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

Stanislaw Raczynski, Emmanuel Vincent

To cite this version:

Stanislaw Raczynski, Emmanuel Vincent. Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2014, 22 (3), pp.672-681. 10.1109/TASLP.2014.2300344 . hal-00804567v2

HAL Id: hal-00804567
https://inria.hal.science/hal-00804567v2

Submitted on 8 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Stanisław A. Raczyński, Emmanuel Vincent

***Abstract*—In this work we present a new Bayesian topic model: latent hierarchical Pitman-Yor process allocation (LHPYA), which uses hierarchical Pitman-Yor process priors for both word and topic distributions, and generalizes a few of the existing topic models, including the latent Dirichlet allocation (LDA), the 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 perplexity on a dataset of musical genre and harmony annotations *3 Genre Database* (3GDB) and by measuring its ability to predict musical genre from chord sequences. In terms of perplexity, for a 262-chord dictionary we achieve a value of 2.74, compared to 18.05 for trigrams and 7.73 for a unigram topic model. In terms of genre prediction accuracy with 9 genres, the proposed approach performs about 33% better in relative terms than genre-dependent n-grams, achieving 60.4% of accuracy.

***Index Terms*—topic models, hierarchical Pitman-Yor process, Chinese restaurant process, musical genre recognition, music information retrieval, chord model, genre model**

I. Introduction

Probabilistic music models are means of incorporating prior knowledge about music into the algorithms used in music information retrieval (MIR). This is done by jointly modelling multiple musical variables (like notes, beats, *etc.*) with the goal of increasing the accuracy of estimating all of them [1]. These models have already been applied to such MIR problems as: composer identification [2], [3], audio chord transcription [4]–[6] and symbolic chord transcription [7], [8], automatic harmonization [8]–[10], automatic composition [11], music segmentation [12], [13] and polyphonic pitch estimation [7], [14]. 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 the language models for its own purposes: for example, the n-gram model is nowadays commonly used in MIR [3], [8]–[11], [13]. That simple model, however, does not take into account the fact that the characteristics of music differ between genres. A more flexible, genre-dependent model is therefore sought for. 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 [15]. The same genre-dependent n-gram model was also used by Lee in [6] for automatic chord transcription from audio. However, this approach does not account for any overlap between genres, that is the parameters of the n-grams are not shared between models.

In this work we propose to use *topic models*, known from the field of NLP, which allow documents to be mixtures of latent topics. The topics, contrary to the genres, which are fixed, adapt to the data, and different genres can share topics, so such a model can fit the data much closer. In a typical topic model each observed word (in our case: chord) is assumed to be generated from a topic-dependent distribution, and each document (in our case: song) defines a distribution over those topics. Later, for the purpose of genre recognition, we will further assume that the genres are mixtures of topics. The most basic probabilistic topic model is the latent Dirichlet allocation (LDA) proposed by Blei *et al.* in [16], which was later generalised as the hierarchical Pitman-Yor topic model (HPYTM) [17] and the bigram topic model (BTM) by Wallach in [18]. 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 [19]. Spiliopoulou and Storkey derived a more complex n-gram-based topic model for probabilistic modelling of melodic sequences in [20], however that method did not include advanced smoothing. Without smoothing, models with many degrees of freedom (for n-grams their number grows exponentially with the value of n) will overfit to the training data, *i.e.*, they will have poor predictive power and perform poorly on test data [21].

In this work we propose a model that generalises all of the above models, that we refer to as latent hierarchical Pitman-Yor process allocation (LHPYA). LHPYA uses hierarchical Pitman-Yor process priors and is therefore capable of using n-grams and advanced smoothing with both the word and the topic posteriors. We evaluate our model by means of cross-entropy calculated on chord sequences, which is a common method for language model evaluation [22], and by applying it to chord-based musical genre recognition (MGR), which is one of the fundamental tasks in MIR.

This paper is organised as follows. Section II introduces the LHPYA model, while Section III explains the inference procedure for this model. Symbolic evaluation using cross-entropy is presented in Section IV and MGR results are discussed in Section V. Finally, a conclusion is given in Section VI.

This work was conducted while E. Vincent was with Inria Rennes and was supported by Inria under the Associate Team Program VERSAMUS (http://versamus.inria.fr/). S. Raczyński is with Gdańsk Universit of Technology, Faculty of Electronics, Telecommunications and Informatics, ul. Narutowicza 11/12 Gdańsk (e-mail: staraczy@pg.gda.pl). E. Vincent is with Inria, 54600 Villers-lès-Nancy, France (e-mail: emmanuel.vincent@inria.fr).

II. Latent hierarchical Pitman-Yor process allocation

A. State-of-the-art n-gram and topic models

The currently most popular topic model is the LDA [16], which is a Bayesian generalization of the probabilistic latent semantic analysis (PLSA) from [23]. In topic modelling, each document j, where $j = 1, \dots, J$, is treated as a sequence of observed symbols $w_{j,t}$ from a dictionary W, indexed by $i \in \{1, \dots, |W|\}$, where $t = 1, \dots, T$ is the position in the sequence. Typically the symbols $w_{j,t}$ correspond to words. In MIR they can be chords, pitches, rhythm words, *etc.*, but in this work we focus on sequences of chords for illustration purposes. In LDA-like models, each document is modelled as a categorical distribution over K hidden topics $z_{j,t}$. Each topic, in turn, defines a different categorical distribution over the observed symbols $w_{j,t}$. The topic and symbol distributions are unknown and treated as random variables $\boldsymbol{\theta} = \{\theta_{k,j}\}$ and $\boldsymbol{\phi} = \{\phi_{i,k}\}$:

$$\begin{aligned} \mathrm{P}(z_{j,t} = k|\boldsymbol{\theta}) &= \theta_{k,j}, & (1)\\ \mathrm{P}(w_{j,t} = i|z_{j,t} = k, \boldsymbol{\phi}) &= \phi_{i,k}. & (2) \end{aligned}$$

In LDA, the topic and symbol posteriors are given Dirichlet priors:

$$\begin{aligned} \boldsymbol{\theta}_j &\sim \mathrm{Dir}(\alpha \mathbf{n}), & (3)\\ \boldsymbol{\phi}_k &\sim \mathrm{Dir}(\beta \mathbf{m}), & (4) \end{aligned}$$

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

LDA makes the so called *bag-of-words* assumption, which means that the order of the observed symbols does not matter. We know, however, that the order is very important in many cases. For chords, this order is called the chord progression and is an important genre discriminant [15]. This limitation of LDA can be removed by using better, contextual symbol distributions, such as the popular n-gram model. n-grams are already commonly used to model chord progressions with a context length of $R = 1$ [3]–[5], [8]–[11], [14] and $R > 1$ [6], [7], [15], [21].

The number of parameters that we need to train in the n-gram models grows extremely quickly—with the power of n—which quickly results in overfitting. A common solution to deal with is to use smoothing [24]. The best known n-gram smoothing technique to date is the modified Kneser-Ney smoothing, which interpolates a high-order model with models of lower order and additionally introduces discounting [22], [25]. This kind of model interpolation can also be achieved in a fully Bayesian framework, by using a hierarchy of *Dirichlet processes* (DPs) as the prior for the symbol distributions in [26]. A Dirichlet process $\mathrm{DP}(d, \gamma, \boldsymbol{\phi}_{0,k})$, where γ the *concentration parameter* and $\boldsymbol{\phi}_{0,k}$ is the *base distribution*, is a non-parametric generalisation of the Dirichlet distribution to infinite dimensionality. The additional discounting found in the Kneser-Ney smoothing can be obtained by replacing the DPs with *Pitman-Yor processes* (PYs) and in fact Teh has proven that the interpolated Kneser-Ney smoothing is only an approximation to his model, which is based on hierarchical PYs [27]. His model, called the hierarchical Pitman-Yor process language model (HPYLM), is a hierarchy of PYs, where the symbol distribution $\boldsymbol{\phi}_{k,n}$ for a context of length n is drawn from its parent distribution $\boldsymbol{\phi}_{k,n-1}$ for the same context, but of length $n-1$:

$$\boldsymbol{\phi}_{k,n} \sim \mathrm{PY}(d_n, \gamma_n, \boldsymbol{\phi}_{k,n-1}), \tag{5}$$

where d_n are the *discount* parameters, γ_n the *strength* parameters (which correspond to the concentration parameters of the DP) and $n = 1, \dots, R$. The unigram distribution, that is the distribution $\boldsymbol{\phi}_{0,k}$ for an empty context, is drawn from the another PY:

$$\boldsymbol{\phi}_{0,k} \sim \mathrm{PY}(d_0, \gamma_0, \mathrm{U}), \tag{6}$$

where $\mathrm{U}(w_{j,t}) = \frac{1}{|W|}$ stands for a uniform symbol distribution and $|W|$ is the size of the symbol vocabulary.

The idea of using n-grams in topic models was already explored by Girolami and Kabán in [28], who introduced a LDA-inspired Markov chain mixture model with Dirichlet priors over the mixing weights, but they did not use any smoothing. Later, in [18], Wallach replaced the Dirichlet symbol prior of the LDA with a hierarchical Dirichlet language model (HDLM) of bigrams, developed earlier by MacKay and Peto in [29], and experimented with what she called the BTM. The resulting model was limited to context lengths of $R = 1$. Later, Sato and Nakagawa proposed another extension of LDA with a unigram ($R = 0$) hierarchical PY symbol prior in [17], but their formulation did not take into account the symbol order.

B. Proposed model

Since the proposed model is applied to chord sequences, we will refer to the word posterior as the "chord posterior" from now on. Our model extends the work of [17], [18], [30] by placing hierarchical PYs (HPYs) on both the chord and the topic posteriors:

$$\begin{aligned} \boldsymbol{\theta}_j &\sim \mathrm{PY}(d_1, \gamma_1, \boldsymbol{\theta}_0), & (7)\\ \boldsymbol{\theta}_0 &\sim \mathrm{PY}(d_0, \gamma_0, \mathrm{V}), & (8)\\ \boldsymbol{\phi}_{k,\mathbf{u}(n)} &\sim \mathrm{PY}(d_{k,n}, \gamma_{k,n}, \boldsymbol{\phi}_{k,\mathbf{u}(n-1)}), & (9)\\ \boldsymbol{\phi}_{k,\mathbf{u}(0)} &\sim \mathrm{PY}(d_{k,0}, \gamma_{k,0}, \mathrm{U}), & (10) \end{aligned}$$

where V is a uniform topic distribution and $\mathbf{u}_{j,t}(n) = (w_{j,t-1}, w_{j,t-2}, \dots, w_{j,t-n+1})$ is the context of the chord $w_{j,t}$ of size $n-1$ ($\mathbf{u}_{j,t}(1) = \emptyset$), *i.e.*, the $n-1$ previous chords. Using HPYs for both distributions results in smoothing applied to both chord and topic posteriors. This model is visualised using the plate notation in Fig. 1.

Setting the discount parameters to zero turns PYs into DPs and effectively disables discounting in the smoothing of the model parameters. If we further set $R = 0$ (unigrams) and set γ_0 and $\gamma_{k,0}$ to zeros (then $\boldsymbol{\theta}_0 \to \mathrm{V}$ and $\boldsymbol{\phi}_{k,\mathbf{u}(0)} \to \mathrm{U}$ then our model will use additive (Laplace) smoothing and will therefore be equivalent to LDA with symmetric priors [24]. Similarly, if we use $R = 1$ and no discounting, we will have the BTM model from [18].

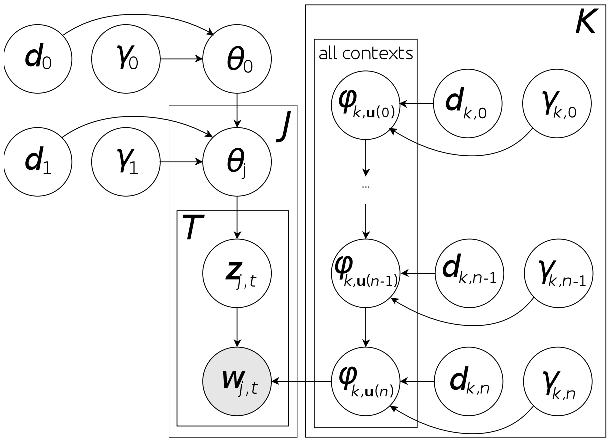


Fig. 1: Plate notation for the proposed model. The hyperparameters are skipped. Variables on plates are repeated the number of times indicated on the plate.

III. 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 hyper-parameters: 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 [17], [27] akin to the collapsed Gibbs sampling approach used with LDA [18], [31], [32]. A Gibbs sampler samples from the joint distribution of all the variables by sequentially sampling each variable from its distribution conditioned on the values of all the others. The individual topics are drawn from the conditional distribution

$$\mathrm{P}(z_{j,t}|\mathbf{w}, \mathbf{z}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{\neg t}) \propto \mathrm{P}(z_{j,t}, w_{j,t}|\mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{\neg t}), \tag{11}$$

where $\neg t = \{1, 2, \ldots, t-1, t+1, \ldots, T\}$ is a set of time indices excluding the current time index t, while $\boldsymbol{\theta}_{\neg t}$ 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:

$$\begin{aligned}&\mathrm{P}(z_{j,t}, w_{j,t}|\mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\theta}_{\neg t}, \boldsymbol{\phi}_{k,\neg t})\\ &= \mathrm{P}(w_{j,t}|z_{j,t}, \mathbf{z}_{\neg t}, \mathbf{w}_{\neg t}, \boldsymbol{\phi}_{k,\neg t})\mathrm{P}(z_{j,t}|\mathbf{z}_{\neg t}, \boldsymbol{\theta}_{\neg t})\end{aligned} \tag{12}$$

These can be sampled as seating arrangements in a corresponding Chinese restaurant with the t-th customer removed [27].

A. Chinese restaurant analogy

Chord unigram models can be drawn from the Pitman-Yor process using a procedure called Chinese restaurant process (CRP) [33]. A Chinese *restaurant* represents a sample of the chord posterior $\boldsymbol{\phi}_{k,\mathbf{u}(0)}$. There is one such restaurant for every topic k. Every restaurant consists of the seating arrangement of customers seated at an unbounded number of *tables*. Each arriving *customer* $w_{j,t}$ is randomly seated to a table and served the *dish* i assigned to this table ($w_{j,t} = i$) that has been drawn independently from U (*cf.* (10)). In our case, we know what the customers are eating, but do not know their seating arrangement. By sampling the seating arrangement, we can use it to calculate a sample of the chord posterior. Similarly, topic posteriors can be represented by seating arrangements in restaurants that are serving topics k to customers $z_{j,t}$.

The CRP technique can also be used for sampling in hierarchical Pitman-Yor processes, and therefore sampling for n-gram models, by introducing a hierarchy of restaurants [27], [30], [34]. In this hierarchy, the child restaurant $\boldsymbol{\phi}_{k,\mathbf{u}(n)}$ for the context $\mathbf{u}(n)$ and topic k is seating a customer, but also sending him/her to an appropriate parent restaurant $\boldsymbol{\phi}_{k,\mathbf{u}(n-1)}$.

In a Chinese restaurant sampling process, each arriving customer is seated either to 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 re-sampling the hierarchical seating arrangements, which is achieved by sequentially *adding* and *removing* customers to the hierarchy of restaurants. The chord posterior can be calculated recursively from the current seating arrangement in a hierarchy of chord-serving Chinese restaurants as

$$\begin{aligned}\phi_{i,k,\mathbf{u}(n)} \quad = \quad & \frac{N_{i,k,\mathbf{u}(n)} - d_{k,n}L_{i,k,\mathbf{u}(n)}}{\gamma_{k,n} + N_{k,\mathbf{u}(n)}} + \\ & + \frac{\gamma_{k,n} + d_{k,n}M_{k,\mathbf{u}(n)}}{\gamma_{k,n} + N_{k,\mathbf{u}(n)}}\phi_{i,k,\mathbf{u}(n-1)},\end{aligned} \tag{13}$$

where $L_{i,k,\mathbf{u}(n)}$ is the number of tables at which dish i is served in the restaurant for topic k and context $\mathbf{u}(n)$, $N_{i,k,\mathbf{u}(n)}$ is the number of people eating this dish in that restaurant; $M_{k,\mathbf{u}(n)}$ is the number of occupied tables in this restaurant and $N_{k,\mathbf{u}(n)}$ is the total number of people in that restaurant. For the parameters, Teh described an efficient sampling scheme using auxiliary variables [30], which we adopt in our implementation, but the same results can be achieved using other samplers, *e.g.*, slice sampling [35].

The hierarchical CRP sampling procedure for the proposed model is a straightforward adaptation of that proposed in [30]. It is outlined in Algorithm 1, while the functions for adding/removing customers and for calculating the posterior are detailed in Algorithm 2.

B. Hyper-parameters

This procedure assumes hyper-prior on the discount and strength parameters. As in [30], we place a beta hyper-prior on the discount parameters and a gamma hyper-prior on the strengths. The hyper-parameters are tied between all contexts $\mathbf{u}(n)$ of the same length for each topic k:

$$d_{k,n} \sim \mathrm{Beta}(\widetilde{a}_{k,n}, \widetilde{b}_{k,n}), \tag{14}$$

$$\gamma_{k,n} \sim \mathrm{Gamma}(\widetilde{\aleph}_{k,n}, \widetilde{\beth}_{k,n}), \tag{15}$$

```
Initialise ($\mathbf{z}$);
Initialise ($\mathbf{z}^{(\text{test})}$);
for $s \leftarrow 1$ to $S$ do
    for $j \leftarrow 1$ to $J$ do
        for $t \leftarrow 1$ to $T$ do
            if $s > 1$ then
                chordTrainingRestaurants [$z_{j,t}$, $\mathbf{u}_{j,t}(n)$].RemoveCustomer
                ($w_{j,t}$);
                topicTrainingRestaurants [$j$].RemoveCustomer ($z_{j,t}$);
                for $k \leftarrow$1 to $K$ do
                    $\phi_{w_{j,t},k,\neg t}$ = chordTrainingRestaurants [$k$,
                    $\mathbf{u}_{j,t}(n)$].DishProbability ($w_{j,t}$);
                    $\theta_{j,k,\neg t}$ = topicTrainingRestaurants [$j$].DishProbability ($k$);
                end
                $z_{j,t}$ = SampleCategorical ($\boldsymbol{\phi}_{w_{j,t},\neg t}\boldsymbol{\theta}_{j,\neg t}$);
            end
            chordTrainingRestaurants [$z_{j,t}$, $\mathbf{u}_{j,t}(n)$].AddCustomer ($w_{j,t}$);
            topicTrainingRestaurants [$j$].AddCustomer ($z_{j,t}$);
        end
        UpdateHyperparameters (topicTrainingRestaurants [$\cdot$]);
        for $k \leftarrow 1$ to $K$ do
            UpdateHyperparameters (chordTrainingRestaurants [$k$,$\cdot$]);
        end
        for $\tau \leftarrow 1$ to $T^{(\text{test})}$ do
            if $s > 1$ then
                topicTestRestaurants [$j_\tau$].RemoveCustomer ($z_\tau$);
                for $k \leftarrow$1 to $K$ do
                    $\phi_{w_\tau,k,\neg\tau}$ = chordTrainingRestaurants [$k$,
                    $\mathbf{u}_\tau(n)$].DishProbability ($w_\tau$);
                    $\theta_{j_\tau,k,\neg\tau}$ = topicTestRestaurants [$j_\tau$].DishProbability ($k$);
                end
                $z_\tau$ = SampleCategorical ($\boldsymbol{\phi}_{w_\tau,\neg\tau}\boldsymbol{\theta}_{j_\tau,\neg\tau}$);
            end
            topicTestRestaurants [$j_\tau$].AddCustomer ($z_\tau$);
        end
        UpdateHyperparameters (topicTestRestaurants [$\cdot$]);
    end
end
```

Algorithm 1: The Gibbs sampler for LHPYA. s is the sample number, t is used to index the training data $\mathbf{w}$, $\mathbf{z}$ and $\mathbf{j}$ (song numbers associated with each t) 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 *italics* denote variables, while the normal font functions and methods. The dot notation $[\cdot]$ stands for using the entire array.

with

$$\widetilde{a}_{k,n} = a_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{p=1}^{M_{k,q}-1} (1 - y_{q,p}), \tag{16}$$

$$\widetilde{b}_{k,n} = b_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{i} \sum_{p=1}^{M_{k,q,i}} \sum_{j=1}^{N_{k,q,p,i}-1} (1 - z_{q,i,p,j}), \tag{17}$$

$$\widetilde{\aleph}_{k,n} = \aleph_{k,n} + \sum_{q=1}^{Q_{k,n}} \sum_{p=1}^{M_{k,q}-1} y_{q,p}, \tag{18}$$

$$\widetilde{\beth}_{k,n} = \beth_{k,n} - \sum_{q=1}^{Q_{k,n}} \log x_q, \tag{19}$$

where $Q_{k,n}$ is the number of restaurants with context length $n-1$ in topic k with at least 2 occupied tables, $M_{k,q}$ is the number of occupied tables in such a restaurant q, $M_{k,q,i}$ is the

```
Restaurant:: AddCustomer($i$)
    if $n > 1$ then
        $\mathrm{P}(i|\mathbf{u}_{j,t}(n-1))$ = restaurants [$\mathbf{u}_{j,t}(n-1)$].DishProbability ($i$);
        restaurants [$\mathbf{u}_{j,t}(n-1)$].AddCustomer ($i$);
    else
        $\mathrm{P}(i|\mathbf{u}_{j,t}(n-1))$ = $\mathrm{U}(i)$;
    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
    $(\gamma_n + d_n M)\mathrm{P}(i|\mathbf{u}_{j,t}(n-1))$;
Restaurant:: RemoveCustomer($i$)
    if $n > 1$ then
        restaurants [$\mathbf{u}_{j,t}(n-1)$].RemoveCustomer ($i$);
    end
    remove a customer eating $i$ from table $p$ with probability
    proportional to $N_p$;
Restaurant:: DishProbability($i$)
    if $n > 1$ then
        return $\frac{N_i - d_n L_i}{\gamma_n + N} + \frac{\gamma_n + d_n M}{\gamma_n + N}$ restaurants
        [$\mathbf{u}_{j,t}(n-1)$].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::) for chord-serving restaurants. The DishProbability() method directly implements (13). The methods for topic-serving restaurants are analogous.

number of tables in that restaurant serving dish i with at least 2 customers and $N_{k,q,p,i}$ is the number of customers sitting in that restaurant at table p eating dish i. The auxiliary variables are sampled as follows:

$$x_q \sim \mathrm{Beta}(\gamma_{k,n_q} + 1, N_q - 1), \tag{20}$$

$$y_{q,p} \sim \mathrm{Bernoulli}(\frac{\gamma_{k,n_q}}{\gamma_{k,n_q} + d_{k,n_q} p}), \tag{21}$$

$$z_{q,i,p,j} \sim \mathrm{Bernoulli}(\frac{j-1}{j - d_{k,n_q}}), \tag{22}$$

where N_q 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 [17].

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

IV. Symbolic evaluation

The proposed model was first evaluated by measuring the cross-entropy (its normalised log-likelihood) on unseen (test) data. We have used data from the 3GDB data set [15], [36], a collection of hand-annotated chord labels for 856 songs from 3 genres: popular, jazz and academic music. Each genre had been further divided into 3 sub-genres: blues, celtic and pop (popular), bop, pre-bop and bossa nova (jazz), and baroque, romanticism and classical (academic). Each song has been annotated with both absolute chord labels (C-maj, C♯-maj, *etc.*) and chord degrees (Roman numeral notation: I-maj, ii-maj, *etc.*). We have used the chord degree data 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 key. The chords in 3GDB had been annotated with three levels of detail: full chord labels, triad-level chords (only the first two intervals of a chord) and dyad-level chords (only the first interval of a chord). The corresponding dictionary sizes (the numbers of distinct chord labels present in the database) were 262, 98 and 15. The entire collection of songs was divided in two equal-sized parts: the training and the test datasets.

A. *Cross-entropy*

The cross-entropy on the unseen (test) data $\mathbf{w}^{(\text{test})}$ is defined as [22]:

$$\mathrm{H}(\mathbf{w}^{(\text{test})}) = -\frac{1}{T} \log_2 \mathrm{P}(\mathbf{w}^{(\text{test})}|\mathbf{w}). \tag{23}$$

There are many possible methods for estimating the distribution over test data [32] and here we chose to use the unbiased estimator from [37], which is the harmonic mean of distribution samples collected using the Gibbs sampler from the previous section:

$$\mathrm{P}(\mathbf{w}^{(\text{test})}|\mathbf{w}) \approx \left(\frac{1}{S} \sum_{s=1}^{S} \mathrm{P}(\mathbf{w}^{(\text{test})}|\mathbf{w}, \mathbf{z}_s, \mathbf{z}_s^{(\text{test})})^{-1} \right)^{-1}, \tag{24}$$

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 have chosen the harmonic mean estimator for its simplicity, although Wallach *et al.* suggests it has tendencies for overestimation [32].

Unsurprisingly, we found that the model is sensitive to initialisation, because the topic labels $\mathbf{z}$ are correlated with each other and the Gibbs sampler tends to get stuck in modes of the joint distribution. As a result, sampling the initial topic labels $\mathbf{z}_{s=0}$ from a uniform distribution resulted in a slightly different estimate of the cross-entropy each time. To minimise this variation, we initialised the topic assignments with the topics obtained with the Gibbs-EM algorithm used on the BTM from [18] (we have used "prior 2" from that paper).

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 then waited another $B_2 = 500$ iterations before starting to collect $S = 500$ likelihood samples to calculate the cross-entropy. These values have been determined experimentally.

B. *Results*

Fig. 2 shows a plot of cross-entropies obtained by setting the number of topics to $K = 1$, *i.e.*, for a smoothed n-gram model. For all chord label detail levels we observe a minimum at $R = 2$, after which the models slowly start to overfit to the data. In the following experiments we therefore focus on trigrams.

Cross-entropies for 3-grams and different number of latent topics are plotted in Fig. 3. The cross-entropy is decreasing quickly with the number of latent topics, until about $K = 30$ for dyads, where the model starts to overfit. The curves for triads and full chord labels saturate at larger values of K because of higher data complexity, but do not decrease significantly above $K = 60$.

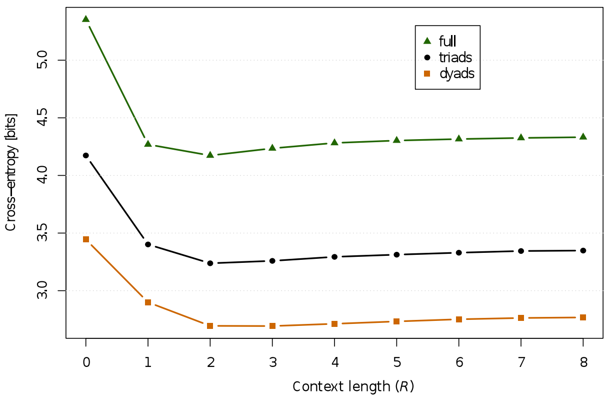


Fig. 2: Cross-entropies for smoothed chord n-grams, calculated on the test data.

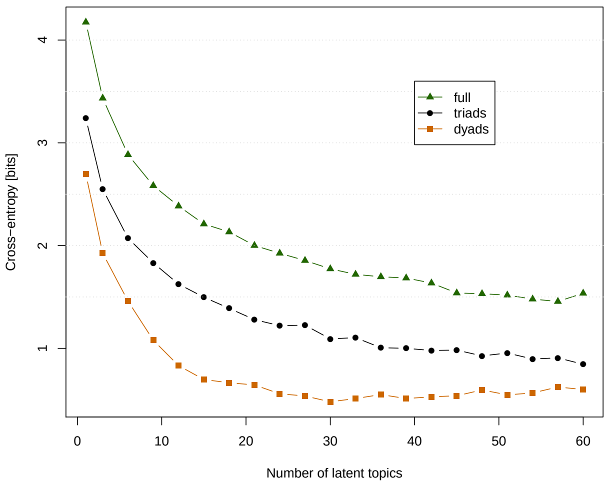


Fig. 3: Cross-entropies of the topic-based 3-grams, as a function of K, calculated on the test data for all three data sets.

For full chord labels the minimal cross-entropy achieved is 1.45 bits per chord, for $K = 57$. Without the topic model, *i.e.*, for trigrams, the cross-entropy is 4.17 bits and without the n-grams, *i.e.*, with a unigram topic model we only achieve 2.95 bits for the optimal $K = 120$. The results are summarised in Table I, together with corresponding values of compression rate (calculated as the ratio of the cross-entropy to the reference cross-entropy of the uniform model) and perplexity (calculated as $2^{\mathrm{H}(\mathbf{w}^{(\text{test})})}$). We can therefore conclude that the chordal context and the latent topics are both factors that significantly and independently reduce the uncertainty about predicted chords.

V. Musical Genre recognition

As a complementary way of evaluating the proposed model, we apply it to chord-based music genre recognition. Assigning

Model	Cross-entropy	Compression rate	Perplexity
Uniform	8.03	0%	262.
Trigrams	4.17	48%	18.05
Unigram topic	2.95	63%	7.73
Trigram topic	1.45	82%	2.74

TABLE I: Cross-entropy, as well as equivalent compression rates and perplexities obtained for three models, compared to the reference uniform model for full chords.

genre labels is a very common way of categorizing music [38] and harmony is one of the key discriminants of musical genre (at least in the Western tonal music). 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 only the features extracted from audio signals [39], mostly because this kind of data is readily available in large amounts. Few of them use extracted chord sequences [40], [41]. However MGR based on such symbolic features is actually more likely to break the glass ceiling of MGR in the long term, because of the semantic information embodied by these features.

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

$$\widehat{g} = \arg\max_{g} \mathrm{P}(g|\boldsymbol{\theta}). \tag{25}$$

We propose two ways to compute the genre posterior: using a generative approach and an approach based on naive Bayes classifiers.

A. Generative approach

In the generative approach, the genre is found as

$$\widehat{g} = \arg\max_{g} \prod_{s=1}^{S} \sum_{k=1}^{K} \mathrm{P}(g|z=k)\mathrm{P}(z=k|\boldsymbol{\theta}_s), \tag{26}$$

where z is a topic for this document, s indexes Gibbs samples of the topic posteriors, $\boldsymbol{\theta}_s$ is a sample of the topic posterior for this document $\theta_{s,k} = \mathrm{P}(z=k|\boldsymbol{\theta}_s)$ and

$$P(g|z=k) = \frac{\mathrm{P}(g, z=k|\mathbf{z}^{(\text{training})})}{\mathrm{P}(z=k|\mathbf{z}^{(\text{training})})}. \tag{27}$$

B. Naive Bayesian classifiers

Another approach is an approach using naive Bayesian classifiers:

$$\widehat{g} = \arg\max_{g} \mathrm{P}(g|\boldsymbol{\theta}, \Lambda) = \arg\max_{g} \prod_{s=1}^{S} \mathrm{P}(\boldsymbol{\theta}_s|g, \Lambda)\mathrm{P}(g), \tag{28}$$

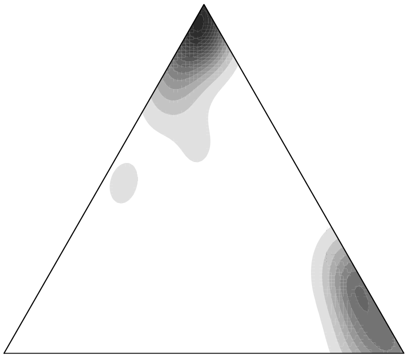


Fig. 4: A topic posterior distribution over a K-simplex for the 'romanticism' sub-genre, where $K = 3$, $L = 9$, $R = 1$ and full chord labels are used. Darker shades represent larger density. The distribution shows strong bimodality.

where $\mathrm{P}(\boldsymbol{\theta}_j^{(\text{test})}|g)$ is a parametric probability density function with parameters Λ that can easily be trained from the training data.

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. 4), we use Dirichlet mixtures [42]:

$$\mathrm{P}(\boldsymbol{\theta}_s|g, \Lambda) = \sum_{d=1}^{D} \lambda_d \mathrm{Dir}(\boldsymbol{\theta}_s; \boldsymbol{\rho}_d), \tag{29}$$

where λ are the mixing coefficients and the Dirichlets are parameterised by K-element vectors $\boldsymbol{\rho}_d$. The mixture coefficients as well as the parameters of the Dirichlets can be found using the EM algorithm (detailed in the Appendix).

C. Results

Experiments were performed for all combinations of the parameters: the number of genres and sub-genres $L \in \{3, 9\}$, the number of topics $K \in \{3, 9, 12, 15, 18, 21, 24, 27, 30, 39, 48\}$ and the context length $R \in \{0, 1, 2, 3, \}$. All three chord detail levels were used (dyads, triads, full chords) and 7 estimation methods: generative, 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. We have used the same division into training and test data as in the symbolic experiments.

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

$$\mathcal{A} = \frac{1}{L} \sum_{g=1}^{L} \frac{NP_g}{NT_g}, \tag{30}$$

where NP_g is the number of correctly identified songs for genre g and NT_g is the total number of songs in that genre.

Fig. 5 shows a plot of accuracies relative to the accuracy for dyads, for all values of L, K, R and all methods. We see that higher detail in chord descriptions translated to an accuracy

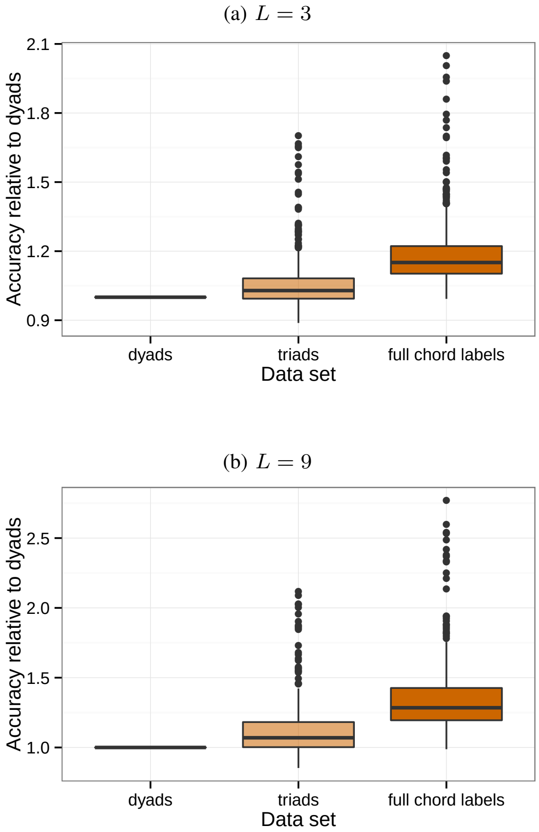


Fig. 5: Boxplot of accuracies obtained on the three data sets, relative to the dyad data set accuracy, for all combinations of K, R and estimation method. The boxes represent the ranges of the second and third quartiles.

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. 6. For $L = 3$, there is little difference between analysis methods if an optimal number of topics K is used (depicted by the bar tops), although mixtures of many Dirichlets are slightly better. For both $L = 3$ and $L = 9$ the best accuracy was generally obtained for $R = 1$ and $R = 2$, which confirms the observation from Fig. 2 that longer contexts do not improve the modelling power, at least for such a small data set. This is also consistent with the results presented in [15], [43].

On the other hand, accuracy increases significantly for $L = 9$ if Dirichlet mixtures are used, by 40% in relative terms (17% absolute) compared to the generative approach. Furthermore, this time the highest accuracy is generally obtained for $R = 2$, so we can conclude that the context is more important in distinguishing between sub-genres than between genres.

Although there is quite some variance in the obtained accuracies (*cf.* Fig.7), we were able to achieve a genre recognition accuracy of 88% for $L = 3$ and 60% for $L = 9$ for the mixture of 24 Dirichlets. As shown in Table II, this is better than the results for n-grams reported both in [15] and [43] by 33% in relative terms (15% absolute) and comparable to the naïve Bayesian classifier with multivariate Bernoulli distribution from [43]. We do not report the results for hierarchical classifiers from [43], because those used both harmony and melodic information.

L	Proposed	[15]		[43]	
		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 II: The best accuracies obtained by the proposed method and the accuracies cited in the reference work by Pérez-Sancho *et al.* in [15] and [43] for the same data (degrees with extensions, full chord labels). * In the case of [15] the number of genres L was 8.

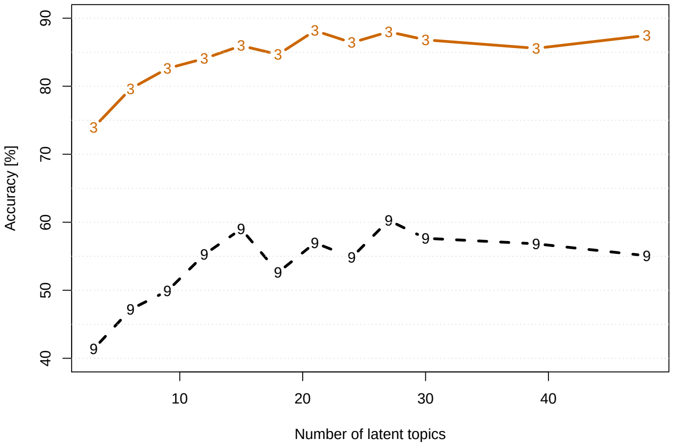
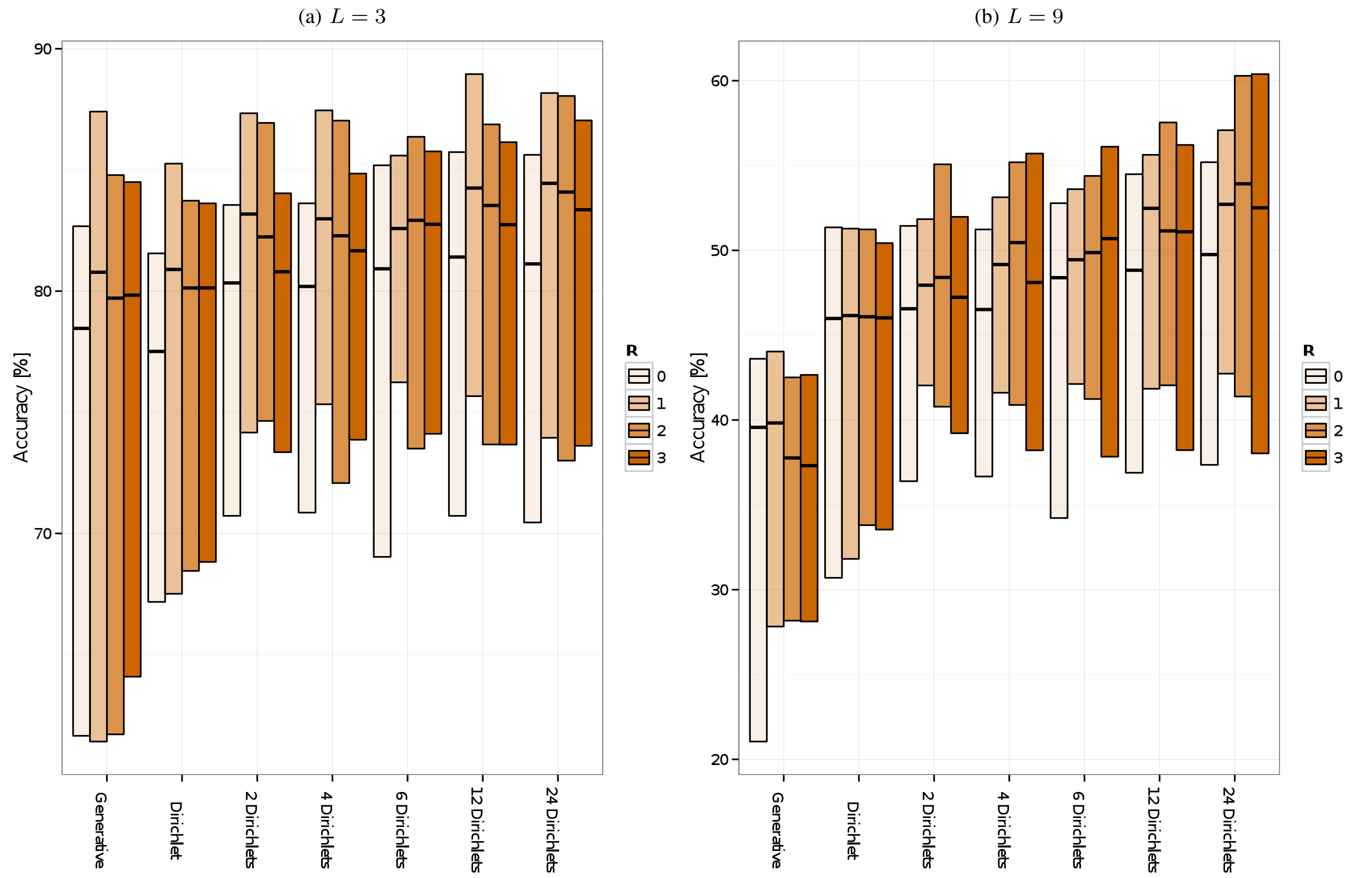


Fig. 7: Genre estimation accuracy as a function of the number of topics K for $L = 3$ (orange solid line) and $L = 9$ (black dashed line). A mixture of 24 Dirichlets was used with $R = 1$ and $R = 2$, respectively.

Fig. 7 shows the accuracies obtained for the best parameter values as a function of K. 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. 8 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 and 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 most confusion is between closely related sub-genres of the same genre, *e.g.*, between bop and pre-bop or between baroque and romantic music.

VI. Conclusion

In this paper we have presented a new smoothed topic model using hierarchical Pitman-Yor process priors and applied it to prediction of chord sequences and to chord-based musical genre recognition. It integrates a topic model with word



(a) $L = 3$

(b) $L = 9$

Fig. 6: Accuracies obtained for all analysis methods (marked on the horizontal axis) and values of R. Bars correspond to minimal, mean and maximal values over all K's for a particular method and value of R.

(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 model by allowing both the songs and the genres to be mixtures of latent topics. Of course, the proposed model is not limited to modelling chords and it would be interesting to apply it to other musical sequences, such as pitches (*e.g.*, melodies or voices), rhythm, *etc*.

In general, this model is potentially useful in all problems where topic models are used and the order of words matters, *e.g.* in text document modelling [18] or in genomic information analysis [44].

References

[1] E. Vincent, S. Raczyński, N. Ono, and S. Sagayama, “A roadmap towards versatile MIR,” in *Proc. 11th International Conference on Music Information Retrieval (ISMIR)*, 2010, pp. 662–664.

[2] 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)*, 2007, pp. 255–258.

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

[4] 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)*. IEEE, 2007, pp. 53–60.

[5] 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)*. IEEE, 2010, pp. 5518–5521.

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

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

[8] S. A. Raczyński, S. Fukayama, and E. Vincent, “Melody harmonisation with interpolated probabilistic models,” INRIA, Research Report RR-8110, October 2012. [Online]. Available: http://hal.inria.fr/hal-00742957

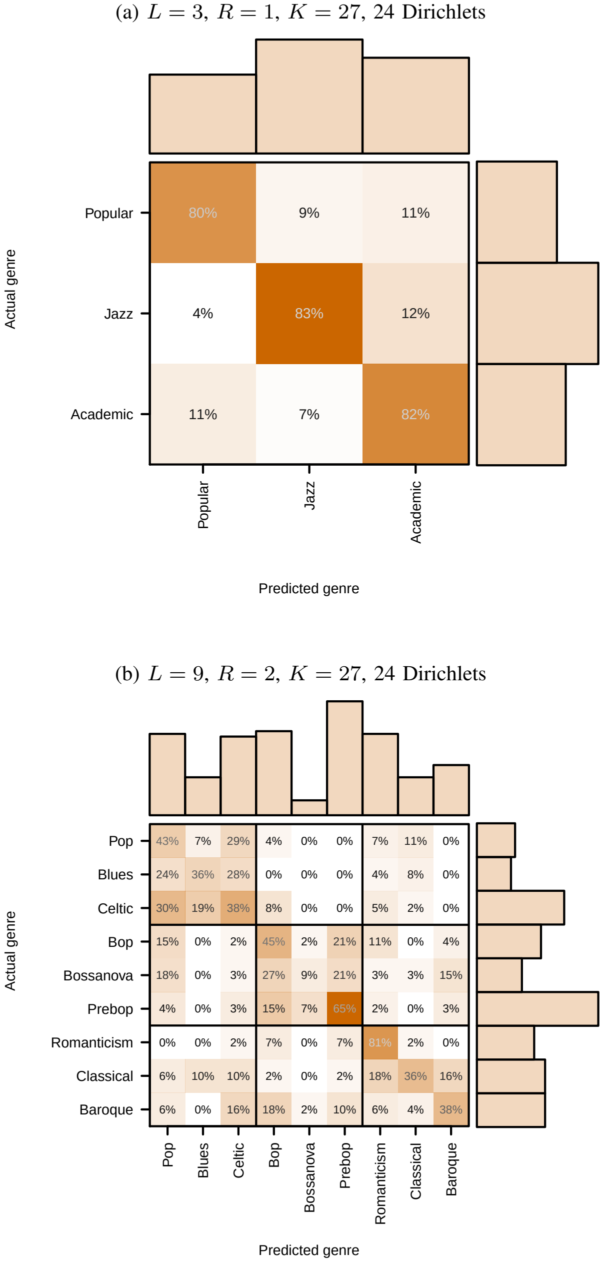
[9] “PG Music Inc. Band-in-a-Box,” http://www.pgmusic.com/, August 2012.

[10] 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*, 2008, pp. 725–734.

[11] 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. 7th Sound and Music Computing Conference (SMC)*, 2010, pp. 299–302.

[12] 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)*, 2011, pp. 483–488.

[13] M. Mauch, K. Noland, and S. Dixon, “Using musical structure to


(a) $L = 3$, $R = 1$, $K = 27$, 24 Dirichlets

(b) $L = 9$, $R = 2$, $K = 27$, 24 Dirichlets

Fig. 8: Confusion matrices for best sets of parameters. Shades correspond to absolute counts of labelling an "actual genre" as a "predicted count". Marginal barplots show total counts for the actual and predicted genres. Additionally, the text labels show counts relative to the total number of songs for a particular actual genre (*i.e.*, rows sum up to one)

enhance automatic chord transcription," in *Proc. 10th International Society for Music Information Retrieval (ISMIR)*, 2009, pp. 231–236.

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

[15] C. Pérez-Sancho, D. Rizo, and J. M. Iñesta, "Genre classification using chords and stochastic language models," *Connection science*, vol. 21, no. 2-3, pp. 145–159, 2009.

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

[17] 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*, 2010, pp. 673–681.

[18] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proc. 23rd International Conference on Machine Learning*. ACM, 2006, pp. 977–984.

[19] 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)*, 2009, pp. 441–446.

[20] A. Spiliopoulou and A. Storkey, "A Topic Model for Melodic Sequences," *ArXiv e-prints*, Jun. 2012.

[21] 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)*. IEEE, 2009, pp. 53–56.

[22] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. 34th annual meeting on Association for Computational Linguistics*. ACL, 1996, pp. 310–318.

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

[24] 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*. ACM, 2001, pp. 334–342.

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

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

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

[28] M. Kaban and G. Ata, "Simplicial mixtures of markov chains: Distributed modelling of dynamic user profiles," *Advances in neural information processing systems*, vol. 16, p. 9, 2004.

[29] D. J. C. MacKay and L. C. B. Peto, "A hierarchical Dirichlet language model," *Natural Language Engineering*, vol. 1, no. 3, pp. 1–19, 1995.

[30] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney," National University of Singapore, School of Computing, Tech. Rep., 2006.

[31] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 (Suppl 1), 2004, pp. 5228–5235.

[32] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proc. 26th International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.

[33] J. Pitman, "Combinatorial stochastic processes," University of California Berkeley, Dept. Statistics, Tech. Rep. 621, 2002.

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

[35] 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*, 2008, pp. 15–20.

[36] "3 Genre Database (3GDB)," http://grfia.dlsi.ua.es/cm/projects/prosemus/database.php, February 2013.

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

[38] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

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

[40] A. Anglade, R. Ramirez, and S. Dixon, "Genre classification using harmony rules induced from automatic chord transcriptions," in *Proc. 10th International Society for Music Information Retrieval (ISMIR)*, 2009, pp. 669–674.

[41] 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)*, 2010, pp. 279–284.

[42] M. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjolander, D. Haussler *et al.*, "Using Dirichlet mixture priors to derive hidden Markov models for protein families," in *Proc. of the First International Conference on Intelligent Systems for Molecular Biology*, 1993, pp. 47–55.

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

[44] X. C., X. H., X. S., and G. Rosen, "Probabilistic topic modeling for genomic data interpretation," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 149–152.

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

[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

A. Learning Dirichlet mixtures

The Dirichlet distribution is defined as [45]:

$$\mathrm{Dir}(\boldsymbol{\theta};\boldsymbol{\rho}) = \frac{1}{B(\boldsymbol{\rho})}\prod_{k=1}^{K}\theta_k^{\rho_k-1}, \tag{31}$$

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

$$B(\boldsymbol{\rho}) = \frac{\prod_{k=1}^{K}\Gamma(\rho_k)}{\Gamma(\sum_{k=1}^{K}\rho_k)}, \tag{32}$$

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 EM algorithm [46]. The parameter likelihood is given by

$$\mathcal{L}(\Lambda;\boldsymbol{\theta},\mathbf{v}) = \mathrm{P}(\boldsymbol{\theta},\mathbf{v}|\Lambda) = \prod_{j=1}^{J}\sum_{d=1}^{D}\mathbf{I}(v_j,d)\lambda_d\mathrm{Dir}(\boldsymbol{\theta}_j;\boldsymbol{\rho}_d), \tag{33}$$

where $\mathbf{I}(v_j,d)$ is a binary indicator function that is equal to one iff $v_j = d$ and otherwise equal to zero.

B. E-step

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

$$Q(\Lambda|\Lambda_l) \;=\; \mathrm{E}_{\mathbf{v}|\boldsymbol{\theta},\Lambda_l}\left[\log\mathcal{L}(\Lambda;\boldsymbol{\theta},\mathbf{v})\right], \tag{34}$$

where l indexes iterations of the EM algorithm. We define

$$\begin{aligned} T_{l,j,d} &= \mathrm{P}(v_j|\boldsymbol{\theta},\Lambda_l) \\ &= \frac{\lambda_d\mathrm{Dir}(\boldsymbol{\theta}_d;\boldsymbol{\rho}_{l,d})}{\sum_{d'=1}^{D}\lambda_{d'}\mathrm{Dir}(\boldsymbol{\theta}_{d'};\boldsymbol{\rho}_{l,d'})}. \end{aligned} \tag{35}$$

Then

$$\begin{aligned} Q(\Lambda|\Lambda_l) = \;& \sum_{j=1}^{J}\sum_{d=1}^{D}T_{l,j,d}\Big(\log\lambda_{l,d} - \log B(\boldsymbol{\rho}_l) + \\ & + \sum_{k=1}^{K}(\rho_{l,d,k}-1)\log\theta_{j,k}\Big). \end{aligned} \tag{36}$$

C. M-step

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

$$\begin{aligned} \lambda_{l+1,d} &= \arg\max_{\lambda}\sum_{d=1}^{D}\log\lambda_d\sum_{j=1}^{J}T_{l,j,d} \\ &= \frac{1}{J}\sum_{j=1}^{J}T_{l,j,d} \end{aligned} \tag{37}$$

$$\begin{aligned} \frac{\partial}{\partial\rho_{l,d,k}}Q(\Lambda|\Lambda_l) &= \Big(\Psi(\sum_{k'=1}^{K}\rho_{l,d,k'}) - \Psi(\rho_{l,d,k})\Big)\sum_{j=1}^{J}T_{l,j,d} + \\ & + \sum_{j=1}^{J}T_{l,j,d}\log\theta_{j,k}, \end{aligned} \tag{38}$$

where Ψ is the digamma function. The new $\boldsymbol{\rho}$ can therefore be found as [45]:

$$\rho_{l+1,d,k} = \Psi^{-1}\left(\Psi\left(\sum_{k'=1}^{K}\rho_{l,d,k'}\right) + \frac{\sum_{j=1}^{J}T_{l,j,d}\log\theta_{j,k}}{\sum_{j=1}^{J}T_{l,j,d}}\right). \tag{39}$$

Stanisław A. Raczyński is an assistant professor at the Gdańsk University of Technology (Poland). He received his M. Eng. degree in automatic control and robotics from Gdańsk University of Technology in 2006 and his Ph. D. degree in information physics and computing from the University of Tokyo (Tokyo, Japan) in 2011. He worked as an assistant professor at the University of Tokyo in 2011 and as postdoctoral research fellow with Inria (Rennes, France) from 2011 to 2013.

His main research interests are in statistical signal processing and its applications to music information retrieval, as well as audio modality in mobile robotics.

Emmanuel Vincent (M'07 - SM'10) is a Research Scientist with Inria (Nancy, France). He received the Ph. D. degree in music signal processing from IRCAM (Paris, France) in 2004 and worked as a Research Assistant with the Centre for Digital Music at Queen Mary, University of London (London, U.K.) from 2004 to 2006.

His research focuses on probabilistic machine learning for speech and audio signal processing, with application to real-world audio source localization and separation, noise-robust speech recognition and music information retrieval. He is a founder of the series of Signal Separation Evaluation Campaigns (SiSEC) and CHiME Speech Separation and Recognition Challenges. Dr. Vincent is an Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing.