

Análisis multivariante

Luis Cayuela

18 de noviembre de 2009

EcoLab, Centro Andaluz de Medio Ambiente, Universidad de Granada – Junta de Andalucía, Avenida del Mediterráneo s/n, E-18006, Granada. E-mail: lcayuela@ugr.es.

Índice

1. Introducción	3
2. Análisis de componentes principales (PCA)	3
2.1. Ejemplo: Modelando la riqueza de plantas exóticas en Reino Unido a partir del clima	4
3. Análisis de la varianza multivariado (MANOVA)	12
3.1. Ejemplo: ¿Qué variables determinan la composición florística en bosques tropicales montanos?	13
4. Escalamiento multidimensional no métrico (NMDS)	15
4.1. Ejemplo: Gradientes de composición florística en bosques tropi- cales montanos	16
5. Análisis de correspondencias canónico (CCA)	22
5.1. Ejemplo: ¿Cómo se relaciona la estructura de comunidades de plantas con las variables ambientales?	23
6. Ejercicios	25

1. Introducción

En un sentido amplio, el análisis multivariante hace referencia a cualquier método estadístico que analice simultáneamente múltiples características en cada uno de los individuos o muestras objeto de la investigación. Una de las dificultades en definir qué es el análisis multivariante reside en el hecho de que el término multivariante (o multivariado) no ha sido usado de manera consistente en la literatura. Algunos investigadores usan el término multivariado simplemente para referirse a las relaciones existentes entre más de dos variables. Sin embargo, para que un análisis sea considerado verdaderamente multivariante, todas las variables deben de ser aleatorias y deben de estar interrelacionadas de tal manera que los diferentes efectos no puedan ser interpretados significativamente de manera independiente. Por ejemplo, si queremos ver el efecto de una variable ambiental sobre las diferentes especies de peces que hay en un río, tiene sentido considerar todas las abundancias de cada una de las especies en su conjunto y no la abundancia de cada una de las especies por separado, ya que las diferentes especies se interrelacionan entre sí por medio de interacciones bióticas (competencia por recursos, predación, etc) y es difícil de separar estos efectos de los efectos puramente ambientales.

Podemos considerar como técnicas multivariantes, entre otras:

- Análisis de componentes principales
- Análisis discriminante
- Análisis cluster (técnica de agrupación)
- Análisis de correspondencias
- Escalamiento multidimensional
- Análisis de correspondencias canónico
- Modelo de ecuaciones estructurales (análisis causal)
- Análisis de la varianza multivariado (incluyendo la regresión multivariada)

En esta sesión veremos algunas de ellas, prestando especial atención al análisis de comunidades biológicas.

2. Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en PCA es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá, pues, que estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

Fases de un análisis de componentes principales:

1. Análisis de la matriz de correlaciones. Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.
2. Selección de los factores. La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. A éstos se les denominará componentes principales. Esta decisión puede ser más o menos arbitraria (p.e. que capturen el 80 % de la variabilidad de los datos) o estar basada en criterios estadísticos. El paquete `nFactors` ofrece una serie de funciones para la selección de factores (ver <http://www.statmethods.net/advstats/factor.html>).
3. Análisis de la matriz factorial. Una vez seleccionados los componentes principales, se representan en forma de matriz. Cada elemento de ésta representa los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales). La matriz tendrá tantas columnas como componentes principales y tantas filas como variables.
4. Interpretación de los factores. Para que un factor sea fácilmente interpretable debe tener las siguientes características, que son difíciles de conseguir:
 - Los coeficientes factoriales deben ser próximos a 1.
 - Una variable debe tener coeficientes elevados sólo con un factor.
 - No deben existir factores con coeficientes similares.
5. Cálculo de las puntuaciones factoriales. Son las puntuaciones que tienen los componentes principales para cada caso, que nos permitirán su análisis posterior y su representación gráfica.

2.1. Ejemplo: Modelando la riqueza de plantas exóticas en Reino Unido a partir del clima¹

En este ejemplo queremos modelar la riqueza de especies exóticas en el Reino Unido utilizando variables climáticas. Para ello se ha dividido todo el Reino

¹Datos cedidos por Fabio Suzart, Universidad de Alcalá. Estos datos no pueden ser usados para otros fines que no sean docentes sin permiso del autor.

Unido en celdas de 10 x 10 kms y se han utilizado los registros de colecciones botánicas para contar el número de especies exóticas. Las variables climáticas se han extraído del WorldClim (<http://www.worldclim.org/>).

Los datos están accesibles en la siguiente dirección

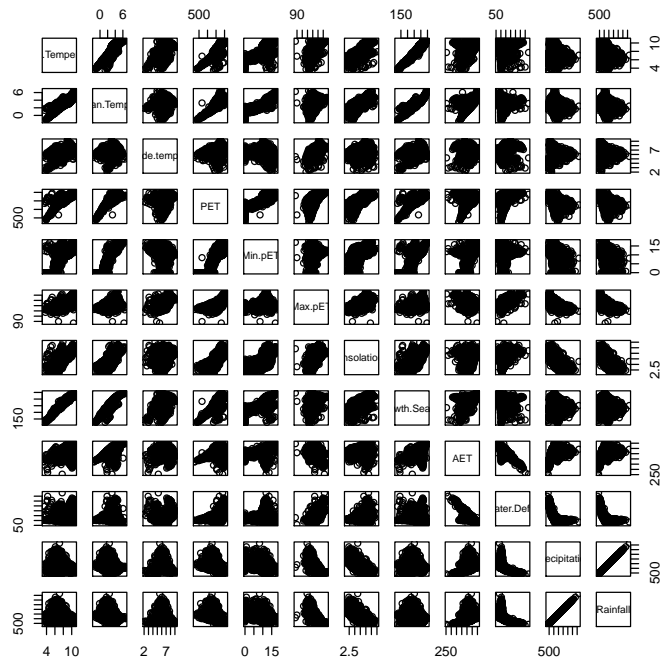
<http://tinyurl.com/yan3b9j>. Vamos a leer los datos directamente de la dirección web con la función `url()`.

```
> clima <- read.table(url("http://tinyurl.com/yan3b9j"), header = T,
+      sep = "\t")
> str(clima)
```

```
'data.frame':      2243 obs. of  13 variables:
 $ Alien          : int  23 32 25 46 35 89 38 46 40 4 ...
 $ Mean.Temperature : num  6.86 7.39 5.3 7.71 7.39 ...
 $ Mean.Jan.Temperature: num  3.27 3.46 2.29 3.31 2.91 ...
 $ Rango.de.temperatura: num  4.84 6 3.98 6.46 6.53 ...
 $ PET            : num  518 600 592 607 601 ...
 $ Min.pET        : num  8.44 13.89 12.98 12.7 11.82 ...
 $ Max.pET        : num  89.9 101.8 101.5 105.4 105.5 ...
 $ Insolation      : num  2.79 2.8 3.04 3.28 3.2 ...
 $ Growth.Season   : num  282 291 205 275 263 ...
 $ AET            : num  459 484 434 459 451 ...
 $ Water.Defcit    : num  58.4 115.6 158 148.8 150.4 ...
 $ Precipitation    : num  1392 1605 855 959 958 ...
 $ Rainfall        : num  1392 1605 855 959 958 ...
```

La primera variable sería la variable respuesta en nuestro modelo y el resto de variables serían variables explicativas. Sin embargo, al ser todas las variables explicativas variables climáticas es muy posible que haya mucha colinealidad (es decir, correlación entre variables), lo que haría cualquier modelo estadístico basado en dichas variables muy inestable. Vamos a ver si realmente existe correlación entre las variables explicativas con la función `cor()` y/o `pairs()`.

```
> pairs(clima[, -1])
```



Así que vemos que realmente existe mucha correlación entre las variables explicativas. Una solución a este problema sería utilizar análisis de componentes principales para reducir la dimensionalidad de los datos y luego utilizar los factores principales que nos resumen los datos para modelar la riqueza de especies exóticas. Para ello podemos utilizar varias funciones, como `prcomp()`, `princomp()` o `factanal()`. El paquete `psych` tiene otras funciones relacionadas con el análisis de componentes principales como los PCA jerárquicos.

```
> pca1 <- prcomp(clima[, -1], scale = T)
> summary(pca1)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.644	1.718	0.9815	0.7172	0.4772	0.4000	0.3572	0.20319
Proportion of Variance	0.582	0.246	0.0803	0.0429	0.0190	0.0133	0.0106	0.00344
Cumulative Proportion	0.582	0.828	0.9088	0.9516	0.9706	0.9839	0.9946	0.99800
	PC9	PC10	PC11	PC12				
Standard deviation	0.12922	0.08389	0.01386	0.00734				
Proportion of Variance	0.00139	0.00059	0.00002	0.00000				
Cumulative Proportion	0.99939	0.99998	1.00000	1.00000				

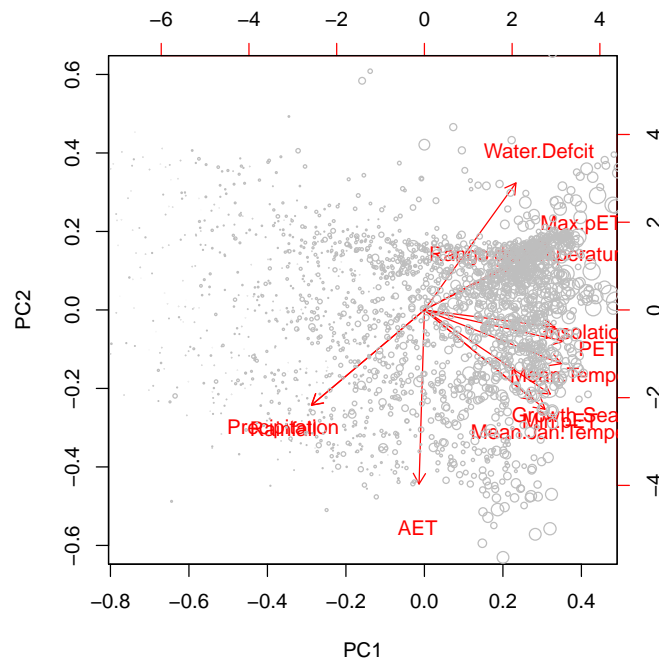
Como podemos ver, los dos primeros factores recogen cerca del 83 % de la variabilidad de las variables climáticas utilizadas. Tomaremos estos dos componentes para representar la variabilidad en el clima. Ahora es importante interpretar qué significan estos componentes principales. Para ello podemos utilizar la matriz de correlación de las variables climáticas con los factores.

```
> pca1$rotation[, 1:2]
```

	PC1	PC2
Mean.Temperature	0.34852153	-0.16985773
Mean.Jan.Temperature	0.30722684	-0.31362840
Rango.de.temperatura	0.21576733	0.13711343
PET	0.35433070	-0.09847938
Min.pET	0.27654607	-0.28400149
Max.pET	0.31976844	0.21453683
Insolation	0.33246966	-0.05442222
Growth.Season	0.32063663	-0.26539819
AET	-0.01362093	-0.54991427
Water.Defcit	0.23318824	0.40121923
Precipitation	-0.28774698	-0.30042944
Rainfall	-0.28741001	-0.30094185

También es conveniente dibujar los componentes seleccionados del PCA en un gráfico. Para ello utilizaremos la función `biplot()`.

```
> biplot(pca1, cex = c(0.01, 1), scale = 0.5, ylim = c(-0.6, 0.6))
> points(x = pca1$x[, 1], y = pca1$x[, 2], cex = clima[, 1]/300,
+       col = "grey")
```



Lo que hemos hecho ha sido, por un lado, representar la relación de las variables climáticas con los dos primeros componentes del PCA. Pero además, hemos representado en este gráfico cada una de las celdas de 10 x 10 km con un tamaño (cex) que es proporcional a su riqueza de especies exóticas. De esta manera podemos interpretar el significado de los ejes y empezar a vislumbrar si existe alguna relación entre estos ejes y nuestra variable respuesta. Tanto el gráfico como las correlaciones de las variables con los ejes parecen apuntar a que el primer componente está relacionado con la temperatura (Mean.Temperature, Mean.Jan.Temperature), la evapotranspiración potencial (PET, Max.PET) y la duración de la estación de crecimiento (Growth.Season), mientras que el segundo componente está relacionado fundamentalmente con la evapotranspiración real (AET) y el déficit hídrico (Water.Deficit). Por tanto podríamos decir que el primer componente está vinculado a la entrada de energía en el sistema y el segundo al déficit hídrico (ya que ésta y la AET están correlacionadas negativamente). Además, vemos que la riqueza de especies nativas parece estar asociada positivamente con el eje 1 (entrada de energía en el sistema).

Vamos a ajustar ahora el modelo estadístico para explicar la riqueza de especies nativas está realmente explicada por estas dos nuevas variables.

```
> lm.exoticas <- lm(clima$Alien ~ pca1$x[, 1:2])
> summary(lm.exoticas)
```

Call:

```
lm(formula = clima$Alien ~ pca1$x[, 1:2])
```

Residuals:

	Min	1Q	Median	3Q	Max
	-144.568	-43.123	-7.342	32.691	365.614

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	155.9568	1.3266	117.562	<2e-16 ***
pca1\$x[, 1:2]PC1	29.8974	0.5019	59.567	<2e-16 ***
pca1\$x[, 1:2]PC2	-0.3346	0.7722	-0.433	0.665

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

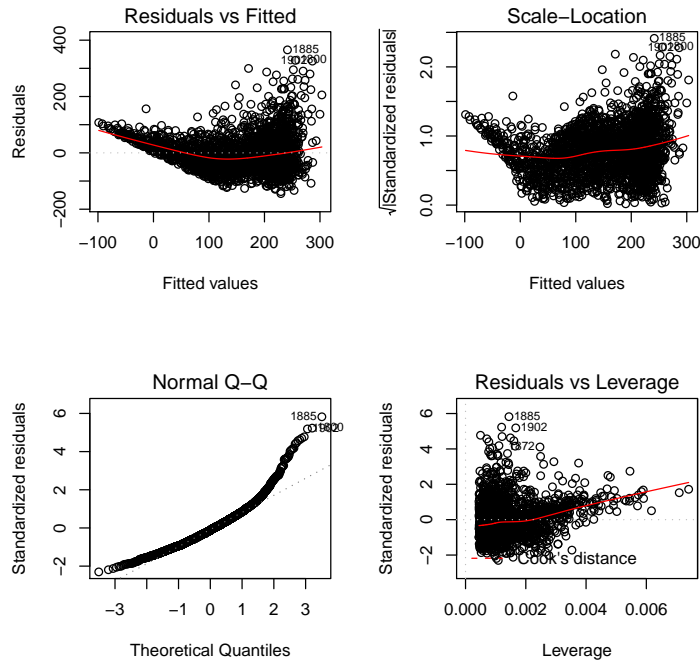
Residual standard error: 62.83 on 2240 degrees of freedom

Multiple R-squared: 0.613, Adjusted R-squared: 0.6127

F-statistic: 1774 on 2 and 2240 DF, p-value: < 2.2e-16

Vemos que la primera variable es significativa y positiva y que el modelo explica cerca del 60 % de la variabilidad de la riqueza de exóticas. Vamos a revisar los residuos del modelo.

```
> par(mfcol = c(2, 2))
> plot(lm.exoticas)
```



No parece que el modelo sea muy idóneo. Es claramente homocedástico y no lineal. Además, tengamos en cuenta que la variable respuesta es un conteo y, por tanto, predicciones que no sean enteros o con valores por debajo de 0 (que son posibles asumiendo una distribución de errores normal) no tienen sentido. Probemos un modelo Poisson.

```
> glm.exoticas <- glm(clima$Alien ~ pca1$x[, 1:2], family = poisson)
> summary(glm.exoticas)
```

Call:

```
glm(formula = clima$Alien ~ pca1$x[, 1:2], family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.8684	-3.4042	-0.6294	2.4196	20.9990

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	4.8574493	0.0020704	2346.14	<2e-16 ***
pca1\$x[, 1:2]PC1	0.2581541	0.0008915	289.56	<2e-16 ***
pca1\$x[, 1:2]PC2	-0.0303083	0.0010245	-29.59	<2e-16 ***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

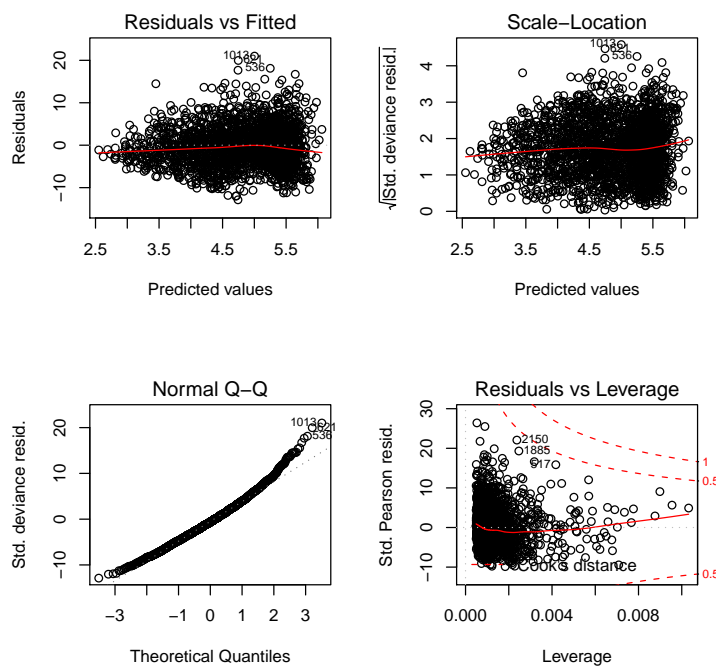
```
(Dispersion parameter for poisson family taken to be 1)
```

```
Null deviance: 153463 on 2242 degrees of freedom
Residual deviance: 45873 on 2240 degrees of freedom
AIC: 60635
```

```
Number of Fisher Scoring iterations: 4
```

Ahora las dos variables son significativas. La primera, relacionada con la entrada de energía en el sistema, se relaciona positivamente con la riqueza de nativas. Y la segunda, que es una indicadora del déficit hídrico, lo está negativamente. Así que a mayor déficit hídrico, menor riqueza de especies exóticas. Vamos a ver si esta vez los residuos son adecuados.

```
> par(mfcol = c(2, 2))
> plot(glm.exoticas)
```



Si vemos los residuos observaremos que el modelo, aunque no es perfecto, es bastante más adecuado que el modelo normal.

3. Análisis de la varianza multivariado (MANOVA)

El Análisis de la Varianza Multivariante (MANOVA) es una extensión del análisis de la varianza (ANOVA) que permite cubrir los casos donde hay más de una variable dependiente que no pueden ser combinadas de manera simple. Por tanto, frente al ANOVA o la regresión, en donde tendríamos la siguiente formulación del modelo:

$$y \sim x_1 + x_2 + \dots + x_n$$

en el MANOVA el modelo quedaría formulado de la siguiente forma:

$$y_1 + y_2 + \dots + y_k \sim x_1 + x_2 + \dots + x_n$$

Por lo general, se ha aceptado la terminología de MANOVA para referirse a análisis que contemplan varias variables respuesta continuas, pero sin prestar mucha atención a si las variables explicativas son continuas o discretas. En un sentido estricto, si las variables explicativas fueran continuas tendríamos una regresión múltiple multivariante, si fueran discretas estaríamos ante un caso de análisis de la varianza multifactorial multivariante, y si fueran de ambos tipos el análisis sería del tipo ANCOVA multivariante. Sin embargo, es muy común referirse a cualquiera de ellos como MANOVA, y está será la terminología usada aquí. El MANOVA, al igual que los modelos lineales, se basa en una serie de supuestos:

- las muestras son independientes entre sí;
- cada variable tiene una distribución normal;
- en conjunto las k variables dependientes tienen la distribución normal conjunta;
- las varianzas de cada variable son iguales al compararlas de tratamiento a tratamiento;
- las correlaciones entre dos variables de un mismo grupo son las mismas de grupo a grupo.

Estos supuestos son muchas veces difíciles de cumplir. Por ello, una alternativa eficiente al MANOVA es el MANOVA semi-paramétrico, que utiliza las distancias entre cada par de observaciones para obtener una matriz de distancia sobre la que luego se calcula la significación de las variables explicativas con simulaciones de Monte Carlo. Este tipo de enfoque es muy similar al del escalamiento multidimensional no métrico (NMDS), en tanto que la partición de la varianza se hace utilizando una matriz de distancias, por lo que ambos métodos se complementan bastante bien.

Hay que considerar que la interpretación de un MANOVA (ya sea paramétrico o semi-paramétrico) es bastante más compleja que la de un ANOVA o una

regresión. Por medio de este análisis sólo es posible saber si la(s) variables explicativa(s) tienen un efecto sobre el conjunto de las variables respuesta, pero difícilmente sabremos cómo es este efecto a no ser que utilicemos otras técnicas complementarias como el NMDS. Por tanto, al realizar un análisis de este tipo nos fijaremos en la significación de los coeficientes y, cuando sea posible, en la variabilidad explicada por cada una de las variables explicativas.

En R hay, por lo menos, dos funciones que nos permiten ajustar un MANOVA. La función `manova()` se encuentra dentro del paquete `stats` y ajusta MANOVAs paramétricos, por lo que es importante evaluar la idoneidad del modelo mirando los residuos. La función `adonis()`, dentro del paquete `vegan`, permite ajustar MANOVAs semi-paramétricos, por lo que la evaluación de los residuos del modelo no es necesaria. Nos centraremos en esta última para el análisis de comunidades biológicas.

3.1. Ejemplo: ¿Qué variables determinan la composición florística en bosques tropicales montanos?²

Se quiere investigar qué variables ambientales afectan la composición florística de árboles en parcelas de 0.1 hectáreas muestreadas en distintos tipos de bosques tropicales en los Altos de Chiapas, México (bosque de pino-encino (POF), bosque de encino (OF), bosque de pino (PF), bosque nublado (MCF) y bosque transicional a selva baja caducifolia (TF)). El tipo de bosque es el resultado de factores ambientales (clima) y el uso humano.

Para este caso de estudio se han seleccionado las 86 especies más abundantes sobre un total de 231 en 204 parcelas de 0.1 hectáreas. Para cada especie tenemos su abundancia total en cada parcela. Queremos construir un modelo en dónde la composición de árboles quede en función por un lado del tipo de bosque y, por otro, de la productividad (medida a partir del índice de vegetación NDVI obtenido de una imagen Landsat del año 2000) y la elevación.

La matriz de parcelas (filas) x especies (columnas) está disponible en la siguiente dirección <http://tinyurl.com/yalcxwx>. Las variables ambientales para las parcelas muestreadas (tipo de bosque, productividad, elevación) están disponibles en la siguiente dirección <http://tinyurl.com/yjn7ahv>.

Vamos primero a cargar la matriz de parcelas x especies y los datos ambientales en R.

```
> bio <- read.table(url("http://tinyurl.com/yalcxwx"), header = T,
+   sep = "\t")
> env <- read.table(url("http://tinyurl.com/yjn7ahv"), header = T,
+   sep = "\t")
```

Ahora vamos a ajustar un MANOVA en dónde la composición de especies (bio) va a estar en función de las variables que hay en el arreglo de datos env (Forest type, Productivity, Elevation).

²Cayuela, L., Golicher, D.J., Rey Benayas, J.M., González-Espinosa, M. & Ramírez-Marcial, N. 2006. Fragmentation, disturbance and tree diversity conservation in tropical montane forests. *Journal of Applied Ecology* 43: 1172-1181

```
> library(vegan)
> attach(env)
> manova1 <- adonis(bio ~ Forest.type + Productivity + Elevation)
> manova1
```

Call:

```
adonis(formula = bio ~ Forest.type + Productivity + Elevation)
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Forest.type	4.00000	14.48614	3.62153	12.32310	0.1838	0.001 ***
Productivity	1.00000	1.01881	1.01881	3.46673	0.0129	0.001 ***
Elevation	1.00000	5.40130	5.40130	18.37914	0.0685	0.001 ***
Residuals	197.00000	57.89472	0.29388		0.7347	
Total	203.00000	78.80096			1.0000	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Los resultados muestran que todas las variables son significativas. Las sumas de cuadrados (SumsOfSqs) nos dicen qué cantidad de variabilidad está explicada por cada una de las variables y la variabilidad residual (esto es, no explicada por el modelo). En este ejemplo podemos ver que la composición de árboles en bosques tropicales montanos está explicada fundamentalmente por el tipo de bosque ($14.49/78.80 = 18\%$), pero también por la productividad ($1.018/78.80 = 1\%$) y la elevación ($5.40/78.80 = 7\%$). Es decir, que dependiendo del tipo de bosque vamos a encontrar distintas especies. Pero además existe un gradiente altitudinal que condiciona en parte la composición de estos bosques. Podría ser interesante explorar si este gradiente altitudinal afecta de manera distinta a los distintos tipos de bosque. Para ello vamos a incluir la interacción entre estas variables en un nuevo modelo.

```
> manova2 <- adonis(bio ~ Forest.type + Productivity + Elevation +
+   Forest.type:Elevation)
> manova2
```

Call:

```
adonis(formula = bio ~ Forest.type + Productivity + Elevation + Forest.type:Elevation)
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Forest.type	4.00000	14.48614	3.62153	13.04712	0.1838	0.001 ***
Productivity	1.00000	1.01881	1.01881	3.67041	0.0129	0.001 ***
Elevation	1.00000	5.40130	5.40130	19.45898	0.0685	0.001 ***
Forest.type:Elevation	4.00000	4.32304	1.08076	3.89360	0.0549	0.001 ***
Residuals	193.00000	53.57168	0.27757		0.6798	
Total	203.00000	78.80096			1.0000	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Y vemos que, efectivamente, el cambio en la composición de especies a lo largo del gradiente altitudinal va a ser distinto según el tipo de bosque (y explica

cerca de un 5 % de la variabilidad en la composición de especies). Esto podría indicar, por ejemplo, que algunos tipos de bosque no van a sufrir ningún cambio en la composición de especies a lo largo del gradiente altitudinal y otros sí. Sin embargo, no es posible conocer el sentido de esta interacción a partir únicamente de los resultados de este análisis. Podríamos hacer MANOVAS individuales para cada uno de los tipos de bosque o podríamos utilizar otras técnicas multivariantes que nos van a ayudar a interpretar estos resultados visualmente, como veremos en la siguiente sección.

4. Escalamiento multidimensional no métrico (NMDS)

El escalamiento multidimensional no métrico (NMS, MDS, NMDS o NMDS) es una técnica multivariante de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones las proximidades existentes entre un conjunto de objetos. El NMDS es un método de ordenación adecuado para datos que no son normales o que están en una escala discontinua o arbitraria. Una ventaja del NMDS frente a otras técnicas de ordenación es que, al estar basada en rangos de distancias, tiende a linealizar la relación entre las distancias ambientales y las distancias biológicas (esto es, calculadas a partir de una matriz de sitios x especies). Una de las desventajas de esta técnica es la dificultad para alcanzar una solución estable única. A pesar de ello, el NMDS es una técnica ampliamente utilizada en ecología para detectar gradientes en comunidades biológicas.

El NMDS se implementa de la siguiente forma:

1. Se calcula la matriz de disimilaridad X a partir de la matriz de datos de sitios x especies. Esta matriz nos indica cómo de iguales son cada par de sitios utilizando para ello la similaridad entre sus especies. Supongamos que tenemos tres especies (sp1, sp2, sp3) y tres sitios (A, B, C). El sitio A tiene sp1 = 3, sp2 = 0 y sp3 = 8. El sitio B tiene sp1 = 3, sp2 = 0 y sp3 = 6. El sitio C tiene sp1 = 0, sp2 = 5 y sp3 = 1. Por tanto, podemos calcular una matriz de disimilaridad que nos indique con números que los sitios A y B son muy iguales, mientras que los sitios A y C y B y C son muy distintos entre sí. Cuando se trata de datos biológicos la distancia más usada es la distancia de Sorensen (Bray-Curtis) en vez de la distancia Euclídea.
2. Se asignan los sitios (unidades muestrales) a una configuración inicial aleatoria en un espacio k -dimensional (dónde k es el número de especies), aunque en realidad, la ordenación se va a realizar principalmente sobre unas pocas dimensiones (2 o 3).
3. Se calculan las distancias sobre este nuevo espacio geométrico y se calcula una matriz de distancia Y .
4. Se comparan las matrices de distancia X e Y y se mide cómo son de parecidas entre ellas (stress).

5. A partir de la configuración inicial, se reasignan los sitios (unidades muestrales) para reducir las distancias con la matriz X .
6. Se repite este proceso de manera iterativa hasta que se consigue una solución óptima en donde la matriz de distancias Y es muy parecida a la matriz de distancias X . Esto es, se minimiza el **stress**.

La ventaja del NMDS es que nos permite, al igual que el PCA, reducir la dimensionalidad de nuestros datos originales. El resultado de la ordenación se puede visualizar en un gráfico de ordenación. Posteriormente podemos relacionar los ejes resultantes de dicha ordenación con distintas variables ambientales para determinar de manera indirecta el efecto de éstas sobre la matriz de sitios x especies.

Aunque en ecología se utiliza típicamente esta técnica para analizar datos de comunidades biológicas (matriz de sitios x especies) también se puede aplicar a otro tipo de datos, como por ejemplo múltiples variables físico-químicas medidas en distintos cuerpos de agua (ríos, embalses, pantanos). Esta técnica se utiliza también mucho en otras disciplinas, como la psicología o la economía. En R tenemos una implementación de esta función (**metaMDS**) en el paquete **vegan**.

4.1. Ejemplo: Gradientes de composición florística en bosques tropicales montanos³

Al igual que en ejemplo anterior, se quiere investigar qué variables ambientales afectan la composición florística de árboles en parcelas de 0.1 hectáreas muestreadas en distintos tipos de bosques tropicales en los Altos de Chiapas, México. El tipo de bosque es el resultado de factores ambientales (clima) y el uso humano. Se seleccionaron las 86 especies más abundantes sobre un total de 231 en 204 parcelas de 0.1 hectáreas. Para cada especie tenemos su abundancia total en cada parcela. Queremos construir un modelo en donde la composición de árboles quede en función por un lado del tipo de bosque y, por otro, de la productividad (medida a partir del índice de vegetación NDVI obtenido de una imagen Landsat del año 2000) y la elevación.

Los objetivos concretos son:

1. Explorar visualmente cómo son de similares o distintas las parcelas muestreadas en función de las especies que contienen.
2. Investigar la relación entre esta ordenación y las variables ambientales por medio de correlaciones de dichas variables con los ejes de ordenación y el ajuste de superficies de tendencia.

La matriz de parcelas (filas) x especies (columnas) está disponible en <http://tinyurl.com/yalcxwx>. Las variables ambientales para las parcelas

³Cayuela, L., Golicher, D.J., Rey Benayas, J.M., González-Espinosa, M. & Ramírez-Marcial, N. 2006. Fragmentation, disturbance and tree diversity conservation in tropical montane forests. *Journal of Applied Ecology* 43: 1172-1181

muestreadas (tipo de bosque, productividad, elevación) están disponibles en <http://tinyurl.com/yjn7ahv>.

Al igual que en el caso anterior es necesario cargar la matriz de parcelas x especies y los datos ambientales en R. Si se ha realizado el ejercicio anterior en esta misma sesión se puede saltar este paso.

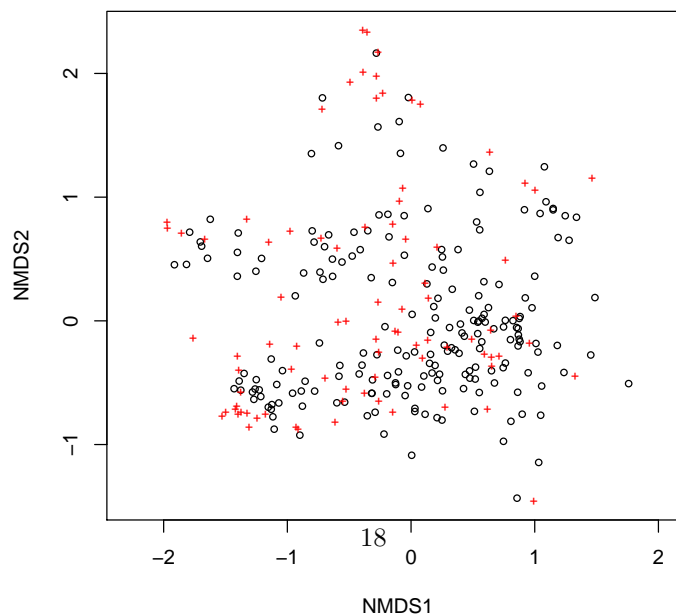
```
> bio <- read.table(url("http://tinyurl.com/yalcxwx"), header = T,  
+   sep = "\t")  
> env <- read.table(url("http://tinyurl.com/yjn7ahv"), header = T,  
+   sep = "\t")
```

Vamos ahora a realizar el escalamiento multidimensional no métrico. Como la configuración inicial de las parcelas es aleatoria, cada vez que realicemos el NMDS vamos a tener un resultado ligeramente distinto. Para evitar esto vamos a utilizar el comando `set.seed()` que genera unos datos “semilla” a partir de los cuales se establece la configuración inicial de las parcelas en los ejes del NMDS. De esta manera, cada vez que realicemos el análisis obtendremos el mismo resultado.

```
> set.seed(0)
> nmds1 <- metaMDS(bio)

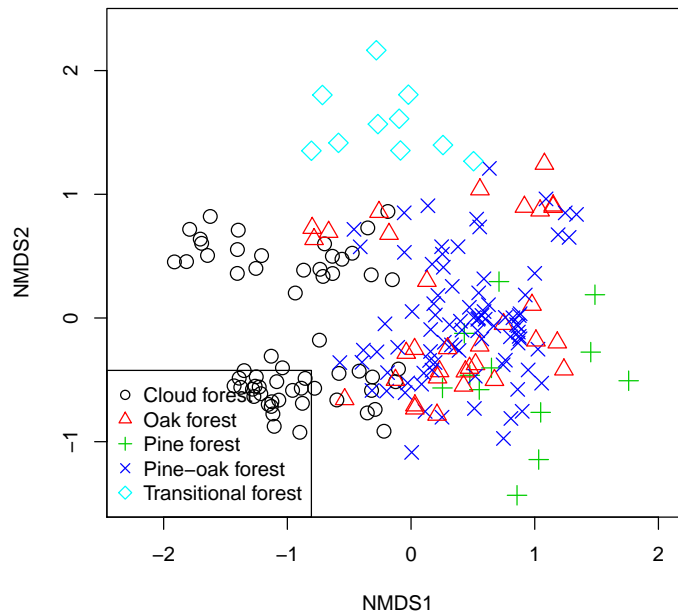
Square root transformation
Wisconsin double standardization
Using step-across dissimilarities:
Too long or NA distances: 3643 out of 20706 (17.6%)
Stepping across 20706 dissimilarities...
Run 0 stress 20.58713
Run 1 stress 21.49227
Run 2 stress 22.13124
Run 3 stress 22.2231
Run 4 stress 24.14967
Run 5 stress 21.73649
Run 6 stress 20.77451
Run 7 stress 23.69372
Run 8 stress 20.98569
Run 9 stress 22.35428
Run 10 stress 21.94549
Run 11 stress 21.27711
Run 12 stress 21.64029
Run 13 stress 21.26395
Run 14 stress 22.31659
Run 15 stress 21.74069
Run 16 stress 22.03471
Run 17 stress 21.23971
Run 18 stress 21.90118
Run 19 stress 21.30491
Run 20 stress 21.26796

> plot(nmds1)
```



Este gráfico no es muy informativo. Vamos a personalizarlo para poder obtener más información sobre los tipos de bosque.

```
> plot(nmds1, type = "n")
> points(nmds1$points, pch = as.numeric(env$Forest.type), col = as.numeric(env$Forest.type)
+       cex = 1.5)
> legend(x = "bottomleft", legend = c("Cloud forest", "Oak forest",
+   "Pine forest", "Pine-oak forest", "Transitional forest"),
+       pch = c(1:5), col = c(1:5))
```

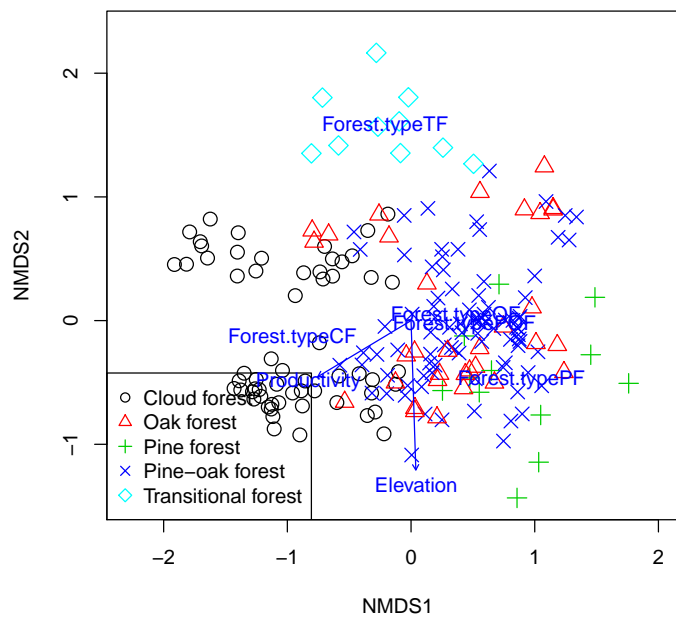


Vemos que los distintos tipos de bosque se diferencian bastante bien en cuanto a la composición de especies que los componen. Algunos grupos son más compactos, como los bosques transicionales, y otros más heterogéneos, como los bosques de niebla (que parece que forman dos subgrupos) y los bosques de encino y pino-encino. Vamos a insertar en la gráfica los vectores de las variables ambientales utilizando para ello la función `envfit()` del paquete `vegan`.

```

> plot(nmds1, type = "n")
> points(nmds1$points, pch = as.numeric(env$Forest.type), col = as.numeric(env$Forest.type),
+       cex = 1.5)
> legend(x = "bottomleft", legend = c("Cloud forest", "Oak forest",
+   "Pine forest", "Pine-oak forest", "Transitional forest"),
+       pch = c(1:5), col = c(1:5))
> ef <- envfit(nmds1, env, permu = 1000)
> plot(ef)

```



Vemos los centroides de los distintos tipos de bosque. También observamos que la elevación está relacionada con el eje 2 y la productividad con ambos ejes marcando un gradiente desde la parte superior derecha de la gráfica (menor productividad) a la parte inferior izquierda (mayor productividad). Sin embargo, las respuestas multivariantes a variables ambientales rara vez son lineales. Por ello vamos a utilizar otra técnica que nos va a permitir ajustar superficies de tendencia para las variables continuas.


```

> plot(nmds1, type = "n")
> points(nmds1$points, pch = as.numeric(env$Forest.type), col = as.numeric(env$Forest.type)
+       cex = 1.5)
> legend(x = "bottomleft", legend = c("Cloud forest", "Oak forest",
+   "Pine forest", "Pine-oak forest", "Transitional forest"),
+       pch = c(1:5), col = c(1:5))
> ordisurf(nmds1, env$Productivity, add = T)

```

This is mgcv 1.5-6 . For overview type ``help("mgcv-package")``.

Family: gaussian
Link function: identity

Formula:
 $y \sim s(x_1, x_2, k = \text{knots})$

Estimated degrees of freedom:
8.0797 total = 9.079708

GCV score: 0.0126316

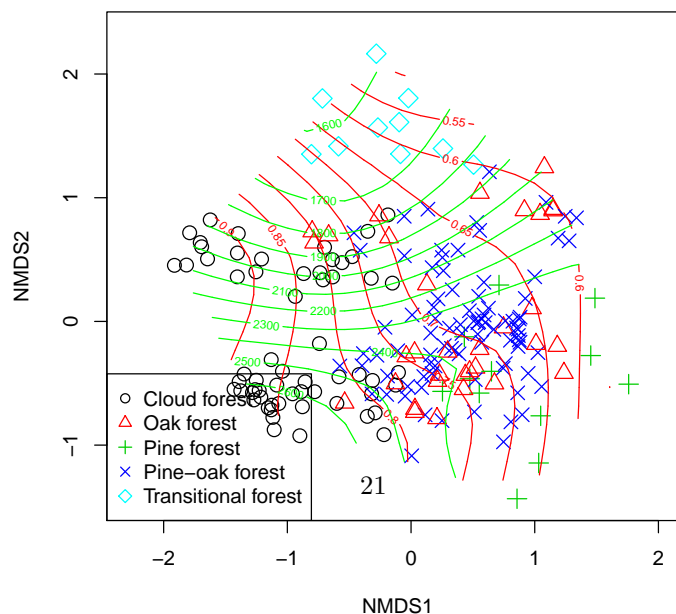
```
> ordisurf(nmds1, env$Elevation, add = T, col = "green")
```

Family: gaussian
Link function: identity

Formula:
 $y \sim s(x_1, x_2, k = \text{knots})$

Estimated degrees of freedom:
8.7164 total = 9.716417

GCV score: 10561.37



Ahora tenemos una visión mucho más completa de qué está pasando. Vemos que las zonas de mayor altitud van a determinar la presencia de bosque nublado, pero no de bosque de pino, como parecía indicar la gráfica anterior. Por otro lado la productividad va a condicionar (en mucha menor medida como vimos en el ejemplo anterior) la formación de bosques transicionales y pinares. Los bosques de encino y pino-encino muestran una heterogeneidad bastante amplia en cuanto a su respuesta a la productividad y la elevación y, finalmente, los bosques de niebla son los que más productividad tienen (por algo son bosques siempre-verdes frente al resto -excepto los bosques de pino- que son mixtos caducifolios).

5. Análisis de correspondencias canónico (CCA)

¿Qué es el análisis de correspondencias canónico? El análisis de correspondencias canónico (CCA) es una técnica multivariante que permite representar en un espacio geométrico de pocas dimensiones las proximidades existentes entre un conjunto de objetos condicionado por una serie de variables predictoras. El CCA es una técnica de ordenación restringida (*constrained ordination*), lo que significa que la ordenación de los objetos representa solamente la estructura de los datos que maximiza la relación con una segunda matriz de variables predictoras. Normalmente el CCA relaciona dos matrices: la matriz de variables dependientes (p.e. una matriz de sitios x especies) y la matriz de variables independientes (p.e. una matriz de variables ambientales). La relación entre ambas matrices se hace por medio de técnicas de regresión multivariante.

Cuando se utiliza CCA es importante tener en cuenta lo siguiente:

1. El CCA incluye la aplicación de técnicas de regresión y, por tanto, todas los supuestos y consideraciones de los modelos lineales han de ser tenidos en cuenta.
2. A medida que el número de variables ambientales aumenta con respecto al número de observaciones (muestras), el resultado del CCA se hace más dudoso, independientemente de que las relaciones observadas sean aparentemente fuertes.
3. Los usuarios de esta técnica han de tener en cuenta que su interpretación no supone una descripción de los datos de la matriz de variables dependientes *per se*, sino más bien de la parte de la estructura de los datos que está relacionada con las variables predictoras.

En el CCA, la variabilidad explicada por los ejes de ordenación está representada por el término inercia (*Inertia*). Hay una inercia total que representaría la variabilidad total de los datos (como la devianza del modelo nulo en GLM) y una devianza de la ordenación restringida (*constrained inertia*) que informa de la parte de la variabilidad total explicada por las variables

predictoras en el CCA. Asimismo es interesante ver qué proporción de dicha variabilidad queda explicada por cada uno de los ejes del CCA, teniendo en cuenta que habrá tantos ejes como variables predictoras incluyamos en el modelo, si bien generalmente la mayor parte de la variabilidad va a quedar resumida en los 2 o 3 primeros ejes.

5.1. Ejemplo: ¿Cómo se relaciona la estructura de comunidades de plantas con las variables ambientales?⁴

Siguiendo con el ejemplo anterior (ver secciones 3.1 y 4.1) queremos seguir profundizando en la relación entre las variables ambientales y la composición de árboles en bosques tropicales montanos. Los objetivos específicos de este caso de estudio son:

1. Investigar cuál es la relación entre especies y sitios explicada por variables ambientales;
2. Visualizar los datos con distintas funciones gráficas y entender los resultados de un CCA.

Los datos son los mismos que hemos utilizado en los ejemplos 3.1 y 4.1.

```
> cca1 <- cca(bio ~ Forest.type + Productivity + Elevation, data = env)
> cca1
```

```
Call: cca(formula = bio ~ Forest.type + Productivity + Elevation, data
= env)
```

```

              Inertia Rank
Total          12.775
Constrained      2.288    6
Unconstrained  10.487   85
Inertia is mean squared contingency coefficient
```

```
Eigenvalues for constrained axes:
      CCA1    CCA2    CCA3    CCA4    CCA5    CCA6
0.73472 0.58627 0.51578 0.24928 0.12219 0.08012
```

```
Eigenvalues for unconstrained axes:
      CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8
0.6702 0.5871 0.4999 0.4946 0.4819 0.4276 0.3761 0.3420
(Shown only 8 of all 85 unconstrained eigenvalues)
```

⁴Cayuela, L., Golicher, D.J., Rey Benayas, J.M., González-Espinosa, M. & Ramírez-Marcial, N. 2006. Fragmentation, disturbance and tree diversity conservation in tropical montane forests. *Journal of Applied Ecology* 43: 1172-1181

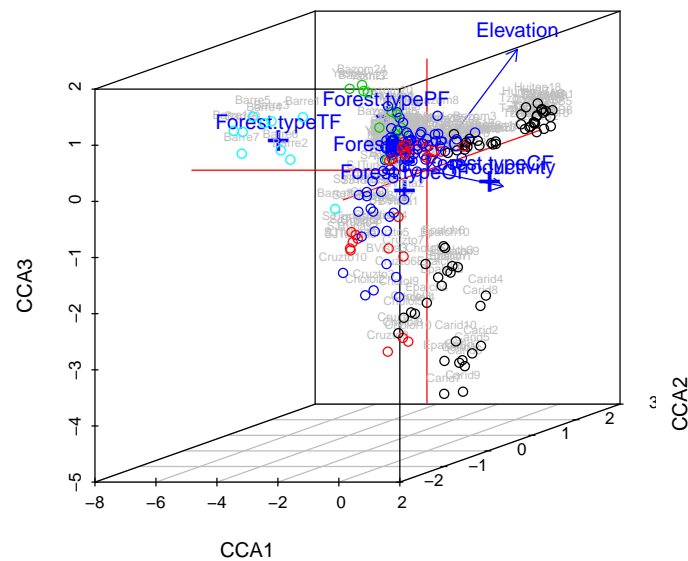
Figure 2 is a CCA ordination plot showing the relationship between forest types and environmental variables. The x-axis is labeled CCA1 and the y-axis is labeled CCA2. The plot includes a legend for forest types: Cloud forest (open circle), Oak forest (open triangle), Pine forest (plus sign), Pine-oak forest (cross), and Transitional forest (open diamond). Environmental variables are represented by vectors: Precip, Temp, Forest type TE, Forest type P, Forest type O, and Forest type PO. The plot shows a clear separation between forest types along the CCA1 axis, with environmental variables also showing distinct patterns.

24

```

> library(scatterplot3d)
> op <- ordiplot3d(cca1, angle = 25, type = "n")
> text(op, "points", col = "grey", pos = 3, cex = 0.6)
> text(op, "arrows", col = "blue", pos = 3)
> text(op, "centroids", col = "blue", pos = 3)
> points(op, "points", col = as.numeric(env$Forest.type))

```



Por último, podemos utilizar las gráficas interactivas del paquete `rgl` para representar los resultados del CCA.

```

> library(rgl)
> ordirgl(cca1, display = "sites")

```

6. Ejercicios

A continuación se proponen cuatro ejercicios. Para resolverlos, la clase se dividirá en cuatro grupos y, dentro de cada grupo, los alumnos se dispondrán por parejas para resolver uno de los cuatro casos de estudio. Para ello hay 45 minutos. El código que se genere deberá guardarse en un archivo de texto. Una vez resuelto el ejercicio, todas las parejas dentro del mismo grupo pondrán los resultados en común y lo resolverán de manera conjunta en la wiki del curso (<http://curso-r-ceama2009.wikispaces.com/>). En esta página se encontrará la información detallada sobre los ejercicios.