Internship Experience Report

Daniel Basso Ribas

08 June 2016 - 16 August 2016

1 INTRODUCTION

This report aims to describe the activities carried out during the Internship I participated in Trintiy College Dublin, with the supervision of Dr. Jason Wyse (Assistant Professor in the School of Computer Science and Statistics). The main goal of the internship was to learn about Latent Class Analysis and how Bayesian Inference could be used to perform it. Secondary goals included working on the code of a R package called **`BayesLCA`**, trying to improve the existing functions.

2 DESCRIPTION OF THE INTERNSHIP

The internship can be divided in two phases:

1. Learning about Bayesian Latent Class Analysis;
2. Applying the acquired knowledge on code of **`BayesLCA`** R package.

The fist phase was accomplished by reading two articles about the subject: *Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler* (White, Wyse, and Murphy 2016); and *BayesLCA: An R Package for Bayesian Latent Class Analysis* (White and Murphy 2014). During this period, Jason asked me to deduce the posterior distribution of certain variables of interest. This took approximately one month until moving to the next phase. A brief literature review about the subject will be presented on section 2.1.

After that, I was presented with the **`BayesLCA`** package, and asked to try to generalize the function **`gibbs.blca()`** to work with a wider range of datasets, based on what I've learned. More details of the work carried are presented on section 2.2.

2.1 Literature Review

Given N individuals that were observed M categorical variables, Latent Class Analysis can be used to group these individuals into homogeneous G groups. Each category $\{1,2,\dots,C_m\}$ has its own probability θ of occurring, that is different for each variable and for each group, and will be denoted θ_{gmc} (probability of variable m assuming category c for the group g). To calculate these probabilities and the prevalence (τ) of each group, both articles describe on section 2 use Bayesian approaches. White and Murphy (2014) describes three algorithms to compute the parameters of interest for dicotomial variables: Expectation Maximization, Gibbs Sampling and Variation Bayes. The work done on section 2.2 focused on the Gibbs Sampler algorithm. This algorithm involves calculating the conditional probability distributions of the variables of interest and sampling from them one at a time, using the sample value in the next iteration. The conditional probabilities necessary to use it are described below:

$$\theta_{gm}^{(k+1)} \sim \text{Beta}\left(\sum_{i=1}^{N} X_{im} Z_{ig}^{(k)} + \alpha_{gm}, \sum_{i=1}^{N} Z_{ig}^{(k)} (1 - X_{im}) + \beta_{gm}\right) \tag{1}$$

$$\tau_{g}^{(k+1)} \sim \text{Dirichlet}\left(\sum_{i=1}^{N} Z_{i1}^{(k)} + \delta_1, \cdots, Z_{iG}^{(k)} + \delta_G\right) \tag{2}$$

$$Z_{i}^{(k+1)} \sim \text{Multinomial}\left(1, \frac{\tau_1^{(k+1)} p\left(X_i \mid \theta_1^{(k+1)}\right)}{\sum_{h=1}^{G} \tau_h^{(k+1)} p\left(X_i \mid \theta_h^{(k+1)}\right)}, \cdots, \frac{\tau_G^{(k+1)} p\left(X_i \mid \theta_G^{(k+1)}\right)}{\sum_{h=1}^{G} \tau_h^{(k+1)} p\left(X_i \mid \theta_h^{(k+1)}\right)}\right) \tag{3}$$

White, Wyse, and Murphy (2016) describes an alternative Gibbs Sampler algorithm, that can work with categorical data instead of just dicotomical one. It involves using a Dirachlet prior instead of a Beta to sample from θ. The full conditional distributions for θ, τ and Z aren't describle in this article and Jason asked me to calculate them. This are the results:

$$\theta_{gmc}^{(k+1)} \sim \text{Dirichlet}\left(\sum_{i=1}^{N} Z_{ng}^{(k)} I(X_{nm} = c), \cdots, Z_{nG}^{(k)} I(X_{nm} = c)\right) \tag{4}$$

$$\tau_{g}^{(k+1)} \sim \text{Dirichlet}\left(\sum_{i=1}^{N} Z_{in1} + \alpha, \cdots, \sum_{i=1}^{N} Z_{inG} + \alpha\right) \tag{5}$$

$$Z_{i}^{(k+1)} \sim \text{Multinomial}\left(1, \frac{\tau_1^{(k+1)} \prod_{c=1}^{C} \theta_{gmc}^{I(X_{nm}=c)}}{\sum_{h=1}^{G} \tau_h^{(k+1)} \prod_{c=1}^{C} \theta_{gmc}^{I(X_{nm}=c)}}, \cdots, \frac{\tau_G^{(k+1)} \prod_{c=1}^{C} \theta_{Gmc}^{I(X_{nm}=c)}}{\sum_{h=1}^{G} \tau_h^{(k+1)} \prod_{c=1}^{C} \theta_{gmc}^{I(X_{nm}=c)}}\right) \tag{6}$$

These equations are essential to conduct the work described on section 2.2, as follows.

2.2 Work Carried Out

After the deducting the posterior distributions, I received the task to try to modify the function **gibbs.blca()** from the package **BayesLCA**. This function draw samples from the conditional probabilities (1), (2) and (3) to compute the parameters of interest (and also their empirical distribution). This means that it will only work for dicotomical variables. The goal was to make it work for categorical data. To do this, the "trick" is changing the posterior for θ, from the Beta (1) to Dirichlet (4). This function code has 230 lines, so the first step was to identify where were the priors and Gibb Sampler were located. After some time of inspection, I found them, as shown below:

```r
# Priors

Z<-unMAP(sample(1:G,size=N,replace=TRUE)) # Z prior
tau<-rdirichlet(1,delta+colSums(Z)) # Tau prior
for(g in 1:G) theta[g,]<-rbeta(M,alpha+colSums(Z[,g]*X), beta+colSums(Z[,
g]*(1-X))) # Theta prior

# Gibbs sampler
```

```
for(g in 1:G)   W[,g]<-tau[g]*apply(theta[g,]^t(X) * (1-theta[g,])^t(1-X)
,2,prod)
Z<-Zsamp(W, counts.n) # Z sample
tau<-rdirichlet(1,delta+colSums(Z)) # Tau Sample
for(g in 1:G) theta[g,]<-rbeta(M,alpha+colSums(Z[,g]*X), beta+colSums(Z[,
g]*(1-X))) # Theta Sample
```

Like mentioned before, for this to work with more categories, θ sample must be drawn from a Dirachlet. Finding a way to properly doing this took some time. After trying different approaches, the solution below came out:

```
for(c in 1:C){
  for (g in 1:G){
    c.store[c, ,g] <- t(Z[,g])%*%(X == c)
  }
}

for(m in 1:M) {
  for (g in 1:G){
    theta[,m,g] <- rdirichlet(1,beta + c.store[,m,g])
  }
}
```

Other changes had to be made in order to the code to work, but this was the main one. To test it, a dataset with four variables, three categories for each variable and two groups. The values for the θ_{gmc}'s and τ_g's are the following:

```
true.tau # Tau values (prob for each group)

## [1] 0.3 0.7

true.theta # Theta valuels for each group

## , , 1
##
##      [,1] [,2] [,3] [,4]
## [1,]  0.5  0.3  0.1  0.1
## [2,]  0.3  0.4  0.7  0.1
## [3,]  0.2  0.3  0.2  0.8
##
## , , 2
##
##      [,1] [,2] [,3] [,4]
## [1,]  0.2  0.7  0.5  0.5
## [2,]  0.2  0.1  0.1  0.3
## [3,]  0.6  0.2  0.4  0.2
```

After running the modified code, the calculated values were:

```
x$classprob # Tau values (prob for each group)
```

```
## [1] 0.2754265 0.7245735

x$itemprob # Theta valuels for each group

## , , 1
## 
##           [,1]      [,2]       [,3]       [,4]
## [1,] 0.3898422 0.2212835 0.09390383 0.06735767
## [2,] 0.3336630 0.4731658 0.71583296 0.14951111
## [3,] 0.2764949 0.3055507 0.19026320 0.78313123
## 
## , , 2
## 
##           [,1]       [,2]      [,3]      [,4]
## [1,] 0.2058791 0.68486671 0.4811691 0.5172480
## [2,] 0.2337836 0.09024182 0.1203611 0.2718266
## [3,] 0.5603373 0.22489147 0.3984698 0.2109254
```

It is possible to see that the code is working for this particular example. But, sometimes, a problem called label switching occurs. Label switching means that the estimation of a particular parameter gets its group switched between iterations. Because the final estimative is the mean of all estimated parameters at each iteration, this can became a problem. During the work, sometimes label switching happened, but I could not generate an example of it using the last updated code. An example of it can be found in White and Murphy (2014).

3 CONCLUSION

This internship showed how difficult it is and how much effort needs to be made to in order to create an R package. During the three months spent on the project, little progress was made in relation to creating new code in R, because a lot of time had to be spent learning how the algorithm desired to be implemented worked. Nevertheless, the main objective was achieved: Learn more about Bayesian Latent Class Analysis and how it can be useful to simultaneously group individuals and calculate the variables probability off occuring.

4 REFERENCES

White, Arthur, and Thomas Murphy. 2014. “BayesLCA: An R Package for Bayesian Latent Class Analysis.” *Journal of Statistical Software* 61 (1): 1–28. doi:10.18637/jss.v061.i13.

White, Arthur, Jason Wyse, and Thomas Brendan Murphy. 2016. “Bayesian Variable Selection for Latent Class Analysis Using a Collapsed Gibbs Sampler.” *Statistics and Computing* 26 (1): 511–27. doi:10.1007/s11222-014-9542-5.