

MANOVA

by Maria T. Mendoza-Marcillo

Abstract An abstract of less than 150 words.

Introduction

En muchos diseños de experimentos se dispone de más de una variables de respuesta y se quiere buscar las diferencias entre grupos, un análisis adecuado para este tipo de situación es el análisis multivariante de la varianza, MANOVA (Amaro et al., 2004), que debe cumplir los siguientes supuestos paramétricos: las observaciones son independientes entre sí, es decir que la muestra sea completamente aleatoria, las variables dependientes tienen una distribución normal multivariante en grupo y la homogeneidad de las varianzas. En esta prueba, dado que hay más de una variable de 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 de respuesta. (Ates et al., 2019). Luego para valorar la significación estadística de las diferencias multivariantes entre los grupos, se encuentran cuatro estadísticos a partir de los cuales contrastar la hipótesis nula de igualdad de vectores de medias: Lambda de Wilks considera todas las raíces características dado que compara si los grupos son diferentes sin estar afectados por que los grupos difieren en al menos una combinación lineal de las variables de respuesta, Máximo valor propio de Roy es válido cuando las variables de respuesta están altamente correlacionadas, Traza de Lawley-Hotelling y Traza de Pillai son similares al Lambda de Wilks ya que así mismo utiliza todas las raíces características (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 la prueba estadística multivariante no paramétrica PERMANOVA (Ebner, 2018). 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). Se extiende a la partición de datos multivariantes en respuesta a diseño de experimentos complejos, como en el que puede haber más variables de respuesta que unidades de muestreo, es así que PERMANOVA ahora ha ganado popularidad en medio del análisis de varianza multivariante, siendo aplicando en campos como química, ciencias sociales, agricultura, genética, psicología, y más. (Anderson, 2017). Sin embargo para obtener una representación, en Porcel et al. (2016) se utilizó un análisis de componentes principales (PCA) para visualizar patrones multivariantes e identificar posibles agrupaciones. Al realizar una revisión a los supuestos de la prueba no paramétrica, notamos que no hace suposiciones explícitas respecto a las distribuciones de las variables originales, dado que solo asume intercambialidad de unidades permutables bajo una hipótesis nula verdadera. (Anderson, 2017). PERMANOVA es muy resistente a la heterogeneidad para diseños equilibrados pero no desequilibrados.

Algunos métodos no paramétricos, como las pruebas de permutación para experimentos, dan resultados similares en problemas simples a las inferencias bayesianas con distribuciones previas no informativas, si el modelo bayesiano se construye para ajustarse razonablemente bien a los datos, estos problemas simples incluyen diseños equilibrados sin datos faltantes. Al estimar varios parámetros a la vez o incluir variables explicativas en el análisis o información previa sobre los parámetros, el método de la teoría de la permutación no da una respuesta directa, y esto a menudo proporciona un incentivo práctico considerable para pasar a un enfoque bayesiano basado en modelos. (Gelman et al., 2014)

Para efectuar inferencias sobre los parámetros desconocidos del modelo es necesario hacer algunas suposiciones, respaldadas por conocimiento previo. (Press, 1980). La característica esencial de los modelos bayesianos es su uso explícito de probabilidad de calificar la incertidumbre en inferencias basadas en análisis de datos estadísticos, adoptando el diseño p-dimensional, unidireccional (clasificación), modelo de efectos fijos, hay observaciones sobre q poblaciones, cada p-dimensional, con matriz de covarianza común.

Aunque la teoría de MANOVA clásico no proporciona errores para la estimación de la varianza y componentes de la covarianza. (Jiang and Skorupski, 2018). El propósito aquí no es descartar los métodos no paramétricos clásicos, sino proponer una segunda opción, ante la imposibilidad de aplicar estadística multivariante paramétrica clásica, como lo es un contexto bayesiano.

Como complemento del análisis de varianza multivariante se aportara a la presentación usando método de Biplot, y así se podría determinar si las variables de respuesta se modifican por la manipulación de las variables independientes (French et al., 2008).

El Manova-Biplot ayuda a la interpretación de las diferencias-emejanzas entre grupos y variables

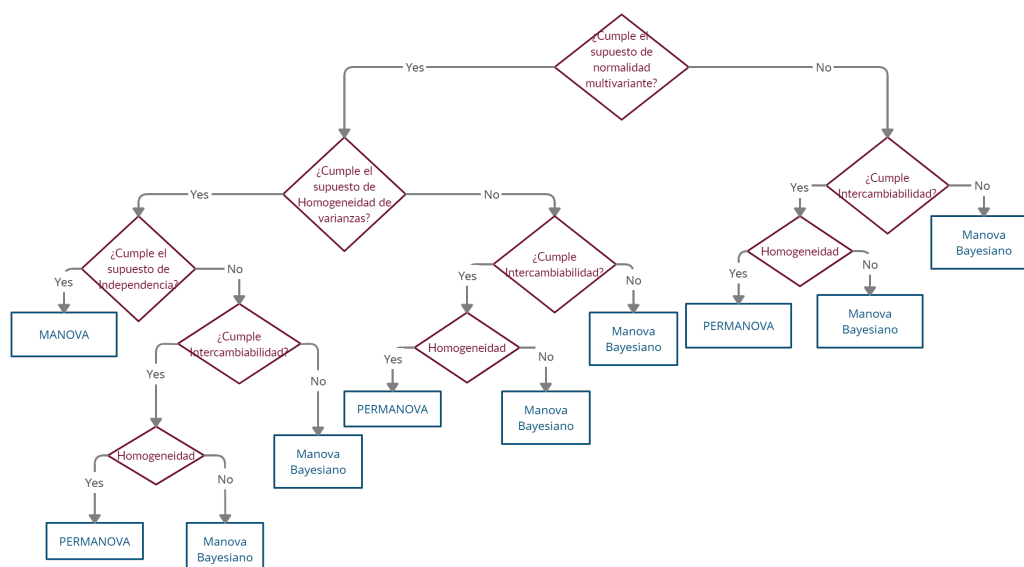


Figure 1: Diagrama de flujo

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). Podemos tener un MANOVA de dos vías, en la cual se pueden contrastar varias hipótesis, para los efectos principales y para la interacción(Amaro et al., 2004), con las cuales se puede construir distintas representaciones como puede ser 4 tipos de Biplot: Biplot Total, Biplot de interacción, de filas y de columnas.

Dadas las alternativas de análisis de varianza multivariante, el trabajo tiene como finalidad implementar un paquete que sea de uso fácil y para analizar datos donde se cumplan o no los supuestos, debido a datos dispersos, tamaño de muestras desiguales, mediante coeficientes de disimilitud o pruebas de permutación.

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	Shapiro-Wilk	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 estos no se cumple, se sigue a la verificación de intercambiabilidad, en el que si se cumple se pregunta si se cumple el supuesto de homogeneidad dando así paso a la práctica de PERMANOVA y si esto no sucede se practicara MANOVA Bayesiano. A Continuación se presentan los modelos necesarios que serán aplicados al marco de datos.

Materiales y métodos

Modelo lineal general Multivariante.

Se tiene n observaciones independientes de p variables observables Y_1, \dots, 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

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

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

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:

$$H_0 = \mu_1 = \dots = \mu_q$$

Este contraste de hipótesis de igualdad se puede llevar a cabo por el método Lambda 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 $n = a \times b$ 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 sería:

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

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:

$$(A) H_0 : \alpha_i = 0, i = 1, \dots, a$$

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

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

MANOVA BAYESIANO

Intercambiabilidad

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

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

Posterior

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

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

,

Siendo,

$\pi(\theta)$ Probabilidad previa y

$p(y | \theta)$ Probabilidad marginal

Factor de Bayes

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

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

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 sería verdad para $FB_{01} < 1$

MÉTODO BIPLLOT

El método Biplot proporciona en la presentación gráfica 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-BIPLLOT

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-emejanzas 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:

$$R^{-1/2} \hat{D} E^{-1/2} = U D_{\lambda} V$$

Siendo

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

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

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 ($70, \geq 70$ años). 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 usará el paquete MaBY el cual se describe a continuación.

El paquete

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 `RJwrapper.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 be modified. 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. [p2, 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. *Interiencia*, 29:26 – 32, 01 2004. ISSN 0378-1844. URL http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0378-18442004000100009&nrm=iso. [p1, 2]
- 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. Konietzschke, 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]
- Z. Jiang and W. Skorupski. A bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior research methods*, 50(6):2193–2214, 2018. [p1]
- J. M. Marin. *Tema 1: Introducción a la Estadística Bayesiana*. BAYES, 2014. [p3]
- 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]
- M. Porcel, J. Pålsson, G. Andersson, M. Tasin, B. Rämert, et al. Influence of agricultural management on canopy-dwelling predatory and herbivorous arthropod communities in swedish apple orchards. *IOBC-WPRS bulletin*, 2016. [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]

Maria 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