Bayesian LDA for categorical mixture model clustering.

Pedro H. M. Albuquerque
University of Bras'{i}lia

Denis Ribeiro do Valle
University of Florida

Abstract

The goal of this paper is to describe the Bayesian LDA model for fuzzy clustering based on different types of data (i.e., Multinomial, Bernoulli, and Binomial entries) and provide some examples of the use of this model in R. These types of data frequently emerge in fields likeas disparate as ecology, remote sensing, marketing, and finance and others fields and, as a result, we believe this package will be of broad interest for pattern recognition, especially fuzzy clustering for categorical data.

Keywords: LDA, fuzzy clustering.

1. Introduction.

The Latent Dirichlet Allocation model (LDA), first proposed by Blei, Ng, and Jordan (2003), has been extensively used for text-mining in multiple fields. Tsai (2011) used LDA to study the tag pattern representation in blogs to construct clusters of tags that represents the most common topics. Lee et al. (2010) compared LDA against three other text mining methods that are frequently used: latent semantic analysis, probabilistic latent semantic analysis and the correlated topic model. The authors described the limitations of each method, asserting that the LDA limitations were essentially two: the method does not consider relationships between topics and cannot allocate a word to multiple topics . However, beside these limitations the LDA continues to be used in multiple disciplines. For instance, Griffiths and Steyvers (2004) used LDA to identify the main scientific topics in a large corpus of Proceedings of the National Academy of Science (PNAS) articles. In conservation biology, LDA has been used to identify research gaps in the conservation literature (Westgate et al. 2015). LDA has also been proposed as a promising method for the automatic annotation of remote sensing imagery (Lienou, Maître, and Datcu 2010). In marketing, LDA has been used to extract information from product reviews across 15 firms in five markets over four years, enabling the identification

of the most important latent dimensions of consumer decision making in each studied market (Tirunillai and Tellis 2014). In finance, a stock market analysis system based on LDA is used to analyze financial news items together with market data to identify and characterize major events that impact the market. This system is then used to predict whether the stock market will fall or rise, based on news items identified by LDA (Mahajan, Dey, and Haque 2008).

Despite its success in text mining across multiple fields, LDA is a model that need not be restricted to text-mining. More specifically, LDA can be viewed as a mixture model since each element in the sample can belong to more than one cluster (or state) simultaneously. There are a few examples of LDA being used for other purposes than text-mining. For instance, a modified version of LDA has been extensively used on genetic data to identify populations and admixture probabilities of individuals (Pritchard, Stephens, and Donnelly 2000). Similarly, LDA has been used in Ecology to identify plant communities from tree data for the eastern United States and from a tropical forest chronosequence (Valle et al. 2014).

The aim of this paper is to describe a Bayesian LDA model for mixture model clustering based on different types of data (i.e., multinomial, Bernoulli and Binomial), illustrating its use in a diverse set of examples. The innovative features of this model are that it generalizes LDA for other types of categorical data that commonly appear in Life and Social Science and it enables the selection of the optimal number of clusters based on a truncated stick-breaking prior approach which regularizes model results. Finally, we provide a package to estimate the quantities of interest.

This paper is organized as follows. Section 2 describes the mathematical formulation for the Bayesian LDA model for fuzzy clustering analysis and section 3 shows how the model was implemented in R. Sections 4 and 5 present examples of the use of the package and the conclusions, respectively.

2. Methods.

In the Bayesian LDA model for fuzzy clustering we postulate that exists a latent state for which each element can be allocated (not exclusively) in one cluster. Specifically, consider a latent matrix $\mathbf{Z}$ with dimension equals to $L \times C$ where each row represent a sampling unit $(l = 1, \ldots, L)$ and each column a possible state or cluster $(c = 1, \ldots, C)$. The Data Generating Process postulated for this latent matrix is given by:

$$Z_{lc} \sim Multinomial(n_l, \boldsymbol{\theta}_l) \tag{1}$$

where n_l is total number of elements drawn for location l and $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lC})$ is a vector of parameters representing the probability of allocation in each cluster. Following Occam's razor, we intend to create the least number of clusters as possible, which is achieved by assuming a truncated stick-breaking prior:

$$\theta_{lc^*} = V_{lc^*} \prod_{c < c^*}^{C} (1 - V_{lc}) \tag{2}$$

where $V_{lc} \sim Beta(1, \alpha)$ for $c = 1, \ldots, C - 1$ and $V_{lC} = 1$ by definition. This truncated stick-breaking prior will force the elements to be aggregated in the minimum number of clusters, given that θ_{lc^*} is stochastically exponentially decreasing.

In the second hierarchical level, we consider a matrix $\mathbf{Y}$ with dimension equals to $L \times S$ where each row represents a sampling unit (e.g., locations, firms, individuals, plots) and each column a variable that describes these elements. In the Bayesian LDA model for fuzzy clustering, after integrating over the latent vector $\mathbf{Z}_l$, Y_{ls} can follow one of these distributions:

$$
\begin{cases}
\mathbf{Y}_{l\cdot} \sim Multinomial(n_l, \boldsymbol{\theta}_i^t \Phi) \\
Y_{ls} \sim Bernoulli(\boldsymbol{\theta}_i^t \boldsymbol{\phi}_s) \\
Y_{ls} \sim Binomial(n_{ls}, \boldsymbol{\theta}_i^t \boldsymbol{\phi}_s)
\end{cases}
\tag{3}
$$

for $l = 1, \ldots, L$ and $s = 1, \ldots, S$ possible states. Y_{ls} represents a random variable, n_l is the total number of elements in sampling unit l and variable s. In these models, $\boldsymbol{\phi}_s = (\phi_{1s}, \ldots, \phi_{Cs})$ is a vector of parameters, while:

$$
\Phi =
\begin{bmatrix}
\phi_{11} & \phi_{12} & \ldots & \phi_{1S} \\
\phi_{21} & \phi_{22} & \ldots & \phi_{2S} \\
\vdots & \vdots & \ddots & \vdots \\
\phi_{C1} & \phi_{C2} & \ldots & \phi_{CS}
\end{bmatrix}
$$

is a $C \times S$ matrix of parameters.

In the last step, we specify the priors for ϕ_{cs}. For the multinomial model, we adopt a Dirichlet prior (i.e. $\boldsymbol{\phi}_c \sim Dirichlet(\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)$ is the hyperparameter vector). For the Bernoulli and Binomial representations, we assume that ϕ_{cs} comes from a Beta distribution, (i.e., $\phi_{cs} \sim Beta(\alpha_0, \alpha_1)$).

These models are fit using Gibbs Sampling where parameters draws are iteratively made from each full conditional distribution. From a conceptual perspective, all of these models assume the following matrix decomposition:

$$
\mathbb{E}[\mathbf{Y}_{L \times S}] = \mathbf{K} \circ [\Theta_{L \times C} \Phi_{C \times S}]
\tag{4}
$$

where $\mathbf{K}$ is a matrix of constants and $\circ$ is the Hadamard product. Sparseness is ensured by forcing large c in the $\Theta_{L \times C}$ matrix to be close to zero. For the multinomial model, the $\mathbf{K}$ matrix contains the total number of elements in each row whereas for the Bernoulli model, this matrix is equal to the identity matrix. Finally, for the Binomial model, the $\mathbf{K}$ matrix has the total number of trials of each binomial distribution (i.e., n_{ls}). Although there are many ways matrices can be decomposed, the key characteristic of this particular form of matrix decomposition is that each row of $\mathbf{K}$ is comprised of probabilities that sum to one. As a result, one can interpret $\Phi_{C \times S}$ as the matrix that contain the "pure" features of the data, which are then mixed by the matrix $\Theta_{L \times C}$ and multiplied by $\mathbf{K}$ to generate the expected data.

2.1. Full Conditional Distributions - FCD.

Bernoulli model.

The probability of community membership status Z_{ls} is given by:

$$p(Z_{lc} = c^*|Y_{ls} = y_{ls}, \phi_{c^*s}, \boldsymbol{\theta}_l) = \frac{\theta_{lc^*}\phi_{c^*s}^{y_{ls}}(1-\phi_{c^*s})^{1-y_{ls}}}{\displaystyle\sum_{c=1}^{C}\theta_{lc}\phi_{cs}^{y_{ls}}(1-\phi_{cs})^{1-y_{ls}}} \tag{5}$$

Therefore, Z_{ls} can be drawn from a categorical distribution. The FCD for ϕ_{cs} is given by:

$$p(V_{lc^*}|\mathbf{Y}) = Beta(N_{lc^*} + 1, N_{l(c>c^*)} + \gamma)$$

where N_{lc^*} is the total number of elements in location l classified into cluster c^*, and $N_{l(c>c^*)}$ is the total number of elements in location l classified in clusters larger than c^*. These quantities are given by $N_{lc^*} = \sum_{s=1}^{S}\mathbb{1}(z_{ls} = c^*)$ and $N_{l(c>c^*)} = \sum_{s=1}^{S}\sum_{c=c^*+1}^{C}\mathbb{1}(z_{ls} = c^*)$, respectively.

Binomial model.

For this model, we have n_{ls} elements for each sampling unit l and community z. The i-th element is denoted as Z_{ils} and its probability is similar to the one for the Bernoulli model:

$$p(Z_{lc} = c^*|Y_{ls} = y_{ls}, \boldsymbol{\phi}_s, \boldsymbol{\theta}_l) = \frac{\theta_{lc^*}\phi_{c^*s}^{x_{ils}}(1-\phi_{c^*s})^{1-x_{ils}}}{\displaystyle\sum_{c=1}^{C}\theta_{lc}\phi_{cs}^{y_{ls}}(1-\phi_{cs})^{1-y_{ls}}} \tag{6}$$

where x_{ils} are binary random variables such that $\sum_{i=1}^{n_{ls}} x_{ils} = y_{ls}$. Therefore, Z_{ils} can be drawn from a multinomial distribution. The FCD is given by:

$$p(\phi_{cs}|\mathbf{Z}, \mathbf{Y}) = Beta\left(\sum_{l=1}^{L}\sum_{i=1}^{n_{ls}}\mathbb{1}(x_{ils} = 1, z_{ils} = c) + \alpha_0, \sum_{l=1}^{L}\sum_{i=1}^{n_{ls}}\mathbb{1}(x_{ils} = 0, z_{ils} = c) + \alpha_1\right) \tag{7}$$

Finally, the final FCD for V_{lc^*} is given by:

$$p\left(V_{lc^*}|Z_{lc^*}\right) = Beta(N_{lc^*} + 1, N_{l(c>c^*)} + \gamma) \tag{8}$$

where N_{lc^*} is the total number of elements in location l classified into cluster c^* and $N_{l(c>c^*)}$ is the total number of elements in location l classified in clusters larger than c^*. These quantities are given by $N_{lc^*} = \sum_{s=1}^{S}\sum_{i=1}^{n_{ls}}\mathbb{1}(z_{ils} = c^*)$ and $N_{l(c>c^*)} = \sum_{s=1}^{S}\sum_{i=1}^{n_{ls}}\sum_{c=c^*+1}^{C}\mathbb{1}(z_{ils} = c^*)$.

Multinomial model.

For the Multinomial case, if unit i in location l is associated with variable s (i.e., $x_{il} = s$ such that $y_{ls} = \sum_{i=1}^{n_l}\mathbb{1}(x_{il} = s)$), we have that:

$$p(Z_{lc} = c^*|\mathbf{Y}_{ls} = s, \boldsymbol{\phi}_s, \boldsymbol{\theta}_l) = \frac{\theta_{lc^*}\phi_{sc^*}}{(\theta_{1l}\phi_{s1} + \cdots + \theta_{Cl}\phi_{sC})} \tag{9}$$

Therefore, Z_{il} can be sampled from a categorical distribution. Since we assumed a conjugate prior for ϕ_{c^*} with $c^* \in \{1, \ldots, C\}$ the Full Conditional Distribution for this vector of parameters is a straight-forward Dirichlet distribution:

$$p(\phi_{c^*}|\mathbf{Y}) = Dirichlet([n_{c^*1} + \beta_1, \ldots, n_{c^*S} + \beta_S]) \tag{10}$$

where n_{c^*s} is the total number of observations classified in cluster c^* in all locations for the s-th variable (i.e., $n_{c^*s} = \sum_{l=1}^{L} \sum_{i=1}^{n_l} \mathbb{1}(z_{il} = c^*, x_{il} = s)$).

Finally, the FCD for V_{lc^*} is given by:

$$p(V_{lc^*}|Z_{lc^*}) = Beta(N_{lc^*} + 1, N_{l(c>c^*)} + \gamma) \tag{11}$$

where N_{lc^*} is the total number of elements in observation l classified into cluster c^* and $N_{l(c>c^*)}$ is the total number of elements in observation l classified in clusters larger than c^*. These quantities are given by $N_{lc^*} = \sum_{i=1}^{n_l} \mathbb{1}(z_{il} = c^*)$ and $N_{l(c>c^*)} = \sum_{i=1}^{n_l} \sum_{c=c^*+1}^{C} \mathbb{1}(z_{il} = c)$.

3. The package.

The package is not the only one package that can be able to perform the Latent Dirichlet Allocation model. We found four packages that propose model a Latent Dirichlet Allocation structure in R. Hornik and Grün (2011) developed the **topicmodels** package for which there are two LDA implementations, one using the variational inference (as described in Blei, Ng, and Jordan (2003)) and the other implementation using Gibbs Sampling based on Phan and Nguyen (2013). Similarly, Jones (2016) proposed the **textmineR** which has a Gibbs Sampling method to estimate the topics in a corpus structure.

Chang (2012) developed the **lda** package which includes the mixed-membership stochastic blockmodel (Airoldi et al. 2008), supervised Latent Dirichlet Allocation - sLDA (Mcauliffe and Blei 2008) and Correspondence-Latent Dirichlet Allocation - corrLDA(Blei and Jordan 2003).

More recently, Roberts, Stewart, and Tingley (2014) created the **stm** which has some unsupervised functions to determine the optimal number of clusters. This unsupervised method is based on likelihood metrics obtained by the EM Algorithm, which are then used for a backward model selection approach.

None of these packages adopt the truncated stick-breaking prior and none of them use others distributions besides the Multinomial outcome for the dependent variable. Thus, our package complements current LDA approaches already available in R.

4. Examples.

In this section we present some applications for Latent Dirichlet Allocation in marketing, finance and ecology.

4.1. Marketing.

The first application considers the classical LDA for a Multinomial entry in the field of Marketing. More specifically, we are interested in characterizing firms based on their consumers' complaints.

It is well known that attracting a new customer is often considerably more costly than keeping current customer (Kotler and Armstrong 2006). For this reason, firms can better retain their customers if they pay careful attention to their consumers' complaints and work to solve them in a satisfactory way.

The data used is from the 2015 Consumer Complaint Database which are complaints received for the **Bureau of Consumer Financial Protection** in US about financial products and services. Specifically in this example we work with only credit card complaint's and it contains information on the number of complaints for each firm ($L = 226$), categorized according to the specific type of issue ($S = 30$). Examples of issues include: Billing disputes, Identity theft / Fraud, Unsolicited issuance of credit card and so on. Each sampling unit in this case represents a firm and each variable the total number of complaints'complaints associated with each issue.

The characterization of firms provided by our analysis can be useful to reveal communalities and differences across different firms. This can then be used by managers to identify and potentially adopt the solutions that are employed by other firms to deal with these issues.

To use the **Rlda** package for the Multinomial entry is necessary to create a matrix where each cell represents the total number of cases observed for each variable and sampling unit.

```
library(Rlda)
#Read the Complaints data
data(complaints)
#Create the abundance matrix
library(reshape2)
mat1<-dcast(complaints[,c("Company","Issue")],
           Company~Issue, length,
           value.var="Issue")
#Create the rowname
rownames(mat1)<-mat1[,1]
#Remove the ID variable
mat1<-mat1[,-1]
```

To use the `GibbsSamplingAbundance` method we need to specify some arguments:

```
#Set seed
set.seed(9292)
#Hyperparameters for each prior distribution
beta<-rep(1,ncol(mat1))
gamma<-0.01
#Execute the LDA for the Multinomial entry
res<-GibbsSamplingAbundance(mat1,30,beta,gamma,1000,TRUE,FALSE)
```

In the above code we defined the maximum number of clusters as 30 and 1000 Gibbs Sampling iterations with the likelihood computed with priors and not showing the progress bar.

We can visually evaluate the convergence Figure 1:

```
#Get the logLikelihood
ll<-res[["logLikelihood"]]
#Plot the log-likelihood
plot(ll,type="l",xlab="Interations",
                   ylab="Log-likelihood")
```
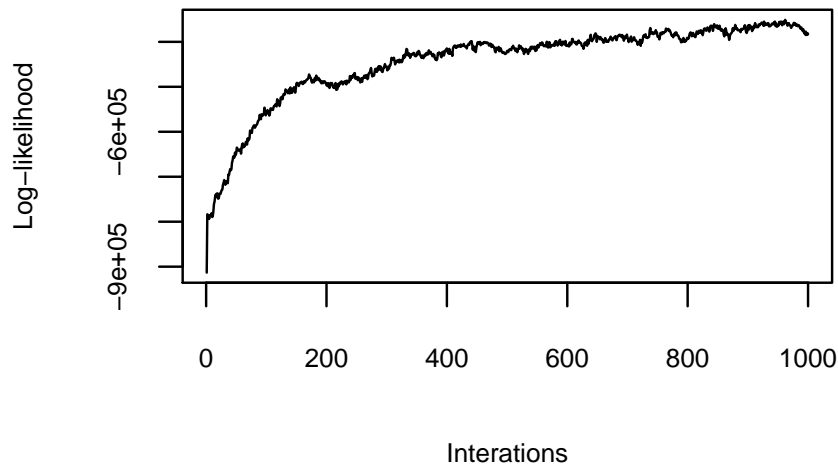

Figure 1: Log-likelihood interations.

The Bayesian estimate is given by the `Theta` and `Phi` matrix. Each line for those objects represent a Gibbs iteration, in this way we can find estimates based on the mean of the matrices only for those iterations after the *burn-in phase*.

```
#Get the Theta Estimate
Theta<-res[["Theta"]]
#Burnout
Theta<-colMeans(Theta[300:1000,])
#Create the matrix
Theta<-matrix(Theta,nrow = nrow(mat1),ncol=30)
#Rownames
rownames(Theta)<-rownames(mat1)
```

The `Theta` has a sparse structure since our truncated stick-breaking prior tends reduce the total number dominant clusters. We can think that each probability estimated in each cell represent the probability of the l-th firm be allocated in cluster s, in this case $s = 1, \ldots, S$.

To describe each cluster, we can look the `Phi` matrix:

```
#Get the Phi Estimate
Phi<-res[["Phi"]]
#Burnout
Phi<-colMeans(Phi[300:1000,])
#Create the matrix
Phi<-matrix(Phi,nrow =30,ncol=ncol(mat1))
#Colnames
colnames(Phi)<-colnames(mat1)
#Rownames
rownames(Phi)<-paste0("Cluster ",seq(1,30))
#Get the most likely issues
ids<-which(Phi > 0.2, arr.ind = TRUE)
```

The most likely issues and its clusters are summarized as follow:

Cluster	Issue
7	Advertising and marketing
7	Rewards
17	Closing/Cancelling account
26	Unsolicited issuance of credit card

Table 1: Clusters description.

Based on those informations we can now analyze the firms with highest probability to belong these clusters:

Firm	Cluster	Probability
U.S. Bancorp	Cluster 7	0.5239
Amex	Cluster 7	0.1878
Barclays PLC	Cluster 7	0.0974
Discover	Cluster 7	0.0511
PayPal Holdings, Inc.	Cluster 7	0.0153
U.S. Bancorp	Cluster 17	0.0000
Amex	Cluster 17	0.0000
Barclays PLC	Cluster 17	0.0000
Discover	Cluster 17	0.2957
PayPal Holdings, Inc.	Cluster 17	0.0088
U.S. Bancorp	Cluster 26	0.0000
Amex	Cluster 26	0.0000
Barclays PLC	Cluster 26	0.0000
Discover	Cluster 26	0.0000
PayPal Holdings, Inc.	Cluster 26	0.3958

Table 2: Probability of belonging for each company.

4.2. Finance.

In Financial Market is usually common analyze firms based on their liquidity, i.e., liquidity of a firm refers to its ability to meet short-term obligations using firm's assets can be quickly converted to cash (Audretsch and Elston 2002).

In this example we studied how firms can be clustered based on two main factor: The firms' free float[1] and the negotiated volume[ˆVolume is number of shares that traded on the trade

---

[1]Free float is the number of outstanding shares that are available for trading by the public.

date.].

We can approximate the volume random variable for each firm in each day as Binomial trial, since the volume should be less than free float[^In fact this is just an approximation, since one share can be negotiated multiple times in the same day] for the same firm and day.

In this example we worked with all daily transactions for 46 firms of the S&P500 Index in 2015 obtained by the Reuters Reuters Terminal, the data was treated to represent the excepted value of daily transactions for a free float equal 100:

```
library(Rlda)
#Read the SP500 data
data(sp500)
#Create size
spSize<-as.data.frame(matrix(100,
                      ncol=ncol(sp500),
                      nrow=nrow(sp500)))
```

In this cases since we are postulated a Binomial trial we should use the `GibbsSamplingBinomial`, for which we need to specify the follow arguments:

```
#Set seed
set.seed(5874)
#Hyperparameters for each prior distribution
gamma <-0.01
alpha0<-0.01
alpha1<-0.01
#Execute the LDA for the Binomial entry
res<-GibbsSamplingBinomial(sp500,spSize,10,
                      alpha0,alpha1,gamma,
                      5000,TRUE,FALSE)
```

The firms can be represented by the `Phi` matrix which presents the probability of belonging (not exclusively) by cluster.

```
#Get the Phi Estimate
Phi<-res[["Phi"]]
#Bayes Estimation
Phi<-colMeans(Phi)
#Create the matrix
Phi<-matrix(Phi,nrow =10,ncol=ncol(sp500))
#Colnames
colnames(Phi)<-colnames(sp500)
#Rownames
rownames(Phi)<-paste0("C",seq(1,10))
```

Hence, one way to summarize the results is showing which cluster has the highest probability to appear for each possible firm, in this case only for clusters with probability greater than 50%:
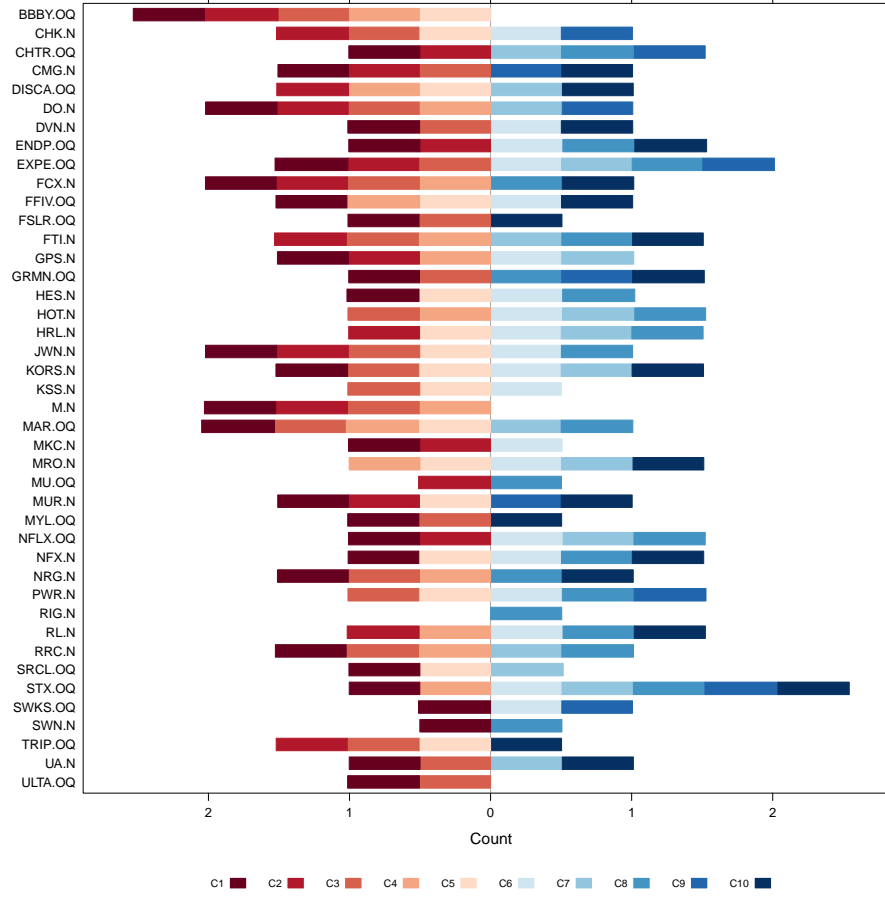
Figure 2: Acumulated probability of belonging.

Financial investors can, for example, using the information presented in Figure 2 to construct portfolios based on these clusters aiming to diversify their assets.

4.3. Ecology.

One of the most challenge situation in study ecological data is the fact the in some applications, it is not possible to observe the total number of species in each sampled plot. However, it is possible to know if some is species is presented (i.e. overserved) or absente (i.e. not observed.) (Pearce and Boyce 2006).

For this kind of data, we can use the the `GibbsSamplingPresence`. Specifically, in this example we used XXXXXXXXXX which is referred to the Species observed (value equal 1) and not observed or absent (value equal 0) in some region. Each observation in this case represents a surveyed location and each variable one type of specie.

5. Conclusion.

Appendix.

Here in the Appendix we will show how the Full Conditional Distributions were constructed. Consider initially the Full Conditional Distributions for $Z_{lc} = c^*$ in all three cases the kernel of the distribution derived is the same:

$$
\begin{aligned}
p(Z_{lc} = c^* | Y_{ls} = s, \boldsymbol{\phi}_{c^*}, \boldsymbol{\theta}_l) &= \kappa \left[\phi_{c^*1}^{\mathbb{1}(Y_{l1}=s, Z_{lc}=c^*)} \times \cdots \times \phi_{c^*S}^{\mathbb{1}(Y_{lS}=s, Z_{lc}=c^*)} \right] \\
&\quad \times \left[\theta_{l1}^{\mathbb{1}(Z_{l1}=c^*)} \times \cdots \times \theta_{lC}^{\mathbb{1}(Z_{lC}=c^*)} \right] \\
&= \kappa \phi_{c^*s} \theta_{lc^*}
\end{aligned}
$$

where $\mathbb{1}(Y_{l1} = s, Z_{lc} = c^*)$ is the indicator function which assumes one only for the l-th observation, s-th variable and has been identified as belong to cluster c^*. In a similar way, $\mathbb{1}(Z_{lC} = c^*)$ assumes one only for observations classified as belong to cluster c^*.

Since Z_{lc} is a categorical random variable with support in $\mathscr{Z} = (1, 2, \ldots, C)$ the sum of all probabilities for all elements must one, then the constant κ is given by:

$$
\kappa = (\theta_{l1}\phi_{1s} + \cdots + \theta_{lC}\phi_{Cs})^{-1}
$$

Then, each category for Z_{lc} can be draw from a categorical distribution with probabilities equal to $\kappa \phi_{c^*s}\theta_{lc^*}$ with $c^* \in \mathscr{Z}$. For the Binomial and Bernoulli cases we have:

$$
p(Z_{lc} = c^* | Y_{ls} = s, \boldsymbol{\phi}_{c^*}, \boldsymbol{\theta}_l) \propto \theta_{lc^*} \phi_{c^*s}^{y_{ls}} (1 - \phi_{c^*s})^{1-y_{ls}}
$$

where each cluster label can be draw from a categorical distribution with probabilities equal to:

$$
p(Z_{lc} = c^* | Y_{ls} = s, \boldsymbol{\phi}_{c^*}, \boldsymbol{\theta}_l) = \frac{\theta_{lc^*} \phi_{c^*s}^{y_{ls}} (1 - \phi_{c^*s})^{1-y_{ls}}}{\displaystyle\sum_{c=1}^{C} \theta_{lc} \phi_{cs}^{y_{ls}} (1 - \phi_{cs})^{1-y_{ls}}}
$$

For the $\boldsymbol{\phi}_{c^*}$ its Full Conditional is given by:

$$
\begin{aligned}
p\left(\boldsymbol{\phi}_{c^*} | \mathbf{Z}_{\cdot c^*}, \{\mathbf{Y}_{l\cdot} | Z_{lc} = c^*\}, \boldsymbol{\beta}\right) &\propto \prod_{l=1}^{L} \phi_{c^*1}^{\mathbb{1}(Y_{l1}=1, Z_{lc}=c^*)} \times \cdots \times \phi_{c^*S}^{\mathbb{1}(Y_{l1}=S, Z_{lc}=c^*)} \\
&\quad \times \phi_{c^*1}^{\beta_1-1} \times \cdots \times \phi_{c^*S}^{\beta_S-1} \\
&\propto \prod_{s=1}^{S} \phi_{c^*s}^{n_{c^*s}+\beta_s-1}
\end{aligned}
$$

where n_{c^*s} is the total number of elements classified in community c^* and independent variable s. In the Binomial or Bernoulli case the Full Conditional is given by:

$$
\begin{aligned}
p\left(\phi_{c^*s} | \mathbf{Z}_{\cdot c^*}, \{\mathbf{Y}_{l\cdot} | Z_{lc} = c^*\}, \boldsymbol{\beta}\right) &\propto \prod_{l=1}^{L} \phi_{c^*s}^{\mathbb{1}(Y_{ls}=s, Z_{lc}=c^*)} \phi_{c^*s}^{\alpha_0-1} (1 - \phi_{c^*s})^{\alpha_1-1} \\
&\propto \phi_{c^*s}^{n_{c^*s}+\alpha_0-1} (1 - \phi_{c^*s})^{n-c^*s+\alpha_1-1}
\end{aligned}
$$

where n_{-c^*s} is the total number of elements that **not** belong to either community c^* and indepenent variable s.

The last Full Conditional described here is the Full Conditional for $p\left(V_{lc}|Z_{lc}\right)$, in this case we have:

$$
\begin{aligned}
p\left(V_{lc}|Z_{lc}\right) &\propto Binom(n_{lc}|n = n_{lc} + n_{l(c^*>c)}, V_{lc})Beta(V_{lc}|1,\gamma) \\
&\propto Beta(1 + n_{lc}, \gamma + n_{l(c^*>c)}) \\
&= \tfrac{\Gamma(n_{lc}+1+n_{l(c^*>c)}+\gamma)}{\Gamma(n_{lc}+1)\Gamma(n_{l(c^*>c)}+\gamma)}V_{lc}^{n_{lc}}\left(1 - V_{lc}^{n_{l(c^*>c)}+\gamma}\right)
\end{aligned}
$$

References.

Airoldi, Edoardo M, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. "Mixed Membership Stochastic Blockmodels." *Journal of Machine Learning Research* 9 (Sep): 1981–2014.

Audretsch, David B, and Julie Ann Elston. 2002. "Does Firm Size Matter? Evidence on the Impact of Liquidity Constraints on Firm Investment Behavior in Germany." *International Journal of Industrial Organization* 20 (1). Elsevier: 1–17.

Blei, David M, and Michael I Jordan. 2003. "Modeling Annotated Data." In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 127–34. ACM.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chang, Jonathan. 2012. "Lda: Collapsed Gibbs Sampling Methods for Topic Models. R Package Version 1.4.2."

Griffiths, Thomas L, and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (suppl 1). National Acad Sciences: 5228–35.

Hornik, Kurt, and Bettina Grün. 2011. "Topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13). American Statistical Association: 1–30.

Jones, Thomas W. 2016. "TextmineR: Functions for Text Mining and Topic Modeling. R Package Version 2.0.2."

Kotler, Philip, and Gary Armstrong. 2006. "Principles of Marketing Management." Englewood Cliffs, NJ: Pearson Prentice-Hall.

Lee, Sangno, Jeff Baker, Jaeki Song, and James C Wetherbe. 2010. "An Empirical Comparison of Four Text Mining Methods." In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 1–10. IEEE.

Lienou, Marie, Henri Maître, and Mihai Datcu. 2010. "Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation." *IEEE Geoscience and Remote Sensing Letters* 7 (1). IEEE: 28–32.

Mahajan, Anuj, Lipika Dey, and Sk Mirajul Haque. 2008. "Mining Financial News for Major Events and Their Impacts on the Market." In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, 1:423–26.

IEEE.

Mcauliffe, Jon D, and David M Blei. 2008. "Supervised Topic Models." In *Advances in Neural Information Processing Systems*, 121–28.

Pearce, Jennie L, and Mark S Boyce. 2006. "Modelling Distribution and Abundance with Presence-Only Data." *Journal of Applied Ecology* 43 (3). Wiley Online Library: 405–12.

Phan, Xuan-Hieu, and Cam-Tu Nguyen. 2013. "GibbsLDA++, AC/C++ Implementation of Latent Dirichlet Allocation (LDA) Using Gibbs Sampling for Parameter Estimation and Inference."

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2). Genetics Soc America: 945–59.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014. "Stm: R Package for Structural Topic Models." *R Package* 1: 12.

Tirunillai, Seshadri, and Gerard J Tellis. 2014. "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation." *Journal of Marketing Research* 51 (4). American Marketing Association: 463–79.

Tsai, Flora S. 2011. "A Tag-Topic Model for Blog Mining." *Expert Systems with Applications* 38 (5). Elsevier: 5330–5.

Valle, Denis, Benjamin Baiser, Christopher W Woodall, and Robin Chazdon. 2014. "Decomposing Biodiversity Data Using the Latent Dirichlet Allocation Model, a Probabilistic Multivariate Statistical Method." *Ecology Letters* 17 (12). Wiley Online Library: 1591–1601.

Westgate, Martin J, Philip S Barton, Jennifer C Pierson, and David B Lindenmayer. 2015. "Text Analysis Tools for Identification of Emerging Topics and Research Gaps in Conservation Science." *Conservation Biology* 29 (6). Wiley Online Library: 1606–14.

Affiliation:

Pedro H. M. Albuquerque
University of Bras'{i}lia
Faculdade de Economia, Administraç˜{a}o e Contabilidade, Building A-2 - Office A1-54/7,
Brasilia, DF 70910-900.
E-mail: pedroa@unb.br
URL: http://pedrounb.blogspot.com/

Denis Ribeiro do Valle
University of Florida
408 McCarty Hall C, PO Box 110339, Gainesville, FL 32611-0410.
E-mail: drvalle@ufl.edu
URL: http://denisvalle.weebly.com/