ENSAE 2018/2019
Statistique Bayésienne
Hierarchical Dirichlet Processes

Skander Karkar, Gabriel Romon, Salomé Do, Ulrich Goué

We present hierarchical Dirichlet processes and their application to topic modeling, as presented in article [1]. The other sources are [2, 3]. In section 1, we review the definitions of the Dirichlet distribution and the Dirichlet process, as well as their useful properties, especially when it comes to mixture models. We introduce hierarchical Dirichlet processes in section 2 and show how to sample from their posterior in section 3. In section 4, we present topic modeling as an application of hierarchical Dirichlet processes.

1 Introduction

1.1 The Dirichlet distribution

The ***Dirichlet distribution*** $\texttt{Dir}(k, \alpha_1, ...\alpha_k)$, with each $\alpha_i > 0$, is a probability distribution over the (k-1)-simplex $S_{k-1} = \{x \in \mathbb{R}_+^k \mid x_1 + ... + x_k = 1\}$ such that its density with respect to the Lebesgue measure is

$$f(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

where $B(\alpha)$ is a normalizing constant. Given its support, the Dirichlet distribution can be seen as a law over the set of discrete probabilities over k points. A useful property of the Dirichlet distribution is that

$$\begin{aligned} &\text{If } (X_1, ..., X_k, Y_1, ..., Y_m) \sim \texttt{Dir}(k+m, \alpha_1, ..., \alpha_k, \beta_1, ..., \beta_m) \\ &\text{Then } \frac{1}{X_1 + ... + X_k}(X_1, .., X_k) \sim \texttt{Dir}(k, \alpha_1, ...\alpha_k) \end{aligned} \tag{P}$$

1.2 Mixture models

In a ***mixture model***, each observation X_i belongs to a unique cluster k. Let $L_i = k$ if and only if observation X_i belongs to cluster k and P_k be the distribution of observations from cluster k, then the observations are i.i.d. from the distribution

$$P_X(.) = \mathbb{P}(X \in .) = \sum_{k \in \mathbb{N}} c_k P_k(.)$$

where $c_k = \mathbb{P}(L = k) = \mathbb{P}(X \text{ in cluster } k)$ so necessarily $\sum_k c_k = 1$ and $c_k \geq 0$. We assume that P_k depends on a parameter $\theta_k \in \Theta$, i.e. $P_k = P_{\theta_k}$. If each P_{θ_k} has density $p(.|\theta_k)$ then X has density

$$p(x) = \sum_{k \in \mathbb{N}} c_k p(x|\theta_k) \tag{1}$$

Let ϕ be a discrete probability over Θ, i.e. for a sequence (c_k) in $\mathbb{R}_+$ that sums to 1 and a sequence (θ_k) in Θ

$$\phi(.) = \sum_{k \in \mathbb{N}} c_k \delta_{\theta_k}(.) \tag{2}$$

If these are the same sequences (c_k) and (θ_k) in the mixture density (1), then

$$p(x) = \int_\Theta p(x|\theta) \,\mathrm{d}\phi(\theta) \tag{3}$$

We can see then that given the choice of p, the parameters of the mixture are summarized by a discrete measure ϕ over Θ. The mixture model is finite with K components if we impose that ϕ has K atoms.

A ***Bayesian mixture model*** is therefore a mixture model with a prior distribution on ϕ. In the finite case, ϕ has atoms $(\theta_1, .., \theta_K)$ with associated weights $(c_1, .., c_K)$. The atoms can be i.i.d. from a distribution H on Θ. Since the weight vector $(c_1, .., c_K)$ has to be in S_{K-1}, it can follow a Dirichlet distribution, independently from the atoms.

1.3 The Dirichlet process

To define a prior in the case of an infinite number of clusters, we use the ***Dirichlet process*** (DP). A Dirichlet process is a probability distribution over the space of probability measures on Θ, that induces finite-dimensional Dirichlet distributions when the data are grouped.

Formally, we write $P \sim \mathtt{DP}(\alpha, H)$, with H a probability distribution over Θ and $\alpha > 0$, and say that P (itself a probability distribution over Θ) follows a Dirichlet process with base distribution H and scaling (or concentration) parameter α, if for every finite measurable partition $\{A_1, .., A_r\}$ of Θ,

$$(P(A_1), ..., P(A_r)) \sim \texttt{Dir}(r, \alpha H(A_1), ..., \alpha H(A_r))$$

That a process satisfying this condition consistently for every partition exists has been shown, and the Dirichlet process can for example be constructed from the gamma process, as shown in [4].

An important property of the Dirichlet process is that the probability P sampled from $\texttt{DP}(\alpha, H)$ is almost surely discrete, even when H isn't. This can be seen through the ***Stick-breaking construction*** of the Dirichlet process. If

$$\theta_i \sim H \text{ i.i.d. for } i \geq 1 \tag{4}$$

$$\beta'_i \sim \texttt{Beta}(1, \alpha) \text{ i.i.d. for } i \geq 1 \text{ and independent from } (\theta_i)_i \tag{5}$$

$$\beta_i = \beta'_i \prod_{j=1}^{i-1} (1 - \beta'_i) \text{ for } i \geq 1 \tag{6}$$

$$P = \sum_{k=1}^{\infty} \beta_i \delta_{\theta_i} \tag{7}$$

then $P \sim \texttt{DP}(\alpha, H)$. Note that $(\beta_i)_i$ as defined in (5) and (6) sums to 1 almost surely. We write $\beta = (\beta_i)_i \sim \texttt{Stick}(\alpha)$ for a sequence generated in this manner.

It is well known that if $P \sim \texttt{DP}(\alpha, H)$ and $\theta_1, \theta_2, ...$ are i.i.d. draws from P then

$$\theta_i | (\theta_1, .., \theta_{i-1}, \alpha, H) \sim \sum_{l=1}^{i-1} \frac{1}{i - 1 + \alpha} \delta_{\theta_l} + \frac{\alpha}{i - 1 + \alpha} H \tag{8}$$

This shows that there is a positive reinforcement effect: the more a point is sampled, the more likely it is to be sampled again in the future. Also, we see that a larger α means more new draws from H and therefore a less concentrated (more spread out) process. To make this clearer, define ϕ_1,.., ϕ_K to be the distinct values taken by θ_1,..,θ_{i-1} and $m_k = \texttt{Card}(\{\theta_l = \phi_k : l = 1, ., i-1\})$. Equation (8) becomes

$$\theta_i | (\theta_1, .., \theta_{i-1}, \alpha, H) \sim \sum_{k=1}^{K} \frac{m_k}{i - 1 + \alpha} \delta_{\phi_k} + \frac{\alpha}{i - 1 + \alpha} H \tag{9}$$

A useful representation of the DP is known as the ***Chinese restaurant process***. Imagine a restaurant with an unbounded number of tables. Each θ_i represents the dish that a customer who comes in the restaurant eats. Each table serves a unique dish. The values $\phi_1, .., \phi_K$ represent the distinct dishes served at the occupied tables at the restaurant. The ith customer sits at an already occupied table with probability proportional to the number of customers already sat at it, which means that he eats an already served dish ϕ_k with probability proportional to the number of customers already eating that dish, in which case $\theta_i = \phi_k$. The customer sits at a new table with probability proportional to α and eats a dish $\theta_i \sim H$, and the table serves only that dish in the future. If it is a new dish, it is added to the list of dishes as ϕ_{K+1}.

A Dirichlet process can be used in a Bayesian mixture model as a prior on the discrete probability ϕ that appears in (2) and (3). More precisely, we would then be assuming that the observed X_i are generated in the following manner:

$$\begin{aligned} &G \sim \texttt{DP}(\alpha, H) \\ &\theta_i | G \sim G \\ &X_i | \theta_i \sim P_{\theta_i} \end{aligned}$$

2 Hierarchical Dirichlet processes

As we have seen, the Dirichlet process is useful when a component of our model is a discrete random variable of unbounded cardinality. The ***hierarchical Dirichlet process*** (HDP) is useful when we have groups of data, each incorporating a discrete random variable of unbounded cardinality, and we want to link these variables across groups.

Say we have J groups indexed by $\mathcal{J}$. A hierarchical Dirichlet process is a distribution over a set, indexed by $\mathcal{J}$, of random probability measures over Θ. A random probability measure G_j is sampled from a Dirichlet process for each $j \in \mathcal{J}$. To tie these probability measures, we use the same base distribution, itself sampled from a Dirichlet process. So that:

$$\begin{aligned} &G_0 | (\gamma, H) \sim \texttt{DP}(\gamma, H) \\ &G_j | (\alpha, G_0) \sim \texttt{DP}(\alpha, G_0) \text{ i.i.d. for } j \in \mathcal{J} \end{aligned}$$

The hyperparameters of the HDP are the base distribution H and the scaling parameters γ and α. H provides a prior on the atoms of all the G_j. G_0 varies around H with variability governed by γ. If we want to model different variability around G_0 in the groups, we can use a separate scaling parameter α_j for each group j. Paper [1] uses gamma priors on γ and α.

If for each group j of data we observe X_{ji}, with $i = 1, .., n_j$, then our model for data generation is

$$\begin{aligned} &G_0 | (\gamma, H) \sim \texttt{DP}(\gamma, H) \\ &G_j | (\alpha, G_0) \sim \texttt{DP}(\alpha, G_0) \text{ i.i.d. for } j \in \mathcal{J} \\ &\theta_{ji} | G_j \sim G_j \text{ for } i = 1, .., n_j \text{ for } j \in \mathcal{J} \\ &X_{ji} | \theta_{ji} \sim P_{\theta_{ji}} \end{aligned} \tag{HDP1}$$

From equation (8), we see that the atoms of $G_0 \sim \texttt{DP}(\gamma, H)$ are all atoms of H. And since $G_j \sim \texttt{DP}(\alpha, G_0)$, all G_j inherit their atoms from G_0. We can look deeper into this using the stick-breaking construction of the Dirichlet process. First, we know that

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where $\phi_k \sim H$ i.i.d. and $\beta = (\beta_k)_k \sim \texttt{Stick}(\gamma)$ are independent. Each G_j will also have atoms among $\phi = (\phi_k)_k$ with new weights $\pi_j = (\pi_{j,k})_k$ and can be written

$$G_j = \sum_{k=1}^{\infty} \pi_{j,k} \delta_{\phi_k}$$

Let $(A_1, .., A_r)$ be a measurable partition of Θ and $K_l = \{k : \phi_k \in A_l\}$ for $l = 1, .., r$. $(K_1, .., K_r)$ is then a partition of $\mathbb{N}^*$. If H is non-atomic, then the ϕ_k's are almost surely distinct, and from any partition of $\mathbb{N}^*$ we can find a partition of Θ that respects it. We have then

$$(G_j(A_1), .., G_j(A_r)) \sim \texttt{Dir}(r, \alpha G_0(A_1), .., \alpha G_0(A_r)) \tag{10}$$

which means

$$(\sum_{k \in K_1} \pi_{j,k}, .., \sum_{k \in K_r} \pi_{j,k}) \sim \texttt{Dir}(r, \alpha \sum_{k \in K_1} \beta_k, .., \sum_{k \in K_r} \beta_k) \tag{11}$$

We can write this $\pi_j \sim \texttt{DP}(\alpha, \beta)$, where we see π_j and β as distributions on $\mathbb{N}^*$. Since $\theta_{ji} \sim G_j$, $\theta_{ji} = \phi_k$ with probability $\pi_{j,k}$. If z_{ji} is the indicator variable such that $\theta_{ij} = \phi_{z_{ij}}$, then the model in equations (HDP1) becomes

$$\begin{aligned}
\beta &\sim \texttt{Stick}(\gamma) \\
(\phi_k)_k &\sim H \text{ i.i.d.} \\
\pi_j &\sim \texttt{DP}(\alpha, \beta) \\
z_{ji} &\sim \pi_j \\
\theta_{ij} &= \phi_{z_{ij}} \\
X_{ji} &\sim P_{\theta_{ij}}
\end{aligned} \tag{HDP2}$$

We can find the relation between the new weights $(\pi_{j,k})_k$ and the global weights $(\beta_k)_k$. Consider the partition $(K_1, K_2, K_3) = (\{1, .., k-1\}, \{k\}, \{k, k+1, ...\})$. Then equation (11) becomes

$$(\sum_{l=1}^{k-1} \pi_{j,l}, \pi_{j,k}, \sum_{l=k+1}^{\infty} \pi_{j,l}) \sim \texttt{Dir}(3, \alpha \sum_{l=1}^{k-1} \beta_l, \alpha \beta_k, \alpha \sum_{l=k+1}^{\infty} \beta_l) \tag{12}$$

From the property (P) of the Dirichlet distribution, we have

$$\frac{1}{\pi_{j,k} + \sum\limits_{l=k+1}^{\infty} \pi_{j,l}} (\pi_{j,k}, \sum_{l=k+1}^{\infty} \pi_{j,l}) \sim \texttt{Dir}(2, \alpha \beta_k, \alpha \sum_{l=k+1}^{\infty} \beta_l) \tag{13}$$

Since π_j is a distribution on $\mathbb{N}^*$

$$\pi_{j,k} + \sum_{l=k+1}^{\infty} \pi_{j,l} = 1 - \sum_{l=1}^{k-1} \pi_{j,l}$$

For the same reason

$$\sum_{l=k+1}^{\infty} \beta_l = 1 - \sum_{l=1}^{k} \beta_l$$

Since the Dirichlet distribution with 2 categories is identical to the Beta distribution, and defining

$$\pi'_{j,k} = \frac{\pi_{j,k}}{1 - \sum\limits_{l=1}^{k-1} \pi_{j,l}} \tag{14}$$

(13) becomes

$$\pi'_{j,k} \sim \texttt{Beta}(\alpha\beta_k, \alpha(1 - \sum_{l=1}^{k} \beta_l)) \tag{15}$$

From the definition of $\pi'_{j,k}$ in (14) it is easy to see that

$$\pi_{j,k} = \pi'_{j,k} \prod_{l=1}^{k-1} (1 - \pi'_{j,l}) \tag{16}$$

Indeed,

$$1 - \pi'_{j,l} = \frac{1 - \sum\limits_{h=1}^{l-1} \pi_{j,h} - \pi_{j,l}}{1 - \sum\limits_{h=1}^{l-1} \pi_{j,h}} = \frac{1 - \sum\limits_{h=1}^{l} \pi_{j,h}}{1 - \sum\limits_{h=1}^{l-1} \pi_{j,h}}$$

So that

$$\prod_{l=1}^{k-1} (1 - \pi'_{j,l}) = 1 - \sum_{h=1}^{k-1} \pi_{j,h}$$

because all intermediate terms simplify. Finally

$$\pi'_{j,k} \prod_{l=1}^{k-1} (1 - \pi'_{j,l}) = \frac{\pi_{j,k}}{1 - \sum\limits_{l=1}^{k-1} \pi_{j,l}} (1 - \sum_{h=1}^{k-1} \pi_{j,h}) = \pi_{j,k}$$

Put together, (15) and (16) give a way of constructing the K first new weights $\pi_{j,k}$ of group j for each group, knowing the K first original weights β_k, obtained by stopping the sampling from $\texttt{Stick}(\gamma)$ after K sampled values.

We can extend the Chinese restaurant process to define the ***Chinese restaurant franchise***. This new metaphor will be useful to represent the hierarchical Dirichlet process.

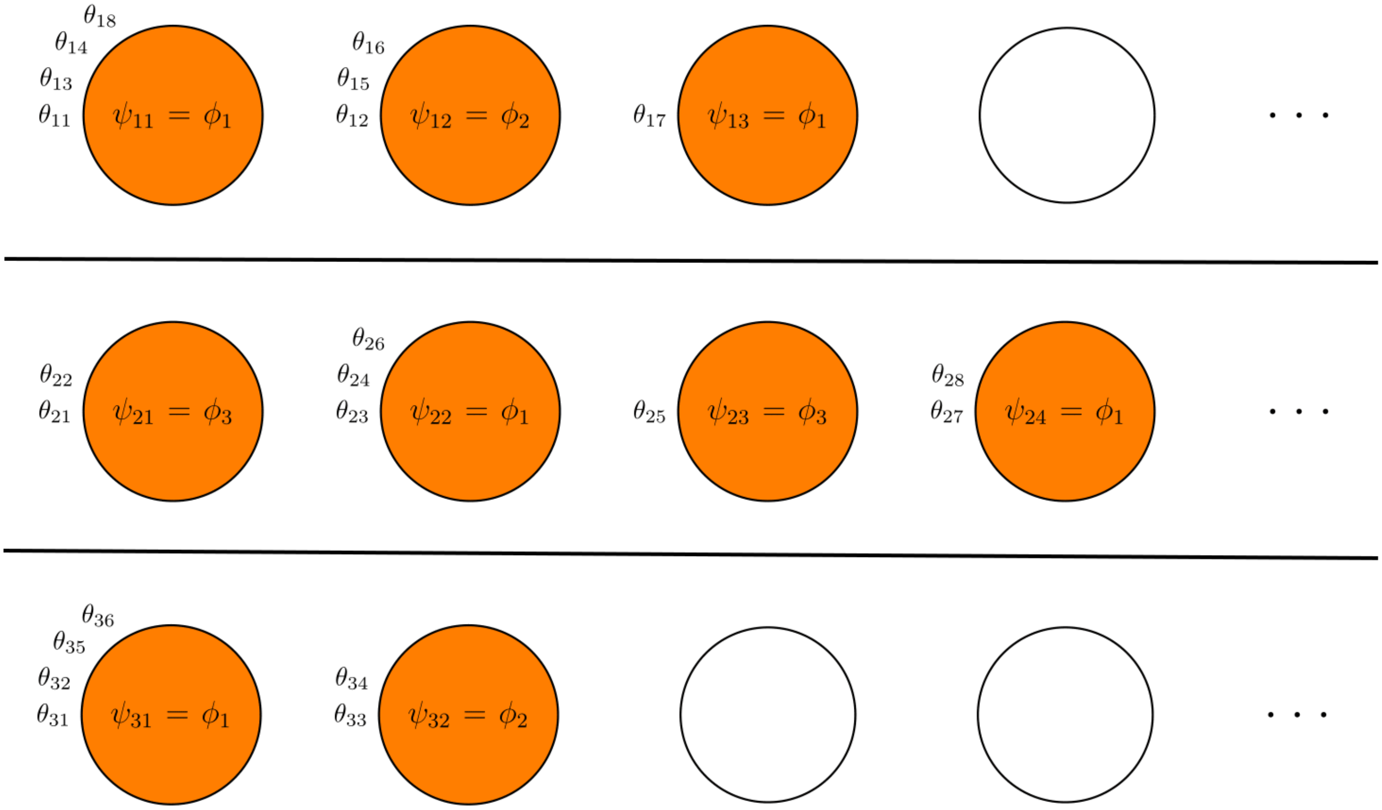

Figure 1: A Chinese restaurant franchise with 3 restaurants.

A restaurant franchise has $\texttt{Card}(\mathcal{J})$ restaurants that serve the same global menu. The restaurants represent the groups j and the global menu represents H. So $\phi_1, \ldots, \phi_K$ which are i.i.d. from H are the distinct dishes being served at one moment in all restaurant. The dish that the ith customer in restaurant j eats represents θ_{ji}, which is decided randomly inside restaurant j according to G_j as in the Chinese restaurant process, independently from other restaurants. Let t_{ji} be the index of the table that customer i in restaurant j sits at. Let ψ_{jt} be the dish served at table t in restaurant j, sampled from G_0, and k_{jt} be the index of that dish in the list $\phi_1, \ldots, \phi_K$ of dishes, so that

$$\psi_{jt} = \phi_{k_{jt}} \tag{17}$$

This means that customer i in restaurant j eats the dish

$$\theta_{ji} = \psi_{jt_{ij}} = \phi_{k_{jt_{ij}}} \tag{18}$$

Equations (8) and (9), along with (HDP1), explain how the franchise works. Customer i in restaurant j follows G_j which itself follows $\texttt{DP}(\alpha, G_0)$ and so samples from the atoms of G_0. The customer either eats a dish already generated by G_j from G_0, or he sits at a new table and a new dish is sampled from G_0. When sampling a dish, since G_0 follows $\texttt{DP}(\gamma, H)$, it either returns one that has already been generated (in any restaurant, since all restaurants call G_0) or generates a new dish according to its base distribution H. The franchise therefore tends to reoffer popular dishes.

To find analogs of the conditional distributions (8) and (9) for a HDP, we need to keep counts of customers and tables. Let n_{jtk} be the number of customers in restaurants j at table i eating the dish indexed by k (they all eat the same dish, but we need this notation), and m_{jk} be the number of tables in restaurant j serving dish k. We denote marginal counts by dots. So

- $n_{jt.}$ is the number of customers in restaurant j at table t
- $n_{j.k}$ is the number of customers in restaurant j eating dish k
- $m_{j.}$ is the number of tables in restaurant j
- $m_{.k}$ is the number of tables across all restaurants serving dish k
- $m_{..}$ is the number of (occupied) tables in all restaurants

With these notations, (9) gives for $\theta_{ji} \sim G_j$, since G_j has base distribution G_0

$$\theta_{ji}|(\theta_{j1}, .., \theta_{j,i-1}, \alpha, G_0) \sim \sum_{t=1}^{m_{j.}} \frac{n_{jt.}}{i-1+\alpha} \delta_{\psi_{jt}} + \frac{\alpha}{i-1+\alpha} G_0 \tag{19}$$

If an old table t from the sum is chosen, we set $t_{ji} = t$ and $\theta_{ji} = \psi_{jt}$ (the dish served at that table) and increment $n_{jt.}$ by one. If a new draw from G_0 is generated, we increment $m_{j.}$ by one and set $n_{jm_{j.}.} = 1$, $\theta_{ji} = \psi_{jm_{j.}} \sim G_0$ and $t_{ji} = m_{j.}$.

Since G_0 has base distribution H, draws of the ψ_{jt} variables from it can also be expressed via (9). Noticing that $m_{..}$ is the number of previous calls to G_0, we need to condition by all previous draws in all restaurants/groups

$$\psi_{jt}|(\psi_{11}, \psi_{12}, .., \psi_{21}, \psi_{22}, .., \psi_{j1}, .., \psi_{j,t-1}, \gamma, H) \sim \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\gamma} \delta_{\phi_k} + \frac{\gamma}{m_{..}+\gamma} H \tag{20}$$

If an already served dish k from the sum is chosen, we set $k_{jt} = k$ and $\psi_{jt} = \phi_k$ and increment $m_{.k}$ by one. If a new draw from H is needed, we increment K by one and set $\psi_{jt} = \phi_K \sim H$ and $k_{jt} = K$ (if H is non-atomic, we almost surely generate a new dish).

We can use equation (19) to sample the θ_{ji}. When a draw from G_0 is needed, we use equation (20). All we need to know are the hyperparameters H, γ and α.

3 Sampling the posterior of a HDP

We present a Gibbs sampler based on the Chinese restaurant franchise representation of the hierarchical Dirichlet process. In Article [1], the authors assume in Section 5 that the values of α and γ are fixed, that H is conjugate to the data distribution P_θ, that P_θ has density $p(.|\theta)$ and H density h.

We denote the observed data by x_{ji}, which we assume are generated independently from $P_{\theta_{ji}}$. We use the same notations as above, adding the variable $z_{ji} = k_{jt_{ji}}$, which represents the mixture component of observation x_{ij}, i.e. the index of the parameter θ_{ij} that generated it in the list of distinct parameters.

We also define the vectors

- All observed data, $\mathbf{x} = (x_{ji} : \forall\ j, i)$
- Customers sat at table t in restaurant j, $\mathbf{x}_{jt} = (x_{ji} : \forall\ i \text{ such that } t_{ji} = t)$
- Customers eating from dish k, $\mathbf{x}_k = (x_{ji} : \forall\ j, i \text{ such that } k_{jt_{ji}} = k)$
- The table indexes of all customers, $\mathbf{t} = (t_{ji} : \forall\ j, i)$
- The dish indexes of all tables, $\mathbf{k} = (k_{jt} : \forall\ j, t)$
- The dish indexes of all customers, $\mathbf{z} = (z_{ji} : \forall\ j, i)$
- Counts of tables serving each dish by restaurant, $\mathbf{m} = (t_{jk} : \forall\ j, k)$
- The menu of distinct dishes, $\boldsymbol{\phi} = (\phi_1, .., \phi_K)$

$\mathbf{x}^{-ji}$ means that the element x_{ji} is eliminated from the vector; likewise for the other vectors. $n_{jt.}^{-ji}$ means that the element is not considered in the count; likewise for other counts.

We want to find the density of x_{ij}, knowing that it comes from mixture component k and knowing all other values in $\mathbf{x}^{-ji}$. First, we have by independence

$$p(x_{ji}, \mathbf{x}^{-ji}|k, \phi_k) \propto p(x_{ji}|\phi_k) \prod_{\substack{(j',i')\neq(j,i) \\ z_{j'i'}=k}} p(x_{j'i'}|\phi_k) \quad \text{(proportional in } x_{ji}) \tag{21}$$

Using the formula $p(a) = \int p(a|b)p(b)\,\mathrm{d}b$, we can integrate ϕ_k out

$$f_k^{-x_{ji}}(x_{ji}) = p(x_{ji}|k, \mathbf{x}^{-ji}) = \frac{\int p(x_{ji}|\phi_k) \prod_{(j',i')\neq(j,i), z_{j'i'}=k} p(x_{j'i'}|\phi_k) h(\phi_k)\,\mathrm{d}\phi_k}{\int \prod_{(j',i')\neq(j,i), z_{j'i'}=k} p(x_{j'i'}|\phi_k) h(\phi_k)\,\mathrm{d}\phi_k} \tag{22}$$

as factors from other mixture components omitted in (21) simplify outside of the integrals.

We will use Gibbs sampling to sample from the posterior of the index variables t_{ji} and k_{jt}. The variable θ_{ji} and ψ_{jt} can be reconstructed using (17) and (18).

We want to sample from the posterior $p(\mathbf{t}, \mathbf{k}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{t}, \mathbf{k})p(\mathbf{t}, \mathbf{k})$. Gibbs sampling is a MCMC algorithm that does this by repeatedly sampling from each coordinate conditionally on the others, starting from an initialization of $(\mathbf{t}, \mathbf{k})$. Updating the vector $(\mathbf{t}, \mathbf{k})$ in this manner gives a sequence approximately distributed according to the posterior. Precisely, we need to sample from

$$p(t_{ji}|\mathbf{x}, \mathbf{t}^{-ji}, \mathbf{k}) \propto p(x_{ji}|t_{ji}, \mathbf{t}^{-ji}, \mathbf{k})p(t_{ij}|\mathbf{t}^{-ji}, \mathbf{k}) \text{ for all } j, i \tag{23}$$

$$p(k_{jt}|\mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt}) \propto p(\mathbf{x}_{jt}|k_{jt}, \mathbf{k}^{-jt}, \mathbf{t})p(k_{jt}|\mathbf{t}, \mathbf{k}^{-jt}) \text{ for all } j, t \tag{24}$$

where we use the fact that t_{ji} only influences x_{ji} and k_{jt} only influences $\mathbf{x}_{jt}$. First, we look at $p(x_{ji}|t_{ji}, \mathbf{t}^{-ji}, \mathbf{k})$. Note that the variables t_{ji} and k_{jt} are exchangeable, as they are inherited from θ_{ji} and ψ_{jt}. We use this to consider any t_{ji} as the last table being sampled in (19) and (20). If t_{ji} is equal to an already used t then the likelihood of observing x_{ji} is

$$p(x_{ji}|t_{ji} = t, \mathbf{t}^{-ji}, \mathbf{k}) = p(x_{ji}|k_{jt}, \mathbf{x}^{-ji}) = f_{k_{jt}}^{-x_{ji}}(x_{ji})$$

which we found in (22). The likelihood of observing x_{ji} from a new table $t_{\text{new}} = m_{j.} + 1$ is given by (20)

$$p(x_{ji}|t_{ji} = t_{\text{new}}, \mathbf{t}^{-ji}, \mathbf{k}) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \gamma} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{m_{..} + \gamma} f_{k_{\text{new}}}^{-x_{ji}}(x_{ji}) \tag{25}$$

where $f_{k_{\text{new}}}^{-x_{ji}}(x_{ji}) = \int p(x_{ji}|\phi)h(\phi)\,\mathrm{d}\phi$ is the prior density of x_{ji}, i.e. the probability of generating x_{ji} from a new draw of a mixture parameter.

As for the prior $p(t_{ji}|\mathbf{t}^{-ji}, \mathbf{k})$ of t_{ji}, we can read it from (19). The probability of picking an old table t is proportional to $n_{jt.}^{-ji}$ and the probability of sitting to a new table is proportional to α. Finally we get our sampling probability for t_{ji}:

$$p(t_{ji} = t|\mathbf{x}, \mathbf{t}^{-ji}, \mathbf{k}) \propto \begin{cases} n_{jt.}^{-ji} f_{k_{jt}}^{-x_{ji}}(x_{ji}) \text{ if } t \text{ previously used} \\ \alpha p(x_{ji}|t_{ji} = t_{\text{new}}, \mathbf{t}^{-ji}, \mathbf{k}) \text{ if } t = t_{\text{new}} \end{cases} \tag{26}$$

If the sampled value of t_{ji} is t_{new}, we sample the index $k_{jt_{\text{new}}}$ of the dish served at this new table from (25)

$$p(k_{jt_{\text{new}}} = k|\mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt_{\text{new}}}) \propto \begin{cases} m_{.k} f_k^{-x_{ji}}(x_{ji}) \text{ if } k \text{ previously used} \\ \gamma f_{k_{\text{new}}}^{-x_{ji}}(x_{ji}) \text{ if } k = k_{\text{new}} \end{cases} \tag{27}$$

As for sampling k_{jt}, from (24) and (25) we see that

$$p(k_{jt} = k|\mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt}) \propto \begin{cases} m_{.k}^{-jt} f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) \text{ if } k \text{ previously used} \\ \gamma f_{k_{\text{new}}}^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) \text{ if } k = k_{\text{new}} \end{cases} \tag{28}$$

where $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$ is the conditional density of $\mathbf{x}_{jt}$ (all customers at table t at restaurant k) given all other data from component k, and

$$f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) = \prod_{x_{ji} \in \mathbf{x}_{jt}} f_k^{-x_{ji}}(x_{ji})$$

After each update, the counts n and m need to be updated. If by updating t_{ji} a table t in restaurant j becomes unoccupied (i.e. $n_{jt.} = 0$), then the probability that a customer will sit at this table in the future is zero, since it is proportional to $n_{jt.}$. We can then erase k_{jt} from memory. If deleting k_{jt} means that mixture component k has no more associated data points, we can also delete it.

A variant of this method also samples $\boldsymbol{\phi}$. This makes the equations simpler as conditioning also by the mixture parameters, $f_k^{-x_{ji}}(x_{ji})$ becomes $p(x_{ji}|\phi_k)$. Sampling a new $\phi_{k_{\mathrm{new}}}$ happens according to (20). Sampling ϕ_k is from

$$p(\phi_k|\mathbf{x}, \mathbf{t}, \mathbf{k}, \boldsymbol{\phi}^{-k}) \propto h(\phi_k) \prod_{x_{ji} \in \mathbf{x}_k} p(x_{ji}|\phi_k) \tag{29}$$

4 Topic modeling

Topic modeling is the problem of finding underlying themes in a set of documents. A simple strategy would be to assign each document to a cluster representing a single idea or topic. But text can be about more than one topic. For example, an article about sports economics should be labeled as an overlap of the *sports* and *economics* topics. We also want to quantify the extent to which an article belongs to a given topic, so we are looking to assign a probabilistic distribution over topics to each document.

The *vocabulary* W is a fixed set of words, for example all words in the documents we have. We define a *topic* as a probability distribution over this vocabulary. We represent each *document* as a bag-of-words, that is to say is a list of words from W, with no particular order between them. Note that even if order is dropped, multiplicity of words is kept and a bag-of-words is actually a multiset.

If only K topics exist, we can imagine the following model for words in a document. Let the document be a probability vector $\pi \in S_{K-1}$ over the K topics. Sample a topic $T \sim \pi$. Then, sample a word from T viewed as a distribution over the vocabulary W. Repeat the two sampling steps to sample more words from the document. π then represents the expected share of words in the documents from each topic. We get the ***latent Dirichlet allocation*** (LDA) model by giving a symmetric Dirichlet prior to π. The Dirichlet distribution $\mathtt{Dir}(K, \alpha_1, ..., \alpha_K)$ is symmetric if $\alpha_1 = ... = \alpha_K = \alpha$. It is a useful prior when we do not have information favoring any component/topic.

Hierarchical Dirichlet processes are clearly well suited to modeling the nonparametric case when the number of topics is unknown and unbounded. If H is the prior distribution over topics, P_θ is a distribution over words in W (so represents a topic) and x_{ji} is the ith word in document j, then the HDP-LDA model for words in a set $\mathcal{J}$ of documents is

$$\begin{aligned}
&G_0|(\gamma, H) \sim \mathtt{DP}(\gamma, H)\\
&G_j|(\alpha, G_0) \sim \mathtt{DP}(\alpha, G_0) \text{ i.i.d. for each document } j \in \mathcal{J}\\
&\theta_{ji}|G_j \sim G_j \text{ for each word } i = 1, .., n_j \text{ in each document } j \in \mathcal{J}\\
&x_{ji}|\theta_{ji} \sim P_{\theta_{ji}} \qquad \text{(HDP-LDA)}
\end{aligned}$$

The topics θ_{ji} are all atoms of G_0 and therefore the same topics can appear in multiple documents.

Due to lack of time and technical difficulties, we were not able to write a working and reliable implementation of Gibbs sampling. Instead, we turned to the model available in the Python package `gensim`, which implements the HDP through online variational inference, as described in [5]. We applied the method on the Grolier dataset which gathers around 27.000 encyclopedia articles in English. The articles have already been preprocessed: very common words of the English language have been removed and the documents come in a bag-of-words format. The algorithm found in [5] is quite different from Gibbs sampling since it truncates the number of topics both at the corpus and document levels to fixed upper bounds values, respectively K and T. The values of these upper bounds on the number of topics have to be provided as parameters to the `gensim` HDP function, and we simply used the default settings $K = 15$ and $T = 150$. The algorithm will then return K
In Table 1 we report the probability distributions over words for the first six topics returned by the model. The first topic may be interpreted as articles related to history of the United States, while the second is about American art, the third about physics, and so on.

Topic number	Distribution over words (7 highest probabilities)
1	$0.006 \cdot$ war $+ 0.004 \cdot$ government $+ 0.004 \cdot$ united$+$ $0.003 \cdot$ world $+ 0.003 \cdot$ american $+ 0.003 \cdot$ century $+ 0.003 \cdot$ political
2	$0.006 \cdot$ art $+ 0.005 \cdot$ century $+ 0.003 \cdot$ music$+$ $0.003 \cdot$ american $+ 0.003 \cdot$ life $+ 0.003 \cdot$ world $+ 0.003 \cdot$ style
3	$0.004 \cdot$ called $+ 0.003 \cdot$ water $+ 0.003 \cdot$ system$+$ $0.003 \cdot$ form $+ 0.003 \cdot$ energy $+ 0.002 \cdot$ time $+ 0.002 \cdot$ earth
4	$0.007 \cdot$ km $+ 0.007 \cdot$ mi $+ 0.006 \cdot$ population$+$ $0.006 \cdot$ sq $+ 0.006 \cdot$ south $+ 0.005 \cdot$ north $+ 0.005 \cdot$ deg
5	$0.012 \cdot$ species $+ 0.009 \cdot$ family $+ 0.005 \cdot$ cm$+$ $0.005 \cdot$ found $+ 0.005 \cdot$ plants $+ 0.005 \cdot$ ft $+ 0.004 \cdot$ north
6	$0.028 \cdot$ city $+ 0.009 \cdot$ mi $+ 0.009 \cdot$ km$+$ $0.009 \cdot$ center $+ 0.009 \cdot$ century $+ 0.008 \cdot$ population $+ 0.008 \cdot$ river

Table 1: Probability distribution over words for the first 6 topics returned by the model

One may also retrieve without difficulty the probability distribution over topics for each document. As an example, for the fourth article in the corpus, the model returns a very sparse distribution: $0.93 \cdot \text{Topic } 2 + 0.07 \cdot \text{Topic } 6$.The content of this article is indeed vastly about art.

References

[1] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.

[2] Nils Lid Hjort, Chris Holmes, Peter Muller, and Stephen G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[3] Peter Orbanz. *Lecture Notes on Bayesian Nonparametrics*. http://stat.columbia.edu/ porbanz/papers.html, 2014.

[4] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[5] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.