

Task 1

Introduction

In the present part of the report, we will investigate to what extent we will be able to classify respondents in their country, and then we will compare the performance of different classifiers.

Data

The data have been obtained from the 6th Wave of the World Value Survey, which was carried out between 2010 and 2013. The data include the standardized scores of 3929 respondents of 3 countries on 32 variables, that have been summarized with 7 factors obtained using exploratory factor analysis with oblique rotation. The 7 factors related to the 32 variables are:

1. **Rights**, that it's related to homosexuality, prostitution, abortion, divorce, sex before marriage, suicide;
2. **Steal**, that it's related to claiming benefits, avoiding fare, stealing property, cheating taxes, accept a bribe;
3. **Crime**, that it's related to robberies, alcohol, police-military, racist behavior, drug sale;
4. **Religion**, that it's related to attend religious services, pray, the importance of God;
5. **Realize self**, that it's related to creative, rich, spoil oneself, be successful, exciting life;
6. **Do good**, that it's related to security, do good, behave properly, protect environment, tradition;
7. **Violence**, that it's related to beat wife, parents beating children, violenc.

Methodology

To investigate the possibility to classify the respondents in their country based on the 7 factors we have used the canonical discriminant analysis. We have applied the linear regression function with 7 predictors and 1 dependent variable, the Country. Then to the output, we have applied the Canonical Discriminant Analysis.

```
lm.out<-lm(cbind(F_rights, F_steal, F_crime,F_religion,F_realizeself,F_dogood,
                 F_violence)~as.factor(country), data=dwvs)
candisc.out<-candisc(lm.out)
print(candisc.out)
```

Canonical Discriminant Analysis for as.factor(country):

	CanRsq	Eigenvalue	Difference	Percent	Cumulative
1	0.80691	4.17882	3.5622	87.142	87.142
2	0.38142	0.61661	3.5622	12.858	100.000

Test of H0: The canonical correlations in the current row and all that follow are zero

```

      LR test stat approx F numDF denDF   Pr(> F)
1      0.11944  1059.53    14  7834 < 2.2e-16 ***
2      0.61858   402.65     6  3918 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

As we can see both the Square Canonical Correlation are significant, but the discriminating power to separate between the groups is higher for the first than for the second discriminant function: 0.81 and 0.38 respectively. The LR test indicates that the discriminant analysis is meaningful. The first test's null hypothesis is $H_0 : \lambda_1 = \lambda_2 = 0$ and this hypothesis as we can see from the *p-value* is rejected. The *null* hypothesis of the first test it's equivalent to the test for $H_0 : \mu_{Netherlands} = \mu_{Nigeria} = \mu_{Philippines}$.

The second LR test indicates that $H_0 : \lambda_2 = 0$, and also this null hypothesis is rejected. So even if the second discriminant function has less discriminant power cannot be omitted and it's statistically meaningful.

On our analysis, we have also applied two different tests for centroids and to test the equal covariance.

To see if the three-country has different centroids and confirm the results of the canonical discriminant analysis we have applied on the linear regression the function *Manova*:

```

res_t1_2 <- summary(Manova(lm.out), test="Wilks")

summary.default(Manova(lm.out), test="Wilks")

```

	Length	Class	Mode
SSP	1	-none-	list
SSPE	49	-none-	numeric
df	1	-none-	numeric
error.df	1	-none-	numeric
terms	1	-none-	character
repeated	1	-none-	logical
type	1	-none-	character
test	1	-none-	character

The *p-value* is small, and the test confirms that the analysis is meaningful and that at least there is a pair of centroids that differs significantly. The function *Manova* in *r* doing the *Wilks Lambda test* uses the Rao approximation. To test the assumption on equal population covariance we have applied to the linear regression the function *boxM*:

```
boxM(lm.out)
```

Box's M-test for Homogeneity of Covariance Matrices

```

data: Y
Chi-Sq (approx.) = 5479.2, df = 56, p-value < 2.2e-16

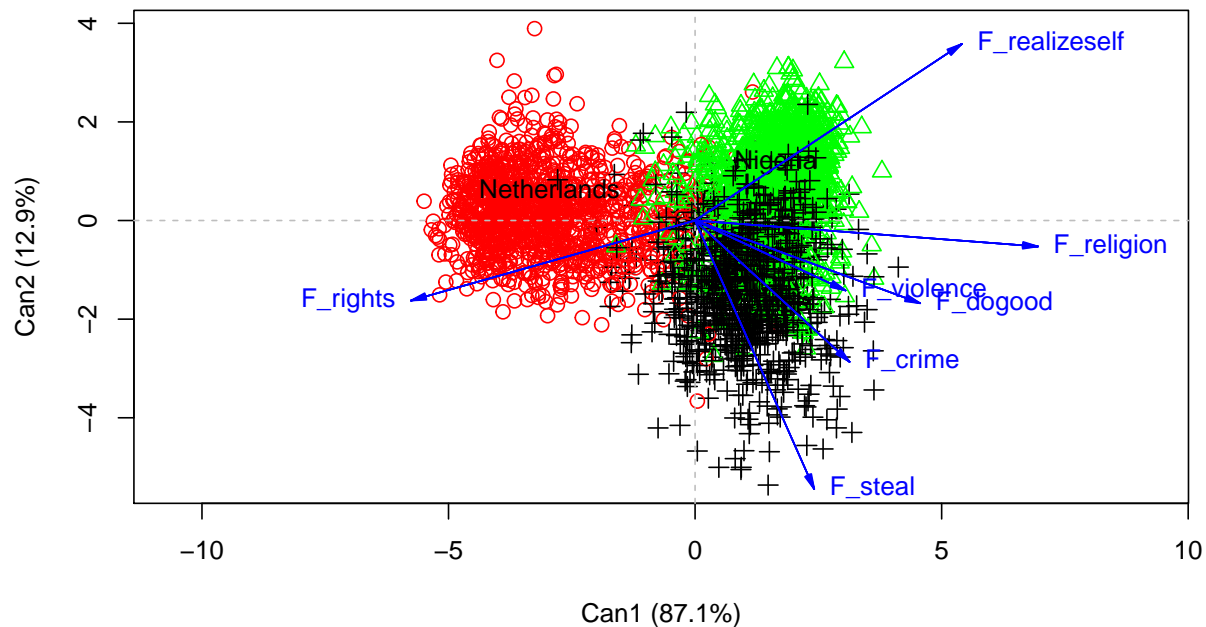
```

The test of Box indicates that H_0 of equal covariance matrices across groups is not supported by data.

Plot

To complete the Canonical Discriminant analysis, we have plotted the three countries and the 7 variables.

Vector scale factor set to 7.827



We can see that the group of individuals in *red* are Netherlands citizens, the group of individuals in *green* are Nigerians citizens and the group of individuals in *black* are Philippines citizens. In blue we can see the 7 explanatory variables. The plot shows a clear separation between Netherlands and the other two countries on the first discriminant function while the second discriminant function could help to separate Nigeria and the Philippines. The first discriminant function especially correlates with the factors: rights, religion, realize self, and do good; whereas the second discriminant function correlates with the factors: steal and realize self. The two-factor crime and violence have a lower correlation on the two factors and so has in this analysis lower importance on separate the 3 countries.

Compare the performance of different classifiers

We are going now to compare the performance of different classifiers to classify respondents in their country based on the 7 factors. To be able to do that we are going to compute the training error and the leave-one-out cross-validation error. The classification method that we are going to compare are: - Linear discriminant analysis; - Quadratic discriminant analysis; - K-nearest neighbors with k ranging from 1 to 100; - High Dimensional Discriminant Analysis;

Linear discriminant analysis

This method aims to separate in the clearest possible way different groups using the linear combination of observed independent variables. The linear discriminant analysis method assumes that the covariance

structure of the independent variable is the same across groups. In our analysis, we know from the Box test previously computed that this assumption is not supported by the data. It will be an interesting test if in this case the Quadratic discriminant analysis, where the assumption on the equality of covariance matrix is relaxed, will perform better. In the linear discriminant analysis, we have applied the method of Fisher correcting for the different prior probability.

```
lda.out1<-lda(country ~ F_rights + F_steal + F_crime + F_religion + F_realizeself +
              F_dogood + F_violence, data=dwvs)
#print(lda.out1)
pred.train1 <- predict(lda.out1,dwvs, prior=c(1,1,1)/3)
```

```
tab1 <- table(dwvs$country,pred.train1$class)
#print(tab1)
kbl(tab1)
```

	Netherlands	Nigeria	Philippines
Netherlands	1145	39	76
Nigeria	10	1327	241
Philippines	17	220	851

```
#training hit rate
kbl(sum(diag(tab1))/sum(tab1))
```

x
0.8464086

```
#classify test observations using LDA
pred.loocv2<-lda(country~F_rights+F_steal+F_crime+F_religion+F_realizeself+F_dogood+
                 F_violence,data=dwvs, prior=c(1,1,1)/3, CV=TRUE)
tab2<-table(dwvs$country,pred.loocv2$class)
print(tab2)
```

	Netherlands	Nigeria	Philippines
Netherlands	1145	39	76
Nigeria	11	1326	241
Philippines	17	224	847

```
#LOOCV hit rate
kbl(sum(diag(tab2))/sum(tab2))
```

x
0.845135

We can see that in that case, the difference between the performance for training error and LOOCV error is really small, so there is no evidence for overfitting.

Quadratic discriminant analysis

The second method that we have applied is Quadratic discriminant analysis. It should perform better considering the difference in the covariance matrix for the different groups. QDA even if has a lower bias

with a different covariance matrix, has a larger variance, and as in our case with a small dataset can be problematic.

Even in this case, we have applied the method of Fisher correcting for prior probabilities.

```
qda.out3<-qda(country ~ F_rights + F_steal + F_crime + F_religion + F_realizeself +
              F_dogood + F_violence, data = dwvs)

pred.train3<-predict(qda.out3,dwvs, prior=c(1,1,1)/3)

tab3<-table(dwvs$country,pred.train3$class)
#print(tab3)
kbl(tab3)
```

	Netherlands	Nigeria	Philippines
Netherlands	1210	21	29
Nigeria	40	1319	219
Philippines	41	219	828

```
#training hit rate
kbl(sum(diag(tab3))/sum(tab3))
```

x
0.8550688

```
#classify test observations using QDA
pred.test4 <- qda(country ~ F_rights + F_steal + F_crime + F_religion + F_realizeself +
                  F_dogood + F_violence, data=dwvs,prior=c(1,1,1)/3,CV=TRUE)
tab4<-table(dwvs$country,pred.test4$class)
#print(tab4)
kbl(tab4)
```

	Netherlands	Nigeria	Philippines
Netherlands	1210	21	29
Nigeria	40	1315	223
Philippines	43	223	822

```
#LOOCV hit rate
kbl(sum(diag(tab4))/sum(tab4))
```

x
0.8525217

In that case there is also no evidence of overfitting. We can see from the results that the QDA performs better than the LDA but not with a significant improvement.

K-nearest Neighbors

The third model that we had analyzed is the K-nearest Neighbors. We have computed the model using all the 3926 observations, and to choose which is the correct number k of parameters to use, we have compared the training error with the Leave one out cross-validation error.

```
#str(dwvs)
table(dwvs$country)
```

```
Netherlands    Nigeria Philippines
      1260         1578         1088
```

```
set.seed(9850) # -> random number generator
gp<-runif(nrow(dwvs))
dwvs2<-dwvs[order(gp),]
#str(dwvs)
#str(dwvs2)
#head(dwvs)
#head(dwvs2)

hitratknn<-function(observed,predicted){
  tab<-table(observed,predicted)
  hitratknn<-sum(diag(tab))/sum(tab)
  return(hitratknn)
}

knnmax<-100
err<-matrix(rep(0,knnmax*2), nrow=knnmax)

for(j in 1:knnmax) {
  predknn.train<-knn(dwvs2[,2:8], dwvs2[,2:8], dwvs2$country, k=j)
  err[j,1]<-hitratknn(dwvs2$country,predknn.train)
}

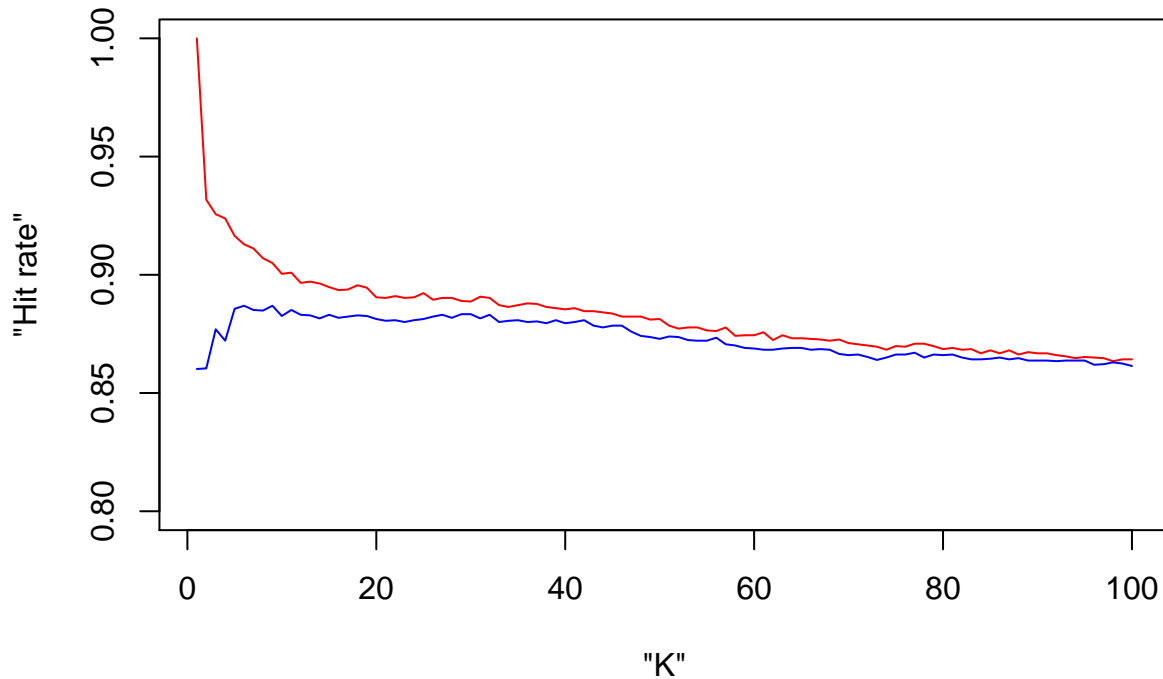
for(j in 1:knnmax) {
  predknn.train<-knn.cv(dwvs2[,2:8], dwvs2$country, k=j)
  err [j,2]<-hitratknn(dwvs2$country,predknn.train)
}

plot('K', 'Hit rate',xlim=c(1,knnmax),ylim=c(0.8,1))
```

Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion

Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion

```
lines(c(1:knnmax),err[,1],col="red") # -> training error
lines(c(1:knnmax),err[,2],col="blue")
```



We can see that with $K=1$ the model is flexible and by definition, we have training hit rate (red line) of 1, but the LOOCV hit rate (blue line) is higher in this case, while with model less flexible, as with $k=98$, the two errors are similar. Since both the errors increase if we increase the parameter K , probably the model that describes the dataset better is the model with $K=30$ or $K=66$.

Table 1:

K	TRAINING HIT RATE	LOOCV HIT RATE
1	1	0.8601630
30	0.8886908	0.8833418
66	0.8728986	0.8683138
100	0.8642384	0.8614366

High Dimensional Discriminant Analysis

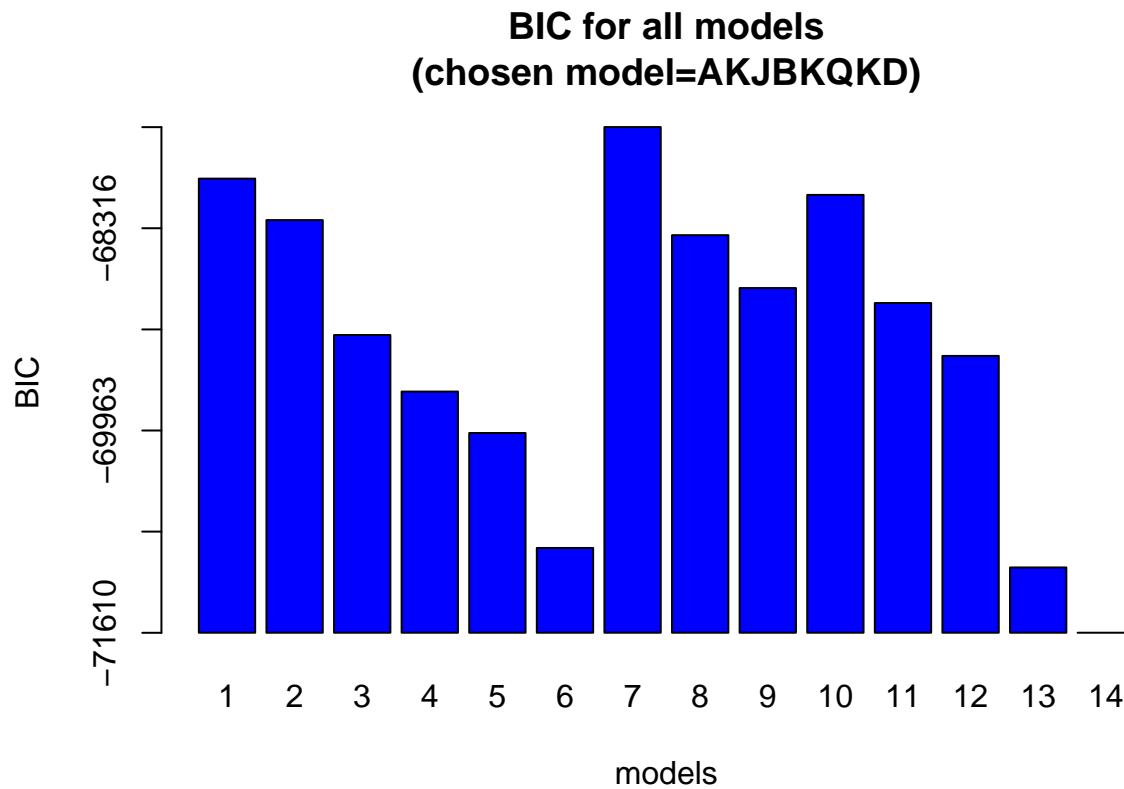
The fourth method that we have used to discriminate between different groups is the HDDA method. This method could be useful while the number of parameters is high compared to the number of data.

```
w <- dwvs[, -1]
cls <- dwvs[, 1]
#HDDA on the learning dataset:
hdda.out7 <- hdda(w, cls, scaling=TRUE, model="all", d="BIC", graph=TRUE, show=TRUE)
```

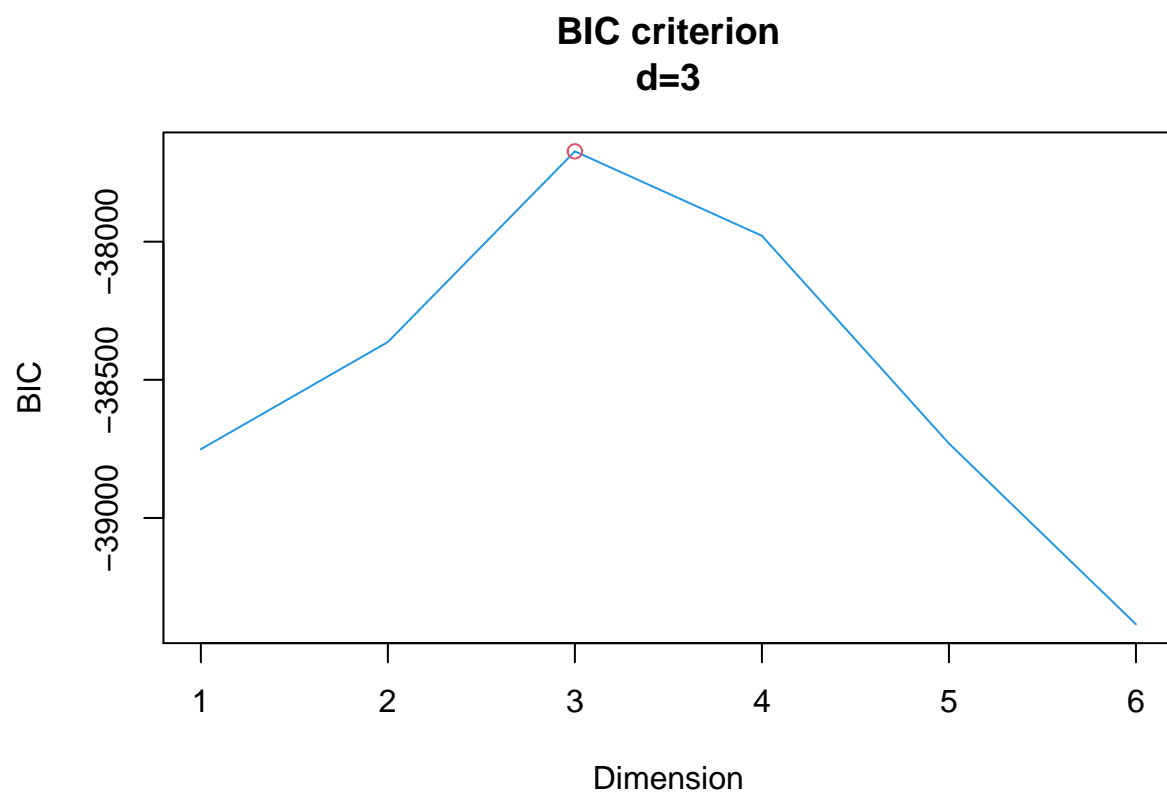
```
# :      Model      BIC
```

1 :	AKJBKQDK	-67912.32
2 :	AKBKQDK	-68249.76
3 :	ABKQDK	-69185.72
4 :	AKJBQDK	-69646.01
5 :	AKBQDK	-69983.46
6 :	ABQDK	-70919.41
7 :	AKJBKQKD	-67492.38
8 :	AKBKQKD	-68372.73
9 :	ABKQKD	-68803.13
10 :	AKJBQKD	-68044.89
11 :	AKBQKD	-68925.24
12 :	ABQKD	-69355.63
13 :	AJBQD	-71078.02
14 :	ABQD	-71609.34

SELECTED: Model AKJBKQKD, BIC=-67492.38.

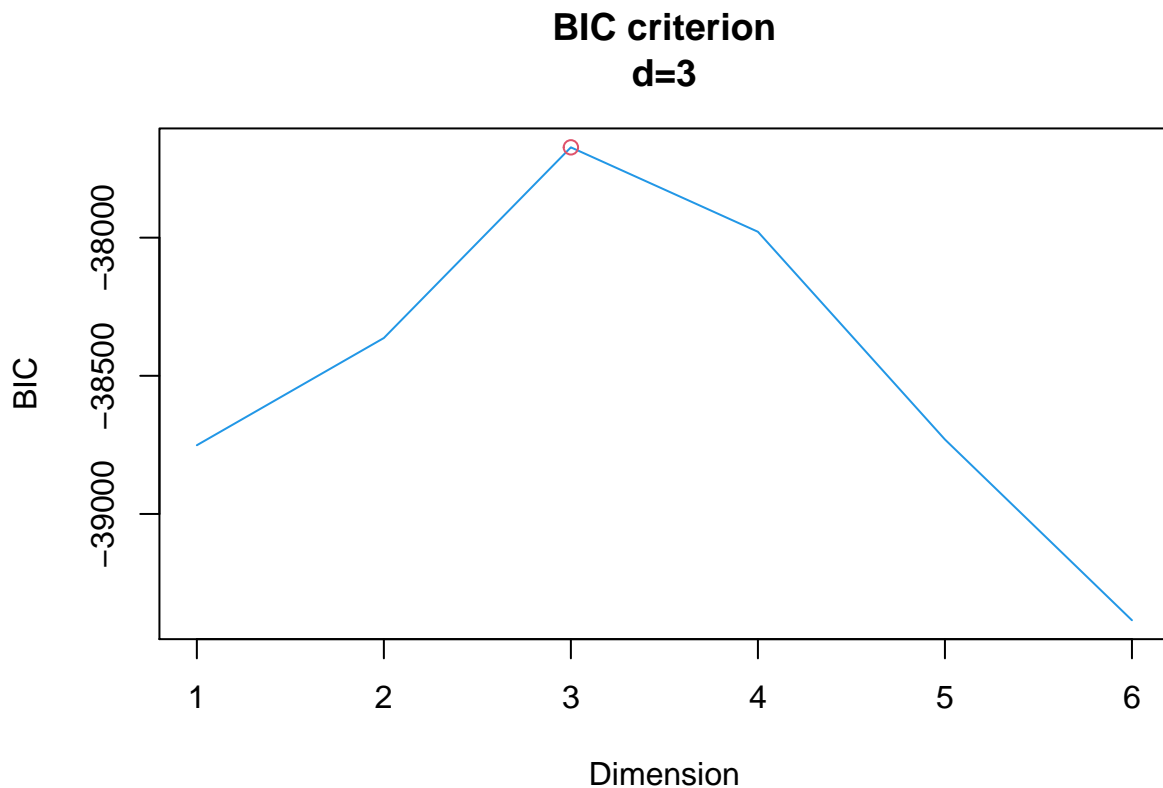


```
plot(hdda.out7)
```

The model used from the HDDA analysis applying the BIC criterion is model 7($A_{kj}K_jB_kQ_kd$). The decision following the *BIC* criteria is to choose the model with the lowest value, in that case, the reference system is negative, so the lowest value is for model 7.

```
plot(hdda.out7,method="BIC")
```



The dimension choose for the model is 3. The model chosen by the *BIC* criteria has the same number of principal components for all the three different classes.

```
pred.train7<-predict(hdda.out7,w,cls)
```

Correct classification rate: 0.8530311.

	Initial class		
Predicted class	Netherlands	Nigeria	Philippines
Netherlands	1196	20	40
Nigeria	26	1356	251
Philippines	38	202	797

```
#print(tab7)
tab7<-table(dwvs$country,pred.train7$class)
#training hit rate
kbl(sum(diag(tab7))/sum(tab7))
```

x
0.8530311

```
pred.loocv8 <- hdda(w, cls,scaling=TRUE, d="BIC", L00=TRUE)
tab8<-table(cls,pred.loocv8$class)
#print(tab8)
kbl(tab8)
```

	Netherlands	Nigeria	Philippines
Netherlands	1197	26	37
Nigeria	22	1378	178
Philippines	44	285	759

#LOOCV hit rate

```
kbl(sum(diag(tab8))/sum(tab8))
```

x
0.8492104

Also with the *HDDA* model, there is rather small evidence of overfitting, but the *HDDA* model does not perform better than the other models.

Error comparison for the different models

We have computed for all 4 models the hit rate, for the comparison in the table we will present the training and LOOCV errors computing $1 - \text{hit rate}$:

Table 2:		
MODEL	TRAINING ERROR	LOOCV ERROR
LDA	0.1535914	0.154865
QDA	0.1449312	0.1474783
KNN (K=30)	0.1113092	0.1166582
HDDA	0.1469689	0.1507896

In all the present models there is little evidence of overfitting, the two errors computed are in all the cases similar. The K-nearest Neighbors is a good model to compare the others and we can see that even if it performs better it has not a huge difference. The model that performs better between the other 3 is the Quadratic discriminant analysis, it is the most complex one with the highest number of parameters used.

Confronting the results, we can say that even if there is a difference between the models, no one of the computed ones has outstanding results.

Table of *QDA LOOCV* rate:

Table 3:			
-	Netherlands	Nigeria	Philippines
Netherlands	1210	21	29
Nigeria	40	1315	223
Philippines	43	223	822

As we were expecting in the analysis of the canonical discriminant analysis, the model has a high ability to differentiate between Netherlands and the two other countries, while it has a high error rate discriminating between Nigeria and the Philippines.

Multinomial logistic regression model

```
m1<- multinom(country~F_rights+F_steal+F_crime+F_religion+F_realizeself+F_dogood +
               F_violence, family=multinomial, data=dwvs, maxit=3926, hess=TRUE)
```

```
# weights: 27 (16 variable)
initial value 4313.151845
iter 10 value 1376.724330
iter 20 value 1352.584896
final value 1346.617148
converged
```

```
t1_15_result <- summary (m1)
t1_15_result
```

Call:

```
multinom(formula = country ~ F_rights + F_steal + F_crime + F_religion +
          F_realizeself + F_dogood + F_violence, data = dwvs, family = multinomial,
          maxit = 3926, hess = TRUE)
```

Coefficients:

	(Intercept)	F_rights	F_steal	F_crime	F_religion	F_realizeself
Nigeria	0.6472240	-2.122889	0.5537755	0.5976742	2.887712	2.8922393
Philippines	0.9959826	-1.466848	1.5556448	1.0662700	1.932117	0.9680894
	F_dogood	F_violence				
Nigeria	-0.01110756	1.2844737				
Philippines	1.16765772	0.7560728				

Std. Errors:

	(Intercept)	F_rights	F_steal	F_crime	F_religion	F_realizeself
Nigeria	0.1612100	0.1642741	0.1735127	0.1334749	0.2143403	0.1695936
Philippines	0.1510987	0.1475679	0.1685630	0.1296852	0.1866007	0.1556415
	F_dogood	F_violence				
Nigeria	0.1218099	0.1534901				
Philippines	0.1197516	0.1470007				

Residual Deviance: 2693.234

AIC: 2725.234

We can see that there are 2 different regression model estimates: the first one compares the probability of Nigeria to the probability of the Netherlands, the second model compares the probability of the Philippines to the probability of the Netherlands. All the parameters are significant except F_dogood for Nigeria, which's not significantly different from 0. The sign of the parameters is the same for all the parameters in both the regressions, except for F_dogood where the Nigeria coefficient is not significant. This could be explained because we have analyzed previously Netherlands strongly differs from the other two countries, while Nigeria and the Philippines have not a clear separation. This analysis shows that sample data is more likely to belong to Nigeria and the Philippines than to the Netherlands when it has a higher positive value in the parameters of steal, crime, violence, religion, and a lower negative value in rights. These coefficients could be probably well explained from the fact that the Netherlands is one of the most developed countries in all the world, the statistics of the Human Development Index published by the United Nations Development Programme places it in the 8th place in the world, while Nigeria and the Philippines are both considered developing states (161 and 107 respectively in the HDI ranking).

```
#### compute hitrate training data####
train.pred<-predict(m1,newdata=dwvs)
tab<-table(dwvs$country, train.pred)
kbl(sum(diag(tab))/sum(tab))
```

	x
0.8571065	

Error rate: 0.1428935

```
#####compute LOOCV
nobs<- 3926
hit<-rep(0,nobs)
for (i in 1:nobs){
  train<-c(1:nobs)
  mod<- multinom(country ~ F_rights + F_steal + F_crime + F_religion +
                  F_realizeself + F_dogood + F_violence, data=dwvs,
                  subset=train[-i], print= FALSE,maxit=3926)
  pred<- predict(mod, newdata=dwvs[i,])
  hit[i]<-ifelse(pred==dwvs$country[i],1,0)
}
```

```
#hitrate
mean(hit) ##### LOOCV
```

Error rate: 0.1444218

The Multinomial logistic regression model has slightly better error values than the other models, except for KNN.