Sesión 6: Análisis de Componentes Principales

Alejandra Lelo de Larrea Ibarra

10 de febrero de 2019.

Índice

1. Introducción

Se tienen datos con los tipos de cambio reales (respecto a USD) de varias economias (periodo 1970-2010). El objetivo es implementar el procedimiento inferencial PCA considerando distribuciones iniciales no informativas para (μ, Λ) y contestar lo siguiente:

- ¿Qué economía tiene el mayor peso esperado en la descomposicion PCA?
- ¿Qué economía tiene la mayor consistencia en estimacion de los cjs correspondientes?

2. Datos

```
# Se cargan los paquetes
library("fields")
library("mnormt")
library("MCMCpack")
library("actuar")
library("ggplot2")
library("kernlab")
library("tidyverse")
library("readr")
library("psych")
library("mvtnorm")
library("MASS")
library("xlsx")
library("knitr")
```

```
# Función para extraer modas
getmode <- function(v) {
   uniqv <- unique(v)
```

```
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Leemos los datos correspondientes a los tipos de cambio de distintas economías. Se tienen 492 observaciones mensuales para 80 economías.

```
# Cargamos los datos)
data<-read.xlsx("../01_Notas_Ovando/est46114_s06_data.xls",sheetName = 'RealXR_Data')

# Obtenemos las dimensiones de los datos
dim(data)
```

```
## [1] 492  81
```

```
# Extraemos las fechas
fechas<-data$Date

data<-select(data,-Date)
```

Vemos que países están en la muestra:

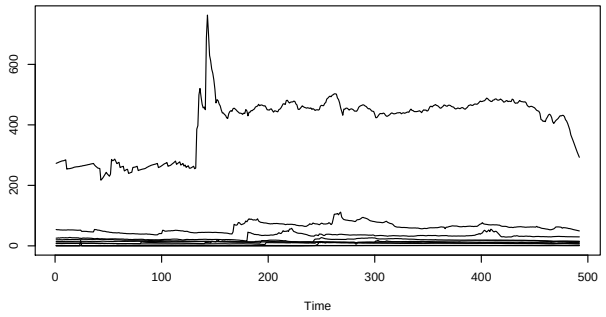
```
# Vemos la lista de países
colnames(data)
```

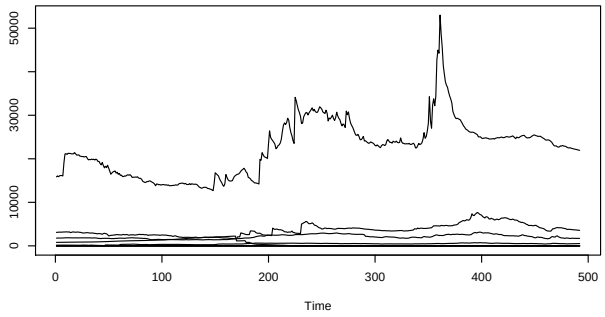
```
##  [1] "Canada"        "Mexico"        "Guatemala"     "El.Salvador"
##  [5] "Honduras"      "Nicaragua"     "Costa.Rica"    "Panama"
##  [9] "Jamaica"       "Dominican.Rep" "Trin.Tobago"   "Colombia"
## [13] "Venezuela."    "Ecuador"       "Peru"          "Chile"
## [17] "Brazil."       "Paraguay"      "Uruguay"       "Argentina"
## [21] "EU12"          "Sweden"        "Norway"        "Finland"
## [25] "Denmark"       "U.K."          "Ireland"       "Luxembourg"
## [29] "Netherlands"   "France"        "Germany"       "Austria"
## [33] "Czech.Rep"     "Hungary"       "Switzerland"   "Poland"
## [37] "Russia"        "Spain"         "Portugal"      "Italy"
## [41] "Greece"        "Turkey"        "Syria"         "Israel"
## [45] "Jordan"        "Kuwait"        "Saudi.Arabia"  "India"
## [49] "Pakistan"      "Bangladesh"    "Sri.Lanka."    "Thailand"
## [53] "Malaysia"      "Singapore"     "Indonesia"     "Philippines"
## [57] "China.PR"      "Korea"         "Hong.Kong"     "Taiwan"
## [61] "Japan"         "Australia"     "New.Zealand"   "Morocco"
## [65] "Algeria"       "Tunisia"       "Egypt"         "Cameroon"
## [69] "Senegal"       "Sierra.Leone"  "Cote.d.Ivoire" "Ghana"
## [73] "Nigeria"       "Benin"         "Congo"         "Kenya"
## [77] "Tanzania"      "Mozambique"    "South.Africa"  "Zambia"
```

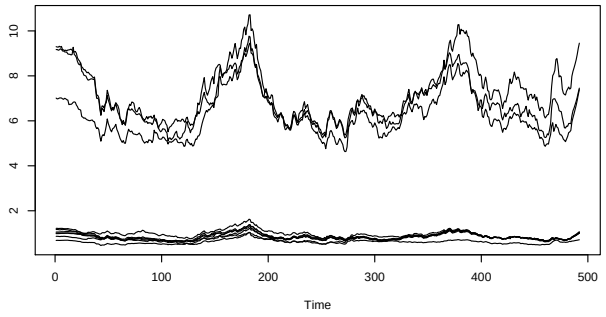
Graficamos las series de tiempo de los países para tener una idea de qué esté pasando.

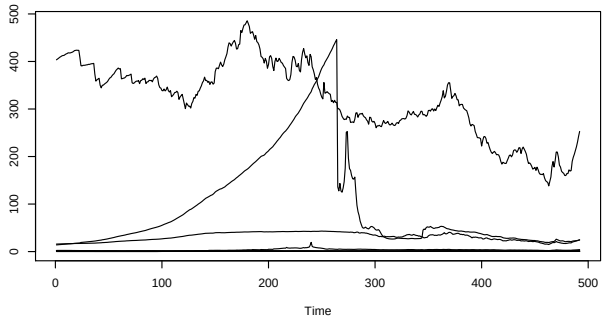
```
# SE parte el plot en 8 pedazos
par(mfrow=c(4,2))

# Se grafican series de tiempo de los tipos de cambio
ts.plot(as.ts(data[,1:10],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,11:20],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,21:30],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,31:40],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,41:50],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,51:60],start=c(1970,1),end=c(2010,12),frequency=12))
```

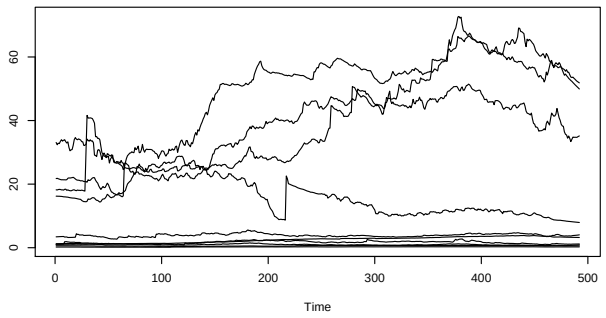
```
ts.plot(as.ts(data[,61:70],start=c(1970,1),end=c(2010,12),frequency=12))
ts.plot(as.ts(data[,71:80],start=c(1970,1),end=c(2010,12),frequency=12))
```

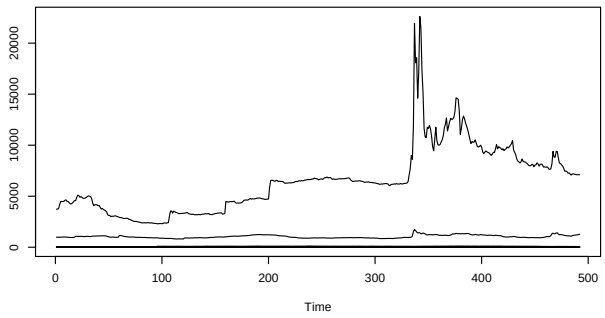


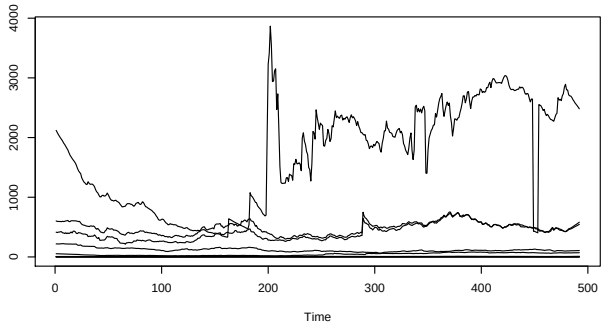


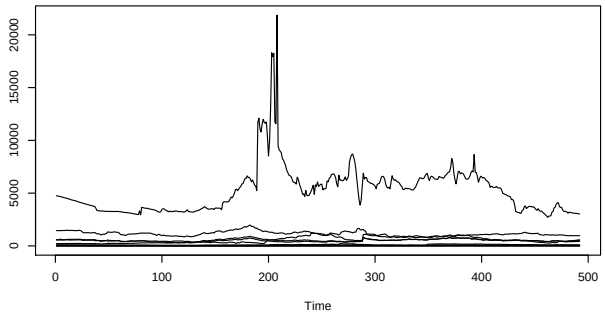












3. Inferencia de SPCA

Pensemos que para cada observación mensual se tiene que

$$\boldsymbol{X}_{j\cdot} \sim N_p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}),$$

con $j = 1, \ldots, 80$ donde μ y Λ son desconocidos.

3.1. Actualizacion bayesiana de parámetros

El desconocimiento acerca de $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ se expresa como

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{m}_0, s_0, a_0, \boldsymbol{B}_0),$$

donde $\boldsymbol{m}_0 = 0, s_0 = 1/2, a_0 = 1, \boldsymbol{B}_0 = I$; es decir, se tiene una distribución inicial no informativa.

De esta manera, la distribución posterior de (μ, Λ) a partir de la consolidación de la información contenida en los datos y la información complementaria está dada por

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}) \sim \text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n),$$

con

$$\begin{aligned}
\boldsymbol{m}_n &= \frac{s_0 m_0 + n\bar{x}}{s_n}, \\
s_n &= s_0 + n, \\
a_n &= a_0 + \frac{n}{2}, \\
\boldsymbol{B}_n &= \boldsymbol{B}_0 + \frac{n}{2}\left[S + \frac{s_0}{s_n}(\bar{x} - m_0)(\bar{x} - m_0)'\right]
\end{aligned}$$

```
# Fijar hiperparámetros
a0 <- 1
s0 <- 1/2
m0 <- matrix(0,ncol=1,nrow=ncol(data))
B0 <- diag(1,ncol=ncol(data),nrow=ncol(data))

# Función para calcular la posterior
gaussian.posterior <- function(data,m0,s0,a0,B0){

  # Media de los datos
  xbar <- as.matrix(colMeans(data))

  # Mat. de var y cov de los datos.
  S <- cov(data)

  # No. de obs.
  n <- nrow(data)

  # No. de variables.
```

```
  p <- ncol(data)

  # Parámetros actualizados de la posterior.
  sn <- s0 + n
  an <- a0 + n/2
  mn <- (s0*m0 + n*xbar)/sn
  Bn <- B0 + (n/2)*(S + (s0/sn)*(xbar-m0)%*%t(xbar-m0))

  # Salida (parámetros actualizados)
  output <- list(mn=mn,sn=sn,an=an,Bn=Bn)
  return(output)
}

# Calculamos los hiperparams actualizados para la posterior
output <- gaussian.posterior(data,m0,s0,a0,B0)
```

3.2. Simulación

Una vez realizada la actualización bayesiana de los hiperparámetros, simulamos M observaciones para (μ, Λ) como

$$\begin{aligned} \boldsymbol{\Lambda}^{(m)} &\sim \text{Wi}(\boldsymbol{\Lambda}|a_n, \boldsymbol{B}_n), \\ \boldsymbol{\mu}^{(m)}|\boldsymbol{\Lambda}^{(m)} &\sim \text{N}(\boldsymbol{\mu}|\boldsymbol{m}_n, s_n\boldsymbol{\Lambda}^{(m)}). \end{aligned}$$

Con estos parámetros, obtenemos simulaciones de los eigenvalores y eigenvectores $(e_j^{(m)}, \boldsymbol{v}_j^{(m)})_{j=1}^p$, de la matriz de varianzas y covarianzas de los datos simulada en el paso anterior (inverso de $\Lambda^{(m)}$.

Por último, obtenemos simulaciones de las componentes principales como $\boldsymbol{c}_{j\cdot}^{(m)} = \boldsymbol{X}\boldsymbol{v}_j^{(m)}$.

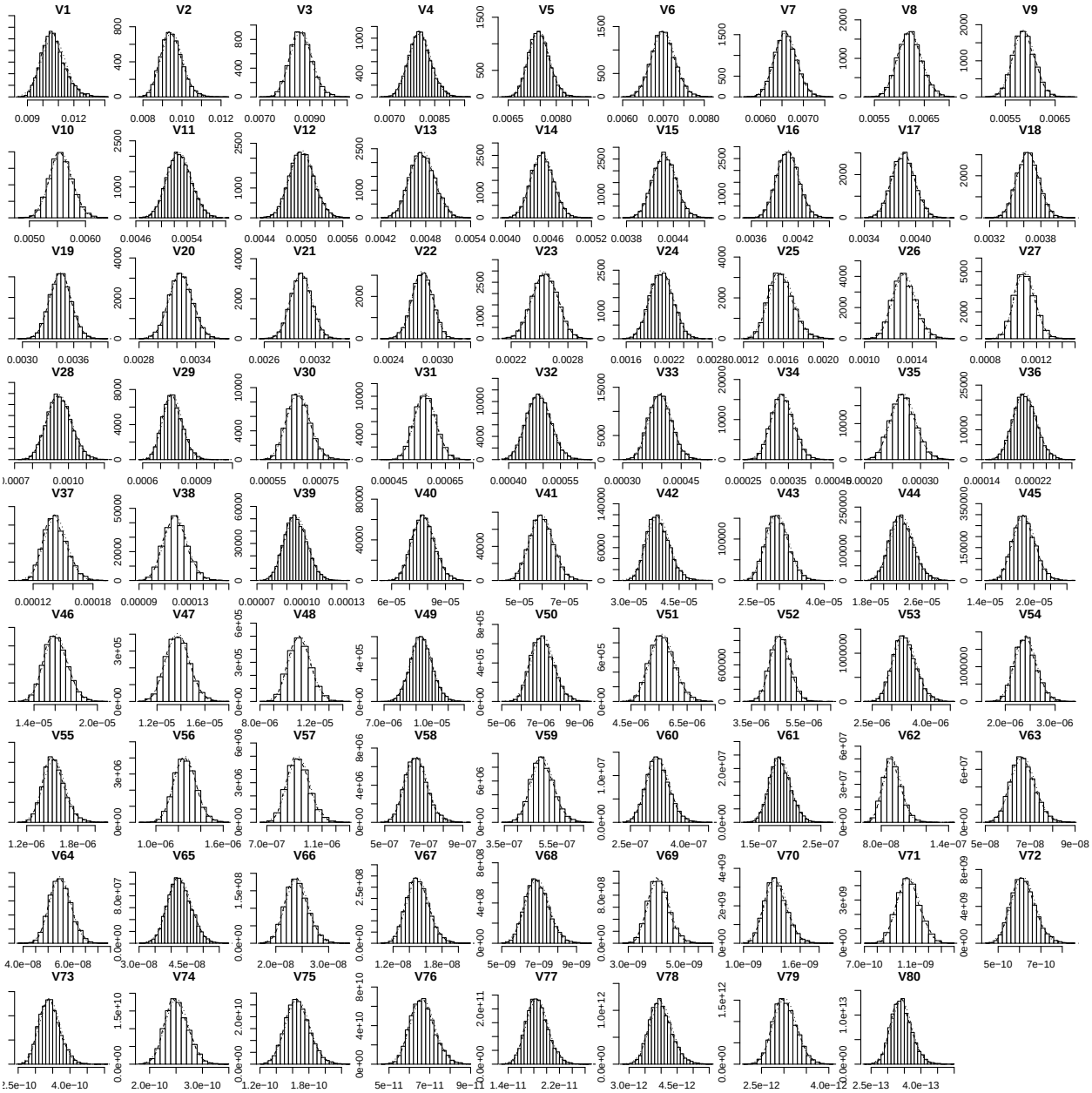
```
# Se fija el no de simulaciones
M <- 10000

# Matriz para guardar medias
mu.sim <- matrix(NA,nrow=M, ncol=ncol(data))

# Arreglo para guardar precisiones
Lambda.sim <- array(NA,dim=c(M,ncol(data),ncol(data)))

# Matriz para guardar valores propios.
e.sim <- matrix(NA,nrow=M, ncol=ncol(data))

# Arreglo para gaurdar vectores propios
V.sim <- array(NA,dim=c(M,ncol(data),ncol(data)))

# Arreglo para guardar componentes principales
C.sim <- array(NA,dim=c(M,nrow(data),ncol(data)))

# Se convierten los datos a matriz.
X <- as.matrix(data)

# En cada iteración:
```

```r
for(m in 1:M){

  # Se simulan valores (mu,Lambda)
  Lambda.sim[m,,] <- rWishart(1, output$an, output$Bn)
  mu.sim[m,] <- mvrnorm(1, mu=output$mn, Sigma=solve(output$sn*Lambda.sim[m,,]), tol = 1e-6)

  # Simulacion de eigenvalores y eigenvectores (e,V)
  eigen_aux <- eigen(solve(Lambda.sim[m,,]))
  e.sim[m,] <- eigen_aux$values
  V.sim[m,,] <- eigen_aux$vectors

  # Simulación de componentes principales.
  C.sim[m,,] <- X %*% V.sim[m,,]
}
```

4. Resultados

4.1. Eigenvalores

```r
multi.hist(e.sim)
```



4.2. ¿Qué economía tiene el mayor peso esperado en la descomposición PCA?

Los pesos de las componentes están representados por los eigenvectores. Utilizamos la moda de la distribución para obtener la matriz de eigenvectores promedio.

```
# Estimador puntual de la matriz de eigenvectores.
V.mode<-matrix(NA,ncol(data),ncol(data))

# Se extrae la moda para cada elemento de la matriz de eigenvectores
for(i in 1:ncol(data)){

  for(j in 1:ncol(data)){
```

```
    V.mode[i,j]<-getmode(V.sim[,i,j])
  }
}

colnames(V.mode)<-paste("PC",1:ncol(data),sep="_")
rownames(V.mode)<-colnames(data)

# Se busca el peso más grande en cada una de las componentes.
peso_max<-data_frame(CP=colnames(V.mode),
                     Economia=apply(V.mode,2,function(x){rownames(V.mode)[which.max(x)]}))

kable(cbind(peso_max[1:20,],peso_max[21:40,],peso_max[41:60,],peso_max[61:80,]),
      type='latex',
      caption='Economia con mayor peso por componente principal')
```

Tabla 1: Economia con mayor peso por componente principal

CP	Economia	CP	Economia	CP	Economia	CP	Economia
PC_1	Greece	PC_21	U.K.	PC_41	South.Africa	PC_61	Cameroon
PC_2	Ireland	PC_22	Portugal	PC_42	Honduras	PC_62	Japan
PC_3	EU12	PC_23	Australia	PC_43	Syria	PC_63	Congo
PC_4	Austria	PC_24	Ghana	PC_44	Thailand	PC_64	Japan
PC_5	Jordan	PC_25	Denmark	PC_45	Syria	PC_65	Cote.d.Ivoire
PC_6	EU12	PC_26	Turkey	PC_46	Mexico	PC_66	Benin
PC_7	Tunisia	PC_27	Denmark	PC_47	India	PC_67	Costa.Rica
PC_8	Portugal	PC_28	New.Zealand	PC_48	Taiwan	PC_68	Chile
PC_9	Tunisia	PC_29	Morocco	PC_49	India	PC_69	Colombia
PC_10	Spain	PC_30	Brazil.	PC_50	Syria	PC_70	Korea
PC_11	Portugal	PC_31	Brazil.	PC_51	Mexico	PC_71	Korea
PC_12	Panama	PC_32	Israel	PC_52	Philippines	PC_72	Benin
PC_13	Finland	PC_33	Trin.Tobago	PC_53	Pakistan	PC_73	Colombia
PC_14	France	PC_34	Venezuela.	PC_54	Pakistan	PC_74	Peru
PC_15	Tunisia	PC_35	Egypt	PC_55	Jamaica	PC_75	Tanzania
PC_16	Austria	PC_36	Egypt	PC_56	Dominican.Rep	PC_76	Paraguay
PC_17	Switzerland	PC_37	Argentina	PC_57	Jamaica	PC_77	Indonesia
PC_18	Singapore	PC_38	Argentina	PC_58	Algeria	PC_78	Zambia
PC_19	Luxembourg	PC_39	Argentina	PC_59	Cameroon	PC_79	Ecuador
PC_20	Australia	PC_40	South.Africa	PC_60	Jamaica	PC_80	Peru

```
# Tabla de frecuencias para pesos por país
frec_peso_max<-table(peso_max$Economia)
frec_peso_max<-data_frame(Economia=names(frec_peso_max),
                          Frecuencia=frec_peso_max)
kable(cbind(frec_peso_max[1:ceiling(nrow(frec_peso_max)/2),],
            frec_peso_max[(ceiling(nrow(frec_peso_max)/2)+1):nrow(frec_peso_max),]),
      type='latex',
      caption='Frecuencias de Economías con Mayor Peso')
```

Tabla 2: Frecuencias de Economas con Mayor Peso

Economia	Frecuencia	Economia	Frecuencia
Algeria	1	Japan	2
Argentina	3	Jordan	1
Australia	2	Korea	2
Austria	2	Luxembourg	1
Benin	2	Mexico	2
Brazil.	2	Morocco	1
Cameroon	2	New.Zealand	1
Chile	1	Pakistan	2
Colombia	2	Panama	1
Congo	1	Paraguay	1
Costa.Rica	1	Peru	2
Cote.d.Ivoire	1	Philippines	1
Denmark	2	Portugal	3
Dominican.Rep	1	Singapore	1
EU12	2	South.Africa	2
Ecuador	1	Spain	1
Egypt	2	Switzerland	1
Finland	1	Syria	3
France	1	Taiwan	1
Ghana	1	Tanzania	1
Greece	1	Thailand	1
Honduras	1	Trin.Tobago	1
India	2	Tunisia	3
Indonesia	1	Turkey	1
Ireland	1	U.K.	1
Israel	1	Venezuela.	1
Jamaica	3	Zambia	1

4.3. ¿Qué economía tiene la mayor consistencia en estimacion de los cjs correspondientes?

```
# Varianza de los eigenvectores.
V.var<-matrix(NA,ncol(data),ncol(data))

# Se calcula la varianza de cada elemento de los eigenvectores
for(i in 1:ncol(data)){

  for(j in 1:ncol(data)){

    V.var[i,j]<-var(C.sim[,i,j],na.rm=TRUE)
  }
}

colnames(V.var)<-paste("PC",1:ncol(data),sep="_")
rownames(V.var)<-colnames(data)

# Se busca el menor varianza en cada una de las componentes.
var_min<-data_frame(CP=colnames(V.var),
```

```
                    EcoVarMin=apply(V.var,2,function(x){rownames(V.var)[which.min(x)]}))

kable(cbind(var_min[1:20,],var_min[21:40,],var_min[41:60,],var_min[61:80,]),
      type='latex',
      caption='Economia con Menor Varianza por Componente Principal')
```

Tabla 3: Economia con Menor Varianza por Componente Principal

CP	EcoVarMin	CP	EcoVarMin	CP	EcoVarMin	CP	EcoVarMin
PC_1	Netherlands	PC_21	Austria	PC_41	France	PC_61	Italy
PC_2	Netherlands	PC_22	Austria	PC_42	Senegal	PC_62	Pakistan
PC_3	Netherlands	PC_23	Austria	PC_43	Cameroon	PC_63	Pakistan
PC_4	Netherlands	PC_24	Netherlands	PC_44	Bangladesh	PC_64	Honduras
PC_5	Netherlands	PC_25	Netherlands	PC_45	Taiwan	PC_65	Guatemala
PC_6	Netherlands	PC_26	Netherlands	PC_46	Sri.Lanka.	PC_66	Kuwait
PC_7	Netherlands	PC_27	Netherlands	PC_47	Sierra.Leone	PC_67	Honduras
PC_8	Netherlands	PC_28	U.K.	PC_48	Guatemala	PC_68	Tunisia
PC_9	Netherlands	PC_29	France	PC_49	El.Salvador	PC_69	Senegal
PC_10	Netherlands	PC_30	Syria	PC_50	Ecuador	PC_70	Nicaragua
PC_11	Netherlands	PC_31	Syria	PC_51	Chile	PC_71	Egypt
PC_12	Netherlands	PC_32	France	PC_52	Peru	PC_72	Syria
PC_13	Netherlands	PC_33	Turkey	PC_53	Australia	PC_73	Syria
PC_14	Netherlands	PC_34	Turkey	PC_54	Luxembourg	PC_74	Syria
PC_15	Netherlands	PC_35	Singapore	PC_55	Pakistan	PC_75	Tunisia
PC_16	Netherlands	PC_36	Saudi.Arabia	PC_56	U.K.	PC_76	Canada
PC_17	Netherlands	PC_37	China.PR	PC_57	Malaysia	PC_77	Thailand
PC_18	Austria	PC_38	Korea	PC_58	Switzerland	PC_78	Greece
PC_19	Austria	PC_39	China.PR	PC_59	South.Africa	PC_79	Costa.Rica
PC_20	Austria	PC_40	Luxembourg	PC_60	Malaysia	PC_80	Zambia

```
# Tabla de frecuencias para pesos por país
frec_var_min<-table(var_min$EcoVarMin)
frec_var_min<-data_frame(Economia=names(frec_var_min),
                         Frecuencia=frec_var_min)
kable(cbind(frec_var_min[1:ceiling(nrow(frec_var_min)/2),],
            frec_var_min[(ceiling(nrow(frec_var_min)/2)+1):nrow(frec_var_min),]),
      type='latex',
      caption='Frecuencia de Economías con menor varianza')
```

Tabla 4: Frecuencia de Economas con menor varianza

Economia	Frecuencia	Economia	Frecuencia
Australia	1	Malaysia	2
Austria	6	Netherlands	21
Bangladesh	1	Nicaragua	1
Cameroon	1	Pakistan	3
Canada	1	Peru	1
Chile	1	Saudi.Arabia	1
China.PR	2	Senegal	2
Costa.Rica	1	Sierra.Leone	1
Ecuador	1	Singapore	1
Egypt	1	South.Africa	1

Economia	Frecuencia	Economia	Frecuencia
El.Salvador	1	Sri.Lanka.	1
France	3	Switzerland	1
Greece	1	Syria	5
Guatemala	2	Taiwan	1
Honduras	2	Thailand	1
Italy	1	Tunisia	2
Korea	1	Turkey	2
Kuwait	1	U.K.	2
Luxembourg	2	Zambia	1