Modelling Digital Media Objects

Troelsgaard, Rasmus; Hansen, Lars Kai; Larsen, Jan

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Modelling Digital Media Objects

Rasmus Troelsgaard

Kongens Lyngby 2016

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

The goal of this thesis is to investigate two relevant issues regarding computational representation and classification of digital multi-media objects. With a special focus on music, a model for representation of objects comprising multiple heterogeneous data types is investigated. Necessary to this work are considerations regarding integration of multiple diverse data modalities and evaluation of the resulting concept representation.
Regarding modelling of data exhibiting certain sequential structure, a number of theoretical and empirical results are presented. These are results related to model parameter estimation and the use of sequence models in a classification scenario. The latter being of importance in various digital multimedia navigation and retrieval tasks.

In the fields of *topic modelling* and *multi-modal integration*, we formulate a model to describe entities composed of multiple aspects. The particular aspects considered in the publications are sound, song lyrics, and user-provided metadata. This model integrates the diverse data types comprising the objects and defines concrete unified representations in a joint "semantic" space. Within the context of this model, general measures of similarity between such multi-modal objects are investigated.

In the fields of *method of moments* and *sequence modelling*, we increase practical applicability of a certain moment based parameter estimation method for Hidden Markov models by showing how to use full-length sequences in the estimation process. Consequently, this impacts the quality of the estimated model parameters.

Subsequently, we show how to perform time series classification using a composite likelihood formulated from third order moments defined by the Hidden Markov model. Compared to the conventional likelihood based method, our contribution is less computationally expensive, while retaining the level of classification performance.

Summary (Danish)

Denne afhandling undersøger to relevante problemstillinger i forbindelse med repræsentation og kategorisering af digitale medieobjekter. Med særligt fokus på musik undersøges et system til repræsentation af objekter udgjort af flere forskelligartede datatyper. Dette indebærer overvejelser om kombination af heterogene data, og om evaluering af den resulterende objekt-repræsentation.
I forbindelse med modellering af sekventielle data præsenteres der en række teoretiske og empiriske resultater, der knytter sig til henholdsvis model-estimering og brugen af sekvensmodeller i forbindelse med kategorisering eller "tagging". Sidstnævnte kan facilitere søgning i digitale mediedatabaser.

Inden for emnerne *emnemodellering* og *multimodal integration* formulerer vi en model til at beskrive enheder bestående af flere aspekter, i vores tilfælde lyd, sangtekst, og metadata indtastet af musikforbrugere.
Dette gøres ved at lade hver enkelt aspekt bidrage til en fælles beskrivelse af objektet som helhed. Samme model bliver benyttet til at undersøge generelle mål for similaritet mellem sådanne multi-modale objekter.

Inden for emnerne *momentmetoder* og *tidsserieanalyse* viser vi, hvordan det ved brugen af tredje-ordens momenter til estimation af skjulte Markov-modeller er muligt at benytte den fulde længde af observerede datasekvenser. Dette betyder helt praktisk, at man kan øge præcisionen af estimaterne med samme antal tilgængelige datasekvenser.
Herefter viser vi, hvordan det er muligt at udføre kategorisering af datasekvenser ved hjælp af skjulte Markov-modeller og empiriske tredje-ordens momenter. Sammenlignet med en traditionel likelihood-baseret metode resulterer vores bidrag i reduceret beregningstid ved klassifikation af lange sekvenser.

Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science (DTU Compute), Technical University of Denmark (DTU) in partial fulfilment of the requirements for acquiring the degree of PhD.

The project work was carried out between December 2012 and October 2016, and was supervised by professor Lars Kai Hansen (DTU Compute) and co-supervised by professor Jan Larsen (DTU Compute) for the full project period. The project was funded in part by DTU and in part by the Innovation Fund Denmark under the CoSound project, case number 0603-00475B. This publication only reflects the authors' views.

The thesis reflects the main parts of the research conducted during the project period, and it consists of published and submitted research papers accompanied by a summary report. All included publications have been prepared during the project period.

The reader is assumed to possess a basic level of knowledge in the field of statistics, probability theory and their typical use in machine learning and signal processing applications. Hence this thesis will not provide details of widely known concepts and methods, but rather cite relevant literature. The summary report should not be regarded as rigorous and detailed exposition but rather be seen as an attempt to place the contributions in the general context of modelling abstract multimedia concepts.

Rasmus Troelsgaard
Lyngby, 22-October-2016

Dissemination

Papers (peer-reviewed)

C Rasmus Troelsgaard and Lars Kai Hansen. "Spectral Learning of Hidden Markov Models in Non-stationary Data". In: *[submitted]* (2016)

D Rasmus Troelsgaard and Lars Kai Hansen. "Sequence Classification Using Third Order Moments". In: *[submitted]* (2016)

A Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen. "Towards a universal representation for audio information retrieval and analysis". In: *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2013. DOI: 10.1109/ICASSP.2013.6638242

Contribution to Workshops (peer-reviewed)

B Rasmus Troelsgaard, Bjørn Sand Jensen, and Lars Kai Hansen. "A Topic Model Approach to Multi-Modal Similarity". In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. 2013, pp. 1–5. arXiv: arXiv:1405.6886v1

Papers that are not part of this thesis

- Morten Hertzum, Haakon Lund, and Rasmus Troelsgaard. "Retrieving Radio News Broadcasts in Danish: Accuracy and Categorization of Un-

recognized Words". In: *28th Australian Conference on Human-Computer Interaction (OzCHI)*. 2016, pp. 1–4

Contents

Chapter 1

Introduction

We have gradually got used to the vast amount of content made available through the internet, and people have come to rely heavily on information retrieval services such as search engines. Until recently, the ability to find an item based on an abstract thought described through search query terms has been confined by whether or not another human being has explicitly provided a matching description for that particular item.
This issue arises because abstractions form the basis of human understanding and communication of ideas and thoughts.
Abstractions facilitate reasoning about complex issues without considering all the details while doing so. Hence, search engine performance is in many cases limited by the fact that the description of every item in the index has to originate from a cognitive process. This limitation has escalated with enormous amounts of multimedia constantly being co-produced and uploaded by people around the world, who do not care or do not have time to provide proper descriptions in natural language. Moreover, any natural language description of an abstract entity will be imperfect due to the mere nature of generality of an abstraction.
A natural question to ask in this situation is whether it is possible to generate proxies for cognitive representations with little or no human interaction. This is one of the main questions driving the machine learning research field, and considerable progress has been made by academia and industry in learning representations for tasks such as object recognition in images and machine translation of natural language.

In this thesis, we use the term "abstraction" to signify a representation of an entity produced from a combination or grouping of physical measurements or lower level abstractions, creating a many-to-one mapping between entities at the lower level and entities at the higher level. Hence, an increase in level of abstraction corresponds to an increase in generality and variability and a decrease in specificity and detail. The inherent polysemy in abstractions is extremely difficult to represent in a formalised model. This is hardly surprising considering that ambiguities often lead to misunderstandings in human-to-human communication.

Tightly coupled with the notion of abstractions is the notion of similarity. The sense of similarity of abstract concepts at some level of abstraction can define groups of similar or dissimilar items. A group of items perceived as similar at one abstraction level might give rise to a new abstract concept representing the group as a whole at a higher level of abstraction. To formally represent abstract concepts and measure similarity between them, we can create mathematical models. The model itself may be described as a set of rules for interaction between variables, and the values of model parameters can be viewed as the strengths of these interactions. The question of how to measure similarity is obviously deeply dependent on the chosen model. Some models might express certain sequential or spatial structures of interest e.g. when dealing with time series or image data. Other models might represent objects as entities composed of un-ordered collections of sub-objects.
Common to all such models is the aim of creating abstract representations comprehensible to human beings, hence the value of such systems increases with their ability to produce output aligned with human cognition.

If we accept the premise that the ultimate goal of modelling is to produce abstractions exhibiting similar properties to human perception of concepts, a limit likely exists to which level in the hierarchy of abstraction one can expect to represent using only physical inputs. i.e. excluding any product of human cognitive processes. This limit is caused by user-provided information not present in the observed data object, taking part in the production of higher level abstractions, and it is often referred to as the semantic gap [6]. Fortunately, parts of the semantic gap can be bridged by inclusion of user-specific data such as preference, descriptions of perception and experience, and categorisation with the purpose of communication based on a more or less common inter-human understanding of the terms. User-specific data can be used in various ways to reason and make decisions. The classical example of models based solely on user preference data is that of collaborative filtering [7]. In general, supervised learning schemes are often defined using target variables that are aligned with common human concepts, and the objective is to predict the values of unobserved target values

as closely as possible. Unsupervised learning methods seek to construct useful representations by learning or imposing certain structure on both physical measurements and user provided data. Unsupervised learning is often applied for exploratory data analysis and modelling of data in cases where an exact task has not or cannot be defined. Combinations of the two approaches are often termed semi-supervised learning techniques and usually rely on a few items for which the target variables are known, and a lot of items for which they are not. The intuition behind semi-supervised learning is that the labelled items guide the learned representation, and the unlabelled items improve generalisation performance.

One of the important components of most machine learning techniques and in particular unsupervised learning, is the transformation of observations into representations on higher levels of abstraction. An example of a process increasing the abstraction level is what is commonly known as feature extraction. Most feature extraction techniques process physical measurements and represent them in terms of usually very low level concepts. E.g. when modelling western tonal music tracks, a commonly used feature is the abstraction of fundamental frequency, from which further abstractions such as chords and keys can be modelled [8, 9].

This thesis focuses on a few models for creating representations of abstract entities. It does so in the form of an admixture model, a type of mixture model allowing for each item to be represented as a convex combination of multiple components. In contrast, the basic mixture model assumes that each item is generated by a single component.
The second model considered is the Hidden Markov model which effectively is a mixture model of sequence data with a dependency structure between neighbouring observations. Both models make use of what is often referred to as latent variables. Latent variables are unobserved variables assumed to follow some specific probability distributions making sense from a human perspective, and consequently they are the main component in obtaining abstract representations. Hence abstract entities can effectively be represented by inferred parameters which in turn can be compared to other items' parameters according to some measure of similarity.

Having formulated a mathematical model believed to enable representation of a certain type of abstractions, one has to estimate its parameters or make inference about the parameter distributions. A great variety of methods exists for doing so ranging from numerical optimisation of heuristic loss functions to methods with roots in probability theory. In this thesis two different approaches to parameter learning have been applied. In contribution A, Markov Chain Monte

Carlo (MCMC) is used to generate samples from the posterior distribution of the model parameters. In contribution C, a variant of the method of moments is used for parameter estimation in the Hidden Markov model (HMM).

The contributions of this thesis aim to address some of the sketched issues by dealing with a few very specific models for representation of abstract concepts. The treated methods are examples of unsupervised learning used to represent higher level structure in the data.

To summarise, the contributions of this thesis are as follows. An admixture model for representation of music tracks using multiple heterogeneous data types is treated in A. Furthermore, alignment with user-provided data is evaluated and discussed. C and D treat modelling of temporal structure using Hidden Markov models and the work builds upon a recently proposed parameter estimation method based on sub-sequences of length 3 [10]. Contribution C theoretically justifies application of the method with data sequences of lengths >3 for which stationarity cannot be assumed, and shows empirical evidence for improved performance. Contribution D formulates a distance measure between estimated models and observed sequences for use in model based classification scenarios. The proposed distance measure is based on a composite likelihood formulated from third order moments.

The remaining chapters of this thesis consist of specific introductions to the scientific contributions and their related areas of research. Finally, a conclusion summarises the key findings.

Chapter 2

Representation of Multi-Modal Objects

Higher level abstractions are often multi-modal in the sense that they are constituted by heterogeneous collections of other abstract objects. In applications depending on representations of such entities, we therefore need models describing the contributions of each of the sub-components and the possible interplay between them. While a unified representation of multi-modal objects is an intriguing thought, it is an extremely hard problem to solve because of the combinatorial nature of the problem. However the idea opens up for opportunities e.g. to integrate multiple observed object properties with social and behavioural measurements such as buying- and usage patterns, and utterances or ratings possibly relating physical objects to human cognitive concepts. This link to perception is a crucial component in construction of systems designed with human interaction in mind [11].

Because a measure of similarity is key to form relations between concepts, the pertinence of a common latent representation is evident for any multi-modal concept for which a sufficiently large number of people can agree on a general sense of similarity. In such situations, the multiple views constituting a concept has to contribute to the common representation. The general question of how to integrate multiple heterogeneous sources of information for modelling purposes has been investigated heavily, and with more and more multi-modal data sets

available, the question is of increasing relevance. Works such as [12] and [13] give thorough reviews on the variety of methodologies. An important concept is the level of fusion of the modalities. *Early* fusion denotes the process of integration of modalities at the feature level and is often performed by concatenation of feature vectors, whereas *late* fusion describes the situation of integration at the "semantic" (decision) level. Naturally, integration at different levels can be performed as well and is referred to as *hybrid* fusion [13].
Given a problem concerning input data consisting of multiple modalities, the choice of fusion level is important both in terms of performance and model properties. If features in different modalities live on different time scales or are otherwise not directly compatible, late fusion is generally easier to implement. This argument is used in [14] where audio-video integration for classification with HMMs is investigated. Relying on late integration might however miss possible correlation structure among features of different modalities. This is noted by both [15] and [16] who formally link late fusion to the assumption of conditionally independent modalities. The severity of this effect is obviously controlled by the modality specific components of the model.
In [17] the features of each modality are transformed non-linearly to presumably retain their individual distributional properties, while at the same time making them compatible with the other transformed modalities, thereby allowing fusion. In these new representations of the individual modalities, correlations might still be present and exploited in the construction of a joint representation of image-text pairs using Deep Boltzmann Machines. Fusion at a common level as opposed to hybrid fusion, implies that modalities are treated symmetrically which, depending on the task at hand, can be more or less appropriate in such multi-modal generative models.

An alternative take on the fusion of transformed modalities is that of fusion of systems. This view allows fusion to be done in a pipeline or hierarchical fashion where some of the systems may be constructed independently of the others, and subsequently be used to guide the remaining systems e.g. by providing prior beliefs. This fusion of systems approach has recently been applied within music research, where an instrument detection model was used to guide and automatic music transcription system [18]. Another example, applied but not limited to music classification is the idea of Bag of Systems (BoS) presented in [19]. BoS represents objects as counts of an un-ordered set of prototypes being generative models themselves.
Interestingly, the traditional pipeline approach: pre-processing $\rightarrow$ feature extraction $\rightarrow$ modelling can also be seen as a type of system fusion.

One very intuitive way of describing relations between multiple measurements is through probabilistic latent variable models (LVM). This class of models allows

for defining complex conditional dependency structures between observed and un-observed variables via relatively simple distributional arguments, reasoning and common sense. One limitation of this kind of model building methodology is that it only allows for interactions between variables if specifically expressed by the designer. On the other hand, LVMs are often quite easy to interpret because of their explicit structure.

One of the obstacles to overcome is that data from different modalities are often of diverse data types, non-synchronous, and come in different quantities for a single object. This raises the question of how to control the relative weighting of different modalities. Contribution A applies a latent variable model to obtain representations of music tracks. The specific model used is a multi-modal variant of the formerly very popular Latent Dirichlet Allocation model [20]. The model implements late symmetric fusion of information derived from the music itself as well as user generated text data and category labels to obtain a joint semantic space along the lines of [17] and [21].
In the mmLDA model, the influence of each modality is decided by its abundance, and there is no obvious principled way to control this influence. On the other hand, this very same construction ensures the existence of a representation of an object even if only a single modality is observed. Related to this question, another approach to multi-modal similarity is presented in [22], where a multi-modal PLSA model is obtained by optimising a combined likelihood of individual modalities, and used to assess music track similarity. This construction allows for control of the relative weighting of modalities. The general problem of modelling multi-modal similarity for music tracks in particular has been investigated multiple times during the last decade. Weston et al. [23] define a multitask learning problem with joint semantic space of different modalities. This enables comparison of heterogeneous concepts such as audio, tags and artists in a non-probabilistic setting. McFee et al. [24] define a unified embedding space based on kernel matrices of individual modalities. Their algorithm is evaluated on similarities of music artists, and also shows that some music genre structure is revealed by the method. Additionally, it is noted that the audio itself is the weakest of the modalities in terms of artist similarity.

2.1 A Note on Features

To enable meaningful representations of highly abstract concepts in general, the building blocks necessarily need to contain all the relevant information. For this reason, representations of music tracks are often created from lower level

acoustic features including MFCC, chroma and beat structure, believed to capture information related to timbre, melodic content, and tempo respectively. This traditional pipeline procedure has previously proven useful in other music similarity tasks such as cover song identification [25], and this approach is also followed in contribution A.
It should however be noted that while explicit feature extraction enhances transparency of a model, it can obviously only benefit from the included, well defined features. Consequently, as a model designer, one can only do so much with a given set of features. With the resurrection of interest in artificial neural networks, so-called end-to-end systems have become increasing popular [26]. One of the reasons is that the inputs to this type of system are often only slightly pre-processed versions of original measurements. This enables the modelling objective to be taken into account from the very beginning of the modelling process. In principle this type of system should be able to perform task specific feature extraction, and retain only information relevant to the objective, but comes at the cost of lost transparency otherwise offered by explicit feature extraction and selection.

2.2 Multi-Modal Topic Modelling

This section describes a multi-modal version of Latent Dirichlet Allocation (mmLDA) and its relation to other multi-modal topic models. For simplicity, we will use the standard concepts known from the topic modelling literature, such that a *corpus* refers to a collection of *documents*, a *document* refers to an entity composed of an un-ordered set of *words* (often referred to as a bag-of-words representation), a *word* refers to the smallest object of interest in the modelling framework, and finally a *vocabulary* refers to the set of all possible *words*.

Numerous different ways of incorporating multiple data sources have been proposed in the topic modelling literature. In addition to the works mentioned earlier in relation to general multi-modality, we now review a few specific ideas. In the Dirichlet Multinomial Regression Topic Model (DMRTM) Mimno et al. [27] integrate arbitrary document level metadata with the word content of documents. Mimno et al. distinguish between downstream and upstream integration. These terms indicate whether observed metadata variables, according to the generative process of a model, are generated conditional on the latent variables or not. The DMRTM model is able to obtain performance similar to that of special purpose models: the Author-Topic Model [28] and the Topics-Over-Time model (TOT) [29]. A newer idea along the same lines is presented in [30] as the Inverse Regression Topic Model (IRTM). Like DMRTM, the IRTM is proposed

for modelling document level metadata, but lets the metadata directly affect the words of the document in contrast to the DMRTM where words and metadata are conditionally independent given the latent topic variables. While the models work well for annotation purposes where it makes sense to have an asymmetric relationship between modalities, we now turn to the more classical symmetrical multi-modal representation in the context of topic models.
The mmLDA treats all modalities symmetrically and can thus be characterised as a downstream, conditionally independent model of multi-modal documents. The model is quite similar to the GM-LDA model described in [31] and in fact identical to the poly-lingual topic model proposed by Mimno et al. [32], although their focus is explicitly on modelling multilingual text corpora. For modelling multimedia objects with the goal of classification and auto-annotation of TV clips, Putthividhya et al. [33] uses a multi-modal version of un-smoothed LDA, i.e. no prior distribution is specified for the topic-word distributions. Hence contribution A broadens the application areas of the smoothed mmLDA model by modelling collections of more general entities consisting of multiple modalities with assumed correspondence across modalities. Thus we use the mmLDA to obtain document representations in a joint semantic space and investigate properties of the induced similarity.
In mmLDA, correspondence of topics across modalities is assumed by letting the document-topic distribution be shared among modalities. According to [31] this model structure limits its use for image annotation because its mainly models the joint distribution of modalities and does not represent conditional probabilities that well. To improve representation of the relationship between modalities, the Correspondence LDA model is proposed in [31]. This improves modelling of conditional probabilities, but breaks the symmetric representation of modalities. The same issue is mentioned by Virtanen et al. [34], who suggest that successful application of mmLDA is quite dependent on the 1-1 relationship between topics in different modalities implied by the assumptions of shared document-topic distributions. This point of view is supported by the findings in [32], where multiple languages constitute the modalities. The approach suggested in [34] is to provide alternative means of correspondence by allowing topics to be correlated using the main idea of the Correlated Topic Model (CTM) [35]. Additionally, one Hierarchical Dirichlet Process (HDP) [36] per modality provides an on/off switch for each topic's contribution to the document. This effectively allows some topics to be private to a few or even a single modality.
In [37] an alternative way to deal with the strong correspondence across modalities is proposed. Here, all documents are split into their modality specific sub-documents, and a Markov Random Field with edges between modalities of the original documents is used to model the relationships between modalities.

We now proceed with a formal description of the mmLDA. In the multi-modal LDA model, a corpus consist of a D documents, containing data from M dif-

ferent modalities. Each modality has its own vocabulary and hence its own multinomial topic-word distributions parametrised by $\boldsymbol{\Phi}^{(m)}$. Each topic in each modality is assumed to be a multinomial distribution generated from a Dirichlet prior with parameters $\boldsymbol{\beta}^{(m)}$.
A document collection is modelled as being generated by the following steps

- For each modality m
  - For each topic k
    - Draw a multinomial parameter vector $\boldsymbol{\phi}_k^{(m)}$ randomly from a Dirichlet distribution with parameters $\boldsymbol{\beta}^{(m)}$. This represents the k^{th} topic's distribution over words.
- For each document d
  - Draw a multinomial parameter vector $\boldsymbol{\theta}_d$ randomly from a Dirichlet distribution with parameters $\boldsymbol{\alpha}$. This represents the d^{th} document's distribution over topics.
  - For each modality m
    - For each word $w_{d,n}^{(m)}$
      - Draw the topic $z_{d,n}^{(m)}$ from $\text{Cat}(\boldsymbol{\theta}_d)$
      - Draw the word $w_{d,n}^{(m)}$ from $\text{Cat}(\boldsymbol{\phi}_{z_{d,i}^{(m)}}^{(m)})$

The conditional dependencies of mmLDA can be represented as the graphical model shown in Fig. 2.1.

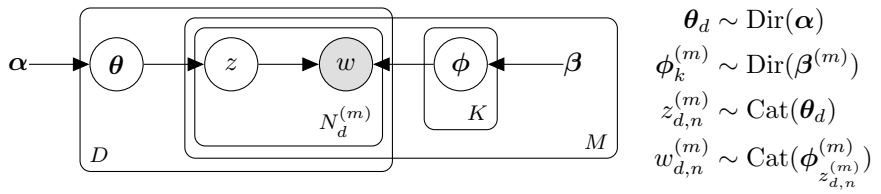


Figure 2.1: Graphical representation of the multi-modal Latent Dirichlet allocation model showing the conditional dependencies in the joint distribution. Each circular node represents a real random variable. The model is represented using plates, describing the presence of multiple instances of the variables shown in the plate. The number in the corner of each plate denotes the number of instances of the variables in the plate. The dark nodes represent variables that are observed.

The joint distribution of the random variables in the model can be written as

$$
\begin{aligned}
&p(\mathbf{w}, \mathbf{z}, \mathbf{\Phi}, \mathbf{\Theta} | \boldsymbol{\alpha}, \boldsymbol{B}) \\
&= p(\mathbf{\Theta}|\boldsymbol{\alpha}) \prod_{m=1}^{M} p(\mathbf{w}^{(m)}|\mathbf{z}^{(m)}, \mathbf{\Phi}^{(m)}) p(\mathbf{z}^{(m)}|\mathbf{\Theta}) p(\mathbf{\Phi}^{(m)}|\boldsymbol{\beta}^{(m)}) \\
&= \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{m=1}^{M} p(\mathbf{\Phi}^{(m)}|\boldsymbol{\beta}^{(m)}) \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{n=1}^{N_d^{(m)}} p(w_{d,n}^{(m)}|z_{d,n}^{(m)}, \boldsymbol{\phi}^{(m)}) p(z_{d,n}^{(m)}|\boldsymbol{\theta}_d) \\
&= \prod_{d=1}^{D} \mathrm{Dir}(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\phi}_k^{(m)}|\boldsymbol{\beta}^{(m)}) \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{n=1}^{N_d^{(m)}} \phi^{(m)}_{z_{d,n}^{(m)}, w_{d,n}^{(m)}} \theta_{d, z_{d,n}^{(m)}} \\
&= \prod_{d=1}^{D} \mathrm{Dir}(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\phi}_k^{(m)}|\boldsymbol{\beta}^{(m)}) \prod_{m=1}^{M} \prod_{j=1}^{V^{(m)}} \prod_{k=1}^{K} (\phi_{k,j}^{(m)})^{c_{m,k,j}} \prod_{d=1}^{D} \prod_{k=1}^{K} (\theta_{d,k})^{v_{d,k}}
\end{aligned}
$$

where $\mathbf{\Theta} = \{\theta_{d,k}\}^{[d=1:D]\times[k=1:K]}$, $\mathbf{\Phi} = \{\boldsymbol{\phi}_k^{(m)}\}^{[k=1:K]\times[m=1:M]}$, $V^{(m)}$ is the size of the vocabulary of modality m, and $\boldsymbol{\phi}_k^{(m)} = \{\phi_{k,j}^{(m)}\}^{j=1:V^{(m)}}$.
As exact inference in this model is computationally intractable, the parameter distributions are approximated using Gibbs sampling. Because the Categorical parameter variables $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are conveniently assumed to be Dirichlet distributed, it is relatively straightforward to integrate them out. This results in what is often referred to as a collapsed likelihood from which the equations defining a collapsed Gibbs sampler for the model can be derived.

$$
p(z_{d,n}^{(m)} = t | w_{d,n}^{(m)} = b, \mathbf{z}^{\setminus mdn}, \mathbf{w}^{\setminus mdn}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{(c_{m,t,b}^{\setminus mdn} + \beta_b)(v_{d,t}^{\setminus mdn} + \alpha_t)}{\sum_{j=1}^{V}(c_{m,t,j}^{\setminus mdn} + \beta_j)} \tag{2.1}
$$

where $c_{m,t,j}$ is the number of times the word j in mode m has been assigned to topic t, and $v_{d,t}$ is the number of words in document d that has been assigned to topic t (across all modalities). The superscript $^{\setminus mdn}$ means that the topic assignment currently being sampled (at position m, d, n) is not included.

2.2.1 Hyper Parameter Optimisation

The parameters of the Dirichlet priors on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, are in a classical Bayesian setting fixed at some values considered to represent reasonable prior knowledge of the model parameters. However, as suggested in [38, 39], when dealing with large data sets, the hyper-parameters in question are very well-defined, and resorting to direct optimisation of the likelihood is an acceptable alternative to the more Bayesian approach of hyper-priors. Wallach et al.

[39] recommend the use of an asymmetrical Dirichlet distribution as prior for $\boldsymbol{\theta}$ and a symmetrical Dirichlet distribution as prior for $\boldsymbol{\phi}$. This allows some topics to be more prevalent than others, and often results in a topics being dominated by ubiquitous words[1]. In contribution A we follow the described suggestions of regarding choice of prior, and optimise the hyper-parameters using fixed point iterations due to [40]. For alternative iterative methods we refer to [38].

2.2.2 Model Estimation

In general probabilistic modelling, multiple approaches to approximate parameter inference exist. Numerous different inference procedures have been applied to the plain LDA model including Variational Bayes (VB) [20], expectation propagation (EP) [41], a collapsed version of VB (CVB)[42], various combinations of (C)VB and (GS/CGS) [43, 44, 45], online VB using mini-batches [46], Sequential Monte Carlo [47], Method of Moments (MOM) [48], Particle Mirror Descent [49] and collapsed Gibbs sampling (CGS) [50]. In Gibbs Sampling (GS), which is a variant of Markov Chain Monte Carlo (MCMC), each variable in the model is sampled from its conditional distribution of the variable given the values of all other variables. If the model definition allows (as it does in LDA because of conjugacy), some variables can be integrated out, and application of GS in this context is referred to as Collapsed Gibbs Sampling.
The results reported in contribution A are produced using CGS.

One of the general drawbacks of using MCMC methods is the often quite lengthy sampling time. The following works suggest different ideas to reducing the sampling time used for model estimation in LDA in the specific case of CGS.
In [51] it is proposed to decrease the sampling time by only sampling part of the words in the training data. In [52], Porteous et al. improve performance by adaptively approximating the normalisation constant for the sampling distribution. By exploiting structure in the sampling equations and updating the count variables more effectively than standard collapsed Gibbs sampling, Yao et al. [53] observe significant speed gains over the naïve Gibbs sampling implementation. In short, it is proposed to split the sampling probability mass into three "buckets", where the size of each bucket is depending on different counts of assignments of words to topics. Sorting of the buckets and using a 32-bit encoding of the word-topic counts are the main reasons for the speed-up. The benefits of the last two methods become more distinct when the number of topics is increased because sparsity in the topic-word counts is exploited. To be able to produce the results is contribution A in a reasonable amount of time,

[1] In the text modelling domain, these words are often referred to as stop-words

we applied the sparse sampling method presented in [53]. Further improvement might be possible by following the ideas in [54], where GS is used in a setting where the non-zero probability of a variable being re-sampled is varied in an online fashion.

2.2.3 Maximum A Posteriori Estimation

For some practical applications it can be convenient to work with point estimates of model parameters instead full distributions e.g. represented by posterior samples. In contribution A, we obtained approximate maximum a posteriori (MAP) point estimates of parameters. Such a MAP estimate was obtained by, after 1950 Gibbs sampling iterations, choosing the sample with the highest marginal likelihood among the next 50 iterations.
An alternative approach could have been to consider the use of Iterated Conditional Modes (ICM) to generate the MAP estimate. ICM has previously proven successful in similar applications in topic modelling using LDA [55] and NMF [56]. In ICM, the stochastic sampling step of the Gibbs sampler is replace by a deterministic sampling step; choosing the maximum of the conditional distribution to be sampled. For the LDA model, the relevant conditional distribution is a Categorical distribution, which is readily available in an un-normalised form defined by the counts of assignments of topics to words. Hence the maximum can be calculated extremely efficiently. The ICM algorithm was not used for production of the results in [3], but was implemented in a later version of the mmLDA++ software[2].

2.2.4 Multi-Modal Similarity Revisited

One way to analyse the latent representation induced by a model, and thereby what similarity calculations are based on, is to measure the alignment of the presence of particular words to the topic representation. The alignment between a specific word λ in a modality not included in the model estimation process and the grouping defined by the model in terms of topics, is measured using the average Normalised Mutual Information across all documents (avgNMI). Specifically, we let $p(\omega_\lambda = 0|z = k) = 1 - p(\omega_\lambda = 1|z = k) = 1 - \phi^*_{k,\lambda}$ where the asterisk denotes a parameter point estimate, and ω_λ is an indicator variable for the occurrence of λ. For a specific document d, the mutual information between

[2]The mmLDA++ software is available for download at http://people.compute.dtu.dk/rast/

the presence of λ and the topic variable z is then

$$\begin{aligned}\mathrm{MI}(\omega_\lambda, z|d) &= \mathrm{KL}\left(p\left(\omega_\lambda, z|d\right) \| p\left(\omega_\lambda|d\right) p(z|d)\right) \\ &= \sum_{i\in\{0,1\}} \sum_{k=1}^{K} \left(p\left(\omega_\lambda = i|z=k\right) p(z=k|d) \log \frac{p\left(\omega_\lambda = i|z=k\right)}{\sum_{k'=1}^{K} p\left(\omega_\lambda = i|z=k'\right) p(z=k'|d)} \right)\end{aligned}$$

It is of course possible to assess the uncertainty of the avgNMI for λ by repetition of the calculation for multiple samples from the Gibbs chain. For the results presented in contribution A, a single sample considered the MAP estimate was used.

In contribution A we suggested to evaluate the latent grouping defined by a model estimated using only audio features and lyrics for the set of training songs. This was done by measuring the alignment, as measured by avgNMI, between the model and the user provided tags, which was considered the single available modality best describing expressed human perception of the music. The results showed that the user provided tags best aligned with the model representation were mainly names of music genres. This means that the mmLDA model is able find structure in just the audio and the lyrics that to some degree aligns with highly abstract concepts often used to communicate music taste among humans.

To illustrate an application of the music track representation obtained from the mmLDA model in contribution A, the classical task of genre-classification was performed using both tags, lyrics, and audio modalities from the Million Song Data set (MSD) [57, 58]. The MSD is benchmark data set of songs which includes audio features from the.echonest.com API and metadata regarding artists, lyrics, tags and user-song play counts. Because of its comprehensiveness it has been used to evaluate numerous MIR tasks such as a multi-modal approach to artist identification [59], and artist, genre and key recognition using convolutional neural networks [60].

Music genre classification is a popular problem in the MIR community [61], and the fact that music genres are abstract and subjective makes the classification task interesting but also a very hard. Ambiguities arise from the generality induced by high level abstractions, and the need for consideration of both the music and the listeners seems obvious in order to successfully predict music genre.

2.3 Measuring Similarity by Correlations

One issue with intractable latent variable models is that parameter estimates obtained via approximate inference techniques are likely to represent local minima of the objective function. This necessarily affects the measure of similarity implied by the model parameters. Furthermore, how to actually calculate similarities using parameter point estimates can be addressed in various ways. These questions, stated in the context of the mmLDA model, are the main drivers for the work described in contribution B. Several ways to measure similarity between two documents A and B in the LDA model have been proposed, of which most only consider the related document-topic distributions $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$. For an image retrieval task, Hörster et al. [62] compare cosine distance (2.2), ℓ_1-norm (2.3), Jensen-Shannon divergence (2.4) and a measure based on the likelihood of the topic distribution of one document given contents of the other document (2.6). The distances/similarities are evaluated by human scoring of retrieval results, and in that particular study, the likelihood based measure is preferred in terms of quality. This seems sensible as none of the competing measures consider the estimated topic-word distributions $\boldsymbol{\Phi}$. With the purpose of visualising topic models, Chaney et al. [63] propose to use (2.5) to describe document similarity.

$$\cos(A,B) = \frac{\boldsymbol{\theta}_A^\top \boldsymbol{\theta}_B}{\|\boldsymbol{\theta}_A\| \|\boldsymbol{\theta}_B\|} \tag{2.2}$$

$$\ell_1(A,B) = \sum_{k=1}^{K} \left|\boldsymbol{\theta}_{Ak} - \boldsymbol{\theta}_{Bk}\right| \tag{2.3}$$

$$\mathrm{JS}(A,B) = \frac{1}{2}\left(\mathrm{KL}\left(\boldsymbol{\theta}_A, \frac{\boldsymbol{\theta}_A + \boldsymbol{\theta}_B}{2}\right) + \mathrm{KL}\left(\boldsymbol{\theta}_B, \frac{\boldsymbol{\theta}_A + \boldsymbol{\theta}_B}{2}\right)\right) \tag{2.4}$$

$$\mathrm{CB}(A,B) = \sum_{k=1}^{K} \left|\log(\boldsymbol{\theta}_{Ak}) - \log(\boldsymbol{\theta}_{Bk})\right| \tag{2.5}$$

$$\mathrm{LL}(A,B) = \frac{\log p(\boldsymbol{w}_A | \boldsymbol{\theta}_B^s, \boldsymbol{\Phi}^s)}{\sum_{m=1}^{M} N_A^{(m)}} \tag{2.6}$$

To assess the stability of defined document similarities across multiple point estimates the correlation between similarity matrices can be calculated. Contribution B applies Spearman's rank correlation r_s, as the preservation of rank-relations seems sensible and consideration of only linear correlation might be too restrictive. Figure 2.2 illustrates the stability of the different similarity and distance measures listed above. It is clear that the likelihood is superior to the

others. This emphasises the importance of including the topic-word distributions in the distance measure.

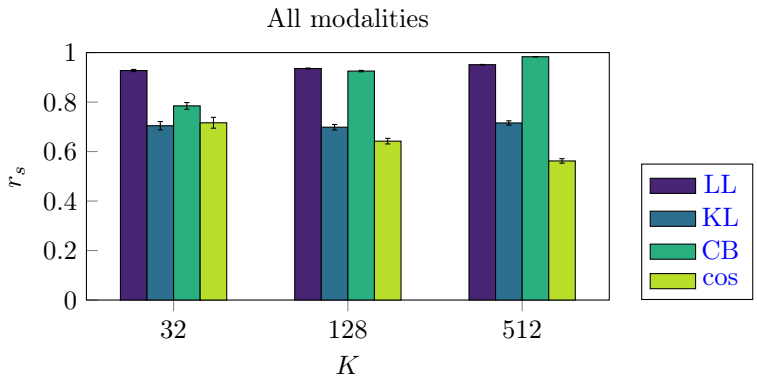


Figure 2.2: Mean and standard deviations of Spearman correlations between all pairs of similarity matrices resulting from 5 random re-initialisations, using 10-fold cross validation. Models are estimated from all modalities considered in contribution B. The differences in correlations suggest that the choice of similarity method is an important issue, and that the likelihood based method as expected seems superior.

The modalities considered in the models of contribution B are listed below and represent both very low level features of the audio, a textual dimension of the tracks via the lyrics, and information provided by music consumers. Together they provide information derived directly from the music tracks, and information related to the tracks through cognitive and sociological processes [64].

- Open vocabulary tags provided by the users of the service last.fm.
- Track lyrics provided by musiXmatch.com.
- Editorial artist tags provided by allmusic.com.
- Artist tags provided by the MusicBrainz project.
- User listening history provided by the.echonest.com
- Genre and style tags provided by allmusic.com
- Various audio features from the.echonest.com

The given audio features are continuous and to include them in the mmLDA model, some kind of discretisation is needed to obtain an *audio word* representation [65, 66, 67, 68]. The presented results rely on k-means clustering of timbre, chroma, loudness and tempo features obtained via a now-closed audio analyzer API provided by the.echonest.com.

In contribution B, also correlations between similarity matrices based on different modalities are calculated. This should be seen as yet another way to measure alignments between the similarities defined by different modalities. Our paper states that there is a significant positive correlation between the similarities defined by the audio modality alone and the similarities defined by the remaining non-audio modalities. The relevance of this result is however questionable as the correlation is very close to 0.

Chapter 3

Time Series Modelling and Classification with Hidden Markov Models and Method of Moments

While the topic model in the previous section used a bag-of-words representation of documents, we now turn to modelling of data that exhibit sequential structure. This discipline is therefore often referred to as time series modelling and explicitly encode system dynamics in the model parameters. Problems where the ordering of observations is important to the end goal have numerous times been approached with models possessing the ability to describe progression. This includes general control theory, human activity recognition, chord progression in polyphonic music, and weather forecasting just to mention a few. Time series models seek to represent abstract entities by estimating parameters related to dynamical properties of the data. Hence time series modelling can be a valuable tool for comparing and grouping sequential data. Recently, recurrent artificial neural network architectures such as the LSTM [69] have become increasingly popular and are widely used in e.g. speech recognition [70, 71] and natural language processing [72, 73]. We will however limit this discussion to one of the more classical probabilistic approaches to sequence modelling, the Hidden Markov model, as this is the main topic of contributions C and D.

3.1 The HMM and its Estimation by Method of Moments

In the Hidden Markov Model, observed data is assumed to be generated conditional on an underlying sequence of unobserved discrete *state variables* denoted $z_n \in [1:S]$, $n \in [1:N]$,, and that there is a dependency structure between these latent state variables at the different time-steps. The term *Markov* is due to a restriction in this dependency structure, such that given the value of a particular state variable, the state variable at the next time step is conditionally independent of all previous state variable and observations:

$$p(z_{n+1}|\boldsymbol{x}_n, \boldsymbol{z}_n) = p(z_{n+1}|z_n, z_{n-1}, \ldots, z_1) = p(z_{n+1}|z_n) \tag{3.1}$$

The dependency structure of observed variables are as follows: Given the current state, the current observation is conditionally independent of all previous state variables and observations:

$$p(x_n|\boldsymbol{x}_{n-1}, \boldsymbol{z}_n) = p(x_n|z_n) \tag{3.2}$$

This process can be illustrated using a graphical model such as the one shown in Fig. 3.1

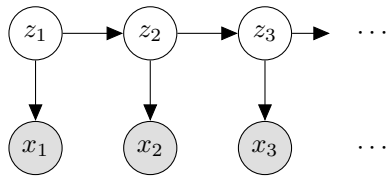


Figure 3.1: Graphical representation of the Hidden Markov Model

The dynamic behaviour of the model is expressed through a transition probability matrix $\boldsymbol{T} \in \mathbb{R}^{S\times S}$ where the i^{th} column represents a categorical distribution describing the probability of transitioning from state i to any of the S states, see (3.1). To start the process, an initial distribution over states usually denoted $\boldsymbol{\pi}^{(1)} \in \mathbb{R}^S$ is defined. For simplicity we limit this presentation to the discrete HMM, such that the probabilities of observations can be represented by a stochastic matrix $\boldsymbol{O} \in \mathbb{R}^{K\times S}$. The i^{th} column of $\boldsymbol{O}$ then fully describes the probability of observing any of the K symbols if the current state is i, see. (3.2).

A common way to estimate the parameters in the HMM is using the maximum likelihood principle either via direct optimisation using standard constraint optimisation tools [74] or the more common choice in the field of machine learning, the Expectation-Maximisation (EM) algorithm (also known as Baum-Welch in the specific case of HMM) [75]. Obviously, it is also possible to take a more Bayesian approach and provide prior information for the model parameters. This line of thought has been explored by Goldwater et al. [76] for Part-of-Speech tagging. Using ideas from Approximate Bayesian Computation (ABC) and method of moments for HMM estimation, which is also trated in this thesis, Bonnevie et al. [77] suggest a method for fast sampling of posterior HMM parameters in a large data setting.
Recently, another estimation method has gained attention specifically within latent variable modelling, hence including (ad-)mixture models and Hidden Markov Models. The general idea of the method is to relate model parameters to empirical data moments through a set of equations, which is why it is often termed the Method of Moments (MOM). The original idea is quite old, dating back to Karl Pearson [78].

Traditionally, application of MOM depended on estimates of moments of the same order as the number of parameters to be estimated. For more complex models, this limited the applicability of the method, because of the increased uncertainty of higher order moment estimates. The recent interest in the field seems to be driven by convergence proofs involving spectral decomposition of quantities related to moments of relatively low order [79]. Specifically for the HMM, the common starting point seems to be an observable operator formulation due to Jaeger et al. [80]. Later Hsu et al. [81] show how to calculate likelihoods under the HMM without recovering the traditional parametrisation presented above. An algorithm for recovery of the parameters $\boldsymbol{T}$ and $\boldsymbol{O}$ is subsequently given in [10]. Both methods rely on spectral decomposition of quantities calculated from empirical third order moments subject to the mild conditions of $\text{rank}(\boldsymbol{O}) = \text{rank}(\boldsymbol{T}) = S$. The moments considered in this context are expectations of the first three consecutive observations of sequences (3.3).

$$P_{1,2,3} = \mathbb{E}\left[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3\right] \qquad \boldsymbol{x}_i \in \mathbb{R}^K \tag{3.3}$$

$$P_{1,3,2}(\boldsymbol{\eta}) = \mathbb{E}[(\boldsymbol{x}_1 \otimes \boldsymbol{x}_3)\langle \boldsymbol{\eta}, \boldsymbol{x}_2 \rangle] \qquad \boldsymbol{\eta} \in \mathbb{R}^K \tag{3.4}$$

While the above-mentioned works apply spectral decompositions to expressions involving matrix "slices" of (3.3), defined in (3.4), also methods directly applying orthogonal tensor decompositions have been proposed and seems to dominate the field [82]. Very recent work proves learnability of over-complete models [83] which was not covered by previous methods.

Some variations and extensions of the results for basic HMMs have been inves-

tigated in [84] (reduced-rank HMM), [85] (mixture of HMMs), and [86] (contrastive learning of HMMs). For the specific estimation algorithms and their convergence properties we refer to the original works [81, 10]. One apparent limitation of cited methods is the restriction to use only the first three observations in a number of i.i.d. HMM data sequences. The main reason for this limitation seems to be avoidance of unnecessarily complicated convergence proofs also considering correlation between triplets in the sequences. It is however not obvious if the framework is valid for non-stationary sequences of lengths greater than three. This matter is handled in contribution C where it is shown that the algorithms remain valid when using expectations of moments across all time steps. This shows not only to be true for assumed stationary HMMs, but also for ergodic non-stationary HMMs.

3.2 Sequence Similarity via HMM Classification

While the previous section dealt with estimation of model parameters in the Hidden Markov Model, this section turns to an application leveraging the obtained abstract representation of the dynamics in observed sequence data. The problem of calculating similarities between time series and hence also represent groups of data sequences exhibiting similar intra-group dynamics has often been treated with the use of HMMs. Grouping of data can often be divided into the two main areas; clustering and classification. In a classification problem a set predefined classes exist, the goal is to be able to classify unseen data correctly. In classical model based classification this translates into finding good representatives of the given classes. This is typically done by estimating class conditional models using exemplar data for which the class label is known. In clustering, we also assume the existence of groups (clusters) of data, but in contrast to classification, there are no predefined labels. This adds to the complexity of clustering by also requiring discovery of a partitioning of data. Early work in sequence clustering using HMMs includes [87, 88, 89] which has laid ground for later work formulating embedding procedures for sequence data [90]. More recently, spectral clustering using probability product kernels (PPK) for HMMs was explored in [91]. In the classification setting, some of the early work was done by Lawrence Rabiner [75] with applications to speech recognition. Additionally, HMM based classification has be applied in diverse fields such as software virus detection [92], 2D object shape classification [93] and symbolic folk music classification [94].

In a model based classification setting where each class prototype is an HMM, data sequences can be assigned to classes based on how well they represent the particular class according to a specific cost function. Due to the probabilistic

nature of the model, the cost function is often chosen to be the negative log-likelihood of the model parameters given the data. However, inspired by the advances in parameter estimation using MOM, contribution D proposes to use a cost function based on third order moments and triplets of observations. The presented idea is inspired by the use of third order moments as a kind of sufficient statistics for HMM estimation in the MOM framework. In contribution D, we show how the Kullback-Leibler divergence can be used to derive a cost function relating model parameters to an observed data sequence. This is done by representing the HMM in terms of its third order moments at every time step:

$$P_{n,n+1,n+2}(\cdot,k,\cdot) = \boldsymbol{O}\,\mathrm{diag}(\boldsymbol{T}^{n-1}\boldsymbol{\pi}^{(1)})\boldsymbol{T}^{\top}\,\mathrm{diag}(\boldsymbol{O}(k,\cdot))\boldsymbol{T}^{\top}\boldsymbol{O}^{\top} \qquad k \in [1;K] \tag{3.5}$$

Further analysis reveals that the proposed cost constitutes a composite likelihood of the HMM re-parametrised in the form of moments. A composite likelihood is a product of likelihoods due to sub-components of the data [95], and generalises the term pseudo-likelihood, specifically used for conditional component likelihoods [96], to also include marginal component likelihoods.
The proposed distance score (3.6) collects the elements of the third order moments based on model parameters $\mathcal{M} = \{\boldsymbol{\pi}^{(1)}, \boldsymbol{T}, \boldsymbol{O}\}$ (3.5) that correspond to the observed triples in the sequence $\boldsymbol{x}$ at each time step.

$$\mathcal{D}(\boldsymbol{x},\mathcal{M}) = \frac{1}{N-2}\sum_{n=1}^{N-2} -\log\left(P_{n,n+1,n+2}(x^{(n)},x^{(n+1)},x^{(n+2)})\right) \tag{3.6}$$

Using the described composite likelihood method for classification of time series data, we are able to obtain results equal to the likelihood based method in terms of quality, but at a reduced computational complexity of $\mathcal{O}(N)$ per class compared to $\mathcal{O}(NS^2)$ for the classical method.
While the described approach seems to perform well, the memory requirements of a naïve implementation grows with the lengths of the sequences to be classified. This can be handled by assuming stationarity after a certain number of time steps. The approximation is controlled by specifying a maximum allowed total variation distance between the initial state distribution and the stationary state distribution, which can be converted into an upper bound on the convergence time to approximate stationarity.

Chapter 4

Summary & Conclusions

With the overall aim to investigate possible representations of abstract concepts in a way that can assist humans by improving search experience and facilitating navigation of large scale multimedia content databases, this thesis treated two latent variable models.

The problem of representation of multi-modal concepts was treated by integration of the diverse information of different modalities using a multi-modal version of the Latent Dirichlet Allocation topic model. This resulted in concrete approximate representations of abstract concepts. The model was applied to represent music tracks composed of song lyrics, audio content and user provided tags. The obtained representation from using only the lyrics and audio modalities was evaluated by measuring the alignment with the tag modality. Finding that the best aligned tags are dominated by common western music genres and styles suggests that the latent representation, defined in an unsupervised manner, may to some degree be an acceptable proxy for certain "cognitive" aspects of music.
Applied to classification tasks of music genre and style, as expected, representations based on all three modalities performed best. Combining audio features and lyrics improved classification results compared to using either of the two.
The question of measuring similarity between multi-modal concepts was further investigated in contribution B, where different ways of defining similarity specif-

ically in topic models were evaluated.
Returning to the issues of modern information retrieval, the presented results indicate the viability of the mmLDA model as a means to assess similarity between abstract concepts comprising heterogeneous data types. This is done by integration of the different modalities into unified computational representations. These representations potentially improves peoples' experiences of music navigation and retrieval.

The problem representing data exhibiting sequential structure was treated using the well known Hidden Markov model. Two aspects regarding such systems were investigated, namely parameter estimation and classification by measuring similarity.
The first result described in detail in contribution C regarded improving the practical applications of a recently proposed parameter estimation method based on third order moments. While not limited to the case of a stationary HMM, the main finding was a proof of validity of using empirical expectations of third order moments across time steps.
The second result was related to the practical application of computational systems for classification of sequence data. The main idea of contribution D was an alternative distance score for describing relationships between models and sequences. The proposed distance score was a composite likelihood of a sequence of third order moments derived from the HMM. While retaining classification performance the proposed method showed less costly than exact likelihood calculations in terms of computational complexity.
This result has the potential to expand certain application areas of HMM based classification by increasing computational feasibility for large collections of data to be classified.

Appendix A

Towards a Universal Representation for Audio Information Retrieval and Analysis

---

Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen. "Towards a universal representation for audio information retrieval and analysis". In: *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2013. DOI: 10.1109/ICASSP.2013.6638242

TOWARDS A UNIVERSAL REPRESENTATION FOR AUDIO INFORMATION RETRIEVAL AND ANALYSIS

Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen

DTU Compute
Technical University of Denmark
Asmussens Allé B305, 2800 Kgs. Lyngby, Denmark
{bjje,rast,janla,lkai}@dtu.dk

ABSTRACT

A fundamental and general representation of audio and music which integrates multi-modal data sources is important for both application and basic research purposes. In this paper we address this challenge by proposing a multi-modal version of the Latent Dirichlet Allocation model which provides a joint latent representation. We evaluate this representation on the Million Song Dataset by integrating three fundamentally different modalities, namely tags, lyrics, and audio features. We show how the resulting representation is aligned with common 'cognitive' variables such as tags, and provide some evidence for the common assumption that genres form an acceptable categorization when evaluating latent representations of music. We furthermore quantify the model by its predictive performance in terms of genre and style, providing benchmark results for the Million Song Dataset.

Index Terms— Audio representation, multi-modal LDA, Million Song Dataset, genre classification.

1. INTRODUCTION

Music representation and information retrieval are issues of great theoretical and practical importance. The theoretical interest relates in part to the close interplay between audio, human cognition and sociality, leading to heterogenous and highly multi-modal representations in music. The practical importance, on the other hand, is evident as current music business models suffer from the lack of efficient and user friendly navigation tools. We are interested in representations that directly support interactivity, thus representations based on latent variables that are well-aligned with cognitively (semantic) relevant variables [1]. User generated tags can be seen as such 'cognitive variables' since they represent decisions that express reflections on music content and context. Clearly, such tags are often extremely heterogenous, high-dimensional, and idiosyncratic as they may relate to any aspect of music use and understanding.

Moving towards broadly applicable and cognitively relevant representations of music data is clearly contingent on the ability to handle multi-modality. This is reflected in current music information research that use a large variety of representations and models, ranging from support vector machine (SVM) genre classifiers [2]; custom latent variable models models for tagging [3]; similarity based methods for recommendation based on Gaussian Mixture models [4]; and latent variable models for hybrid recommendation [5]. A significant step in the direction of flexible multi-modal representations was taken in the work of Law *et al.* [6] based on the probabilistic framework of Latent Dirichlet Allocation (LDA) topic modeling. Their topic model representation of tags allows capturing rich cognitive semantics as users are able to tag freely without being constrained by a fixed vocabulary. However, with a strong focus on automatic tagging Law *et al.* refrained from developing a universal representation - symmetric with respect to all modalities. A more symmetric representation is pursued in recent work by Weston *et al.* [7]; however, without a formal statistical framework it offers less flexibility, e.g., in relation to handling missing features or modalities. This is often a challenge encountered in real world music applications.

In this work we pursue a multi-modal view towards a unifying representation, focusing on latent representations informed symmetrically by all modalities based on a multi-modal version of the Latent Dirichlet Allocation model. In order to quantify the approach, we evaluate the model and representation in a large-scale setting using the million song dataset (MSD) [8], and consider a number of models trained on combinations of the three basic modalities: user tags (top-down view), lyrics (meta-data view) and content based audio features (bottom-up view). First, we show that the latent representation obtained by considering the audio and lyrics modalities is well aligned—in an unsupervised manner - with 'cognitive' variables by analyzing the mutual information

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. Bob L. Sturm, Aalborg University Copenhagen is acknowledged for suggestion of relevant references in music interpretation.

between the user generated tags and the representation itself. Secondly, with knowledge obtained in the first step, we evaluate auxiliary predictive tasks to demonstrate the predictive alignment of the latent representation with well-known human categories and metadata information. In particular we consider genre and styles provided by [9], none of which is used to learn the latent semantics themselves. This leads to benchmark results on the MSD and provides insight into the nature of generative genre and style classifiers.

Our work is related to a rich body of studies in music modeling, and multi-modal integration. In terms of non-probabilistic approaches this includes the already mentioned work of Weston *et al.* [7]. McFee *et al.* [10] showed how hypergraphs (see also [11]) can be used to combine multiple modalities with the possibilities to learn the importance of each modality for a particular task. Recently McVicar *et al.* [12] applied multi-way CCA to analyze emotional aspects of music based on the MSD.

In the topic modelling domain, Arenas-García *et al.* [13] proposed multi-modal PLSA as a way to integrate multiple descriptors of similarity such as genre and low-level audio features. Yoshii *et al.* [5, 14] suggested a similar approach for hybrid music recommendation integrating subject taste and timbre features. In [15], standard LDA was applied with audio words for the task of obtaining low-dimensional features (topic distributions) applied in a discriminative SVM classifier. For the particular task of genre classification *et al.* [16] applied the pLSA model as a generative genre classifier. Our work is a generalization and extension of these previous ideas and contributions based on the multi-modal LDA, multiple audio features, audio words and a generative classification view.

2. DATA & REPRESENTATION

The recently published million song dataset (MSD) [8] has highlighted some of the challenges in modern music information retrieval; and made it possible to evaluate top-down and bottom-up integration of data sources on a large scale. Hence, we naturally use the MSD and associated data sets to evaluate the merits of our approach. In defining the latent semantic representation, we integrate the following modalities/data sources.

The tags, or top-down features, are human annotations from `last.fm` often conveying information about genre and year of release. Since users have consciously annotated the music in an open vocabulary, such tags are considered an expressed view of the users cognitive representation. The metadata level, i.e., the lyrics, is of course nonexistent for for majority of certain genres, and in other cases simply missing for individual songs which is not a problem for the proposed model. The lyrics are represented in a *bag-of-words* style, i.e., no information about the order in which the terms occurs is included. The content based or bottom up features are derived from the audio itself. We rely on the Echonest feature extraction[1] already available in for the MSD, namely timbre, chroma, loudness, and tempo. These are orginally derived in event related segments, but we follow previous work [17] by beat aligning all features obtaining an meaningful alignment with music related aspects.

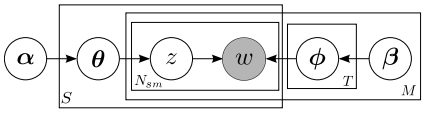


Fig. 1: Graphical model of the multi-modal LDA model

In order to allow for practical and efficient indexing and representation, we abandon the classic representation of using for example a Gaussian mixture model for representing each song in its respective feature space. Instead we turn to the so-called *audio word* approach (see e.g. [18, 19, 3, 17]) where each song is represented by a vector of counts of (finite) number of audio words (feature vector). We obtain these *audio words* by running a randomly initiated K-means algorithm on a 5% random subset of the MSD for timbre, chroma, loudness and tempo with 1024, 1024, 32, and 32 clusters, respectively. All beat segments in a all songs are then quantized into these audio words and the resulting counts, representing the four different audio features, are concatenated to yield the audio modality.

3. MULTI-MODAL MODEL

In order to model the heterogeneous modalities outline above, we turn to the framework of topic modeling. We propose to use a multi-modal modification of the standard LDA to represent the latent representation in a symmetric way relevant to many music applications. The multi-modal LDA, mmLDA, [20] is a straight forward extension of standard LDA topic model [21], as shown in Fig. 1. The model and notation is easily understood by the way it generates a new song by the different modalities, thus the following generative process defines the model:

- For each topic $z \in [1;T]$ in each modality $m \in [1;M]$
  Draw $\boldsymbol{\phi}_z^{(m)} \sim$ *Dirichlet*$(\boldsymbol{\beta}^{(m)})$.
  This is the parameters of the z^{th} topic's distribution over vocabulary $[1;V^{(m)}]$ of modality m.
- For each song $s \in [1;S]$
  - Draw $\boldsymbol{\theta}_s \sim$ *Dirichlet*$(\boldsymbol{\alpha})$.
    This is the parameters of the s^{th} song's distribution over topics $[1;T]$.
  - For each modality $m \in [1;M]$
    - For each word $w \in [1;N_{sm}]$
      - Draw a specific topic $z^{(m)} \sim$ *Categorical*$(\boldsymbol{\theta}_s)$
      - Draw a word $w^{(m)} \sim$ *Categorical*$(\boldsymbol{\phi}_{z^{(m)}}^{(m)})$

[1] `http://the.echonest.com`

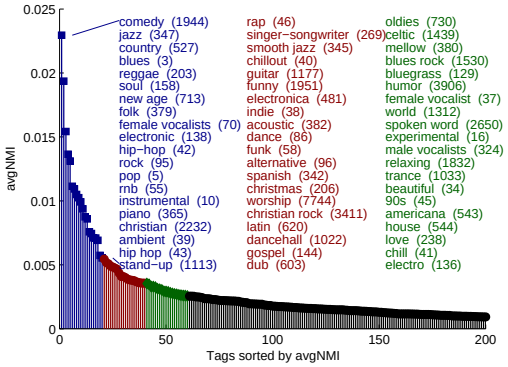


Fig. 2: Normalized average mutual information (avgNMI) between the latent representation defined by audio and lyrics for $T = 128$ topics and the 200 top-ranked tags. avgNMI is computed on the test set in each fold. The popularity of each tag is indicated in parenthesis.

A main characteristic of mmLDA is the common topic proportions for all M modalities in each song, s, and separate word-topic distributions $p(w^{(m)}|z)$ for each modality, where z denotes a particular topic. Thus, each modality has its own definition of what a topic is in terms of its own vocabulary.

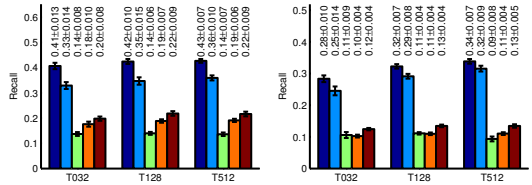
Model inference is performed using a collapsed Gibbs sampler [22] similar to the standard LDA. The Gibbs sampler is run for a limited number of complete sweeps through the training songs, and the model state with the highest model evidence within the last 50 iterations is regarded as the MAP estimate. From this MAP sample, point estimates of the topic-song distribution, $\hat{p}(z|s)$, and the modality, m, specific word-topic distribution, $\hat{p}(w^{(m)}|z)$, can be computed based on the expectations of the corresponding Dirichlet distributions.

Evaluation of model performance on a unknown test song, s^*, is performed using the procedure of fold-in [23, 24] by computing the point estimate of the topic distribution, $\hat{p}(z|s^*)$ for the new song, by keeping the all the word-topic counts fixed during a number of new Gibbs sweeps. Testing on a modality, not included in the training phase, requires a point estimate of the word-topic distribution, $p(w^{(m^*)}|z)$, of the held out modality, m^*, of the training data. This is obtained by fixing the song-topic counts while updating the word-topic counts for that specific modality. This is similar to the fold-in procedure used for test songs.

4. EXPERIMENTAL RESULTS & DISCUSSION

4.1. Alignment

The first aim is to evaluate the latent representation's alignment with a human 'cognitive' variable, which we previously argued could be the open vocabulary tags. We do this by including only the lower level modalities of audio and lyrics



(a) Genre (b) Style

Fig. 3: Classification accuracy for $T \in \{32, 128, 512\}$. Dark blue: Combined model; Light Blue: Tags; Green: Lyrics; Orange: Audio; Red: Audio+Lyrics.

when estimating the model. Then the normalized mutual information between a single tag and the latent representations, i.e., the topics, is calculated for all the tags.

Thus for a single tag, ${w_i}^{(tag)}$ we can compute the mutual information between the tag and the topic distribution for a specific song, s as:

$$\mathrm{MI}\left({w_i}^{(tag)}, z|s\right) = \mathrm{KL}\left(\hat{p}\left({w_i}^{(tag)}, z|s\right) || \hat{p}\left({w_i}^{(tag)}|s\right)\hat{p}\left(z|s\right)\right), \quad (1)$$

where $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence. We normalize the MI to be in $[0; 1]$, i.e,

$$\mathrm{NMI}\left({w_i}^{(tag)}, z|s\right) = 2\frac{\mathrm{MI}\left({w_i}^{(tag)}, z|s\right)}{H\left({w_i}^{(tag)}|s\right) + H\left(z|s\right)},$$

where $H(\cdot)$ denotes the entropy. Finally, we compute the average over all songs to arrive at the final measure of alignment for a specific tag, given by $\mathrm{avgNMI}({w_i}^{(tag)}) = \frac{1}{S}\sum_s \mathrm{NMI}\left({w_i}^{(tag)}, z|s\right)$.

Fig. 2 shows a sorted list of tags, where tags with high alignment with the latent representation have higher average NMI (avgNMI). It is notable that the combination of the audio and lyrics modality, in defining the latent representation, seems to align well with genre-like and style-like tags. On the contrary, emotional and period tags are relatively less aligned with the representation. Also note that the alignment is not simply a matter of the tag being the most popular as can be seen from Fig. 2. Less popular tags are ranked higher by avgNMI than very popular tags, suggesting that some are more specialized in terms of the latent representation than others.

The result gives merit to the idea of using genre and styles as proxy for evaluating latent representation in comparison with other open vocabulary tags, since we - from lower level features, such as audio features and lyrics - can find latent representations which align well with high-level, 'cognitive' aspects in an unsupervised way. This is in line with many studies in music informatics on western music (see e.g. [25, 26, 27]) which indicate coherence between genre and tag categories and cognitive understanding of music structure. In

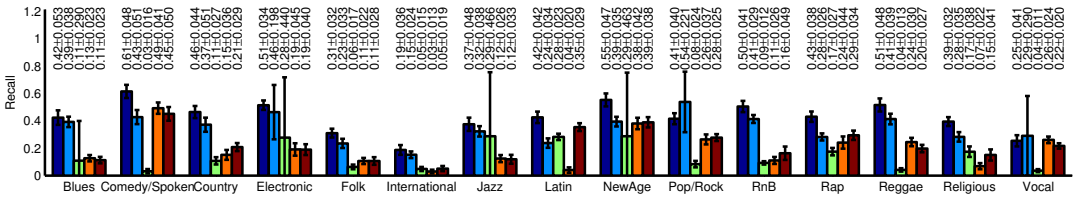
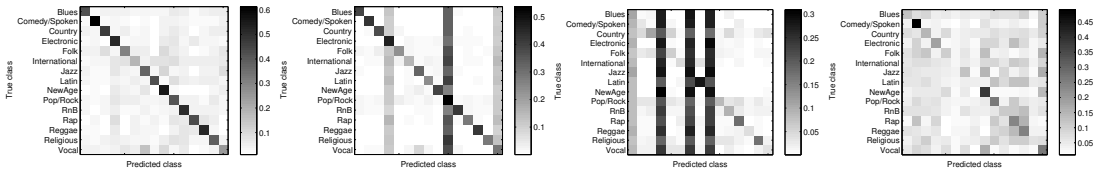


Fig. 4: Dark blue: Combined model, Light Blue: Tags, Green: Lyrics, Orange: Audio, Red: Audio+Lyrics, genre, $T = 128$.



(a) Combined Model (b) Tag Model (c) Lyrics Model (d) Audio Model

Fig. 5: Confusion matrices for genre and 128 topics. The color level indicates the classification accuracy.

summary, the ranking of tag alignment using our modeling approach on the MSD provides some evidence in favor of such coherence.

4.2. Prediction

Given the evidence presented for genre and style being the relatively most appropriate human categories, our second aim is to evaluate the predictive performance of the multi-modal model for genre and style, and we turn to the recently published extension of the MSD [9] for reference test/train splits and genre and style labels. In particular, we use the balanced splits defined in [9].

For the genre case, this results in 2000 labeled examples per genre and 15 genres, thus resulting in $30,000$ songs. We estimate the predictive genre performance by 10-fold cross-validation. Fig. 4 shows the per-label classification accuracy (perfect classification equals 1). The total genre classification performance is illustrated in Fig. 3a. The corresponding result for style classification, based on a total of $50,000$ labeled examples, is shown in Fig. 3b. Both results are generated using $T = 128$ topics, 2000 Gibbs sweeps and predicting using the MAP estimate from the Gibbs sampler.

We first note that the combination of all modalities performs the best and significantly better than random as seen from Fig. 3, which is encouraging, and support the multi-modal approach. It is furthermore noted that the tag modality is able to perform very well. This indicates that despite the possibly noisy user expressed view, the model is able to find structure in line with the taxonomy defined in the reference labels of [9]. More interesting is perhaps the audio and lyric modalities and the combination of the two. This shows that lyrics performs the worst for genre, possibly due to the missing data in some tracks, while the combination is significantly better. For style there is no significant difference between audio and lyrics.

Looking at the genre specific performance in Fig. 4 we find a significant difference between the modalities. It appears that the importance of the modalities is partly in line with the fundamentally different characteristics of each specific genre. For example 'latin' is driven by very characteristic lyrics. Further insight can be obtained by considering the confusion matrices which show some systematic pattern of error in the individual modalities, whereas the combined model shows a distinct diagonal structure, highlighting the benefits of multi-modal integration.

5. CONCLUSION

In this paper, we proposed the multi-way LDA as a flexible model for analyzing and modeling multi-modal and heterogeneous music data in a large scale setting. Based on the analysis of tags and latent representation, we provided evidence for the common assumption that genre may be an acceptable proxy for cognitive categorization of (western) music. Finally, we demonstrated and analyzed the predictive performance of the generative model providing benchmark result for the Million Song Dataset, and a genre dependent performance was observed. In our current research, we are looking at purely supervised topic models trained for, e.g. genre prediction. In order to address truly multi-modal and multi-task scenarios such as [7], we are currently pursuing an extended probabilistic framework that include correlated topic models [28], multi-task models [29], and non-parametric priors [30].

6. REFERENCES

[1] L.K. Hansen, P. Ahrendt, and J. Larsen, "Towards cognitive component analysis," in *AKRR05-International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[2] C. Xu, N.C. Maddage, and X. Shao, "Musical genre classification using support vector machines," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 429–432, 2003.

[3] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," *Proc. of ISMIR*, pp. 369–374, 2009.

[4] F. Pachet and J.J. Aucouturier, "Improving timbre similarity: How high is the sky?," *Journal of negative results in speech and audio*, pp. 1–13, 2004.

[5] Y. Kazuyoshi, M. Goto, K. Komatani, R. Ogata, and H.G. Okuno, "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR*, 2006, pp. 296–301.

[6] E. Law, B. Settles, and T. Mitchell, "Learning to tag from open vocabulary labels," *Machine Learning and Knowledge Discovery in Databases*, pp. 211–226, 2010.

[7] J. Weston, S. Bengio, and P. Hamel, "Multi-Tasking with Joint Semantic Spaces for Large- Scale Music Annotation and Retrieval Multi-Tasking with Joint Semantic Spaces for Large-Scale Music Annotation and Retrieval," *Journal of New Music Research*, , no. November 2012, pp. 37–41, 2011.

[8] T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[9] A. Schindler, R. Mayer, and A. Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset," in *13th International Conference on Music Information Retrieval (ISMIR 2012)*. 2012.

[10] B. McFee and G. R. G. Lanckriet, "Hypergraph models of playlist dialects," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard Müller, Eds. 2012, pp. 343–348, FEUP Edições.

[11] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music Recommendation by Unified Hypergraph: Combining Social Media Information and Music Content," pp. 391–400, 2010.

[12] M. Mcvicar and T. de Bie, "CCA and a Multi-way Extension for Investigating Common Components between Audio , Lyrics and Tags .," in *CMMR*, 2012, number June, pp. 19–22.

[13] J. Arenas-García, A. Meng, K.B. Petersen, T. Lehn-Schiøler, L.K. Hansen, and J. Larsen, *Unveiling Music Structure Via PLSA Similarity Fusion*, pp. 419–424, IEEE, 2007.

[14] K. Yoshii and M. Goto, "Continuous pLSI and smoothing techniques for hybrid music recommendation," *International Society for Music Information Retrieval Conference*, pp. 339–344, 2009.

[15] S. K., S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," pp. 2–5, 2009.

[16] Zhi Zeng, Shuwu Zhang, Heping Li, W. Liang, and Haibo Zheng, "A novel approach to musical genre classification using probabilistic latent semantic analysis model," in *IEEE International Conference on Multimedia and Expo (ICME), 2009*, 2009, pp. 486–489.

[17] T. Bertin-Mahieux, "Clustering beat-chroma patterns in a large music database," in *International Society for Music Information Retrieval Conference*, 2010.

[18] Y. Cho and L.K. Saul, "Learning dictionaries of stable autoregressive models for audio scene analysis," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8, 2009.

[19] K. Seyerlehner, G. Widmer, and P. Knees, "Frame level audio similarity-a codebook approach," *Conference on Digital Audio Effects*, pp. 1–8, 2008.

[20] D.M. Blei and M.I. Jordan, "Modeling annotated data," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, 2003.

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

[22] T.L. Griffiths and M. Steyvers, "Finding scientific topics.," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 5228–35, Apr. 2004.

[23] H.M. Wallach, I. Murray, Ruslan Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, , no. d, pp. 1–8, 2009.

[24] T. Hofmann, "Probabilistic latent semantic analysis," *Proc. of Uncertainty in Artificial Intelligence, UAI*, p. 21, 1999.

[25] J.H. Lee and J..S Downie, "Survey of music information needs, uses, and seeking behaviours: Preliminary findings," in *Proc. of ISMIR*, 2004, pp. 441–446.

[26] J. Frow, *Genre*, Routledge, New York, NY, USA, 2005.

[27] E. Law, "Human computation for music classification," in *Music Data Mining*, T. Li, M. Ogihara, and G. Tzanetakis, Eds., pp. 281–301. CRC Press, 2011.

[28] S. Virtanen, Y. Jia, A. Klami, and T. Darrell, "Factorized Multi-Modal Topic Model," *auai.org*, 2010.

[29] A. Faisal, J. Gillberg, J. Peltonen, G. Leen, and S. Kaski, "Sparse Nonparametric Topic Model for Transfer Learning," *dice.ucl.ac.be*.

[30] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, 2004.

Appendix B

A Topic Model Approach to Multi-Modal Similarity

Rasmus Troelsgaard, Bjørn Sand Jensen, and Lars Kai Hansen. "A Topic Model Approach to Multi-Modal Similarity". In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. 2013, pp. 1–5. arXiv: arXiv:1405.6886v1

A Topic Model Approach to Multi-Modal Similarity

Rasmus Troelsgård, Bjørn Sand Jensen and Lars Kai Hansen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Matematiktorvet 303B, 2800 Kgs. Lyngby
{rast,bjje,lkai}@dtu.dk

Abstract

Calculating similarities between objects defined by many heterogeneous data modalities is an important challenge in many multimedia applications. We use a multi-modal topic model as a basis for defining such a similarity between objects. We propose to compare the resulting similarities from different model realizations using the non-parametric Mantel test. The approach is evaluated on a music dataset.

1 Introduction

Calculating similarity between objects linked to multiple data sources is more urgent than ever. A prime example is the typical multimedia application of music services where users face a virtually infinite pool of songs to choose from. Here choices are based on many different information sources including the audio/sound, meta-data like genre, and social influences [1], hence, attempts of modeling the geometry of music navigation have taken on a multi-modal perspective. In fusing heterogeneous modalities like audio, genre, and user generated tags it is both a challenge to establish a combined model in a 'symmetric' manner so that one modality do not dominate others and it is challenging to evaluate the quality of the resulting geometric representation. Here, we focus on the latter issue by testing the consistency of derived inter-song (dis-)similarity by means of direct comparison between similarities using the Mantel permutation test.

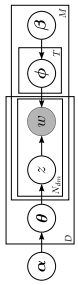
Topic models have previously been used to infer geometry in the image and music domain, e.g. by [2] combining audio features and listening histories. In [3] images and tags were analyzed, also by means of a multi-modal topic model. In [4] music similarity is inferred with a nonparametric Bayesian model, and [5] describe multiple multi-modal extensions to basic LDA models and evaluate the models on an image information retrieval task. Furthermore, topic model induced similarities among documents have been put to use in a navigation application [6], and different similarity estimates are also discussed in relation to a content-based image retrieval problem [7].

2 Model & Inference

To be able to measure similarities between objects, a representation of these objects is needed. In this work we use a version of Latent Dirichlet Allocation that incorporates multiple sources of information into a joint object representation similar to [5]. In [8], this model was applied to a multilingual corpus. Each object is represented by a multinomial distribution over topics which is common for all of the modalities composing the object. Each topic is defined by a set of multinomial distributions over features, each of which is defined on the vocabulary specific for a modality. To explain the characteristics of the model, the assumed generative process for objects is outlined in figure 1 together with a graphical representation of the model. The difference from a number of individual LDA models, each defined on a separate modality, is that each object is described by a single, shared distribution over topics, which potentially induces strong dependencies between the feature distributions representing the same topic in the individual modalities.

- For each topic indexed by $t \in [1; T]$ in each modality indexed by $m \in [1; M]$
  Draw $\boldsymbol{\phi}_t^{(m)} \sim \textit{Dirichlet}(\boldsymbol{\beta}^{(m)})$
  This is the parameters of the t^{th} topic's distribution over vocabulary $[1; V^{(m)}]$ of modality m.
- For each document indexed by $d \in [1; D]$
  - Draw $\boldsymbol{\theta}_d \sim \textit{Dirichlet}(\boldsymbol{\alpha})$
    This is the parameters of the d^{th} documents's distribution over topics $[1; T]$.
  - For each modality $m \in [1; M]$
    - For each word w in the m^{th} modality of document d
      - Draw a specific topic $z^{(m)} \sim \textit{Categorical}(\boldsymbol{\theta}_d)$
      - Draw a word $w^{(m)} \sim \textit{Categorical}(\boldsymbol{\phi}_{z^{(m)}}^{(m)})$

(a) Generative process



(b) The multi-modal Latent Dirichlet Allocation model represented as a probabilistic graphical model.

Figure 1

Performing inference in the model amounts to estimation of the posterior distributions over the latent variables. We use a Gibbs sampler inspired by the sparsity improvements proposed by [9]. For evaluation (see section 4), we use point estimates $\boldsymbol{\theta}^s$ and $\boldsymbol{\phi}^s$ derived from a sample $\mathbf{z}^s$ from the Markov chain, by taking the expectations of the respective posterior Dirichlet distributions defined by $\mathbf{z}^s$. In this work we choose the state of the chain with the highest model evidence within the last 50 out of 4000 iterations. Hyper-parameters are optimized using fixed point updates [10, 11]. The prior on the document topic distributions is an asymmetric Dirichlet with parameter $\boldsymbol{\alpha}$, and the priors over the vocabularies of the respective modalities are symmetric Dirichlet distributions with parameters $\boldsymbol{\beta}^{(m)}$.

3 Similarities in Topic Models

As already hinted, there are many ways to define and calculate similarities in topic models; both between topics and documents. In this paper we focus on the latter. Most methods in literature are based solely on the distributions of topics in the documents, $\boldsymbol{\theta}$, e.g. [4] measures the Kullback-Leibler divergence between two such distributions, while [7] also mentions inner products and cosine similarities as candidates. With focus on visualization, [6], introduces the yet another dissimilarity measure based on topic proportions. [7] promotes a measure based on the predictive likelihood of the document contents, and this approach is the basis of the method chosen here; The similarity of two documents A and B is given by the mean per-word log-likelihood of the words of document A given the topic distribution of document B (and the vocabulary distributions).

$$\frac{\log p(\mathbf{w}_A|\boldsymbol{\theta}_B^s, \boldsymbol{\phi}^s)}{\sum_{m=1}^{M} N_A^{(m)}}, \quad \text{where} \quad p(\mathbf{w}_A|\boldsymbol{\theta}_B^s, \boldsymbol{\phi}^s) = \prod_{m=1}^{M} \prod_{i=1}^{N_A^{(m)}} \sum_{t=1}^{T} (\boldsymbol{\phi}_{t,w_{Ai}^{(m)}}^{(m)})^\top \boldsymbol{\theta}_{t,B} \tag{1}$$

We use this approach to calculate a non-symmetric similarity matrix between all objects in the held-out cross-validation fold, for which the topic proportions have been estimated using "fold-in". [1] While this similarity measure is more computationally demanding than e.g. the KL-divergence, when the number of topics T used in the model increases, it might happen that some topics have vocabulary distributions that are very alike and only differ on a few words. Thus two documents with mainly the same type of content may have large proportions of different topics, causing them to be very dissimilar according to a topic proportion based measure. For a non-parametric topic model such as [4], this might not be a large concern, however, for parametric topic models, this should be taken into consideration. Generally, most of the discussed similarity measures are not proper metrics

[1] For the few held-out documents that do not contain any words in the modalities used for model estimation, we chose to simulate a uniform distribution of words in such an empty document by one occurrence of every word in the vocabulary.

in the geometric sense, but for (dis-)similarity purposes the exact properties might not be important, depending on the application.

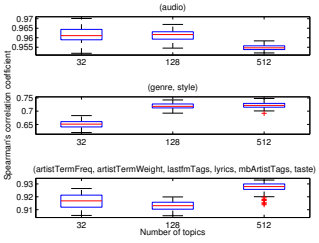
Comparing Similarities - the Mantel test
An important aspect of this work is the ability to assess the relations between different similarities induced by models estimated from multiple, possibly different, heterogeneous data sources. To compare such similarities we look at the correlation between the defined similarities. For testing the significance of the correlations we can apply a Mantel style test [12]. The Mantel test is a non-parametric test to assess the relation between two (dis-)similarity matrices. The null hypothesis is that the two matrices are unrelated, and the null distribution is approximated by calculating the test statistic for a large number of random permutations of the two matrices (excluding the diagonal elements); permuting rows and columns together to maintain the distribution of (dis-)similarities for each object. In this work we use Spearman's correlation coefficient as the test statistic.

4 Experimental Results: Music Similarity

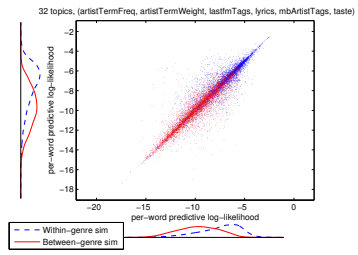
In this preliminary study we examine induced similarities in a subset of the Million Song Dataset [13], consisting of 30.000 tracks with equal proportions of 15 different genres. Each track is composed of data from a number of different sources: Open vocabulary tags from users (last.fm), Lyrics (musiXmatch.com), Editorial artist tags (allmusic.com), Artist tags (musicBrainz), User listening history (echonest), Genre and style (allmusic), and Audio Features (echonest). All modalities—besides the audio features—are naturally occurring as counts of words and for the audio we turn to an *audio word* approach, where the continuous features are vector quantized into a total of 2144 words. For this pilot study we estimate topic models on combinations of groups of modalities from the mentioned list, respectively consisting of the first 5, the genre and style labels, and the audio. To be able to assess the model stability of the similarities, we estimate each model five times from different random initialisations of the Markov chain. This is done for every training set of a 10-fold cross-validation split. The correlations between all combinations of the 5 similarity matrices resulting from each held-out fold are then calculated, and the resulting distributions of correlation coefficients are shown in figure 2a. Figure 3a shows the distributions of correlations between similarities based on audio and on the larger modality group. The correlations are evidently much smaller than for identical models, but a Mantel test with 100 permutations suggest that the null hypothesis of no correlation can be rejected at a significance level of at least 1% for all three model complexities.

5 Discussion & Conclusion

The issue of stability is relevant for similarities induced by topic models using approximate inference techniques. The correlations between similarities from identical but randomly initialized models,

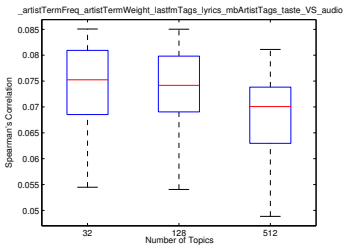


(a) Boxplots of Spearman's correlation between similarity matrices obtained using the same parameters and data, but different random initialisation. Each row correspond to a modality group.

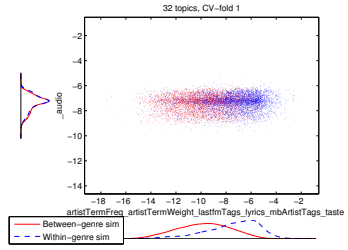


(b) Scatter plot of two specific similarities obtained by random initialisations of the same model estimated from the larger group of modalities. The colors indicate within- and between-genre similarities.

Figure 2



(a) Boxplots of Spearman's correlation between similarity matrices obtained using the larger modality group and the audio.



(b) Scatter plot of the two examples of similarities obtained from topic models with same parameters, but two mutually exclusive modalities of data.

Figure 3

can be used as a tool to gain some insight into this matter. From the preliminary results on the music example we find the induced similarities (fig. 2) to be highly stable. Furthermore, inspecting the similarities obtained from different data types; figure 3, we observe that while the audio model in itself does not seem to provide higher intra- than inter-genre similarity, it is still significantly positively correlated to the other modality group which does possess some discriminative power in terms of genre labels. Moreover, it seems that an increasing number of topics causes the correlation between similarities from models estimated on different modality groups to decrease. We speculate that this is linked to the specific topic model variant, for which [5] also note that the model describes the joint distribution of different modalities well, but does not model the relations between them.

In conclusion, we have proposed the multi-modal LDA as a method to define similarities in multimedia applications with multiple heterogeneous data sources based on the predictive-likelihood. This was extended with the Mantel test allowing direct evaluation of the consistency and correspondence of the resulting similarities.

Acknowledgment

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

References

[1] M.J. Salganik, P. Sheridan Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[2] Kazuyoshi Yoshii, M. Goto, K. Komatani, R. Ogata, and H.G. Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR*, pages 296–301, 2006.

[3] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer pLSA for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 9:1–9:8, New York, New York, USA, 2009. ACM Press.

[4] M. Hoffman, D. Blei, and P Cook. Content-based musical similarity computation using the hierarchical dirichlet process. *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, pages 349–354, 2008.

[5] D.M. Blei and M.I. Jordan. Modeling annotated data. *Annual ACM Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[6] A.J.B. Chaney and D.M. Blei. Visualizing Topic Models. *International AAAI Conference on Social Media and Weblogs*, 2012.

[7] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 17–24, New York, New York, USA, 2007. ACM Press.

[8] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, 2009.

[9] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 937–946, New York, NY, USA, 2009. ACM.

[10] T. P. Minka. Estimating a Dirichlet distribution. *Annals of Physics*, 2000(8):1–14, 2012.

[11] H. M Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[12] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

[13] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.

Appendix C

Spectral Learning of Hidden Markov Models in Non-stationary Data

---

Rasmus Troelsgaard and Lars Kai Hansen. "Spectral Learning of Hidden Markov Models in Non-stationary Data". In: *[submitted]* (2016)

Spectral Learning of Hidden Markov Models in Non-stationary Data

Rasmus Troelsgaard and Lars Kai Hansen

***Abstract*—In this letter we estimate the parameters of Hidden Markov Models via spectral estimation using empirical third order moments based on full sequences. Recently, parameter estimation techniques based on the method of moments have been proposed for Hidden Markov Models. These methods were either restricted to estimation based on the first three observations in the observed data sequences or forced to assume that the hidden Markov chain's initial state was drawn from the stationary distribution. We propose a method for estimation of moments involving all observed data without assumed stationarity. The scheme is based the observation that the specific formulations of the original methods remain valid when applied to averages of moments with different initial distributions, hence different positions in the observed sequences. The potential gains are illustrated in a set of numerical experiments.**

***Index Terms*—Method of Moments, Spectral Estimation, Hidden Markov Model**

I. Introduction

SPECTRAL methods for parameter estimation in statistical latent variable models have gained great interest in the signal processing community. This includes work on the Hidden Markov model (HMM)[1], [2], reduced-rank HMM [3], mixture of HMMs [4], discriminative training of HMMs [5], overcomplete latent variable models [6], [7], and on the use of spectral algorithms as a means to initialisation of maximum likelihood methods [8]. Additional perspectives on spectral methods for time series models were provided in [9], where the estimation problem is turned into a regularized optimisation problem involving functions over strings represented as Hankel matrices.

The main tools of spectral learning are matrix factorization techniques. In the particular case of the HMM, the model parameters can be recovered by applying the singular value decomposition to certain empirical third order moments of the data and performing basic linear algebra. The methods are valid under mild rank conditions. [1], [2] require full column rank of observation and transition matrices, while [3] enables estimation of models with lower rank than number of states.

The surge in the field of spectral learning has been fed by proofs of global convergence and low computational complexity. For convenience, the theoretical work make additional assumptions that formally constrain estimators to be based on the first three observations from the data sequences. The limitation to the first triplet simplifies the proven bounds because observed triplets in this case can be regarded i.i.d. and the correlation of triplets due to serial correlation can be ignored. While this is useful from a theoretical point of view relatively few papers discuss these constraints. [1] states that all available data in principle ought to be used, and [3] mentions that by assuming stationarity it is possible to form moments from as little as a single observed sequence. However, for most applications, the stationarity assumption is not suitable, hence we are left with the engineering issue: How can the spectral estimation method can be applied to general HMM problems and how can we use all observed triplets? In this letter we establish such tools to take advantage of full set of observed data sequences without assuming stationarity. We focus on the particular spectral estimation framework of [2] although the presented results are equally valid for all presentations sharing the same components of third order moments and spectral decompositions. Our work could in colloquial terms be regarded as improving the "practical statistical efficiency" of the existing method of [2].

Rasmus Troelsgaard and Lars Kai Hansen are with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

II. Hidden Markov Models and Spectral Estimation

This section briefly recapitulates definitions and model formulations necessary for description and further analysis of the use of empirical moments for estimation of parameters in a Hidden Markov Model.

A. Hidden Markov Models

The Hidden Markov Model is a widely used probabilistic latent variable model for sequence data, where observations are drawn from distributions conditioned on unobserved states. The unobserved states are assumed to form a Markov chain. We parametrize the HMM as follows:

$\boldsymbol{T} \in \mathbb{R}^{S\times S}$: Transition probability matrix
$\boldsymbol{T}_{g,h} = P(z_n = g | z_{n-1} = h) \qquad n \geq 2$

$\boldsymbol{O} \in \mathbb{R}^{K\times S}$: Observation probability matrix
$\boldsymbol{O}_h = \mathbb{E}[\boldsymbol{x}_t | z_n = h] \qquad n \geq 1$

$\boldsymbol{\pi}^{(1)} \in \mathbb{R}^{S}$: Initial state probability vector
$\boldsymbol{\pi}^{(1)}_h = P(z_1 = h)$

for $g, h \in \{1, 2, \cdots, S\}$.

B. Spectral Learning of HMM Parameters

Following [2], we define the second and third order moments as:

$$P_{n,n+2} = \mathbb{E}[\boldsymbol{x}_n \otimes \boldsymbol{x}_{n+2}] \tag{1}$$

$$P_{n,n+2,n+1} = \mathbb{E}[\boldsymbol{x}_n \otimes \boldsymbol{x}_{n+2} \otimes \boldsymbol{x}_{n+1}] \tag{2}$$

And let the third order moment $P_{n,n+2,n+1}$ act as a linear operator on the vector $\boldsymbol{\eta} \in \mathbb{R}^K$:

$$P_{n,n+2,n+1}(\boldsymbol{\eta}) = \mathbb{E}[(\boldsymbol{x}_n \otimes \boldsymbol{x}_{n+2})\langle \boldsymbol{\eta}, \boldsymbol{x}_{n+1}\rangle] \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product.
Let $\boldsymbol{\pi}^{(n)} = \boldsymbol{T}^{n-1}\boldsymbol{\pi}^{(1)}$ denote the expectation of the state distribution at time $n > 0$, then for $n \in \{1, 2, \ldots, N-2\}$ (3) is related to the Hidden Markov model parameters by:

$$P_{n,n+2,n+1}(\boldsymbol{\eta}) = \boldsymbol{O} diag(\boldsymbol{\pi}^{(n)}) \boldsymbol{T}^\top diag(\boldsymbol{\eta}^\top \boldsymbol{O}) \boldsymbol{T}^\top \boldsymbol{O}^\top \tag{4}$$

Similarly, the second order moments are related to the model parameters by

$$P_{n,n+2} = \boldsymbol{O} diag(\boldsymbol{\pi}^{(n)}) \boldsymbol{T}^\top \boldsymbol{T}^\top \boldsymbol{O}^\top \tag{5}$$

The procedure described in [2] is based on eigendecompositions of the following observable operator.

$$\begin{aligned} B_{1,3,2}(\boldsymbol{\eta}) &= (\boldsymbol{U}_1^\top P_{1,3} \boldsymbol{U}_3)^{-1} \boldsymbol{U}_1^\top P_{1,3,2}(\boldsymbol{\eta}) \boldsymbol{U}_3 \\ &= (\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3)^{-1} diag(\boldsymbol{\eta}^\top \boldsymbol{O}) (\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3) \end{aligned} \tag{6}$$

provided that $\boldsymbol{T}$ and $\boldsymbol{O}$ have column rank S, and $\boldsymbol{U}_1 \in \mathbb{R}^{K \times S}$ and $\boldsymbol{U}_3 \in \mathbb{R}^{K \times S}$ are matrices of orthonormal column vectors such that $\boldsymbol{O}^\top \boldsymbol{U}_1$ and $\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3$ are invertible. Thus by replacing $P_{1,3}$ and $P_{1,3,2}$ with emperical estimates $\hat{P}_{1,3}$ and $\hat{P}_{1,3,2}$, the eigendecomposition of (6) can then be used to recover the HMM parameters following the procedure of [2]. In practical applications, for convenience $\boldsymbol{U}_1$ and $\boldsymbol{U}_3$ are chosen to be the left and right singular vectors of $\hat{P}_{1,3}$ respectively, corresponding to the S largest singular values.

Note, by letting the columns of $\boldsymbol{O}$ represent state-conditional expectations, the above formulation supports both disctrete and continous observations [2], [10]. E.g. $P(\boldsymbol{x}_n | z_n)$ could be assumed to be a Poisson, Gaussian, or multinomial distribution.

III. Using Averages of Moments

We now present the main result of our letter, namely a simple proof that even without assuming stationarity, one can exploit all observations in a given set of sequences assumed to stem from the same HMM.

Condition 1. *$\sum_{n=1}^{N-2} diag(\boldsymbol{\pi}^{(n)})$ has rank S*

Condition 1 can be met in the following three ways.

1) $\boldsymbol{\pi}_i^{(1)} > 0 \quad \forall\, i \in \{1, 2, \ldots, S\}$
2) $\boldsymbol{T}_{i,j} > 0 \quad \forall\, i, j \in \{1, 2, \ldots, S\}$ and $N > 1$
3) ergodicity of the Markov chain and that $N \geq S$

1) seems a bit restrictive as it prevents us from dealing with HMMs certain to start in a particular state. 3) which is the least restrictive, both $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{T}$ are allowed to contain zeros as long as $\boldsymbol{T}$ is ergodic. In the worst case of a maximally sparse $\boldsymbol{\pi}^{(1)}$ (one-hot) and a sparse, yet ergodic $\boldsymbol{T}$, $\boldsymbol{\pi}^{(n)}$ will become increasingly dense for each multiplication by $\boldsymbol{T}$.

Theorem 1. *Assume Condition 1. Let $\tilde{P}_{1,2,3} = \frac{1}{N-2}\sum_{n=0}^{N-3} P_{n+1,n+2,n+3}$ denote the sum over expectations of all triples in a sequence, then the observable operator $\tilde{B}_{1,3,2}(\boldsymbol{\eta}) = \boldsymbol{U}_c^\top \tilde{P}_{3,1,2}(\boldsymbol{\eta}) \boldsymbol{U}_a (\boldsymbol{U}_c^\top \tilde{P}_{3,1} \boldsymbol{U}_a)^{-1}$ can be used to estimate model parameters $\boldsymbol{O}$ and $\boldsymbol{T}$ via the method described in [2].*

We can regard $\tilde{P}_{1,2,3} \in \mathbb{R}^{K \times K \times K}$ as a linear operator on a vector $\boldsymbol{\eta}$ defined in the same way as $P_{1,2,3}(\boldsymbol{\eta})$.
Theorem 1 states a result for a single sequence, but it is just as valid for multiple sequences (assumed to be generated by the same model).

Proof of Theorem 1.

$$\begin{aligned} \tilde{P}_{1,3,2}(\boldsymbol{\eta}) &= \frac{1}{N-2}\sum_{n=0}^{N-3} \mathbb{E}[(\boldsymbol{x}_{n+1} \otimes \boldsymbol{x}_{n+3})\langle \boldsymbol{\eta}, \boldsymbol{x}_{n+2}\rangle] \\ &= \frac{1}{N-2}\sum_{n=0}^{N-3} \boldsymbol{O} diag(\boldsymbol{\pi}^{(n+1)}) \boldsymbol{T}^\top diag(\boldsymbol{\eta}^\top \boldsymbol{O}) \boldsymbol{T}^\top \boldsymbol{O}^\top \\ &= \boldsymbol{O}\left(\frac{1}{N-2}\sum_{n=0}^{N-3} diag(\boldsymbol{\pi}^{(n+1)})\right) \boldsymbol{T}^\top diag(\boldsymbol{\eta}^\top \boldsymbol{O}) \boldsymbol{T}^\top \boldsymbol{O}^\top \end{aligned}$$

$$\begin{aligned} \tilde{P}_{1,3} &= \frac{1}{N-2}\sum_{n=0}^{N-3} \mathbb{E}[(\boldsymbol{x}_{n+1} \otimes \boldsymbol{x}_{n+3})] \\ &= \frac{1}{N-2}\sum_{n=0}^{N-3} \boldsymbol{O} diag(\boldsymbol{\pi}^{(n+1)}) \boldsymbol{T}^\top \boldsymbol{T}^\top \boldsymbol{O}^\top \\ &= \boldsymbol{O}\left(\frac{1}{N-2}\sum_{n=0}^{N-3} diag(\boldsymbol{\pi}^{(n+1)})\right) \boldsymbol{T}^\top \boldsymbol{T}^\top \boldsymbol{O}^\top \end{aligned}$$

$$\begin{aligned} \tilde{B}_{3,1,2}(\boldsymbol{\eta}) &= \boldsymbol{U}_c^\top \tilde{P}_{3,1,2}(\boldsymbol{\eta}) \boldsymbol{U}_a (\boldsymbol{U}_c^\top \tilde{P}_{3,1} \boldsymbol{U}_a)^{-1} \\ &= \boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T} diag(\boldsymbol{\eta}^\top \boldsymbol{O}) (\boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T})^{-1} \quad \times \\ &\quad \boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T}\boldsymbol{T}\left(\frac{1}{N-2}\sum_{n=0}^{N-3} diag(\boldsymbol{\pi}^{(n+1)})\right) \boldsymbol{O}^\top \boldsymbol{U}_a \quad \times \\ &\quad \left(\boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T}\boldsymbol{T}\left(\frac{1}{N-2}\sum_{n=0}^{N-3} diag(\boldsymbol{\pi}^{(n+1)})\right) \boldsymbol{O}^\top \boldsymbol{U}_a\right)^{-1} \\ &= (\boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T}) diag(\boldsymbol{\eta}^\top \boldsymbol{O}) (\boldsymbol{U}_c^\top \boldsymbol{O}\boldsymbol{T})^{-1} \end{aligned} \tag{7}$$

(7) has the same form as (6) and can thus be used to recover the HMM parameters as described in [2]. □

Condition 1 ensures that $\tilde{P}_{a,c}$ is invertible.

IV. Empirical Evaluation/Simulation

To illustrate the potential improvements of using all data, we conduct a simulation study using Hidden Markov models with discrete observations.

The experiment illustrates how the empirical moment estimates as described in table I, perform under various combinations of parameters influencing the difficulty of the estimation problem.

Table I
DESCRIPTIONS OF THE THREE TYPES OF EMPIRICAL MOMENTS USED IN THE EXPERIMENT.

$(1,2,3)$	Third order moments are estimated from first triplet in all of the N_c sequences.
(full)	Third order moments are estimated from the full sequences using only the first N_c triplets. Let N_{full} denote the number of sequences used for the estimation. This quantity is roughly $\frac{N_c}{l}$ when sequence lengths are $\sim Pois(l)$.
$(1,2,3_{N_{\text{full}}})$	Third order moments are estimated from the first triplet from N_{full} sequencess.

We now define a scalar quantity to quantify how well a particular empirical third order moment estimates the parameters of a particular HMM. Let M be an HMM. Let $\hat{\boldsymbol{P}}_{1,3}$ and $\hat{\boldsymbol{P}}_{1,2,3}$ be particular second and third order moments calculated from a finite set of sequences generated by M. Let $\boldsymbol{U}_3$ be the matrix of right singular vectors of $\hat{\boldsymbol{P}}_{1,3}$ corresponding to the S larges singular values, then according to (6)

$$\boldsymbol{\eta}^\top \boldsymbol{O} = diag\left((\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3)\boldsymbol{B}_{1,3,2}(\boldsymbol{\eta})(\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3)^{-1}\right) \quad (8)$$

From the empirical observable operator

$$\hat{\boldsymbol{B}}_{1,3,2}(\boldsymbol{\eta}) = (\boldsymbol{U}_1^\top \hat{\boldsymbol{P}}_{1,3}\boldsymbol{U}_3)^{-1}\boldsymbol{U}_1^\top \hat{\boldsymbol{P}}_{1,\boldsymbol{\eta},3}\boldsymbol{U}_3 \quad (9)$$

we can compute a related quantity

$$\boldsymbol{w}(\boldsymbol{\eta}) = diag\left((\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3)\hat{\boldsymbol{B}}_{1,3,2}(\boldsymbol{\eta})(\boldsymbol{T}^\top \boldsymbol{O}^\top \boldsymbol{U}_3)^{-1}\right) \quad (10)$$

We calculate each of these quantities for all columns of $\boldsymbol{H} = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_S]$ and construct the matrix $\boldsymbol{W} = [\boldsymbol{w}(\boldsymbol{\eta}_1), \boldsymbol{w}(\boldsymbol{\eta}_2), \cdots, \boldsymbol{w}(\boldsymbol{\eta}_S)]^\top$.
We now define the scalar quantity of interest as the normalized Frobenius norm of the difference between the eigenvalues of the model and eigenvalues produced by the empirical moments

$$D(\boldsymbol{H}^\top \boldsymbol{O}, \boldsymbol{W}) = \frac{\left\|\boldsymbol{H}^\top \boldsymbol{O} - \boldsymbol{W}\right\|_{\text{Fro}}}{\|\boldsymbol{H}^\top \boldsymbol{O}\|_{\text{Fro}} + \|\boldsymbol{W}\|_{\text{Fro}}}. \quad (11)$$

The quantity is bounded between 0 and 1 for interpretation.

The difficulty of the estimation problem is controlled by the number of observed sequences N_c, by T_{diag} which determines the Markov chain's mixing, and finally the size of the latent space S.
The transition matrix $\boldsymbol{T}$ is constructed from a stochastic matrix $\boldsymbol{P} \in \mathbb{R}^{S\times S}$ with uniformly distributed elements by:

$$\boldsymbol{T} = \frac{1}{1 + T_{\text{diag}}}(\boldsymbol{P} + \boldsymbol{I}_s T_{\text{diag}})$$

Varying T_{diag}, is a way of controlling the auto correlation and thereby also the convergence time of the Markov chain. High values of T_{diag} result in higher correlation between neighbouring observations, which means a lower 'effective' sample size. In order to relate the results to the convergence times of the Markov chains, we also report an upper bound for the half-life on the total variation distance from the stationary distribution. We can derive this bound from the error bound given by [11, Theorem 2.7]. The bound depends on β_1, which denotes the second largest eigenvalue of the multiplicative reversibilization of $\boldsymbol{T}$:

$$M(\boldsymbol{T}) = \boldsymbol{T}\tilde{\boldsymbol{T}} \quad (12)$$

where

$$\tilde{\boldsymbol{T}}_{j,i} = \frac{\hat{\boldsymbol{\pi}}_j \boldsymbol{T}_{i,j}}{\hat{\boldsymbol{\pi}}_i} \quad (13)$$

The resulting upper bound on the total variation half-life is

$$t_{\frac{1}{2}} \leq \frac{-2\log 2}{\log \beta_1} \quad (14)$$

A. Simulation Results

For each combination of N_c, T_{diag} and S, we construct an HMM with $K = 10$ and generate N_c sequences with random lengths $\sim Pois(30)$.
From the generated sequences we form the three different empirical third order moments (table I) and compare their performaces in terms of (11).
Figure 1 summarizes the results of 500 repetitions of the described experiment. The structure of transition matrices ranges from uniform to very diagonal, and the number of observed sequences ranges from 10 to 10^4. The model complexity is varied by performing the experiment for $S \in \{2, 5\}$.
Mean values of $t_{\frac{1}{2}}$ and standard deviations hereof, calculated from the randomly generated HMMs, are listed in Table II.

We note a generaly improved performance for both increasing number of observations and lower model complexity. The other obvious feature of the results in Fig. 1 is that the performances of $(1,2,3)$ and (full) are very similar and in all cases they clearly outperform $(1,2,3_{N_{\text{full}}})$. Fig. 1 demostrates that observations beyond the first triplet can contain critical information about the HMM parameters.
Looking more closely at the performance as $t_{\frac{1}{2}}$ increases, we observe that $(1,2,3)$ and $(1,2,3_{N_{\text{full}}})$ generally improve due to the 'easier' estimation problem (see [2] for errors bounds), while the performance of (full) actually starts to decrease when $t_{\frac{1}{2}}$ gets beyond 30 (mean sequence length). This is an illustration of decreased effective sample size due to auto correlation in the Markov chains.

Table II
$t_{\frac{1}{2}}$ CORRESPONDING TO FIG. 1. MEAN±STD. DEV.

S	T_{diag}			
	0	1	10	100
2	0.66 ± 0.03	1.23 ± 0.05	9.36 ± 0.51	86.60 ± 2.86
5	0.78 ± 0.01	1.52 ± 0.01	9.96 ± 0.07	93.69 ± 0.54

V. CONCLUSION

We provided a proof that certain spectral estimation schemes can take advantage of all observed triplets for general Hidden

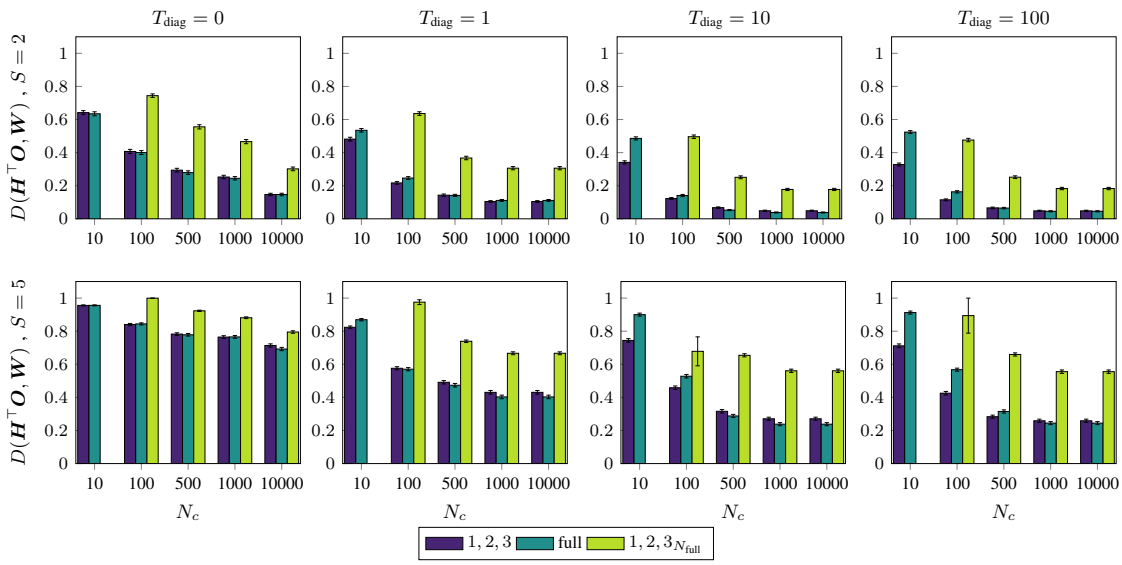


Figure 1. Normalized Frobenius norm of differences between eigenvalues of the model-based and the empirical observable operator (11). The height of a bar represents the mean value of $D(\boldsymbol{H}^\top \boldsymbol{O}, \boldsymbol{W})$ (11) calculated from 500 repetitions of the experiment. Each errorbar denotes the standard deviation of the estimated mean value. For $N_c = 10$ error values for $(1, 2, 3_{N_{\text{full}}})$ are missing due to the fact that none of 500 empirical second order moments had rank S because they were calculated from the first three observations in a single chain.

Markov model parameter estimation. A simulation demonstrated that performance is generally improved by this mechanism.

ACKNOWLEDGMENT

This work was supported in part by the Innovation Fund Denmark under the CoSound project, case number 0603-00475B. This publication only reflects the authors' views.

REFERENCES

[1] D. Hsu, S. M. Kakade, and T. Zhang, "A Spectral Algorithm for Learning Hidden Markov Models", *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012. arXiv: arXiv:0811.4413v6.

[2] A. Anandkumar, D. Hsu, and S. Kakade, "A method of moments for mixture models and hidden Markov models", *ArXiv preprint arXiv:1203.0683*, pp. 1–31, 2012. arXiv: arXiv:1203.0683v3. [Online]. Available: http://arxiv.org/abs/1203.0683.

[3] S. Siddiqi, B. Boots, and G. Gordon, "Reduced-rank hidden Markov models", in *Thirteenth International Conference on Artificial Intelligence and Statistics May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy*, vol. 9, 2010, pp. 741–748. [Online]. Available: http://arxiv.org/abs/0910.0902.

[4] C. Subakan, J. Traa, and P. Smaragdis, "Spectral Learning of Mixture of Hidden Markov Models", in *Advances in Neural Information ...*, 2014, pp. 1–9. [Online]. Available: http://papers.nips.cc/paper/5518-spectral-learning-of-mixture-of-hidden-markov-models.

[5] A. Nazábal and A. Artés-Rodríguez, "DISCRIMINATIVE SPECTRAL LEARNING OF HIDDEN MARKOV MODELS FOR HUMAN ACTIVITY RECOGNITION", in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 1966–1970, ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[6] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions", *Proceedings of the 46th ...*, pp. 594–603, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2591881.

[7] A. Anandkumar, A. A. Edu, R. Ge, and M. Janzamin, "Learning Overcomplete Latent Variable Models through Tensor Methods", *Proceedings of the Conference on Learning Theory (COLT), Paris, France*, vol. 40, pp. 1–77, 2015.

[8] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing", *Journal of Machine Learning Research*, vol. 17, pp. 1–44, 2016.

[9] B. Balle, A. Quattoni, and X. Carreras, "Local Loss Optimization in Operator Models: A New Insight into Spectral Learning", *International Conference on Machine Learning (ICML)*, 2012.

[10] A. Anandkumar, R. Ge, and D. Hsu, "Tensor decompositions for learning latent variable models", *The Journal of Machine ...*, vol. 15, pp. 2773–2832, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2697055.

[11] J. Fill, “Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process”, *The annals of applied probability*, 1991. [Online]. Available: http://www.jstor.org/stable/2959625.

Appendix D

Sequence Classification Using Third Order Moments

Rasmus Troelsgaard and Lars Kai Hansen. "Sequence Classification Using Third Order Moments". In: *[submitted]* (2016)

Sequence Classification Using Third Order Moments

Rasmus Troelsgaard[1]

rast@dtu.dk

Lars Kai Hansen[1]

lkai@dtu.dk

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

Keywords: Method of Moments, Time Series, Classification, Hidden Markov Model

Abstract

Model based classification of sequence data using a set of Hidden Markov models is a well known technique. The involved score function, which is often based on the class-conditional likelihood, can however be computationally demanding, especially for long data sequences. Inspired by recent theoretical advances in spectral learning of Hidden

Markov Models, we propose a score function based on third order moments. In particular we propose to use the Kullback-Leibler divergence between theoretical and empirical third order moments for classification of sequence data with discrete observations. The proposed method provides lower computational complexity at classification time than the usual likelihood based methods.

In order to demonstrate the properties of the proposed method, we perform classification of both synthetic data, and empirical data from a human activity recognition study.

1 Introduction

Classification and clustering of sequences into categories is essential to human interpretation of the data. Different methodologies have been proposed to deal with this problem, and (Xing, Pei, & Keogh, 2010) gives a brief and general overview of the field, including model based classification. The general approach in model based classification is to represent each class by a generative model, hence there are two main components in a model based classification system; the first is the formulation of the statistical model representing each of a given set of classes, and the second is a measure of distance between observed data and models. For probabilistic models, the obvious and common choice is to use a distance measure derived from the class-conditional likelihoods.

Each model is estimated using a set of exemplar data sequences (training set) representing a specific class. Hence the problem can be stated as follows: Given L trained models and a held out, observed sequence of length N, find the model that best fits the

observation. One classical approach to this problem is to use the (log-)likelihood of each class-conditional model given the test sequence as a score for the model-sequence pair. Usually the test sequence is assigned to the class model for which the (log-)likelihood is the highest.

In this letter we consider class-conditional model based classification for sequential data using Hidden Markov models.

Classification using Hidden Markov models in particular has previously been applied in a variety of contexts. (Oates, Firoiu, & Cohen, 1999) takes the classical model based approach to clustering of sequence data using one HMM per cluster. An original HMM based representation of images is explored in (Mouret, Solnon, & Wolf, 2009). In (Wong & Stamp, 2006) HMMs are used to represent software virus families, and a log-likelihood threshold is used for binary classification of benign software vs. malware. Another practical example is found in (Wang, Mehrabi, & Kannatey-Asibu, 2002) where HMM based classification is applied for tool wear monitoring in industrial machinery. In (Bicego, Murino, & Figueiredo, 2004), the similarities between sequences and models are used as features in a discriminatively trained classifier. One HMM is estimated for each training example, and all sequences are the embedded in the space of estimated HMMs using log-likelihood. This line of thought is also explored in (García-García, Emilio, & Díaz-de-María, 2009), where a KL-divergence based similarity measure is proposed.

Recently, methods based on spectral decomposition of observed data moments have been developed for parameter estimation in models for sequential data (Hsu, Kakade, & Zhang, 2012; Anandkumar, Hsu, & Kakade, 2012). While these methods provide

exciting results regarding both global convergence and computational complexity of the parameter estimation problem, the complexity of likelihood calculations which is of particular interest when performing model based sequence classification is unchanged. In settings where the amount of data to be classified is vast, and time spent on model estimation is of minor importance, we find ourselves in the need for a fast approximation to the likelihood that do not require the calculation of matrix products for every observation in a given sequence. The advances in spectral learning using moments enables us to view the third order moments as sufficient statistics under the model assumptions of (Hsu, Kakade, & Zhang, 2012; Anandkumar, Hsu, & Kakade, 2012). Based on this interpretation, we propose a simple framework for classification of sequences of discrete observations, using only observed third order moments. The distance measure we propose to substitute for likelihood calculations is based on Kullback-Leibler divergence between empirical and theoretical third order moments, and we show that it has lower computational complexity at classification time, while achieving indistinguishable performance.

This rest of this letter is organised as follows. Section 2 introduces the proposed distances score in the context of both stationary and non-stationary HMMs, and relates it to a particular composite likelihood. Next, we compare the computational complexity of the proposed method to the likelihood based approach. Finally a upper bound on the convergence time of a Markov chain is exploited to reduce memory requirements for the proposed method. Section 3 sketches an approach to sequence embedding wherein the distance score for sequence-model pairs plays a central role. In sections 4 and 5 we present classification results of both simulated and real world data sets respectively.

2 KL-Divergence Of Third Order Moments

In this section we use the interpretation of third order moments as sufficient statistics to develop a distance measure defined directly in terms of these moments.

The main idea is to utilise the third order moments of observed discrete sequences as multinomial probability distributions. Because the third order moments in the general case are dependent on the initial state distribution $\boldsymbol{\pi}^{(1)}$ we start by describing the simplified case of assumed stationarity of the HMM ($\boldsymbol{\pi}^{(1)} = \hat{\boldsymbol{\pi}}$).

2.1 Stationary Markov Processes

Let $\bar{P}_{1,2,3}$ be the empirical third order moment of the observed sequence, and let $P_{1,2,3}$ be the corresponding theoretical third order moment due to model parameters.

$$P_{1,2,3}(\cdot, k, \cdot) = \boldsymbol{O}\,\mathrm{diag}(\hat{\boldsymbol{\pi}})\boldsymbol{T}^{\top}\,\mathrm{diag}(\boldsymbol{O}(k,\cdot))\boldsymbol{T}^{\top}\boldsymbol{O}^{\top} \qquad k \in [1, K]$$

We can then use the KL-divergence of $P_{1,2,3}$ from $\bar{P}_{1,2,3}$ as measure of difference between a model and an observed sequence.

$$\mathrm{KL}(\bar{P}_{1,2,3} \| P_{1,2,3}) = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} \bar{P}_{1,2,3}(i,j,k) \log \frac{\bar{P}_{1,2,3}(i,j,k)}{P_{1,2,3}(i,j,k)}$$

Note that this is also valid in the non-stationary case, if observations from a suitable burn-in period are discard (see section 2.5). This however, requires that the length of the test sequence is at least as long as the maximum convergence time of the all class models, which might limit the practical usefulness of the method.

With the aim to avoid discarding burn-in data for classification in the non-stationary case, we now present the main contribution of this letter.

2.2 Non-stationary Markov Processes

If stationarity **cannot** be assumed, the expectation of the state distribution changes along the underlying Markov chain. Hence we have to consider the third order moments for each triplet in the observed sequence separately. Let $\bar{P}_{n,n+1,n+2}$ be the empirical third order moment of the triplet starting at position n in the sequence, and let $P_{n,n+1,n+2}$ be the corresponding theoretical third order moment due to model parameters

$$P_{n,n+1,n+2}(\cdot,k,\cdot) = \boldsymbol{O}\,\mathrm{diag}(\boldsymbol{T}^{n-1}\boldsymbol{\pi}^{(1)})\boldsymbol{T}^{\top}\,\mathrm{diag}(\boldsymbol{O}(k,\cdot))\boldsymbol{T}^{\top}\boldsymbol{O}^{\top} \qquad k \in [1,K]$$

We can then for an arbitrary position n calculate the KL-divergence of $P_{n,n+1,n+2}$ from $\bar{P}_{n,n+1,n+2}$:

$$\begin{aligned}\mathrm{KL}^{(n)} &= \mathrm{KL}(\bar{P}_{n,n+1,n+2} \| P_{n,n+1,n+2}) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} \bar{P}_{n,n+1,n+2}(i,j,k)\log\frac{\bar{P}_{n,n+1,n+2}(i,j,k)}{P_{n,n+1,n+2}(i,j,k)} \end{aligned} \tag{1}$$

Each $\mathrm{KL}^{(n)}$ can then interpreted as a cost describing how well $\bar{P}_{n,n+1,n+2}$ approximates the theoretical third order moment of that particular triplet $P_{n,n+1,n+2}$.

Note that in the typical classification scenario, the cost is calculated for a single sequence $\boldsymbol{x} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$. Thus for any given $n \in \{1,2,\ldots,N-2\}$, (1) reduces to $-\log\left(P_{n,n+1,n+2}(x^{(n)}, x^{(n+1)}, x^{(n+2)})\right)$. To obtain a cost using the full sequence, we calculate the arithmetic mean across all triplets, which is exactly equivalent to considering the joint discrete probability distribution of all triplets in the sequence.

$$\frac{1}{N-2}\sum_{n=1}^{N-2}\mathrm{KL}^{(n)} = \frac{1}{N-2}\sum_{n=1}^{N-2} -\log\left(P_{n,n+1,n+2}(x^{(n)}, x^{(n+1)}, x^{(n+2)})\right) \tag{2}$$

The described procedure requires the calculation of powers of $\boldsymbol{T}$ up to the N^{th} power (N being the length of the observed candidate sequence), which can be demanding in terms

of memory. However, the Markov chain converges to its stationary distribution, and for a given allowed distance ε from this stationary distribution it is possible to derive a bound on the convergence time for the chain. This can be exploited to limit the maximum power of $\boldsymbol{T}$ to calculate. In section 2.5 such a convergence time bound is derived.

Let $c_{i,j,k} \geq 0$ be the number of occurrences of the triplet (i, j, k) in the stationary part of the sequence $\boldsymbol{x}$, and let $c_s = \sum_i^K \sum_{j=1}^K \sum_{k=1}^K c_{i,j,k}$ be the number of triplets beyond the convergence time. We then simply calculate the KL-divergence from the stationary distribution and use the weighted arithmetic mean.

$$
\begin{aligned}
&\frac{1}{N-2} \sum_{n=1}^{N-2-c_s} \mathrm{KL}^{(n)} + \frac{c_s}{N-2} \mathrm{KL}^{\text{stationary}} \\
&\qquad = \frac{1}{N-2} \sum_{n=1}^{N-2-c_s} -\log\left(P_{n,n+1,n+2}(x^{(n)}, x^{(n+1)}, x^{(n+2)})\right) \\
&\qquad\quad + \frac{c_s}{N-2} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{c_{i,j,k}}{c_s} \log \frac{\frac{c_{i,j,k}}{c_s}}{\hat{P}_{1,2,3}(i,j,k)} \qquad (3) \\
&\qquad = \frac{1}{N-2} \sum_{n=1}^{N-2} -\log\left(P_{n,n+1,n+2}(x^{(n)}, x^{(n+1)}, x^{(n+2)})\right) \\
&\qquad\quad + \frac{1}{N-2} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} c_{i,j,k} \log \frac{c_{i,j,k}}{c_s}
\end{aligned}
$$

where $\hat{P}_{1,2,3}$ denotes the stationary third order moment. We observe that (3) is just (2) plus the additional term on the last line of (3), which is due to the Shannon entropy of the empirical stationary third order moment.

2.3 Interpretation as Composite Likelihood

An empirical moment estimated from a single triplet is clearly a very crude approximation, and contrast the usual practice of method of moments, where averaging over a huge number of samples is exploited. The intuition behind using the 1-sample approximations

along the chain is that each of the terms $\mathrm{KL}^{(n)} = -\log\left(P_{n,n+1,n+2}(i,j,k)\right)$ on average is lower for a matching pair of sequence and model than for non-matching pairs.

Furthermore, by viewing the model-based third order moments along a Markov chain as a re-parametrisation of the HMM, when disregarding the entropy term, (3) corresponds to a negative per-sample composite log-likelihood of this model given the observations (triplets). The pseudo-likelihood was introduced in (Besag, 1975) as a product of possibly correlated local conditional likelihood terms. Later, under the term *composite likelihood*, (Lindsay, 1988) generalised the concept to also include marginal likelihood terms of sub-components. This interpretation of the KL-divergence based distance further justifies the proposed approach. Based on the above analysis, we propose the following distance measure for model based classification using HMMs:

$$\mathcal{D}(\boldsymbol{x}, \mathcal{M}) = \frac{1}{N-2} \sum_{n=1}^{N-2} -\log\left(P_{n,n+1,n+2}(x^{(n)}, x^{(n+1)}, x^{(n+2)})\right) \tag{4}$$

2.4 Computational Complexity

We will now compare the computational complexity of the proposed method and the classical likelihood based approach. The cost of estimating the L class HMMs is disregarded as we focus solely on the classification step.

We start be examining the total complexity of scoring a single observed sequence by L estimated models. The likelihood calculations scale with $\mathcal{O}(LNS^2)$ thus we obtain a mean per-class complexity of $\mathcal{O}(NS^2)$.

In the stationary case, the third order moment of the test sequence can be calculated in $\mathcal{O}(N)$, and because it is independent of the number of classes it has to be calculated

only once. Comparison of the third order moment of a test sequence to moments of all the trained class models take $\mathcal{O}(\min(N, K^3)L)$. Here it is exploited that the cost function only depends on the N triplets that are actually observed. This means that in the stationary case, the total computational complexity of the moment comparison becomes $\mathcal{O}(\min(N, K^3)L + N)$, and $\mathcal{O}(\min(N, K^3) + \frac{N}{L})$ for the per-class complexity. In the non-stationary case we have to consider all triplets in the test sequence separately resulting in a total complexity of $\mathcal{O}(NL)$, and per-class complexity $\mathcal{O}(N)$. This analysis shows that the classification task in theory can be performed faster when using third order moments compared to the classical likelihood approach.

Although the computational complexity remains unchanged, the memory requirements will of course increase compared to the stationary situation as we have to store third order moments for all possible positions in a chain (in theory infinitely many). Section 2.5 outlines a method to limit the amount of required memory based on an upper bound on the convergence time of a Markov chain (to the stationary distribution).

2.5 Estimating Convergence Time for a Markov Chain

This section describes how to calculate a upper bound on the convergence time of an ergodic Markov chain given an upper bound on the total variation distance a any given time instance t. We begin by stating a bound for the slightly simpler case of a reversible Markov chain, and then proceed to the more general case of a non-reversible chain.

The convergence time of a reversible irreducible Markov chain with transition probability matrix $\boldsymbol{T}$ and stationary distribution $\hat{\boldsymbol{\pi}}$ can be bounded using an upper bound

on the relative point-wise distance $\Delta(t)$. This quantity is larger than the total variation distance $\Delta(t) = \max_{i,j} \left| \frac{\boldsymbol{T}^t_{i,j}}{\hat{\boldsymbol{\pi}}_i} - 1 \right|$ for which the following bound exists: $\Delta(t) \leq \frac{\beta_1(\boldsymbol{T})^t}{\hat{\boldsymbol{\pi}}_{min}}$ where $\beta_1(\cdot)$ denotes the second largest eigenvalue (Durrett, 2007, p. 161).

For a the general non-reversible Markov chain a similar result exists for the multiplicative reversibilisation of $\boldsymbol{T}$, $M(\boldsymbol{T}) = \boldsymbol{T}\tilde{\boldsymbol{T}}$, where $\tilde{\boldsymbol{T}}_{j,i} = \frac{\hat{\boldsymbol{\pi}}_j \boldsymbol{T}_{i,j}}{\hat{\boldsymbol{\pi}}_i}$ (Fill, 1991).

Let $\mathcal{X}_0^2 = \sum_{x=1}^{S} \frac{(\boldsymbol{\pi}_x^{(1)} - \hat{\boldsymbol{\pi}}_x)^2}{\hat{\boldsymbol{\pi}}_x}$, then according to (Fill, 1991) the upper bound on the total variation distance at time step t is

$$\left\| \boldsymbol{T}^t \boldsymbol{\pi}^{(1)} - \hat{\boldsymbol{\pi}} \right\|_{\text{TV}} = \frac{1}{2} \sum_{x=1}^{S} \left| (\boldsymbol{T}^t \boldsymbol{\pi}^{(1)})_x - \hat{\boldsymbol{\pi}}_x \right| \leq \frac{(\beta_1(M(\boldsymbol{T})))^{\frac{t}{2}}}{2} \mathcal{X}_0$$

from which we can construct a upper bound on t given an acceptable total variation distance $\varepsilon \in [0, \min\left(1, \frac{\mathcal{X}_0}{2}\right)]$:

$$\begin{aligned} & \varepsilon \leq \frac{(\beta_1(M(\boldsymbol{T})))^{\frac{t}{2}}}{2} \mathcal{X}_0 \\ \iff \quad & t \leq 2 \frac{\log\left(\frac{2\varepsilon}{\mathcal{X}_0}\right)}{\log \beta_1\left(M(\boldsymbol{T})\right)} \end{aligned} \tag{5}$$

2.6 Exploiting Approximate Convergence

We now show an example of how classification performance can be affected by the size of ε.

We illustrate the effect by analysing a synthetic 5-class problem using KL-divergence as the distance score in the classifier (as described in section 2.2). For the purpose of illustration, all class models share parameters $\boldsymbol{T}$ ans $\boldsymbol{S}$ but differ by their initial distributions $\boldsymbol{\pi}^{(1)}$. Thus all class conditional models have identical stationary distributions and identical stationary third order moments. We assess the classification performance using the well-known F_1-measure. Figure 1 shows how the classification performance

decreases when the accepted distance to the stationary distribution is increased.

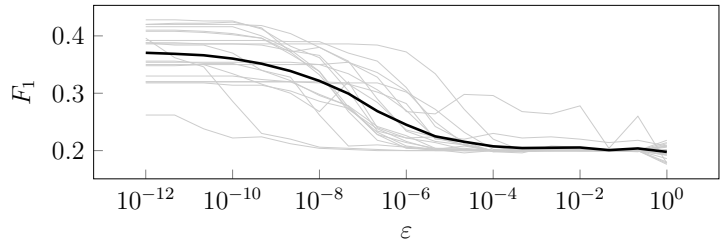


Figure 1: Classification performance using KL divergence as a function of ε. (grey) repetitions of the experiment, (black) mean classification score of the repetitions.

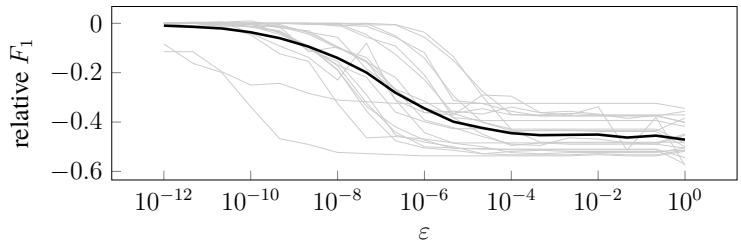


Figure 2: Classification performance using KL divergence as a function of ε. This plot shows the performance relative to using $\varepsilon = 10^{-20}$ (not assuming convergence). (grey) repetitions of the experiment, (black) mean classification score of the repetitions.

3 "Embedding" Sequences for Classification

To improve on the classical model based classification approach, several authors have suggested to "embed" the observed test sequences in a space spanned by the training sequences (García-García, Emilio, & Díaz-de-María, 2009; Bicego, Murino, & Figueiredo, 2004). An arbitrary discriminatively trained classifier can then be applied leveraging this new representation of sequences.

The main idea is to estimate a model for each training example and the embedding

then defined from the model-sequence distance score to all pairs of training examples. Similar to the work in (García-García, Emilio, & Díaz-de-María, 2009), for a single sequence we normalise its scores relating it to the training sequences, such that it sums to 1. This allows us to use the Jensen-Shannon divergence as the similarity score in the embedding space. Given a test sequence to be classified, one has to evaluate the distance score for all trained models. Hence the distance score remains a central component of the classification procedure. The procedure is described in Algorithm 1 in Appendix B.

We include this classification strategy to provide an alternative evaluation of the proposed KL-divergence based distance score. For the results presented in sections 4 and 5, we used a K-nearest-neighbour classifier where K was chosen via 5-fold cross-validation on the training sequences.

4 Classification of Simulated Time Series

This section seeks to illustrate how the proposed KL-divergence based score, $\mathcal{D}$, compares to the negative log-likelihood, ℓ, under different simulated conditions such as lengths of the observed sequences, diagonality of the transition matrices, and how interrelated the class-conditional models are.

For estimation of the class-conditional models we rely on the classical Baum-Welch/EM algorithm (Baum, Petrie, Soules, & Weiss, 1970; Dempster, Laird, & Rubin, 1977). Although alternatives such as spectral estimation techniques presented by (Anandkumar, Hsu, & Kakade, 2012; Anandkumar, Ge, & Hsu, 2014; Troelsgaard & Hansen, 2016) in

principle could be used as well, in order not to unintentionally favour the moment based classification scheme the likelihood based estimation is preferred.

The number of symbols in the training and test sequences are $\sim \text{Poisson}(\bar{N}), \quad \bar{N} \in \{10, 50, 200, 1000\}$. The numbers of training and test sequences per class are fixed at 30 and 50 respectively.

The diagonality is controlled by the parameter $T_{\text{diag}} \in]0, 1[$. $\rho \in]0; 1]$ controls the variance of the elements of $\boldsymbol{T}$, and is used as a means to generate sets of more or less interrelated HMMs. For a detailed description of the construction of the synthetic HMMs used in these classification experiments, we refer the reader to Appendix A.

We consider a synthetic classification problem with $L = 5$ classes, where each class-conditional model is a $S = 4$ state HMM with $K = 15$ discrete observation symbols

4.1 Results

The performance is reported in terms of the F_1-measure. The reported evaluation quantities are mean values over all classes. Each experiment was repeated 20 times to quantify variation in performance. The error bars denote the standard deviations of the estimated mean values.

Figure 3 shows how classification performance is improved by longer observed sequences. Furthermore, class-conditional models closer to each other are harder to distinguish between. These observations hold for both ℓ, and $\mathcal{D}$. The performances of the two methods are virtually indistinguishable with the exception that for long sequences ($\bar{N} \gtrsim 1000$) and class-conditional models quite close to each other ($\rho \lesssim 0.05$), $\mathcal{D}$ seem

to be superior. To better illustrate the minor differences, Fig. 4 shows the mean of the pairwise relative performances relative to ℓ. Hence the results for ℓ are constant at 1.

In total we performed 428 experiments with different combinations of parameters. With the null-hypothesis that $F_1(\ell) \geq F_1(\mathcal{D})$ we can calculate p-values for the experiment by applying Bonferroni correction to paired-samples binomial sign tests. Hence we calculate the probability of observing the experiment results or more extreme results under the null-hypothesis. For a couple of the classification problems with low values of ρ shown in the lower plots of Fig. 4 ($\bar{N} = 1000\,, T_{\text{diag}} \in \{0.25, 0.7, 0.95\}$) we find (corrected) p-values in the range $[0.0008, 0.0327]$ indicating that the null-hypothesis is very unlikely for these particular classification problems.

For the null-hypothesis $F_1(\ell) \leq F_1(\mathcal{D})$, The 3 lowest obtained p-value were 0.0620, 0.1722, and 0.3118 indicating no general tendency to rejecting the null-hypothesis.

In summary, the statistical tests indicate that using $\mathcal{D}$ as the distance score in HMM based classification of sequence data provides equally good results compared to the classical likelihood score ℓ.

4.2 Classification Results for Sequence Embedding

Using the sequence embedding procedure described in 3 we now compare performance to the classical model-based approach. Figure 5 shows that ℓ and $\mathcal{D}$ perform equally well in all the synthetic classification problems. Furthermore, we observe that the embedding improves performance slightly for moderate to long sequences ($\bar{N} \in 50, 200, 1000$) when class-conditional models are quite different and have a dominating diagonal structure ($T_{\text{diag}} = 0.95$). On the contrary, the embedding seems to have a negative impact on

performance under the conditions of more interrelated class-conditional models and less diagonal transition matrices. These performances of ℓ and $\mathcal{D}$ for high values of ρ are significantly better than without the embedding cf. the significance test in previous section.

Fig. 6 shows the time spent on classification relative to the time of ℓ. The figure clearly illustrates the gains of the reduced computational complexity of using $\mathcal{D}$ over ℓ, for everything but very short sequences.

4.3 Conclusion of Experiment with Simulated Data

For short sequences, using the classical log-likelihood approach is both faster and more accurate in terms of F_1 score. For increased sequence lengths in addition to being faster, the performance of the KL-divergence based method catches up and under certain conditions even seems to outperform the log-likelihood approach. Embedding test sequences in the space of training sequences seems to be most beneficial for long sequences ($\gtrsim 50$) as long as class-conditional models are quite dissimilar.

5 Classification of Human Activities

We now turn to application of the proposed method on non-simulated sequence data. We use the UCI HAR benchmark data set (Anguita, Ghio, Oneto, Parra, & Reyes-Ortiz, 2013), which is a human activity recognition data set consisting of inertial measurements from a waist-mounted mobile device during 6 different activities. We perform 5-fold cross-validation on the training set (21 persons) for finding optimal number of states S,

for each class. Table 1 shows the class labels and the optimal number of hidden states for each of the six classes. As input we used body acceleration and angular velocity, and all variables were scaled to unit variance, whitened and quantised into 50 "symbols" using K-means clustering (Elkan, 2003).

In this experiment we assume that the boundaries of activities are known in advance such that every training and test sequence only contains data from a single activity. Thus the task is to provide a label for each test segment.

Table 1: Optimal number of hidden states obtained via 5-fold cross validation on the training data set.

WALKING	13
WALKING-UPSTAIRS	6
WALKING-DOWNSTAIRS	4
SITTING	4
STANDING	3
LAYING	3

Using the values in table 1, we estimated one HMM per class using the full training set, and then classified the sequences corresponding to the 9 left-out persons. The confusion matrix shown in Table 2 is obtained from 80 random repetitions of the experiment. Thus variation in results are due to random initialisations of HMM parameters and of cluster centres in the quantisation process.

The mean (micro-averaged)F_1 scores for ℓ and $\mathcal{D}$ respectively, are $\frac{1}{80}\sum_{i=1}^{80} F_1(\ell^{(i)}) =$

0.8152 and $\frac{1}{80}\sum_{i=1}^{80} F_1(\mathcal{D}^{(i)}) = 0.8303$. Although these two numbers seem very close, both distance scores are applied to the same test-data and class-conditional models at each random repetition. Hence we are dealing with paired-samples of F_1 which enables us to evaluate the pairwise differences instead of two separate measures.

Table 2: Mean confusion tables from 80 repetitions of the UCI HAR classification task.

15.71	0.74	1.55	0	0	0
0.13	21.65	1.23	0	0	0
0.15	0.38	24.48	0	0	0
0	0	0	11.43	3.05	3.53
0	0	0	3.25	11.11	3.64
0	0	0	2.23	2.33	13.45

(a) Distance score: ℓ

15.8	0.88	1.33	0	0	0
0.06	22.04	0.9	0	0	0
0.1	0.24	24.66	0	0	0
0	0	0	11.44	3.98	2.59
0	0	0	3.61	12.66	1.73
0	0	0	2.29	2.68	13.04

(b) Distance score: $\mathcal{D}$, $\epsilon = 0.0001$

5.1 Paired-Samples Binomial Sign Test

To assess whether the difference is significant, we perform a one-sided paired samples sign test with the null-hypothesis $F_1(\ell) \geq F_1(\mathcal{D})$ and the alternative hypothesis that $F_1(\ell) \leq F_1(\mathcal{D})$. The number of pairs where $F_1(\ell^{(i)}) < F_1(\mathcal{D}^{(i)})$ is 49, and the opposite is 12. This results in a p-value of $9.85 \cdot 10^{-7}$ i.e. the probability of observing 12 or less negative differences if the null hypothesis is true. The result of this test suggests that $\mathcal{D}$ performs slightly better than ℓ for this particular classification problem.

6 Conclusion

We have proposed a new distance score for use in Hidden Markov model-base classification problems, dominated by long sequences of discrete observations. The score is based on expectations of triplets along a Markov chain, and can be interpreted as a composite likelihood for a moment-based Hidden Markov model representation. We show how the memory requirements of the proposed method can be controlled by considering the convergence time of Markov chains. Finally, we show that the proposed score performs at least on par with the commonly used likelihood-based score, but at a substantially reduced computation time in classification of long data sequences.

References

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*. Retrieved from http://dl.acm.org/citation.cfm?id=1882478

Oates, T., Firoiu, L., & Cohen, P. R. (1999). Clustering Time Series with Hidden Markov Models and Dynamic Time Warping. *Proceedings of the IJCAI-99 workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*.

Mouret, M., Solnon, C., & Wolf, C. (2009). Classification of Images Based on Hidden Markov Models. *2009 Seventh International Workshop on Content-Based Multimedia Indexing*. doi:10.1109/CBMI.2009.22

Wong, W. & Stamp, M. (2006). Hunting for metamorphic engines. *Journal in Computer Virology*, *2*(3), 211–229. doi:10.1007/s11416-006-0028-7

Wang, L., Mehrabi, M. G., & Kannatey-Asibu, E. (2002). Hidden Markov Model-based Tool Wear Monitoring in Turning. *Journal of Manufacturing Science and Engineering*, *124*(3), 651. doi:10.1115/1.1475320

Bicego, M., Murino, V., & Figueiredo, M. a. T. (2004). Similarity-based classification of sequences using hidden Markov models. *Pattern Recognition*, *37*(12), 2281–2291. doi:10.1016/j.patcog.2004.04.005

García-García, D., Emilio, H. P., & Díaz-de-María, F. (2009). A new distance measure for model-based sequence clustering. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *31*(7), 1325–1331. Retrieved from http://e-archivo.uc3m.es/handle/10016/8978

Hsu, D., Kakade, S. M., & Zhang, T. (2012). A Spectral Algorithm for Learning Hidden Markov Models. *Journal of Computer and System Sciences*, *78*(5), 1460–1480. arXiv: arXiv:0811.4413v6

Anandkumar, A., Hsu, D., & Kakade, S. (2012). A method of moments for mixture models and hidden Markov models. *JMLR: Workshop and Conference Proceedings*, *23*, 1–31. arXiv: arXiv:1203.0683v3. Retrieved from http://arxiv.org/abs/1203.0683

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(3), 179–195.

Lindsay, B. G. (1988). Composite Likelihood Methods. *Contemporary mathematics*, *80*(1), 221–239.

Durrett, R. (2007). *Random graph dynamics*. Cambridge University Press. Retrieved from http://www.math.duke.edu/~rtd/RGD/RGD.pdf

Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The annals of applied probability*, 62–87. Retrieved from http://www.jstor.org/stable/2959625

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, *41*(1), 164–174. doi:10.1214/09-STS284

Dempster, a. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. doi:10.1.1.133.4884

Anandkumar, A., Ge, R., & Hsu, D. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine ...* *15*, 2773–2832. Retrieved from http://dl.acm.org/citation.cfm?id=2697055

Troelsgaard, R. & Hansen, L. K. (2016). Spectral Learning of Hidden Markov Models in Non-stationary Data. *[submitted]*.

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (April), 24–26. Retrieved from http://www.i6doc.com/en/livre/?GCOI=28001100131010

Elkan, C. (2003). Using the Triangle Inequality to Accelerate -Means. In *International conference on machine learning (icml)* (pp. 147–153).

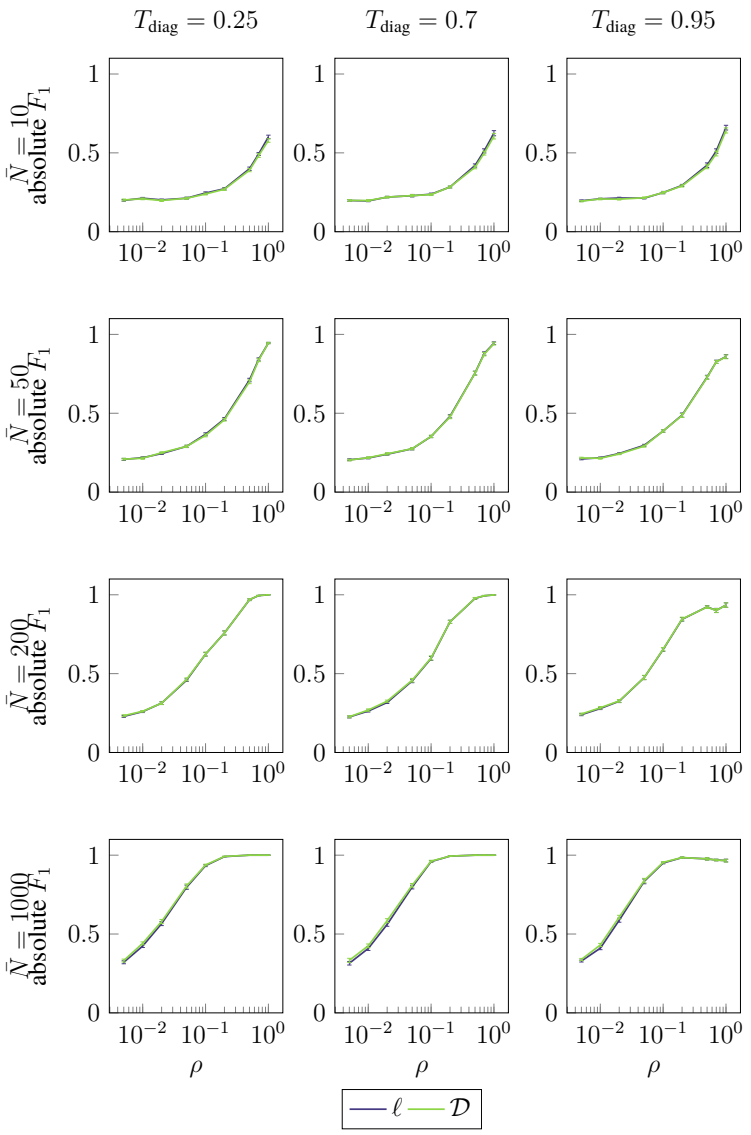


Figure 3: This figure shows how the performances of ℓ and $\mathcal{D}$ vary for different amounts of diagonality, and the parameter ρ. The results are reported in terms of F_1 using $\varepsilon = 0.001$ in the calculation of the bound on the convergence time.

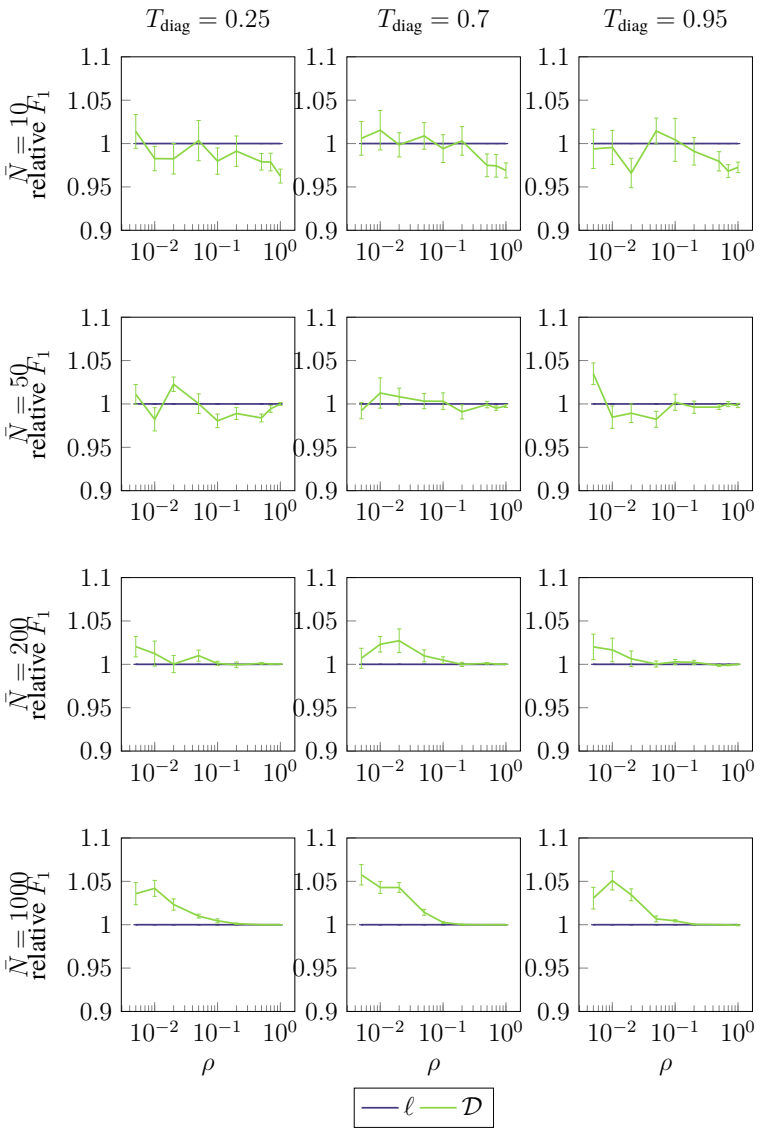


Figure 4: This figure shows how the relative performances of ℓ and $\mathcal{D}$ vary for different amounts of diagonality, and the parameter ρ. See Fig. 3 for absolute performance. The results are reported in terms of F_1 relative to the score of ℓ. In the calculation of the bound on the convergence time we set $\varepsilon = 0.001$

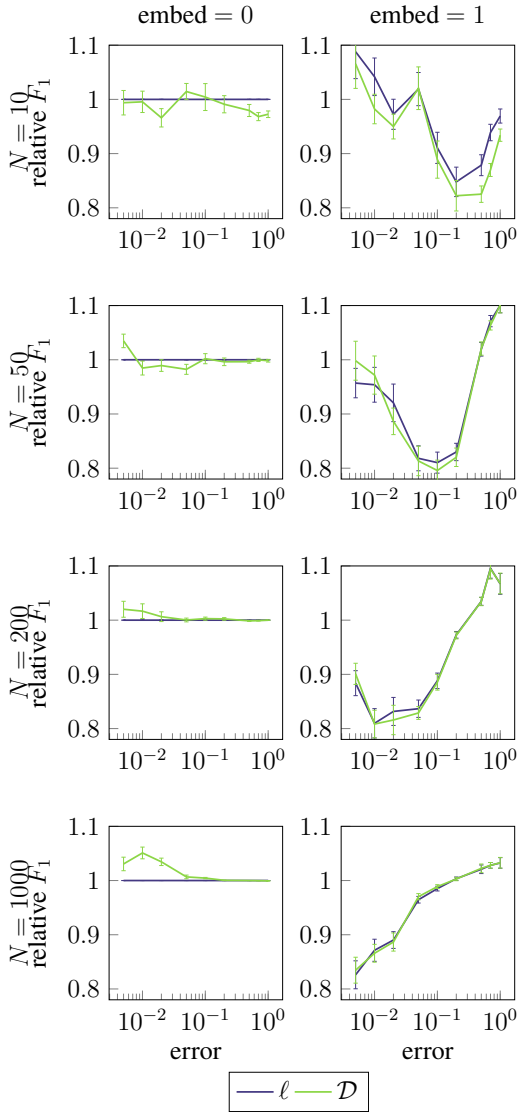


Figure 5: Relative performance when using sequence embedding for classification. The presented results are conditioned on $\varepsilon = 0.001$ and $T_{\text{diag}} = 0.95$.

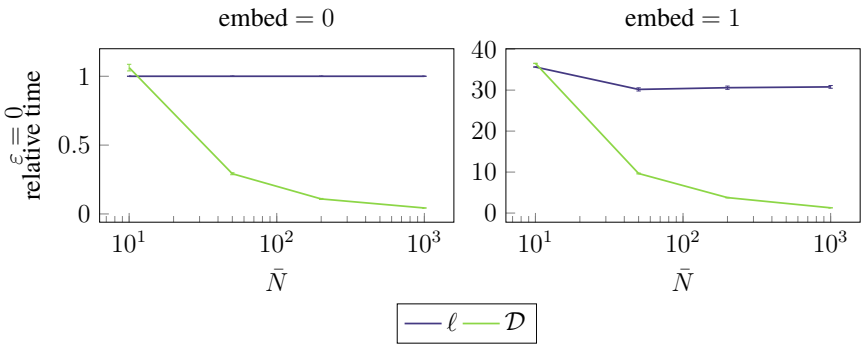


Figure 6: Mean timing factor (relative to ℓ). For long sequences, the proposed composite likelihood based method is superior to the log-likelihood calculations. It is also evident that the embedding procedure is quite costly because of the higher number of model estimations and cost evaluations.

A Construction of Synthetic HMM Classification Problems

Each column in the transition matrices are constructed by a single draw from a Dirichlet distribution with base measure $\boldsymbol{\alpha} = \sum_{j=1}^{S} \alpha_j$ and concentration parameter σ. To be able to control the diagonal structure of the transition matrices, the distribution of the i^{th} column, $\boldsymbol{T}_i$, is sampled from Dirichlet distribution with base measure given by

$$\alpha_j = \begin{cases} T_{\text{diag}} & j = i \\ \frac{1-T_{\text{diag}}}{S-1} & j \neq i \end{cases}$$

where $T_{\text{diag}} \in]0, 1[$

The interpolation parameter $\rho \in]0, 1]$ controls the variance of the simulated multinomial elements. We let ρ determine the relative size of the variance to a maximum variance v_{max} which is determined by a given minimal concentration parameter σ_{min}.

Even though the variances of diagonal- and off-diagonal elements in general are different, the relation between σ and ρ is independent of the base measure, and is given by

$$\sigma = \frac{\sigma_{\text{min}} + 1}{\rho} - 1$$

In the current experiment, the columns of $\boldsymbol{T}$ are generated using $\sigma_{\text{min}} = S$. Hence, the set of most unrelated models are drawn using $\sigma = S$ which is obtained by setting $\rho = 1$, and for $\rho \to 0$, the Dirichlet distribution becomes the Dirac delta function: $\delta(\boldsymbol{\alpha})$

B Embedding Algorithm

Algorithm 1 Classification via sequence embedding

Input: Training and test sequences: $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_{\text{train}}}\}$ and $\{\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_{N_{\text{test}}}\}$. Training labels: $(y_1, y_2, \cdots, y_{N_{\text{train}}})$. Distance score function $D(\boldsymbol{x}, \mathcal{M})$ relating a sequence $\boldsymbol{x}$ to a model $\mathcal{M}$. Classification algorithm C.

Output: Test labels: $(\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_{N_{\text{test}}})$

---

for $i = 1$ **to** N_{train} **do**

  Estimate an HMM $\mathcal{M}_i$ from $\boldsymbol{x}_i$

end for

for $i = 1$ **to** N_{train} **do**

  for $j = 1$ **to** N_{train} **do**

    Calculate $D(\boldsymbol{x}_i, \mathcal{M}_j)$

  end for

end for

Train C using the distance scores of the training sequences for all estimated models as features and $(y_1, y_2, \cdots, y_{N_{\text{train}}})$ as labels

for $i = 1$ **to** N_{test} **do**

  for $j = 1$ **to** N_{train} **do**

    Calculate $D(\bar{\boldsymbol{x}}_i, \mathcal{M}_j)$

  end for

  $\bar{y}_i = C(\bar{\boldsymbol{x}}_i)$

end for

Acknowledgements

This work was supported in part by the Innovation Fund Denmark under the CoSound project, case number 0603-00475B. This publication only reflects the authors' views.

Bibliography

[1] Rasmus Troelsgaard and Lars Kai Hansen. "Spectral Learning of Hidden Markov Models in Non-stationary Data". In: *[submitted]* (2016).

[2] Rasmus Troelsgaard and Lars Kai Hansen. "Sequence Classification Using Third Order Moments". In: *[submitted]* (2016).

[3] Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen. "Towards a universal representation for audio information retrieval and analysis". In: *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2013. DOI: 10.1109/ICASSP.2013.6638242.

[4] Rasmus Troelsgaard, Bjørn Sand Jensen, and Lars Kai Hansen. "A Topic Model Approach to Multi-Modal Similarity". In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. 2013, pp. 1–5. arXiv: arXiv:1405.6886v1.

[5] Morten Hertzum, Haakon Lund, and Rasmus Troelsgaard. "Retrieving Radio News Broadcasts in Danish: Accuracy and Categorization of Unrecognized Words". In: *28th Australian Conference on Human-Computer Interaction (OzCHI)*. 2016, pp. 1–4.

[6] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. "Content-based image retrieval at the end of the early years". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.12 (2000), pp. 1349–1380. ISSN: 01628828. DOI: 10.1109/34.895972. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=895972.

[7] J. Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. "Collaborative Filtering Recommender Systems". In: *The Adaptive Web* (2007), pp. 291–324. ISSN: 1551-3955. DOI: 10.1007/978-3-540-72079-9_9.

arXiv: 1202.1112. URL: http://www.springerlink.com/content/t87386742n752843.

[8] M.a. Bartsch and G.H. Wakefield. "To catch a chorus: using chroma-based representations for audio thumbnailing". In: *Workshop on the Applications of Signal Processing to Audio and Acoustics* October (2001), pp. 15–18. DOI: 10.1109/ASPAA.2001.969531.

[9] Juan P Bello and Jeremy Pickens. "A Robust Mid-level Representation for Harmonic Content in Music Signals". In: *ISMIR*. 2005, pp. 304–311.

[10] Animashree Anandkumar, Daniel Hsu, and SM Kakade. "A method of moments for mixture models and hidden Markov models". In: *JMLR: Workshop and Conference Proceedings* 23 (2012), pp. 1–31. arXiv: arXiv:1203.0683v3. URL: http://arxiv.org/abs/1203.0683.

[11] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. "Linguistic alignment between people and computers". In: *Journal of Pragmatics* 42.9 (2010), pp. 2355–2368. ISSN: 03782166. DOI: 10.1016/j.pragma.2009.12.012.

[12] Chang Xu, Dacheng Tao, and Chao Xu. "A Survey on Multi-view Learning". In: (2013), pp. 1–59. arXiv: 1304.5634. URL: http://arxiv.org/abs/1304.5634.

[13] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. *Multimodal fusion for multimedia analysis: A survey*. Vol. 16. 6. 2010, pp. 345–379. ISBN: 0053001001. DOI: 10.1007/s00530-010-0182-0.

[14] Mark Barnard, J-M Odobez, and Samy Bengio. "Multi-modal audio-visual event recognition for football analysis". In: *Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on*. 2003, pp. 469–478.

[15] Lizhong Wu Lizhong Wu, S.L. Oviatt, and P.R. Cohen. "Multimodal integration-a statistical view". In: *IEEE Transactions on Multimedia* 1.4 (1999), pp. 334–341. ISSN: 1520-9210. DOI: 10.1109/6046.807953.

[16] Martin C. Axelsen, Nikolaj Bak, and Lars Kai Hansen. "Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data". In: *Proceedings - 2015 International Workshop on Pattern Recognition in NeuroImaging, PRNI 2015* (2015), pp. 37–40. DOI: 10.1109/PRNI.2015.20.

[17] Nitish Srivastava and Ruslan Salakhutdinov. "Multimodal Learning with Deep Boltzmann Machines". In: *Advances in neural information processing systems (NIPS)* (2012), pp. 2222–2230. ISSN: 10495258. DOI: 10.1109/CVPR.2013.49.

[18] Dimitrios Giannoulis, Emmanouil Benetos, Anssi Klapuri, and Mark D Plumbley. "Improving instrument recognition in polyphonic music through system integration". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 5222–5226. ISBN: 9780754679707. DOI: 10.1016/j.cpr.2010.02.004.

[19] Katherine Ellis, Emanuele Coviello, Antoni B. Chan, and Gert Lanckriet. "A bag of systems representation for music auto-tagging". In: *IEEE Transactions on Audio, Speech and Language Processing* 21.12 (2013), pp. 2554–2569. ISSN: 15587916. DOI: 10.1109/TASL.2013.2279318.

[20] David M Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet allocation". In: *The Journal of Machine Learning Research* 3 (2003), pp. 993–1022. URL: http://www.mendeley.com/research/latent-dirichlet-allocation-1/.

[21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. "Multimodal Deep Learning". In: *Proceedings of The 28th International Conference on Machine Learning (ICML)* (2011), pp. 689–696. DOI: 10.1145/2647868.2654931. arXiv: 1502.07209.

[22] Jernimo Arenas-García, Anders Meng, K. B. Petersen, T. L. Schiøler, Lars Kai Hansen, and Jan Larsen. "Unveiling music structure via plsa similarity fusion". In: *IEEE International Workshop on IEEE International Workshop on Machine Learning for Signal Processing*. Ed. by Konstantinos Diamantaras, Tülay Adali, Ioannis Pitas, Jan Larsen, Theophilos Papadimitriou, and Scott Douglas. 2007, pp. 419–424. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4414343.

[23] Jason Weston, Samy Bengio, Philippe Hamel, and L G May. "Large-Scale Music Annotation and Retrieval: Learning to Rank in Joint Semantic Spaces". In: *CoRR* abs/1105.5 (2011). arXiv: arXiv:1105.5196v1.

[24] Brian McFee and Gert Lanckriet. "Learning multi-modal similarity". In: *The Journal of Machine Learning Research* 12 (2011), pp. 491–523. URL: http://dl.acm.org/citation.cfm?id=1953063.

[25] Jesper Højvang Jensen, Mads G. Christensen, Daniel P W Ellis, and Søren Holdt Jensen. "A tempo-insensitive distance measure for cover song identification based on chroma features". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2008), pp. 2209–2212. ISSN: 15206149. DOI: 10.1109/ICASSP.2008.4518083.

[26] Y LeCun, U Muller, J Ben, E Cosatto, and B Flepp. "Off-road obstacle avoidance through end-to-end learning". In: *Advances in neural information processing systems* 18 (2006), p. 739. ISSN: 1049-5258. URL: http://books.nips.cc/papers/files/nips18/NIPS2005_0742.pdf?q=07751.

[27] D Mimno and A McCallum. "Topic models conditioned on arbitrary features with dirichlet-multinomial regression". In: *Uncertainty in Artificial Intelligence*. 2008, pp. 411–418.

[28] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. "Probabilistic author-topic models for information discovery". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '04. New York, NY, USA: ACM, 2004, pp. 306–315. ISBN: 1-58113-888-1. DOI: 10.1145/1014052.1014087. URL: http://doi.acm.org/10.1145/1014052.1014087.

[29] Xuerui Wang and Andrew McCallum. "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends". In: (2006), pp. 424–433. ISSN: 1595933395. DOI: 10.1145/1150402.1150450. URL: http://people.cs.umass.edu/$%5Csim$mccallum/papers/tot-kdd06.pdf.

[30] Maxim Rabinovich and David Blei. "The Inverse Regression Topic Model". In: *Proceedings of The 31st International Conference on Machine Learning* 32 (2014), pp. 199–207.

[31] DM Blei and MI Jordan. "Modeling annotated data". In: *Proceedings of the 26th annual international ACM ...* 2003, pp. 127–134. ISBN: 1581136463. URL: http://dl.acm.org/citation.cfm?id=860460.

[32] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. "Polylingual topic models". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Vol. 2. 2009. URL: http://dl.acm.org/citation.cfm?id=1699627.

[33] D Putthividhya, H T Attias, S S Nagarajan, and T W Lee. "Probabilistic Graphical Model for Auto-Annotation , Content-Based Retrieval , and Classification of TV Clips Containing Audio , Video , and Text". In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. 2007, pp. II–789–792.

[34] Seppo Virtanen, Yangqing Jia, Arto Klami, and Trevor Darrell. "Factorized Multi-Modal Topic Model". In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2012, pp. 843–851. URL: http://www.auai.org/uai2012/papers/272.pdf.

[35] David Blei and John Lafferty. "Correlated Topic Models". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y Weiss, B Schölkopf, and J Platt. Cambridge, MA: MIT Press, 2006, pp. 147–154. URL: http://books.nips.cc/papers/files/nips18/NIPS2005_0774.pdf.

[36] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Sharing Clusters Among Related Groups : Hierarchical Dirichlet Processes". In: *Advances in Neural Information Processing Systems* 1 (2005), pp. 1385–1392. ISSN: 0162-1459. DOI: 10.1198/016214506000000302.

[37] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. "Learning cross-modality similarity for multinomial data". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. Ieee, Nov. 2011, pp. 2407–2414. ISBN: 978-1-4577-1102-2. DOI: 10.1109/ICCV.2011.6126524. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6126524.

[38] Hanna M Wallach. "Structured Topic Models for Language". PhD thesis. University of Cambridge, 2008. URL: http://people.cs.umass.edu/$%5Csim$wallach/theses/wallach_phd_thesis.pdf.

[39] Hanna Wallach, David Mimno, and Andrew McCallum. "Rethinking LDA: Why Priors Matter". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y Bengio, D Schuurmans, J Lafferty, C K I Williams, and A Culotta. Curran Associates, Inc., 2009, pp. 1973–1981. URL: http://dblp.uni-trier.de/db/conf/nips/nips2009.html#WallachMM09.

[40] Thomas P Minka. *Estimating a Dirichlet distribution*. Internet manuscript 8. 2012, pp. 1–14. URL: http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/.

[41] Thomas Minka and John Lafferty. "Expectation-Propagation for the Generative Aspect Model". In: *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. 2002, pp. 352–359. URL: http://research.microsoft.com/en-us/um/people/minka/papers/aspect/minka-aspect.pdf%20http://dl.acm.org/citation.cfm?id=2073918.

[42] Arthur Asuncion, Max Welling, Padhraic Smyth, Yee-Whye Teh, and Asuncion E T Al. "On Smoothing and Inference for Topic Models". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09 Ml. Arlington, Virginia, United States: AUAI Press, 2009, pp. 27–34. ISBN: 978-0-9749039-5-8. URL: http://dl.acm.org.globalproxy.cvt.dk/citation.cfm?id=1795114.1795118.

[43] David Mimno, M Hoffman, and D Blei. "Sparse stochastic inference for latent Dirichlet allocation". In: *arXiv preprint arXiv:1206.6425* (2012). URL: http://arxiv.org/abs/1206.6425.

[44] Max Welling, YW Teh, and H Kappen. "Hybrid variational/Gibbs collapsed inference in topic models". In: *CoRR* abs/1206.3 (2012). URL: http://arxiv.org/abs/1206.3297.

[45] James Foulds, Levi Boyles, Christopher Dubois, Padhraic Smyth, and Max Welling. "Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 446–454. arXiv: arXiv:1305.2452v1. URL: http://arxiv.org/abs/1305.2452.

[46] Matthew Hoffman, David M Blei, and Francis Bach. "Online Learning for Latent Dirichlet Allocation". In: *NIPS*. 2010. URL: http://www.cs.princeton.edu/$%5Csim$blei/papers/HoffmanBleiBach2010b.pdf.

[47] KR R Canini, Lei Shi, and TL L Griffiths. "Online inference of topics with latent Dirichlet allocation". In: *Proceeings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by David van Dyk and Max Welling. Vol. 5. 2009. 2009, pp. 65–72. URL: http://cocosci.berkeley.edu/tom/papers/topicpf.pdf.

[48] Anima Anandkumar, Yi-kai Liu, and DJ Hsu. "A spectral algorithm for latent dirichlet allocation". In: *Advances in Neural ...* (2012), pp. 1–9. URL: http://papers.nips.cc/paper/4637-a-spectral-algorithm-for-latent-dirichlet-allocation.

[49] Bo Dai, Niao He, Hanjun Dai, and Le Song. "Scalable Bayesian Inference via Particle Mirror Descent". In: (2015). arXiv: arXiv:1506.03101v1.

[50] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics." In: *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl.October (Apr. 2004), pp. 5228–35. ISSN: 0027-8424. DOI: 10.1073/pnas.0307752101. URL: http://www.pnas.org.globalproxy.cvt.dk/content/101/suppl.1/5228.full.pdf%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=387300&tool=pmcentrez&rendertype=abstract.

[51] Han Xiao and Thomas Stibor. "Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation". In: *Journal of Machine Learning Research - Proceedings Track* 13 (2010), pp. 63–78.

[52] Ian Porteous, David Newman, A Ihler, Arthur Asuncion, P Smyth, and Max Welling. "Fast collapsed gibbs sampling for latent dirichlet allocation". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 569–577. ISBN: 9781605581934.

[53] Limin Yao, David Mimno, and Andrew McCallum. "Efficient methods for topic model inference on streaming document collections". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '09. New York, NY, USA: ACM, 2009, pp. 937–946. ISBN: 978-1-60558-495-9. DOI: http://doi.acm.org/10.1145/1557019.1557121. URL: http://www.cs.umass.edu/$%5Csim$mimno/papers/fast-topic-model.pdf.

[54] Mirwaes Wahabzada and Kristian Kersting. "Larger residuals, less work: Active document scheduling for latent Dirichlet allocation". In: *Machine Learning and Knowledge Discovery ...* Vol. 1. August 2010. 2011, pp. 1–16. URL: http://www.springerlink.com/index/J23845623H5U6G25.pdf.

[55] David Mimno. "Topic regression". PhD thesis. 2012.

[56] Mikkel N Schmidt, Ole Winther, and Lars Kai Hansen. "Bayesian non-negative matrix factorization". In: *Independent Component Analysis and Signal Separation* (2009), pp. 540–547. ISSN: 0302-9743. DOI: 10.1007/978-3-642-00599-2_68. URL: http://link.springer.com/chapter/10.1007/978-3-642-00599-2_68.

[57] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The million song dataset". In: *the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. 2011. URL: http://academiccommons.columbia.edu/catalog/ac:148381.

[58] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. "Facilitating comprehensive benchmarking experiments on the million song dataset". In: *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*. Ismir. 2012, pp. 469–474.

[59] Kamelia Aryafar and Ali Shokoufandeh. "Multimodal music and lyrics fusion classifier for artist identification". In: *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014* (2014), pp. 506–509. DOI: 10.1109/ICMLA.2014.88.

[60] Sander Dieleman, P Brakel, and Benjamin Schrauwen. "Audio-based music classification with a pretrained convolutional network". In: *... International Society for Music ...* Ismir (2011), pp. 669–674. URL: https://biblio.ugent.be/publication/1989534.

[61] Jean-julien Aucouturier and François Pachet. "Representing Musical Genre : A State of the Art". In: *Journal of New Music Research* 32.February 2015 (2003), pp. 83–93. ISSN: 0929-8215. DOI: 10.1076/jnmr.32.1.83.16801.

[62] Eva Hörster, Rainer Lienhart, and Malcolm Slaney. "Image retrieval on large-scale image databases". In: *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*. New York, New York, USA: ACM Press, 2007, pp. 17–24. ISBN: 9781595937339. DOI: 10.1145/1282280.1282283. URL: http://dl.acm.org/citation.cfm?doid=1282280.1282283%20http://dl.acm.org/citation.cfm?id=1282283.

[63] AJB Chaney and DM Blei. "Visualizing Topic Models." In: *ICWSM*. 2012. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4645/5021.

[64] William G. Roy and Timothy J. Dowd. "What Is Sociological about Music?" In: *Annual Review of Sociology* 36.1 (2010), pp. 183–203. ISSN: 0360-0572. DOI: 10.1146/annurev.soc.012809.102618. arXiv: arXiv:1011.1669v3. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.soc.012809.102618.

[65] T Bertin-Mahieux. "Clustering beat-chroma patterns in a large music database". In: *International Society for Music Information Retrieval Conference*. 2010. URL: http://www.ee.columbia.edu/$%5Csim$ronw/pubs/ismir2010-beatchromapatterns.pdf.

[66] Matthew D Hoffman, David M Blei, and Perry R Cook. "Easy As CBA: A Simple Probabilistic Model for Tagging Music". In: *ISMIR*. 2009, pp. 369–374.

[67] K Seyerlehner, G Widmer, and P Knees. "Frame level audio similarity-a codebook approach". In: *Conference on Digital Audio Effects* (2008), pp. 1–8. URL: https://www.acoustics.hut.fi/dafx08/papers/dafx08_61.pdf.

[68] Stephanie Pancoast and Murat Akbacak. "Bag-of-Audio-Words Approach for Multimedia Event Classification." In: *Interspeech* September (2012), pp. 1–4. URL: http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/interspeech_2012/i12_2105.pdf.

[69] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. URL: http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf.

[70] Haşim Sak, Andrew Senior, and Françoise Beaufays. "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling". In: *Interspeech 2014* September (2014), pp. 338–342. arXiv: arXiv:1402.1128v1. URL: http://193.6.4.39/$%5Csim$czap/letoltes/IS14/IS2014/PDF/AUTHOR/IS141304.PDF.

[71] A. Graves, A.-R. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks". In: 3 (2013), pp. 6645–6649. ISSN: 1520-6149. DOI: 10.1109/ICASSP.2013.6638947. arXiv: arXiv:1303.5778v1. URL: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6638947&url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%3D%26arnumber%3D6638947.

[72] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (2015), pp. 1556–1566. ISSN: 9781941643723. DOI: 10.1515/popets-2015-0023. arXiv: 1503.0075. URL: http://arxiv.org/abs/1503.0075.

[73] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah a Smith. "Transition-Based Dependency Parsing with Stack Long Short-Term Memory". In: *Acl* (2015), pp. 334–343. DOI: 10.3115/v1/P15-1033. arXiv: arXiv:1505.08075v1. URL: http://www.aclweb.org/anthology/P15-1033.

[74] Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series: An Introduction using R*. 2009, p. 265. ISBN: 9781420010893. URL: https://books.google.com/books?hl=es&lr=&id=LDDzvCsdVs8C&pgis=1.

[75] Lawrence Rabiner. "A Tutorial on HMM & Selected Applications in speech Recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[76] Sharon Goldwater and Tom Griffiths. "A fully Bayesian approach to unsupervised part-of-speech tagging". In: *ACL'07 - 45th Annual Meeting of the Association of Computational Linguistics* (2007), pp. 744–751. ISSN: 0736587X. URL: http://www.aclweb.org/anthology/P07-1094.

[77] Rasmus Bonnevie and Lars Kai Hansen. "Fast sampling from a Hidden Markov Model posterior for large data". In: *IEEE International Workshop on Machine Learning for Signal Processing, MLSP* (2014). ISSN: 21610371. DOI: 10.1109/MLSP.2014.6958859.

[78] Karl Pearson. "Contributions to the Mathematical Theory of Evolution". In: *Philosophical Transactions of the Royal Society of London* 185 (1894), pp. 71–110.

[79] Elchanan Mossel and Sébastien Roch. *Learning nonsingular phylogenies and hidden Markov models*. Vol. 16. 2. 2006, pp. 583–614. ISBN: 1050516060. DOI: 10.1214/105051606000000024. arXiv: 0502076 [cs].

[80] Herbert Jaeger and Herbert. "Observable Operator Models for Discrete Stochastic Time Series". In: *Neural Computation* 12.6 (June 2000), pp. 1371–1398. ISSN: 0899-7667. DOI: 10.1162/089976600300015411. URL: http://www.mitpressjournals.org/doi/abs/10.1162/089976600300015411.

[81] Daniel Hsu, Sham M Kakade, and Tong Zhang. "A Spectral Algorithm for Learning Hidden Markov Models". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1460–1480. arXiv: arXiv:0811.4413v6.

[82] Animashree Anandkumar, Rong Ge, and Daniel Hsu. "Tensor decompositions for learning latent variable models". In: *The Journal of Machine ...* 15 (2014), pp. 2773–2832. URL: http://dl.acm.org/citation.cfm?id=2697055.

[83] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. "Smoothed analysis of tensor decompositions". In: *Proceedings of the 46th ...* (2014), pp. 594–603. URL: http://dl.acm.org/citation.cfm?id=2591881.

[84] SM Siddiqi, Byron Boots, and GJ Gordon. "Reduced-rank hidden Markov models". In: *Thirteenth International Conference on Artificial Intelligence and Statistics May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy*. Vol. 9. 2010, pp. 741–748. URL: http://arxiv.org/abs/0910.0902.

[85] C Subakan, Johannes Traa, and Paris Smaragdis. "Spectral Learning of Mixture of Hidden Markov Models". In: *Advances in Neural Information ...* 2014, pp. 1–9. URL: http://papers.nips.cc/paper/5518-spectral-learning-of-mixture-of-hidden-markov-models.

[86] James Y Zou, Daniel Hsu, David C Parkes, and Ryan P Adams. "Contrastive Learning Using Spectral Methods". In: *Advances in Neural Information Processing Systems 26* (2013), pp. 2238–2246. URL: http://papers.nips.cc/paper/5007-contrastive-learning-using-spectral-methods.pdf.

[87] Lawrence R Rabiner, CH Lee, BH Juang, and JG Wilpon. "Hmm clustering for connected word recognition.pdf". In: *International Conference on Acoustics, Speech, and Signal Processing. ICASSP-89*. 1989.

[88] Tim Oates, Laura Firoiu, and Paul R Cohen. "Clustering Time Series with Hidden Markov Models and Dynamic Time Warping". In: *Proceedings of the IJCAI-99 workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning* (1999).

[89] Padhraic Smyth. "Clustering sequences with hidden Markov models". In: *Advances in neural information processing systems* (1997). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.3648&rep=rep1&type=pdf.

[90] Manuele Bicego, Vittorio Murino, and Mário A. T. Figueiredo. "Similarity-Based Clustering of Sequences Using Hidden Markov Models". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. 2003, pp. 86–95. ISBN: 3-540-40504-6. DOI: 10.1007/3-540-45065-3_8. URL: http://link.springer.com/chapter/10.1007/3-540-45065-3_8%5Cnhttp://link.springer.com/10.1007/3-540-45065-3_8.

[91] Tony Jebara, Yingbo Song, and Kapil Thadani. "Spectral Clustering and Embedding with Hidden Markov Models". In: *18th European Conference on Machine Learning, ECML 2007, Proceedings of* 4701 (2007), pp. 164–175. ISSN: 03029743. DOI: 10.1007/978-3-540-74958-5_18.

[92] Wing Wong and Mark Stamp. "Hunting for metamorphic engines". In: *Journal in Computer Virology* 2.3 (2006), pp. 211–229. ISSN: 17729890. DOI: 10.1007/s11416-006-0028-7.

[93] Manuele Bicego and Vittorio Murino. "Investigating Hidden Markov Models' Capabilities in 2D Shape Classification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2 (2004), pp. 281–286. ISSN: 01628828. DOI: 10.1109/TPAMI.2004.1262200.

[94] W Chai and B Vercoe. "Folk music classification using hidden Markov models". In: *Proceedings of International Conference on Artificial Intelligence*. Vol. 6. 6.4. 2001. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.206&rep=rep1&type=pdf.

[95] Bruce G. Lindsay. "Composite Likelihood Methods". In: *Contemporary mathematics* 80.1 (1988), pp. 221–239.

[96] Julian Besag. "Statistical Analysis of Non-Lattice Data". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 24.3 (1975), pp. 179–195.