MANOVA

by María T. Mendoza-Marcillo

Abstract An abstract of less than 150 words.

1. Introducción

En muchos diseños de experimentos se dispone de más de una variables respuesta y se busca las diferencias entre grupos, un análisis adecuado para este tipo de situación es el análisis multivariante de la varianza, MANOVA (Amaro et al., 2004). El cual debe cumplir los siguientes supuestos paramétricos: (1)las observaciones son independientes entre sí, es decir que la muestra sea completamente aleatoria, (2)las variables respuestas tienen una distribución normal multivariante en grupo y (3)la homogeneidad de las varianzas. En este análisis, dado que hay más de una variable respuesta, no solo se debe asegurar la igualdad de las varianzas entre los grupos sino que también se debe mantener la igualdad de la covarianza entre las variables respuesta (Ates et al., 2019). Para la valoracion estadística de MANOVA, por lo general se utiliza cuatro estadísticos a partir de los cuales se contrasta la hipótesis nula de igualdad del vector de medias: (1)Lamdba de Wilks expresa la proporción de variablididad total que no es ocasionada por las diferencias entre grupos, lo que permite contrastar la hipótesis nula de que las medias multivariantes de los grupos son iguales, es decir, considera todas las raíces características pues compara si los grupos son de algún modo diferentes sin estar afectados por el hecho de que los grupos difieran en al menos una combinación lineal de las variables predictoras, (2)Máximo valor propio de Roy mide las diferencias solamente sobre la primer raíz canónica, siendo el más indicado cuando las variables respuestas están fuertemente interrelacionadas en una sola dimensión aunque es la medida que se ve más afectada por el incumplimiento de los supuestos paramétricos, (3)Traza de Lawley-Hotelling y (4)Traza de Pillai son similares al Lambda de Wilks pues utiliza todas las raíces características y se aproxima a ellas a partir de un estadístico F (Catalá and Jaume, 2001).

Si al realizar el diagnostico de MANOVA, nos lleva a rechazar los supuestos paramétricos, en especial el supuesto de homocedasticidad, un camino es hacer uso de pruebas estadísticas multivariantes no paramétricas, como PERMANOVA (Ebner, 2018). Dicha prueba, se utiliza para comparar grupos de objetos y probar la hipótesis nula de que los centroides y la dispersión de los grupos, definidos por el espacio de medida, son equivalentes para todos los grupos. (Ebner, 2018). Además, PERMANOVA a diferencia de MANOVA tradicional, no es sensible a las diferencias en la estructura de correlación entre los grupos (Anderson, 2017). Es así que PERMANOVA ahora ha ganado popularidad en medio del análisis de varianza multivariante, siendo aplicado en campos como química, ciencias sociales, agricultura, genética, psicología,etc(Anderson, 2017). Los métodos no paramétricos no hacen suposiciones explicitas respecto a las distribuciones de las variables originales, dado que solo asume intercambialidad de unidades permutables bajo una hipótesis nula verdadera. (Anderson, 2017), por lo que, PERMANOVA es muy resistente a la heterogeneidad para diseños equilibrados pero no desequilibrados.

Ciertos métodos no paramétricos, como las pruebas de permutación para experimentos, dan resultados similares a los métodos bayesianos con previas no informativas(Press, 1980). Es así, que si las pruebas de permutaciones no da una respuesta directa y significativa al problema planteado, es deseable y practico pasar a un enfoque bayesiano basado en modelos.(Gelman et al., 2014)

En los modelos bayesianos las inferencias sobre los parámetros desconocidos del modelo, deben ser respaldadas por conocimiento previo.(Press, 1980). La característica esencial de los modelos bayesianos es su uso explícito de la probabilidad para cuantificar la incertidumbre en el análisis de datos estadísticos, con lo cual se consigue modelizar efectos fijos y aleatorios mediante una estructura jerárquica de los parámetros a estimar.

Las ilustraciones académicas de los modelos paramétricos han incluido un pequeño número de observaciones en el estudio, sin realizar las validaciones de los supuestos antes mencionados. Otros problemas están relacionados con la elección del modelo no paramétrico y la falta de claridad a la hora de elegir el modelo bayesiano para el análisis de varianza(Páez et al., 2011).

En el presente estudio, utilizamos modelos parámetricos, no paramétricos y bayesianos para realizar un análisis de varianza multivariante acompañado del Manova Biplot(Amaro et al., 2004), el cual, permitirá determinar de manera visual, si las variables respuestas son influenciadas por las variables predictoras (French et al., 2008). El Manova-Biplot ayuda a la interpretación de las diferencias-semejanzas entre grupos (variables) y proporciona medidas de la calidad de representación, tanto de medias como de grupos de variables, permitiendo una mejor interpretación de resultados.(Amaro et al., 2008). A través de un diagrama de decisión, la implementación de un paquete estadístico con el

software estadístico R y el reanálisis de un conjunto de datos previamente publicado, evaluamos los métodos paramétricos, no paramétricos y bayesianos previamente mencionados.

En la sección 2 describimos y evaluamos cumplimiento de los supuestos del conjunto de datos, así mismo, se introducen los conceptos más importantes de los distintos modelos y se propone la selección del modelo que mejor se ajusta a la naturaleza de los datos. Tras ofrecer una visión general de los modelos mencionados, describimos los resultados y salidas mas importante del paquete en la Sección 3. Por último, en la sección 4 se analizan los resultados y las recomendaciones para seleccionar un modelo de análisis de varianza adecuado en la práctica.

2. Materiales y Métodos

Para el estudio se considera el conjunto de datos EEG del paquete MANOVA.RM: En el que constan 160 pacientes que fueron diagnosticados con enfermedad de Alzheimer (AD), deterioro cognitivo leve (MCI o SCC), según diagnósticos neuropsicológicos(Bathke et al., 2016). A continuación se realiza una evaluación de las pruebas estadísticas de los datos, presentados en la tabla 1, en el que se considera $\alpha = 0.05$

Method	Supuestos	Prueba	Valor-p	Decision
MANOVA	Normalidad	Mardia	2.2e-16	no cumple
MANOVA	Homogeneidad	Prueba MdeBox	2.2e-16	no cumple
PERMANOVA	Homogeneidad	permutest	0.07	no cumple

Tabla 1: evaluacion de supuestos

En la tabla 1 se evidencia que el marco de datos no cumple los supuestos. Para poder decidir qué camino seguir se propone el siguiente diagrama de decisión, el cual proporciona información para seleccionar una técnica estadística en función del cumplimiento o no de los supuestos.

La figura 1 presenta el diagrama de flujo, en el que se nombran los supuestos y se disponen los caminos a seguir debido a su cumplimiento, finalizando en la técnica que se debe aplicar con los datos. Comenzando por el supuesto de normalidad si este se cumple sigue al supuesto de homogeneidad y así continua al cumplimiento del supuesto de independencia y si es favorable se puede realizar MANOVA. Ahora si el supuesto de normalidad no se cumple, se pregunta si las observaciones intercambiabiables son independientes, y si este se cumple se realiza PERMANOVA y si esto no sucede se practicara MANOVA Bayesiano. Si el supuesto de homogeneidad de varianzas no se cumple se realizara PERMANOVA mientras sean independientes dado que es muy resistente a la heterogeneidad. A Continuación se presentan los modelos necesarios que serán aplicados al marco de datos.

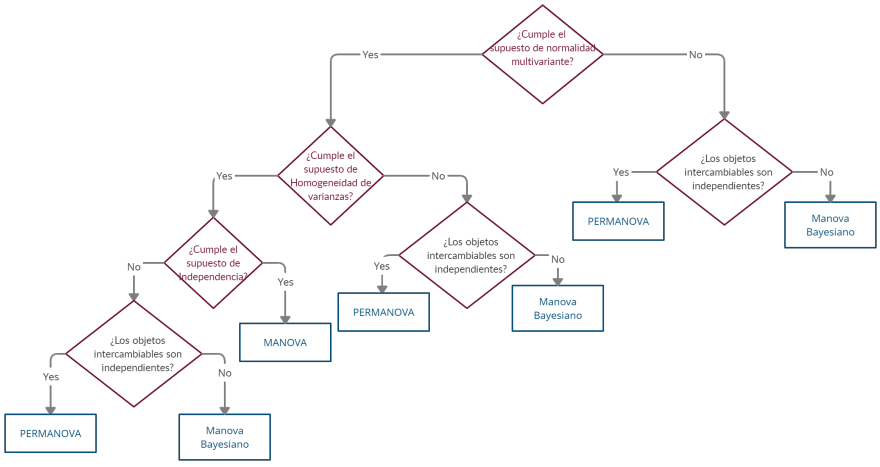


Figure 1: Diagrama de flujo

Modelos

Modelo lineal general Multivariante.

Se tiene n observaciones independientes de p variables observables $Y_1, ..., Y_p$ obtenida de diversas condiciones experimentales. El modelo lineal multivariante es:

$$Y_{np} = X_{nq}\beta_{qp} + E_{np}$$

Siendo:

X la matriz de diseño

β La matriz de parámetros de regresión desconocidos

E la matriz de desviaciones aleatorias (errores)

MANOVA de una vía

El modelo que relaciona las observaciones con los parámetros μ_i es de la forma

$$Y_{ij} = \mu_i + E_{ij}$$

$E_{ij} \sim N_p(0, \Sigma), i = 1, \ldots, q; j = 1, \ldots, n_i$

Y_{ij} Es el vector de valores que toma la v.a. multivariante estudiada para el caso j-esimo sujeto al tratamiento. μ_i Es un vector de medias general.

Ahora se plantea el contraste de la igualdad de medias:

$H_0 = \mu_1 = ... = \mu_q$

Este contraste de hipótesis de igualdad se puede llevar a cabo por el método Lamdba de Wilks, Máximo valor propio de Roy, Traza de Lawley-Hotelling y Traza de Pillai.

MANOVA de dos vías.

La diferencia en un MANOVA de dos vías es que pueden contrastarse varias hipótesis, para los efectos principales y para la interacción Suponemos un conjunto de datos $n = axb$ donde las filas representan los niveles del primer factor y las columnas los niveles del segundo factor, denotados por A y B respectivamente. El modelo que los relaciona seria:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

El parámetro μ representa la media global, los parámetros α_i representan el efecto principal del factor A, los parámetros β_j representan el efecto principal del factor B y los parámetros γ_{ij} representan la interacción de los factores A y B. (Consideremos et al.) Para esto se plantean las siguientes hipótesis:

(A) $H_0 : \alpha_i = 0, i = 1, ..., a$
(B) $H_0 : \beta_j = 0, j = 1, ..., b$ (AB) $H_0 : \gamma_{ij} = 0, i = 1, ..., a; j = 1, ..., b$

PERMANOVA

La prueba es especialmente apropiada para el análisis de datos con un tamaño de muestra pequeño y un gran número de característica(Tang et al., 2016). El pseudo estadístico F de PERMANOVA para probar la hipótesis nula que indica que no hay diferencias en las posiciones de los centroides del grupo en el espacio de la medida de disimilitud elegida viene dada por:

$$F = \frac{tr(HGH)}{tr((I - H)G(I - H)}$$

Donde, $H = X(X^TX)^{-1}X^T$ y $G = \frac{-1}{2}(I - \frac{11^T}{n})D^2(I - \frac{11^T}{n})$, siendo I una matriz identidad n x n, 1 es un vector de unos n x 1, D es una matriz de distancias que cuantifica la disimilitud entre muestras basadas en Y, y X es la matriz de diseño.

Intercambiabilidad

Sea $p(y_1, ..., y_n)$ la distribución conjunta de $Y_1, ..., Y_n$. Si $p(y_1, ..., y_n) = p(y_{\pi 1}, ..., y_{\pi n})$ para todas las permutaciones π de $1, ..., n$, entonces $Y_1, ..., Y_n$ se dice que son intercambiables. Intuitivamente, se dice que $Y_1, ..., Y_n$ son intercambiables si las etiquetas de los subíndices no contienen información sobre el resultado obtenido.(Marin, 2014)

MANOVA BAYESIANO

El enfoque bayesiano, además de especificar un modelo para los datos observados $y = (y_1, ..., y_n)$ dado un vector de parámetros desconocidos $\theta = (\theta_1, ..., \theta_k)$, usualmente en forma de densidad condicional $p(y \mid \theta)$, supone que θ es aleatorio y que tiene un densidad a priori $p(\theta \mid n)$, donde n es un vector de hiper-parámetros.(Rojas, 2010)

Posterior

Considerando las hipótesis de θ se cuestiona cual es la probabilidad posterior para cada modelo dado y por el teorema de bayes se tiene que:

$$p(\theta \mid y) \propto p(y \mid \theta)\pi(\theta)$$

,

Siendo,

$\pi(\theta)$ Probabilidad previa y

$p(y \mid \theta)$ Probabilidad marginal

Factor de Bayes

La razón de las probabilidades de los modelos dados se denomina factor de Bayes(Trotta, 2007)

$$FB_{01} = \frac{p(y \mid \theta_0)}{p(y \mid \theta_1)}$$

Si el $FB_01 > 1$, este favorece a θ_0, es decir se preferiría $\frac{\theta_0}{\theta_1}$ con probabilidad FB_{01}, de manera contraria seria verdad para $FB_{01} < 1$

MÉTODO BIPLOT

El método Biplot proporciona en la presentación grafica de un gran marco de datos basado en matrices. Siendo su característica mostrar las distancias o agrupamientos que existen entre individuos y correlación entre variables. (Narváez et al., 2020)

MANOVA-BIPLOT

El Manova-Biplot es una técnica estadística multivariante utilizada en situaciones experimentales donde se dispone de varias variables respuesta y se quiere buscar las diferencias entre varios grupos.(Amaro et al., 2008). En el Manova-Biplot, además de interpretar las diferencias-semejanzas entre los grupos; también podemos interpretar las relaciones entre las variables; y las relaciones entre grupos y variables.(Amaro et al., 2008)

El Manova-Biplot se construye a partir de la descomposición en valores singulares generalizada:

$$R^{-1/2}\widehat{D}E^{-1/2} = UD_\lambda V$$

Siendo

$R = C(A'A)^{-1}C'$

$\widehat{D} = C\widehat{B} = C(A'A)^{-1}A'X$

3. Resultados

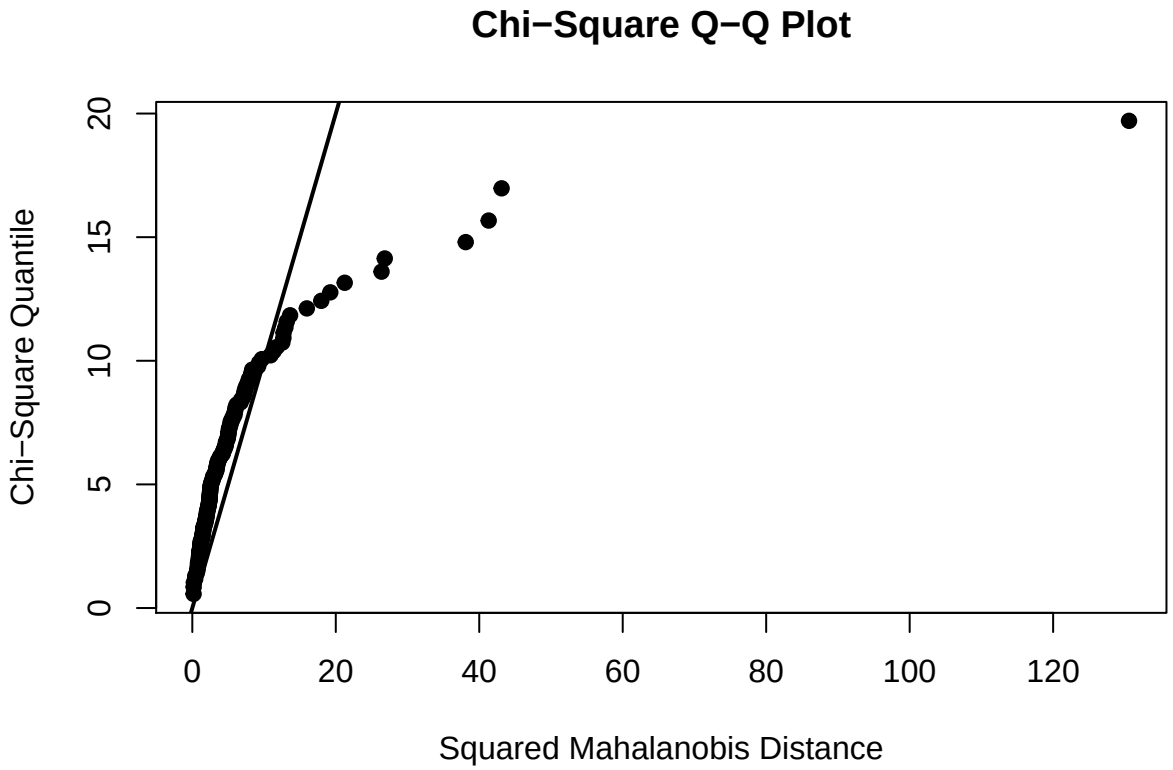
Datos

El conjunto de datos para el estudio contiene Z-score para la frecuencia cerebral y Complejidad, cada una medida en las posiciones de los electrodos frontal, temporal y central y promediada a través de hemisferios. Los tres factores que se consideraron para la parcela completa fueron sexo (hombres, mujeres), diagnóstico (EA, DCL, SCC) y edad $(70, \geq 70 aos)$. Además, la región de factores de la subparcela (frontal, temporal, central) y característica (tasa cerebral, complejidad) estructuran el vector de respuesta. Para realizar el análisis de los datos, se usara el paquete MaBY el cual se describe a continuación.

El paquete

```
#> Registered S3 method overwritten by 'GGally':
#>   method from
#>   +.gg   ggplot2

#> sROC 0.1-2 loaded

#> Loading required package: car

#> Loading required package: carData

#>
#> Attaching package: 'car'

#> The following object is masked from 'package:psych':
#>
#>     logit

#>                   Df  Pillai approx F num Df den Df    Pr(>F)
#> diagnostico        2 0.29851   4.3861     12    300 1.863e-06 ***
#> sexo               1 0.11204   3.1333      6    149  0.006386 **
#> diagnostico:sexo   2 0.10375   1.3678     12    300  0.180272
#> Residuals        154
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#>                   Df   Wilks approx F num Df den Df   Pr(>F)
#> diagnostico        2 0.71226   4.5915     12    298 7.95e-07 ***
#> sexo               1 0.88796   3.1333      6    149 0.006386 **
#> diagnostico:sexo   2 0.89778   1.3757     12    298 0.176383
#> Residuals        154
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Normalidad




```
#> $multivariateNormality
#>              Test        Statistic p value Result
#> 1 Mardia Skewness 2834.90226596294       0     NO
#> 2 Mardia Kurtosis  84.124773217597       0     NO
#> 3             MVN             <NA>    <NA>     NO
#>
#> $univariateNormality
#>           Test            Variable Statistic   p value Normality
#> 1 Shapiro-Wilk brainrate_temporal     0.9875   0.166      YES
#> 2 Shapiro-Wilk  brainrate_frontal     0.9867  0.1315      YES
#> 3 Shapiro-Wilk  brainrate_central     0.9876  0.1703      YES
#> 4 Shapiro-Wilk complexity_temporal    0.8534  <0.001      NO
#> 5 Shapiro-Wilk complexity_frontal     0.8273  <0.001      NO
#> 6 Shapiro-Wilk complexity_central     0.6068  <0.001      NO
```

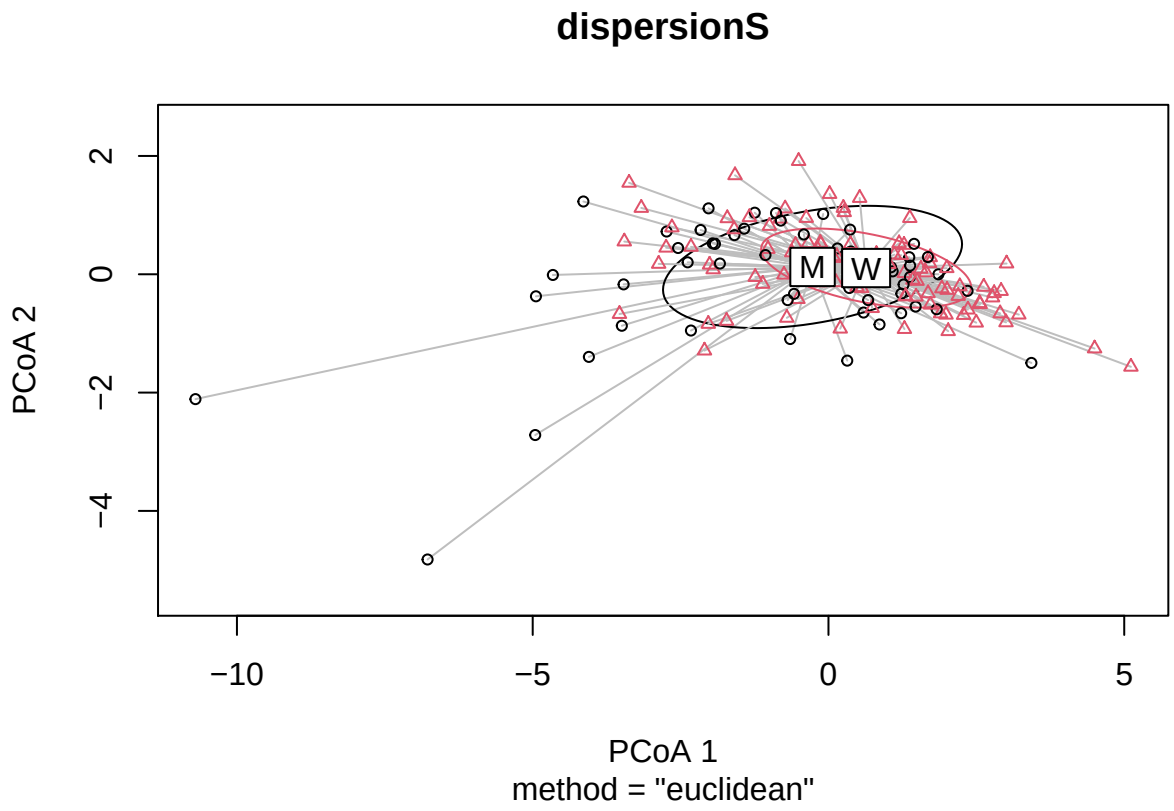
Para evaluar el supuesto de normalidad, se usa el test de Mardia basada en las medida de asimetria y kurtosis, este test nos indica que no se esta cumpliendo la normalidad

Homocedasticidad

```
#validar el supuesto de igual de matrices de varianzas y covarianzas

boxM(as.matrix(Y)~diagnostico *  sexo, data=EEGwide)

#>
#>  Box's M-test for Homogeneity of Covariance Matrices
#>
#> data:  Y
#> Chi-Sq (approx.) = 470.12, df = 105, p-value < 2.2e-16

#Se viola el supuesto de Igualdad, se recomienda usar Pillai's
```

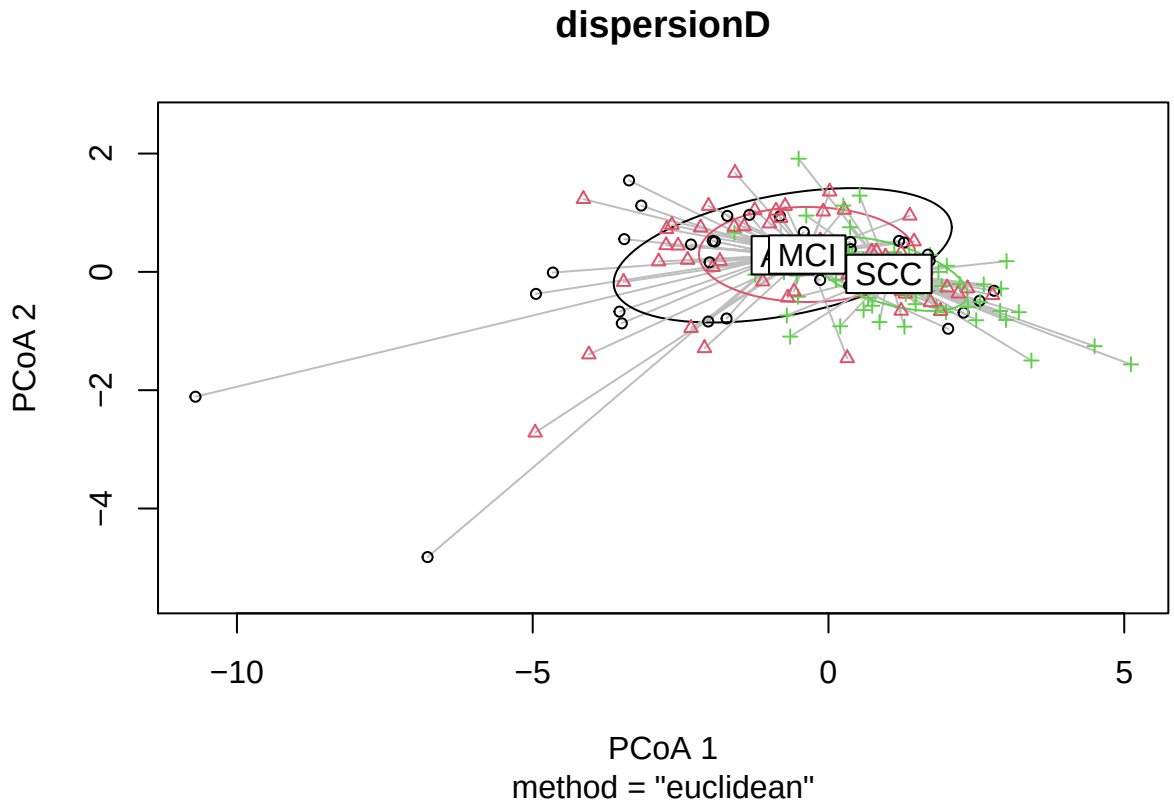
Intercambiabilidad

```
#> Loading required package: permute

#> Loading required package: lattice

#> This is vegan 2.5-7

#> Permutation test for adonis under reduced model
#> Terms added sequentially (first to last)
#> Permutation: free
#> Number of permutations: 999
#>
#> adonis2(formula = DistanciasY ~ diagnostico * sexo, data = EEGwide, permutations = 999)
#>                   Df SumOfSqs      R2       F Pr(>F)
#> diagnostico        2   149.23 0.15643 15.6546  0.001 ***
#> sexo               1    48.48 0.05082 10.1718  0.001 ***
#> diagnostico:sexo   2    22.27 0.02334  2.3361  0.057 .
#> Residual         154   734.02 0.76941
#> Total            159   954.00 1.00000
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#>
#> Permutation test for homogeneity of multivariate dispersions
#> Permutation: free
#> Number of permutations: 999
#>
#> Response: Distances
#>            Df Sum Sq Mean Sq      F N.Perm Pr(>F)
#> Groups      1   6.28  6.2773 3.0878    999  0.073 .
#> Residuals 158 321.20  2.0329
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
#>
#> Permutation test for homogeneity of multivariate dispersions
```

```
#> Permutation: free
#> Number of permutations: 999
#>
#> Response: Distances
#>           Df  Sum Sq Mean Sq      F N.Perm Pr(>F)
#> Groups     2  36.988 18.4939 11.881    999  0.001 ***
#> Residuals 157 244.382  1.5566
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



4.

About this format and the R Journal requirements

`rticles::rjournal_article` will help you build the correct files requirements:

- A R file will be generated automatically using `knitr::purl` - see https://bookdown.org/yihui/rmarkdown-cookbook/purl.html for more information.
- A tex file will be generated from this Rmd file and correctly included in `RJwapper.tex` as expected to build `RJwrapper.pdf`.
- All figure files will be kept in the default rmarkdown `*_files` folder. This happens because `keep_tex = TRUE` by default in `rticles::rjournal_article`
- Only the bib filename is to modifed. An example bib file is included in the template (`RJreferences.bib`) and you will have to name your bib file as the tex, R, and pdf files.

Bibliography

I. Amaro, J. L. Vicente Villardon, and M. P. GALINDO VILLARDÓN. Contribution to manova-biplots: alternative confidence regions. *Revista Investigación Operacional*, 29, 01 2008. [p1, 4]

I. R. Amaro, V.-V. Luis, and M. P. Galindo-Villardón. Manova Biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, 29:26 – 32, 01

2004. ISSN 0378-1844. URL http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0378-18442004000100009&nrm=iso. [p1]

M. J. Anderson. *Permutational Multivariate Analysis of Variance (PERMANOVA)*, pages 1–15. American Cancer Society, 2017. ISBN 9781118445112. doi: https://doi.org/10.1002/9781118445112.stat07841. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat07841. [p1]

C. Ates, O. Kaymaz, H. E. Kale, and M. A. Tekindal. Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-i error rates. *Computational and mathematical methods in medicine*, 2019, 2019. [p1]

A. Bathke, S. Friedrich, F. Konietschke, M. Pauly, W. Staffen, N. Strobl, and Y. Höller. Using eeg, spect, and multivariate resampling methods to differentiate between alzheimer's and other cognitive impairments. *arXiv preprint arXiv:1606.09004*, 2016. [p2]

R. M. Catalá and M. J. R. Jaume. *Estadística informática: casos y ejemplos con el SPSS*. Publicaciones de la Universidad de Alicante, 2001. [p1]

I. Consideremos, Y. YI, and Y. nI de una población Nd. Tema 4. análisis multivariante de la varianza. [p3]

J. N. Ebner. Permutational multivariate analysis of variance (permanova) in r. *Archetypal Ecology*, 2018. [p1]

A. French, M. Macedo, J. Poulsen, T. Waterson, and A. Yu. Multivariate analysis of variance (manova), 2008. [p1]

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian data analysis (vol. 2), 2014. [p1]

J. M. Marin. *Tema 1: Introducción a la Estadística Bayesiana*. BAYES, 2014. [p4]

M. G. Narváez, O. R. Barzola, and A. N. Librero. Análisis multivariante: Un recorrido por las técnicas de reducción de dimensiones. *Matemática*, 18(2), 2020. [p4]

L. O. M. Páez, M. R. Lozano, and J. A. R. Davila. Descripción general de la inferencia bayesiana y sus aplicaciones en los procesos de gestión. *La simulación al Servicio de la Academia*, 2:1–28, 2011. [p1]

S. J. Press. 4 bayesian inference in manova. *Handbook of Statistics*, 1:117–132, 1980. [p1]

H. A. G. Rojas. Modelamiento bayesiano conjunto de media y varianza en modelos lineales mixtos. 2010. [p4]

Z.-Z. Tang, G. Chen, and A. V. Alekseyenko. Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, 32(17):2618–2625, 2016. [p3]

R. Trotta. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378(1):72–82, 05 2007. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2007.11738.x. URL https://doi.org/10.1111/j.1365-2966.2007.11738.x. [p4]

María T. Mendoza-Marcillo
Facultad de Ciencias Naturales y Matematicas
ESPOL, Escuela Superior Politecnica del Litoral
Guayaquil, Ecuador
https://journal.r-project.org
ORCiD: 0000-0002-9079-593X
mtmendoz@espol.edu.ec