Master in *Statistical Sciences*

Lab 3 *Latent variable models*

Exercise 1
Latent class analysis on the carcinoma data set

The carcinoma data from Agresti (2002) consist of seven dichotomous variables that represent the ratings by seven pathologists of 118 slides on the presence or absence of carcinoma. The purpose of studying these data is to model 'interobserver agreement' by examining how subjects might be divided into groups depending upon the consistency of their diagnoses.

1. Load and display the data set `carcinoma` (library `poLCA`).

```
library(poLCA)
data(carcinoma)
?carcinoma
head(carcinoma) carcinoma
```

2. Use the function `poLCA` to fit the models with 2, 3 e 4 latent classes to the data.

```
?poLCA
formula<-cbind(A,B,C,D,E,F,G)~ 1
m.2<-poLCA(formula,carcinoma,nclass=2,nrep=10,verbose=FALSE)
m.3<-poLCA(formula,carcinoma,nclass=3,nrep=10,verbose=FALSE)
m.4<-poLCA(formula,carcinoma,nclass=4,nrep=10,verbose=FALSE)
```

3. Compare the estimates of the expected frequencies for the *response pattern*obtained with the three models.

```
freq.estim<-data.frame(m.2$predcell[1:9],m.3$predcell[9],m.4$predcell[9])
freq.estim
```

4. Compare the three models considering the following measures: log-likelihood, number of free parameters, likelihood ratio test, Pearson chi-square test (and associated $p-value$), AIC, BIC.

```
K<-c("2", "3", "4")
llik<-c(m.2$llik,m.3$llik,m.4$llik)
npar<-c(m.2$npar,m.3$npar,m.4$npar)
Gsq<-round(c(m.2$Gsq,m.3$Gsq,m.4$Gsq),3)
Chisq<-round(c(m.2$Chisq,m.3$Chisq,m.4$Chisq),3)
df<-c(m.2$resid.df,m.3$resid.df,m.4$resid.df)
p-value<-round(1-pchisq(Chisq,df),4)
AIC<-round(c(m.2$aic,m.3$aic,m.4$aic),3)
```

BIC<-round(c(m.2$bic,m.3$bic,m.4$bic),3)
summary<-data.frame(K,llik,npar,Gsq,Chisq,df,p-value,AIC,BIC)
summary

5. Which model shows the best fit to the observed data?

6. Display the parameter estimates of the three-class model and compare this solution with those obtained with your colleagues. Which problem arises?

round(m.3$P,4)
lapply(m.3$probs,round,4)

7. Use the function `poLCA.reorder` to solve the indeterminacy of the ranking of the latent classes.

probs.start.m3<-m.3$probs.start
new.probs.start.m3<-poLCA.reorder(probs.start.m3,order(m.3$P))
m.3.ord<-poLCA(formula,carcinoma,nclass=3,probs.start=new.probs.start.m3
,verbose=FALSE)
round(m.3.ord$P,4)

8. Interpret the three latent classes.

9. Compute the estimate of the probability of the agreement of the pathologists in the negative diagnosis.

```
p.dnc1<-m.3$P[1]*m.3$probs$A[1,1]*m.3$probs$B[1,1]*m.3$probs$C[1,1]*m.3
$probs$D[1,1]*m.3$probs$E[1,1]*m.3$probs$F[1,1]*m.3$probs$G[1,1]
p.dnc2<-m.3$P[2]*m.3$probs$A[2,1]*m.3$probs$B[2,1]*m.3$probs$C[2,1]*m.3
$probs$D[2,1]*m.3$probs$E[2,1]*m.3$probs$F[2,1]*m.3$probs$G[2,1]
p.dnc3<-m.3$P[3]*m.3$probs$A[3,1]*m.3$probs$B[3,1]*m.3$probs$C[3,1]*m.3
$probs$D[3,1]*m.3$probs$E[3,1]*m.3$probs$F[3,1]*m.3$probs$G[3,1]
p.dn<-p.dnc1+p.dnc2+p.dnc3
round(p.dn,4)
```

10. Compute the estimate of the number of tissues for which the pathologists agree in the negative diagnosis.

round(p.dn*m.3$N,4)

11. Display the posterior probability estimates of the *response pattern* (1,1,1,1,1,1,1) for the three latent classes. In which class the sample of tissue correspondent to this *response pattern* is allocated?

post.1<-m.3.ord$posterior[1,]
round(post.1,5)

12. Display the latent class in which the samples of tissue are allocated according to highest posterior probability.

m.3.ord$predclass
table(m.3.ord$predclass)

13. Select the samples of tissues allocated to the class of inconsistent diagnosis and evaluate the correspondent *response pattern*.

sel<-carcinoma[m.3.ord$predclass==1,]
sel

Exercise 2
Women's mobility data

The data are from the Bangladesh Fertility Survey of 1989. The questionnaire contains a number of items that indicate whether a woman living in rural Bangladesh could engage in various activities alone. Following the steps of the analysis of the previous example perform a latent class analysis in order to identify groups of women with similar patterns of mobility

1. Load the data set `Mobility` (library ltm).

   ?Mobility

2. Recode the data set properly in order to use the function `poLCA`.

   names(Mobility)<-c("Item1","Item2","Item3","Item4",
   "Item5","Item6","Item7","Item8")
   Mobility[Mobility==1]<-2
   Mobility[Mobility==0]<-1

3. Use the function `poLCA` to fit the latent class models to the data. Estimate models with $K = 2, 3, 4, 5$ classes and select the model that represents the best compromise between goodness of fit and parsimonious.

3. Comment obtained the results .