Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Bayesian LDA for mixed-membership clustering analysis: The Rlda package.

Pedro Albuquerque
University of Brasília

Denis Ribeiro do Valle
University of Florida

Daijiang Li
University of Florida

Abstract

The goal of this paper is to present the **Rlda** package for mixed-membership clustering analysis based on different types of data (i.e., Multinomial, Bernoulli, and Binomial entries), present the theory behind the developed method and also provide some examples of the use of this package in R. These types of data frequently emerge in fields as disparate as ecology, remote sensing, marketing, and finance, for example. As result, we believe this package will be of broad interest for unsupervised pattern recognition, particularly mixed-membership clustering analysis for categorical data.

Keywords: LDA, mixed-membership clustering, categorical data, pattern recognition.

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 construct clusters of tags that represent the most common topics in blogs. Lee, Baker, Song, and Wetherbe (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 major limitation of LDA, as identified by these authors, was that the method does not consider relationship between topics as a mixed-membership clustering approach does (Erosheva and Fienberg 2005). Despite these limitations, however, 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 the Proceedings of the National Academy of Science articles. In conservation biology, LDA has been used to identify research gaps in the conservation literature (Westgate, Barton, Pierson, and Lindenmayer 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). Finally, in finance, a stock market analysis system based on LDA was used to combine financial news items together with stock market data to identify and characterize major events that impact the market. This system was then used to make predictions regarding stock market behavior 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 mixed membership models 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 admixed 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, Baiser, Woodall, and Chazdon 2014).

The aim of this paper is to present the **Rlda** package for mixed-membership clustering analysis and describe this novel Bayesian model 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 twofold. First generalizes LDA for other types of commonly encountered categorical data. Second it enables the selection of the optimal number of clusters based on a truncated stick-breaking prior approach regularizing model results.

This paper is organized as follows. Section 2 and section 3 describe the mathematical formulation for the Bayesian LDA mixed-membership cluster model. Section 4 justifies our **Rlda** package by reviewing current available R packages and their limitations. Sections 5 and 6 present examples of the use of the package and the conclusions, respectively.

2. Methods.

In the Bayesian LDA mixed-membership cluster model we postulate that each element is allocated to a single cluster, represented by a latent state variable. Specifically, consider a latent matrix $\mathbf{Z}$ with dimension equals to $L \times C$ where each row represents 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:

$$\boldsymbol{Z}_{l\cdot} \sim Multinomial(n_l, \boldsymbol{\theta}_l) \tag{1}$$

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:

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

where $V_{lc} \sim Beta(1, \gamma)$ 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 equal 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 mixed-membership clustering, after integrating over the latent vector $\mathbf{Z}_{l\cdot}$, 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 variables. Y_{ls} represents a random variable, $\boldsymbol{Y}_{l\cdot}$ is a vector with these random variables for location l, n_l is the total number of elements in sampling unit l, n_{ls} 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 $\boldsymbol{\Phi}$ is a $C \times S$ matrix of parameters, given by:

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

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 parameter 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 the form of matrix decomposition we choose is that each row of $\boldsymbol{\Theta}$ 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.

3. Full Conditional Distributions - FCD.

3.1. Bernoulli model.

The probability of community membership status for each element W_{ls}, where $Z_{lc} = \sum_{s=1}^{S} \mathbb{1}(W_{ls} = c)$, is given by:

$$p(W_{ls} = c^*|...) = \frac{\theta_{lc^*}\phi_{c^*s}^{y_{ls}}(1-\phi_{c^*s})^{1-y_{ls}}}{\sum_{c=1}^{C}\theta_{lc}\phi_{cs}^{y_{ls}}(1-\phi_{cs})^{1-y_{ls}}} \tag{5}$$

Therefore, W_{ls} can be drawn from a categorical distribution. The FCD for V_{lc} is given by:

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

where z_{lc} is the total number of elements in location l classified into cluster c, and $z_{l(c^*>c)}$ is the total number of elements in location l classified in clusters larger than c. This latter quantity is given by $z_{l(c^*>c)} = \sum_{s=1}^{S}\sum_{c^*=c+1}^{C}\mathbb{1}(w_{ls} = c^*)$. Finally, the FCD for ϕ_{cs} is given by:

$$p(\phi_{cs}|...) = Beta(q_{cs}^{(1)} + \alpha_0, q_{cs}^{(0)} + \alpha_1)$$

where $q_{cs}^{(j)}$ is the number of elements assigned to group c and for which $y_{ls} = j$ (i.e., $q_{cs}^{(j)} = \sum_{l=1}^{L}\mathbb{1}(w_{ls} = c, y_{ls} = j)$).

3.2. Binomial model.

For this model, we have n_{ls} elements for each sampling unit l and variable s. The community membership status of the i-th element is denoted by W_{ils}, where $Z_{lc} = \sum_{s=1}^{S}\sum_{i=1}^{n_l}\mathbb{1}(W_{ils} = c)$, and its probability is similar to the one for the Bernoulli model:

$$p(W_{ils} = c^*|...) = \frac{\theta_{lc^*}\phi_{c^*s}^{x_{ils}}(1-\phi_{c^*s})^{1-x_{ils}}}{\sum_{c=1}^{C}\theta_{lc}\phi_{cs}^{x_{ils}}(1-\phi_{cs})^{1-x_{ils}}} \tag{6}$$

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

$$p(\phi_{cs}|...) = Beta\left(q_{cs}^{(1)} + \alpha_0, q_{cs}^{(0)} + \alpha_1\right) \tag{7}$$

where, similar to the Bernoulli model, $q_{cs}^{(j)} = \sum_{l=1}^{L}\sum_{i=1}^{n_{ls}}\mathbb{1}(x_{ils} = j, w_{ils} = c)$.
Finally, the FCD for V_{lc} is given by:

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

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

3.3. 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:

$$p(W_{il} = c^*|...) = \frac{\theta_{lc^*}\phi_{sc^*}}{(\theta_{1l}\phi_{s1} + \cdots + \theta_{Cl}\phi_{sC})} \tag{9}$$

In this equation, W_{il} is the group assignment of element i in location l, such that $Z_{lc} = \sum_{i=1}^{n_l} \mathbb{1}(W_{il} = c)$, and it can be sampled from a categorical distribution. Since we assumed a conjugate prior for $\boldsymbol{\phi}_c$ with $c \in \{1, \ldots, C\}$, the Full Conditional Distribution for this vector of parameters is a straight-forward Dirichlet distribution:

$$p(\boldsymbol{\phi}_c|...) = Dirichlet([q_{c1} + \beta_1, \ldots, q_{cS} + \beta_S]) \tag{10}$$

where $q_{cs} = \sum_{l=1}^{L} \sum_{i=1}^{n_l} \mathbb{1}(w_{il} = c, x_{il} = s)$.
Finally, the FCD for V_{lc} is given by:

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

where $z_{l(c^*>c)}$ is the total number of elements in observation l classified in clusters larger than c. This quantity is given by $z_{l(c^*>c)} = \sum_{i=1}^{n_l} \sum_{c^*=c+1}^{C} \mathbb{1}(w_{il} = c^*)$.

4. The Rlda package.

We found five other packages that can fit the Latent Dirichlet Allocation model. Hornik and Grün (2011) developed the **topicmodels** package for which has two LDA implementations: one uses the variational inference (as described in Blei *et al.* (2003)) and the other uses Gibbs Sampling based on Phan and Nguyen (2013). Similarly, Jones (2016) proposed the **textmineR** which relies on Gibbs Sampling to estimate the topics in a corpus structure. Chang (2012) developed the **lda** package which includes the mixed-membership stochastic blockmodel (Airoldi, Blei, Fienberg, and Xing 2008), supervised Latent Dirichlet Allocation - sLDA (Mcauliffe and Blei 2008) and Correspondence-Latent Dirichlet Allocation - corrLDA (Blei and Jordan 2003). Finally, 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 relies on the EM Algorithm and uses a backward model selection approach to determine the best number of groups. Finally, the last package we found was the **LDAvis**. This package presents a tool to create an interactive web-based visualization which can help users interpret the topics that result from LDA.

None of these packages adopt the truncated stick-breaking prior which enables the selection of the optimal number of clusters and regularizes model results. Furthermore, the fact that our model is fit within the Bayesian paradigm can be useful when the user already expects

certain "topics" or "groups" but is unsure about the other ones. This information can potentially be incorporated through the priors adopted for the analysis (Garthwaite, Kadane, and O'Hagan 2005). Furthermore, none of them use other distributions besides the Multinomial outcome for the dependent variable. Thus, our **Rlda** package complements current LDA approaches already available in R. The package **Rlda** is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=Rlda. The returned objects are S3 class that enables the use of the common `plot` and `summary` methods to facilitate interpretation of model results.

5. Examples.

In this section we provide examples of the Multinomial, Binomial and Bernoulli entries models focused on Marketing, Remote Sensing and Ecology fields, respectively. In each subsection we first start with some motivation about the specific problem and then we show how the **Rlda** package can be used to analyze the corresponding dataset.

5.1. Marketing.

It is well known that attracting a new customer is often considerably costlier 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. Therefore, our first application using the **Rlda** package considers the LDA for Multinomial entries applied to the field of Marketing. Specifically, we are interested in characterizing firms based on their consumers' complaints.

The data came from the 2015 Consumer Complaint Database and consist of complaints received by the **Bureau of Consumer Financial Protection** in US regarding financial products and services. In this example, we work only with credit card complaints. This dataset 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, and unsolicited issuance of credit card. In this case, each sampling unit represents a firm and each variable represents an issue.

The characterization of firms provided by our analysis can be useful to reveal commonalities 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, it is first necessary to create a matrix where each cell represents the total number of cases observed for each sampling unit and type of complaint.

```
R> library(Rlda)
R> #Read the Complaints data
R> data(complaints)
R> #Create the abundance matrix
R> library(reshape2)
R> mat1<- dcast(complaints[ ,c("Company","Issue")],
         +       Company~Issue, length,
         +       value.var="Issue")
```

```
R> #Create the rowname
R> rownames(mat1)<- mat1[,1]
R> #Remove the ID variable
R> mat1<- mat1[,-1]
```

To use the `rlda.multinomial` method we need to specify several arguments. Specifically, we need to pass the matrix with the Multinomial data (`data`), the number of clusters (`n_community`), the hyperparameters associated with our priors (`beta` and `gamma`), the number of gibbs iterations (`n_gibbs`), if the log-likelihood should be summed to the log priors distributions (`ll_prior`) and the last argument specifies if the progress bar should be presented or not (`display_progress`). In this problem we set the maximum number of clusters to 30 and the number of Gibbs Sampling iterations to 1000.

```
R> #Set seed
R> set.seed(9292)
R> #Hyperparameters for each prior distribution
R> beta<- rep(1,ncol(mat1))
R> gamma<- 0.01
R> #Execute the LDA for the Multinomial entry
R> res<- rlda.multinomial(data=mat1, n_community=30, beta, gamma,
                        + n_gibbs=1000, ll_prior=TRUE, display_progress=FALSE)
```

In the **Rlda** package the three main methods `rlda.multinomial`, `rlda.binomial` and `rlda.bernoulli` return a `Rlda` S3 object which can be used in a straight forward fashion with the `plot` and `summary` functions. For instance, we can visually evaluate the convergence by examining Figure 1:

```
R> #Get the logLikelihood
R> ll<- res$logLikelihood
R> #Plot the log-likelihood
R> plot(ll, type="l", xlab="Iterations",
                  +   ylab="Log(likel.)+log(prior)")
R> abline(v=700,col='grey')
```

The `plot.rlda` method also outputs the mean of the posterior distribution for the matrices Θ and Φ. To this end, the user has to define the percentage of the Gibbs iterations that must be used as burn in. For instance `plot(res, burnin= 0.1)` determines that 10% of the first Gibbs iterations are eliminated before plotting the results of `Rlda`.

Samples of our parameter estimates are given in the `Theta` and `Phi` matrices, where each line in these matrices contains the result of a Gibbs iteration. Thus, parameter estimates can be obtained by averaging the results in each column of these matrices after discarding the burn-in iterations. This can be quickly done using the function `summary`, as illustrated below:

```
R> #Get the Theta Estimate
R> Theta<-summary(res, burnin= 0.1, silent= TRUE)$Theta
```

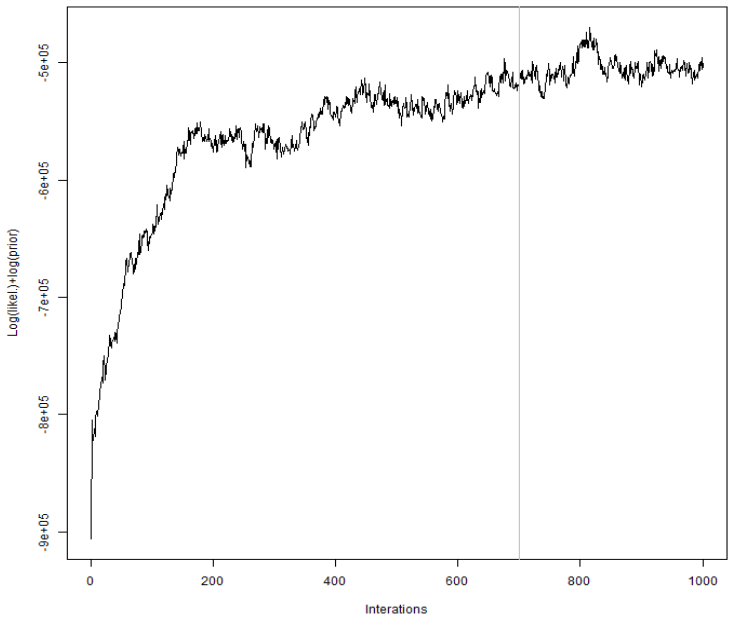

Figure 1: Algorithm convergence. Log(likelihood)+log(prior) (y-axis) is displayed as a function of algorithm iterations (x-axis). Vertical gray lines depict the iteration after which the algorithm is judged to have converged.

Similar to the `plot` method, in the `summary` S3 method we require the user to define the percentage of the Gibbs iterations that must be used as burn in (`burnin` argument) and if the text summary should be shown (`silent` argument).

The `Theta` matrix has a sparse structure since our truncated stick-breaking prior tends to reduce the total number of dominant clusters. In this matrix, each cell contains the estimated probability of the l-th firm being allocated to cluster c. A useful way to explore the results from this matrix is using interactive 3D graphics:

```
R> library('rgl')
R> library('car')
R> scatter3d(x=Theta[,'Cluster 1'], y=Theta[,'Cluster 2'], z=Theta[,'Cluster 3'],
             + surface=F, xlab='Cluster 1', ylab='Cluster 2', zlab='Cluster 3',
             + labels=rownames(Theta), id.n=20)
```

In particular, the firms that best represent each of these clusters (i.e., top firms in each cluster with probability greater than 0.3) are given in Table 1.

Cluster	Firms
1	Portfolio Recovery Associates, Inc., Encore Capital Group, Sterling Jewelers Inc.
2	Comerica, PNC Bank N.A., Barclays PLC
3	Continental Finance Company, LLC, Citibank, Amex
4	Synchrony Financial, Regions Financial Corporation, TD Bank US Holding Company
6	Discover
7	U.S. Bancorp
27	PayPal Holdings, Inc.

Table 1: Representative firms for each cluster

Interestingly, although there are several firms with a high proportion of complains arising from clusters 1-4, complaints associated with clusters 6, 7, and 27 arise from a very small subset of companies. We display the complaint profiles associated with the main clusters by examining the `Phi` matrix. Differently from clusters 1-4, clusters 6, 7, and 27 have substantially different complaint profiles (Figure 3). In particular, Cluster 6 (represented by Discover) had a very high proportion of complaints arising from "Closing:Cancelling accounts" while Cluster 7 (represented by US Bancorp) had a high proportion of "Advertising and Marketing" and "Rewards" complaints. Finally, Cluster 27 (represented by PayPal Holdings, Inc) had a high proportion of "Unsolicited credit card" complaints. These last results are likely due to a change in PayPal operations. Standard PayPal accounts were changed to revolving credit accounts but several costumers claimed that they were unaware of these changes.

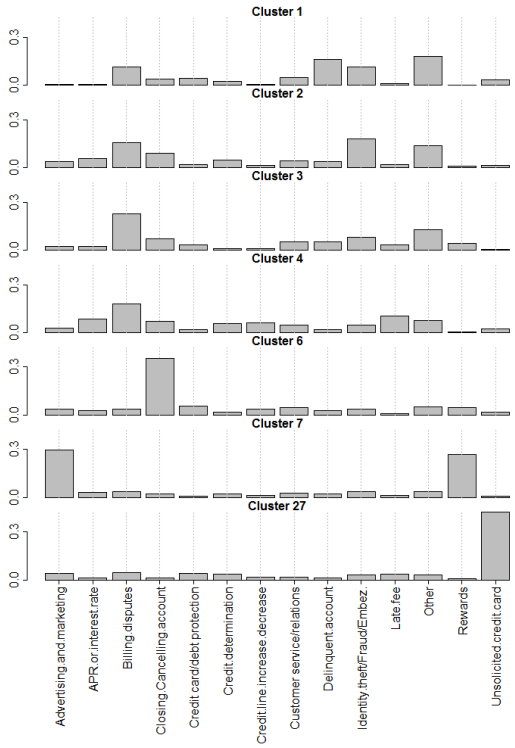

Figure 2: Complaint profiles for the main clusters. Only the main types of complaints are displayed.

5.2. Remote Sensing.

Because pixels in remote sensing imagery are often large enough to encompass different types of material within each pixel, there has been great interest in the development of methods that enable researchers to estimate the proportion of the constituent materials (often called endmembers in the remote sensing literature). Indeed, numerous spectral unmixing algorithms have already been developed in the literature, with multiple approaches used for the dimension reduction, endmember determination, and inversion stages (Keshava 2003).

The key characteristics of the method that we propose here is that it is an unsupervised method (i.e., it does not require the a priori determination of endmembers) that enforces parsimony through our truncated stick-breaking prior. Furthermore, differently from many of the currently existing methods for spectral unmixing, our model explicitly acknowledges the discreteness of the digital numbers used in remote sensing systems and the range of possible values these numbers can take.

In our example, we rely on Landsat TM 5 imagery from 2010 of the Iquitos-Nauta road in the Peruvian Amazon. This area has multiple land-use land-cover (LULC) types and is the site in which we have studied the effect of these different LULC types on the malaria vector *Anopheles darlingi* (Valle *et al.* 2014). The first step of our analysis consists of choosing a subset of the pixels for the estimation of the endmembers:

```
R> #Loading the Landsat data
R> data(Landsat)
R> #Total number of bands
R> nbands<- ncol(Landsat)-2
R> #Temporary dataset
R> nomes<- paste('b',1:nbands,sep='')
R> dat1<- data.matrix(Landsat[,nomes])
R> #let's change the range of our data to start at zero.
R> #Variation outside this range does not help us explain much
R> tmp<- apply(dat1,2,range)
R> min1<- tmp[1,]
R> dat2<- dat1-matrix(min1,nrow(dat1),nbands,byrow=T)
R> tmp<- apply(dat2,2,range)
R> max1<- tmp[2,]
R> #select a sample of the most different
R> #pixels that I have (10*30 pixels for each band)
R> dat3<- unique(dat2)
R> dat4<- numeric()
R> for (i in 1:nbands){
R>   seq1<- seq(from=tmp[1,i],to=tmp[2,i],length.out=10)
R>   nome<- paste('b',i,sep='')
R>   for (j in 1:10){
R>     dist<- abs(dat3[,nome]-seq1[j])
R>     ind<- order(dist)[1:30]
R>     dat4<- rbind(dat4,dat3[ind,])
R>   }
R> }
```

```
R> Landsat_sample<- unique(dat4)
R> Landsat_sample<- Landsat_sample[,c('b7','b6','b5','b4','b3','b2','b1')]
```

Then, we identify the endmembers using a subset of the image (matrix `Landsat_sample`). To do this, we need to supply to the algorithm the maximum number of elements for the binomial distribution, given by the matrix `max2`, and the prior hyper-parameters `a.phi` and `b.phi`.

```
R> tmp            <- apply(Landsat_sample,2,max)
R> npix           <- nrow(Landsat_sample)
R> nbands         <- ncol(Landsat_sample)
R> max2           <- as.data.frame(matrix(tmp,npix,nbands,byrow=T))
R> #Define the hyperparameters
R> a.phi  <- 1
R> b.phi  <- 1
R> gamma=1
R> ngibbs <- 10000
R> ncomm  <- 5
R> #Execute the Binomial LDA
R> z <- rlda.binomial(data= Landsat_sample, pop= max2, n_community=ncomm,
                  + alpha0= a.phi, alpha1= b.phi, gamma= gamma,
                  + n_gibbs= ngibbs, ll_prior=TRUE, display_progress=TRUE)
```

In a similar way as presented in the Marketing example, we can get the `Phi` and `Theta` matrices after discarding the burn-in iterations. Using these average parameter estimates, the next step is to make predictions of the proportion of these endmembers for the rest of the image (i.e., matrix "Landsat"). Using the results from `rlda.binomial` we can obtain some predictions for the whole dataset using `predict.rlda`:

```
R> #Make the prediction
R> Landsat2<- Landsat[,c(-1,-2)]
R> #Create a matrix with all possible combinations of proportions
R> res<- predict.rlda(object=z, data=Landsat2,
R>         +nclus=5, burnin=0.1, places.round=0)
R> pred<- cbind(Landsat,res)
R> colnames(pred)[1:2]<- c('x','y')
```

The `predict.rlda` has five arguments, the S3 object from function `rlda.binomial` through the argument `object=z`, while the data to be used for prediction is given by the argument `data=Landsat2`. The argument `nclus=5` represents the number of clusters used in the prediction, but, if the user chooses `nclus=NA`, all available clusters will be used. Similar to the argument in the other functions described before, `burnin=0.1` represents the percentage of burn-in observations. Finally, `places.round=0` determines how many of the digits will be truncated in the prediction dataset. Truncating digits in the prediction dataset can significantly speed up the prediction process but will generate only approximate results.

Finally, we display the proportion of each cluster throughout the landscape:

```
R> #Make the plot
R> library('gplots')
R> seq1<- seq(from=0,to=1,by=0.05)
R> n<- length(seq1)
R> par(mfrow=c(1,5),mar=c(1,1,4,1),oma=c(2,2,0,0))
R> for (i in 1:5){
R>   var1<- pred[,paste('prop',i,sep='')]
R>   ind<- 1+var1/0.05
R>   palette(rev(rich.colors(n)))
R>   plot(pred$x,pred$y,pch=15,col=ind,main=paste('Cluster',i),
R>        + xlim=c(671000,675000),cex=1,xlab='',ylab='',
R>        + ylim=c(-453000,-443000),xaxt='n',yaxt='n',cex.main=3)
R> }
```

Figure 3 reveals that clusters 1 and 3 are associated with deforested areas and bare soil. Cluster 3 delimits very well small water bodies that exist in the region. Cluster 4 is substantially more diffuse, potentially representing forests. These results suggest that the proposed method can be used to classify regions in an unsupervised way to solve pixel unmixing problems.

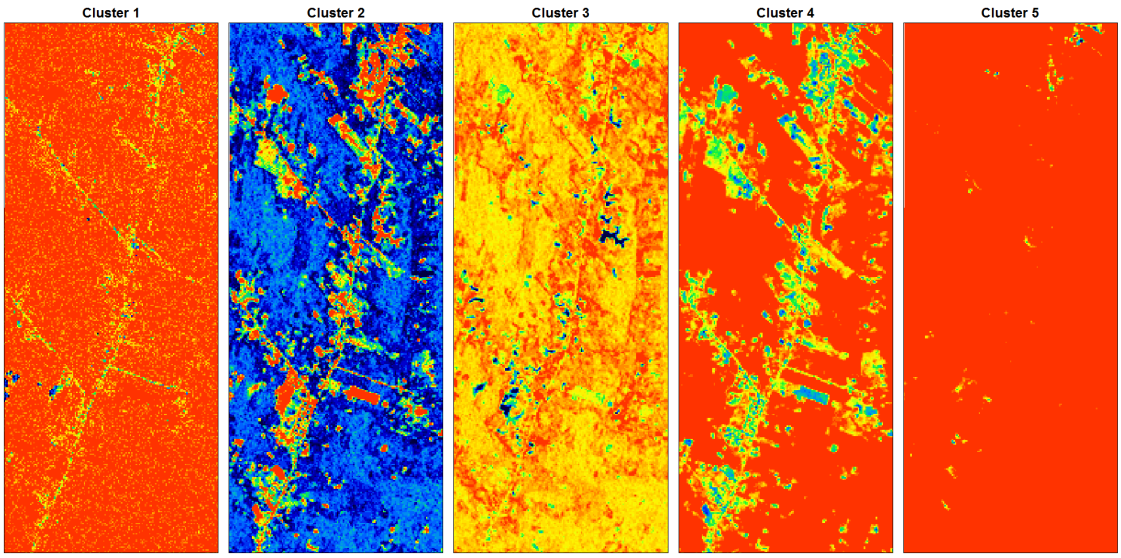

Figure 3: Clusters: Iquitos-Nauta road. Hot to cold colors indicate low to high probabilities, respectively

5.3. Ecology.

LDA has been proposed to analyze biodiversity data in Ecology (Valle *et al.* 2014) given its advantages over traditional clustering methods. For example, LDA provides the probabilities of a sample unit (e.g., a site) to be part of multiple component communities. However, traditional clustering methods allow a sample unit to be part of only one component community. This is the same for species: LDA allows each species to be part of multiple component communities. In addition, LDA accommodates missing values and provides coherent estimates

of uncertainty (Valle *et al.* 2014). However, the method proposed by Valle *et al.* (2014) only applies to abundance data. In many ecological studies, it is not possible to determine the total number of individuals per species in each sampling unit. As a result, these data are often summarized into binary presence/absence matrices (1 and 0, respectively) (Pearce and Boyce 2006). In this paper, we updated Valle *et al.* (2014)'s method to analyze presence/absence data, which significantly broadened the scope of the use of LDA in Ecology.

In this example we used `data("presence")`, which includes presence/absence information on 13 species at 386 forested locations (Moisen, Freeman, Blackard, Frescino, Zimmermann, and Edwards 2006). We analyze these data using the `rlda.bernoulli` S3 method.

```
R> #Load data
R> data(presence)
R> #Set seed
R> set.seed(9842)
R> #Hyperparameters for each prior distribution
R> gamma <-0.01
R> alpha0<-0.01
R> alpha1<-0.01
R> #Execute the LDA for the Binomial entry
R> res<-rlda.bernoulli(presence, 10, alpha0, alpha1, gamma,
                       +       5000, TRUE, FALSE)
```

We can visually evaluate the cluster distribution across species, after the *burn-in phase* in Figure 4: in this type of graph each slice size is proportional to the probability of belonging, and it is possible to note that some species of trees are more or less associated with some clusters. For example, QUGA species is more associated with Clusters 7 and 2 as well ACGR3 species. More detailed examples of using LDA in ecological studies can be found at Valle *et al.* (2014).

6. Conclusion.

The goal of this paper was to describe the Bayesian LDA model for mixed membership clustering based on different types of discrete data. We have demonstrated how to use the model for Multinomial entry in the Marketing example, Binomial trial using the Landsat dataset, and Bernoulli trial using ecological data.

One of the main properties of the model presented here is the fact that this model adopts the truncated stick-breaking prior which enables the selection of the optimal number of clusters by regularizing model results. The next step in the development of the model presented here is the possibility to work with explanatory variables within our algorithms, which can be useful to make inference on the drivers of the probability of each cluster.

Appendix.

In this Appendix we provide the derivation for the Full Conditional Distributions associated with each model.

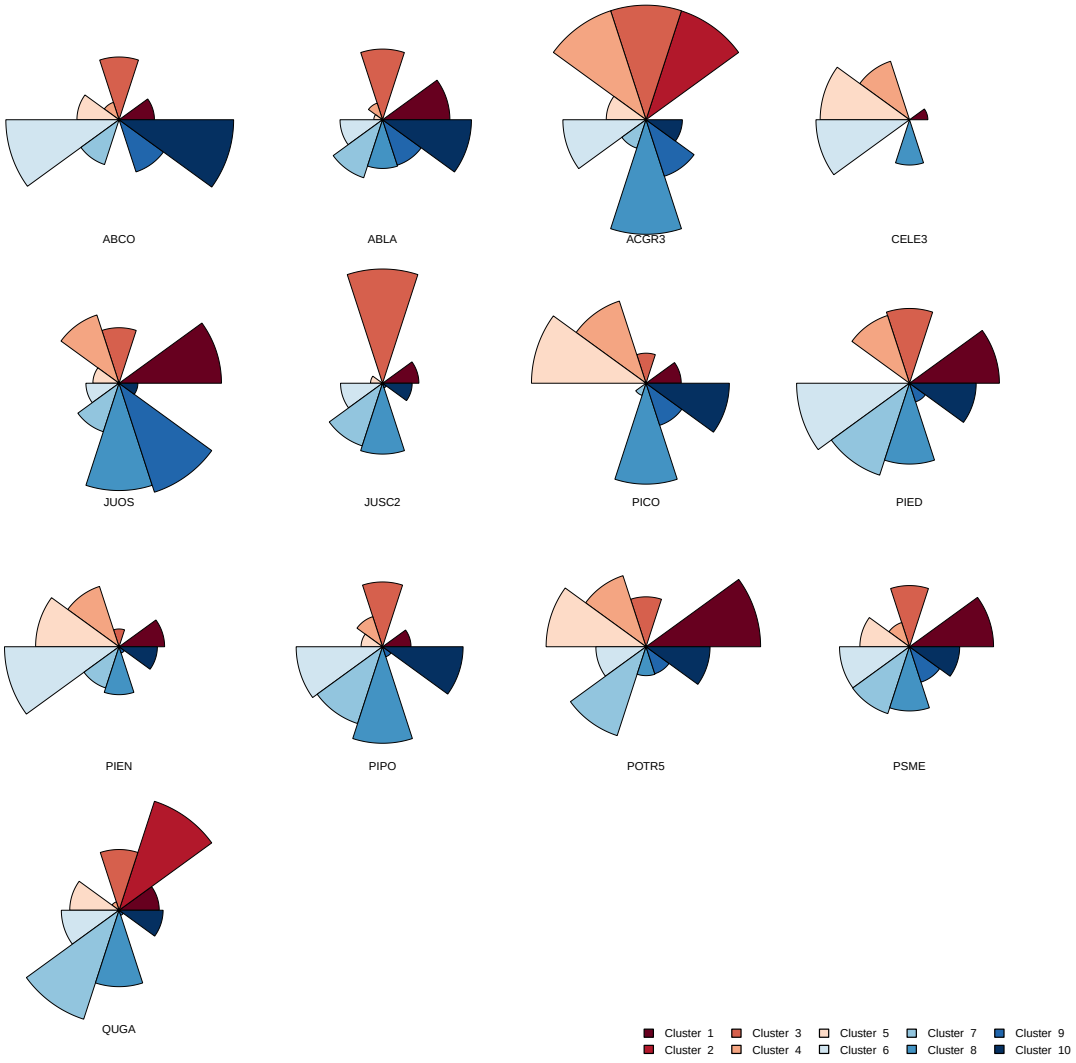

Figure 4: Cluster distribution across species.

Bernoulli model.

For W_{ls}:

$$\begin{aligned} p(W_{ls} = c^*|...) &= k \times Cat(W_{ls} = c^*|\boldsymbol{\theta_l}) \times Bernoulli(y_{ls}|\phi_{c^*s}) \\ &= k \times \theta_{lc^*} \times \phi_{c^*s}^{y_{ls}}(1-\phi_{c^*s})^{1-y_{ls}} \end{aligned}$$

Since W_{ls} is a categorical random variable with support in $\mathscr{Z} = (1, 2, \ldots, C)$, the sum of the probabilities for all elements must equal one. Therefore, the constant k is given by:

$$k = \sum_{c=1}^{C} \theta_{lc} \times \phi_{cs}^{y_{ls}}(1-\phi_{cs})^{1-y_{ls}}$$

As a result, W_{ls} can be sampled from a categorical distribution.

For V_{ls}:

$$\begin{aligned} p(V_{ls}|...) &\propto Binomial(z_{lc}|z_{lc} + z_{l(c^*>c)}, V_{lc}) \times Beta(V_{lc}|1, \gamma) \\ &\propto V_{lc}^{z_{lc}}(1-V_{lc})^{z_{l(c^*>c)}} \times (1-V_{lc})^{\gamma-1} \\ &\propto V_{lc}^{(z_{lc}+1)-1}(1-V_{lc})^{(z_{l(c^*>c)}+\gamma)-1} \\ p(V_{ls}|...) &= Beta(z_{lc}+1, z_{l(c^*>c)}+\gamma) \end{aligned}$$

For ϕ_{cs}:

$$\begin{aligned} p(\phi_{cs}|...) &\propto [\prod_{l=1}^{L} Bernoulli(y_{ls}|\phi_{cs})^{\mathbb{1}(w_{ls}=c)}] \times Beta(\phi_{cs}|\alpha_0, \alpha_1) \\ &\propto [\prod_{l=1}^{L} \phi_{cs}^{\mathbb{1}(w_{ls}=c, y_{ls}=1)}(1-\phi_{cs})^{\mathbb{1}(w_{ls}=c, y_{ls}=0)}] \times \phi_{cs}^{\alpha_0-1}(1-\phi_{cs})^{\alpha_1-1} \\ &\propto \phi_{cs}^{\sum_{l=1}^{L}\mathbb{1}(w_{ls}=c, y_{ls}=1)+\alpha_0-1}(1-\phi_{cs})^{\sum_{l=1}^{L}\mathbb{1}(w_{ls}=c, y_{ls}=0)+\alpha_1-1} \\ p(\phi_{cs}|...) &= Beta(q_{cs}^{(1)}+\alpha_0, q_{cs}^{(0)}+\alpha_1) \end{aligned}$$

where $q_{cs}^{(j)} = \sum_{l=1}^{L} \mathbb{1}(w_{ls} = c, y_{ls} = j)$.

Binomial model.

For W_{ils}:

$$\begin{aligned} p(W_{ils} = c^*|...) &= k \times Cat(W_{ils} = c^*|\boldsymbol{\theta_l}) \times Bernoulli(x_{ils}|\phi_{c^*s}) \\ &= k \times \theta_{lc^*} \times \phi_{c^*s}^{x_{ils}}(1-\phi_{c^*s})^{1-x_{ils}} \end{aligned}$$

Since W_{ils} is a categorical random variable with support in $\mathscr{Z} = (1, 2, \ldots, C)$, the sum of the probabilities for all elements must equal one. Therefore, the constant k is given by:

$$k = \sum_{c=1}^{C} \theta_{lc} \times \phi_{cs}^{x_{ils}}(1-\phi_{cs})^{1-x_{ils}}$$

As a result, W_{ils} can be sampled from a categorical distribution.
For V_{ls}:

$$\begin{aligned} p(V_{ls}|...) &\propto Binomial(z_{lc}|z_{lc}+z_{l(c^*>c)}, V_{lc}) \times Beta(V_{lc}|1,\gamma) \\ &\propto V_{lc}^{z_{lc}}(1-V_{lc})^{z_{l(c^*>c)}} \times (1-V_{lc})^{\gamma-1} \\ &\propto V_{lc}^{(z_{lc}+1)-1}(1-V_{lc})^{(z_{l(c^*>c)}+\gamma)-1} \\ p(V_{ls}|...) &= Beta(z_{lc}+1, z_{l(c^*>c)}+\gamma) \end{aligned}$$

For ϕ_{cs}:

$$\begin{aligned} p(\phi_{cs}|...) &\propto [\prod_{l=1}^{L}\prod_{i=1}^{n_{ls}} Bernoulli(x_{ils}|\phi_{cs})^{\mathbb{1}(w_{ils}=c)}] \times Beta(\phi_{cs}|\alpha_0,\alpha_1) \\ &\propto [\prod_{l=1}^{L}\prod_{i=1}^{n_{ls}} \phi_{cs}^{\mathbb{1}(w_{ils}=c,x_{ils}=1)}(1-\phi_{cs})^{\mathbb{1}(w_{ils}=c,x_{ils}=0)}] \times \phi_{cs}^{\alpha_0-1}(1-\phi_{cs})^{\alpha_1-1} \\ &\propto \phi_{cs}^{\sum_{l=1}^{L}\sum_{i=1}^{n_{ls}}\mathbb{1}(w_{ils}=c,x_{ils}=1)+\alpha_0-1}(1-\phi_{cs})^{\sum_{l=1}^{L}\sum_{i=1}^{n_{ls}}\mathbb{1}(w_{ils}=c,x_{ils}=0)+\alpha_1-1} \\ p(\phi_{cs}|...) &= Beta(q_{cs}^{(1)}+\alpha_0, q_{cs}^{(0)}+\alpha_1) \end{aligned}$$

where $q_{cs}^{(j)} = \sum_{l=1}^{L}\sum_{i=1}^{n_{ls}} \mathbb{1}(w_{ls}=c, y_{ls}=j)$.

Multinomial model.

For W_{il}:

$$\begin{aligned} p(W_{il}=c^*|...) &= k \times Cat(W_{il}=c^*|\boldsymbol{\theta_l}) \times Cat(x_{il}=s|\phi_{sc^*}) \\ &= k \times \theta_{lc^*}\phi_{sc^*} \end{aligned}$$

Since W_{il} is a categorical random variable with support in $\mathscr{Z} = (1, 2, \ldots, C)$, the sum of the probabilities for all elements must equal one. Therefore, the constant k is given by:

$$k = \sum_{c=1}^{C} \theta_{lc} \times \phi_{cs}$$

As a result, W_{il} can be sampled from a categorical distribution.
For V_{ls}:

$$\begin{aligned} p(V_{ls}|...) &\propto Binomial(z_{lc}|z_{lc}+z_{l(c^*>c)}, V_{lc}) \times Beta(V_{lc}|1,\gamma) \\ &\propto V_{lc}^{z_{lc}}(1-V_{lc})^{z_{l(c^*>c)}} \times (1-V_{lc})^{\gamma-1} \\ &\propto V_{lc}^{(z_{lc}+1)-1}(1-V_{lc})^{(z_{l(c^*>c)}+\gamma)-1} \\ p(V_{ls}|...) &= Beta(z_{lc}+1, z_{l(c^*>c)}+\gamma) \end{aligned}$$

For $\boldsymbol{\phi_c}$:

$$p(\boldsymbol{\phi_c}|...) \propto [\prod_{l=1}^{L}\prod_{i=1}^{n_l} Cat(x_{il}|\boldsymbol{\phi_c})^{\mathbb{1}(w_{il}=c)}] \times Dirichlet(\boldsymbol{\phi_c}|\boldsymbol{\beta})$$

$$\propto [\prod_{l=1}^{L}\prod_{i=1}^{n_l} \phi_{1c}^{\mathbb{1}(x_{il}=1,w_{li}=c)} \times ... \times \phi_{Sc}^{\mathbb{1}(x_{il}=S,w_{li}=c)}] \times \phi_{1c}^{\beta_1-1} \times ... \times \phi_{Sc}^{\beta_S-1}$$

$$\phi_{1c}^{\sum_{l=1}^{L}\sum_{i=1}^{n_l}\mathbb{1}(x_{il}=1,w_{li}=c)+\beta_1-1} \times ... \times \phi_{Sc}^{\sum_{i=1}^{n_l}\mathbb{1}(x_{il}=S,w_{li}=c)+\beta_S-1}$$

$$p(\phi_c|...) = Dirichlet([q_{c1}+\beta_1, ..., q_{cS}+\beta_S])$$

where $q_{cs} = \sum_{i=1}^{n_l} \mathbb{1}(x_{il} = s, w_{li} = c)$.

Acknowledgements.

We thank the numerous comments and suggestions provided by Justin Millar. This work was partly supported by the US Department of Agriculture National Institute of Food and Agriculture McIntire Stennis project 1005163 and by the US National Science Foundation award 1458034 to DV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008). "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research*, **9**(Sep), 1981–2014.

Blei DM, Jordan MI (2003). "Modeling annotated data." In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134. ACM.

Blei DM, Ng AY, Jordan MI (2003). "Latent dirichlet allocation." *Journal of machine Learning research*, **3**(Jan), 993–1022.

Chang J (2012). "lda: Collapsed Gibbs sampling methods for topic models. R package version 1.4.2."

Erosheva EA, Fienberg SE (2005). "Bayesian mixed membership models for soft clustering and classification." In *Classification - The Ubiquitous Challenge*, pp. 11–26. Springer.

Garthwaite PH, Kadane JB, O'Hagan A (2005). "Statistical methods for eliciting probability distributions." *Journal of the American Statistical Association*, **100**(470), 680–701.

Griffiths TL, Steyvers M (2004). "Finding scientific topics." *Proceedings of the National academy of Sciences*, **101**(suppl 1), 5228–5235.

Hornik K, Grün B (2011). "topicmodels: An R package for fitting topic models." *Journal of Statistical Software*, **40**(13), 1–30.

Jones TW (2016). "textmineR: Functions for Text Mining and Topic Modeling. R package version 2.0.2."

Keshava N (2003). "A survey of spectral unmixing algorithms." *Lincoln Laboratory Journal*, **14**(1), 55–78.

Kotler P, Armstrong G (2006). "Principles of marketing management."

Lee S, Baker J, Song J, Wetherbe JC (2010). "An empirical comparison of four text mining methods." In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pp. 1–10. IEEE.

Lienou M, Maître H, Datcu M (2010). "Semantic annotation of satellite images using latent dirichlet allocation." *IEEE Geoscience and Remote Sensing Letters*, **7**(1), 28–32.

Mahajan A, Dey L, Haque SM (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*, volume 1, pp. 423–426. IEEE.

Mcauliffe JD, Blei DM (2008). "Supervised topic models." In *Advances in neural information processing systems*, pp. 121–128.

Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards TC (2006). "Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods." *ecological modelling*, **199**(2), 176–187.

Pearce JL, Boyce MS (2006). "Modelling distribution and abundance with presence-only data." *Journal of applied ecology*, **43**(3), 405–412.

Phan XH, Nguyen CT (2013). "GibbsLDA++, AC/C++ implementation of latent dirichlet allocation (LDA) using Gibbs sampling for parameter estimation and inference."

Pritchard JK, Stephens M, Donnelly P (2000). "Inference of population structure using multilocus genotype data." *Genetics*, **155**(2), 945–959.

Roberts ME, Stewart BM, Tingley D (2014). "stm: R package for structural topic models." *R package*, **1**, 12.

Tirunillai S, Tellis GJ (2014). "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation." *Journal of Marketing Research*, **51**(4), 463–479.

Tsai FS (2011). "A tag-topic model for blog mining." *Expert Systems with Applications*, **38**(5), 5330–5335.

Valle D, Baiser B, Woodall CW, Chazdon R (2014). "Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method." *Ecology letters*, **17**(12), 1591–1601.

Westgate MJ, Barton PS, Pierson JC, Lindenmayer DB (2015). "Text analysis tools for identification of emerging topics and research gaps in conservation science." *Conservation Biology*, **29**(6), 1606–1614.

Affiliation:

Pedro Albuquerque
Faculdade de Economia, Administração e Contabilidade
University of Brasília
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
Institute of Food and Agricultural Sciences
University of Florida
408 McCarty Hall C
PO Box 110339
Gainesville, FL 32611-0410.
E-mail: drvalle@ufl.edu
URL: http://denisvalle.weebly.com/

Daijiang Li
Institute of Food and Agricultural Sciences
University of Florida
408 McCarty Hall C
PO Box 110339
Gainesville, FL 32611-0410.
E-mail: daijianglee@gmail.com
URL: http://daijiang.name/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd