Actividad 1.7 Análisis Factorial I

Raúl Correa Ocañas

2024-08-21

El problema del Lago

Se trata de varias variables limnológicas de varios lagos Neotropicales. Se hace un estudio para saber el grado de productividad potencial del lago (concentración de nutrientes y carbono orgánico disuelto) y la adecuación del hábitat en lo que se refiere a sus condiciones para la vida (profundidad, pH, conductividad, oxígeno disuelto y temperatura). Estos dos factores, productividad y hábitat, podrían explicar razonablemente la correlación observada entre las distintas variables. Se trata de hacer un análisis factorial en este contexto para comprobar si este modelo responde razonablemente a la realidad y es, por lo tanto, adecuado para explicar los siguientes datos.

Importación de Datos.

```
data_path = "../data/datoslago.csv"
df = read.csv(data_path)
head(df,5)
```

```
##   Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## 1    7.5        13.8      65.9     2.8    4.188138 2.014903
## 2    5.8        13.8      20.3     2.8    3.010621 1.757858
## 3    8.0        13.8     136.6     2.8    4.917057 2.079442
## 4    8.0        13.8      70.5     2.8    4.255613 2.079442
## 5   10.0        13.8     117.8     2.8    4.768988 2.302585
```

Realicen el análisis factorial, discutiendo y comentando los resultados obtendidos de:

Ejercicio 1.

Obtener la matriz de correlaciones y la matriz de valores p de significancia por pares.

```
corr.test(df,adjust="none")
```

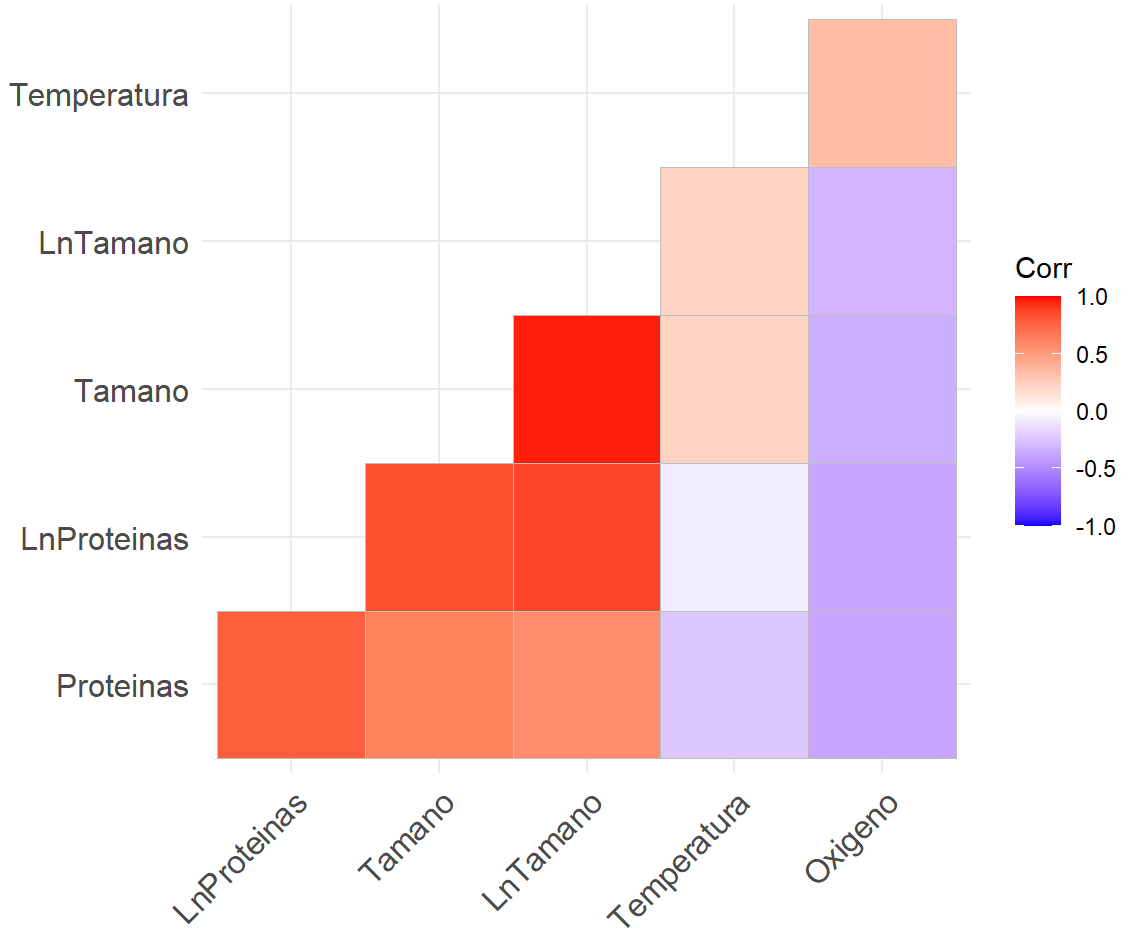
```
## Call:corr.test(x = df, adjust = "none")
## Correlation matrix
##             Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## Tamano        1.00        0.23      0.63   -0.34        0.84     0.97
## Temperatura   0.23        1.00     -0.22    0.34       -0.08     0.22
## Proteinas     0.63       -0.22      1.00   -0.37        0.78     0.57
## Oxigeno      -0.34        0.34     -0.37    1.00       -0.38    -0.31
## LnProteinas   0.84       -0.08      0.78   -0.38        1.00     0.86
## LnTamano      0.97        0.22      0.57   -0.31        0.86     1.00
## Sample Size
## [1] 90
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##             Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## Tamano        0.00        0.03      0.00       0        0.00     0.00
## Temperatura   0.03        0.00      0.04       0        0.46     0.04
## Proteinas     0.00        0.04      0.00       0        0.00     0.00
## Oxigeno       0.00        0.00      0.00       0        0.00     0.00
## LnProteinas   0.00        0.46      0.00       0        0.00     0.00
## LnTamano      0.00        0.04      0.00       0        0.00     0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Según lo observado como resultado de salida para este comando, se tienen las correlaciones por combinación de variables y sus respectivos valores p. Estos valores indican la significancia estadística de los resultados obtenidos. Según lo observado en la segunda tabla, se acepta una hipotesis nula para la correlación de las variables Temperatura y LnProteinas, indicando que su correlación no es estadísticamente significativa. Se observa que las variables con mejor correlaciones son: LnTamaño y Tamaño, LnTamaño y Proteina, y LnProteina y Proteina.

Ejercicio 2.

Hacer una gráfica de la matriz de correlaciones.

```
mat_cor = hetcor(df)$correlations
ggcorrplot(mat_cor,type="lower",hc.order = T)
```



En esta gráfica podemos ver de manera visual la correlación de las variables, mientras más rojas con mayor correlacion positiva y mientras más azules más correlacion negativa. En la gráfica se excluyen la repetición de las combinaciones de las variables, por lo que permite un mejor analisis de que correlaciones tienen mayor intensidad. El gráfico sugiere que Ln Tamaño y Tamaño tienen una fuerte correlación positiva, y después hay otros tres pares de combinaciones que visualmente no se puede apreciar la diferencia de correlaciones entre estos, por lo que es más complicado observar cual tendría mayor o menor correlación.

Ejercicio 3.

Aplicar una prueba de correlación conjunta a los datos para verficar si es aplicable el Análisis Factorial y concluir.

H_0: Las variables son ortogonales (matriz identidad) H_A: Las variables no son ortogonales (difiere de matriz identidad)

```
b = check_sphericity_bartlett(df)

b
```

```
## # Test of Sphericity
##
## Bartlett's test of sphericity suggests that there is sufficient significant correlation in th
e data for factor analysis (Chisq(15) = 557.86, p < .001).
```

```
b$chisq
```

```
## [1] 557.8585
```

```
b$p
```

```
## [1] 3.135832e-109
```

```
b$dof
```

```
## [1] 15
```

Según los resultados obtenidos en la prueba de esfericidad de Bartlet, se tiene un valor de p menor a 0.001, por lo que se puede inferir que existe suficiente evidencia estadística y correlación significante para poder realizar un análisis factorial.

Ejercicio 4.

Otra prueba para, para comprobar si el análisis factorial es viable, y muy citada, es la prueba KMO. Aplíquela a estos datos, ¿contradice los resutados del inciso anterior?

```
cor_df = cor(df)

K = KMO(cor_df)
cat("El valor del estadístico es: ", K$MSA)
```

```
## El valor del estadístico es:  0.6297281
```

Según los resultados obtenidos con la prueba KMO, los resultados caerían en el intervalo 0.60 a 0.69, lo cual se considera mediocre según los autores. Aunque utilizar descripciones de este estilo no enriquecen mucho la información que se tiene de los datos, ciertamente contradice la certeza que se tiene en la prueba de Chi Cuadrada, la cual con mucha certeza ($p < 0.001$) indica que los datos son aptos para un análisis factorial. Sin embargo, un estadístico de 0.5 indica para terminos prácticos una aceptación de la correlación conjunta. Esto significa que se tiene argumentos para sustentar que ambas pruebas concuerdan en el resultado.

Ejercicio 5.

Si los datos pasaron la prueba de los puntos anteriores 3 y 4, hacer un análisis factorial usando el criterio de máxima verosimilitud y el de mínimo residuo.

```
# Análisis de máxima verosimilitud
fa_mle = fa(cor(df), nfactors = 2, rotate = "none", fm = "mle")

# Análisis de mínimo residuo
fa_minres = fa(cor(df), nfactors = 2, rotate = "none", fm = "minres")
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
# Comunalidades
communalities = data.frame(MLE = fa_mle$communality, MINRES = fa_minres$communality)
communalities = communalities[order(communalities$MLE, decreasing = TRUE), ]
communalities
```

```
##                   MLE    MINRES
## LnTamano    0.9950455 0.9238582
## Tamano      0.9558283 0.9748176
## LnProteinas 0.9181230 0.9030553
## Proteinas   0.7708383 0.6029645
## Temperatura 0.3963254 0.8453996
## Oxigeno     0.2071029 0.2835476
```

```
eigenvalues = data.frame(MLE = fa_mle$values, MINRES = fa_minres$values)
eigenvalues = eigenvalues[order(eigenvalues$MLE, decreasing = TRUE), ]
eigenvalues
```

```
##           MLE      MINRES
## 1  3.41569944  3.37445056
## 2  0.88778561  1.15919198
## 3  0.13495076  0.12203164
## 4  0.03652320  0.01513825
## 5 -0.02272192 -0.04791798
## 6 -0.20897353 -0.08925167
```

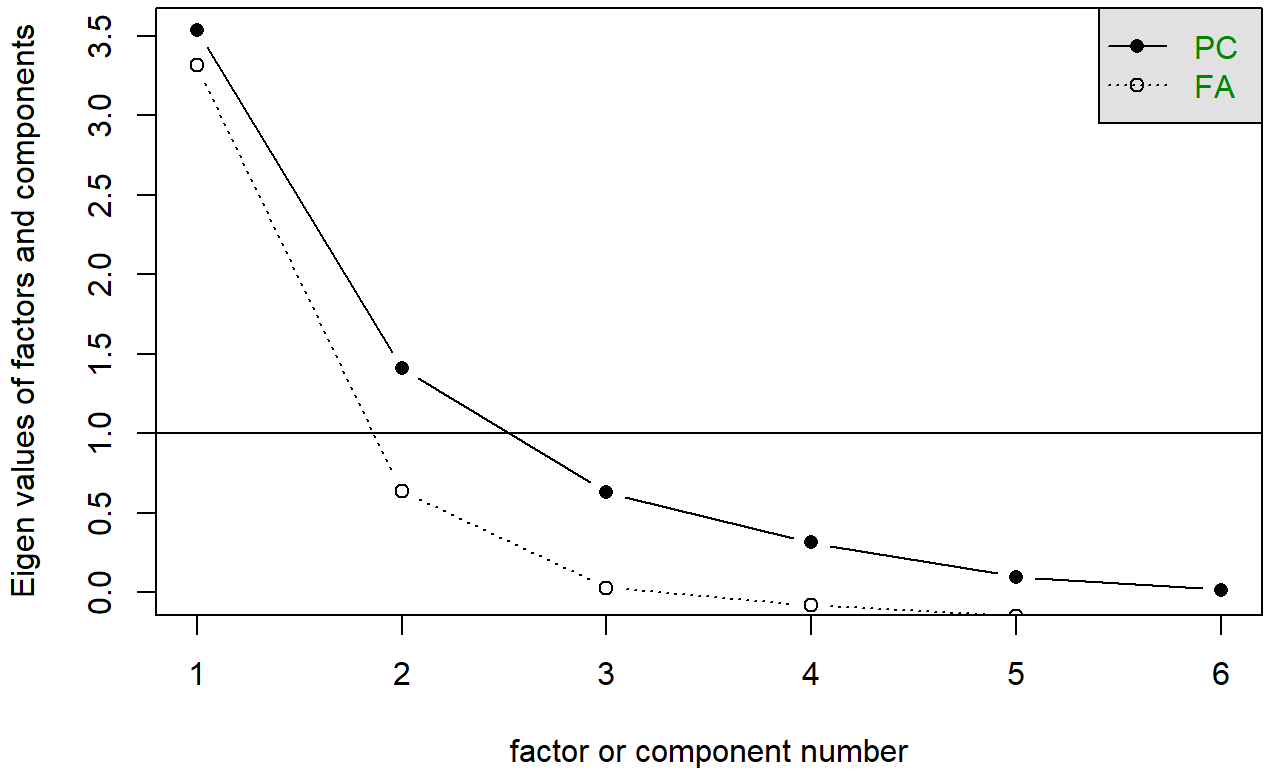
Se muestran las comunalidades de las variables, que indican qué proporción de la varianza de cada variable es explicada por los factores extraídos. En ambos modelos, LnTamaño, Tamaño y LnProteinas son explicadas a al menos el 90% de sus variabilidades con los factores propuestos. La variable Proteinas pueden ser explicadas a al menos un 77% con los factores propuestos en el método de maxima verosimilitud, y al menos un 60% en el método de mínimo residuo. La temperatura es altamente explicada utilizando minimos residuos, pero su explicación es pésima en el método de máxima verosimilitud. La variable Oxígeno tanto para ambos modelos no es adecuada, por lo que ambos métodos no explican con los dos factores esta variable.

Ejercicio 6.

Determine el número de factores adecuado según el criterio del gráfico de Cattell

```
scree(cor_df)
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

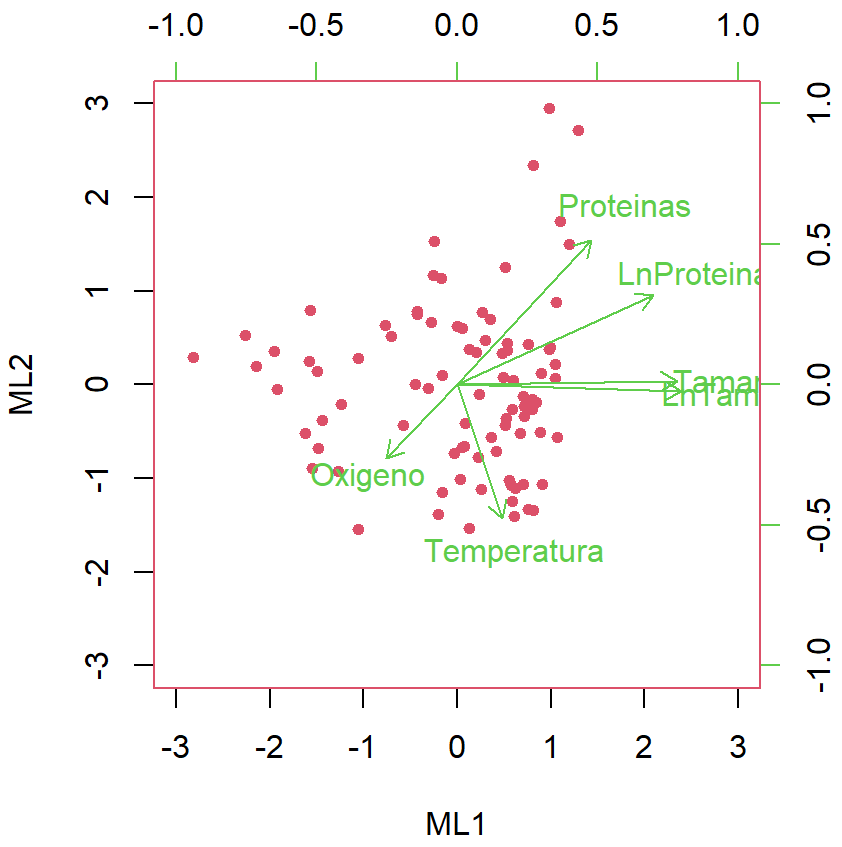


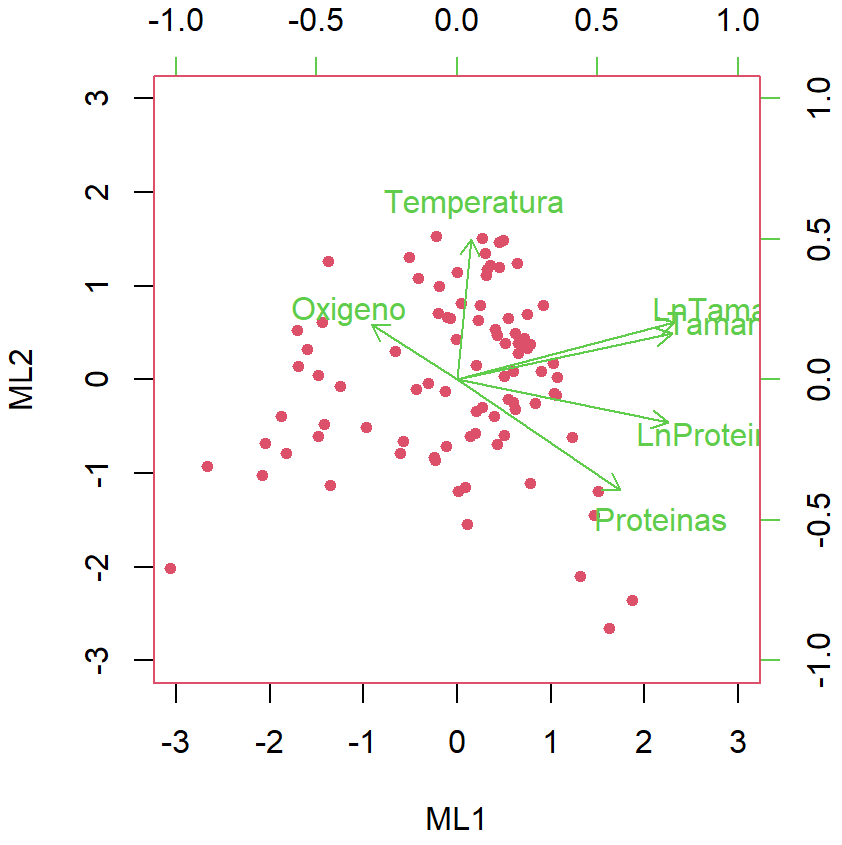
Tanto para el método de MLE como MINRES, se puede decidir escoger entre 2 y 3 variables para el análisis factorial. Seleccionar más de 3 variables implicaría poca explicación de los datos a costo de un aumento de dimensionalidad, por lo que en este caso se puede considerar aceptable explicar los datos con dos factores, teniendo como mejor opción el método de MINRES.

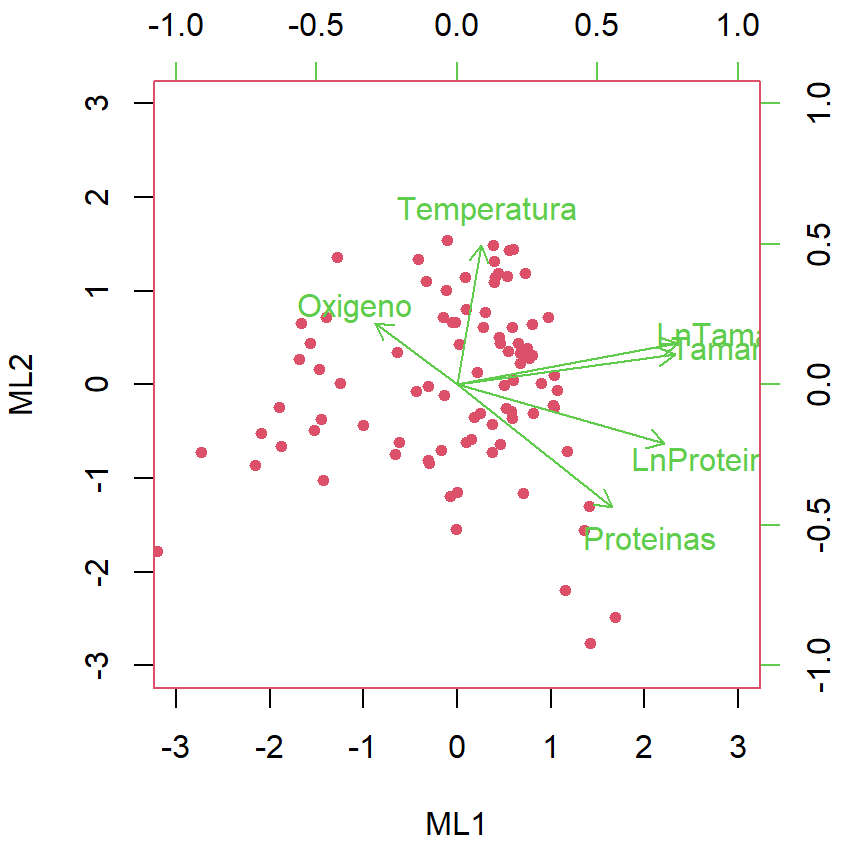
Ejercicio 7.

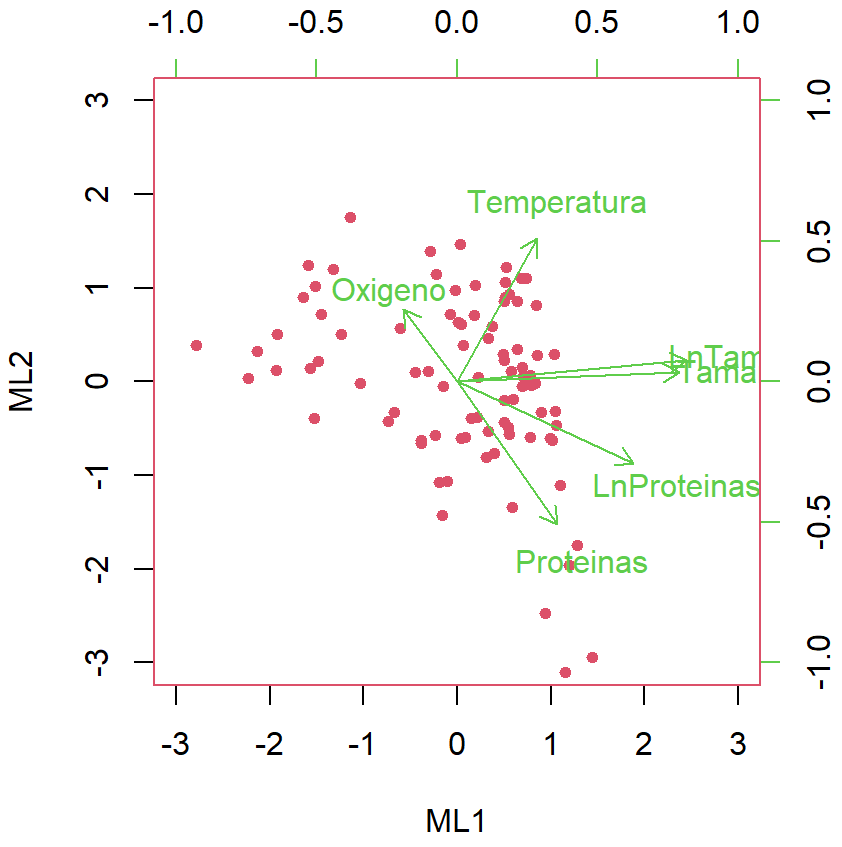
Realicen los gráficos correspondientes a la rotación Varimax y quartimax de los datos e interpreten en equipo los resultados.

```
rot = c("none", "varimax", "quartimax", "oblimin")
bi_mod = function(tipo){
biplot.psych(fa(df, nfactors = 2, fm="mle", rotate = tipo), main = "",col=c(2,3,4),pch = c(21,1
8))  }
sapply(rot,bi_mod)
```


-1.0
-0.5
0.0
0.5
1.0
3
2
1
0
-1
-2
-3
ML2
Proteinas
LnProteina
Tamar
LnTam
Oxigeno
Temperatura
ML1


-1.0
-0.5
0.0
0.5
1.0
3
2
1
0
-1
-2
-3
ML2
Temperatura
Oxigeno
LnTama
Tamar
LnProtei
Proteinas
ML1

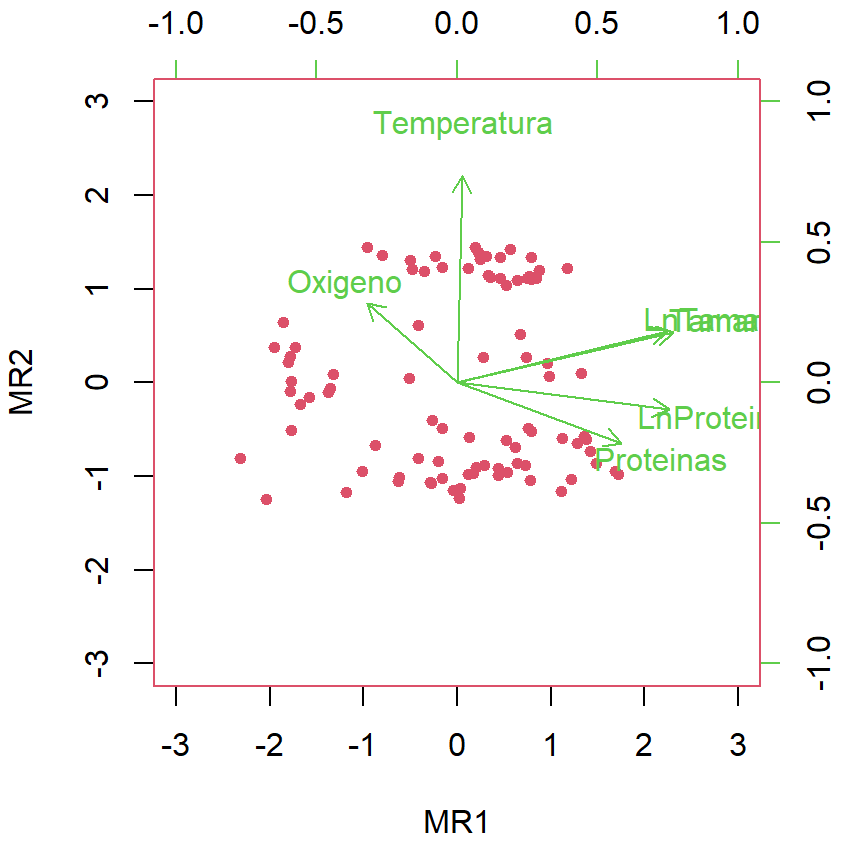

-1.0
-0.5
0.0
0.5
1.0
3
2
1
0
-1
-2
-3
ML2
Temperatura
Oxigeno
LnTama
Tamaı
LnProtein
Proteinas
1.0
0.5
0.0
-0.5
-1.0
-3
-2
-1
0
1
2
3
ML1


-1.0
-0.5
0.0
0.5
1.0
3
2
1
0
-1
-2
-3
ML2
Temperatura
Oxigeno
LnTam
Tama
LnProteinas
Proteinas
1.0
0.5
0.0
-0.5
-1.0
-3
-2
-1
0
1
2
3
ML1

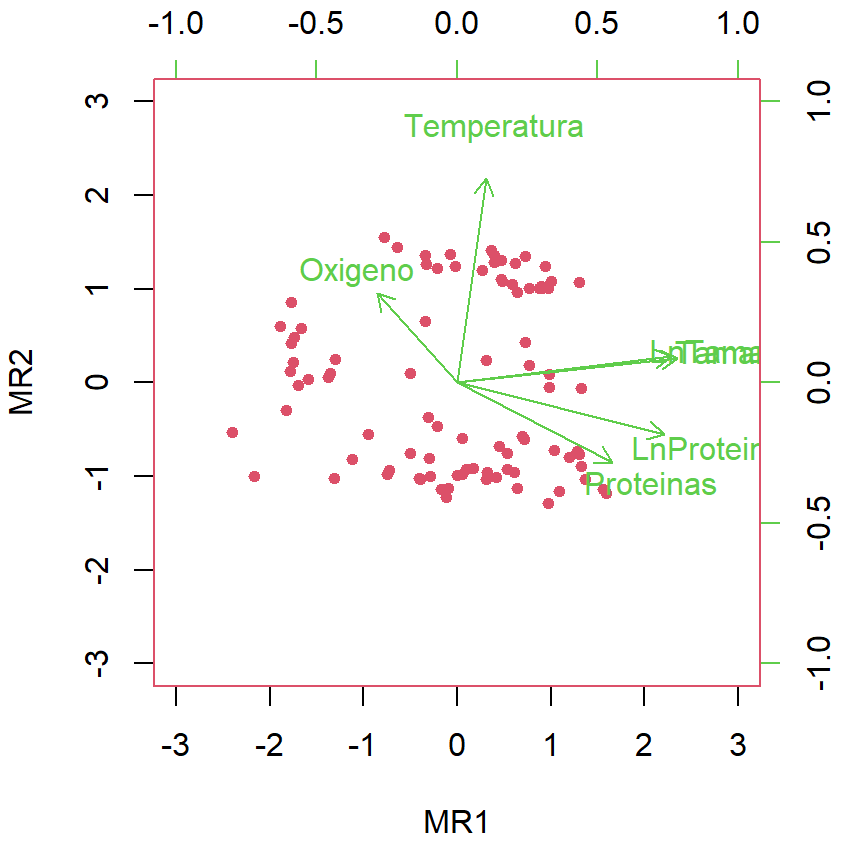
```
## $none
## NULL
##
## $varimax
## NULL
##
## $quartimax
## NULL
##
## $oblimin
## NULL
```

```
rot = c("none", "varimax", "quartimax", "oblimin")
bi_mod = function(tipo){
biplot.psych(fa(df, nfactors = 2, fm="minres", rotate = tipo), main = "",col=c(2,3,4),pch = c(2
1,18))  }
sapply(rot,bi_mod)
```

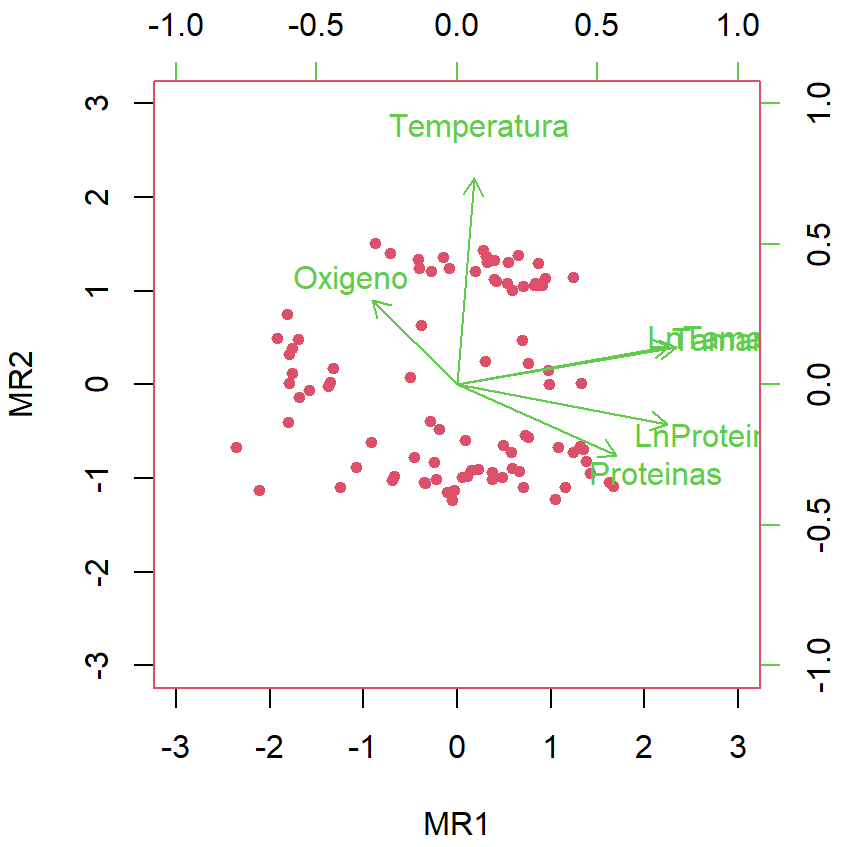
```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

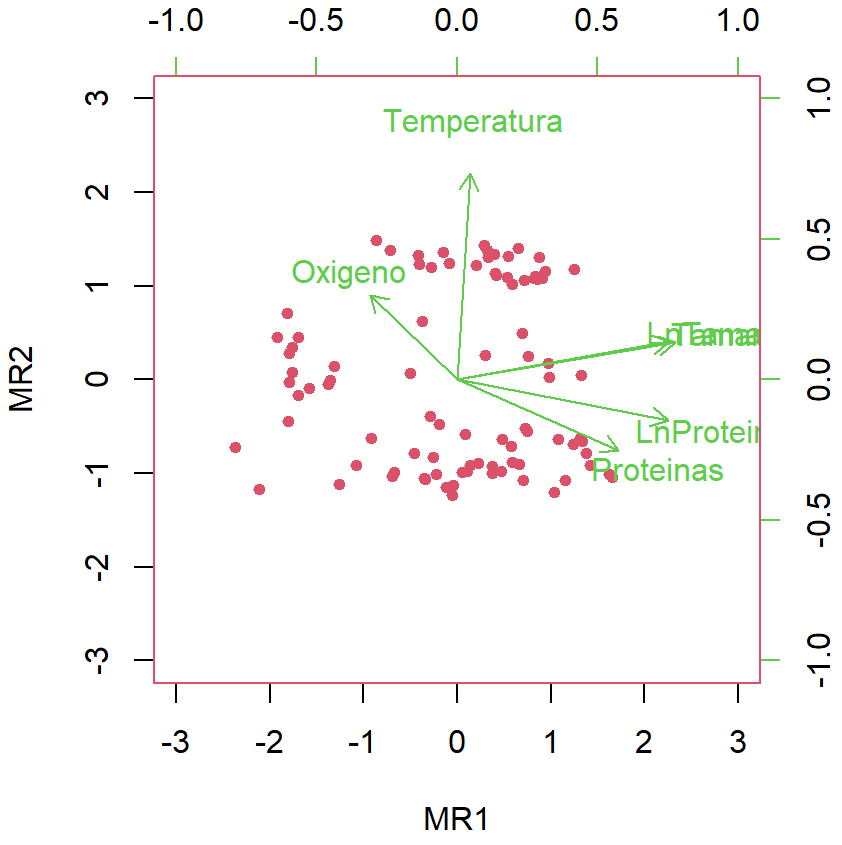



```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```




```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```


-1.0
-0.5
0.0
0.5
1.0
Temperatura
Oxigeno
LnTamaa
LnProtein
Proteinas
MR2
MR1
-3
-2
-1
0
1
2
3


-1.0
-0.5
0.0
0.5
1.0
Temperatura
Oxigeno
LnTamaa
LnProtein
Proteinas
MR2
MR1
-3
-2
-1
0
1
2
3

```
## $none
## NULL
##
## $varimax
## NULL
##
## $quartimax
## NULL
##
## $oblimin
## NULL
```

```
fa_mle = fa(cor(df), nfactors = 2, rotate = "none", fm = "mle")
fa_mle_varimax = fa(cor(df), nfactors = 2, rotate = "varimax", fm = "mle")
fa_mle_quartimax = fa(cor(df), nfactors = 2, rotate = "quartimax", fm = "mle")
fa_mle_oblimin = fa(cor(df), nfactors = 2, rotate = "oblimin", fm = "mle")

fa_minres = fa(cor(df), nfactors = 2, rotate = "none", fm = "minres")
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
fa_minres_varimax = fa(cor(df), nfactors = 2, rotate = "varimax", fm = "minres")
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
fa_minres_quartimax = fa(cor(df), nfactors = 2, rotate = "quartimax", fm = "minres")
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
fa_minres_oblimin = fa(cor(df), nfactors = 2, rotate = "oblimin", fm = "minres")
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
# communalities_mle = data.frame(NONE = fa_mle$communalities, VARIMAX = fa_mle_varimax$communalities, QUARTIMAX = fa_mle_quartimax$communalities, OBLIMIN = fa_mle_oblimin$communalities)
# communalities_minres = data.frame(NONE = fa_minres$communalities, VARIMAX = fa_minres_varimax$communalities, QUARTIMAX = fa_minres_quartimax$communalities, OBLIMIN = fa_minres_oblimin$communalities)
#
# eigenvalues_mle = data.frame(NONE = fa_mle$values, VARIMAX = fa_mle_varimax$values, QUARTIMAX = fa_mle_quartimax$values, OBLIMIN = fa_mle_oblimin$values)
# eigenvalues_minres = data.frame(NONE = fa_minres$values, VARIMAX = fa_minres_varimax$values, QUARTIMAX = fa_minres_quartimax$values, OBLIMIN = fa_minres_oblimin$values)
```

```
# communalities_mle
# communalities_minres
#
# eigenvalues_mle
# eigenvalues_minres
```

```
print("NONE")
```

```
## [1] "NONE"
```

```
fa_mle$loadings
```

```
##
## Loadings:
##             ML1    ML2
## Tamano       0.978
## Temperatura  0.202 -0.596
## Proteinas    0.597  0.644
## Oxigeno     -0.318 -0.326
## LnProteinas  0.872  0.398
## LnTamano     0.997
##
##                  ML1   ML2
## SS loadings    3.207 1.036
## Proportion Var 0.535 0.173
## Cumulative Var 0.535 0.707
```

```
print("VARIMAX")
```

```
## [1] "VARIMAX"
```

```
fa_mle_varimax$loadings
```

```
##
## Loadings:
##             ML1    ML2
## Tamano       0.956  0.204
## Temperatura         0.626
## Proteinas    0.727 -0.493
## Oxigeno     -0.383  0.246
## LnProteinas  0.939 -0.192
## LnTamano     0.964  0.255
##
##                  ML1   ML2
## SS loadings    3.404 0.839
## Proportion Var 0.567 0.140
## Cumulative Var 0.567 0.707
```

```
print("QUARTIMAX")
```

```
## [1] "QUARTIMAX"
```

```
fa_mle_quartimax$loadings
```

```
##
## Loadings:
##             ML1    ML2
## Tamano       0.969  0.133
## Temperatura  0.108  0.620
## Proteinas    0.688 -0.545
## Oxigeno     -0.364  0.274
## LnProteinas  0.922 -0.261
## LnTamano     0.981  0.183
##
##                  ML1   ML2
## SS loadings    3.368 0.876
## Proportion Var 0.561 0.146
## Cumulative Var 0.561 0.707
```

```
print("OBLIMIN")
```

```
## [1] "OBLIMIN"
```

```
fa_mle_oblimin$loadings
```

```
##
## Loadings:
##             ML1    ML2
## Tamano       0.989
## Temperatura  0.356  0.636
## Proteinas    0.444 -0.637
## Oxigeno     -0.240  0.322
## LnProteinas  0.785 -0.364
## LnTamano     1.021
##
##                  ML1   ML2
## SS loadings    3.019 1.056
## Proportion Var 0.503 0.176
## Cumulative Var 0.503 0.679
```

```
print("NONE")
```

```
## [1] "NONE"
```

```
fa_minres$loadings
```

```
##
## Loadings:
##             MR1    MR2
## Tamano       0.961  0.226
## Temperatura         0.919
## Proteinas    0.728 -0.270
## Oxigeno     -0.399  0.352
## LnProteinas  0.943 -0.119
## LnTamano     0.934  0.228
##
##                  MR1   MR2
## SS loadings    3.374 1.159
## Proportion Var 0.562 0.193
## Cumulative Var 0.562 0.756
```

```
print("VARIMAX")
```

```
## [1] "VARIMAX"
```

```
fa_minres_varimax$loadings
```

```
##
## Loadings:
##             MR1    MR2
## Tamano       0.981  0.110
## Temperatura  0.131  0.910
## Proteinas    0.691 -0.354
## Oxigeno     -0.355  0.397
## LnProteinas  0.922 -0.230
## LnTamano     0.954  0.116
##
##                  MR1   MR2
## SS loadings    3.343 1.190
## Proportion Var 0.557 0.198
## Cumulative Var 0.557 0.756
```

```
print("QUARTIMAX")
```

```
## [1] "QUARTIMAX"
```

```
fa_minres_quartimax$loadings
```

```
##
## Loadings:
##             MR1    MR2
## Tamano       0.973  0.166
## Temperatura         0.916
## Proteinas    0.710 -0.314
## Oxigeno     -0.377  0.376
## LnProteinas  0.934 -0.177
## LnTamano     0.946  0.171
##
##                  MR1   MR2
## SS loadings    3.366 1.167
## Proportion Var 0.561 0.195
## Cumulative Var 0.561 0.756
```

```
print("OBLIMIN")
```

```
## [1] "OBLIMIN"
```

```
fa_minres_oblimin$loadings
```

```
##
## Loadings:
##             MR1    MR2
## Tamano       0.970  0.164
## Temperatura         0.916
## Proteinas    0.717 -0.316
## Oxigeno     -0.385  0.377
## LnProteinas  0.938 -0.179
## LnTamano     0.942  0.168
##
##                  MR1   MR2
## SS loadings    3.373 1.169
## Proportion Var 0.562 0.195
## Cumulative Var 0.562 0.757
```

Varimax:

En cuanto al método de MLE, Tamano y LnTamano tienen cargas altas en el primer factor, lo que sugiere que estos factores están altamente relacionados con el tamaño. La variable temperatura carga fuertemente en el segundo factor, lo que indica que este factor representa principalmente la temperatura. En cuanto al método de MINRES, las cargas son similares a las de Varimax con mle, con Tamano y LnTamano fuertemente cargando en el primer factor y Temperatura en el segundo.

Quartimax: Los resultados son similares a los de Varimax, pero con cargas más repartidas, dificultando la interpretación pero confirmando lo observado para Varimax.

Ejercicio 8.

¿Qué pueden concluir? ¿Resultó razonable para este caso el modelo de análisis factorial? Expliquen.

El análisis factorial realizado es razonable para este caso, debido a que la prueba de esfericidad de Bartlett indicó que las correlaciones entre las variables son suficientemente fuertes para el análisis. Confirmando con la prueba KMO, se sugiere que la adecuación no es óptima pero son aceptables. Los resultados de comunalidades y eigenvalues muestran que las variables principales, como LnTamaño y Tamaño, están bien representadas por los factores extraídos, aunque variables como Oxígeno tienen una explicación deficiente. La elección de dos factores, respaldada por el gráfico de Cattell, parece adecuada para evitar la sobrecomplicación del modelo. Las rotaciones Varimax y Quartimax facilitan la interpretación de los factores, revelando una estructura clara relacionada con el tamaño y la temperatura. Al observar los loadings de cada factor, se confirma que hace sentido que los factores sean productividad y adecuación del habitat. Un factor definitvamente está más cargado hacia la temperatura, las proteinas y el oxigeno, la cual está asociada a la productividad. El resto de las variables están asociadas con la adecuacion del habitat.