Supplementary Material

for the project *Unsupervised structure analysis of an audio file using Hierarchical Dirichlet Process Hidden Markov Models*

Jose Luis Diez Antich
Mattia Paterna

Aalborg University Copenhagen

Table of Contents

Introduction

This report works as a complementary material for the project Unsupervised structure analysis of an audio file using Hierarchical Dirichlet Process Hidden Markov Models [1]. The purpose of this report is to explain basic concepts that were not covered in the project.

1 Basics

Random Variable :Uncertain, numerical (i.e., with values in R) quantity.

Probability Distribution : The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.

Stochastic Process : Family of Random Variables[1].

2 Bayesian Probability

Bayesian probability belongs to the category of evidential probabilities; to evaluate the probability of a hypothesis, prior probability is specified, and is then updated to a posterior probability in the light of new, relevant data[2].

Prior Distribution : a prior $p(\theta)$ the probability distribution that would express beliefs about this quantity before some evidence is taken into account. The Prior is combined with the probability distribution of new data to yield the posterior distribution[3].

Likelihood :The likelihood $p(x|\theta)$ of a set of parameter values, θ, given outcomes x, is equal to the probability of those observed outcomes given those parameter values.

Posterior Distribution : Given observed events and a model, it gives the probability of the parameters that may explain the observed data, $p(\theta|x)$.

Conjugate Prior : if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions[4].

3 Bayesian nonparametric models

Model selection is the task of selecting a statistical model from a set of candidate models, given data. Model selection metrics usually include two terms. The first term measures how well the model fits the data. The second term, a complexity penalty, favors simpler models.

[1] `www.ma.utexas.edu/users/gordanz/notes/introduction_to_stochastic_processes.pdf`
[2] `aleph0.clarku.edu/~djoyce/ma218/bayes1.pdf`
[3] `www.stat.columbia.edu/~gelman/research/published/p039-_o.pdf`
[4] `www.people.fas.harvard.edu/~plam/teaching/methods/conjugacy/conjugacy_print.pdf`

Bayesian nonparametric (BNP) models approach to model selection is to fit a single model that can adapt its complexity to the data. BNP models allow the complexity to grow as more data are observed.

Given an observed data set, data analysis is performed by posterior inference, computing the conditional distribution of the hidden variables given the observed data. What distinguishes Bayesian nonparametric models from other Bayesian models is that the hidden structure is assumed to grow with the data. Its complexity, e.g., the number of mixture components or the number of factors, is part of the posterior distribution.Rather than needing to be specified in advance, it is determined as part of analyzing the data [4].

4 Beta distribution

Beta is a distribution over Binomials[5]. The Beta density function is a very versatile way to represent outcomes like proportions or probabilities. It is defined on the continuum between 0 and 1. There are two parameters which work together to determine if the distribution has a mode in the interior of the unit interval and whether it is symmetrical.

4.1 Mathematical Definition

The standard $Beta$ distribution gives the probability density of a value x on the interval (0,1):

$$Beta(\alpha, \beta): \ f(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{1}$$

where $\alpha > 0$, $\beta > 0$ and B is the beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

The Beta function B in the denominator assures that the total area under the density curve equals 1[6].

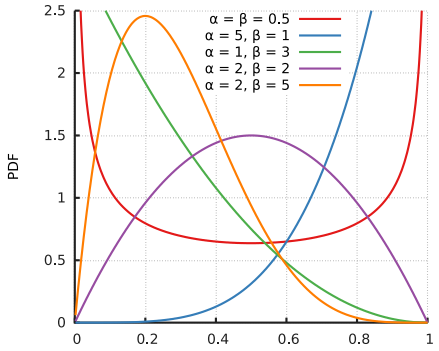
5 Dirichlet distribution

Dirichlet is a distribution over Multinomials. The Dirichlet distribution is a multivariate distribution, meaning that a single outcome is actually a vector of N numbers, a L-tuple. The elements in this vector are are all between 0 and 1. The L values in this vector represent the probabilities of L different mutually exclusive events and they sum to one.

The Dirichlet distribution gives a formula which tells how likely we are to observe a particular L-tuple.

[5] www.itl.nist.gov/div898/handbook/eda/section3/eda366h.htm

[6] pj.freefaculty.org/guides/stat/Distributions/DistributionWriteups



Explanation from [3]: Let $Q = [Q_1, Q_2, ..., Q_k]$ be a random pmf, that is $Q_i \geq 0$ for $i = 1, 2, ..., k$ and $\sum_{i=1}^{k} Q_i = 1$. In addition, suppose that $\alpha = [\alpha_1, \alpha_2, ..., \alpha_k]$, with $\alpha_i > 0 \ \forall i$, and let $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Then, Q has a Dirichlet distribution with parameter α, $Q \sim Dir(\alpha)$, if it has $f(q; \alpha) = 0$ if q is not a pmf, and if q is a pmf then

$$f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1} \tag{2}$$

where $\Gamma(s)$ denotes the gamma function, a generalization of the factorial function.

6 Dirichlet Process

The Dirichlet process is a random distribution whose realizations are distributions over an arbitrary (possibly infinite) sample space [8].

6.1 Dirichlet Process: explain base distribution

The base distribution is basically the mean of the DP [8].

6.2 Dirichlet Process: explain concentration parameter

The concentration parameter can be understood as an inverse variance.The larger α is, the smaller the variance, and the DP will concentrate more of its mass around the mean. The concentration parameter is also called the strength parameter, referring to the strength of the prior when using the DP as a nonparametric prior over distributions in a Bayesian nonparametric model, and the mass parameter, as this prior strength can be measured in units of sample size (or mass) of observations [8].

7 Gibbs Sampling

We could suppose we have random variables $\theta_1, \theta_2, ..., \theta_k$ and we want to draw sample from their joint distribution[7]. Being this much difficult, we could simulate it in a easier way using the conditional distribution

$$P(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_k) \tag{3}$$

for j = 1,...,K

The Gibbs Sampling alternatively generates from each of these distributions and continues for $t = 1, 2, ...$ until the joint distribution of $\theta_1, \theta_2, ..., \theta_k$ does not change. Using a two-variable example, we could use the Gibbs Sampling in finding the joint probability $P(x, y)$ using a recursive process such as

- take $P(x|y)$
- take $P(y|x)$
- repeat until the process stabilizes

It is finally important to remind that this process has to be repeated for each $k = 1, 2, ..., K$ and, for time t, the variables are generated in the following way:

$$P(\theta_j^{(t)}|\theta_1^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_k^{(t-1)}) \tag{4}$$

that is, we consider the variable at previous time index $t-1$ for the variables with index k larger than the considered one.

8 Regularity

The term *regularity* is introduced in [7] to select the clustering level. For each level, a sub-sequence of symbols is defined with the *appropriate symbols*. First, The histogram of the time differences (CIOIH) between all possible combination of two onsets is computed. This histogram defines a harmonic series of peaks that are more or less prominent according to the self-similarity of the sequence. Then, the autocorrelation of the histogram is computed, which has peaks at multiples of the tempo. With this, the regularity is computed.

9 MFCCs

The Mel-Frequency Cepstral Coefficients (MFCCs) are the dominant features used for speech recognition.

As seen in figure 1, the process of creating MFCC consists in five main steps:

1. Divide the signal into frames
2. Obtain the amplitude spectrum
3. Take the logarithm
4. Convert to Mel Spectrum
5. Take the DCT

[7] based on section 8.6 in [2]

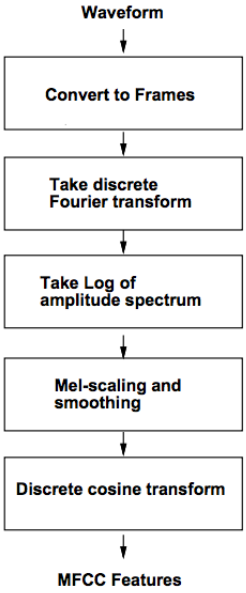


Fig. 1. Taken from [6]

10 From *finite* mixture models to *Hierarchical Dirichlet Process* mixture models

Hierarchical Dirichlet Processes are a particular case of a more general framework based on a Bayesian nonparametric hierarchical approach. The main goal of HDPs are to switch from a classical finite mixture model to an *infinite* mixture model, avoiding any limit in the possible number of clusters. Moreover, the application of HMM to HDPs allows for an unknown number of states, which they are not meant to be kwown in advance. This section aims to explain the basic concepts behind Bayesian hierarchical framework and the transition between finite mixture models (FMMs) and HDPs.

10.1 Bayesian Finite Mixture Models

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulation within an overall population[8]. It can be used, for instance, when a variable that has to be know is taken under two different condition (e.g. distribution of heights in population of adults reflects the mixture of males and females). A finite mixture model can be expressed as

[8] This and the following about Bayesian hierarchical nonparametric approach closely follows [5].

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \tag{5}$$

where G is an underlying measure of the sum of the atom at position ϕ_k and δ_{ϕ_k} is the dirac delta for that location. It is possible to define the process of obtaining a sample x_i from a finite mixture model as follows: for $i = 1, ..., n$

$$\theta_i \sim G \tag{6}$$

$$x_i \sim p(\cdot|\theta_i) \tag{7}$$

where each θ_i is equal to one of the underlying ϕ_k. Therefore, the subset mapped to ϕ_k is the k-th cluster. A graphical representation of a finite mixture model is shown in fig. 2.

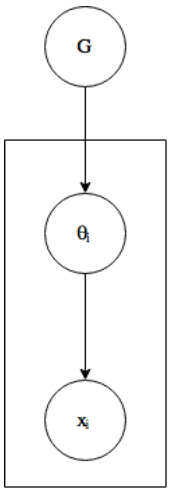


Fig. 2. Finite mixture model

The *Bayesian* approach places a prior on π and ϕ parameters. Considering the latter, such a choice of the prior is model-specific and usually denoted as G_0, while the mixing proportions π is given a Dirichlet prior, which could be described as $Dir(\alpha_0/K, ..., \alpha_0/K)$. This scaling shown in the Dirichlet prior gives α_0 parameter the characteristic of a concentration parameter. Moreover, being the prior symmetric we could change labels of the mixture components without affecting the model. In the cluster assignment, it can be seen how G now is a *random* measure. A graphical representation is shown in fig. 3.

Still, the number of mixtures K has to be inferred. It can be useful therefore put a prior over K and use Bayesian methods. In this approach, Dirichlet process and the Chinese restaurant process provide a nonparametric alternative.

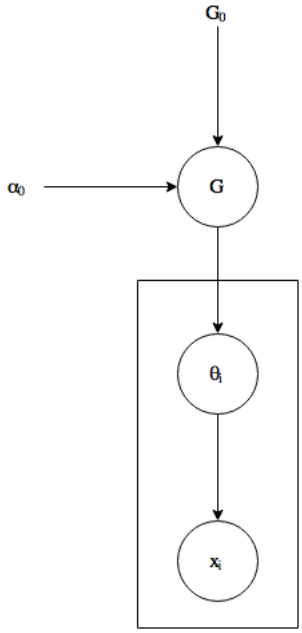


Fig. 3. Bayesian finite mixture model

10.2 Toward a nonparametric approach

A good perspective in defining an infinite mixture model takes into consideration some random processes. First, we could define an infinite sequence of Beta random variables, that is coming from a Beta distribution

$$\beta_k \sim Beta(1, \alpha_0) \tag{8}$$

for $k = 1, 2, ...$ We could then define an infinite sequence of mixing proportion π_k as:

$$\pi_1 = \beta_1 \tag{9}$$

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \tag{10}$$

That means we are breaking off portions of a stick. This is exactly the *stick-breaking process.*

It could be also demonstrated from the equations above that

$$\sum_{k=1}^{\infty} \pi_k = 1 \tag{11}$$

as a consequence of

$$1 - \sum_{k=1}^{K} \pi_k = \prod_{k=1}^{K} (1 - \beta_k) \tag{12}$$

Moreover, convergence is guaranteed

$$\prod_{k=1}^{K} (1 - \beta_k) \to 0 \tag{13}$$

for $K \to \infty$.

Now it is possible to write the equation for an infinite mixture model. G has a clear definition as an *infinite* random measure since the infinite mixing coefficients come π come from a Beta distribution. A graphical representation is shown in fig. 4.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \tag{14}$$

10.3 Hierarchical Bayesian Modeling

As said before, in the mixture model setting θ_i is the parameter associated with the ith data point x_i and it is nor directly observed. The Dirichlet process induces a prior over θ_i, and then the model could be completed introducing a likelihood as in finite mixture model. In this model, sampling from G_0 is a stick-breaking process alike, and from repeated sampling we get the random measure G. This model is named Dirichlet process mixture model, or *infinite* mixture model. A graphical representation is shown in fig. 5.

However, maximum likelihood can be replaced by a hierarchical Bayesian model. In a Bayesian sense, a *hierarchy* means a model in which parameters are treated as random variables. Such a model provides some advantages:

- representations at multiple level of abstraction
- related quantities are linked to each other
- hypotheses space at several levels of abstraction given by *over hypothesis*

A representation of this model is shown in fig. 6

So, it is possible combining both the hierarchical approach and our *infinite* mixture model. A representation of the new model is given in fig. 7

The Bayesian inference coming from this overall model yields to *shrinkage*, i.e. the posterior mean for each θ_k combines date from all the groups. This allows for having less biased data and statistics. Moreover, the parameters θ_i are seen as random variables sampled from an underlying variable θ.

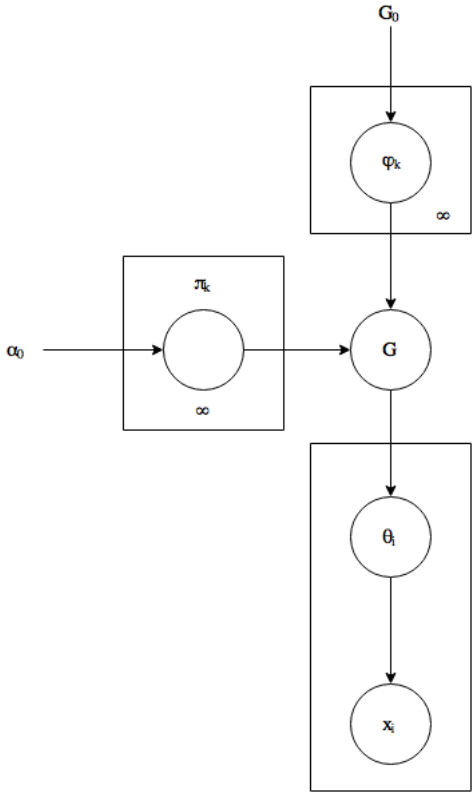


Fig. 4. Bayesian infinite mixture model

10.4 Hierarchical Dirichlet Process Mixtures

In dealing with multiple clustering problems Hierarchical Bayesian methods, and particularly a nonparametric one, could be an optimal solution. It has to be thought what to share in the mixture models and it has to be assumed that we don't know the K_i, i.e. neither the groups nor their number.

We could start from a first draft, that is providing each mixture with its own Dirichlet Process, which are linked to an underlying base distribution G_0. Unfortunately, the atoms in one group will be distinct from the atoms in the other groups, and no sharing is possible at all. This is a consequence of having a *continuous* base distribution. The idea is to let the underlying base distribution be *discrete* and random at the same time.

Doing so, we let the base distribution G_0 be itself a draw from a Dirichlet Process:

$$G_0|\gamma, H \sim DP(\gamma H) \tag{15}$$

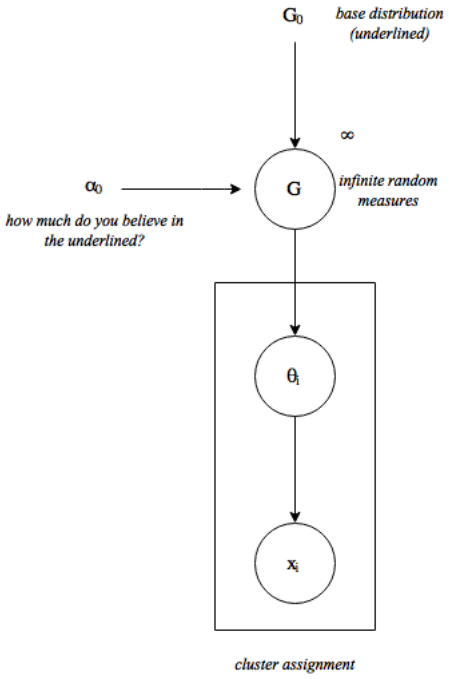


Fig. 5. Dirichlet process, or *infinite* mixture model

Consequently, each measure

$$G_j|\alpha, G_0 \sim DP(\alpha_0 G_0) \tag{16}$$

has as its base measure G_0 a random and *atomic* distribution. That means different samples of G_j will resample from the same atoms, because of their discreteness[9]. Thus, all these atoms are getting shared by everybody inside each mixture. To sum up, we only need to add another level in the Bayesian hierarchical model. What we get is a complete hierarchical Bayesian nonparametric mixture model. We could define it *hierarchical* because all the mixtures participates in the cluster assignment process because of the sharing of the atoms. We could also define it *nonparametric* because we don't have to handle with the base measure parameters (e.g. as it is instead in a Gaussian mixture model, where we have to speculate about μ and σ) since the base measure is *random*. Finally, it is *Bayesian* because we placed a prior on the mixture parameters and made G[10] random. A graphical representation of the complete model is given in fig. 8.

[9] Nevertheless, this is a recursive process

[10] We remind here that G is an underlying measure expressing the underlying distribution of a mixture

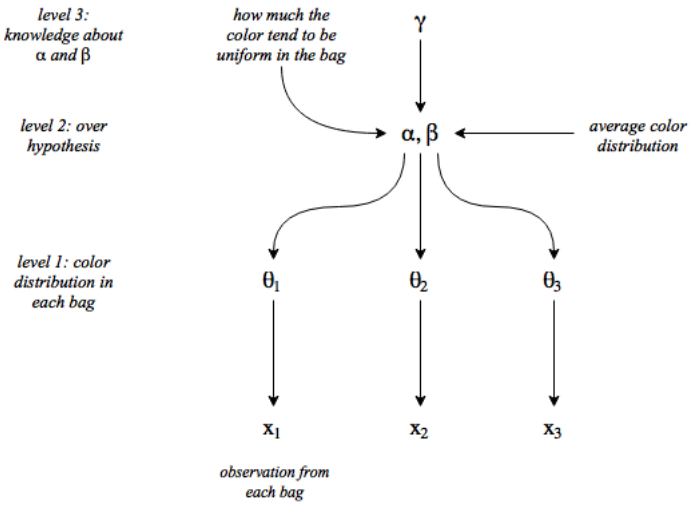


Fig. 6. Example of Hierarchical Bayesian model. In this case, several level of abstraction are given. A bottom level is represented by the distributions in each bag. A mid-level is given by the over hypothesis affecting the distributions mentioned below. The last, top level is represented by γ variable, controlling all the chain.

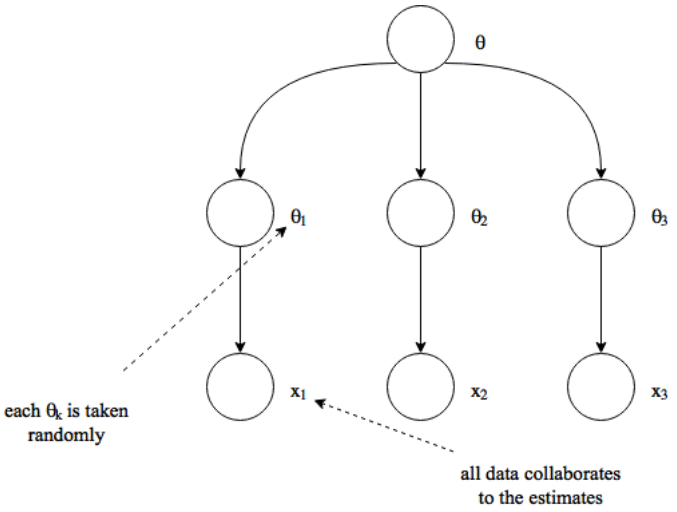


Fig. 7. Hierarchical Bayesian approach. Here, all the parameters θ_k are sampled from the same randon variable θ.

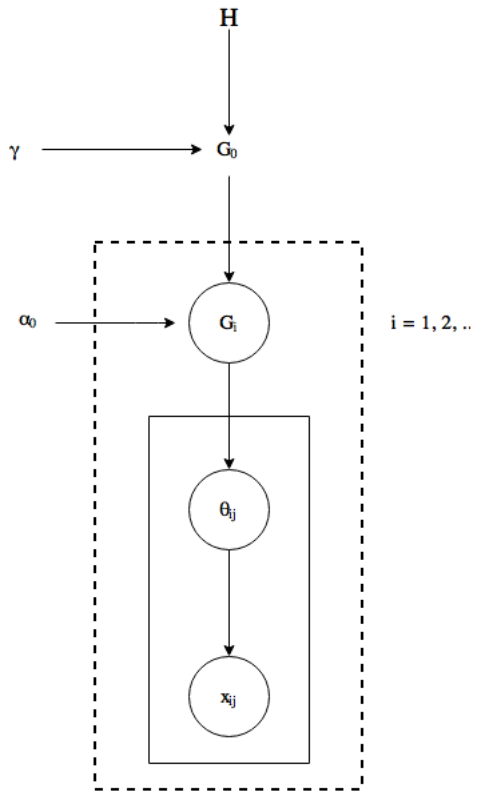


Fig. 8. Hierarchical Dirichlet Process Mixtures.

Bibliography

[1] JL. Diez Antich and Mattia Paterna. Unsupervised structure analysis of an audio file using hierarchical dirichlet process hidden markov models, 2016.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[3] Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, 2010.

[4] Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[5] Michael I Jordan. Dirichlet processes, chinese restaurant processes and all that. In *Tutorial presentation at the NIPS Conference*, 2005.

[6] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[7] Marco Marchini and Hendrik Purwins. Unsupervised analysis and generation of audio percussion sequences. In *International Symposium on Computer Music Modeling and Retrieval*, pages 205–218. Springer, 2010.

[8] Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.