PAC1_Estadística Multivariante

Amelia Martínez Sequera

Las operaciones en R se muestran en el archivo Rmd anexo o en el siguiente enlace:

https://github.com/gititub/An-l.-Multivar.

Ejercicio 1.1

Media para cada gen WT (sólo se muestran los primeros resultados)

```
   NAT2     ADA     CDH2     AKT3     MED6    NR2E3
3.269111 7.256277 4.323965 5.449183 5.378173 3.327165
```

Media para cada gen MUT

```
   NAT2     ADA     CDH2     AKT3     MED6    NR2E3
3.265556 6.411842 4.556572 5.067518 5.553835 3.130331
```

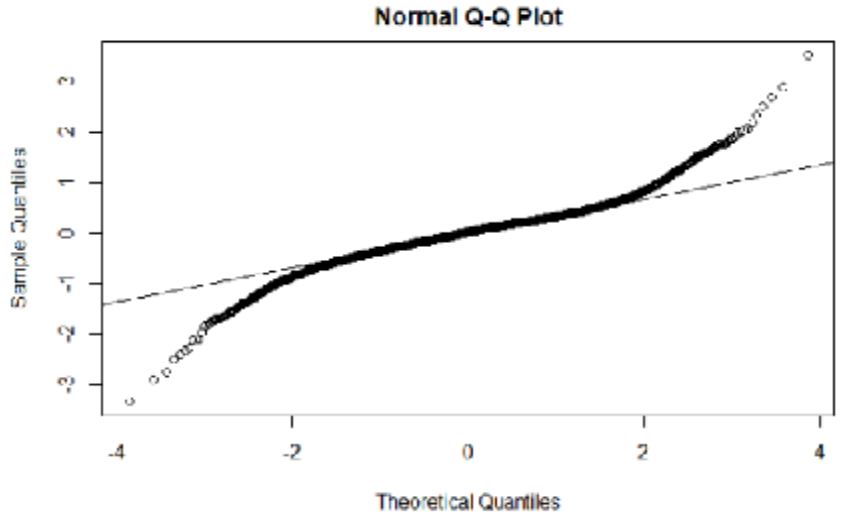
Varianza WT: 5.338128

Varianza MUT: 5.326398

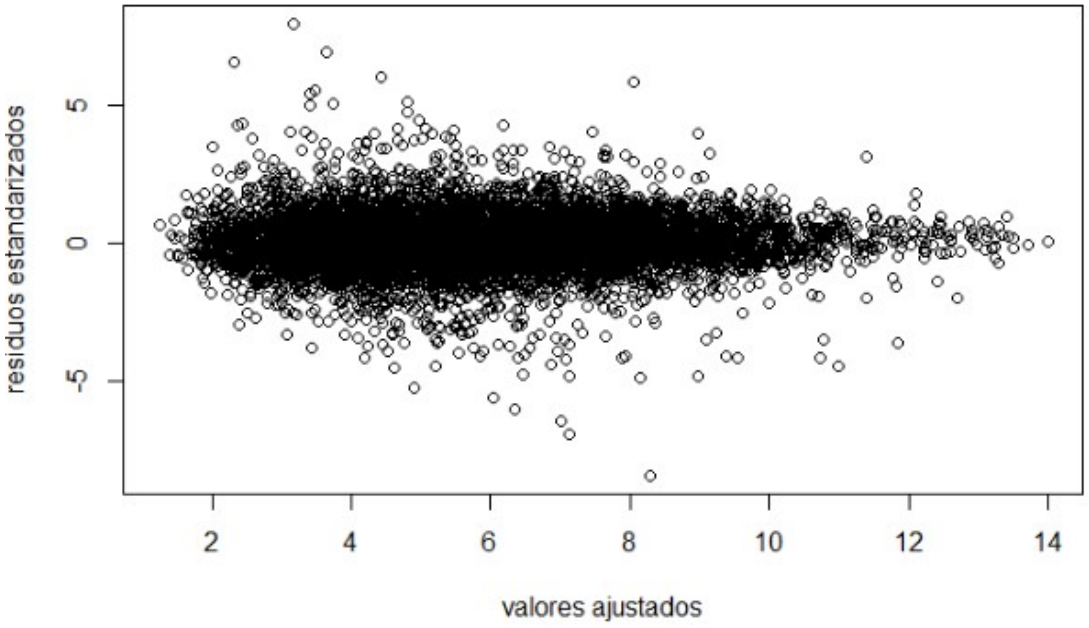
Estadístico t para cada gen (sólo se muestran los primeros):

```
$NAT2

        One Sample t-test

data:  newX[, i]
t = 1838, df = 1, p-value = 0.0003464
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.244746 3.289921
sample estimates:
mean of x
 3.267333


$ADA

        One Sample t-test

data:  newX[, i]
t = 16.186, df = 1, p-value = 0.03928
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  1.469279 12.198840
sample estimates:
mean of x
 6.834059


$CDH2

        One Sample t-test

data:  newX[, i]
t = 38.178, df = 1, p-value = 0.01667
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.962488 5.918049
```

qqplot

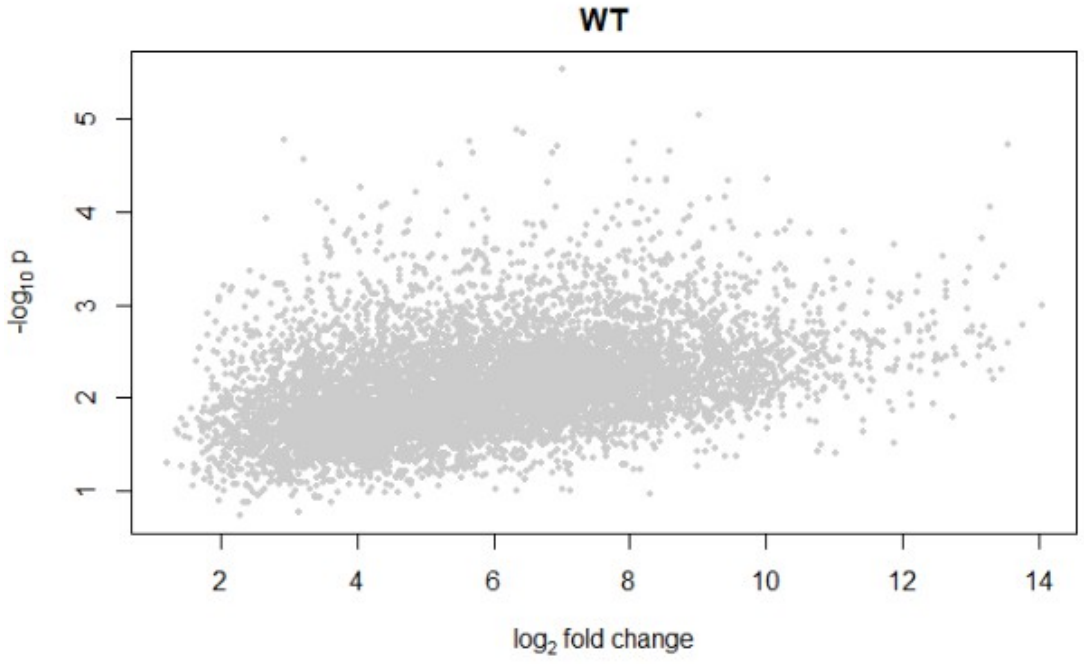


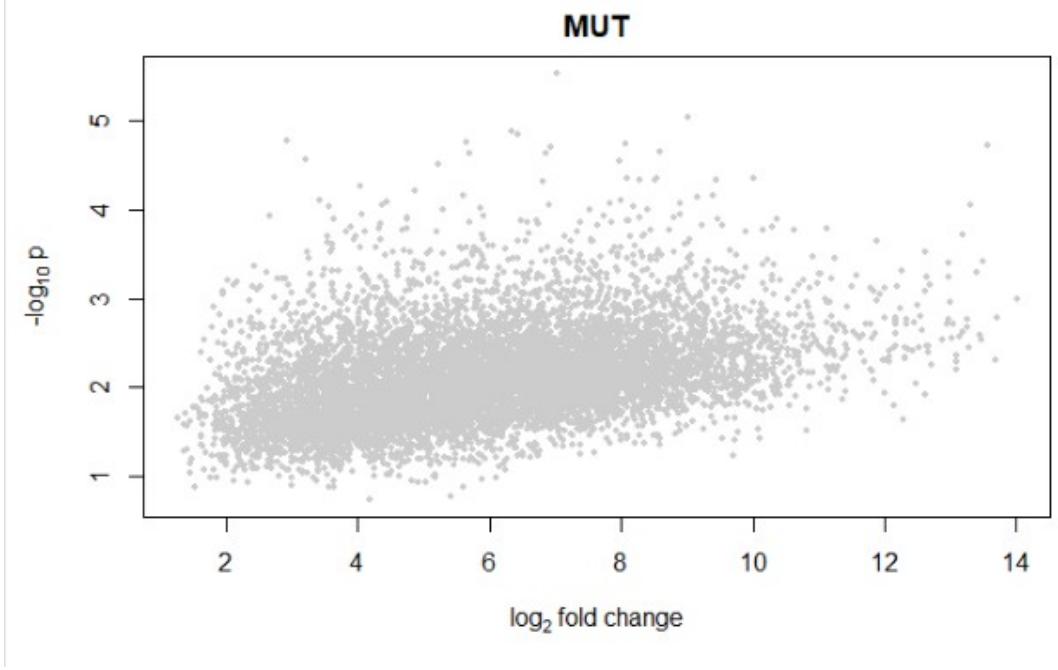
Se ajusta bastante.



Se agrupan alrededor del 0 mayormente.

Volcano plot





Ejercicio 1.2

```
sdg<- apply(p53DataSet,1,sd)
So<- median(sdg)
```

So = 0.6120621

Ejercicio 1.3

Test de Hotelling

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2,$$

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

```
tmut<- t(p53DataSet_mut)
twt <- t(p53DataSet_wt)
S2<- cov.shrink(tmut)
S1<- cov.shrink(twt)
S<- (16*S1+32*S2)/48
dim(S)
invS<- chol2inv(S)

x1<- as.vector(mediawt)
x2<- as.vector(mediamut)
x<- x1-x2
D2<- t(x)%*%invS%*%(x)

T2<-D2%*%(17*33/50)
T2
```

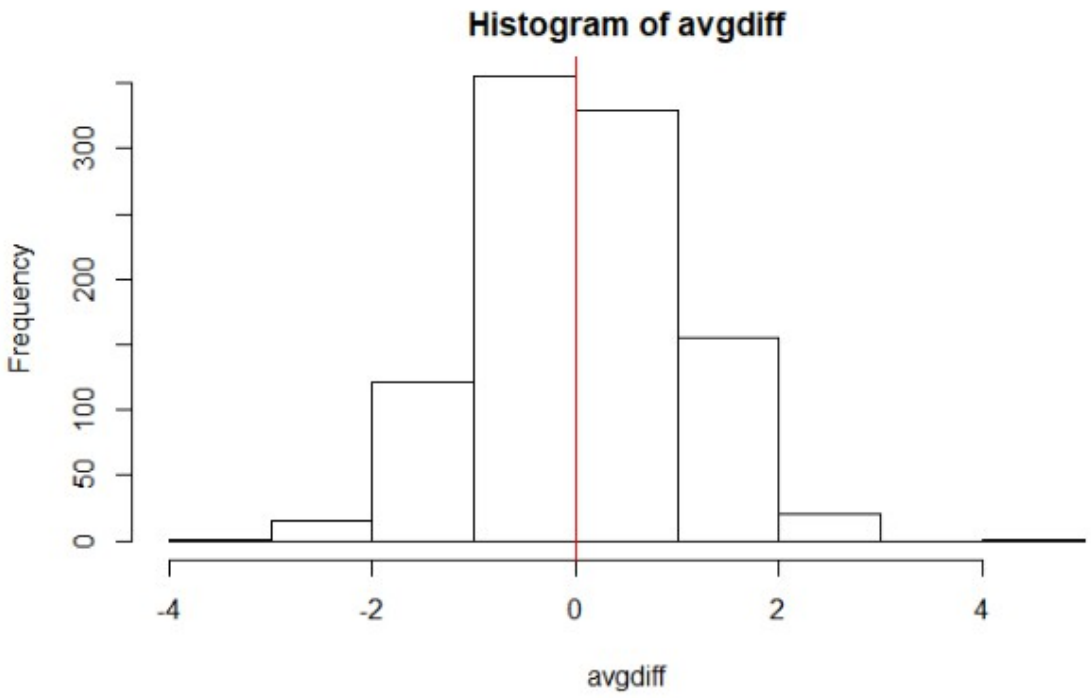
T2 = 28373.35

Ejercicio 1.4

Test de permutaciones

En el histograma se muestra la diferencia entre las medias de las permutaciones. La línia roja muestra la diferencia observada

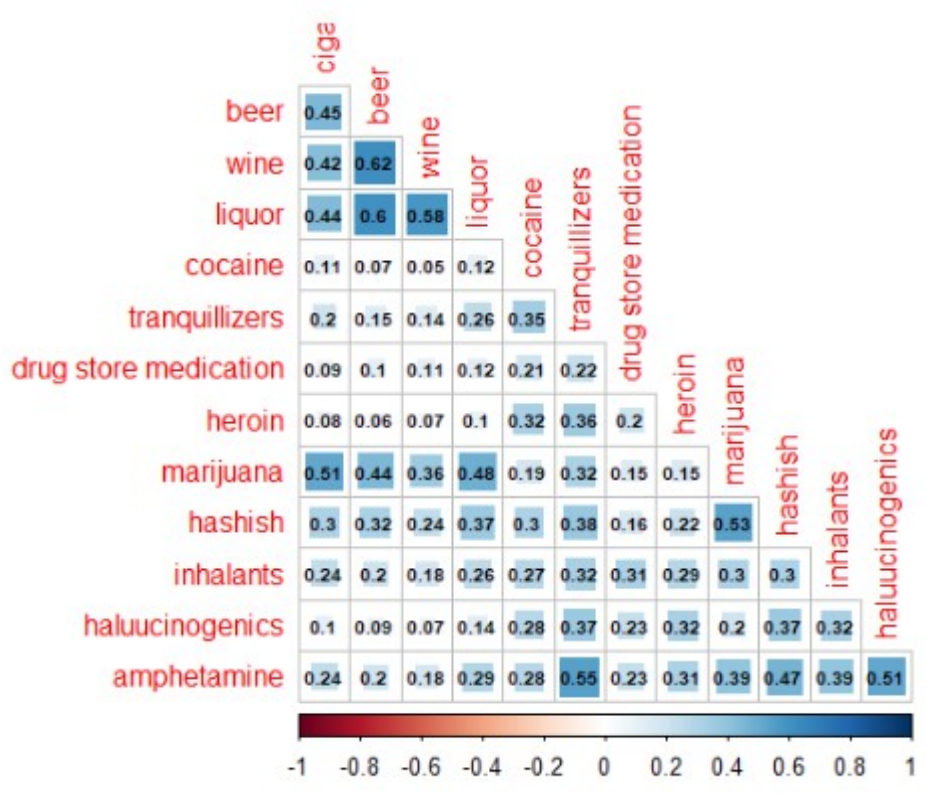
```
set.seed(2020)
obsdif<- mean(mediawt)-mean(mediamut)
N <- 10
avgdiff <- replicate(1000, {
  all <- sample(c(mediawt,mediamut))
  newwt <- all[1:N]
  newmut <- all[(N+1):(2*N)]
  return(mean(newwt) - mean(newmut))
})
hist(avgdiff)
abline(v=obsdif, col="red")
```

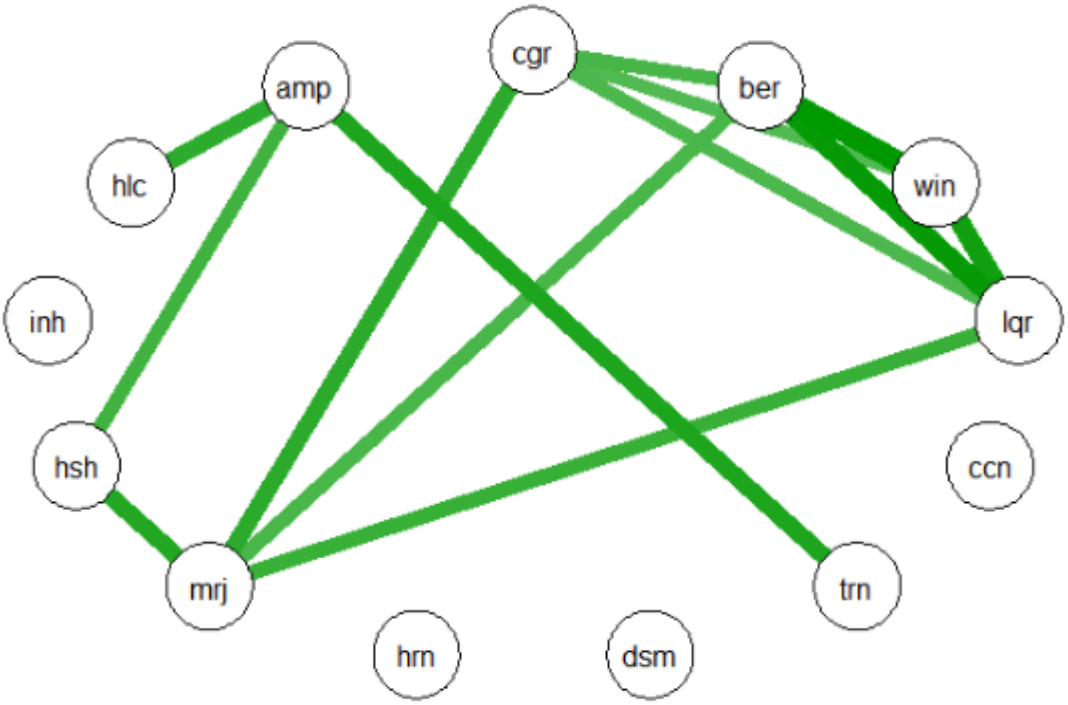


La proporción de medias nulas que sobrepasan el valor observado sería el p-valor.

Ejercicio 2

```
+ )
> isSymmetric(druguse.cor)
[1] FALSE
> newdrug<-(t(druguse.cor)+druguse.cor)/2
> isSymmetric(newdrug)
[1] TRUE
```





Se observa que las variables más correlacionadas són beer, wine y liquor y, algo menos cigarrettes, también éstas con marijuana. Marijuana-hashish y tranquilizers, hashish, marijuana,

inhalants , alucinogens con anphetamines también muestran más correlación.

Test de esfericidad

Test de Barlett

$chisq: 6584.031

$p.value: 0

$df: 78

Podemos rechazar la hipótesis nula de esfericidad (que la matriz de coeficientes de correlación no es significativamente distinta de la matriz identidad). Además, el resultado se puede considerar fiable, debido a que el tamaño de la muestra es grande. En caso contrario, no existirían correlaciones significativas entre las variables, y el modelo factorial no sería pertinente.

El estadístico está basado en el valor del determinante de la matriz de coeficientes de correlación:

-[n-1-(2k+5)/6]ln|R|

donde k= número de variables, n=tamaño de la muestra y R= matriz de correlaciones.

Índice KMO

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = newdrug)
Overall MSA =  0.87
MSA for each item =
          cigarettes                beer                wine               liquor
                0.89                0.84                0.83                0.88
             cocaine       tranquillizers drug store medication             heroin
                0.87                0.87                0.88                0.88
           marijuana             hashish           inhalants      haluucinogenics
                0.87                0.88                0.91                0.85
          amphetamine
                0.86
```

Como los índices son próximos a 1, el ACP se puede hacer (el valor 1 indica que cada variable es perfectamente predicha sin error por las otras variables).

Como partimos de la matriz de correlaciones, los valores ya estan normalizados.

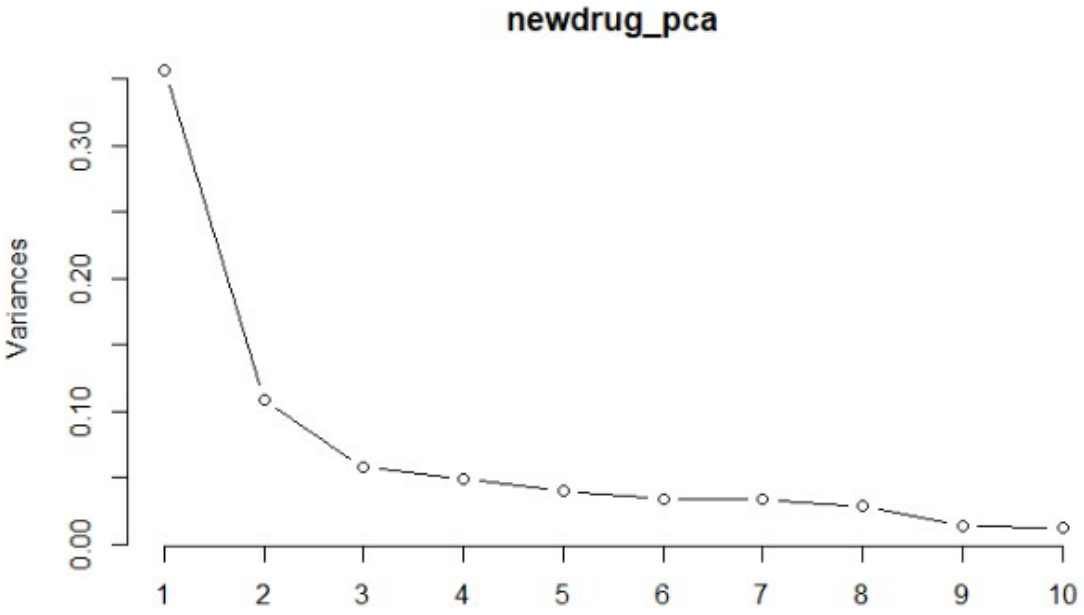
PCA

(Sólo se muestran los 10 primeros)

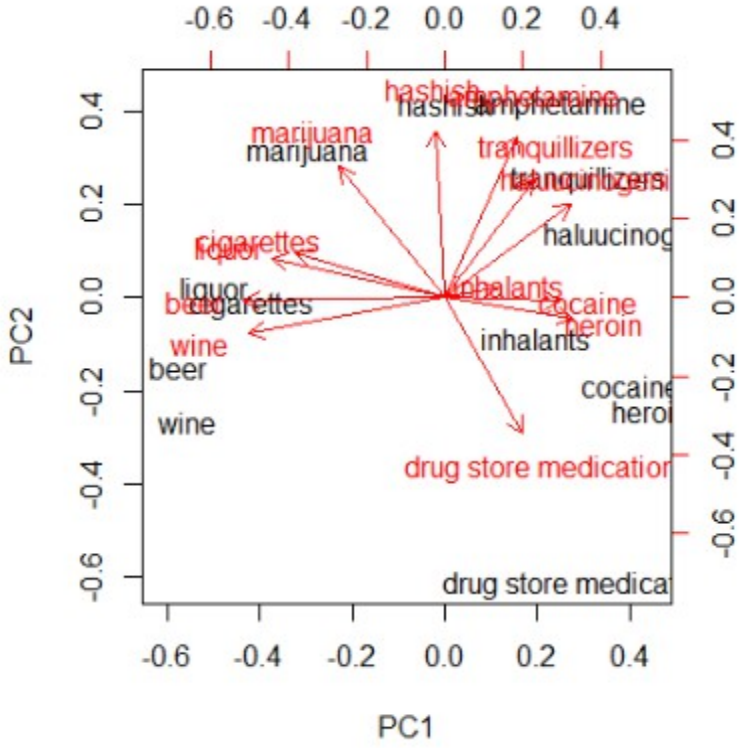
```
Standard deviations (1, .., p=13):
 [1] 5.972548e-01 3.294644e-01 2.407305e-01 2.213785e-01 2.002411e-01 1.840436e-01
 [7] 1.831712e-01 1.694289e-01 1.160626e-01 1.143205e-01 1.080335e-01 1.047561e-01
[13] 1.518572e-17

Rotation (n x k) = (13 x 13):
                           PC1          PC2          PC3         PC4          PC5
cigarettes           -0.32525761  0.128024786 -0.023031017 -0.3260365  0.38226000
beer                 -0.43842196 -0.009832734 -0.030022387  0.1809090 -0.10574943
wine                 -0.42710804 -0.103903230 -0.043988740  0.3131262 -0.09687669
liquor               -0.37807506  0.114225344  0.004969318  0.1805267 -0.08142882
cocaine               0.25241724 -0.010142515 -0.500461025 -0.4352031 -0.30087909
tranquillizers        0.19541705  0.347883749 -0.064431249  0.1242738  0.06000120
drug store medication 0.16943132 -0.394536066  0.581198547 -0.2106017 -0.21885337
heroin                0.28149930 -0.060508895 -0.411669381  0.3088642  0.43935048
marijuana            -0.23070619  0.382764628  0.071739564 -0.3597028  0.10921294
hashish              -0.02008608  0.486430040  0.031582778 -0.2401455 -0.27650675
inhalants             0.11329370  0.029010370  0.361037068 -0.1230152  0.58773267
haluucinogenics       0.27438732  0.273831458  0.214779000  0.3722899 -0.23998572
amphetamine           0.15820204  0.469335105  0.228277575  0.2120135  0.01477508
                           PC6          PC7          PC8         PC9          PC10
cigarettes           -0.063740494  0.29955351 -0.54602818  0.10882103 -0.037592235
beer                 -0.034224606 -0.12796701  0.13433789  0.09041970 -0.007417841
wine                  0.049021641 -0.16396125 -0.02858377  0.26674352  0.386967573
liquor                0.171996211 -0.23075072  0.18027656 -0.52824655 -0.384933176
cocaine              -0.003921203 -0.42433964 -0.12499681 -0.20225121  0.145141294
tranquillizers        0.727771274 -0.06947244 -0.13879101  0.27906636 -0.347035140
drug store medication 0.275260127  0.21954738  0.16520339 -0.07602241  0.037167049
heroin               -0.047340728  0.32504626  0.41905906 -0.09574823  0.062193573
marijuana            -0.006212562  0.23521441  0.23735781 -0.40765507  0.028522249
hashish              -0.191887061  0.06951004  0.47319802  0.49532835 -0.028066824
inhalants            -0.171488043 -0.64125791  0.15408434  0.09811105 -0.049371460
haluucinogenics      -0.505976791  0.04129293 -0.30722290 -0.10897050 -0.304960497
amphetamine           0.181175108 -0.04406820 -0.11965407 -0.25104149  0.675888913
```

Cada columna representa un vector con los coeficientes que, combinados con las variables, dan lugar a las componentes principales (o factores).

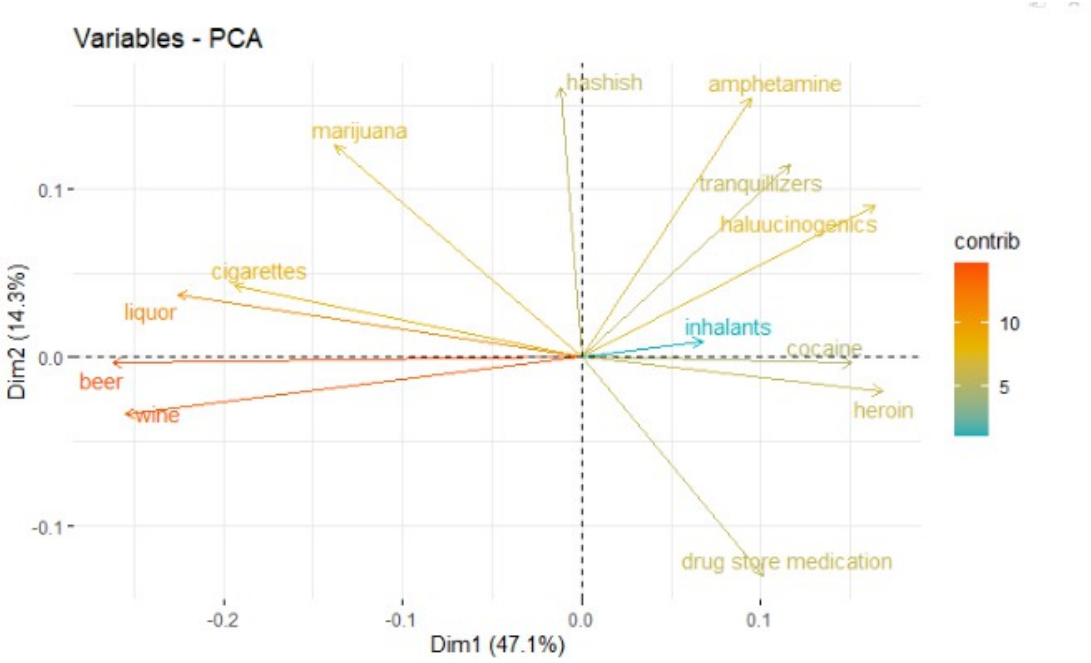
Análisis de los factores:



Visualmente, parece que podríamos escoger PC1-PC3, si miramos los datos de variabilidad acumulada, nos podríamos quedar con PC1-P6 que explicaría un 85% de la variabilidad, o un 90% si nos quedamos con P1-PC7.

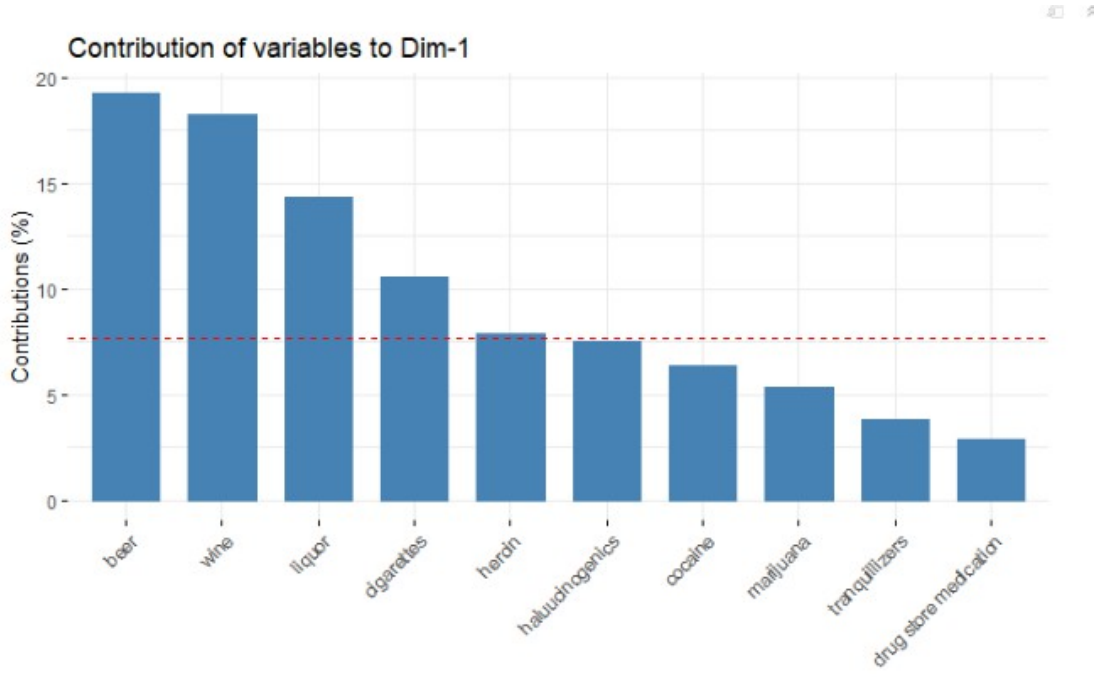


En el gráfico biplot lo que importa son las direcciones. PC1 seria la componente que explicaría más la variable beer (alineada con el 0), y también wine, liquor y cigarrettes, porque tienden más a PC2 (se mueven en el eje horizontal). De la misma manera, para PC2 sería hashis, y luego amphetamine y drugstoremedication. Se obtiene la misma conclusión si observamos los valores absolutos de los coeficientes.

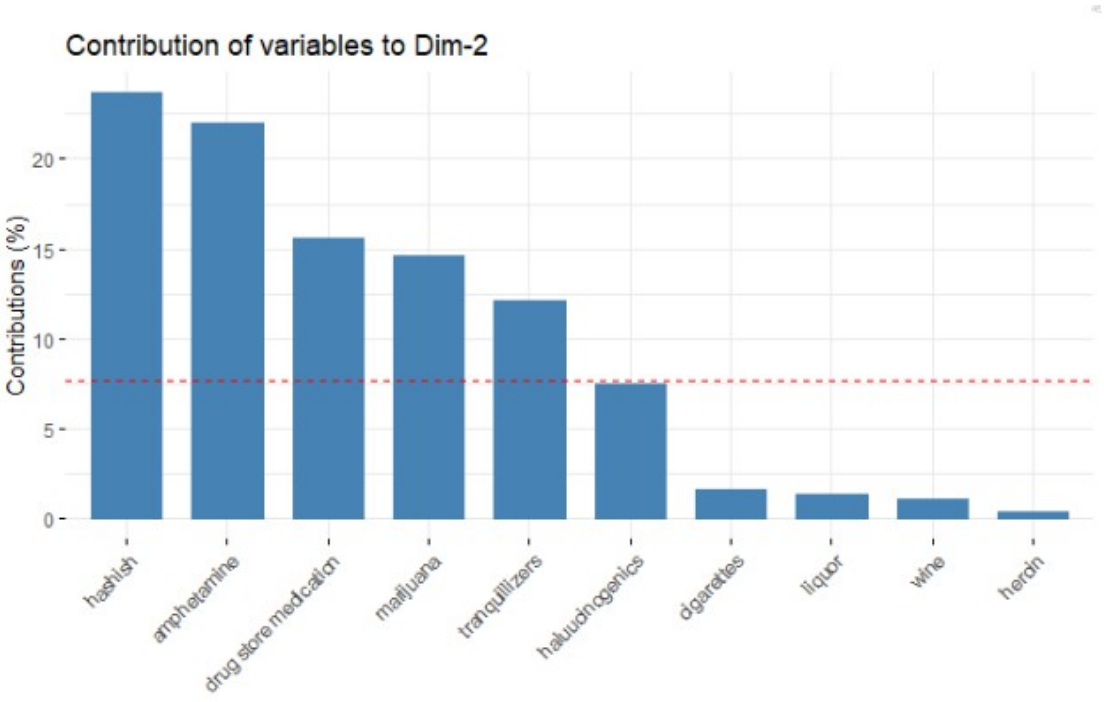


Se puede hacer un análisis por separado de la contribución de cada variable. Se observa que ya sólo con las dos primeras PC se explica una parte más o menos significativa de todas las variables.

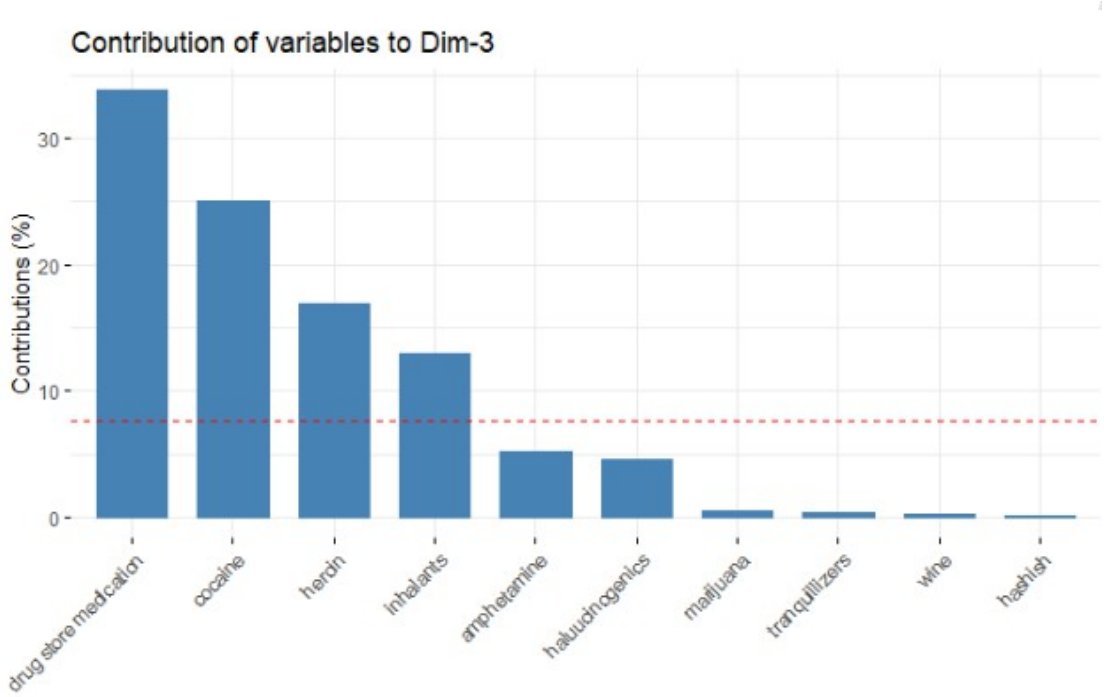
PC1:



PC2:



PC3:



Calculamos las diferencias entre las correlaciones observadas y predichas (6 factores) y observamos que son pequeñas, entonces, 6 factores es una buena elección. Pero si hacemos el cálculo ahora para 3 factores, comprobamos que las diferencias también son bastante pequeñas, también sería una elección acertada (se muestran sólo los primeros valores, ver código R adjunto):

```
pred <- newdrug_pca[[6]]$loadings%*%t(newdrug_pca[[6]]$loadings) +
diag(newdrug_pca[[6]]$uniquenesses)
round(newdrug-pred, digits=3)
```

```
                      cigarettes   beer   wine liquor cocaine tranquillizers
cigarettes                 0.000 -0.001  0.014 -0.018   0.010          0.001
beer                      -0.001  0.000 -0.002  0.004   0.004         -0.011
wine                       0.014 -0.002  0.000 -0.001  -0.001         -0.005
liquor                    -0.018  0.004 -0.001  0.000  -0.007          0.020
cocaine                    0.010  0.004 -0.001 -0.007   0.000          0.001
tranquillizers             0.001 -0.011 -0.005  0.020   0.001          0.000
drug store medication     -0.020 -0.001  0.008 -0.004   0.006          0.009
heroin                    -0.004  0.007  0.008 -0.018   0.004         -0.004
marijuana                  0.002  0.002 -0.004  0.003  -0.004         -0.003
hashish                    0.000  0.000  0.000  0.000   0.000          0.000
inhalants                  0.011 -0.004 -0.007  0.013  -0.003          0.001
haluucinogenics           -0.004  0.005 -0.001 -0.005  -0.008         -0.009
amphetamine                0.000  0.000  0.000  0.000   0.000          0.000
```

Ejercicio 3

Análisis factorial

Función *principal (),* sin rotación, para 6 componentes:

```
Principal Components Analysis
Call: principal(r = newdrug, nfactors = 6, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix

                        PC1  PC2  PC3  PC4  PC5  PC6
SS loadings            4.38 2.04 0.95 0.82 0.77 0.69
Proportion Var         0.34 0.16 0.07 0.06 0.06 0.05
Cumulative Var         0.34 0.49 0.57 0.63 0.69 0.74
Proportion Explained   0.45 0.21 0.10 0.08 0.08 0.07
Cumulative Proportion  0.45 0.67 0.76 0.85 0.93 1.00

Mean item complexity =  2.9
Test of the hypothesis that 6 components are sufficient.

The root mean square of the residuals (RMSR) is  0.07

Fit based upon off diagonal values = 0.95
```

	PC1 <S3: AsIs>	PC2 <S3: AsIs>	PC3 <S3: AsIs>	PC4 <S3: AsIs>	PC5 <S3: AsIs>	PC6 <S3: AsIs>
cigarettes	0.58	-0.40	-0.06	0.01	-0.29	-0.36
beer	0.60	-0.57	0.12	0.09	0.16	0.12
wine	0.56	-0.56	0.21	0.13	0.27	0.13
liquor	0.67	-0.46	0.05	0.06	0.16	0.12
cocaine	0.44	0.41	0.06	0.53	-0.38	0.33
tranquillizers	0.61	0.37	-0.16	0.07	0.11	0.00
drug store medication	0.37	0.27	0.71	-0.32	-0.18	0.22
heroin	0.42	0.45	0.15	0.48	0.27	-0.32
marijuana	0.71	-0.23	-0.23	-0.10	-0.32	-0.11
hashish	0.69	0.07	-0.34	-0.11	-0.21	0.21

Las cargas o loadings dan una idea sobre que peso tiene una variable en cada componente, define además la dirección en el espacio sobre el cual la varianza de los datos es mayor. De manera que observamos resultados similares

En nuestros resultados, SS loadings nos indica la saturación acumulada.

```
Loadings:
                      PC1    PC2    PC3    PC4    PC5    PC6
cigarettes             0.583 -0.400                -0.285 -0.355
beer                   0.599 -0.567  0.124         0.159  0.124
wine                   0.555 -0.560  0.211  0.130  0.275  0.126
liquor                 0.666 -0.464                0.159  0.122
cocaine                0.436  0.413         0.528 -0.380  0.330
tranquillizers         0.613  0.372 -0.164         0.109
drug store medication  0.368  0.271  0.708 -0.319 -0.184  0.221
heroin                 0.422  0.452  0.146  0.482  0.272 -0.324
marijuana              0.710 -0.233 -0.228 -0.102 -0.315 -0.108
hashish                0.688        -0.345 -0.110 -0.211  0.212
inhalants              0.576  0.237  0.313 -0.161 -0.107 -0.422
haluucinogenics        0.517  0.469 -0.123 -0.250  0.325  0.144
amphetamine            0.688  0.333 -0.228 -0.239  0.185

                 PC1   PC2   PC3   PC4   PC5   PC6
SS loadings    4.379 2.044 0.954 0.816 0.767 0.690
Proportion Var 0.337 0.157 0.073 0.063 0.059 0.053
Cumulative Var 0.337 0.494 0.567 0.630 0.689 0.742
```

```
newdrug_ppal$values
```

```
 [1] 4.3791774 2.0436423 0.9540391 0.8155787 0.7668768 0.6900562 0.6365578 0.6173830
 [9] 0.5665840 0.3992407 0.3941099 0.3741506 0.3626034
```

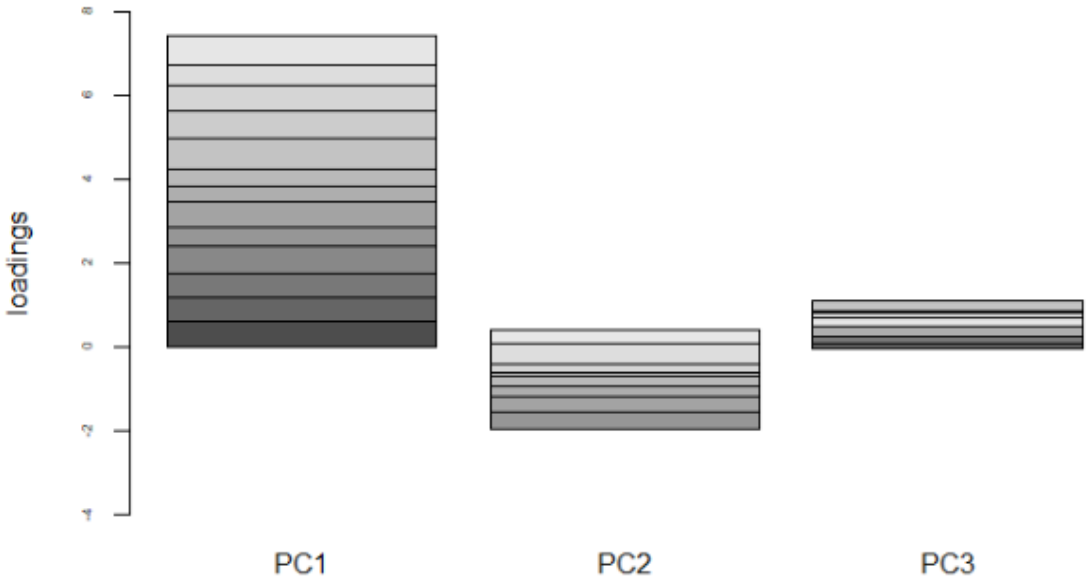
Los loadings son proporcionales a los vectores propios del PCA.

Con la función *fa()*, método: "pa" (principal factor solution), sin rotación:

```
Factor Analysis using method =  pa
Call: fa(r = newdrug, nfactors = 6, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix

                       PA1  PA2  PA3  PA4  PA5  PA6
SS loadings           3.95 1.57 0.45 0.32 0.22 0.17
Proportion Var        0.30 0.12 0.03 0.02 0.02 0.01
Cumulative Var        0.30 0.42 0.46 0.48 0.50 0.51
Proportion Explained  0.59 0.24 0.07 0.05 0.03 0.03
Cumulative Proportion 0.59 0.83 0.89 0.94 0.97 1.00

Mean item complexity =  2.4
Test of the hypothesis that 6 factors are sufficient.

The degrees of freedom for the null model are  78  and the objective function was  4.04
The degrees of freedom for the model are 15  and the objective function was  0.02

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 PA1  PA2  PA3   PA4   PA5   PA6
Correlation of (regression) scores with factors 0.95 0.89 0.72  0.70  0.58  0.51
Multiple R square of scores with factors        0.91 0.79 0.52  0.49  0.33  0.26
Minimum correlation of possible factor scores   0.82 0.57 0.05 -0.02 -0.34 -0.47
```

	PA1 <S3: AsIs>	PA2 <S3: AsIs>	PA3 <S3: AsIs>	PA4 <S3: AsIs>	PA5 <S3: AsIs>	PA6 <S3: AsIs>
cigarettes	0.54	-0.31	-0.08	0.09	-0.15	-0.11
beer	0.58	-0.52	0.13	-0.04	0.08	0.06
wine	0.54	-0.51	0.24	-0.10	0.07	0.01
liquor	0.64	-0.40	0.08	-0.07	0.06	-0.02
cocaine	0.38	0.33	0.10	0.23	0.17	-0.08
tranquillizers	0.57	0.35	0.03	-0.06	0.07	-0.23
drug store medication	0.31	0.19	0.21	0.12	-0.12	0.11
heroin	0.37	0.35	0.18	0.11	0.08	-0.09
marijuana	0.70	-0.19	-0.35	0.17	-0.14	-0.07
hashish	0.67	0.10	-0.32	0.09	0.21	0.19

Gráfico de barras de los 3 primeros factores en función de los loadings:



MLFA.

Tal y como se describe a continuación, los resultados obtenidos con la funciones *factanal()* y *fa()* son muy similares, debido a que se basan en el mismo método de máxima verosimilitud. Este método proporciona las estimaciones de los parámetros que con mayor probabilidad han producido la matriz de correlaciones observada, bajo el supuesto de que la muestra procede de una distribución normal multivariada.

Función *factanal()*, método: "mle"

```
Call:
factanal(factors = i, covmat = newdrug, method = "mle")

Uniquenesses:
           cigarettes                  beer                  wine                liquor
                0.563                 0.368                 0.374                 0.411
              cocaine        tranquillizers drug store medication               heroin
                0.682                 0.521                 0.768                 0.667
            marijuana               hashish             inhalants       haluucinogenics
                0.320                 0.005                 0.571                 0.627
          amphetamine
                0.005

Loadings:
                     Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
cigarettes            0.494                           0.409   0.105
beer                  0.775                           0.112
wine                  0.785
liquor                0.721   0.123   0.104   0.114   0.161
cocaine                       0.518           0.132           0.160
tranquillizers        0.131   0.566   0.320   0.104   0.144
drug store medication         0.252                           0.397
heroin                        0.535   0.101                   0.187
marijuana             0.428   0.159   0.153   0.259   0.608   0.107
hashish               0.244   0.279   0.188   0.881   0.195
inhalants             0.169   0.317   0.158           0.144   0.498
haluucinogenics               0.393   0.340   0.186           0.260
amphetamine           0.151   0.339   0.887   0.143   0.138   0.177

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
SS loadings       2.301   1.432   1.124   0.962   0.679   0.621
Proportion Var    0.177   0.110   0.086   0.074   0.052   0.048
Cumulative Var    0.177   0.287   0.374   0.448   0.500   0.548

The degrees of freedom for the model is 15 and the fit was 0.0141
```

Uniquenesses (unicidad) es el porcentaje de variabilidad que no es explicada por los factores. Es igual a 1 - comunalidad.

En cambio la comunalidad (loadings) es el porcentaje de la variabilidad de la variable explicada por ese factor.

Si todos los factores explican conjuntamente un gran porcentaje de varianza en una variable dada, esa variable tiene una alta comunalidad (y por lo tanto una singularidad baja).

Para facilitar la interpretación del significado de los factores seleccionados, se suele llevar a cabo una rotación de los ejes factoriales. Se utiliza para ajustar la varianza que expli cará el factor.

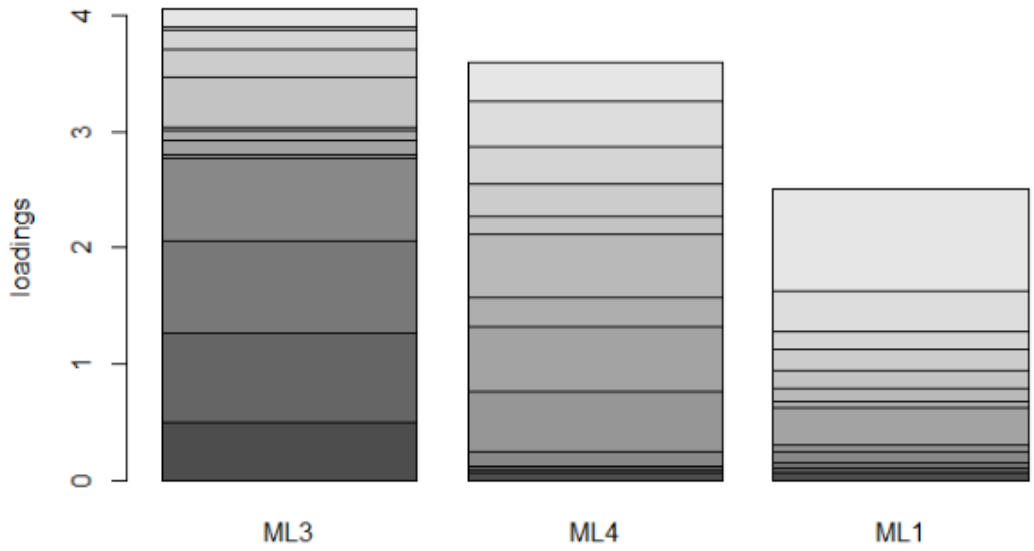
Con la rotacion varimax(rotación ortogonal de los ejes factoriales) se obtiene el mejor resultado, ya que practicamente asimila cada variable con un eje. Este es un estudio comparativo de las marcas, no evaluativo. Pueden ser todas muy buenas o muy malas, pero el estudio determina unicamente las diferencias entre ellas, no el valor; este se aprecia estudiando los valores iniciales.

Función *fa()*, método: "ml" (maximum likelihood factor analysis), rotación *varimax:*

```
Factor Analysis using method =  ml
Call: fa(r = newdrug, nfactors = 6, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix

                       ML3  ML4  ML1  ML2  ML5  ML6
SS loadings           2.30 1.43 1.12 0.96 0.68 0.62
Proportion Var        0.18 0.11 0.09 0.07 0.05 0.05
Cumulative Var        0.18 0.29 0.37 0.45 0.50 0.55
Proportion Explained  0.32 0.20 0.16 0.14 0.10 0.09
Cumulative Proportion 0.32 0.52 0.68 0.82 0.91 1.00

Mean item complexity =  1.8
Test of the hypothesis that 6 factors are sufficient.

The degrees of freedom for the null model are  78  and the objective function was  4.04
The degrees of freedom for the model are 15  and the objective function was  0.01

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 ML3  ML4  ML1  ML2   ML5   ML6
Correlation of (regression) scores with factors 0.90 0.74 0.97 0.98  0.71  0.59
Multiple R square of scores with factors        0.81 0.54 0.94 0.95  0.50  0.35
Minimum correlation of possible factor scores   0.61 0.09 0.88 0.91 -0.01 -0.31
```

	ML3 <S3: AsIs>	ML4 <S3: AsIs>	ML1 <S3: AsIs>	ML2 <S3: AsIs>	ML5 <S3: AsIs>	ML6 <S3: AsIs>
cigarettes	0.49	0.07	0.07	0.07	0.41	0.10
beer	0.78	0.02	0.04	0.10	0.11	0.08
wine	0.79	0.03	0.04	0.03	0.01	0.08
liquor	0.72	0.12	0.10	0.11	0.16	0.07
cocaine	0.02	0.52	0.05	0.13	0.06	0.16
tranquillizers	0.13	0.57	0.32	0.10	0.14	0.08
drug store medication	0.08	0.25	0.07	0.02	0.02	0.40
heroin	0.03	0.53	0.10	0.03	0.01	0.19
marijuana	0.43	0.16	0.15	0.26	0.61	0.11
hashish	0.24	0.28	0.19	0.88	0.19	0.09

Barplot de los tres primeros componentes en función de los loadings:


loadings
0
1
2
3
4
ML3
ML4
ML1