

`Restaurant:: DishProbability(`i`)`
  if $n > 1$ **then**
    return $\frac{N_i - d_0 * L_i}{\gamma_0 + N} + \frac{\gamma_0 + d_0 M}{\gamma_0 + N}$ restaurants $[\mathbf{u}_t(n-1)]$.`DishProbability` (i);
  else
    return $\frac{N_i - d_0 * L_i}{\gamma_0 + N} + \frac{\gamma_0 + d_0 M}{\gamma_0 + N}\mathrm{U}(i)$;
  end

Algorithm 2: Methods of the Chinese restaurant class (`Restaurant::`).

where $Q_{k,n}$ is the number of restaurants with the context length $n-1$ in topic k with at least 2 occupied tables, $M_{k,q}$ is the number of occupied tables in that restaurant, $M_{k,q,i}$ is the number of tables in that restaurant serving dish i with at least two customers, $N_{k,q,i}$ is the number of customers sitting in that restaurant at table p eating dish i and where the auxiliary variables are drawn as follows:

$$x_q \sim \text{Beta}(\gamma_{n_q}+1, N_q-1), \tag{21}$$

$$y_{q,r} \sim \text{Bernoulli}(\frac{\gamma_{n_q}}{\gamma_{n_q}+d_{n_q}r}), \tag{22}$$

$$z_{q,i,p,j} \sim \text{Bernoulli}(\frac{j-1}{j-d_{n_q}}), \tag{23}$$

where N is the total number of customers in restaurant q and n_q is the order of that restaurant. $a_{k,n}$, $b_{k,n}$, $\aleph_{k,n}$ and $\beth_{k,n}$ are hyper-hyper-parameters, which we all set to 1, following our own experience and suggestions from [33].

The same sampler is used for the topic restaurants' hyper-parameters.

4 Symbolic evaluation

The proposed model was first evaluated by measuring the cross-entropy (its normalised log-likelihood) on unseen (test) data. We then analysed its performance on MGR. We have used the chord degree data (chords annotated using the Roman numeral notation) from the 3GDB data set in all experiments, because that way we work with smaller dictionaries, diminishing somewhat the effects of data sparsity, and the model is independent of the tonic (key). The chords had been annotated with three levels of detail: full chord labels, triad-level chords (they take into account only the first two intervals of the chord) and dyad-level chords (only the first interval of the chord). The corresponding dictionary sizes were 262, 98 and 15.

4.1 Cross-entropy

Cross-entropy on the unseen (test) data $\mathbf{w}^{(\text{test})}$ is defined as

$$H(\mathbf{w}^{(\text{test})}) = -\frac{1}{T}\log_2 \text{P}(\mathbf{w}^{(\text{test})}|\mathbf{w}). \tag{24}$$

There are many possible methods for estimating the distribution over test data [45] and here we chose to use the unbiased estimator from [23], which is the harmonic mean of distribution samples collected using the Gibbs sampler from the previous section:

$$\text{P}(\mathbf{w}^{(\text{test})}|\mathbf{w}) \approx \left(\frac{1}{S}\sum_{s=1}^{S}\text{P}(\mathbf{w}^{(\text{test})}|\mathbf{w},\mathbf{z}_s,\mathbf{z}_s^{(\text{test})})^{-1}\right)^{-1}, \tag{25}$$

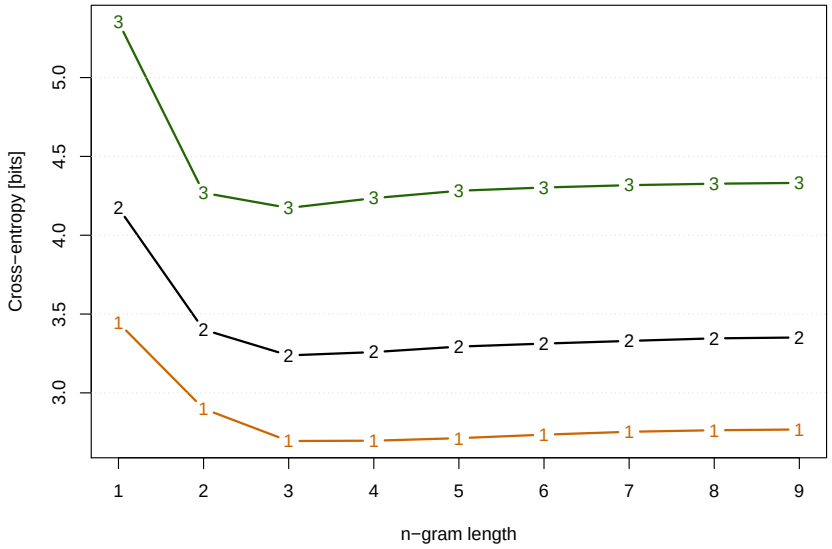


Figure 1: Cross-entropies for smoothed chord n-grams, for dyads (1), triads (2) and full chord labels (3).

where $\mathbf{z}_s$ and $\mathbf{z}_s^{(\text{test})}$ are training and test topic samples from the Gibbs sampler and S is the total number of collected samples.

We found that the model is sensitive to initialisation, because the topic labels $\mathbf{z}$ are highly correlated with each other and the Gibbs sampler tends to gets stuck in modes of the joint distribution. We found then that sampling the initial topic labels $\mathbf{z}_{s=0}$ from a uniform distribution results in a slightly different estimate of the likelihood each time. To minimise this variation, we were bootstrapping the algorithm with the topics obtained with the Gibbs-EM algorithm used on the unigram LDA, which is described in [44] (using prior 2).

Nevertheless, the convergence of the algorithm (to a quasi-stable distribution) is fast. In our experiments, we burned the Gibbs sampler in by sampling for only training data variables for $B_1 = 500$ samples before starting to sample all the test variables as well. We would then wait another $B_2 = 500$ iterations before starting to collect $S = 500$ likelihood samples to calculate the cross-entropy.

4.2 Results

Fig. 1 shows a plot of the cross-entropies obtained by setting the topic number to $K = 1$, *i.e.*, for a smoothed n-gram model. For all chord label detail levels we observe a minimum at $n = 3$, after which the models slowly start to overfit to the data.

Individual cross-entropies are plotted in Fig. 2, the top part of which shows

the cross-entropy decreasing quickly with the number of latent topics and saturates at about 15, 50 and 120 topics for dyads, triads and full chord labels, respectively. The bottom part of Fig. 2b depicts cross-entropies for different values of n and the full chord labels. We see that the decrease in cross-entropy for $n > 2$ is minimal, as was also visible in Fig. 1.

5 Musical Genre recognition

Assigning genre labels is a very common way of categorizing music. It is also very unprecise: there is no general agreement on the musical genre taxonomy, the genres often overlap and some are defined based on purely extrasonic features such as geographical region, historical era or cultural background [34]. Furthermore, the intrasonic genre and sub-genre discriminants belong to different acoustic domains, making it difficult to construct a general automatic MGR system. For example, synthpop can be distinguished from regular pop by its extensive use of synthesizers, while its other characteristics vary greatly between bands; on the other hand, different sub-genres of heavy metal differ mostly in their "aggressiveness", tempo, lyrics and instrument playing practices, while the used instruments are mostly the same.

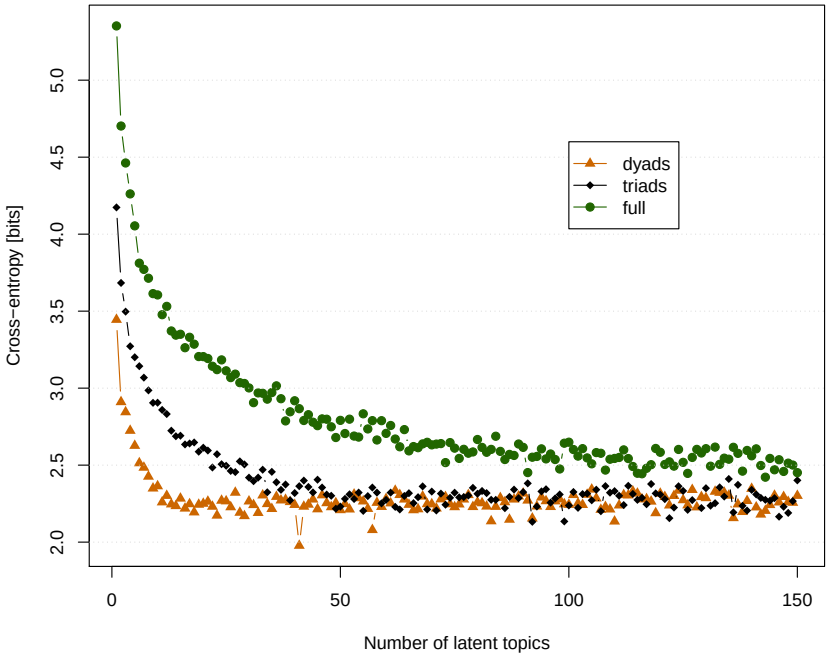
One of the key discriminants of musical genre (at least in the Western tonal music) is the utilised harmony. For instance, the harmonies of the baroque and classical periods are characterised by strict formalisms, while the more modern jazz music employs much more liberal, complex and dissonant chord progressions; at the same time the harmonies of pop music tend to be simplistic and repetitive, a good illustration of which is the commonly used pop-rock chord progression I–V–vi–IV. The majority of MGR methods use features extracted from audio signals [38], mostly because this kind of data is readily available in large amounts, however only few of them use extracted chord sequences [3, 16]. Fortunately, the recent years have produced a steadily increasing stream of hand-annotated musical data sets, including a data set of harmony annotations called 3GDB [27, 2], making room for harmony-based MGR. This work focuses on harmonic data, but naturally the proposed approach can be used with more musical features.

Given a sample of the topic posteriors $\boldsymbol{\theta}$, the genre y for a previously unseen song j can be obtained as a MAP estimate:

$$\widehat{y} = \arg\max_{y} \mathrm{P}(y|\boldsymbol{\theta}). \tag{26}$$

The genre posterior can be obtained with either a generative approach or using discriminative approach based on naïve Bayesian classifiers.

(a) Unigrams



(b) n-grams of full chord labels

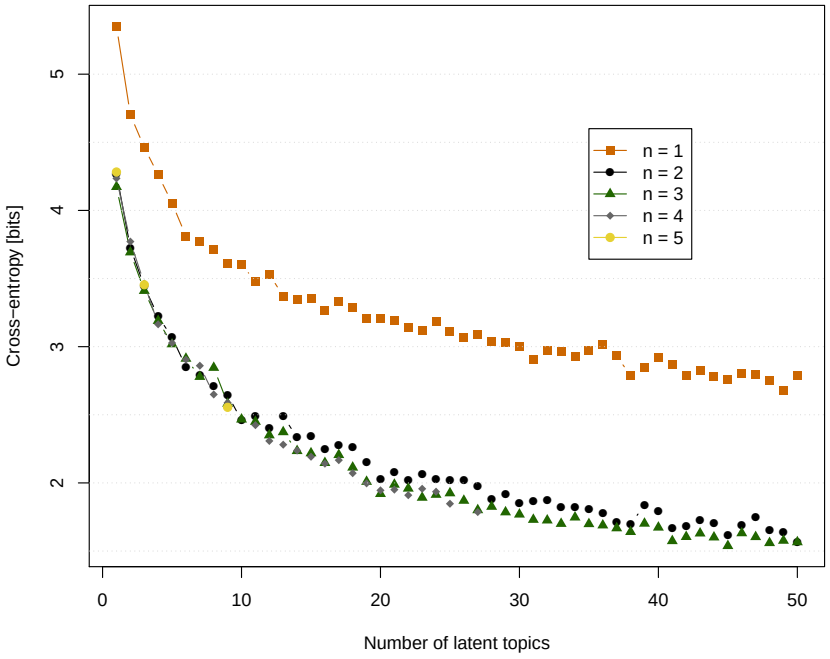


Figure 2: Cross-entropies of the model, calculated on the test data.

5.1 Generative approach

The genre posterior can be found as

$$\widehat{y} = \arg\max_y \sum_{k=1}^{K} \mathrm{P}(y|z=k,\mathbf{A})\mathrm{P}(z=k|\boldsymbol{\theta}), \tag{27}$$

where z is a topic for this document, $\boldsymbol{\theta}$ is a sample of the topic posterior for this document

$$\theta_k = \mathrm{P}(z=k|\boldsymbol{\theta}) \tag{28}$$

and $\mathbf{A} = \{A_{i,k}\}$ is the "transformation matrix"

$$A_{i,k} = \mathrm{P}(y=i|z=k,\mathbf{A}). \tag{29}$$

The transformation matrix is found from the joint distribution of genres and topics of the training data:

$$A_{i,k} = \frac{\mathrm{P}(y, z=k|\mathbf{z}^{(\text{training})})}{\mathrm{P}(z=k|\mathbf{z}^{(\text{training})})}, \tag{30}$$

5.2 Naïve Bayesian classifiers

Another approach is a discriminative approach using naïve Bayesian classifiers:

$$\begin{aligned} \widehat{y} &= \arg\max_y \mathrm{P}(y|\boldsymbol{\theta},\Lambda) = \arg\max_y \mathrm{P}(\boldsymbol{\theta}|y,\Lambda)\mathrm{P}(y) \\ &= \arg\max_y \prod_{s=1}^{S} \mathrm{P}(\boldsymbol{\theta}_s|y,\Lambda)\mathrm{P}(y), \end{aligned} \tag{31}$$

where s indexes Gibbs samples of the topic posteriors and $\mathrm{P}(\boldsymbol{\theta}_j^{(\text{test})}|y)$ is a parametric probability density function with parameters Λ that can easily be trained from the training data.

5.2.1 Gaussian density

The topic posterior density can be modelled with a Gaussian,

$$\mathrm{P}(\boldsymbol{\theta}_s|y,\Lambda) = \mathcal{N}(\boldsymbol{\theta}_s;\Lambda), \tag{32}$$

where $\Lambda = (\boldsymbol{\Sigma}, \overline{\boldsymbol{\theta}})$. The covariance matrices and the mean vectors are estimated on the training data. However because the covariance matrices would be singular since the topic posteriors are normalized (and therefore linearily dependent), we need to only use the first $K-1$ coefficients.

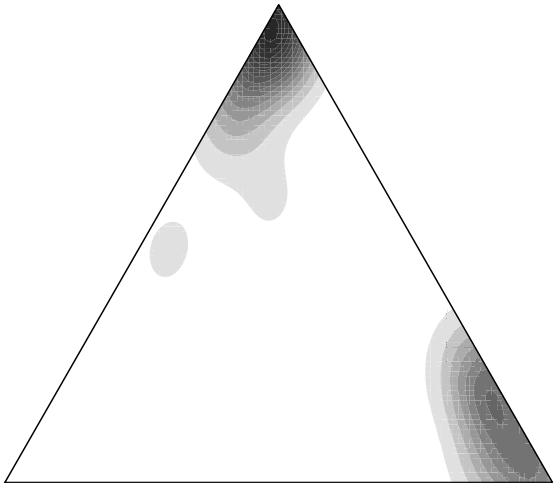


Figure 3: A topic posterior distribution over a K-simplex for one of the genres, where $K = 3$ and $L = 9$, $n = 2$ and full chord labels showing strong bimodality. Darker shades represent larger density.

5.2.2 Dirichlet mixtures

Since $\boldsymbol{\theta}$ is a probability distribution, we would like its distribution to be defined on a K-simplex (a Gaussian is defined on the entire $\mathbb{R}^K$ space). The most popular choice of a prior over categorical distributions (such as the topic posteriors) is the Dirichlet distribution. However, because the topic posterior density appears to be multimodal (see Fig. 3), we will Dirichlet mixtures:

$$\mathrm{P}(\boldsymbol{\theta}_s|y,\Lambda) = \sum_{d=1}^{D} \lambda_d \mathrm{Dir}(\boldsymbol{\theta}_s;\boldsymbol{\rho}_d), \tag{33}$$

where λ are the mixing coefficients and the Dirichlets are parameterised by K-element vectors ρ_d. The mixture coefficients as well as the paremeters of the Dirichlets can be found using the E-M algorithm (detailed in Appendix).

5.3 Results

Experiments were performed for all combinations of the parameters: $L \in \{3, 9\}$, $K \in \{3, 9, 12, 15, 18, 21, 24, 27, 30, 39, 48\}$, $n \in \{1, 2, 3, 4\}$, all three chord label detail levels (dyads, triads, full chord labels) and 8 estimation methods: generative, Gaussian and Dirichlet mixtures of 1, 2, 4, 6, 12 and 24 components. For every set of parameter values, topic posterior samples were collected with $B_1 = 500$, $B_2 = 50$ and $S = 100$ and this was performed 20 times, resulting in 2000 topic posterior samples per song.

L	Proposed	[27]		[26]	
		n-grams	Bayes	n-grams	Bayes
3	89.0%	84.8%	85.3%	$86 \pm 3\%$	$86 \pm 4\%$
9*	60.4%	45.3%	48.4%	$38 \pm 12\%$	$62 \pm 6\%$

Table 1: Accuracies obtained by the proposed method and the reference work by Pérez-Sancho *et al.* in [27] and [26] for the same data (degrees with extensions, full chord labels). * In the case of [27] the number of genres L was 8.

The accuracy of musical genre estimation was calculated as the average over accuracies for each genre:

$$\mathcal{A} = \frac{1}{L} \sum_{l=1}^{L} \frac{NP_l}{NT_l}, \tag{34}$$

where NP_l is the number of correctly identified songs for genre l and NT_l is the total number of songs in that genre.

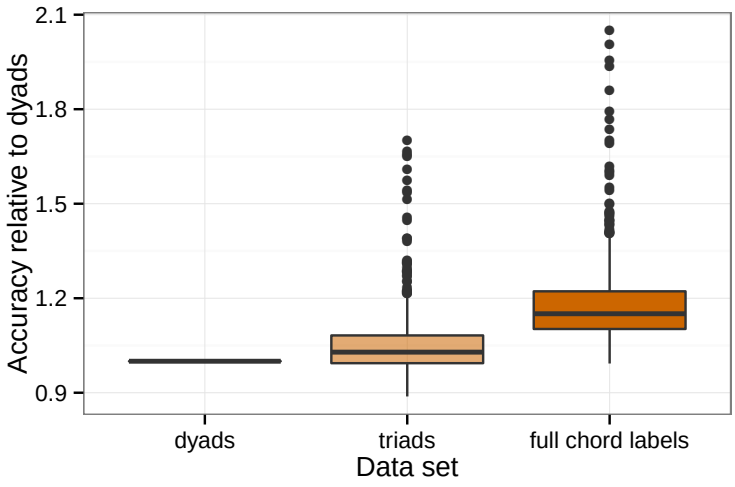
Fig. 4 shows a plot of accuracies relative to the accuracy for dyads, for all values of L, K, n and all methods. We see that higher detail level in chord descriptions translated to an accuracy higher by 20–30% on average, even for longer contexts. This suggests that the effect of the extra information about chords is stronger than that of overfitting the model by using larger vocabulary (which, given the smoothing in our model, should indeed be low). In the following discussion and plots we will therefore present data for full chord labels only.

Accuracies for full chord labels are visualised in Fig. 5. For $L = 3$, there is little difference between analysis methods if sufficiently large number of topics K is used, although mixtures of many Dirichlets are slightly better. Also, the best accuracy was generally obtained for $n = 2$, which confirms the observation from Fig. 1 that longer contexts do not improve the model, at least for such a small data set. This is also consistent with the results presented in [27, 26].

On the other hand, accuracy increases significantly for $L = 9$ the generative approach is significantly (about 15%) lower than for the other methods. Here too mixtures of many Dirichlets performed best. Furthermore, this time the highest accuracy is generally obtained for $n = 3$, so the context is more important in distinguishing between sub-genres than it is between genres.

Although there is quite some variance in the obtained accuracies (*cf.* Fig.6), we were able to achieve genre recognition accuracy of 88% for $L = 3$ and 60% for $L = 9$ for the mixture of 24 Dirichlets. This is better than the results for n-grams reported both in [27] and [26] and comparable to the naïve Bayesian classifier with multivariate Bernoulli distribution from [26] (see Table 1; we do not quote results for hierarchical classifiers from [26], because those used both harmony and melodic information).

(a) $L = 3$



(b) $L = 9$

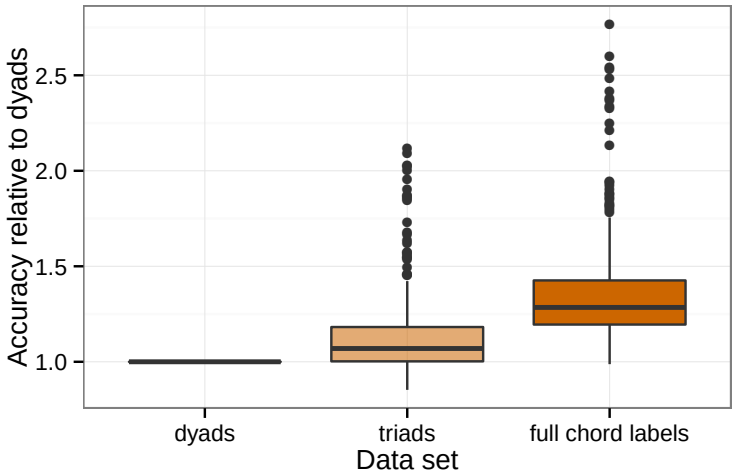


Figure 4: Boxplot of accuracies obtained on the three data sets (1 = dyads, 2 = triads, 3 = full chord labels), relative to the dyad data set accuracy, for all combinations of K, n and estimation method.

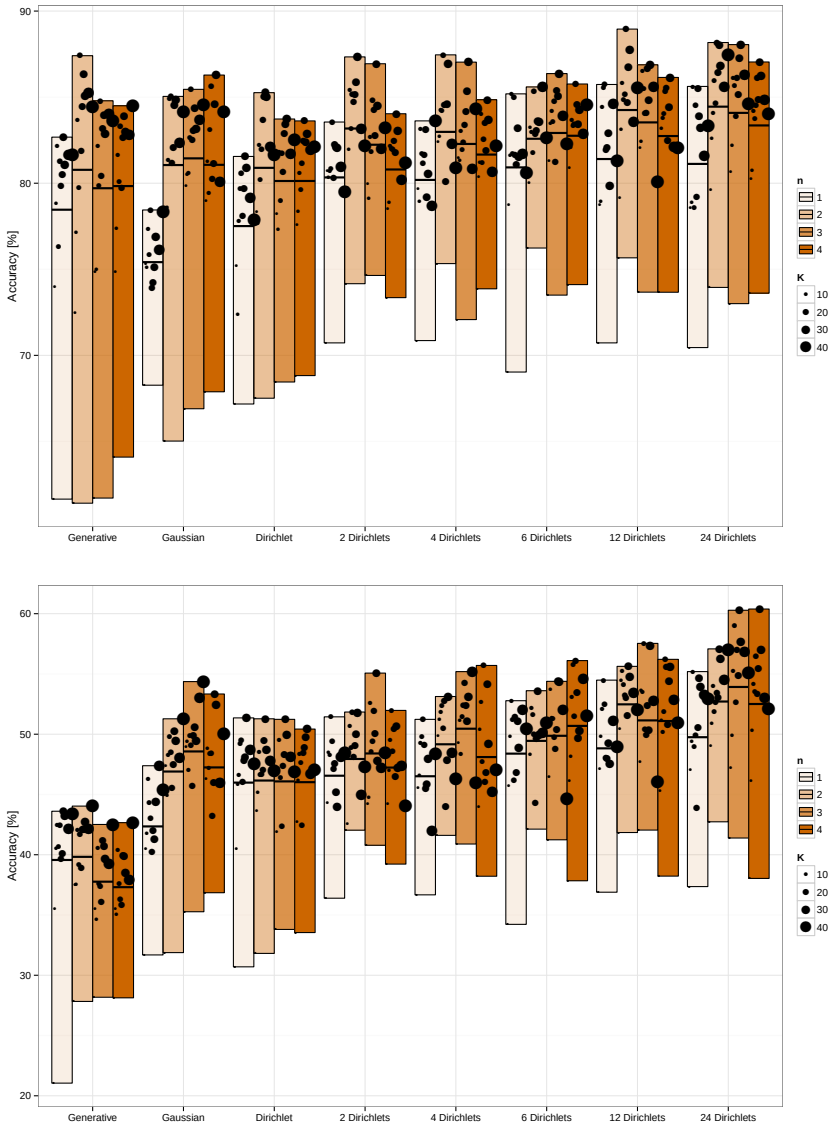


Figure 5: Accuracies obtained for all analysis methods (marked on the horizontal axis) and values of n and K, for $L = 3$ (top) and $L = 9$ (bottom). Bars correspond to minimal, mean and maximal values over all K's for a particular method and value of n. Spots stand for individual accuracy values, whereas their size and relative position within a bar correspond to the value of K.

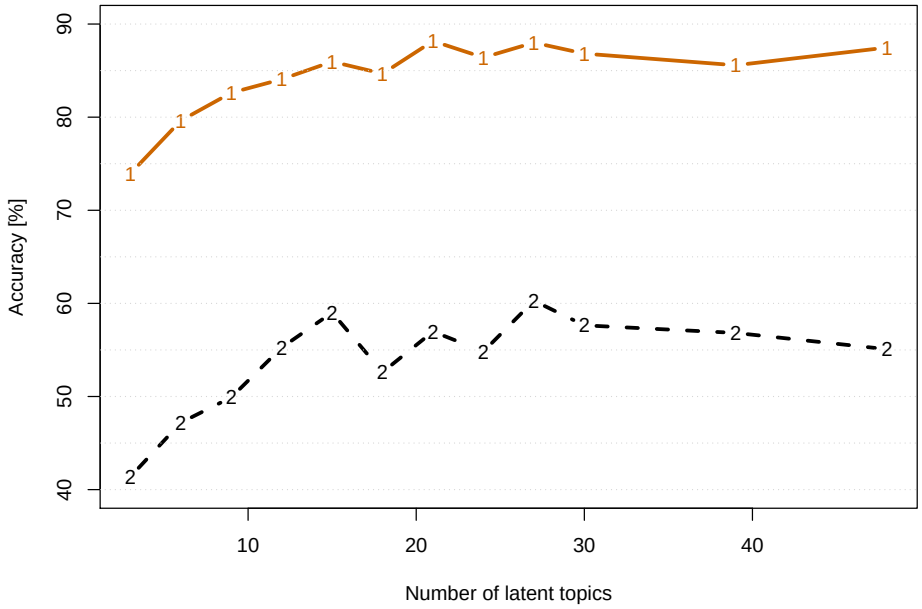


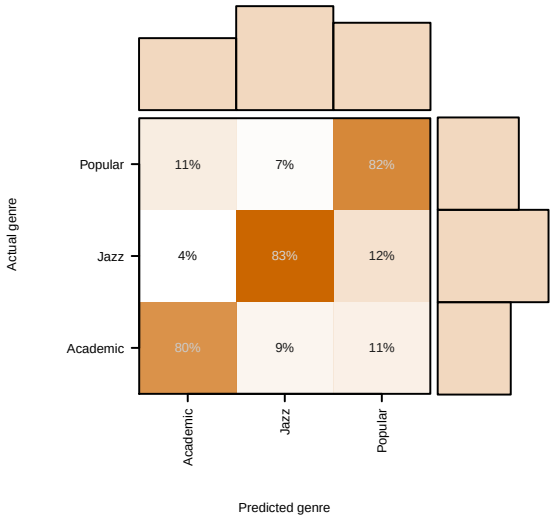
Figure 6: Genre estimation accuracy as a function of the number of topics for $L = 3$ (black solid line) and $L = 9$ (orange dashed line). A mixture of 24 Dirichlets was used with $n = 2$ and $n = 3$, respectively.

Fig. 6 shows accuracies obtained for the best parameter values as a function of K, but the same shape can be observed for all methods: the accuracy increases up to about 20–30 topics and then shows a slow decrease towards larger values of K. Additionally, Fig. 7 depicts the confusion matrices for $K = 27$ and the mixture of 24 Dirichlets. For $L = 3$ there is more confusion between popular and academic genres than between them any jazz, which is to be expected since jazz generally employs more complex harmonies, which makes it more distinct. It is more difficult to detect confusion patterns for $L = 9$, but the most confusion is between closely related sub-genres of the same genre, *e.g.*, between bop and prebop or between baroque and romantic classes.

6 Conclusion

In this paper we have presented a new smoothed topic model using hierarchical Pitman-Yor process priors and applied it to musical genre recognition. It integrates a topic model with word (chord) n-grams and is therefore capable of handling data for which the bag-of-words assumption does not hold, such as chord sequences, where progressions between chords, in addition to the chords themselves, contribute to discrimination between musical genres. Using topic models is more flexible than the previously proposed genre-dependent n-gram

(a) $L = 3$, $n = 2$, $K = 27$, 24 Dirichlets



(b) $L = 9$, $n = 3$, $K = 27$, 24 Dirichlets

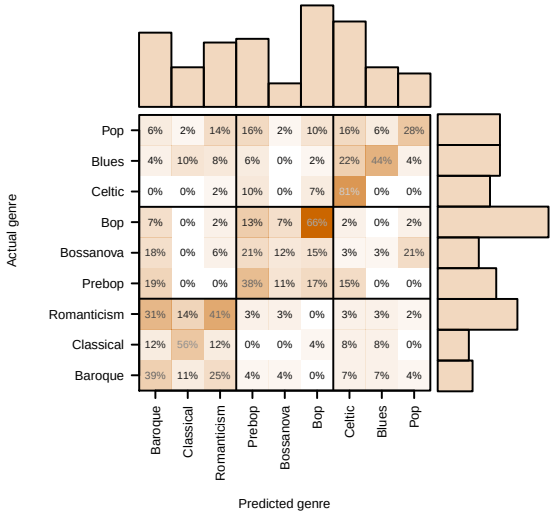


Figure 7: Confusion matrices for best sets of parameters. Shades correspond to absolute counts of labelling an "actual genre" as a "predicted count", while the text labels show counts relative to the total number of songs for a particular actual genre (*i.e.*, rows sum up to one). Marginal barplots show total counts for the actual and predicted genres.

model by allowing both the songs and the genres to be mixtures of latent topics.

The proposed model has the potential to be used in all applications where topic models are used and the order of words matter, such as modelling textual documents [44] or analysis of genomic information [6].

References

[1] PG Music Inc. Band-in-a-Box. `http://www.pgmusic.com/`, August 2012.

[2] 3 Genre Database (3GDB). `http://grfia.dlsi.ua.es/cm/projects/prosemus/database.php`, February 2013.

[3] A. Anglade, R. Ramirez, and S. Dixon. Genre classification using harmony rules induced from automatic chord transcriptions. In *Proc. 10^{th} International Society for Music Information Retrieval (ISMIR)*, pages 669–674, 2009.

[4] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Xin C., Xiaohua H., Xiajiong S., and G. Rosen. Probabilistic topic modeling for genomic data interpretation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 149–152, 2010.

[7] Y.-L. Chang and J.-T. C. Bayesian nonparametric language models. In *Proc. 8^{th} International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 188–192. IEEE, 2012.

[8] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. 34^{th} annual meeting on Association for Computational Linguistics*, pages 310–318. ACL, 1996.

[9] S. Fukayama, K. Nakatsuma, S. Sako, T. Nishimoto, and S. Sagayama. Automatic song composition from the lyrics exploiting prosody of the Japanese language. In *Proc. 7^{th} Sound and Music Computing Conference (SMC)*, pages 299–302, 2010.

[10] D. M. Griffiths, T. L. Blei, M.I. Tenenbaum, and J. B. Jordan. Hierarchical topic models and the nested Chinese restaurant process. In *Proc. Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2004.

[11] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 101 (Suppl 1), pages 5228–5235, 2004.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

[13] D. J. Hu and L. K. Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *Proc. 10th International Society for Music Information Retrieval (ISMIR)*, pages 441–446, 2009.

[14] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE, 1995.

[15] K. Lee. A system for automatic chord transcription from audio using genre-specific hidden markov models. *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pages 134–146, 2008.

[16] T. Lidy, R. Mayer, A. Rauber, P. J. Ponce de León Amador, A. Pertusa Ibáñez, and J. M. Iñesta Quereda. A Cartesian ensemble of feature subspace classifiers for music categorization. In *Proc. 11th International Society for Music Information Retrieval (ISMIR)*, pages 279–284, 2010.

[17] D. J. C. MacKay and L. C. B. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1995.

[18] M. Mauch, S. Dixon, C. Harte, and Q. Mary. Discovering chord idioms through Beatles and Real Book songs. In *Proc. 8th International Society for Music Information Retrieval (ISMIR)*, pages 255–258, 2007.

[19] M. Mauch, K. Noland, and S. Dixon. Using musical structure to enhance automatic chord transcription. In *Proc. 10th International Society for Music Information Retrieval (ISMIR)*, pages 231–236, 2009.

[20] R. Mayer, R. Neumayer, and A. Rauber. Rhyme and style features for musical genre classification by song lyrics. In *Proc. 9th International Society for Music Information Retrieval (ISMIR)*, pages 337–342, 2008.

[21] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proc. 5th International Society for Music Information Retrieval (ISMIR)*, pages 30–35, 2004.

[22] T. Minka. Estimating a dirichlet distribution. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`, 2012.

[23] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

[24] M. Ogihara and T. Li. N-gram chord profiles for composer style representation. In *Proc. 9th International Society for Music Information Retrieval (ISMIR)*, pages 671–676, 2008.

[25] H. Papadopoulos and G. Peeters. Large-scale study of chord estimation algorithms based on chroma representation and HMM. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 53–60. IEEE, 2007.

[26] C. Pérez-Sancho, D. Rizo, and J. M. Iñesta. Stochastic text models for music categorization. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 55–64, 2008.

[27] C. Pérez-Sancho, D. Rizo, and J. M. Iñesta. Genre classification using chords and stochastic language models. *Connection science*, 21(2-3):145–159, 2009.

[28] J. Pitman. Combinatorial stochastic processes. Technical Report 621, University of California Berkeley, Dept. Statistics, 2002.

[29] S. A. Raczyński, S. Fukayama, and E. Vincent. Melody harmonisation with interpolated probabilistic models. Research Report RR-8110, INRIA, October 2012.

[30] S. A. Raczyński and E. Vincent. Dynamic Bayesian networks for symbolic polyphonic pitch modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[31] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagayama. Autoregressive MFCC models for genre classification improved by harmonic-percussion separation. In *Proc. 11th International Society for Music Information Retrieval (ISMIR)*, pages 87–92, 2010.

[32] G. Sargent, F. Bimbot, and E. Vincent. A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs. In *Proc. 12th International Society for Music Information Retrieval (ISMIR)*, pages 483–488, 2011.

[33] I. Sato and H. Nakagawa. Topic models with power-law using pitman-yor process. In *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–681, 2010.

[34] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[35] R. Scholz, E. Vincent, and F. Bimbot. Robust modeling of musical chord sequences using probabilistic N-grams. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56. IEEE, 2009.

[36] I. Simon, D. Morris, and S. Basu. MySong: automatic accompaniment generation for vocal melodies. In *Proc. 26th SIGCHI Conference on Human Factors in Computing Systems*, pages 725–734, 2008.

[37] A. Spiliopoulou and A. Storkey. A Topic Model for Melodic Sequences. *ArXiv e-prints*, June 2012.

[38] B. L. Sturm. A survey of evaluation in music genre recognition. In *Proc. 10^{th} International Workshop Adaptive Multimedia Retrieval (AMR)*, 2012.

[39] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical report, National University of Singapore, School of Computing, 2006.

[40] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. 21^{st} International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. ACL, 2006.

[41] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[42] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5518–5521. IEEE, 2010.

[43] E. Vincent, S.A. Raczyński, N. Ono, and S. Sagayama. A roadmap towards versatile MIR. In *Proc. 11^{th} International Conference on Music Information Retrieval (ISMIR)*, pages 662–664, 2010.

[44] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proc. 23^{rd} International Conference on Machine Learning*, pages 977–984. ACM, 2006.

[45] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proc. 26^{th} International Conference on Machine Learning*, pages 1105–1112. ACM, 2009.

[46] H. M. Wallach, C. Sutton, and A. McCallum. Bayesian modeling of dependency trees using hierarchical Pitman-Yor priors. In *ICML Workshop on Prior Knowledge for Text and Language Processing*, pages 15–20, 2008.

[47] K. Yoshii and M. Goto. Unsupervised music understanding based on nonparametric Bayesian models. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5353–5356. IEEE, 2012.

[48] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342. ACM, 2001.

A Estimating Dirichlet mixtures

The Dirichlet distribution is defined as

$$\mathrm{Dir}(\boldsymbol{\theta};\boldsymbol{\rho}) = \frac{1}{B(\boldsymbol{\rho})}\prod_{k=1}^{K}\theta_k^{\rho_k-1}, \tag{35}$$

where $\boldsymbol{\rho}$ is a vector of K parameters and the normalising constant is defined as

$$B(\boldsymbol{\rho}) = \frac{\prod_{k=1}^{K}\Gamma(\rho_k)}{\Gamma(\sum_{k=1}^{K}\rho_k)}, \tag{36}$$

where $\Gamma()$ is the gamma function.

The parameters of the Dirichlet mixture model can be found by maximising their likelihood on a set of topic posterior samples collected for songs from a particular genre. In order to find these estimates we need to integrate out the latent mixture component selector variables v_j, so we will use the E-M algorithm. The parameter likelihood is given by

$$\mathcal{L}(\Lambda;\boldsymbol{\theta},\mathbf{v}) = \mathrm{P}(\boldsymbol{\theta},\mathbf{v}|\Lambda) = \prod_{j=1}^{J}\sum_{d=1}^{D}\mathbf{I}(v_j=d)\lambda_d\mathrm{Dir}(\boldsymbol{\theta}_j;\boldsymbol{\rho}_d), \tag{37}$$

where $\mathbf{I}(v_j=d)$ is a binary indicator function that is non-zero iff $v_j=d$.

A.1 E-step

The expected value of the log-likelihood function is given by

$$Q(\Lambda|\Lambda_t) \quad = \quad \mathrm{E}_{\mathbf{v}|\boldsymbol{\theta},\Lambda_t}\left[\log\mathcal{L}(\Lambda;\boldsymbol{\theta},\mathbf{v})\right]. \tag{38}$$

Note that here t indexes iterations of the E-M algorithm, not elements (*e.g.*, chords) in a time sequence. We define

$$\begin{aligned} T_{t,j,d} &= \mathrm{P}(v_j|\boldsymbol{\theta},\Lambda_t) \\ &= \frac{\lambda_d\mathrm{Dir}(\boldsymbol{\theta}_d;\boldsymbol{\rho}_{t,d})}{\sum_{d'=1}^{D}\lambda_{d'}\mathrm{Dir}(\boldsymbol{\theta}_{d'};\boldsymbol{\rho}_{t,d'})}. \end{aligned} \tag{39}$$

Then

$$\begin{aligned} Q(\Lambda|\Lambda_t) = \sum_{j=1}^{J}\sum_{d=1}^{D}\big(T_{t,j,d}(\log\lambda_{t,d} - \log B(\boldsymbol{\rho}_t) + \\ + \sum_{k=1}^{K}(\rho_{t,d,k}-1)\log\theta_{j,k})\big). \end{aligned} \tag{40}$$

A.2 M-step

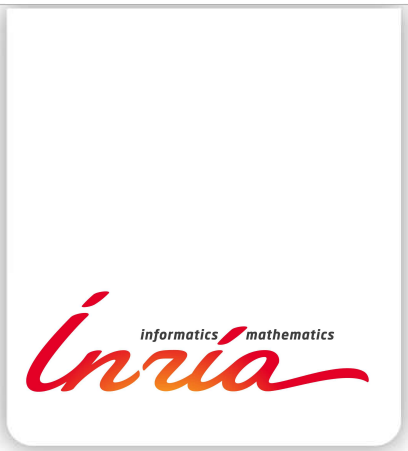
The expectation function Q is a sum of terms that depend on different parameters, so we can maximize it for each of them separately.

$$\begin{aligned} \lambda_{t+1,d} &= \arg\max_{\lambda} \sum_{d=1}^{D} \log \lambda_d \sum_{j=1}^{J} T_{t,j,d} \\ &= \frac{1}{J} \sum_{j=1}^{J} T_{t,j,d} \end{aligned} \tag{41}$$

$$\begin{aligned} \frac{\partial}{\partial \rho_{t,d,k}} Q(\Lambda|\Lambda_t) &= (\Psi(\sum_{k'=1}^{K} \rho_{t,d,k'}) - \Psi(\rho_{t,d,k})) \sum_{j=1}^{J} T_{t,j,d} + \\ &\quad + \sum_{j=1}^{J} T_{t,j,d} \log \theta_{j,k}, \end{aligned} \tag{42}$$

where Ψ is the digamma function. The new $\boldsymbol{\rho}$ can therefore be found as [22]:

$$\rho_{t+1,d,k} = \Psi^{-1}(\Psi(\sum_{k'=1}^{K} \rho_{t,d,k'}) + \frac{\sum_{j=1}^{J} T_{t,j,d} \log \theta_{j,k}}{\sum_{j=1}^{J} T_{t,j,d}}). \tag{43}$$



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399