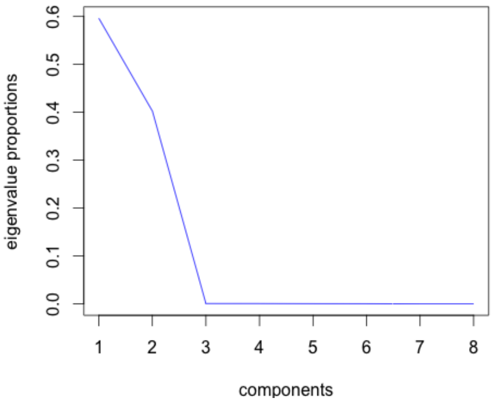
MVA deliverable 2 by Diego Garcia-Olano

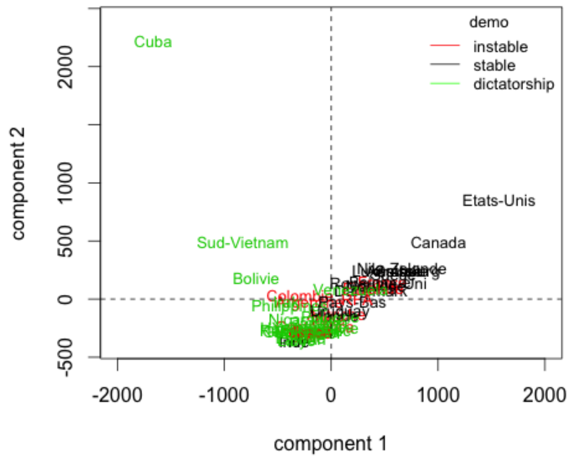
1 PCA on centered data

```
#1. Read again the Russet data set and impute the missing values.
#  Define as X matrix, the one defined by the continuous variables.  (Now, just using matrix operation)
Russet <- read.csv("/Users/diego/Documents/UPC-MIRI/semester2/MultiVariate-Analysis/
      assignment_2/Russet_ineqdata.csv", header = TRUE, quote = "\"", dec = ".", check.names=TRUE)
Russet$demo <- as.factor(Russet$demo)
levels(Russet$demo) <- c("stable","instable","dictatorship")
X.init <- X
X <- Russet[,2:9]  #only take continous variables ( and no id)

summary(Russet)
library(mice)     #impute values using mice for "Rent" variable which has 3 NA's  and for ecks as well
X = as.matrix(complete(mice(X,m=1,seed=8675309)))
Xc.init <- X

N = diag(rep(1/47,47))                   #2.  Define the matrix of weights of individuals.
G = (t(X) %*% N) %*% rep(1,47)           #3.  Compute the centroid of individuals.

ones = matrix(rep(1,nrow(X)*ncol(X)),nrow=nrow(X))
Xc = X - ( ones %*% diag(t(G)[1,]))     #4.  Compute the centered X matrix.

V = t(Xc) %*% N   %*% Xc                #5.  Compute the covariance matrix of X
eigX <- eigen(V)                        #    and diagonalize it.

#6. Do the screeplot of the eigenvalues and define the number of significant dimensions.
#  How much is the retained information?  (with the chosen dimensions, how much inertia is captured )
lambdas = eigX$values
total_inertia = sum(lambdas)
taus = lambdas/sum(eigX$values)       #calculate importance of each component
par(mfrow=c(1,1))
plot(c(1:8),taus,type="l",col="blue",xlab="components",ylab="eigenvalue proportions")
```

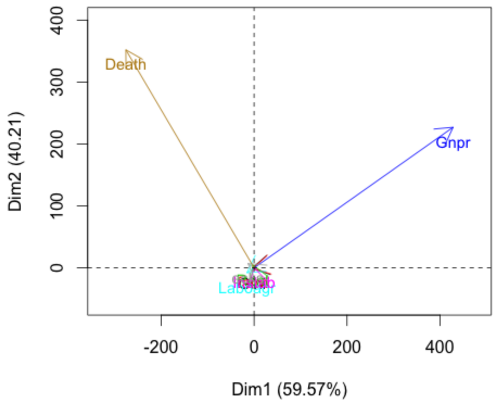


```
#Using the Last Elbow rule from the screeplot, we see that only the first two dimensions are significant.
#The first two components alone explain for 0.5956 + 0.4020 = 0.9977 % of the information in the data!

#7. Compute the projections of individuals in the significant dimensions.
```

```
#  The projections of the individuals on the 1st and 2nd dimension are given by
#  their coordinates for the first two principal components in Psi.
Psi <- Xc %*% eigX$vectors

#8. Plot the individuals in the first factorial plane of R^p and color them according to the "demo" variable.
plot(Psi[,1],Psi[,2],type="n",xlab="component 1",ylab="component 2",xlim=c(-2000,2000),ylim=c(-400,2400))
text(Psi[,1],Psi[,2],labels=as.character(Russet$pais),col=as.integer(Russet$demo),cex=.8)
legend('topright', as.character(unique(Russet$demo)), title="demo",lty=1,
      col=c('red', 'black','green'  ), bty='n', cex=.8)
abline(h=0, v=0, col="black",lty=2)
```



```
#9. Compute the projection of variables in the significant dimensions, use transformation equation.
Pvar = t(matrix(rep(sqrt(eigX$values),8),ncol=8)) * eigX$vectors       #Pvar = sqrt(lambdas) * u

#10. Plot the variables (as arrows) in the first factorial plane of Rn.
par(mfrow=c(1,1))
plot(Pvar[,1],Pvar[,2],type="n",,xlim=c(min(Pvar[,1]) - 50,max(Pvar[,1]) + 50 ),
      ylim=c(min(Pvar[,2]) - 50 ,max(Pvar[,2]) + 50), xlab="Dim1 (59.57%)",ylab="Dim2 (40.21)")
for(i in 1:8){ arrows(0,0,Pvar[i,1],Pvar[i,2],col=i)
    text(Pvar[i,1],Pvar[i,2],labels=colnames(X.init)[i],col=as.integer(i),cex=.9,pos=1)    }
abline(h=0,v=0,col="black",lty=2)
```



```
#11. Compute the correlation of the variables with the significant components and interpret them.

# From our calculation in 9, we see the correlation of each variable (along the rows) to each component.
row.names(Pvar) = colnames(X.init)
colnames(Pvar) = c("Dim1","Dim2","Dim3","Dim4","Dim5","Dim6","Dim7","Dim8")
```

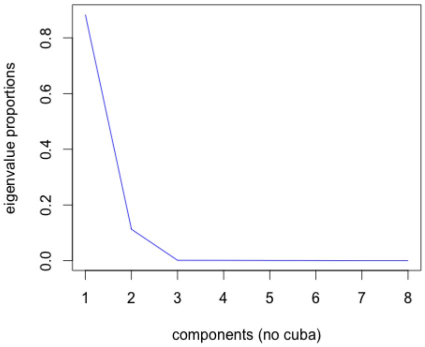
	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
Gini	-4.48	-0.66	-6.49	-2.60	11.49	0.55	-0.14	0.78
farm	-2.42	-0.38	-3.11	-1.26	4.81	-0.01	-0.86	-1.72
Rent	-2.57	4.62	-7.44	-13.91	-4.35	-3.43	0.20	0.02
Gnpr	428.16	226.89	-0.32	0.31	0.10	-0.35	-0.01	0.00
Laboagr	-15.37	-7.92	-2.01	6.18	0.86	-11.13	0.03	0.03
Instab	-0.10	0.30	-0.75	0.41	1.64	0.32	3.89	-0.35
ecks	-13.30	6.94	-14.84	7.51	-4.05	2.98	-0.06	0.02
Death	-276.06	351.97	0.54	-0.04	0.12	-0.04	-0.00	-0.00

We determined Dimension1 and Dimension2 to be the significant components, and from the table we see that the plane composed of Dim1 and Dim2 is dominated by the variables Gnpr, which has a high positive correlation with both Dim1 and Dim2, 428.15 and 226.89 respectively, and Death, which has a high negative correlation with Dim 1 (-276.07) and a high positive correlation with Dim 2 (351.96).

These two variables dominate the analysis because we are not using standardized data and their scale is greater than the others.

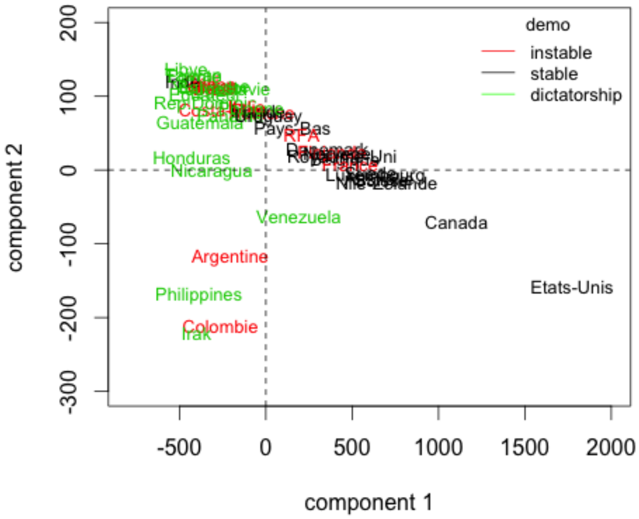
2 PCA on centered data with a weight of 0 for Cuba

```
Nc = diag(rep(1/46,47))                       #12.2.  Define the matrix of weights of individuals.
which(Russet$pais=="Cuba")  #11
Nc[11,11] = 0                                 # Give Cuba weight of zero
Gc = (t(X.init) %*% Nc) %*% rep(1,47)         #12.3.  Compute the centroid of individuals.

ones = matrix(rep(1,nrow(X.init)*ncol(X.init)),nrow=nrow(X.init))
Xcc = X.init - ( ones %*% diag(t(Gc)[1,]))    #12.4.  Compute the centered X matrix.
Xcc <- as.matrix(Xcc)
Vc = t(Xcc) %*% Nc %*% Xcc                    #12.5.  Compute the covariance matrix of X
eigXc <- eigen(Vc)                            # and diagonalize it.
lambdasc = eigXc$values
total_inertiac = sum(lambdasc)
tausc = lambdasc/sum(eigXc$values)            #calculate importance of each component;
par(mfrow=c(1,1))                             #12.6.  Do the screeplot of the eigenvalues
plot(c(1:8),tausc,type="l",col="blue",xlab="components (no cuba)",ylab="eigenvalue proportions")
```

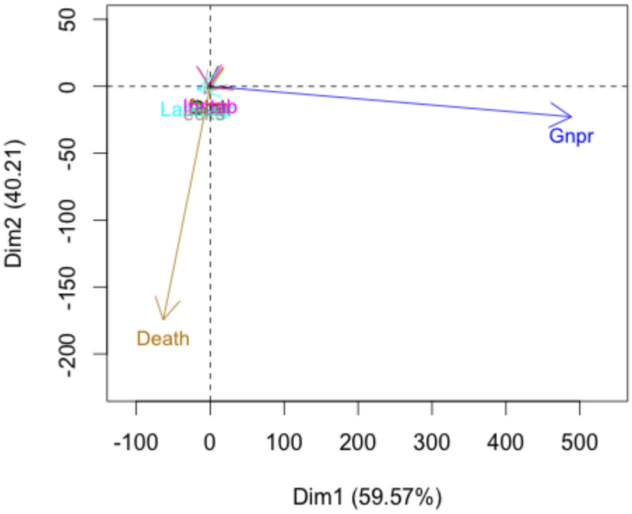


```
#  Using the Last Elbow rule, we see again that only the first two dimensions are significant.
#  This time the first two components explain for 0.8835 + 0.1128 = 0.9963 % of the information in the data!

#12.7. Compute the projections of individuals in the significant dimensions.
#  The projections of the individuals on the first and 2nd dimension are
#  given by their coordinates for the first two principal components in Psi.
Psic <- Xcc %*% eigXc$vectors

#12.8. Plot the individuals in the first factorial plane of and color them according to "demo"
```

```
plot(Psic[,1],Psic[,2],type="n",xlab="component 1",ylab="component 2",xlim=c(-800,2000),ylim=c(-300,200))
text(Psic[,1],Psic[,2],labels=as.character(Russet$pais),col=as.integer(Russet$demo),cex=.8)
legend('topright',as.character(unique(Russet$demo)),title="demo",col=c('red','black','green'),cex=.8)
abline(h=0, v=0, col="black",lty=2)
```



```
#12.9. Compute the projection of variables in the significant dimensions
Pvarc = t(matrix(rep(sqrt(eigXc$values),8),ncol=8)) * eigXc$vectors

#12.10. Plot the variables (as arrows) in the first factorial plane of Rn.
par(mfrow=c(1,1))
plot(Pvarc[,1],Pvarc[,2],type="n",,xlim=c(min(Pvarc[,1]) - 50,max(Pvarc[,1]) + 50 ),
   ylim=c(min(Pvarc[,2]) - 50 ,max(Pvarc[,2]) + 50), xlab="Dim1 (59.57%)",ylab="Dim2 (40.21)")
for(i in 1:8){ arrows(0,0,Pvarc[i,1],Pvarc[i,2],col=i)
  text(Pvarc[i,1],Pvarc[i,2],labels=colnames(X.init)[i],col=as.integer(i),cex=.9,pos=1)}
abline(h=0,v=0,col="black",lty=2)
```



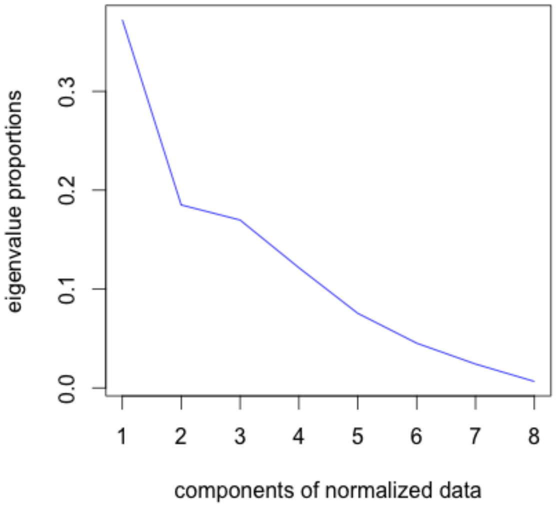
```
#12.11. Compute the correlation of the variables with the significant components and interpret them.
# From our calculation in 9, we see the correlation of each variable (along the rows) to each component.
row.names(Pvarc) = colnames(X.init)
colnames(Pvarc) = c("Dim1","Dim2","Dim3","Dim4","Dim5","Dim6","Dim7","Dim8")
```

	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
Gini	-4.33	-1.56	-6.61	0.40	11.81	0.52	-0.15	0.79
farm	-2.33	-0.56	-3.26	0.13	4.91	0.49	-0.85	-1.72
Rent	-0.18	-2.57	-13.01	-10.97	-4.24	2.06	0.19	0.03
Gnpr	488.37	-22.86	-0.17	0.35	0.04	0.36	-0.01	0.00
Laboagr	-17.63	-2.28	0.14	5.14	-0.89	11.51	0.03	0.04
Instab	0.04	-0.39	-0.48	0.70	1.65	-0.09	3.94	-0.34
ecks	-8.20	-5.79	-10.73	13.03	-3.72	-2.83	-0.05	0.02
Death	-63.44	-174.49	0.64	-0.39	0.07	-0.14	-0.01	-0.00

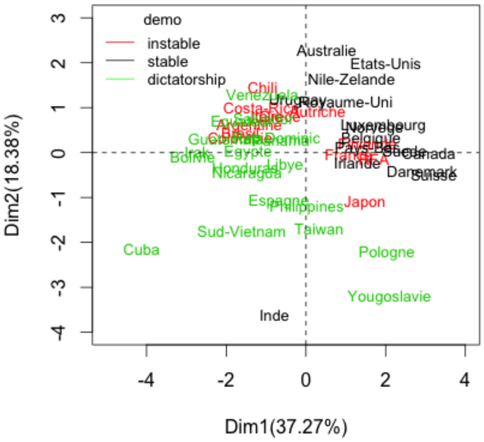
We can see that the first dimension is dominated again by it is high positive correlation with Gnpr and to a lesser extent is negative correlation with Death. Thus countries with higher Gnprs will be farther to the ri and to a lesser extent, those with higher death will be pushed slightly to the left.
The second dimension is then highly negatively correlated with Death, so countries with a high number for De will be pushed to the bottom near the 2nd axis.

3 PCA on standardized data

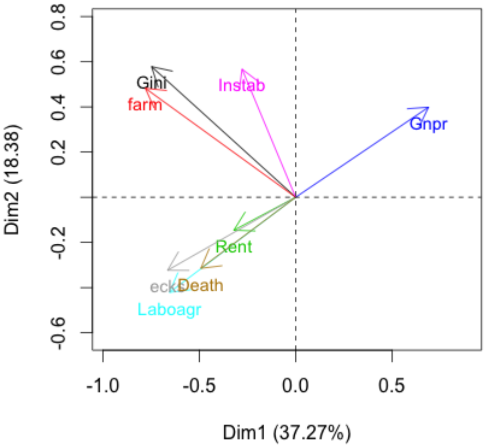
```
13. Redo 2:11, but taking as metric of Rp the inverse of the variances of the variables.
  Nt = diag(rep(1/47,47))                      #13.2.  Define the metric matrix of weights of individuals
  Gt = (t(Xc.init) %*% Nt) %*% rep(1,47)        #13.3.  Compute the centroid of individuals.

  #13.4.  Compute the normalized X matrix according to new metric.  ( centered / sigma^2)
  ones = matrix(rep(1,nrow(X.init)*ncol(X.init)),nrow=nrow(X.init))
  Xt = Xc.init - ( ones %*% diag(t(Gt)[1,]))           #center matrix
  Xtsquared = Xt^2
  variances2 = t(Xtsquared) %*% matrix(rep(1, 47),nrow=47) / 46
  Xtc = Xt * ones %*% diag(1/sqrt(variances))         #divide each cell by standard deviation of column to nor

  Vtc = t(Xtc) %*% Nt %*% Xtc      #13.5.  Compute the correlation matrix of X
  eigXtc <- eigen(Vtc)             # and diagonalize it.

  #13.6. Do the screeplot of the eigenvalues and define the number of significant dimensions.
  lambdasct = eigXtc$values
  total_inertiact = sum(lambdasct)
  tausct = lambdasct/sum(eigXtc$values)    #calculate importance of each component;
  par(mfrow=c(1,1))                       #screeplot
  plot(c(1:8),tausct,type="l",col="blue",xlab="components of normalized data",ylab="eigenvalue proportions")
```



```
# Using the Last Elbow rule from the screeplot, we see that only the first dimension would be significant
# with 37.27% information explained, but the 2nd dimension and 3rd dimension also contain sizeable eigenvalu
# which explain 18.38 and 16.36 percent of the variance respectively.
# This time the first two components explain for .3727 + 0.1838 = 0.5565 % of the information in the data!

#13.7. Compute the projections of individuals in the significant dimensions.
Psit <- Xtc %*% eigXtc$vectors

#13.8. Plot the individuals in the first factorial plane of Rp and color them according to "demo" variable.
plot(Psit[,1],Psit[,2],type="n",xlab="Dim1(37.27%)",ylab="Dim2(18.38%)",xlim=c(-5,4),ylim=c(-4,3))
text(Psit[,1],Psit[,2],labels=as.character(Russet$pais),col=as.integer(Russet$demo),cex=.8)
legend('topleft', as.character(unique(Russet$demo)),title="demo",
    lty=1,col=c('red','black','green'), cex=.8,bty="n")
abline(h=0, v=0, col="black",lty=2)
```



```
#13.9. Compute the projection of variables in the significant dimensions
Pvarct = t(matrix(rep(sqrt(eigXtc$values),8),ncol=8)) * eigXtc$vectors

#13.10. Plot the variables (as arrows) in the first factorial plane of Rn. #plot6
par(mfrow=c(1,1))
plot(Pvarct[,1],Pvarct[,2],type="n",,xlim=c(min(Pvarct[,1]) - .2,max(Pvarct[,1]) + .2 ),
   ylim=c(min(Pvarct[,2]) - .2 ,max(Pvarct[,2]) + .2), xlab="Dim1 (37.27%)",ylab="Dim2 (18.38)")
for(i in 1:8){ arrows(0,0,Pvarct[i,1],Pvarct[i,2],col=i)
  text(Pvarct[i,1],Pvarct[i,2],labels=colnames(X.init)[i],col=as.integer(i),cex=.9,pos=1)}
abline(h=0,v=0,col="black",lty=2)
```



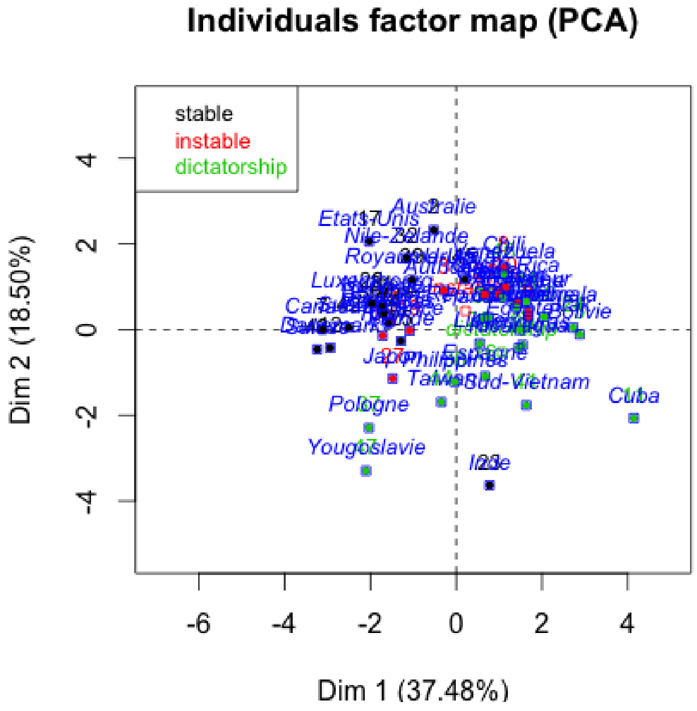
```
#13.11. Compute the correlation of the variables with the significant principal components and interpret the
# From our calculation in 9, we see the correlation of each variable (along the rows) to each component.
row.names(Pvarct) = colnames(X.init)
colnames(Pvarct) = c("Dim1","Dim2","Dim3","Dim4","Dim5","Dim6","Dim7","Dim8")
```

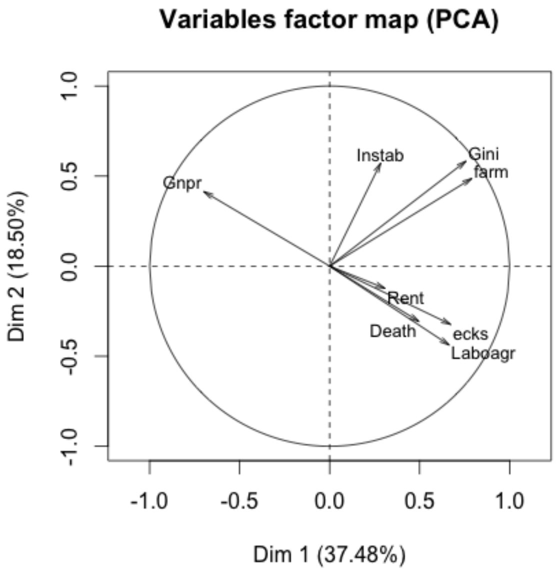
	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
Gini	-0.75	0.58	-0.05	0.15	0.18	0.03	-0.00	0.16
farm	-0.78	0.48	-0.04	0.24	0.23	-0.00	-0.03	-0.16
Rent	-0.32	-0.15	0.59	0.61	-0.43	-0.05	-0.01	0.01
Gnpr	0.69	0.40	0.46	-0.07	0.15	-0.11	-0.31	0.00
Laboagr	-0.65	-0.42	-0.50	-0.02	-0.17	0.03	-0.30	0.01
Instab	-0.28	0.57	0.06	-0.53	-0.54	0.02	0.01	-0.02
ecks	-0.66	-0.32	0.32	-0.34	0.14	-0.44	0.04	0.01
Death	-0.49	-0.31	0.60	-0.30	0.18	0.38	-0.03	-0.00

With the normalized dataset, the analysis is more nuananced than before.
The variables most negatively correlated with the 1st dimension are Gini (-.75) and farm (-.78)
while Gnpr (.69) is the only variable positively correlated with the first dimension.
Thus the first dimension opposes countries with high Gnpr against those with high Gini and farm values.
The second dimension is most positively correlated with variables Instab (.57), Gini (.58), and to
a lesser extent farm (.48). These are not particularly strong but still represent an opposition of
these variables against the others along that axis.

4 normalized PCA with FactoMiner

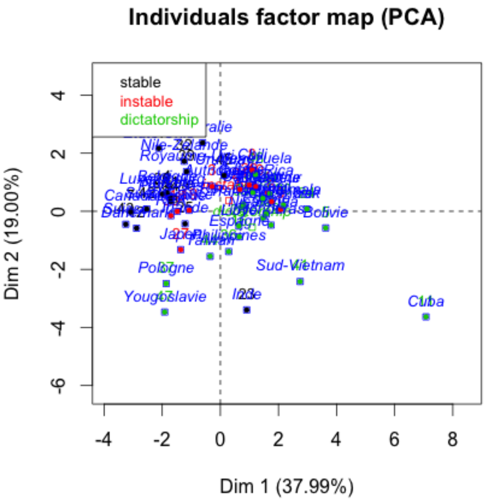
```
#14. Do again the PCA (normalized) with FactoMiner using the "demo" variable as illustrative.
library(FactoMineR)
ds = data.frame(Russet[,1],Xc.init,Russet[,10])
names(ds) = names(Russet)
res.pca <- PCA(ds, quali.sup=c(10,1))
plot.PCA(res.pca,choix="ind",habillage=10,cex=0.8,col.quali="blue") #individuals with demo as qualitative supp
```

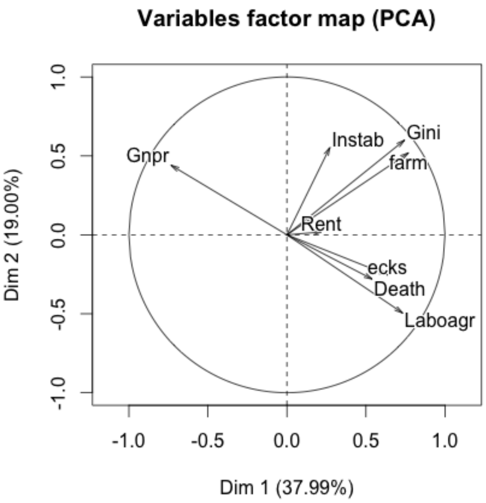




```
plot.PCA(res.pca,choix="var",habillage=10,cex=0.8)  #variable plot


#15. What is the country best represented in the first factorial plane?. And what is the worse?.
ds[which(res.pca$ind$coord[,1] == max(res.pca$ind$coord[,1])),1]     #Cuba
# Cuba is the country best represented in the first factorial plane (ie, that fartherest to the right)
ds[which(res.pca$ind$coord[,1] == min(res.pca$ind$coord[,1])),1]     #Suisse
# Suisse is the country that is worst represented. (ie, farthest to the left)

#16. What are the three countries most influencing the formation of the first principal component?,
#    and what are the three countries most influencing the formation of the second principal component?

# We first look at the contributions made for component one as follows:
head(sort(res.pca$ind$contrib[,1],decreasing=TRUE))
    11        43         7
    12.184449  7.465825  6.874929
# and see that Cuba, Suisse and Canada are the three countries most influencing the formation of it.
# We do similarly for the second component
head(sort(res.pca$ind$contrib[,2],decreasing=TRUE))
    23        47         2
    18.974169 15.609055  7.781159
# and see that Inde, Yougoslavie, and Australie contribute most to the formation of the 2nd component.

#17. What is the variable best represented in the first factorial plane?. And what is the worse?.
#By looking at the correlation between the original data and the Psi values, located in
res.pca$var$cor , we see that the variable best represented in the first factorial plane is Gini
( 0.7572015 + 0.5828405) while the worst variable is Rent (0.3074779 -0.1231920).  These correspond
to the longest and shortest arrows on the variable factor map.

#18. What are the three variables most influencing the formation of the first principal component?,
#    and what are the three variables most influencing the formation of the  second principal component?
#For the first component, we check to see which variables most contribute to the formation of it:
sort(res.pca$var$contrib[,1], decreasing=TRUE)
     farm      Gini      Gnpr      ecks   Laboagr     Death      Rent    Instab
20.782159 19.123980 16.282134 15.085488 14.697366  8.202290  3.153430  2.673153
#Thus the farm, Gini, and Gnpr are the top three variables influencing the formation of component 1.

#For the second component, we see:
sort(res.pca$var$contrib[,2], decreasing=TRUE)
     Gini    Instab      farm   Laboagr      Gnpr      ecks     Death      Rent
```

```
22.949886 22.082214 15.997116 12.946468 11.564069  7.065538  6.369420  1.025288
#Thus Gini, Instab and farm are the three variables which contribute the most to forming component 2.

#19. Which modalities of the variable "demo" are significant in the selected significant principal components?
# The "dictatorship" regime level (in green on individual plot) is most significant on the 1st dimension
# where as the "stable" and "instable" regime levels are significant along the 2nd dimension.
# This is found by checking the v.test value for our categorical supplementary variable, and seeing which
# variables have positive values for which components.
res.pca$quali.sup$v.test
                   Dim.1        Dim.2       Dim.3        Dim.4       Dim.5
stable       -4.26125663  1.355591445  2.54853969 -0.793271626  0.57543198
instable      0.49695784  1.418258720 -0.97748859  1.637896627 -0.87013330
dictatorship  3.57927566 -2.528850493 -1.54072236 -0.696585388  0.22486297

# 20. Now, study the sensibility of the performed PCA respect to considering Cuba as an  outlier.
# Redo the PCA function with weight for Cuba equal to 0.0001, and
# compute the correlations of the obtained significant components with the previous obtained ones.
Cuba is index 11, so
w = (1 - .0000001) / 46   # 0.02173696
Nc = rep(0.02173913,47)                         # initial weights accounting that cuba will be given .0001
Nc[11] = .0000001                                # Give Cuba weight .0000001, sum(Nc) == 1
res.pca.nocuba = PCA(ds, quali.sup=c(10,1), row.w = Nc)
plot.PCA(res.pca.nocuba,choix="ind",habillage=10,cex=0.8,col.quali="blue") #individuals plot
plot.PCA(res.pca.nocuba,choix="var") #var plot
```





We note Cuba has been pushed farther from the origin in the individuals plot and that ecks is less represented than before (ie, its arrow is shorter) and that Laboagr representation has improved slightly.

We can now compare the contribution of Cuba to the prior PCA and this current one to see how Cuba now plays almost no role whatsoever in the formation of the principal components.

```
round(res.pca$ind$contrib[11,],7)
    Dim.1     Dim.2     Dim.3     Dim.4     Dim.5
12.184449  6.128553 42.603325  5.298076  3.753134

round(res.pca.nocuba$ind$contrib[11,],7)
    Dim.1     Dim.2     Dim.3     Dim.4     Dim.5
0.0001680 0.0000888 0.0000769 0.0012265 0.0004395
```

Additionally, we can look at the eigen values for both the prior PCA and this current one.

Eigen values for data set including Cuba outlier

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.00	37.48	37.48
comp 2	1.48	18.50	55.98
comp 3	1.30	16.25	72.23
comp 4	0.95	11.83	84.06
comp 5	0.67	8.35	92.41
comp 6	0.36	4.52	96.94
comp 7	0.19	2.42	99.35
comp 8	0.05	0.65	100.00

Eigen values for data set without Cuba outlier

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.04	38.02	38.02
comp 2	1.52	19.01	57.02
comp 3	1.05	13.07	70.10
comp 4	0.91	11.43	81.53
comp 5	0.66	8.26	89.78
comp 6	0.58	7.22	97.00
comp 7	0.19	2.36	99.36
comp 8	0.05	0.64	100.00

We can also see how the variables correlation with the dimensions has changed.
For the first dimension, we see that without Cuba, Laboagr is more correlated with the component while ecks has lowered slightly.

```
> sort(res.pca.nocuba$var$cor[,1],decreasing=TRUE)
      farm       Gini    Laboagr       ecks      Death     Instab       Rent       Gnpr 
 0.7685349  0.7437941  0.7335405  0.6364685  0.5426388  0.2716081  0.2169563 -0.7341199 

> sort(res.pca$var$cor[,1],decreasing=TRUE)
      farm       Gini       ecks    Laboagr      Death       Rent     Instab       Gnpr 
 0.7893465  0.7572015  0.6725151  0.6638074  0.4958952  0.3074779  0.2830963 -0.6986795 
```

The second dimension correlations have remained relatively the same with Gini and Instab being still the variables with the highest correlations to the component.

We can also check how the contributions for the components have changed from the view of variables:

```
> sort(res.pca.nocuba$var$contrib[,1],decreasing=TRUE)
     farm      Gini      Gnpr   Laboagr      ecks     Death    Instab      Rent 
19.420970 18.190688 17.720570 17.692610 13.319794  9.682007  2.425656  1.547705 
> sort(res.pca$var$contrib[,1],decreasing=TRUE)
     farm      Gini      Gnpr      ecks   Laboagr     Death      Rent    Instab 
20.782159 19.123980 16.282134 15.085488 14.697366  8.202290  3.153430  2.673153 
```

We see again that the importance of Labagr is brought forth and ecks subsides when discarding Cuba.

From the point of view of individuals, we can see how the contributions to the formations of the dimensions has changed as such:

```
> head(sort(res.pca.nocuba$ind$contrib[,1],decreasing=TRUE))
       5       43        7       24       12       41 
9.472526 7.575994 6.808197 6.362239 5.949644 5.487840 
> head(sort(res.pca$ind$contrib[,1],decreasing=TRUE))
       11        43         7        12         5        24 
12.184449  7.465825  6.874929  6.133401  5.885996  5.274324 
```

For the first dimension, Cuba (11) no longer plays any role and Bolivie (5) becomes the most important contributor to the construction of the component.

```
> head(sort(res.pca.nocuba$ind$contrib[,2],decreasing=TRUE))
       47        23        37        41         2        17 
17.227393 16.435248  8.861950  8.447161  7.954311  6.696429 
```

```
> head(sort(res.pca$ind$contrib[,2],decreasing=TRUE))
       23        47         2        37        11        17
18.974169 15.609055  7.781159  7.547597  6.128553  6.076871
```

For the second dimension, Yougoslavie (47) becomes the most important contributor.

Thus overall, we see that removing Cuba as an outlier has a slight albeit little effect on the overall amount of information expressed by the first component plane, but its exclusion does change greatly the representation quality of the variables Laboagr and ecks.