Trabajo práctico N°2

Mariana Vargas V.

July 12, 2017

1 Análisis de Componentes Principales

El Análisis de Componentes Principales es una técnica multivariada que consiste en una transformación ortogonal de una matrix de datos en otra cuyas variables resulten linealmente independientes. Estas, llamadas *componentes principales*, son tales que representan las direcciones de mayor variabilidad de los datos, permitiendo así la posibilidad de reducir la dimensión.

1.1 Nuestros datos

Llevamos adelante el análisis de componentes principales en la base de datos `indice.RData` que tiene información sobre las finanzas de un grupo de empresas. Con el comando `prcomp` de R obtuvimos las componentes principales. La salida de `summary(pc)` sugiere que las dos primeras son las más importantes dado que representan el 99% de la varianza.

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     2.7263 1.5827 0.19559 0.13514
Proportion of Variance 0.7433 0.2505 0.00383 0.00183
Cumulative Proportion  0.7433 0.9938 0.99759 0.99942

    PC5     PC6     PC7      PC8     PC9      PC10
0.07218 0.01957 0.01399 0.006285 0.00238 2.239e-17
0.00052 0.00004 0.00002 0.000000 0.00000 0.000e+00
0.99994 0.99998 1.00000 1.000000 1.00000 1.000e+00
```

Podemos confirmar esto en la figura 1 en donde se muestran un gráfico de barra de las componentes contra la varianza.

En la figura 2 podemos ver a las variables representadas en función de las dos primeras componentes.

Observamos que algunas variables se aproximan a la primera componente, tales como la rentabilidad económica, el margen de explotación y el costo marginal de financiamiento, y otras que se aproximan a la segunda componente, como la inmovilización del activo. También hay variables que mantienen con las componentes una relación negativa, como la inmovilización del patrimonio, el pasivo, y la solvencia. La matriz de correlaciones (tabla 1) confirma esto.

Repetimos el análisis usando la matriz de correlación de los datos. La salida muestra que esta vez las cuatro primeras componentes suman el 95% de la correlación entre las variables, lo cual podemos ver en la figura 3.

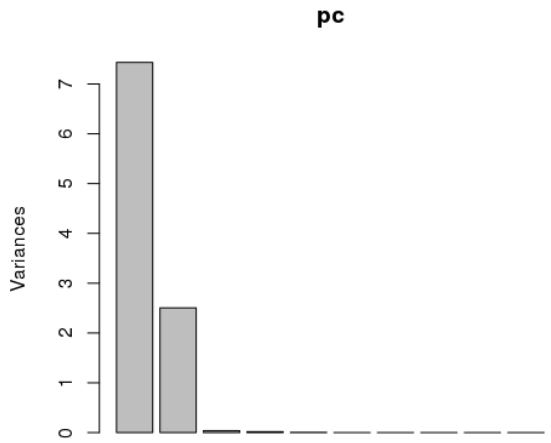

Figure 1: Componentes vs. la varianza.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.4670 1.5222 0.79441 0.71303 0.55569
Proportion of Variance 0.6086 0.2317 0.06311 0.05084 0.03088
Cumulative Proportion  0.6086 0.8403 0.90341 0.95425 0.98513

     PC6     PC7     PC8     PC9      PC10
 0.32151 0.16390 0.13310 0.02710 7.353e-18
 0.01034 0.00269 0.00177 0.00007 0.000e+00
 0.99547 0.99816 0.99993 1.00000 1.000e+00
```

La figura 4 muestra las variables originales en función de las componentes elegidas, que a diferencia del análisis anterior, maximizan la correlación entre variables y no su varianza.

De la matriz de correlación entre componentes y variables muestra una relación positiva entre la inmovilización del activo y del patrimonio, y del pasivo respecto de la primera componente, mientras que la propiedad del activo mantiene una relación negativa.

2 Análisis de correspondencia

Mediante el análisis de correspondencia estudiamos tablas de contingencia de variables clasificatorias. Esta técnica es análoga al análisis de componentes principales.

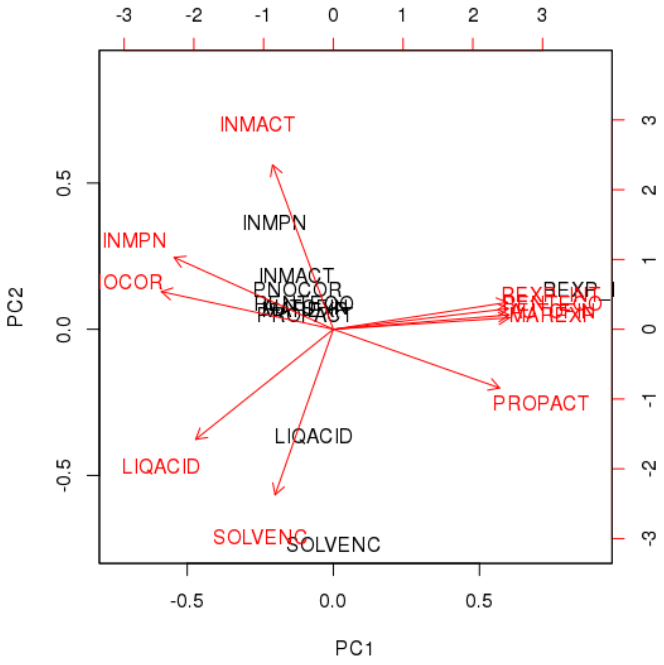

Figure 2: Biplot en las dos primeras componentes.

2.1 Resultados en los datos

Trabajamos con una base de datos que tiene información sobre una encuensta sobre hogares. Analizamos las variables correspondientes al estado civil y al nivel de educación. Con el comando `ca` obtenemos el análisis de correspondencia. La raíz cuadrada de la suma de los autovalores nos da un índice de correlación entre filas y columnas. En este caso es de 0.4, lo que indica una baja correlación. Esto es, el nivel de educación no necesariamente define el estado civil, y viceversa. En la figura 5 podemos visualizar los resultados. Recordemos que puntos correspondientes a las filas que estén más cerca indican perfiles columna asociados más parecidos. Por ejemplo, observar que las personas unidas y casadas tienden a tener el mismo nivel de educación, contrario a lo que ocurre con personas viudas y solteras, que se ven alejadas en el gráfico.

La función `mjca` nos permite hacer un análisis de correspondencia múltiple. En este caso la inercia indica una correlación del 84% entre las variables fila y las variables columna.

2.2 Equivalencia distribucional de la distancia χ^2

Propiedad 1. *Dada una tabla de contingencia, si dos filas i, i' son tales que sus estructuras relativas son iguales, es decir,*

$$\frac{f_{i,j}}{f_{i,.}} = \frac{f_{i',j}}{f_{i',.}}$$

	1	2	3	4	5	6	7	8	9	10
LIQACID	-0.78	-0.62	0.03	0.01	-0.02	-0.01	-0.00	-0.00	-0.00	0.00
SOLVENC	-0.33	-0.94	0.10	0.03	-0.00	0.00	0.00	0.00	-0.00	0.00
PROPACT	0.94	-0.33	0.03	0.03	-0.00	0.01	0.00	-0.00	0.00	-0.00
PNOCOR	-0.97	0.21	-0.03	-0.01	-0.06	0.00	0.00	0.00	0.00	-0.00
AUTOFIN	1.00	0.08	0.04	-0.02	-0.01	-0.01	-0.00	0.00	0.00	-0.00
INMACT	-0.34	0.93	0.08	0.09	-0.00	-0.00	-0.00	0.00	-0.00	0.00
INMPN	-0.90	0.41	0.11	-0.08	0.01	0.00	-0.00	-0.00	0.00	-0.00
RENTECO	0.99	0.12	0.04	-0.01	-0.01	-0.00	0.00	-0.00	-0.00	-0.00
MAREXP	1.00	0.06	0.01	-0.02	-0.03	0.01	-0.01	0.00	0.00	0.00
REXP_INT	0.99	0.15	0.05	-0.01	-0.01	0.00	0.01	-0.00	-0.00	0.00

Table 1: Matriz de correlaciones entre componentes y variables usando la matriz S.

	1	2	3	4	5	6	7	8	9	10
LIQACID	-0.75	-0.65	0.04	0.01	0.00	0.12	0.05	0.04	0.02	0.00
SOLVENC	-0.81	-0.57	-0.06	0.04	-0.04	0.06	0.04	0.03	-0.02	0.00
PROPACT	-0.91	-0.28	-0.05	0.29	-0.01	-0.07	-0.04	-0.01	0.00	-0.00
PNOCOR	0.88	0.07	0.34	0.23	0.04	0.20	-0.05	0.03	-0.00	-0.00
AUTOFIN	-0.74	0.45	-0.18	-0.19	0.42	0.05	-0.03	0.04	-0.00	-0.00
INMACT	0.86	0.03	-0.11	0.44	0.21	-0.09	0.07	0.03	0.00	0.00
INMPN	0.94	0.08	-0.06	-0.30	-0.04	0.04	0.08	0.01	-0.00	-0.00
RENTECO	-0.72	0.65	0.05	0.17	0.02	0.11	0.07	-0.08	-0.00	-0.00
MAREXP	-0.67	0.40	0.61	-0.06	-0.00	-0.11	0.03	0.04	-0.00	-0.00
REXP_INT	-0.37	0.83	-0.29	0.12	-0.28	0.04	-0.00	0.07	0.00	0.00

Table 2: Matriz de correlación de las variables en función de las componentes.

para $j = 1, ..., J$, y son combinadas en una única fila p, entonces,

$$d_{\chi^2}(p, p') = d_{\chi^2}(i, p') = d_{\chi^2}(i', p')$$

para cualquier otra fila p'.

Proof. Tomemos dos filas arbitrarias i y i' tales que cumplen la hipótesis, esto es,

$$\frac{f_{i,j}}{f_{i,.}} = \frac{f_{i',j}}{f_{i',.}} = k_j$$

con $j = 1, ..., J$. Notar que la unión de dos filas consiste en la suma de ambas, lo que resulta en la suma de los totales y de cada conteo. Por lo tanto la fila que resulta de la unión de i e i', p, es tal que

$$\frac{f_{p,j}}{f_{p,.}} = \frac{f_{i,j} + f_{i',j}}{f_{i,.} + f_{i',.}}$$

Para demostrar la propiedad basta con probar que $\frac{f_{p,j}}{f_{p,.}} = k_j$ para todo $j = 1, ..., J$, lo que se deduce de la ecuación:

$$\frac{f_{p,j}}{f_{p,.}} = \frac{f_{i,j} + f_{i',j}}{f_{i,.} + f_{i',.}}$$

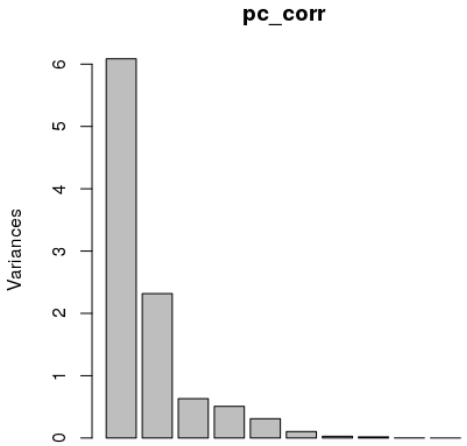

Figure 3: Gráfico de las componentes vs. la varianza, usando la matriz de correlación.

Luego las distancias a otras filas se mantienen invariables. □

3 Análisis factorial

El análisis factorial tiene como fin explicar y representar a los datos observados en términos de variables latentes o no observadas. Esto es, busca inferir un modelo

$$X = \mu + \Lambda f + u$$

en donde f es un vector de factores, μ el vector de medias, Λ una matriz de coeficientes que ponderan el peso de cada factor para cada variable observada, y u un vector de error. Intuitivamente podemos pensar que mediante el análisis factorial estamos modelando la correlación entre las variables observadas, si bien los factores pueden o no ser independientes.

3.1 Análisis factorial en nuestros datos

Nuestra base de datos contiene índices de índole social tales como niveles de educación, empleo, y necesidades, sobre personas distribuidas en diferentes departamentos de la provincia de Córdoba. Para estudiar a qué se deben estos comportamientos y encontrar variables latentes que los expliquen usaremos el comando de R `factanal`. Asumiremos varios escenarios.

1. **Dos factores sin rotación.** En esta configuración el test de hipótesis arroja un p-valor de casi 0 que indica que dos factores son suficientes. En la tabla 3 podemos observar la matriz de cargas. Notar que el primer

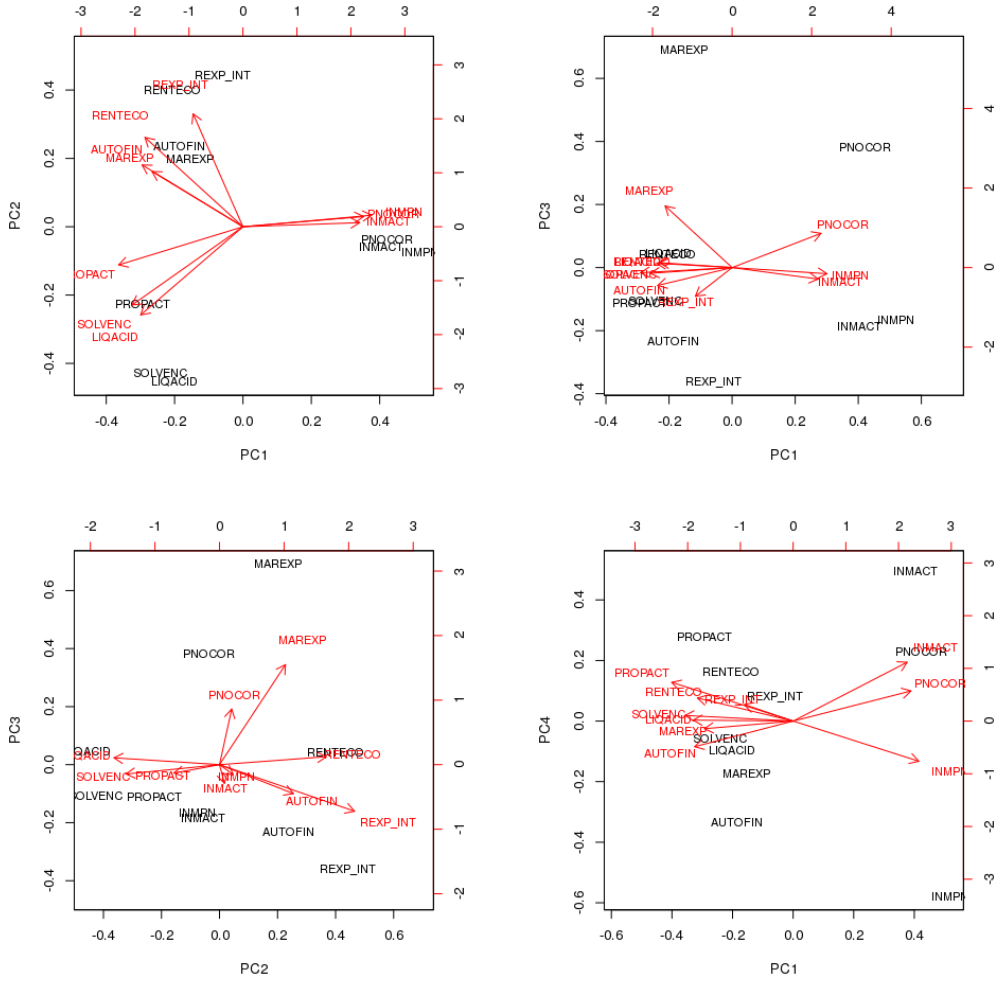

Figure 4: Gráfico de las variables en función de las componentes principales.

factor "agrupa" variables que caen bajo la categoría de necesidades y las contrapone a otras que claramente representan un tipo de ocupación: estudiante, jubilado, etc. Por otra parte el segundo factor, que toma valores menos extremos, parece contrastar combinaciones de estas categorías, y, por ejemplo, personas con alguna ocupación y aquellas que no la tienen. Podemos ver las proyecciones de los datos en función de los dos factores en la figura 7.

2. **Dos factores con rotación.** Para facilitar la interpretación de los factores aplicamos una transformación ortogonal sobre la matriz de carga, es decir, una rotación, que maximice la varianza de los coeficientes que ponderan cada factor en las variables. Obtenemos la matriz que se muestra en la tabla 4. Observar que los valores son apenas más extremos, y que confirman las conclusiones a las que arribamos en el incsiso anterior. En

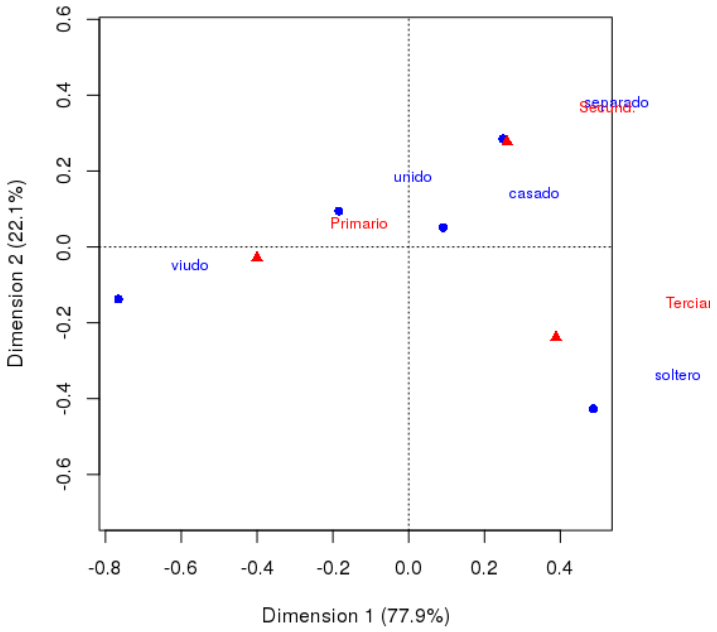

Figure 5: Gráfico de correspondencia simple entre estado civil y educación.

esta configuración tanto como en la anterior podríamos hablar de factores que tienen que ver con el nivel de ocupación y el nivel de necesidades insatisfechas y/o privaciones. En la figura 8 podemos ver el gráfico de las observaciones en función de los factores. Observar que es casi idéntico al gráfico del inciso anterior, excepto por una leve rotación.

3. **Tres factores (con rotación variamax).** Si observamos la tabla 5 veremos que nuevamente el primer factor tiende a agrupar y contraponer niveles de ocupación y educación, con niveles de necesidades insatisfechas. Este es un común denominador en nuestros análisis. El segundo factor vuelve a darnos una noción de personas ocupadas vs. personas no ocupadas, y el tercero pareciera contrastar variables que tienen que ver con la población pasiva, (jubilados, la edad) con las demás, especialmente el porcentaje de personas con una ocupación. En la figura 9 se observan las proyecciones de las observaciones en función de los dos primeros factores.

3.2 Determinación del número máximo de factores

Recordar que asumimos que la matriz de varianzas V tiene la propiedad siguiente:

$$V = \Lambda\Lambda' + \psi \tag{1}$$

en donde Λ es la matriz de carga y, por lo tanto, $\Lambda\Lambda'$ la matriz de varianzas

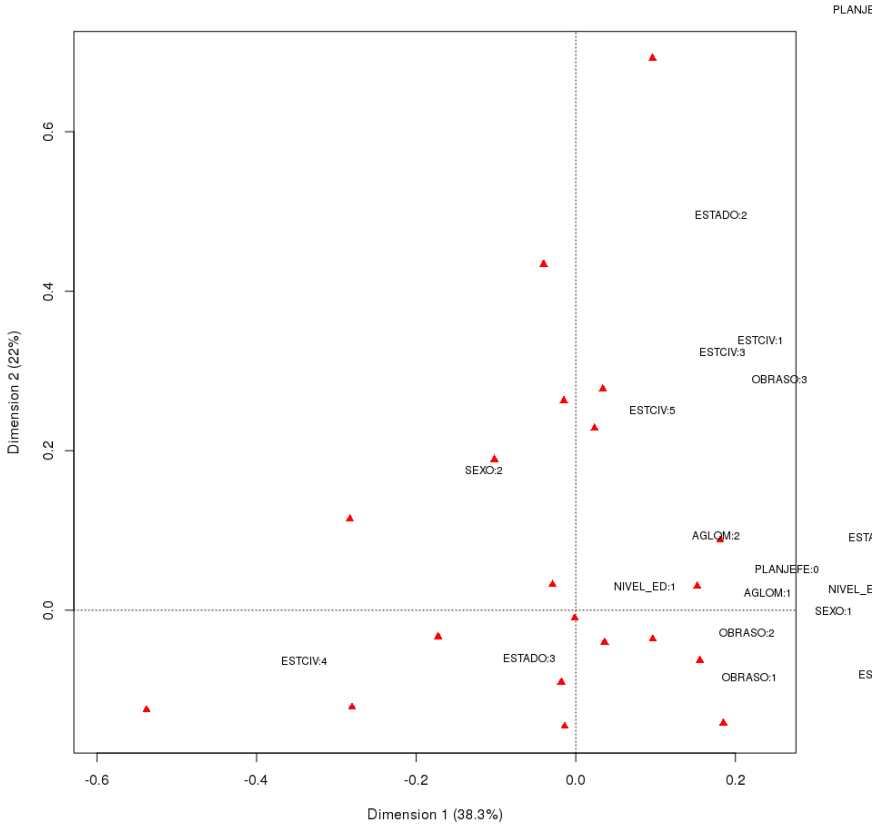

Figure 6: Gráfico de correspondencia múltiple.

asociada a los factores, y ψ una matriz diagonal que representa la variabilidad específica de cada variable. Esto quiere decir que V puede descomponerse como la suma de dos matrices de varianza, cada una de las cuales representa una componente de variabilidad en el modelo. Ahora bien, dado que no conocemos V, la reemplazamos por S en la ecuación (1). Nos queda:

$$S = \Lambda\Lambda' + \psi \tag{2}$$

S es una matriz diagonal $p \times p$ y por lo tanto tiene $p(p+1)/2$ elementos distintos, que definen el mismo número de ecuaciones. En el lado derecho de la ecuación (2) nos queda determinado un número de incógnitas que depende de la cantidad de factores. Sea m tal cantidad. La matriz Λ es $p \times m$ y por lo tanto tiene pm elementos distintos, i.e., incógnitas. Por otro lado, ψ es una matriz diagonal $p \times p$, por lo que suma p incógnitas más. Esto resulta en $pm + p$ incógnitas. Sin embargo, a los fines de identificar Λ de manera única debemos imponer restricciones sobre ella, por ejemplo, requerir

$$\Lambda'\psi^{-1}\Lambda = D$$

en donde D es diagonal $m \times m$. Esta condición restringe el modelo en

	Factor1	Factor2
X.ocupados	-0.46	-0.42
X.desoc	0.13	0.54
X.jubil	-0.36	-0.14
X.estud	-0.28	0.55
X.NBIhog	0.90	-0.06
X.privcte	-0.05	0.68
X.privpat	0.58	-0.45
X.privconv	0.93	0.01
edad	-0.67	-0.38
educ	-0.84	0.17
hijos	0.89	0.12
X.propietarios	-0.05	-0.07

Table 3: Matriz de carga para la configuración 1 de la sección 3.1

	Factor1	Factor2
X.ocupados	-0.47	-0.41
X.desoc	0.15	0.54
X.jubil	-0.36	-0.13
X.estud	-0.26	0.56
X.NBIhog	0.90	-0.08
X.privcte	-0.03	0.68
X.privpat	0.57	-0.46
X.privconv	0.93	-0.01
edad	-0.68	-0.36
educ	-0.83	0.20
hijos	0.89	0.10
X.propietarios	-0.05	-0.07

Table 4: Matriz de carga de la configuración 2 de 3.1

$m(m-1)/2$ elementos, que es la cantidad de términos que hay debajo (y por encima) de la diagonal de una matriz cuadrada $m \times m$ (al ser diagonal la matriz es simétrica). En otras palabras, tenemos $m(m-1)/2$ incógnitas menos. Por lo tanto la cantidad de incógnitas en el lado izquierdo de la ecuación (2) es $pm + p - m(m-1)/2$. Para que el sistema esté determinado debe haber un mayor número de ecuaciones que de incógnitas. Nos queda:

$$\frac{p(p+1)}{2} \geq pm + p - \frac{m(m-1)}{2} \tag{3}$$

Reagrupando obtenemos:

$$(p-m)^2 \geq p + m \tag{4}$$

Luego, m deberá ser tal que la ecuación (4) se cumple.

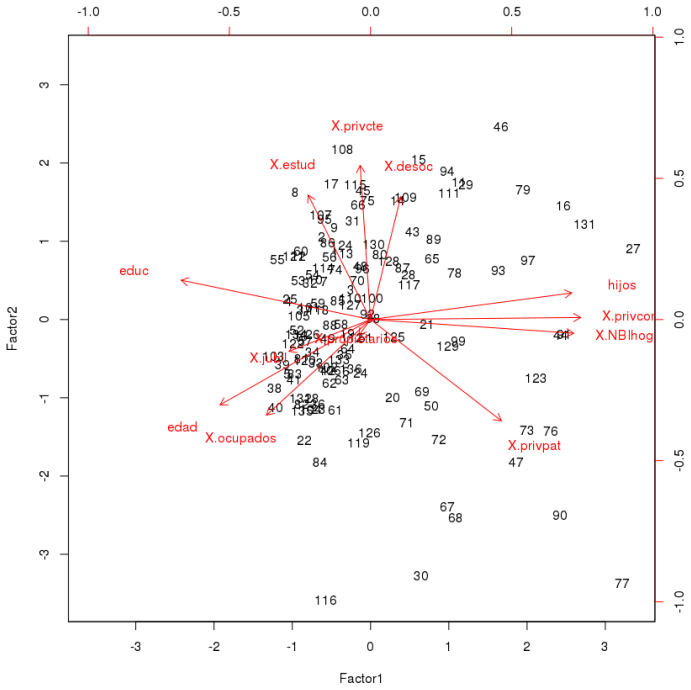

Figure 7: Observaciones en función de los dos factores para la configuración 1 de 3.1.

	Factor1	Factor2	Factor3
X.ocupados	-0.56	-0.74	-0.36
X.desoc	0.06	0.55	-0.15
X.jubil	-0.15	0.06	0.78
X.estud	-0.33	0.47	-0.03
X.NBIhog	0.89	0.03	-0.13
X.privcte	-0.14	0.76	-0.08
X.privpat	0.63	-0.29	0.04
X.privconv	0.91	0.09	-0.27
edad	-0.45	-0.32	0.83
educ	-0.80	0.14	0.25
hijos	0.83	0.16	-0.27
X.propietarios	-0.03	-0.14	0.16

Table 5: Matriz de carga para tres factores con rotación varimax.

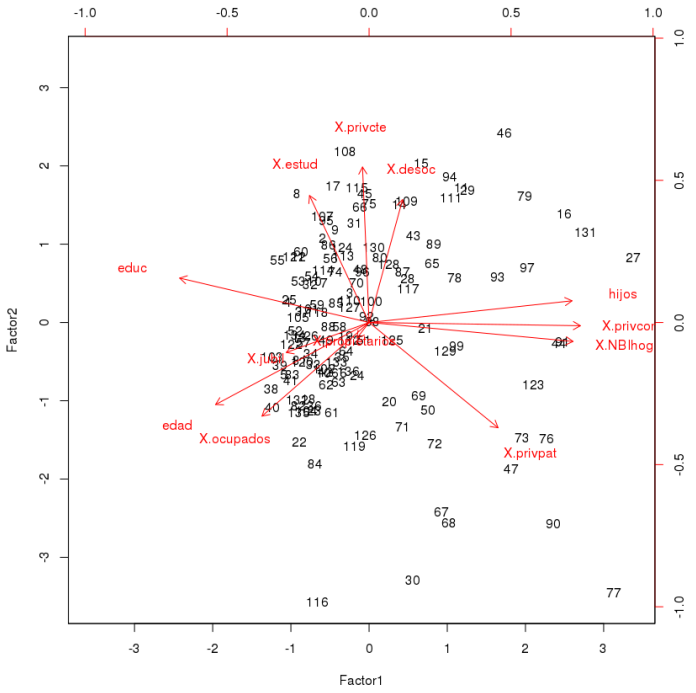

Figure 8: Observaciones en función de los dos factores para la configuración 2 de 3.1.

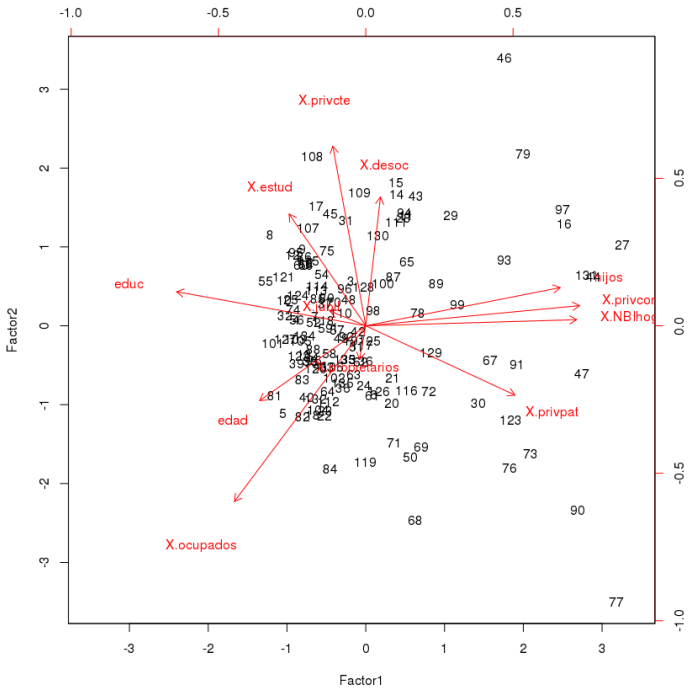

Figure 9: Observaciones en función de los dos primeros factores para la configuración 3 de 3.1.