

UNIVERSIDAD NACIONAL DE CÓRDOBA  
MAESTRÍA EN ESTADÍSTICA APLICADA



MODELOS MIXTOS MULTIVARIADOS PARA RESPUESTAS  
CORRELACIONADAS  
AVANCES ESTADÍSTICOS-COMPUTACIONALES PARA  
PROCESAR INFORMACIÓN EN NEUROCIENCIAS

Tesis para optar por el grado de  
**Magister en Estadística Aplicada**

Autora  
**Dra. Vanesa B. Meinardi**

Directora  
**Dra. Mónica Balzarini**

2021



Modelos mixtos multivariados para respuestas correlacionadas avances estadísticos-computacionales para procesar información en neurociencia por Vanesa Beatriz Meinardi se distribuye bajo una [Licencia Creative Commons Atribución-NoComercial-SinDerivadas 4.0 Internacional](https://creativecommons.org/licenses/by-nc-nd/4.0/).

*Dedicado a  
mi familia*

# Agradecimientos

A **Leo**, mi compañero de ruta, mi amor, que sin su sostén este trabajo no hubiese sido posible. A **Fede y Lara**, la razón de mi vida, gracias por existir. A **mis padres**, que con su amor me formaron como persona y me transmitieron la importancia de estudiar, este logro y muchos otros se lo debo a ellos. A **Betty** que con su amor y enorme paciencia cuidaba de Fede para que pudiese cursar. A **Mónica**, por confiar en mi sin conocerme demasiado y brindarme su apoyo condicional desde el primer día. A **Carina** por estar siempre presente, la vida nos sigue juntando por distintos motivos. A **Juan** por confiar en mi trabajo y contagiarme su pasión por lo que hace. Al **cuerpo docente de la maestría**. A mis **amigos** y a todos los que de una u otra forma estuvieron presentes en este camino.

*¡Muchas gracias a todos!*

# Resumen

La electroencefalografía es una técnica de exploración funcional del sistema nervioso central mediante el cual se obtiene el registro de la actividad eléctrica cerebral en tiempo real. Para poder recoger y registrar una señal de la actividad eléctrica en cada región cerebral a través de la superficie craneal se colocan electrodos que captan la diferencia de potencial entre ellos, sobre determinados canales de registros. Cada canal de registro, mide la diferencia de voltaje entre dos electrodos. Para la obtención del trazado se incluyen técnicas de activación, principalmente la optoestimulación intermitente y la hiperventilación. Estas técnicas de activación están dirigidas a provocar o intensificar la aparición de anomalías o variaciones en la actividad cerebral.

En la interpretación clínica de un electroencefalograma (EEG), se observan canales claves de la deflexión de la señal producto del parpadeo que permiten diferenciar condiciones de vigilia. La diferenciación visual de estas condiciones es difícil de lograr por lo que se recurre a estudios empíricos sobre los grandes volúmenes de datos generados durante el monitoreo cerebral de pacientes bajo condiciones controladas. El análisis de los datos de EEG usual es a través de medidas e índices procedentes de la teoría de la información. Los más usados son la entropía de permutación (HPE) y la complejidad de

permutación Lempel-Ziv (CLZ), que permiten identificar cambios dinámicos en la función cerebral producida por distintos estados electrofisiológicos.

En esta tesis se ajustaron modelos lineales mixtos (MLM) a una base de datos consistente en registros de electroencefalogramas (EEG), los cuales fueron tomados como parte de un proyecto de investigación de una tesis doctoral en Neurociencia de la UNC. Los EEG fueron realizados según el sistema internacional 10–20 IFNC (Federación Internacional de Clínica Neurofisiológica), con 20 electrodos colocados en el cuero cabelludo, sobre determinados canales de registro. La duración de los EEG fue de 60 minutos, durante el cual cada individuo se encontraba en condición de vigilia y dentro de esta condición se describieron 4 estados electrofisiológicos: *vigilia ojos cerrados*; *vigilia ojos abiertos* y dos estados conocidos como técnicas o métodos de activación que son *optoestimulación* e *hiperventilación*. Las 20 señales de cada uno de los 4 segmentos de los registros (estados electrofisiológicos) se cuantificaron con métricas de teoría de la información: entropía de permutación y complejidad de Lempel Ziv.

Los MLM multivariados fueron usados con el fin de testear pruebas de hipótesis estadísticas para identificar cambios funcionales en los EEG mediante experimentos en neurociencia, explotando la estructura de correlación de los datos.

Los modelos estadísticos usados incluyeron en su estructura de media las condiciones que definen el escenario experimental en relación a los factores que interesa comparar, mientras que la estructura de varianza-covarianza es especificada con distintos modelos de correlación entre clúster de datos y he-

terogeneidad de varianzas. Se seleccionó el modelo de mejor ajuste mediante criterios de información y prueba de cociente de verosimilitud. Además se evaluó, mediante simulación estocástica, el desempeño de los diferentes modelos ajustados en relación a la probabilidad estimada de los errores inherentes a las pruebas de hipótesis estadísticas. Finalmente, se implementó el modelo más potente para detectar cambios en la evaluación de EEG sobre el caso ilustrativo de aplicación.

***Palabras claves:*** Modelos Lineales mixtos, respuestas múltiples, datos correlacionados, entropía de permutación, complejidad de permutación, neurociencia.

# Summary

Electroencephalography is a technique of functional exploration of the central nervous system, which is obtained through the electrical activity of the brain in real time. In order to collect and record a signal of electrical activity in each brain region through the cranial surface, electrodes capture the potential difference between them are placed on certain recording channels. Each recording channel measures the voltage difference between two electrodes. Activation techniques are included to obtain the tracing, mainly intermittent optostimulation and hyperventilation. These activation techniques are aimed at causing or intensifying the appearance of abnormalities or variations in brain activity.

In the clinical interpretation of an electroencephalogram (EEG), key channels of the deflection of the signal produced by blinking are observed that allow differentiating conditions of wakefulness. Visual differentiation of these conditions is difficult to achieve, so empirical studies are used on the large volumes of data generated during brain monitoring of patients under controlled conditions. The analysis of the usual EEG data is through measurements and indices from the information theory. The most used are the permutation entropy (HPE) and the Lempel-Ziv permutation complexity (CLZ), which



allow identifying dynamic changes in brain function produced by different physiological states.

In this thesis, linear mixed models (MLM) were adjusted to a database consisting of electroencephalogram (EEG) records, which were taken as part of a research project for a doctoral thesis in Neuroscience at the UNC. The EEGs were performed according to the international 10-20 IFNC system (International Federation of Neurophysiological Clinic), with 20 electrodes placed on the scalp, over certain recording channels. The duration of the EEGs was 60 minutes, during which each individual is in a waking condition and within this condition 4 electrophysiological states are described: *wake eyes closed*; *wake eyes open* and two states known as techniques or methods of activation that are *optostimulation* and *hyperventilation*. The 20 signals from each of the 4 segments of the recordings (electrophysiological states) were quantified with information theory metrics: permutation entropy and complexity of Lempel Ziv.

Multivariate MLMs are used for the purpose to confirmatory hypothesis testing to identify functional changes in EEGs through neuroscience experiments, exploiting the correlation structure of the data.

The statistical models used included in their mean structure the conditions that define the experimental setting in relation to the factors that are of interest to compare, while the variance-covariance structure is specified with different correlation models between cluster of data and heterogeneity of variances. The best fit model was selected using information criteria and the likelihood ratio test. In addition, the performance of the different models

adjusted in relation to the estimated probability of the errors inherent in the hypothetical tests was evaluated by stochastic simulation. Finally, the most powerful model was implemented to detect changes in the EEG evaluation on the illustrative case of application.

***Key words:*** Linear mixed models, multiple responses, correlated data, permutation entropy, permutation complexity, neuroscience.

# Índice general

<b>Agradecimientos</b>	<b>II</b>
Resumen . . . . .	III
<b>Lista de figuras</b>	<b>XII</b>
<b>Lista de tablas</b>	<b>XIII</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Problema motivador . . . . .	3
1.2. Objetivos . . . . .	5
<b>2. Marco teórico</b>	<b>6</b>
2.1. Desde la neurociencia . . . . .	6
2.2. Cuantificación de los EEG . . . . .	8
2.2.1. Entropía de Permutación . . . . .	9
2.2.2. Complejidad de Lempel-Ziv . . . . .	9
2.3. Divergencia de Jensen-Shanon . . . . .	12
2.3.1. Interpretación de $D$ en el marco de la estadística ma- temática . . . . .	13

2.4.	Modelos lineales mixtos . . . . .	16
2.5.	Modelo lineal marginal . . . . .	18
2.6.	Estimación de efectos fijos y predicción de efectos aleatorios en un MLM . . . . .	19
2.7.	Inferencia en modelos lineales mixtos . . . . .	23
2.8.	Criterio de selección de modelos . . . . .	24
2.9.	Prueba de Hipótesis . . . . .	25
<b>3.</b>	<b>Aplicación en la experimentación en Neurociencias</b>	<b>29</b>
3.1.	Análisis exploratorio . . . . .	29
3.2.	Investigación metodológica . . . . .	36
3.2.1.	Propósitos y parametrización de modelos . . . . .	36
3.3.	Procedimientos de evaluación . . . . .	40
3.3.1.	MLM multivariados ajustados para la diferenciación de estados electrofisiológicos . . . . .	40
3.3.2.	Evaluación de modelos . . . . .	48
3.3.3.	Resultados . . . . .	51
3.3.4.	Predicciones en el modelo multivariado . . . . .	54
3.4.	Modelación estadística y divergencia de Jensen-Shanon para diferenciación de estados electrofisiológicos por canales . . . . .	56
3.4.1.	Modelos univariados . . . . .	57
3.4.2.	Aplicación Divergencia de Jensen-Shanon . . . . .	65
<b>4.</b>		<b>70</b>
4.1.	Comentarios finales . . . . .	70

<b>5.</b>	<b>75</b>
5.1. Bibliografía . . . . .	75
5.2. Apéndice A . . . . .	79
5.3. Apéndice B . . . . .	88
5.3.1. Conglomerados . . . . .	88
5.3.2. Modelos ajustados . . . . .	89
5.3.3. LSD Fisher con corrección de Bonferroni . . . . .	90
5.3.4. Error de Tipo I y II . . . . .	109
5.3.5. Predicciones en el modelo multivariado . . . . .	112
5.3.6. Divergencia de Jensen-Shanon . . . . .	113

# Índice de figuras

3.1. Diagrama de cajas de entropía y complejidad particionadas por los diferentes estados . . . . .	30
3.2. Diagrama de cajas de entropía particionadas por individuo . .	32
3.3. Diagrama de cajas de complejidad particionadas por individuo	32
3.4. Diagrama de cajas de entropía particionadas por los diferentes estados e individuos . . . . .	33
3.5. Sectores basados en la media del índice de entropía de permutación . . . . .	34
3.6. Sectores basados en la media del índice de entropía de permutación y complejidad de Lempel Ziv . . . . .	35
3.7. Diagrama de cajas de entropía y complejidad particionadas por sectores . . . . .	35

# Índice de tablas

3.1. Medidas estadísticas de las variables entropía de permutación y complejidad de permutación Lempel-Ziv particionadas por estados. . . . .	30
3.2. Sectores basados en la media del índice de entropía de permutación. . . . .	34
3.3. Sectores basados en la media del índice de entropía de permutación y complejidad de permutación Lempel-Ziv. . . . .	35
3.4. Parámetros de covarianza estimados por un modelo lineal mixto multivariado 3.6 . . . . .	45
3.5. Covarianza de los residuos agrupados por sujetos . . . . .	46
3.6. Errores de tipo I y II inherentes a la pruebas de hipótesis 3.1 para los modelos (3.6)y (3.7) . . . . .	51
3.7. Medias estimadas por el modelo multivariado (3.6) para la variable respuesta entropía y complejidad . . . . .	53
3.8. Medias estimadas por el modelo (3.7) para la variable respuesta entropía . . . . .	53
3.9. Medias estimadas por el modelo (3.7) para la variable respuesta complejidad . . . . .	53

3.10. Predicciones para el modelo (3.6) . . . . .	54
3.11. Predicciones para el modelo (3.6) . . . . .	55
3.12. Predicciones para el modelo (3.6) . . . . .	56
3.13. Ajustes de medias correspondientes al modelo (3.8) . . . . .	61
3.14. Ajustes de medias correspondientes al modelo (3.8) . . . . .	62
3.15. Ajustes de medias correspondientes al modelo (3.8) . . . . .	63
3.16. Mínimo de $\Delta L$ para cada Sujeto (Id) entre todos los canales para cada combinación de a dos de estados fisiológico. . . . .	69
5.1. Medidas estadísticas de la variable complejidad particionada por estado y canal . . . . .	80
5.2. Medidas estadísticas de la variable complejidad particionada por estado y canal . . . . .	81
5.3. Medidas estadísticas de la variable complejidad y entropía par- ticionadas por estado y canal . . . . .	82
5.4. Medidas estadísticas de la variable entropía particionada por estado y canal . . . . .	83
5.5. Medidas estadísticas de la variable entropía particionada por estado y canal . . . . .	84
5.6. Medidas estadísticas de la variable complejidad por individuo .	85
5.7. Medidas estadísticas de la variable entropía por individuo . . .	86
5.8. Medidas estadísticas para las variables entropía y complejidad particionadas por estado y sector . . . . .	87



# Capítulo 1

## Introducción

Los modelos lineales mixtos (MLM)(West et al., 2015) para variables de respuestas continuas son modelos estadísticos que permiten relajar los supuestos tradicionales del modelo lineal general y modelar de manera flexible, estructura complicada de datos. Estos modelos mantienen la suposición de que los errores están normalmente distribuidos pero admiten que estos puedan ser correlacionados o no independientes y que no tengan varianza constante. En la experimentación con humanos, la correlación entre las observaciones es fuertemente inducida por el diseño de muestreo, por ej. porque se miden clúster de personas relacionadas o porque cada persona se mide repetidamente y emergen así clúster de datos de un mismo sujeto. Los MLM se clasifican según el tipo de datos en modelos *multinivel o jerárquico* si los datos están agrupados y; en *longitudinal* o modelos de *medidas repetidas* si los individuos son medidos repetidamente a lo largo del tiempo. También existen los MLM para *datos longitudinales agrupados* que combinan características de datos agrupados y longitudinales. La modelación de la correlación entre

las observaciones en el marco de los MLM puede realizarse mediante modelación directa de la matriz de covarianza de los errores (*modelos marginales*), o por la incorporación de variables aleatorias (*modelos sujeto específico*). El primero es también conocido como modelo de covarianza residual y el segundo como modelo mixto propiamente dicho ya que tendrá efectos fijos y efectos aleatorios distinto al término de error presente en todo modelo estadístico.

En los *modelos marginales* se especifica una estructura de correlación entre las respuestas y se modela la esperanza marginal de la variable respuesta como una función de covariables o efectos de factores de interés. El objetivo de las inferencias recae sobre el efecto promedio de estos factores (usualmente factores tratamiento) en la población (inferencia promedio poblacional). Esto significa que si trabajamos con estructura jerárquica de anidamiento de uno o más factores de clasificación aleatorios, el modelo marginal no tendrá en cuenta el efecto de esta estructura al momento de la inferencia promedio poblacional. En estos modelos, los coeficientes del predictor lineal representan a efectos fijos.

En los *modelos sujetos específicos o modelo mixto* se tienen en cuenta los conglomerados de datos conformados por el muestreo de distintos niveles de uno o mas factor aleatorio. Así se incorporan coeficientes de efectos aleatorios además de los correspondientes a efectos fijos en el predictor lineal para modelar no solo el valor esperado promedio poblacional sino también la variabilidad entre los conglomerados definidos por la estructura de efectos aleatorios. Estos modelan la esperanza condicional de la variable respuesta como función de la covariables incorporadas en el predictor lineal así como

de parámetros específicos a cada conglomerado de datos.

Así la modelación de datos experimentales en el marco teórico de los modelos lineales mixtos brinda la posibilidad de analizar datos con estructuras de correlación o dependencia. Además, por el método de estimación que se usa y el relajamiento del supuesto de independencia, no es necesario que el conjunto de datos sea balanceado en cuanto a cantidad de niveles y observaciones por nivel de los factores que clasifican los datos. Ambos aspectos, la posibilidad de modelar datos correlacionados y desbalanceados, representan beneficios que se pueden obtener con el uso de modelos mixtos. Además, en numerosas situaciones se incrementa la precisión de las estimaciones de los efectos fijos. En otras situaciones, el hecho de contemplar mejor el muestreo y la estructura jerárquica de los datos permite ampliar el espacio de inferencia.

Otro caso menos abordado en la literatura, es aquel donde dos o más variables son medidas sobre una misma unidad de análisis (sujeto), las correlaciones entre ambas variables también afectan el supuesto de independencia cuando se desea analizar un análisis simultáneo de las medidas a una misma unidad. En este contexto es al que nos referimos a modelos multivariados.

## **1.1. Problema motivador**

En esta tesis se trabajó con la modelación estadística sobre una base de datos consistente en registros de EEG, que fueron tomados como parte de un proyecto de investigación para una tesis doctoral en Neurociencia de la UNC, durante 3 años (2017-2020). Para esta investigación se reclutaron 30 individuos con ciertos criterios de inclusión, entre ellos, personas sanas entre

20 y 50 años de edad. Cada participante tuvo dos registros electroencefalográficos con 30 días o más entre los dos registros. Los EEG se realizaron según el sistema internacional 10-20 IFNC (Federación Internacional de Clínica Neurofisiológica), con 20 electrodos colocados en el cuero cabelludo sobre determinados canales de registro. La duración de los registros de EEG fue de 60 minutos. Para esa investigación se cortaron 4 segmentos de registros EEG con 3 minutos de duración, con una frecuencia de muestreo de 65 Hz. Cada segmento corresponde a diferentes condiciones electrofisiológicas, por ejemplo: ojos cerrados, ojos abiertos, hiperventilación y optoestimulación. En el diseño experimental post hoc, las 20 canales de cada uno de los 4 segmentos de los registros se cuantifican con métricas de teoría de la información. Para cada canal se obtuvieron un valor único de entropía de permutación y complejidad de permutación del Lempel Ziv. Las dimensiones de la incrustación y del retardo fueron  $d = 5$  y  $\tau = 1$ , respectivamente. Se usaron los mismos parámetros para cuantificar todas las canales en cada condición electrofisiológica. Para la aplicación de los modelos lineales mixtos multivariados para datos longitudinales, sobre experimentos en neurociencias, se utilizó el software libre R.

Se desea indagar sobre las siguientes preguntas de interés, las cuales serán reformuladas como hipótesis estadísticas:

1. ¿Qué cambios dinámicos se pueden diferenciar en la función cerebral producida por distintos estados electrofisiológicos cuando han sido suprimidos los canales  $F1$ ,  $F2$ ,  $F3$ ,  $F4$  y  $Fz$ ?. Se omiten las cinco canales  $F1$ ,  $F2$ ,  $F3$ ,  $F4$  y  $Fz$ , para evitar la alteración de la señal de EEG

inducida por el parpadeo.

2. ¿Cuál es la diferenciación de los estados electrofisiológicos respecto de cada canal cerebral?

## 1.2. Objetivos

El objetivo general del presente trabajo es desarrollar modelos estadísticos multivariados utilizando dos o más indicadores de la teoría de información en forma conjunta con la finalidad de mejorar la identificación de cambios en sistemas complejos.

Los objetivos específicos son:

Seleccionar desde un conjunto de modelos estadísticos alternativos, el modelo que presente mejor ajuste respecto del análisis de criterios de información, prueba de cociente de verosimilitud y análisis de los residuos.

Evaluar los modelos estadísticos multivariados y univariados en relación a la probabilidad estimada de los errores inherentes a las pruebas de hipótesis estadísticas planteadas y determinar si hay ganancia de información en los modelos multivariados, respecto a los modelos univariados.

Identificar y estimar los modelos más potentes para detectar cambios en la evaluación de EEG en el marco de experimentos en Neurociencia.

Generar protocolos de análisis estadístico-computacional que puedan ser aplicados para fortalecer la experimentación de base cuantitativa en Neurociencia.

# Capítulo 2

## Marco teórico

### 2.1. Desde la neurociencia

Desde la perspectiva de la clínica neurofisiológica, el electroencefalograma (EEG) es un instrumento de observación y monitoreo de la actividad electrofisiológica del cerebro. Tiene las ventajas de ser de bajo costo, posible replicar y no implica mayor riesgo para los pacientes. Se puede usar en cualquier rango de edad, desde bebés a adultos mayores (Janáčková et.al., 2016), además de ser muy sensibles a los cambios temporales en la electricidad dinámica de la corteza cerebral. En la práctica clínica, la lectura y comparación de las grabaciones electroencefalográficas son a veces insuficientes y toman demasiado tiempo. Herramientas procedentes de la teoría de la información como la entropía de permutación (HPE) (Bandt y Pompe, 2002) y la complejidad de permutación Lempel-Ziv (Zozor et.al, 2014) son herramientas usadas para abordar el análisis de datos provenientes de los EEG. Ambas métricas han sido utilizadas en el análisis de señales de EEG para obtener información

relevante que no es posible obtener por otros métodos convencionales como espectro de potencia (Jung et.al, 1997; Adeli et.al, 2003 ). Este tipo de análisis de representación se ha utilizado en caos versus análisis de ruido, análisis de autoría de texto y evolución electrofisiológica de EEG (Roso et.al, 2007; Rosso et.al; 2009, Montani et.al, 2014 ).

El análisis conjunto de ambas métricas permite recopilar información que no es posible obtener cuando se analizan por separado (Mateos et.al, 2017). La rápida, abundante y creciente disponibilidad de información relacionada con experimentos en Neurociencia, supera ampliamente la capacidad que tienen los algoritmos basados en medidas de distancia y demanda el desarrollo de nuevas herramientas para el análisis estadístico eficiente de los datos generados. Sin embargo, el análisis bivariado no se ha abordado aún desde una perspectiva de modelación estadística. Los modelos mixtos constituyen herramientas modernas para el análisis de datos correlacionados (West et al., 2015; Stroup et al. 2012) y pueden ser usados para desarrollar modelos multivariado, i.e. considerando dos o mas variables respuestas correlacionadas como las métricas de actividad cerebral antes mencionadas. Se supone que el desarrollo de herramientas estadístico computacionales basadas en el modelado de datos experimentales permitirá el contraste confiable de las hipótesis de interés en Neurociencia, obteniendo mayor información por unidad de recurso destinado a la investigación.

## 2.2. Cuantificación de los EEG

Para trabajar con los datos continuos proporcionados por un EEG, se utilizó un método de discretización de series continuas introducido por Bandt y Pompe llamado *discretización de permutación* (Band y Pompe, 2002). Este método es basado en los entornos relativos pertenecientes a las series de señales. Considerar una serie de tiempo con valores reales  $\{X_t\}_{t \geq 0}$  que se asume ser un estado de una trayectoria multivariada. Dados dos enteros  $d \geq 2$  y  $\tau \geq 1$ , una trayectoria es definida en el espacio  $d$ -dimensional como

$$Y_t^{(d,\tau)} = [X_{t-(d-1)\tau}, \dots, X_{t-\tau}, X_t], \quad t \geq (d-1)\tau \quad (2.1)$$

donde la dimensión  $d$  es llamada la *dimensión de la incrustación* y  $\tau$  el *retardo*. Se define el *vector permutación*  $\Pi_t^{(d,\tau)}$  correspondiente a la trayectoria  $Y_t^{(d,\tau)}$ , como el vector donde cada componente está dada por el valor relativo de la componente en el vector  $Y_t^{(d,\tau)}$ . La cantidad total de posibles vectores de permutaciones ( $\Pi_t^{(d,\tau)}$ ) está dado por el factorial de la dimensión de la incrustación  $d!$ . Para aclarar estas definiciones, se mostrará un ejemplo. Supongamos que se tiene una serie continua de valores  $\{X_t = 0.45, 0.87, 1.3, 0.35, 7.5\}$  y tomamos los parámetros  $d = 3$  y  $\tau = 1$ . Luego los vectores de trayectoria son definidos como  $Y_1^{(3,1)} = (0.45, 0.87, 1.3)$ ;  $Y_2^{(3,1)} = (0.87, 1.3, 0.35)$ ;  $Y_3^{(3,1)} = (1.3, 0.35, 7.5)$ , y los vectores de permutación respectivos son  $\Pi_1^{(3,1)} = (0, 1, 2)$ ;  $\Pi_2^{(3,1)} = (1, 2, 0)$ ;  $\Pi_3^{(3,1)} = (1, 0, 2)$ .



### 2.2.1. Entropía de Permutación

Bandt y Pompe definen la *entropía de permutación*(HPE) como la entropía de Shannon de la distribución de probabilidad de los vectores de permutación  $P_{\Pi_i} = Pr(\Pi_i^{(d,\tau)})$  para  $i = 1, \dots, d!$  (Bandt y Pompe, 2002). La entropía de permutación normalizada es definida como

$$h_{pe} = -\frac{1}{d!} \sum_{i=1}^{d!} P_{\Pi_i} \ln(P_{\Pi_i}).$$

La *HPE* captura la dinámica del proceso bajo los datos subyacentes. Más precisamente, la idea detrás de la definición de *HPE* es que los vectores de permutación pueden no tener la misma probabilidad de ocurrencia, y por lo tanto esta probabilidad proveer información sobre los datos subyacentes.

### 2.2.2. Complejidad de Lempel-Ziv

Corresponde a una nueva forma de analizar una secuencia, la cual no es basada en la probabilidad de los símbolos sino en como esos símbolos son repetidos a lo largo de la sucesión. Esto está basado en la idea de la complejidad de Kolmogorov, cual es la mínima “información” contenida en la secuencia (Cover et.al, 2006). Lempel y Ziv desarrollaron un método para trabajar con la complejidad de una secuencia restringiendo esta noción a los “programas” basados en sólo dos operaciones recursivas : copiar y pegar. Consideremos una secuencia dada de tamaño finito  $S_n = s_1, \dots, s_n$ , donde los símbolos  $s_i$  están en un alfabeto  $\mathcal{A}$  que contiene  $\alpha$  letras. Se definen las acciones *reproducción y producción* de la siguiente manera.

*Reproducción:* permite ir una secuencia extendida  $S_{1:T}$  a una secuencia  $Q_{1:N}$  vía las operaciones recursivas copiar-pegar, lo cual nos determina una secuencia  $S_{1:T+N} = S_{1:T}Q_{1:N}$ , tal que la primera letra  $Q_1$  está en  $S_{1:T}$ , por lo tanto existe  $i < T$ , tal que  $Q_1 = S_i$ , la segunda letra en  $Q_{1:T}$  es la que sigue en la secuencia extendida de tamaño  $T + 1$  i.e  $Q_2 = S_{i+1}$ , y así continuamos recursivamente. Por lo tanto  $Q_{1:N}$  es una subsecuencia de  $S_{1:T+N-1}$ .

*Producción:* la secuencia extendida  $S_{1:T+N}$  es ahora tal que  $S_{1:T+N-1}$  puede ser reproducida por  $S_{1:T}$ , por lo tanto el último símbolo de la extensión puede seguir una operación recursiva de copiar-pegar (por lo tanto ser una reproducción) o puede ser nuevo. Notar que una reproducción es una producción, pero no viceversa. Se denota una producción por  $S_{1:T} \rightarrow S_{1:N+T}$

Cualquier sucesión puede ser vista como una construcción a través de una secuencia de procesos de producción, lo cual es llamado una historia. Para el ejemplo  $\emptyset \rightarrow s_1 \rightarrow s_1s_2 \rightarrow s_1s_2s_3 \rightarrow \dots$  comenzando con el conjunto vacío tenemos  $n$  operaciones de producción, por lo tanto el tamaño de la historia es  $n$ . Sin embargo, una secuencia no tiene una única historia. Por ejemplo en el caso, en que  $s_2 = s_1$ , uno puede reducir el número de procesos de producción pues  $\emptyset \rightarrow s_1 \rightarrow s_1s_2s_3 \rightarrow \dots$ . Entonces, para una historia dada  $\mathcal{H}_i$  de la secuencia, se define por  $C_{\mathcal{H}_i}(S_n)$  el número de procesos de producción necesarios en una historia dada. Lempel-Ziv definen la *complejidad* de la secuencia como el número mínimo de procesos de producción necesarios para generarla:

$$C_{LZ}(S_n) = \min_{\mathcal{H}_i \in \{\text{historias de } S_n\}} C_{\mathcal{H}_i}(S_n)$$

Se usa el mismo alfabeto que el utilizado para la entropía de permutación,

es decir  $\mathcal{A} = \{\Pi_n^{(d,\tau)}\}$ , donde el tamaño del alfabeto es  $\alpha = d!$ .

Sorprendentemente, aunque se esta analizando una secuencia desde un punto completamente determinístico, parece que  $C_{LZ}(S_n)$  aveces contiene una noción de información en el sentido estadístico. Esto hace que surja una relación entre la complejidad de Lempel-Ziv y la entropía de Shannon (Thakor et.al, 1990; Zhang et.al 1999; Radhakrishnan, et.al, 1998). Usando esta relación en Diaz et.al, 2017, definen el *mapa de complejidad Lempel-Ziv de permutación vs mapa de entropía de permutación*, que consiste en el análisis paralelo de ambas métricas. La idea detrás de este enfoque es que el análisis de una señal por medio de estos mapas da más información que el uso separado de ellas. Un ejemplo puede ser el análisis de convulsiones con estos mapas, donde dependiendo del episodio, el cambio en la complejidad es más notable que en la entropía y viceversa . Otro ejemplo se ve en la caracterización y diferenciación de las señales caóticas y aleatorias, donde el análisis bajo el mapa muestra un mejor resultado que el uso de cada medida por separado (Mateos et.al, 2017). Por otro lado, la visualización de los datos en un gráfico entropía-complejidad permite una mejor comprensión de la resultados, especialmente para personas que no están tan familiarizadas con este tipo de análisis. Por lo tanto se pretende combinar ambas medidas en el análisis de una señal desde el modelado estadístico, para verificar si desde el análisis estadístico el uso conjunto de ellas, también nos brinda más información que el uso por separado de cada una de ellas para detectar cambios en la evaluación de EEG.

## 2.3. Divergencia de Jensen-Shanon

El análisis estadístico de secuencias simbólicas es de importancia central en varios campos de la ciencia, tales como dinámicas simbólicas (Hao, 1989 y 1991), lingüísticas (Shannon, 1949) o análisis de secuencia de ADN (Li, 1990; Hersel y Groze, 1995 y 1997; Ramensky, 2000). Una funcional que se usa para llevar a cabo estos análisis es la métrica conocida como *Divergencia de Jensen-Shanon*  $D$ , (Borokov, 1984; Wong y You, 1985; Lin, 1991; Menéndez et.al, 1997), la cual cuantifica la diferencia entre dos o más distribuciones de probabilidad y puede ser usada para comparar la composición simbólica entre diferentes secuencias. Las razones por las cuales se elige a  $D$  como una medida de divergencia entre distribuciones de probabilidades son:

1.  $D$  está relacionada con otras funcionales de la teoría de la información, tal como la entropía relativa o la divergencia de Kullback, y por lo tanto comparte sus propiedades matemáticas como su interpretabilidad intuitiva.
2.  $D$  puede ser generalizada a una distancia entre más de dos distribuciones.
3. Las distribuciones comparadas pueden ser ponderadas, lo cual permite tener en cuenta las diferentes longitudes de subsecuencias para las cuales las distribuciones de probabilidad son calculadas (Bernaola et.al, 1996)

La divergencia de Jensen-Shanon entre las distribuciones de probabilidad  $\mathbf{p}^{(1)}$  y  $\mathbf{p}^{(2)}$  con pesos  $\Pi^{(1)}$  y  $\Pi^{(2)}$  es definida por (Lin, 1991):

$$D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}] = H[\Pi^{(1)}\mathbf{p}^{(1)} + \Pi^{(2)}\mathbf{p}^{(2)}] - (\Pi^{(1)}H[\mathbf{p}^{(1)}] + \Pi^{(2)}H[\mathbf{p}^{(2)}]) \quad (2.2)$$

donde  $\mathbf{p}^{(1)} = (p_1^{(1)}, p_2^{(1)}, \dots, p_k^{(1)})$  y  $\mathbf{p}^{(2)} = (p_1^{(2)}, p_2^{(2)}, \dots, p_k^{(2)})$  denotan dos distribuciones de probabilidad satisfaciendo las restricciones usuales  $\sum_{j=1}^k p_i^{(j)} = 1$  y  $0 \leq p_i^{(j)} \leq 1$  para todo  $i = 1, 2, \dots, k$  y  $j = 1, 2$ .

$$H[\mathbf{p}] = - \sum_{i=1}^k p_i \log_2 p_i \quad (2.3)$$

denota la entropía de Shannon de la distribución de probabilidad  $\mathbf{p} = (p_1, p_2, \dots, p_k)$

### 2.3.1. Interpretación de $D$ en el marco de la estadística matemática

En esta subsección, siguiendo Grosse et al, 2002, se mostrará que  $D$  puede ser interpretada como un *cociente de log-verosimilitud*. Considerar el problema de estimar las probabilidades  $\mathbf{p} = (p_1, p_2, \dots, p_k)$  de una secuencia  $S$ , simbólica i.i.d. (los símbolos son independentientes e idénticamente distribuidos) de longitud  $N$  en la cual en cada posición un símbolo  $a_i \in \mathcal{A} = \{a_1, a_2, \dots, a_k\}$  es extraído al azar con probabilidad  $p_i$ . El principio de máxima verosimilitud sugiere elegir el vector de probabilidad  $\mathbf{p}$  el cual maximiza la verosimilitud

$$L(S|\mathbf{p}) = \prod_{i=1}^k p_i^{F_i} \quad (2.4)$$

donde  $F_i$  denota el número de ocurrencias del símbolo  $a_i$  en la secuencia  $S$ . Como el logaritmo es una función monótona estrictamente, se puede buscar equivalentemente el  $\mathbf{p}$  que maximiza  $\ln L = \sum_{i=1}^k F_i \ln p_i$ . Se deriva fácilmente usando un multiplicador de Lagrange para la restricción  $\sum_{i=1}^k p_i = 1$  que  $p_i = F_i/N$  maximiza la log-verosimilitud en  $L$ . Por lo tanto, se obtiene como máximo de la log-verosimilitud

$$\ln L_{\max} = N \sum_{i=1}^k f_i \ln f_i = N(\ln 2)H[\mathbf{f}] \quad (2.5)$$

donde  $f_i = F_i/N$  denota la frecuencia relativa de encontrar el símbolo  $a_i$  en la secuencia  $S$  de longitud  $N$ . Ahora considerar un problema más complejo de una secuencia no estacionaria  $S$  de longitud  $N$  consistiendo de  $m$  subsecuencias estacionarias  $S^{(1)}, S^{(2)}, \dots, S^{(m)}$  de longitudes  $n^{(1)}, n^{(2)}, \dots, n^{(m)}$ , donde la probabilidad  $p_i^{(j)}$  de generar el símbolo  $a_i$  en una subsecuencia  $S^{(j)}$  puede variar de subsecuencia a subsecuencia. La verosimilitud de obtener la secuencia entera  $S$  es igual al producto de las verosimilitudes de obtener las  $m$  subsecuencias  $S^{(1)}, S^{(2)}, \dots, S^{(m)}$ . Por lo tanto, el principio de máxima verosimilitud sugiere elegir para cada  $j = 1, 2, \dots, m$  el vector de probabilidad  $\mathbf{p}^{(j)} = (p_1^{(j)}, p_2^{(j)}, \dots, p_k^{(j)})$  que maximiza la verosimilitud

$$L(S^{(j)}|\mathbf{p}^{(j)}) = \prod_{i=1}^k (p_i^{(j)})^{F_i^{(j)}} \quad (2.6)$$

donde  $F_i^{(j)}$  es el número de ocurrencias del símbolo  $a_i$  en la subsecuencia  $S^{(j)}$ . Nuevamente, es fácil derivar usando  $m$  multiplicadores de Lagrange para las  $m$  restricciones  $\sum_{i=1}^k p_i^{(j)} = 1$  que,  $p_i^{(j)} = F_i^{(j)}/n^{(j)}$  maximiza la

log-verosimilitud en  $L^{(j)}$ . Por lo tanto se obtiene como máximo de la log-verosimilitud

$$\ln L_{\max}^{(j)} = n^{(j)} \sum_i f_i^{(j)} \ln f_i^{(j)} = -n^{(j)} (\ln 2) H[\mathbf{f}^{(j)}] \quad (2.7)$$

donde  $f_i^{(j)} = F_i^{(j)}/n^{(j)}$  denota la frecuencia relativa de encontrar el símbolo  $a_i$  en la subsecuencia  $S^{(j)}$  de longitud  $n^{(j)}$ . Como el problema uno de tener una secuencia es un caso especial del problema dos de tener  $m$  subsecuencias, la suma de los máximos log-verosimilitud  $\sum_{j=1}^m \ln L_{\max}^{(j)}$  no puede ser mayor que  $\ln L_{\max}$ . Dado que en el peor de los casos en el cual todas las  $m$  subsecuencias del problema dos fueran idénticas, daría el mismo log-verosimilitud como en el problema uno, Por lo tanto, la cantidad

$$\Delta L = \sum_{j=1}^m \ln L_{\max}^{(j)} - \ln L_{\max} \quad (2.8)$$

es no-negativo y  $\Delta L$  es llamado conumente como el *cociente de log-verosimilitud*. Se sigue de las ecuaciones (2.5), (2.7) y (2.8) que

$$\Delta L = N(\ln 2)D. \quad (2.9)$$

Por lo tanto, en el marco de la estadística matemática  $\Delta L$  puede ser interpretado como el crecimiento del log-verosimilitud cuando una secuencia  $S$ , en lugar de ser modelada como una secuencia generada con un vector de probabilidad simple  $\mathbf{p}$ , es modelada como una concatenación de  $m$  subsecuencias  $S^{(1)}, S^{(2)}, \dots, S^{(m)}$  generadas por los vectores de probabilidad

$\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$ . La desigualdad  $\Delta L \geq 0$  establece que cualquier partición de la secuencia original dentro de  $m$  subsecuencias incrementa la verosimilitud del segundo modelo sobre el primero. En orden a elegir la hipótesis dos ( $m$  subsecuencias) en favor de la hipótesis uno (una sola secuencia), se requiere que  $\Delta L$  sea significativamente mayor que cero. Para poder cuantificar la significancia estadística de los valores observados experimentalmente de  $\Delta L$ , se deberá usar que  $\Delta L$  converge asintóticamente, para grandes valores de  $N$ , a una  $\chi^2$  central con  $\nu = (k - 1)(m - 1)$  grados de libertad

$$\Delta L \simeq \chi_{(k-1)(m-1)}^2 \quad (2.10)$$

## 2.4. Modelos lineales mixtos

Siguiendo a West et al., 2015, en esta sección y las siguientes, se resumen los principales conceptos asociados a modelos lineales mixtos, adecuados para ser usados en situaciones de datos correlacionados con variables respuestas continuas. En la notación se utilizan las letras mayúsculas para representar variables aleatorias y matrices y letras minúsculas para referirse a realizaciones de las variables aleatorias. Un *Modelo Lineal Mixto* (Verbeke y Molenberghs 2000) ni-dimensional o multivariado satisface

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i \quad (2.11)$$

donde  $Y_i$  es el vector respuesta  $n_i$ -dimensional para el sujeto  $i$ ,  $1 \leq i \leq m$ ,  $m$  el número de sujetos,  $X_i, Z_i$  son matrices de covariables de dimensiones



$(n_i \times p)$  y  $(n_i \times q)$  respectivamente,  $\beta$  es el vector  $p$ -dimensional conteniendo los efectos fijos,  $\mu_i$  es el vector  $q$ -dimensional conteniendo los efectos aleatorios y  $\epsilon_i$  es el vector  $n_i$ -dimensional de términos de error. Además se asumen los supuestos:

- $Y_i/u_i \sim N(X_i\beta + Z_i\mu_i, \Sigma_i)$
- $u_i \sim N(0, D)$ ,
- $\epsilon_i \sim N(0, \Sigma_i)$
- $u_i, \epsilon_i$  son independientes

También se deben realizar supuestos sobre las estructuras de correlación de las matrices de covarianza  $D$  y  $\Sigma_i$ . Entre las estructuras para la matriz  $D$  más comúnmente usadas (asociadas a cada sujeto o conglomerado) en la práctica se encuentran:

- Desestructurada:

$$\text{Corr}(u_{i,j}, u_{i,j'}) = \sigma_{u_j, u_{j'}} \quad (\forall 1 \leq j, j' \leq q)$$

- Estructura diagonal:

$$\text{Corr}(u_{i,j}, u_{i,j'}) = 0 \quad (\forall j \neq j')$$

Análogamente, las estructuras de correlación para la matriz  $\Sigma_i$  más utilizadas para estos modelos para cada sujeto (conglomerado) son:

- Estructura de independencia:

$$\text{Corr}(\epsilon_{i,j}, \epsilon_{i,j'}) = 0 \quad (\forall j \neq j')$$

- Estructura intercambiable o de simetría compuesta:

$$\text{Corr}(\epsilon_{i,j}, \epsilon_{i,j'}) = \sigma_\epsilon \quad (\forall j \neq j')$$

- Desestructurada

$$\text{Corr}(\epsilon_{i,j}, \epsilon_{i,j'}) = \sigma_{\epsilon_j, \epsilon_{j'}} \quad (\forall 1 \leq j, j' \leq q)$$

Combinando todos los sujetos en un vector  $Y$ , se tiene que la representación matricial del modelo (2.11) está dada por

$$Y = X\beta + Zu + \epsilon \quad \text{con} \quad Y \sim N(X\beta, V) \quad (2.12)$$

donde  $Y$  es el vector respuesta que contiene las  $N = \sum_{i=1}^m n_i$  mediciones,  $X$  es la matriz  $N \times p$  de variables explicatorias,  $Z$  y  $V$  son matrices en bloques diagonal con bloques  $Z_i$  y  $V_i$ ,  $1 \leq i \leq m$ , respectivamente.

## 2.5. Modelo lineal marginal

La diferencia principal entre el modelo lineal marginal y el MLM propiamente dicho radica en la presencia o ausencia de efectos aleatorios. Específicamente, los efectos aleatorios se usan explícitamente en MLM para explicar

la variación entre sujetos o entre grupos y la correlación que esta variación genera, pero no se utilizan en la especificación de modelos marginales. Esta diferencia implica que con el MLM propiamente dicho es factible hacer inferencia sobre sujetos específicos, mientras que con el modelo marginal se hace inferencia sobre el promedio de la población.

La especificación general para un modelo marginal para el sujeto  $i$  es

$$Y_i = X_i\beta + \epsilon_i^* \quad \text{donde} \quad \epsilon_i^* \sim N(0, V_i^*) \quad (2.13)$$

La matrix  $X_i$  de dimensión  $n_i \times p$  es construida de la misma manera que antes siendo  $\beta$ , un vector de efectos fijos. El vector  $\epsilon_i^*$  representa un vector de términos de error marginales. Los elementos de la matriz de covarianza  $V_i^*$  son usualmente definidos por un conjunto pequeño de parámetros de covarianzas. Todas las estructuras usadas para la matrix  $\Sigma_i$  especificadas anteriormente pueden ser usadas para especificar la estructura de  $V_i^*$ .

## 2.6. Estimación de efectos fijos y predicción de efectos aleatorios en un MLM

El MLM introducido en la ecuación (2.12) implica el siguiente modelo marginal:

$$Y = X\beta + \epsilon^*, \quad \text{donde} \quad \epsilon^* \sim N(0, V) \quad (2.14)$$

y la matriz de covarianza es definida por,

$$V = \Sigma + ZGZ'$$

donde  $G = (I_m \otimes D)$ ,  $I_m$  es la matriz identidad de dimensión  $m \times m$ ,  $\otimes$  denota el producto de Kronecker y  $\Sigma$  es una matriz en bloques diagonales con bloques  $\Sigma_i, 1 \leq i \leq m$ .

Dada la matriz  $V$ , el estimador  $\beta$  se obtiene mediante *Mínimos Cuadrados Generalizados*, minimizando la siguiente función:

$$Q = (y - X\beta)'V^{-1}(y - X\beta),$$

lo que implica,

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y, \quad (2.15)$$

que además se corresponde con el estimador máximo verosímil.

En el caso de efectos aleatorios se habla de predicción y no de estimación, ya que los efectos aleatorios no son parámetros sino variables aleatorias. Hay varias maneras de obtener predictores de  $u$  que tengan la propiedad de ser el mejor predictor lineal insesgado (tener menor error cuadrático medio), una de ellas es lo que se llama *ecuaciones de modelos mixtos de Henderson*. Es un método que permite obtener el mejor estimador insesgado de  $X\beta$  y el mejor predictor lineal insesgado de  $u$ . Se obtiene maximizando la densidad conjunta de  $Y$  y  $u$  :

$$f(y, u) = f(y|u)f(u), \quad Y|u \sim N(X\beta + Zu, \Sigma), \quad u \sim N(0, G), \quad G = I_m \otimes D$$

en términos de verosimilitud se tendría

$$l(\beta, \Sigma, G) \propto -\frac{1}{2}[\ln |\Sigma| + \ln |G| + (y - X\beta - Zu)' \Sigma^{-1} (y - X\beta - Zu) + u' G^{-1} u]$$

derivando respecto a  $\beta$  y  $u$  obtenemos las siguientes ecuaciones:

$$\begin{pmatrix} X' \Sigma^{-1} X & X' \Sigma^{-1} Z \\ Z' \Sigma^{-1} X & Z' \Sigma^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X' \Sigma^{-1} y \\ Z' \Sigma^{-1} y \end{pmatrix}$$

Las soluciones a estas ecuaciones son:

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y \quad (2.16)$$

$$\hat{u} = G Z' V^{-1} (y - X \hat{\beta}) \quad (2.17)$$

donde  $V = ZGZ' + \Sigma$ . Pero  $V$  depende de los parámetros de la varianza  $G$  y  $\Sigma$ . Los métodos más comunes para la estimación de los parámetros de las matrices de covarianza son: Máxima verosimilitud (MV) o Máxima verosimilitud restringida (REML).

### Máxima verosimilitud

Como se ha visto anteriormente,  $\text{Cov}(Y) = V = ZGZ + \Sigma$ , entonces el estimador de máxima verosimilitud de  $V$  está basado en el modelo  $Y \sim N(X\beta, V)$ , por lo tanto

$$l(\beta, V) \propto -\frac{1}{2}[\ln |V| + (y - X\beta)' V^{-1} (y - X\beta)] \quad (2.18)$$

El estimador máximoverosimil de  $\beta$  coincide con (2.16). Sustituyendo esta expresión en (2.18), obtenemos el perfil de verosimilitud para  $V$  :

$$l_p(V) = \frac{1}{2}[\ln |V| + y'V^{-1}(I - X(X'V^{-1}X)^{-1}X'V^{-1})y] \quad (2.19)$$

Los estimadores de MV de los parámetros de  $V$  se obtienen maximizando esa función.

### **Máxima verosimilitud restringida**

La estimación REML es un método alternativo para estimar los parámetros de covarianza fue introducida a principio de 1970 por Patterson y Thompson (1971) como un método de estimar componentes en el diseño de bloques incompletos desbalanceados. REML frecuentemente es preferible a ML estimación, pues produce estimadores insesgados teniendo en cuenta la pérdida de los grados de libertad que resultan de estimar los efectos fijos en  $\beta$ . La idea es estimar los componentes de la varianza basándose en los residuos después de estimar los efectos fijos, es decir  $y - X\hat{\beta}$ . Esto significa que el método maximiza la *verosimilitud marginal o restringida*

$$l_R(V) = l_p(V) - \frac{1}{2} \ln |X'V^{-1}X| \quad (2.20)$$

Si el tamaño de muestra con el que se está trabajando es pequeño, REML dará mejores estimaciones que MV, en el caso de tamaño de muestra grande, no habrá prácticamente diferencia.

## 2.7. Inferencia en modelos lineales mixtos

Como vimos en la sección anterior el vector  $\hat{\beta}$  de efectos fijos es estimado por (2.16) en el cual la matriz  $V$  de covarianza es remplazada por su estimador ML o REML. Siguiendo a Molenberghs y Verbeke 2000 y asumiendo que el modelo ajustado es el apropiado, el estimador  $\hat{\beta}$  es asintóticamente distribuido como una normal cuya media es el vector  $\beta$  y matriz de covarianza la matriz inversa de información de Fisher. Pueden ser utilizados el test de Wald y el de cociente de verosimilitud para probar hipótesis sobre los parámetros estimados. Alternativamente, la prueba  $t$ -student puede ser usada para contrastar hipótesis sobre un parámetro de efecto fijo simple. La prueba  $F$  puede ser usada para contrastar hipótesis lineales sobre múltiples efectos fijos en un MLM. También hipótesis compuestas pueden ser probadas usando la formulación más general del estadístico de Wald el cual es una forma cuadrática estandarizada que se compara con la distribución chi-cuadrado.

Si estamos interesados en hacer inferencias en los parámetros de la covarianza podemos usar las pruebas de Wald y cociente de verosimilitud. La prueba de Wald debe usarse siempre y cuando los parámetros de la varianza bajo la hipótesis nula no se encuentren en la frontera del espacio paramétrico, en cuyo caso se recomienda el uso de pruebas de cociente de verosimilitud. En este caso la distribución asintótica de la prueba es una mezcla entre una  $\chi_p^2$  y una  $\chi_{p-1}^2$ , donde  $p$  es el número de parámetros de la varianza que se hacen cero bajo la hipótesis nula (Stram y Lee 1994).

## 2.8. Criterio de selección de modelos

Para la selección de modelos se usan métricas basadas en el *criterio de información*. El criterio de información penalizado provee una forma de evaluar el ajuste basado en el valor óptimo de log-verosimilitud, después de aplicar una penalización por los parámetros que son estimados en el modelo. En todos los criterios, se considera que el modelo que mejor ajusta es aquel que ofrece menores valores. El *criterio de información Akaike (AIC)* es calculado basado en  $l(\hat{\beta}, \hat{u})$ , la log-verosimilitud (ML o REML), de un modelo fijo como sigue (Akaike, 1973)

$$AIC = -2 \times l(\hat{\beta}, \hat{u}) + 2p$$

donde  $p$  representa el número total de parámetros que son estimados en el modelo tanto de efectos fijos como de efectos aleatorios. El *criterio de información de Bayes (BIC)* es también comúnmente usado y puede ser calculado como sigue:

$$BIC = -2 \times l(\hat{\beta}, \hat{\mu}) + p \times \ln(n)$$

El BIC aplica más penalizaciones para un modelo con más parámetros que el AIC, dado que multiplica el número de parámetros estimados por el logaritmo natural de  $n$ , donde  $n$  es el número total de observaciones usadas en la estimación del modelo.



## 2.9. Prueba de Hipótesis

Siguiendo a Wasserman 2005, en esta sección se introducen los conceptos relacionados con la validación de una prueba de hipótesis.

Sea  $X$  variable aleatoria con distribución  $F(\cdot, \theta)$  parametrizada por  $\theta \in \Theta$ , sean  $\Theta_0$  y  $\Theta_1$  tal que  $\Theta_0 \cup \Theta_1 = \Theta$  unión disjunta. Se dice que  $\delta : \mathbb{R} \rightarrow \{0, 1\}$ , con  $\delta(X) : \Omega \rightarrow \{0, 1\}$  es una *prueba de hipótesis* para contrastar la *hipótesis nula*  $H_0$  ( $\theta \in \Theta_0$ ) versus la *hipótesis alternativa*  $H_1$  ( $\theta \in \Theta_1$ ) si:

$$\delta(x) = \begin{cases} 1, & \text{se rechaza } H_0 \\ 0, & \text{cc.} \end{cases}$$

Se dirá que  $\delta$  es un *prueba determinística* si  $\exists R \subset \mathbb{R}$  llamada *región de rechazo* tal que se cumple:

$$R = \{x \in \mathbb{R} / \delta(X) = 1\}$$

y su complemento

$$(\mathbb{R} - R) = R^c = \{x \in \mathbb{R} / \delta(X) = 0\}$$

Hay dos tipos de errores que se pueden cometer en una prueba de hipótesis:

1. Rechazar  $H_0$  cuando es verdadera: *Error tipo I* ( $\delta(X) = 1 \mid \theta \in \Theta_0$ )
2. No rechazar  $H_0$  cuando es falsa: *Error tipo II* ( $\delta(X) = 0 \mid \theta \in \Theta_1$ )

Se llama *función de potencia de la prueba de hipótesis*  $\delta(X)$  a la función

$$\beta_\delta = P_\theta(\delta(X) = 1), \forall \theta \in \Theta$$

Se define el *tamaño o nivel de significación*  $\alpha$  de una prueba de hipótesis  $\delta(x)$  si se cumple:

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\delta(X) = 1)$$

Diremos que una prueba de hipótesis  $\delta$  es la *prueba más potente de nivel menor o igual que*  $\alpha$  si:

1.  $\sup_{\theta \in \Theta_0} \beta_\delta(\theta) \leq \alpha$  ( $\delta$  tiene nivel  $\leq \alpha$ )
2. Si  $\delta^*$  es otro test de nivel  $\leq \alpha$ , entonces  $\beta_{\delta^*}(\theta_1) \leq \beta_\delta(\theta_1)$ . Es decir la probabilidad de Error de tipo II es menor para  $\delta$ .

Una prueba de hipótesis  $\delta$  se llama *prueba uniformemente más potente (UMP) de nivel menor o igual que*  $\alpha$  si es la más potente de nivel menor o igual que  $\alpha$  para todo  $\theta \in \Theta_1$ .

Usualmente la *región de rechazo* tiene la forma

$$R = \{x \in \mathbb{R} / T(x) \geq c\}$$

donde  $T(X)$  es llamado el *estadístico de la prueba* y  $c$  el *valor crítico de la prueba*. Luego el nivel de significancia es  $\alpha = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq c)$ .

El *menor valor de significancia* con el que rechazaríamos  $H_0$  se denomina

$p$ -valor y se define como:

$$p - \text{valor} = \sup_{\theta \in \Theta_1} P_{\theta}(T(X) \geq T(x)).$$

Todas estas definiciones se generalizan para  $X = (X_1, \dots, X_n)$  vector o variable aleatoria, reemplazando  $\mathbb{R}$  por  $\mathbb{R}^n$ .

El *Test LSD (Least significant difference) de Fisher* (Fisher, 1935) es un test de comparaciones múltiples. Permite comparar las medias de los niveles de un factor después de haber rechazado la hipótesis nula de igualdad de medias mediante la técnica ANOVA.

El Test se basa en la creación de un valor común, un umbral, basado en un test de la  $t$  de Student. Se realizan todas las diferencias entre medias de los niveles. Las diferencias que estén por encima de este umbral indicarán una diferencia de medias significativa y las diferencias que estén por debajo indicarán una diferencia no significativa. Cuando el número posibles de comparaciones es elevado, la aplicación reiterada de este procedimiento, para un nivel de significación  $\alpha$  dado, puede conducir a un número grande de rechazos de la hipótesis nula aunque no existan diferencias reales. El intento de salvar el problema de los falsos rechazo justifica la introducción de otros procedimientos para comparaciones múltiples. Entre estos métodos se encuentra la *corrección de Bonferroni* (Hsu, 1996). La corrección de Bonferroni consiste en dividir el nivel de significancia  $\alpha$  entre el número de comparaciones dos a dos realizadas ( $\alpha^* = \frac{\alpha}{M}$ ,  $M$  número de comparaciones). Así los intervalos

quedarán en la forma

$$\bar{y}_{1m} - \bar{y}_{2m} \pm t_{\frac{\alpha}{2M}, n-1} \sqrt{\hat{S}^2 \left( \frac{1}{n_{1m}} + \frac{1}{n_{2m}} \right)}$$

donde  $\bar{y}_{1m}, \bar{y}_{2m}$  y  $n_{1m}, n_{2m}$  son las medias y los tamaños muestrales correspondientes a la comparación  $m$ -ésima y  $\hat{S}^2$  es la varianza residual estimada.

Todos los antecedentes que se han introducido hasta aquí, ha sido con el propósito de darle un marco teórico al problema de estudio y a los métodos que se aplicaron para abordar las preguntas de interés. A continuación retomamos con la resolución que se le dio al mismo.

# Capítulo 3

## Aplicación en la experimentación en Neurociencias

### 3.1. Análisis exploratorio

En esta sección se implementaron un conjunto de acciones de análisis exploratorio para comprender la estructura subyacente de los datos. Se obtuvieron medidas de resumen de los valores medios de las variables entropía de permutación y complejidad de permutación de Lempel-Ziv, particionadas según estados y canales (Ver Apéndice A). Se usaron diagramas de cajas para comparar las distribuciones de las variables entropía y complejidad por estado. Se observó que las medias para la variable entropía difieren aparentemente entre los estados, ubicándose de menor a mayor: HIPER, OC, OPTO y OA. Mientras que, las medias para la variable complejidad para los estados

HIPER y OC son iguales, ordenándose de menor a mayor de la siguiente manera, HIPER, OA y OPTO ( ver Figura 3.1).

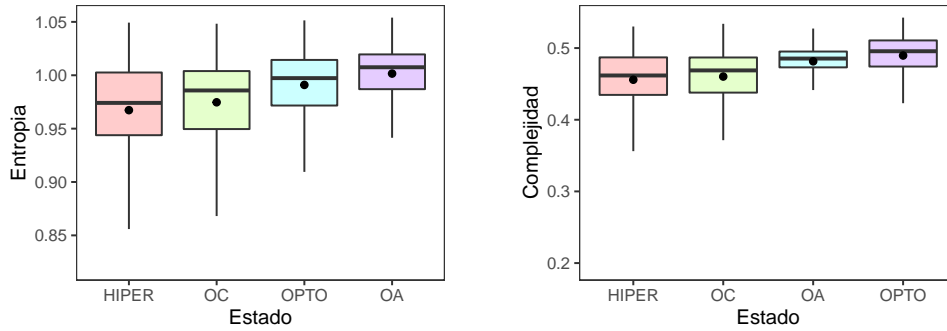


Figura 3.1: Diagrama de cajas de entropía y complejidad particionadas por los diferentes estados

Tabla 3.1: Medidas estadísticas de las variables entropía de permutación y complejidad de permutación Lempel-Ziv particionadas por estados.

Indice	Estado	Variable	n	Media	D.E.	Mín	Máx
Complejidad	HIPER	Respuesta	450	0,460	0,040	0,230	0,530
Complejidad	OA	Respuesta	450	0,480	0,030	0,190	0,530
Complejidad	OC	Respuesta	450	0,460	0,040	0,320	0,530
Complejidad	OPTO	Respuesta	450	0,490	0,030	0,210	0,540
Entropía	HIPER	Respuesta	450	0,967	0,045	0,820	1,049
Entropía	OA	Respuesta	450	1,002	0,025	0,880	1,054
Entropía	OC	Respuesta	450	0,975	0,040	0,821	1,048
Entropía	OPTO	Respuesta	450	0,991	0,033	0,857	1,051

También se obtuvieron los diagramas de cajas de las variables entropía y complejidad por individuo y de las variable entropía por individuo y por estado. La variabilidad de los índices de entropía y complejidad entre los diferentes individuos y entre los individuos y los diferentes estados, parecen diferir ampliamente. Por ejemplo el individuo 20 tiene una variabilidad 5 veces mayor que el individuo 19, tanto para la variable entropía como para la variable complejidad, ver en Figuras 3.2 , 3.3 y 3.4 (se chequeará el supuesto de de varianza constante a través de los individuos y de la combinación de individuo por estado como parte del análisis).

Figura 3.2: Diagrama de cajas de entropía particionadas por individuo

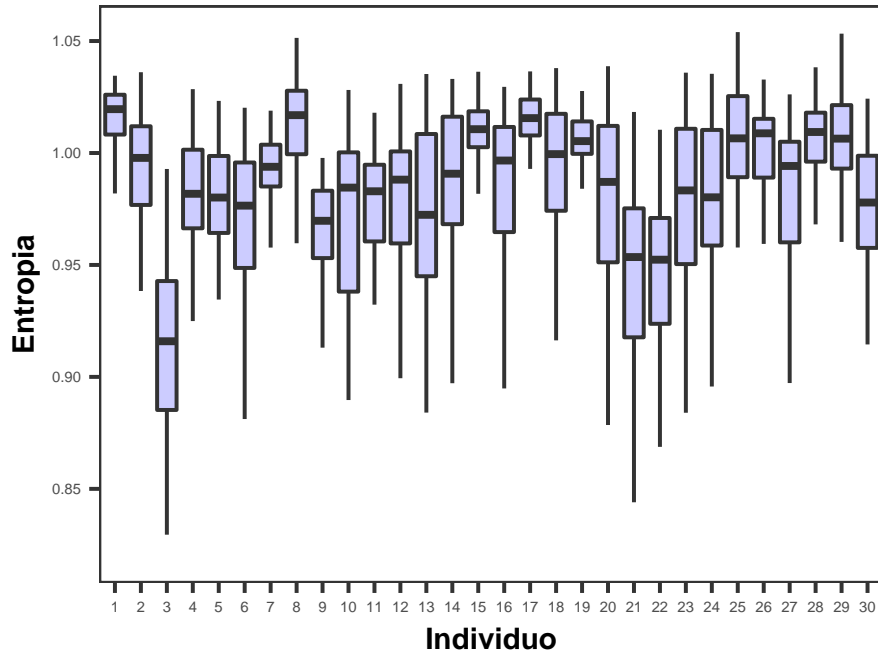
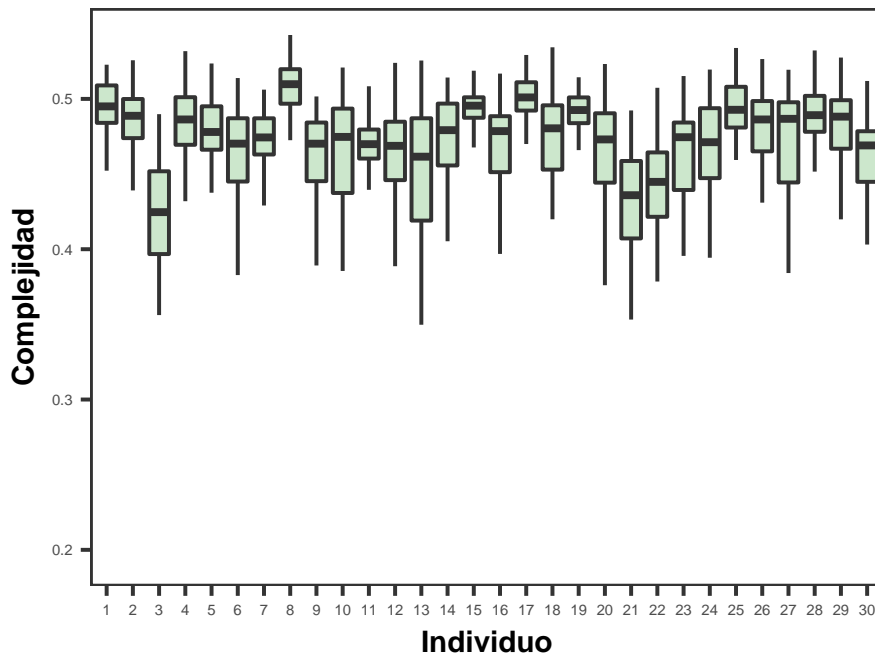


Figura 3.3: Diagrama de cajas de complejidad particionadas por individuo





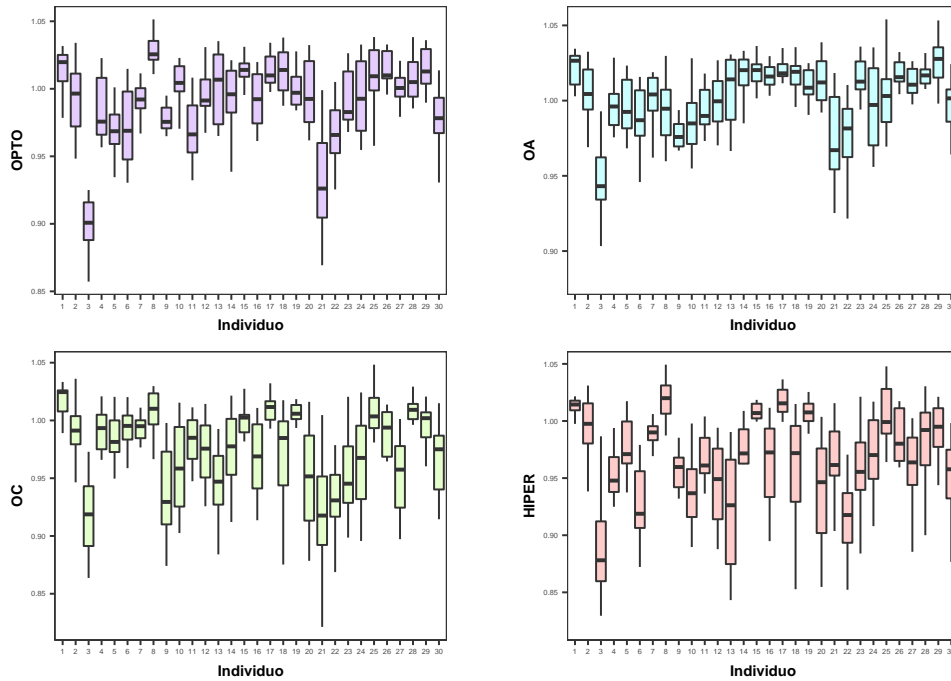


Figura 3.4: Diagrama de cajas de entropía particionadas por los diferentes estados e individuos

Además, se realizó un análisis de conglomerados para identificar grupos homogéneos de canales en base a la media de las variables respuesta univariada entropía y complejidad y de la variable respuesta bivariada entropía y complejidad. Para el análisis de cluster se utilizó el método de Ward y se analizó la reducción de variabilidad relativa tras un agrupamiento adicional para seleccionar el número de conglomerados (Hartigan, 1975). Respecto del análisis de conglomerado en base a la media de la variable bivariada entropía y complejidad, la composición de los sectores cambió levemente respecto de los conglomerados anteriores, ver en la figuras 3.5 y 3.6. Cabe destacar que del análisis de conglomerado para la variable complejidad, se obtuvo como resultado los mismos conglomerados que para la variable bivariada entropía y complejidad. Los sectores se encuentran referenciados con diferentes colo-

res correspondiendo el de color rojo al sector cuyas medias de las variables respuesta entropía (complejidad o la variable respuesta bivariada entropía y complejidad) es mayor que la media del resto de los canales y el azul el sector formado por aquellos canales cuyas medias es menor respecto de las otras medias de los restantes canales. Para analizar la distribución de la entropía y complejidad respecto de los sectores, usamos los diagramas de cajas, ver Figura3.7. Aparentemente se observa una diferencia de variabilidad entre los diferentes estados por sector para la variable respuesta entropía y complejidad. Se chequeará este supuesto en el análisis.

Canal	Medias	SectoresE
O2	0,963	1
Pz	0,967	1
O1	0,97	1
Oz	0,972	2
P4	0,97	2
Cz	0,972	2
C3	0,979	3
P3	0,976	3
C4	0,98	3
T5	0,994	4
T6	0,992	4
F7	0,999	4
T3	1,005	5
T4	1,008	5
F8	1,007	5

Figura 3.5: Sectores basados en la media del índice de entropía de permutación

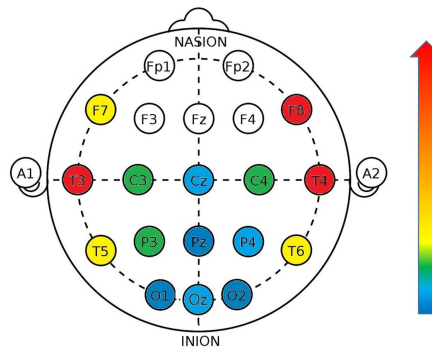


Tabla 3.2: Sectores basados en la media del índice de entropía de permutación.

Canales	Respuesta	Sectores EC
O2	(0,963; 0,455)	1
O1	(0,968; 0,457)	1
Pz	(0,967; 0,46)	1
Oz	(0,972; 0,458)	1
P3	(0,976; 0,457)	2
Cz	(0,972; 0,466)	2
P4	(0,97; 0,464)	2
C3	(0,98; 0,47)	3
C4	(0,98; 0,472)	3
T5	(0,994; 0,487)	4
T6	(0,992; 0,476)	4
F7	(1,00; 0,487)	5
T3	(1,005; 0,489)	5
T4	(1,008; 0,4493)	5
F8	(1,007; 0,486)	5

Figura 3.6: Sectores basados en la media del índice de entropía de permutación y complejidad de Lempel Ziv

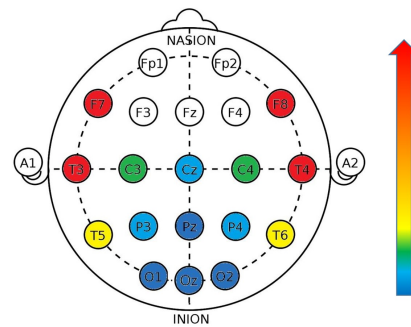


Tabla 3.3: Sectores basados en la media del índice de entropía de permutación y complejidad de permutación Lempel-Ziv.

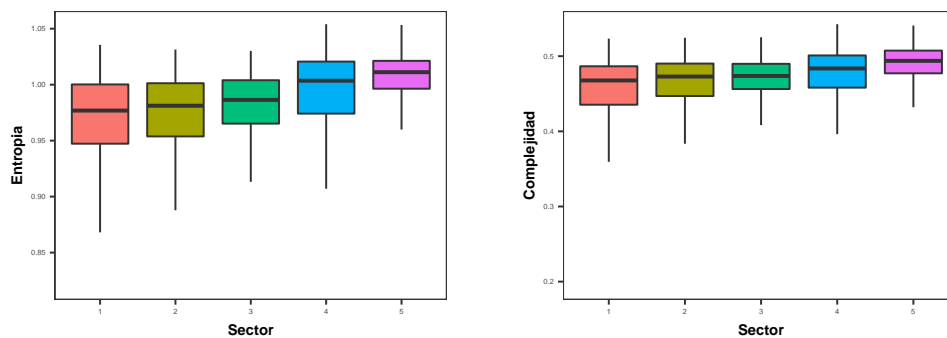


Figura 3.7: Diagrama de cajas de entropía y complejidad particionadas por sectores

Podemos concluir de lo observado en el análisis exploratorio, que hay cierta variabilidad asociada a los individuos (Figuras 3.2 y 3.3), a la combinación de estado por individuo (ver Figura 3.4) y a los sectores (ver Figura 3.7). Además, se presume del diseño de experimento que hay ciertos agrupamientos entre las observaciones dado por el efecto de los individuos que intervienen en el ensayo, la localización espacial de los canales, la naturaleza de las respuesta de acuerdo a la combinación de individuo por estado. Por lo tanto, la estructura de los datos es del tipo longitudinal agrupado, donde los grupos quedaron determinados por los individuos y los sectores, mientras con longitudinal nos referimos a las distintas condiciones (estados electrofisiológicos) en los que se encontraban los individuos durante los registros. Por lo tanto se decidió hacer uso de MLM para modelar el problema y explotar la estructura de correlación subyacente de los datos. Los MLM se utilizaron con el objeto de estimar las medias poblacionales de los distintos estados electrofisiológicos y contrastar las hipótesis de interés.

## **3.2. Investigación metodológica**

### **3.2.1. Propósitos y parametrización de modelos**

Las hipótesis de interés que fueron probadas a través de la modelación estadística son:

$$\begin{aligned}
H_0) \quad & \mu_{OPTO} = \mu_{HIPER} = \mu_{OC} = \mu_{OA} \\
H_1) \quad & \text{algún par de medias difieren}
\end{aligned} \tag{3.1}$$

donde  $\mu_{OPTO}, \mu_{HIPER}, \mu_{OC}, \mu_{OA}$  denotan la medias poblacionales de las variables entropía y complejidad para los estados electrofisiológicos optoestimulación, hiperventilación, ojos cerrados y ojos abiertos, respectivamente, sobre los individuos que se tomaron los registros electroencefalográficos y sobre los 15 canales.

$$\begin{aligned}
H_0) \quad & \mu_{OPTO,C_i} = \mu_{HIPER,C_i} = \mu_{OC,C_i} = \mu_{OA,C_i} \quad \forall i = 1, \dots, 20 \\
H_1) \quad & \text{algún par de medias difieren}
\end{aligned} \tag{3.2}$$

donde  $\mu_{OPTO,C_i}, \mu_{HIPER,C_i}, \mu_{OC,C_i}, \mu_{OA,C_i}$  denotan la medias poblacionales por canal de las variables entropía y complejidad para cada uno los cuatro estados electrofisiológicos optoestimulación, hiperventilación, ojos cerrados y ojos abiertos, respectivamente, sobre el total de pacientes que se realizó la investigación.

Con el fin de evaluar la veracidad de las hipótesis que se contrastan se ajustaron modelos lineales mixtos multivariados y univariados bajo escenarios simulados. Se llevó a cabo la siguiente parametrización de las variables:

- *EC*, variable continua que fue adoptada como variable respuesta en los MLM multivariados y sus valores representan los índices entropía y complejidad de acuerdo a la cuantificación obtenida del registro de cada

individuo en cada uno de los quince canales por cada uno de los cuatro estados electrofisiológicos. Por cada individuo hay cuatro respuesta por cada índice por cada canal, por lo tanto, el conjunto de respuestas de la base de datos que se modelan es ciento veinte veces el número de pacientes control.

- *Entropía*, variable continua que fue adoptada como variable respuesta en los MLM univariado y sus valores representan los índices de entropía.
- *Complejidad*, variable continua que fue adoptada como variable respuesta en los MLM univariado y sus valores representan los índices de complejidad.
- *Indice*, variable indicadora del tipo de respuesta que se obtuvo. Esta variable es una variable clasificatoria de dos niveles designados como *entropía* y *complejidad*.
- *Canales*, variable indicadora del tipo de canal con 20 niveles de clasificación correspondientes a cada uno de los canales de registro  $\{F1, F3, C3, P3, O1, F7, T3, T5, Fz, Cz, Pz, Oz, T6, T4, F8, O2, P4, C4, F4, F2\}$ , los cuales también serán identificados con  $C_i$  con  $i = 1, \dots, 20$ , donde la correspondencia está dada en el orden expuesto previamente.
- *Estado*, variable indicadora de los estados electrofisiológicos con cuatro niveles *OA*, *OPTO*, *HIPER*, *OC*.
- *SectoresE* variable indicadora con 5 niveles donde los sectores se encuentran definidos en la tabla (3.2),

- $(SectorE)_1 = \{O2, Pz, O1\}$
  - $(SectorE)_2 = \{Oz, P4, Cz\}$
  - $(SectorE)_3 = \{C3, P3, C4\}$
  - $(SectorE)_4 = \{F7, T5, T6\}$
  - $(SectorE)_5 = \{T3, T4, F8\}$
- *SectoresEC* variable indicadora con 5 niveles donde los sectores se encuentran definidos en la tabla 3.3,
- $(SectorEC)_1 = \{O1, Pz, O2\}$
  - $(SectorEC)_2 = \{P3, Cz, P4\}$
  - $(SectorEC)_3 = \{C3, C4\}$
  - $(SectorEC)_4 = \{T5, T6\}$
  - $(SectorEC)_5 = \{F7, T3, T4, F8\}$
- *Sexo* variable clasificatoria con 2 niveles *F* y *M* correspondientes a femenino y masculino.
- *CATEdad* variable clasificatoria con 4 niveles, donde cada nivel representa los valores comprendidos de la variable continua Edad particionada por cuartil, explícitamente las categorías quedaron determinadas como:  $C1 = [24, 28)$ ,  $C2 = [28, 34)$ ,  $C3 = [34, 45)$ ,  $C4 = [45, 50)$ .
- *Sujeto* variable clasificatoria con 30 niveles indicando cada uno de los individuos a los que les registró un EEG.

### 3.3. Procedimientos de evaluación

Como se mencionó al comienzo en la descripción de la base de datos, por cada individuo se registraron dos EEG, por lo cual se dividió la base de datos en 2 conjuntos de datos, uno que contenía las mediciones realizadas en una primer ocasión y fue utilizado como base de entrenamiento para ajustar los distintos modelos y el otro conjunto conteniendo las mediciones realizadas en una segunda ocasión que fue utilizado como base de prueba para evaluar los errores de tipo I y II.

#### 3.3.1. MLM multivariados ajustados para la diferenciación de estados electrofisiológicos

Todos los modelos multivariados que se crearon usaron las variables *Indice* y la interacción entre las variables *Estado* e *Indice* como variables explicativas principales. Hay ocho respuestas por sujeto y por canal correspondiente al tipo de índice, entropía y complejidad, y al estado electrofisiológico del paciente al momento del registro. Por lo tanto para cada sujeto  $i$  y canal  $j$  tenemos asociados dos subíndices  $l, k$ , tal que la variable respuesta  $Y_{i,j,l,k}$  adopta el valor:

- $l=1, k=1,2,3, 4$ ; índice de complejidad del sujeto  $i$  en el canal  $j$  para el estado *HIPER*, *OA*, *OC*, *OPTO*, respectivamente;
- $l=2, k=1,2,3, 4$ ; índice de entropía del sujeto  $i$  en el canal  $j$  para el estado *HIPER*, *OA*, *OC*, *OPTO*, respectivamente.



Se comenzó ajustando un modelo con efectos fijos (modelo lineal marginal), donde se adoptó como variable respuesta  $EC$  y como variables explicativas  $Indice$  y la interacción entre  $Estado$  y la variable indicadora  $Indice$ . Luego se fueron incluyendo las interacciones de las variables explicativas  $Sexo$  y  $CATEdad$  con la variable  $Indice$ . Para decidir si eliminar o no a esas variables se fue controlando que los criterios de AIC y BIC fueran disminuyendo lo cual se tomó como indicio de mejora del nuevo ajuste. El modelo marginal con mejor ajuste quedó explícitamente formulado de la siguiente manera,

$$\begin{aligned} Y_{i,j,l,k} &= a_l + b_{k,l} + x_{i,j,l,k} \cdot C + \epsilon_{i,j,l,k} \\ \epsilon_{i,j,l,k} &\simeq N(0, \sigma_\epsilon^2) \end{aligned} \quad (3.3)$$

para,  $i = 1, \dots, 30$ ;  $j = 1, \dots, 15$  (aquí se consideran todos los canales excepto los canales  $F1, F2, F3, F4, Fz$ );  $l = 1, 2$ ;  $k = 1, 2, 3, 4$ ; donde

- $a_l$  son los coeficientes de las variable indicadora  $Indice$ , el modelo asigna a priori, el coeficiente 0 para la variable  $Indice$  cuando  $l = 1$ , que corresponde a la respuesta complejidad
- $b_{k,l}$  son los coeficientes de las variables indicadoras de la interacción entre los estados  $HIPER, OA, OC, OPTO$ , para  $k = 1, 2, 3, 4$  y los índices *entropía y complejidad* para  $l = 1, 2$  respectivamente. El modelo asigna a priori el coeficiente 0 para la variable  $Estado \times Indice$  cuando  $k = 1$ , que corresponde al estado HIPER.
- $C$  vector que contiene los coeficientes de los efectos fijos correspondien-

tes al intercepto y a las variables  $Sexo \times Indice$  y  $CATEdad \times Indice$ .

- $x_{i,j,l,k}$  es la matriz de valores que asumen las variables de clasificación correspondientes a los coeficientes de  $C$  para el individuo  $i$ , en el canal  $j$ , para el índice  $l$  y el estado  $k$ .
- $\epsilon_{i,j,l,k}$  denota la variable aleatoria del error correspondiente al individuo  $i$ , canal  $j$ , índice  $l$  estado  $k$ .

En el modelo (3.3) no se definió ninguna fuente de agrupamiento en los datos; todas las observaciones son condicionalmente independientes, fijado un índice y un canal, dada la condición de intercepto, efecto del Estado, Sexo y CATEdad. Pero la experiencia nos sugiere que es probable que los sujetos esten creando fuentes de agrupamientos, quebrando la independencia condicional entre las observaciones para un sujeto dado. Por lo tanto, se amplió el modelo marginal para dar cuenta de esto, mediante la inclusión de un nuevo término de desplazamiento  $u_i$ , que mide la desviación del intercepto para el sujeto  $i$ . Además del análisis exploratorio realizado, se observó una amplia variabilidad entre sujetos, por lo tanto esto se modeló a través de la matriz de covarianza de los errores con la inclusión del supuesto de heterocedasticidad agrupada por los sujetos. Por lo tanto el modelo expandido resultó:

$$\begin{aligned}
 Y_{i,j,l,k} &= a_l + b_{k,l} + x_{i,j,l,k} \cdot C + u_i + \epsilon_{i,j,l,k} \\
 u_i &\simeq N(0, \sigma_u^2) \\
 \epsilon_{i,j,l,k} &\simeq N(0, \sigma_{\epsilon_i}^2)
 \end{aligned} \tag{3.4}$$

Si bien el modelo (3.4) es preferible al modelo (3.3) este no capta to-

das las dependencias posibles por sujeto en la muestra; la casuística también nos sugiere que los sujetos a menudo pueden variar no solo en sus respuestas generales sino también en la naturaleza de su respuesta de acuerdo al tipo de estado electrofisiológico en el que se encontraba durante los registros. Tenemos múltiples observaciones por combinación de sujeto, canal y tipo de estado, por lo que esta variabilidad en la población también creará agrupaciones en la muestra. Los  $u_i$  no capturan esta variabilidad porque solo permite que los sujetos varíen alrededor del intercepto. Para tener en cuenta esta variabilidad se agregó una pendiente aleatoria  $\rho_{i,k,l}$ . La pendiente  $\rho_{i,k,l}$  captura cuanto se desvía el efecto del sujeto  $i$  del efecto de la media del tratamiento  $k$ , para el índice  $l$ . Notar que tenemos dos efectos aleatorios y entre ellos podría haber correlación, sin embargo dada la cantidad de datos con los que se está trabajando y la limitación computacional que implica modelar diferentes estructuras de correlación para la matriz de covarianza de los efectos aleatorios, estos se supusieron independientes. Sin embargo, se ajustaron diferentes modelos alternativos, donde se fue variando la estructura de correlación de la matriz de covarianza de la pendiente aleatoria  $\rho_{i,k,l}$ , bajo la suposición de que los dos efectos aleatorios son independientes. Se utilizaron las pruebas de cociente de verosimilitud para comparar un modelo con otro y se concluyó que el modelo que mejor ajusta, es el siguiente:

$$\begin{aligned}
Y_{i,j,l,k} &= a_l + b_{k,l} + x_{i,j,l,k} \cdot C + u_i + \rho_{i,k,l} + \epsilon_{i,j,l,k} \\
\begin{pmatrix} \mu_i \\ \rho_{i,k,l} \end{pmatrix} &\simeq N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_\rho^2 \end{pmatrix} \right) \\
\epsilon_{i,j,l,k} &\simeq N(0, \sigma_{\epsilon_i}^2)
\end{aligned} \tag{3.5}$$

Finalmente se incorporó la variabilidad asociada al efecto dado por la interacción  $SectorEC \times Indice$  con la introducción de un nuevo término  $\gamma_{r,l}$ , expandiendo el modelo anterior a un modelo con tres efectos aleatorios, bajo el supuesto que estos efectos son independientes,

$$\begin{aligned}
Y_{i,j,l,k} &= a_l + b_{k,l} + x_{i,j,l,k} \cdot C + u_i + \sum_{r=1}^5 \delta_{(SectorEC)_r}(C_j) \gamma_{r,l} \\
&\quad + \rho_{i,k,l} + \epsilon_{i,j,l,k} \\
\begin{pmatrix} \mu_i \\ \rho_{i,k,l} \\ \gamma_{r,l} \end{pmatrix} &\simeq N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_\rho^2 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{pmatrix} \right) \\
\epsilon_{i,j,l,k} &\simeq N(0, \sigma_{\epsilon_i}^2)
\end{aligned} \tag{3.6}$$

donde  $\delta_{(SectorEC)_r}$  es la función **delta de kronecker** definida por

$$\delta_{(SectorEC)_r}(C_k) = \left\{ \begin{array}{l} 1, \text{ si } C_j \in (SectorEC)_r \\ 0, \text{ cc.} \end{array} \right\}.$$

Los resultados referentes a las componentes de varianza se encuentran en

las tablas 3.4 y 3.5. Del análisis de la prueba del cociente de verosimilitud entre el modelo (3.5) y el modelo (3.6) y el análisis de los residuos, concluimos que el modelo (3.6) mejora el ajuste de todos los modelos anteriores. Además tiene la estructura máxima de efectos aleatorios justificada por nuestro diseño experimental. Por lo tanto, este modelo es el que se selecciona para probar la veracidad de la hipótesis (3.1).

Tabla 3.4: Parámetros de covarianza estimados por un modelo lineal mixto multivariado 3.6

<b>Parámetro de la covarianza</b>	<b>Criterio de clasificación</b>	<b>Desviación estándar</b>
Intercepto	<i>SectorEC × Índice</i>	0.014
Pendiente	<i>OA–entropía</i>	0.014
Pendiente	<i>OPTO–entropía</i>	0.014
Pendiente	<i>OC–entropía</i>	0.014
Pendiente	<i>HIPER–entropía</i>	0.014
Pendiente	<i>OA–entropía</i>	0.014
Pendiente	<i>OPTO–entropía</i>	0.014
Pendiente	<i>OC–entropía</i>	0.014
Pendiente	<i>HIPER–complejidad</i>	0.014
Intercepto	<i>Sujeto</i>	0.019
	<i>Residuo</i>	0.020

Tabla 3.5: Covarianza de los residuos agrupados por sujetos

Parámetro	Covarianza
1	0,02
2	0,01
3	0,022
4	0,01
5	0,011
6	0,022
7	0,023
8	0,01
9	0,02
10	0,023
11	0,01
12	0,021
13	0,031
14	0,025
15	0,008

Parámetro	Covarianza
16	0,015
17	0,008
18	0,039
19	0,007
20	0,013
21	0,028
22	0,024
23	0,022
24	0,018
25	0,013
26	0,009
27	0,02
28	0,06
29	0,021
30	0,023

Luego se ajustaron modelos univariados para la variable respuesta entropía y complejidad respectivamente, repitiendo los mismos pasos que se efectuaron para ajustar los modelos multivariados, adaptando las variables explicativas que se utilizaron para el caso multivariado al caso univariado. Se comenzó con un modelo lineal marginal y se le agregó el efecto aleatorio dado por sujeto a través de la incorporación de una ordenada aleatoria  $u_i$ , una pendiente aleatoria  $\rho_{i,k}$ , que modela la variabilidad dada por la combinación de cada *Estado* y *Sujeto* y una ordenada aleatoria  $\gamma_r$ , que modela el efecto *Sector*, donde la variable *Sector* es la variable *SectorE* cuando la variable respuesta es entropía y es la variable *SectorEC* cuando la variable respuesta es complejidad. Además los residuos se supusieron heterocedasticos, con varianza agrupada por los sujetos. El modelo quedó explícitamente formulado de la siguiente manera,

$$\begin{aligned}
Y_{i,j,k} &= b_k + x_{i,j,k} \cdot C + u_i + \sum_{r=1}^5 \delta_{(Sector)_r}(C_j) \gamma_r \\
&+ \rho_{i,k} + \epsilon_{i,j,k} \\
\begin{pmatrix} \mu_i \\ \rho_{i,k} \\ \gamma_r \end{pmatrix} &\simeq N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_\rho^2 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{pmatrix} \right) \\
\epsilon_{i,j,k} &\simeq N(0, \sigma_{\epsilon_i}^2)
\end{aligned} \tag{3.7}$$

para  $i = 1, \dots, 30$ ;  $j = 1, \dots, 15$ ;  $k = 1, 2, 3, 4$ ; donde

- $b_k$  son los coeficientes de las variables indicadoras de los estados *HIPER*, *OA*, *OC*, *OPTO*, para  $k = 1, 2, 3, 4$  respectivamente. El modelo asigna

a priori el coeficiente 0 para la variable *Estado* cuando  $k = 1$ , que corresponde al estado HIPER.

- $C$  vector que contiene los coeficientes de los efectos fijos correspondientes al intercepto y las variables explicativas *Sexo* y *CATEdad* respectivamente.
- $x_{i,j,k}$  es la matriz de valores que asumen las variables de clasificación correspondientes a los coeficientes de  $C$  para el individuo  $i$ , en el canal  $j$ , en el estado  $k$ .
- $\epsilon_{i,j,k}$  denota la variable aleatoria del error correspondiente al individuo  $i$ , canal  $j$ , estado  $k$ .

Los modelos (3.7) para las variables respuesta entropía y complejidad respectivamente, son los que se seleccionan como modelos univariados para probar la veracidad de la hipótesis (3.1).

### 3.3.2. Evaluación de modelos

Para evaluar el desempeño de los modelos ajustados anteriormente, en esta sección se calculan los errores inherentes a la prueba de hipótesis (3.1) para el modelo (3.6) y el modelo (3.7) para la variable respuesta entropía y complejidad respectivamente.

El *error de tipo I*, se define como la probabilidad de rechazar la hipótesis nula cuando esta es verdadera. La hipótesis nula  $H_0$  en la hipótesis (3.1) establece que las medias poblaciones de las variables entropía y complejidad para los cuatro estados electrofisiológicos son iguales. Por lo tanto para estimar el



error de tipo I se procedió a la simulación de la hipótesis nula una cantidad grande de veces y se contabilizó la cantidad de veces que la hipótesis nula fue rechazada sobre el número total de simulaciones realizadas. Para ello se trabajó con la base de datos considerada como base de prueba que contenía las mediciones realizadas en una segunda ocasión al mismo grupo control al que se realizó las primeras mediciones.

Se siguió el siguiente procedimiento para calcular el error de tipo I para el modelo (3.6), se repitió para el modelo (3.7) para las variables respuesta entropía y complejidad respectivamente (ver algoritmos en el Apéndice B):

1. Se particionó el conjunto de datos en cuatro subconjuntos cada uno determinado por un estado electrofisiológico, *HIPER*, *OC*, *OPTO* y *OA*;
2. Se concatenó tres veces cada uno de estos subconjuntos con la finalidad de tener una cantidad de observaciones que permita tener las repeticiones necesarias para que los modelos ajustados se puedan estimar;
3. Se escogió uno a uno los cuatro subconjuntos obtenidos en el punto anterior y se efectuó un remuestreo con reemplazo 125 veces. Esto produjo  $125 \times 4$  bases diferentes donde la variable *Estado* sólo tenía un valor posible correspondiente al subconjunto que se eligió;
4. En cada una de las 500 bases de datos obtenidas se recodificó la variable *Estado* a través de la generación de valores aleatorios mediante el uso de una distribución uniforme con valores enteros variando entre 1, 2, 3 y 4, representando a los cuatro estados electrofisiológicos respectivamente. El objetivo de este paso es simular la hipótesis nula;

5. Luego se ajustó el modelo (3.6) a cada una de las bases de datos y se aplicó el test de LSD Fisher conjuntamente con la corrección de Bonferroni;
6. Se le asoció un desacierto por cada modelo ajustado, en el que se obtuvo que al menos algún par de medias diferían entre estados electrofisiológicos. Se fueron contabilizando los desaciertos acumulados respecto de cada modelo y se dividieron por el número total de modelos ajustados y este número se tomó como la estimación del error de tipo I.

El *error de tipo II*, se define como la probabilidad de no rechazar la hipótesis nula cuando esta es falsa. Por lo tanto para estimar el error de tipo II se procedió a la simulación de la hipótesis alternativa, una cantidad grande de veces y se contabilizó la cantidad de veces que la hipótesis nula no fue rechazada sobre el número total de simulaciones realizadas. Se siguió el siguiente procedimiento para calcular el error de tipo II para el modelo (3.6), se repitió el mismo procedimiento para el modelo (3.7) para las variables respuesta entropía y complejidad respectivamente (ver algoritmos en el Apéndice B):

1. Se particionó el conjunto de datos en seis subconjuntos cada uno determinado por dos estados. La idea de escoger dos estados diferentes se fundamenta en el objetivo de simular la hipótesis alternativa aprovechando el conocimiento en que estado fue tomado cada registro;
2. Se efectuó un remuestreo con reemplazo 125 veces de las base obtenidas en el paso anterior;

3. Luego se ajustó el modelo (3.6) a cada una de las bases y se aplicó el test de LSD Fisher conjuntamente con la corrección de Bonferroni;
4. En cada uno de los modelos ajustados en los que se obtuvo la igualdad de medias entre los dos estados electrofisiológicos considerados, se apuntó como un desacierto. Se fue contabilizando los desaciertos acumulados y se dividió por el número total de modelos ajustados y este número se tomó como estimación del error de tipo II.

### 3.3.3. Resultados

De la tabla 3.6 observamos que el modelo multivariado (3.6) es más potente que el modelo (3.7) para las variables respuesta entropía y complejidad respectivamente. Con una potencia igual a  $\delta_1 = 1 - 0.15 = 0.85$  mayor que la de los modelo univariados para la respuesta entropía y complejidad  $\delta_2 = 1 - 0.16 = 0.84$ . Mientras que los modelos univariados tienen error más pequeños que el multivariado.

Tabla 3.6: Errores de tipo I y II inherentes a la pruebas de hipótesis 3.1 para los modelos (3.6)y (3.7)

	<b>modelos univariados (3.7)</b>	<b>modelo multivariado(3.6)</b>
<b>error Tipo I</b>	0	0.002
<b>error Tipo II</b>	0.16	0.15
<b>potencia</b>	0.84	0.85

En las tablas 3.7, 3.8 y 3.9 se muestran las estimaciones de las medias obtenidas por el modelo (3.6) y el modelo (3.7) para las variables respuesta entropía y complejidad respectivamente . Además en la columna Dif.Medias de esas tablas, se encuentra el resultado de aplicar el test de LSD Fisher

conjuntamente con la corrección de Bonferroni y se interpreta que las medias que tienen una letra en común no son estadísticamente diferentes para un nivel de significancia del 1 %.

Se concluye que el modelo multivariado (3.6) (Cf. tabla 3.7) distingue las medias para la variable bivariada entropía y complejidad de todos los estados. Es decir las medias de ojos abiertos y optoestimulación, ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación y ojos cerrados, optoestimulación e hiperventilación y ojos cerrados e hiperventilación son diferentes estadísticamente para un nivel de significancia del 1 %, .

Por otro lado, el modelo univariado para la respuesta entropía (3.7)(Cf. tabla 3.8), distingue las medias para la variable entropía correspondientes a los estados, ojos abiertos y optoestimulación, ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación y ojos cerrados. Mientras que las medias correspondientes a los estados ojos cerrados e hiperventilación son estadísticamente iguales.

Finalmente, el modelo univariado para la respuesta complejidad (3.7) (Cf. tabla 3.9), distingue las medias para la variable complejidad correspondientes a los estados, ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación y ojos cerrados. Mientras que las medias correspondientes a los estados ojos abiertos y optoestimulación, ojos cerrados e hiperventilación son estadísticamente iguales.

Tabla 3.7: Medias estimadas por el modelo multivariado (3.6) para la variable respuesta entropía y complejidad

<b>Estado</b>	<b>Medias</b>	<b>E.E.</b>	<b>Dif.Medias</b>
<b>OA</b>	(1,003, 0.4908)	(0.006, 0.006)	(A,A)
<b>OPTO</b>	(0,992, 0.483)	(0.006, 0.006)	(B, B)
<b>OC</b>	(0,976,0.461)	(0.006, 0.006)	(C,C)
<b>HIPER</b>	(0,968,0.457)	(0.006, 0.006)	(D,C)

Tabla 3.8: Medias estimadas por el modelo (3.7) para la variable respuesta entropía

<b>Estado</b>	<b>Medias</b>	<b>E.E.</b>	<b>Dif.Medias</b>
<b>OA</b>	1,005	0,009	A
<b>OPTO</b>	0,995	0,009	B
<b>OC</b>	0,98	0,0089	C
<b>HIPER</b>	0,973	0,009	C

Tabla 3.9: Medias estimadas por el modelo (3.7) para la variable respuesta complejidad

<b>Estado</b>	<b>Medias</b>	<b>E.E.</b>	<b>Dif.Medias</b>
<b>OPTO</b>	0,490	0,004	A
<b>OA</b>	0,483	0,004	A
<b>OC</b>	0,461	0,004	B
<b>HIPER</b>	0,457	0,004	B

### 3.3.4. Predicciones en el modelo multivariado

Una vez ajustados los modelos, estos pueden ser usados para resolver dos tipos de problemas de predicción. Uno es estimar el valor esperado de la respuesta  $Y$  para ciertos valores dados a las variables predictoras del modelo, el otro es predecir una nueva observación. En las tablas 3.10-3.12 se muestran los resultados de las predicciones para cada combinación de la variable *CATEdad*, *Indice*, *Estado* y *Sexo*, para el modelo multivariado (3.6). Se muestran los intervalos de confianza para cada valor esperado de la respuesta  $Y$  y los intervalos de predicción con un nivel del 95 % para ambos (se usa la función `bolker.ci`, ver su descripción en el anexo)

Tabla 3.10: Predicciones para el modelo (3.6)

<b>CATEdad</b>	<b>Indice</b>	<b>Estado</b>	<b>Sexo</b>	<b>pred</b>	<b>ci_l</b>	<b>ci_h</b>	<b>predint_l</b>	<b>predint_h</b>
C1	CLZ	HIPER	M	0,460	0,446	0,473	0,423	0,496
C1	CLZ	HIPER	F	0,463	0,449	0,478	0,427	0,500
C1	CLZ	OPTO	M	0,491	0,477	0,504	0,454	0,527
C1	CLZ	OPTO	F	0,494	0,480	0,509	0,458	0,531
C1	CLZ	OC	M	0,463	0,449	0,476	0,426	0,499
C1	CLZ	OC	F	0,466	0,452	0,481	0,429	0,503
C1	CLZ	OA	M	0,483	0,469	0,497	0,446	0,519
C1	CLZ	OA	F	0,487	0,472	0,501	0,450	0,524
C1	HPE	HIPER	M	0,973	0,960	0,987	0,937	1,010
C1	HPE	HIPER	F	0,981	0,966	0,995	0,944	1,018
C1	HPE	OPTO	M	0,995	0,981	1,008	0,958	1,031
C1	HPE	OPTO	F	1,002	0,987	1,016	0,965	1,039
C1	HPE	OC	M	0,980	0,966	0,993	0,943	1,016
C1	HPE	OC	F	0,987	0,972	1,002	0,950	1,024
C1	HPE	OA	M	1,005	0,991	1,019	0,968	1,041
C1	HPE	OA	F	1,012	0,998	1,027	0,975	1,049

Tabla 3.11: Predicciones para el modelo (3.6)

CATEdad	Indice	Estado	Sexo	pred	ci_l	ci_h	predint_l	predint_h
C2	CLZ	HIPER	M	0,447	0,433	0,461	0,411	0,484
C2	CLZ	HIPER	F	0,451	0,437	0,465	0,414	0,488
C2	CLZ	OPTO	M	0,478	0,464	0,492	0,442	0,515
C2	CLZ	OPTO	F	0,482	0,468	0,496	0,445	0,519
C2	CLZ	OC	M	0,450	0,436	0,464	0,414	0,487
C2	CLZ	OC	F	0,454	0,440	0,468	0,417	0,491
C2	CLZ	OA	M	0,471	0,457	0,485	0,434	0,507
C2	CLZ	OA	F	0,474	0,460	0,488	0,438	0,511
C2	HPE	HIPER	M	0,957	0,943	0,971	0,920	0,994
C2	HPE	HIPER	F	0,964	0,950	0,978	0,928	1,001
C2	HPE	OPTO	M	0,978	0,964	0,992	0,942	1,015
C2	HPE	OPTO	F	0,986	0,971	1,000	0,949	1,022
C2	HPE	OC	M	0,963	0,949	0,977	0,927	1,000
C2	HPE	OC	F	0,971	0,957	0,985	0,934	1,007
C2	HPE	OA	M	0,989	0,974	1,003	0,952	1,025
C2	HPE	OA	F	0,996	0,982	1,010	0,959	1,033
C3	CLZ	HIPER	M	0,449	0,434	0,463	0,412	0,485
C3	CLZ	HIPER	F	0,452	0,439	0,466	0,416	0,489
C3	CLZ	OPTO	M	0,480	0,465	0,494	0,443	0,516
C3	CLZ	OPTO	F	0,483	0,470	0,497	0,447	0,520
C3	CLZ	OC	M	0,452	0,437	0,466	0,415	0,488
C3	CLZ	OC	F	0,455	0,441	0,469	0,419	0,492
C3	CLZ	OA	M	0,472	0,458	0,486	0,435	0,509
C3	CLZ	OA	F	0,476	0,462	0,490	0,439	0,512
C3	HPE	HIPER	M	0,953	0,939	0,967	0,916	0,990
C3	HPE	HIPER	F	0,960	0,947	0,974	0,924	0,997
C3	HPE	OPTO	M	0,974	0,960	0,989	0,937	1,011
C3	HPE	OPTO	F	0,982	0,968	0,995	0,945	1,018
C3	HPE	OC	M	0,959	0,945	0,974	0,922	0,996
C3	HPE	OC	F	0,967	0,953	0,980	0,930	1,003
C3	HPE	OA	M	0,985	0,970	0,999	0,948	1,021
C3	HPE	OA	F	0,992	0,978	1,006	0,955	1,028

Tabla 3.12: Predicciones para el modelo (3.6)

CATEdad	Indice	Estado	Sexo	pred	ci_l	ci_h	predint_l	predint_h
C4	CLZ	HIPER	M	0,473	0,458	0,488	0,436	0,510
C4	CLZ	HIPER	F	0,476	0,462	0,491	0,440	0,513
C4	CLZ	OPTO	M	0,504	0,489	0,519	0,467	0,541
C4	CLZ	OPTO	F	0,507	0,493	0,522	0,471	0,544
C4	CLZ	OC	M	0,476	0,461	0,490	0,439	0,513
C4	CLZ	OC	F	0,479	0,465	0,494	0,443	0,516
C4	CLZ	OA	M	0,496	0,481	0,511	0,459	0,533
C4	CLZ	OA	F	0,500	0,485	0,514	0,463	0,536
C4	HPE	HIPER	M	0,982	0,968	0,997	0,945	1,019
C4	HPE	HIPER	F	0,990	0,976	1,004	0,953	1,027
C4	HPE	OPTO	M	1,004	0,989	1,018	0,967	1,041
C4	HPE	OPTO	F	1,011	0,997	1,025	0,974	1,048
C4	HPE	OC	M	0,989	0,974	1,003	0,952	1,026
C4	HPE	OC	F	0,996	0,982	1,010	0,959	1,033
C4	HPE	OA	M	1,014	0,999	1,029	0,977	1,051
C4	HPE	OA	F	1,021	1,007	1,035	0,984	1,058

### 3.4. Modelación estadística y divergencia de Jensen-Shanon para diferenciación de estados electrofisiológicos por canales

En esta subsección la diferenciación de estados electrofisiológicos por canal se realizó desde dos tipos de enfoques. El primero utilizando la modelación estadística a través de la adaptación de los modelos lineales mixtos univariados como herramienta para evaluar la veracidad de la hipótesis (3.2). El segundo tipo de análisis desde un enfoque clásico que se suele utilizar para esta tipo de problemas haciendo uso de métricas provenientes de la teoría de la información como es la divergencia de Jensen-Shanon. La diferencia entre estos dos tipos de análisis es que mientras el primero permite hacer



inferencia a nivel poblacional, el segundo sólo aporta resultados sólo sobre los individuos en estudio.

### 3.4.1. Modelos univariados

En los modelos univariados que se crearon, la variable respuesta que se adoptó es la *Entropía* y se usaron las variables *Estado*, *Canales* y la interacción entre *Estado* y *Canales* como variables explicativas principales. En estos modelos se incluyeron los 20 canales, dado que se quiere indagar sobre la diferenciación de estados electrofisiológicos en cada uno de los canales de registro. Hay ochenta respuestas por sujeto correspondiente a los cuatro estados electrofisiológicos y a los veinte canales de registro. Por lo tanto para cada sujeto  $i$  tenemos asociados dos subíndices  $j, k$ , tal que la variable respuesta  $Y_{i,j,k}$  adopta el valor:

- $k = 1, j = 1, \dots, 20$ ; índice de entropía del sujeto  $i$  para el estado *HIPER* en el canal  $j$ ;
- $k = 2, j = 1, \dots, 20$ ; índice de entropía del sujeto  $i$  para el estado *OA* en el canal  $j$ ;
- $k = 3, j = 1, \dots, 20$ ; índice de entropía del sujeto  $i$  para el estado *OC* en el canal  $j$ ;
- $k = 4, j = 1, \dots, 20$ ; índice de entropía del sujeto  $i$  para el estado *OPTO* en el canal  $j$ .

Se comenzó ajustando un modelo con efectos fijos (modelo lineal marginal) con las variables explicativas principales mencionadas anteriormente

y luego se fueron incluyendo las las variables explicativas *Sexo* y *CATEDAD*. Para decidir si eliminar o no a esas variables se fueron controlando que los criterios de AIC y BIC fueran disminuyendo lo cual se tomó como indicio de mejora del nuevo ajuste. Una vez evaluado la inclusión de todas las covariables, se le agregó al modelo que quedó determinado, una ordenada aleatoria  $u_i$  para modelar la variabilidad producida por el efecto sujeto y una pendiente aleatoria  $\rho_{i,k}$  para el efecto dado por la interacción entre *Estado* y *Sujeto*. Se supuso que los efectos aleatorios eran independentientes, por la limitación computacional que implica modelar la matriz de covarianzas, dado la cantidad de parámetros que se deben estimar con un gran número de observaciones. Además se incluyó el supuesto de heterocedasticidad de los residuos agrupados por sujetos. El modelo con mejor ajuste quedó explícitamente formulado de la siguiente manera,

$$\begin{aligned}
Y_{i,j,k} &= a_k + b_j + c_{j,k} + x_{i,j,k} \cdot C + u_i + \rho_{i,k} + \epsilon_{i,j,k} \\
\begin{pmatrix} \mu_i \\ \rho_{i,j} \end{pmatrix} &\simeq N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_\rho^2 \end{pmatrix} \right) \\
\epsilon_{i,j,k} &\simeq N(0, \sigma_{\epsilon_i}^2)
\end{aligned} \tag{3.8}$$

para  $i = 1, \dots, 30$ ;  $j = 1, \dots, 20$ ;  $k = 1, 2, 3, 4$ ; donde

- $a_j$  son los coeficiente de las variables indicadoras del tipo de *Estado*, el modelo asigna a priori, el coeficiente 0 para la variable *Estado* cuando  $k = 1$ .
- $b_k$  son los coeficientes de las variables indicadoras *Canales*. El modelo

asigna a priori el coeficiente 0 para la variable *Canales* cuando  $j = 1$ .

- $c_{j,k}$  son los coeficientes de la interacción entre las variables indicadoras *Estado* y *Canales*. El modelo asigna a priori, el coeficiente 0 para la variable *Estado*×*Canales* cuando  $j = 1$ .
- $C$  vector que contiene los coeficientes de los efectos fijos correspondientes al intercepto y las variables *Sexo* y *CATEdad* respectivamente.
- $x_{i,j,k}$  es la matriz de valores que asumen las variables de clasificación correspondientes a los coeficientes de  $C$  para el individuo  $i$ , en el estado  $j$  para el canal  $k$ .
- $\epsilon_{i,j,k}$  denota la variable aleatoria del error correspondiente al individuo  $i$ , estado  $j$ , canal  $k$ .

En las tablas 3.13- 3.15 se presentan las estimaciones de las medias de la variable entropía particionadas por estados y por canal, obtenidas por el modelo univariado (3.8) y en la columna Dif.Medias se muestran los resultados de aplicar el test DGC para un nivel de significancia del 5% conjuntamente con la corrección de Bonferroni. En esta columna, las medias con letra común se interpreta que no son significativamente diferentes. Concluimos que los canales que no distinguen las medias de los estados optoestimulación, ojos abiertos y ojos cerrados son  $Fz, F2$  y diferencian las medias de los estados optoestimulación, ojos abiertos y ojos cerrados de la media hiperventilación. Los canales que no distinguen las medias de los estados optoestimulación, ojos cerrados e hiperventilación son  $C3, C4, F7, T5, T6$  y diferencian la media del estado ojos abiertos de las medias de los estados optoestimulación,

ojos cerrados e hiperventilación. Los canales que no distinguen las medias de los estados ojos cerrados e hiperventilación son  $Pz, P4$  y diferencian las medias de los estados ojos abiertos y optoestimulación, ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación y ojos cerrados, optoestimulación e hiperventilación. Los canales que no distinguen las medias de los estados optoestimulación y ojos cerrados son  $F1, F3, P3, O1, Cz, Oz$  y distinguen las medias de los estados ojos abiertos y optoestimulación, ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación e hiperventilación. Los canales que no distinguen las medias de los estados optoestimulación y ojos abiertos, ojos cerrados e hiperventilación son  $T4, F8, T3, F4$  y distinguen las medias de los estados ojos abiertos y ojos cerrados, ojos abiertos e hiperventilación, optoestimulación y ojos cerrados, optoestimulación e hiperventilación. El canal  $O2$  es el único que distingue las medias de todos los estados.

Tabla 3.13: Ajustes de medias correspondientes al modelo (3.8)

<b>Estados</b>	<b>Canales</b>	<b>Medias</b>	<b>Desv. Estandar</b>	<b>Dif.Medias</b>
OA	01-F1	0,987	0,005	A
OPTO	01-F1	0,985	0,005	B
OC	01-F1	0,977	0,005	B
HIPER	01-F1	0,962	0,005	C
OA	02-F3	0,991	0,005	A
OPTO	02-F3	0,981	0,005	B
OC	02-F3	0,967	0,005	B
HIPER	02-F3	0,961	0,005	C
OA	03-C3	0,998	0,005	A
OPTO	03-C3	0,984	0,005	B
OC	03-C3	0,977	0,005	B
HIPER	03-C3	0,967	0,005	B
OA	04-P3	0,998	0,005	A
OPTO	04-P3	0,980	0,005	B
OC	04-P3	0,974	0,005	B
HIPER	04-P3	0,964	0,005	C
OA	05-O1	0,999	0,005	A
OPTO	05-O1	0,978	0,005	B
OC	05-O1	0,969	0,005	B
HIPER	05-O1	0,952	0,005	C
OA	06-F7	1,009	0,005	A
OPTO	06-F7	1,004	0,005	B
OC	06-F7	0,996	0,005	B
HIPER	06-F7	0,988	0,005	B

Igual letra en la columna Dif.Medias implica que, las medias no son estadísticamente diferentes

Tabla 3.14: Ajustes de medias correspondientes al modelo (3.8)

Estados	Canales	Medias	Desv. Estandar	Dif.Medias
OA	07-T3	1,014	0,005	A
OPTO	07-T3	1,007	0,005	A
OC	07-T3	0,999	0,005	B
HIPER	07-T3	0,999	0,005	B
OA	08-T5	1,012	0,005	A
OPTO	08-T5	1,001	0,005	B
OC	08-T5	0,991	0,005	B
HIPER	08-T5	0,989	0,005	B
OA	09-Fz	0,983	0,005	A
OPTO	09-Fz	0,969	0,005	A
OC	09-Fz	0,969	0,005	A
HIPER	09-Fz	0,955	0,005	B
OA	10-Cz	0,992	0,005	A
OPTO	10-Cz	0,980	0,005	B
HIPER	10-Cz	0,967	0,005	B
OC	10-Cz	0,964	0,005	C
OA	11-Pz	0,991	0,005	A
OPTO	11-Pz	0,976	0,005	B
HIPER	11-Pz	0,960	0,005	C
OC	11-Pz	0,958	0,005	C
OA	12-Oz	0,997	0,005	A
OC	12-Oz	0,972	0,005	B
OPTO	12-Oz	0,970	0,005	B
HIPER	12-Oz	0,959	0,005	C
OA	13-T6	1,016	0,005	A
OPTO	13-T6	1,001	0,005	B
OC	13-T6	0,991	0,005	B
HIPER	13-T6	0,988	0,005	B
OA	14-T4	1,018	0,005	A
OPTO	14-T4	1,009	0,005	A
OC	14-T4	1,003	0,005	B
HIPER	14-T4	1,000	0,005	B
OA	15-F8	1,016	0,005	A
OPTO	15-F8	1,015	0,005	A
HIPER	15-F8	1,004	0,005	B
OC	15-F8	1,003	0,005	B

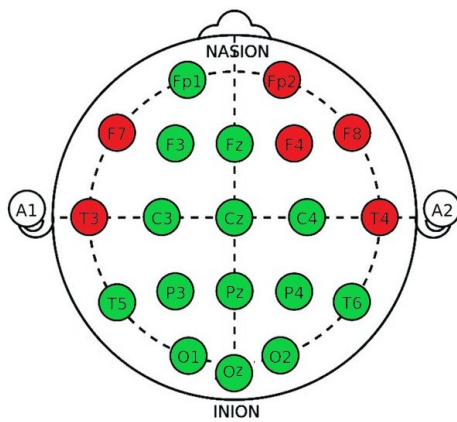
Igual letra en la columna Dif.Medias implica que, las medias no son estadísticamente diferentes

Tabla 3.15: Ajustes de medias correspondientes al modelo (3.8)

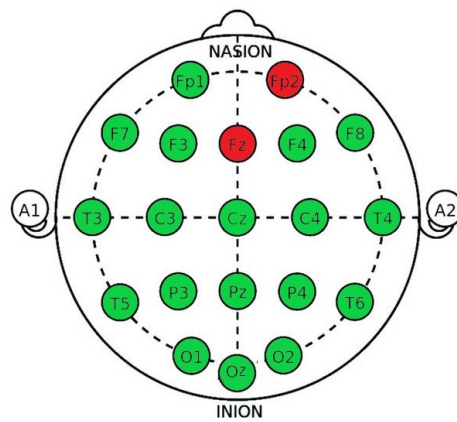
<b>Estados</b>	<b>Canales</b>	<b>Medias</b>	<b>Desv. Estandar</b>	<b>Dif.Medias</b>
OA	16-O2	0,989	0,005	A
OPTO	16-O2	0,974	0,005	B
OC	16-O2	0,956	0,005	C
HIPER	16-O2	0,947	0,005	D
OA	17-P4	0,991	0,005	A
OPTO	17-P4	0,975	0,005	B
OC	17-P4	0,962	0,005	C
HIPER	17-P4	0,960	0,005	C
OA	18-C4	0,995	0,005	A
OPTO	18-C4	0,982	0,005	B
OC	18-C4	0,971	0,005	B
HIPER	18-C4	0,969	0,005	B
OA	19-F4	1,000	0,005	A
OPTO	19-F4	0,989	0,005	A
OC	19-F4	0,979	0,005	B
HIPER	19-F4	0,976	0,005	B
OA	20-F2	0,994	0,005	A
OPTO	20-F2	0,991	0,005	A
OC	20-F2	0,990	0,005	A
HIPER	20-F2	0,975	0,005	B

Igual letra en la columna Dif.Medias implica que, las medias no son estadísticamente diferentes

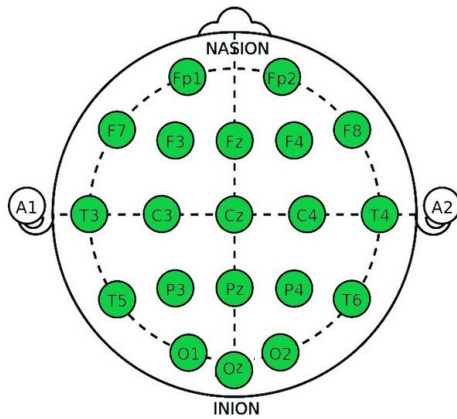
En los siguientes gráficos se plantean cada una de las seis relaciones entre las medias de la variable entropía de los diferentes estados electrofisiológicos, indicándose con color verde aquellos canales donde las medias difieren estadísticamente para un nivel de significancia del 5% y los indicados con color rojo para aquellos canales en los que no hubo evidencia suficiente para probar que las medias difieren estadísticamente.



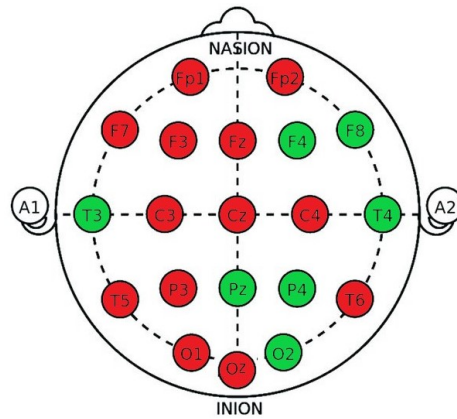
(a) Dif. de la medias OA y OPTO



(b) Dif. de la medias OA y OC



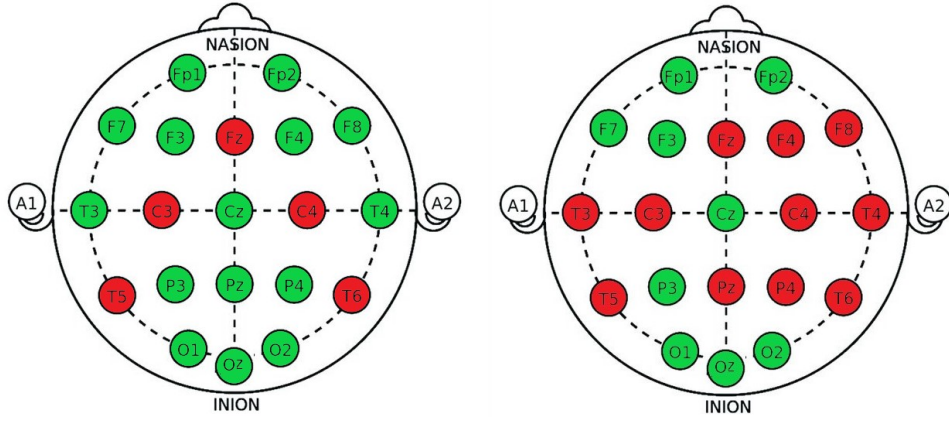
(c) Dif. de la medias OA e HIPER



(d) Dif. de la medias OPTO Y OC

El color verde indica medias difieren estadísticamente para un nivel de significancia del 5%.





(e) Dif. de la medias OPTO e HIPER      (f) Dif. de la medias OC e HIPER

El color verde indica medias difieren estadísticamente para un nivel de significancia del 5 %.

### 3.4.2. Aplicación Divergencia de Jensen-Shanon

Como se mencionó anteriormente, la divergencia de Jensen-Shanon es utilizada para verificar la diferenciación de estados electrofisiológicos en un mismo individuo respecto de cada canal. La misma fue implementada como un cociente de log-verosimilud como se describió en la Sección 2.3.1. Para cada individuo y cada canal fijo el procedimiento fue el siguiente:

- Se escogieron dos estados electrofisiológicos diferentes (esto se repitió para todas las combinaciones posibles de a dos estados) llamados Estado<sub>1</sub> y Estado<sub>2</sub>, respectivamente. Se denotaron como  $\{X_i\}_{0 \leq t \leq N_i}$  con  $i = 1, 2$ ; a las series correspondientes a los registros para los Estado<sub>1</sub> y Estado<sub>2</sub> para cada individuo y canal fijados al comienzo del algoritmo.
- Se determinó la dimensión de la incrustación  $d$  y el retardo  $\tau$ . De acuerdo a la longitud de las series que se disponen, la literatura sugiere que

un  $d = 5$  y  $\tau = 1$  son apropiados.

- Se determinaron los vectores de trayectoria  $\{Y_{it}^{(5,1)}\}_{(d-1)\tau \leq t \leq N_i - (d-1)\tau}$  con  $i = 1, 2$  (ver ecuación (2.1)) correspondientes a los estados electrofisiológicos Estado<sub>1</sub> y Estado<sub>2</sub> respectivamente.
- A cada una de las trayectorias determinadas en el punto anterior se les asoció los vectores de permutación correspondientes, denotados como  $\Pi_{it}^{(5,1)}$ ,  $i = 1, 2$ . (Ver Sección 2.2 Cuantificación de los datos). Los símbolos  $\Pi_{it}^{(d,\tau)}$  están en un alfabeto  $\mathcal{A} = \{\Pi_j^{(d,\tau)}\}$  donde el tamaño del alfabeto es  $d! = 120$ .
- Cada conjunto de vectores de permutación es representado en una secuencia simbólica de acuerdo al orden de aparición en la trayectoria,  $S^{(i)} = (\Pi_{i(d-1)\tau}^{(5,1)}, \Pi_{i((d-1)\tau+1)}^{(5,1)}, \dots, \Pi_{i(N_i - (d-1)\tau)}^{(5,1)})$  con  $i = 1, 2$ .
- Se calculan las entropías de Shanon  $H[f^{(1)}]$  y  $H[f^{(2)}]$  de las series  $S^{(1)}$  y  $S^{(2)}$  respectivamente (dadas por la fórmula (2.3)). Donde  $f^{(i)} = (f_1^{(i)}, f_2^{(i)}, \dots, f_{d!}^{(i)})$  tal que  $f_j^{(i)} = F_j^{(i)}/n^{(i)}$  con  $F_j^{(i)}$  la frecuencia de aparición del vector de permutación  $\Pi_j^{(d,\tau)}$  en la secuencia  $i$  y  $n^{(i)} = N_i - (d-1)\tau$  con  $i = 1, 2$ .
- Se concatenan las secuencias simbólicas  $S^{(1)}$  y  $S^{(2)}$  (i.e  $S^{(1)}$  seguida de  $S^{(2)}$ ) formando una nueva secuencia simbólica  $S = S^{(1)}S^{(2)}$  de longitud  $N = N_1 + N_2 - 2(d-1)\tau$ .
- Se calcula la divergencia de Jensen-Shannon  $H[f]$  de la secuencia simbólica  $S$ , donde  $f = (f_1, f_2, \dots, f_{d!})$ ,  $f_j = F_j/N = \Pi_1(F_j^{(1)}/n^{(1)}) + \Pi_2(F_j^{(2)}/n^{(2)})$ .

Tal que  $F_j$  es la frecuencia de aparición del símbolo  $\Pi_j^{(d,\tau)}$  en la secuencia  $S$  y  $\Pi_i = n^{(i)}/N$  con  $i = 1, 2$ .

- Se calcula el radio de log-verosimilitud  $\Delta L = N(\ln 2)D$  con  $D$  dada por la fórmula (2.2), i.e

$$D[\mathbf{f}^{(1)}, \mathbf{f}^{(2)}] = H[\mathbf{f}] - (\Pi^{(1)}H[\mathbf{f}^{(1)}] + \Pi^{(2)}H[\mathbf{f}^{(2)}])$$

Luego el problema de interés sobre el que se quiere indagar se plantea como las dos secuencias  $S^{(1)}$  y  $S^{(2)}$  corresponden a los mismos estados electrofisiológicos o las dos secuencias corresponden a dos estados electrofisiológicos diferentes. Esto puede ser estadísticamente interpretado como que las dos secuencias son generadas por el mismo vector de probabilidad  $p^{(1)} = p^{(2)}$  (que es equivalente a modelar la secuencia  $S$  con un vector de probabilidad simple  $p = \Pi_1 p^{(1)} + \Pi_2 p^{(2)}$ ) o  $S$  es modelada como concatenación de dos subsecuencias  $S^{(1)}, S^{(2)}$  generadas por los vectores de probabilidad  $p^{(1)}, p^{(2)}$  respectivamente. Por lo tanto la hipótesis estadística queda planteada como:

$$H_0) \quad \Delta L = 0$$

$$H_1) \quad \Delta L > 0$$

donde  $\Delta L$  converge asintóticamente para grandes valores de  $N$  a una  $\chi^2$  central con  $\nu = (d! - 1) = 119$  grados de libertad. Luego la zona de rechazo, para un nivel de significancia  $\alpha = 0.01$  se define como  $R = \{\Delta L \in \mathbb{R}_{\geq 0} : \Delta L > k^*\}$  donde  $k^* = P(X < k^*) = 0,99, X \sim \chi_{119}^2$ . Por lo tanto la *zona de rechazo* es

$$R = \{\Delta L \in \mathbb{R}_{\geq 0} : \Delta L > 157,79\}$$

La tabla 3.16 muestra el mínimo de los  $\Delta L$  obtenido por paciente respecto de todos los canales por cada combinación de dos estados electrofisiológicos. Se observa que el mínimo por canal para cada uno de los pacientes es un resultado superior a 157,79, valor crítico de la zona de rechazo. Por lo tanto, como el mínimo por canal es mayor que este valor, en particular todos los valores del resto de los canales también van a ser más grande que el valor crítico por definición de mínimo. Es decir, para cada paciente del grupo control, cada canal y cada combinación de estados electrofisiológicos dados, el radio de log-verosimilitud  $\Delta L$  cae en zona de rechazo del test. Entonces, podemos concluir que para cada uno de los pacientes de estudio y cada uno de los canales hay evidencia suficiente para afirmar que los estados electrofisiológicos no están generados por los mismos vectores de probabilidad. Es decir que todos los estados se pueden diferenciar por canal y por sujeto.

Tabla 3.16: Mínimo de  $\Delta L$  para cada Sujeto (Id) entre todos los canales para cada combinación de a dos de estados fisiológico.

<b>Id</b>	<b>OA OPTO</b>	<b>OA HIPER</b>	<b>OA OC</b>	<b>OPTO HIPER</b>	<b>OPTO OC</b>	<b>OC HIPER</b>
1	351,29	527,18	336,26	347,42	265,53	485,75
2	403,16	393,85	384,66	376,40	378,30	395,74
3	416,57	369,78	313,96	286,58	313,61	367,20
4	338,67	656,94	328,90	419,83	341,50	561,69
5	353,19	304,56	392,03	348,17	384,05	315,91
6	351,00	599,62	314,42	368,04	357,95	648,30
7	223,30	505,95	462,61	365,71	313,50	280,62
8	517,10	402,58	363,05	394,30	380,38	456,09
9	288,68	366,73	535,57	334,70	443,81	368,50
10	309,48	469,09	392,67	589,76	645,66	327,77
11	490,72	447,76	509,74	349,64	276,55	349,66
12	391,87	467,82	444,18	452,24	393,25	407,40
13	368,67	643,70	758,41	525,08	462,16	366,54
14	367,37	382,40	408,12	408,59	375,40	339,68
15	403,42	396,16	412,00	323,92	358,66	344,13
16	1032,30	2723,15	2314,39	385,47	488,84	315,59
17	448,96	416,38	429,03	335,64	368,02	410,71
18	417,12	540,63	617,16	479,92	423,11	309,08
19	375,02	358,39	415,11	378,57	436,06	375,59
20	372,92	952,50	766,21	627,69	514,45	272,32
21	462,91	417,33	395,34	454,45	309,27	405,87
22	321,11	647,60	420,87	371,86	321,44	351,24
23	458,48	644,83	525,95	397,51	376,87	386,86
24	406,71	386,65	532,10	351,25	425,09	409,27
25	561,66	568,82	609,62	382,87	412,37	402,44
26	406,85	528,21	576,99	377,46	387,75	345,67
27	343,65	563,42	675,05	454,14	580,26	400,35
28	411,30	469,64	505,13	423,01	342,28	416,63
29	454,12	669,59	553,34	476,80	374,82	371,07
30	436,70	506,50	542,67	388,54	364,33	349,67

# Capítulo 4

## 4.1. Comentarios finales

En el tipo de contexto dado como en el caso usado para ilustración de este trabajo, en el que se desea comparar las medias poblaciones para cuatro tratamientos, estados electrofisiológicos, en el cual los participantes (sujetos en estudios) son tratados como un efecto aleatorio, se podría haber aplicado un modelo mixto ANOVA. Pero en este tipo de experimentos en Neurociencia, donde los sujetos a menudo pueden variar no solo sus respuestas generales sino también en la naturaleza de su respuesta de acuerdo a la condición electrofisiológica en la que se encontraban durante el registro, se quiebra el supuesto de independencia condicional subyacente a un modelo mixto ANOVA, el cual considera a los participantes del experimento como el único factor aleatorio. Cada fuente de variación aleatoria que no es contabilizada tenderá a trabajar en contra de lo que se pretende probar. Por un lado, la variación no contabilizada que es ortogonal a nuestro efecto de interés (por ejemplo, la un coeficiente de intercepto aleatorio) tenderá a reducir la potencia de las pruebas. Por el otro lado, la variación que no es tenida en cuenta que se confunde

con nuestro efecto de interés (por ejemplo, una pendiente aleatoria) puede generar diferencia entre medias y por lo tanto aumentar el riesgo de error de tipo I. Por estas razones se propuso trabajar con los datos a través de los modelos lineales mixtos multivariados y explotar la correlación subyacente de los datos. Los MLM tienen numerosas ventajas. Son más flexibles en el sentido de que permiten a los investigadores analizar los efectos de variables continuas dentro y entre sujetos. Tienen la capacidad de incorporar datos faltantes directamente (es decir, no es necesario eliminar los casos incompletos o imputar los valores faltantes). Pueden dar cuenta para múltiples fuentes de no independencia (por ejemplo, cuando los sujetos reaccionan al mismo conjunto de elementos).

En este trabajo, se utilizó un conjunto de datos como ejemplo, consistente en registros de EEG que fueron tomados como parte de un trabajo de doctorado en Neurociencia de la UNC. Para esta investigación 30 individuos fueron reclutados para formar un grupo control. Cada participante tuvo dos electroencefalogramas con 30 días o más entre la grabación de cada uno de ellos. Los EEG fueron realizados de acuerdo al sistema internacional 10-20 IFNC con 20 electrodos en el cuero cabelludo sobre determinados canales de registro. La duración de los EEG fue de 60 minutos, durante el cual el individuo experimentó cuatro estados electrofisiológicos: ojos cerrados, abiertos, optoestimulación e hiperventilación. Las 20 señales de cada uno de los 4 segmentos de los registros (estados electrofisiológicos) se cuantificaron con métricas de teoría de la información, entropía de permutación, HPE y complejidad de Lempel Ziv, CLZ. Por otro lado se encuentra documentado en Diaz et. al 2017, que el uso conjunto de ambas métricas, da más información

que por separado. Por lo tanto se decidió ajustar modelos multivariados mixtos considerando como variable respuesta a ambas métricas con la finalidad de mejorar la identificación de cambios en sistemas complejos.

El software utilizado fue R, se ajustaron modelos multivariados donde la variable respuesta considerada fue entropía y complejidad y modelos univariados donde se consideró cada uno de los índices por separado. La especificación de la estructura de la parte aleatoria de los MLM ajustados, fue determinada por el diseño de experimento. Se modeló la variabilidad generada por el efecto sujeto, la de los canales y la combinación de estado por sujeto a través de la incorporación de ordenadas y pendientes aleatorias y se fue testeando si era necesario o no su incorporación a través de las prueba de cociente de verosimilitud. Además se modeló la heterocedasticidad de los errores agrupada por los sujetos. Se seleccionaron los modelos cuya estructura contenía todos los efectos aleatorios considerados anteriormente y la misma resultó maximal para el diseño de experimento. El modelo multivariado seleccionado encontró que las medias de los cuatro estados se diferenciaban entre sí, mientras que el modelo univariado para la respuesta entropía no pudo distinguir entre las medias correspondientes a los estados de hiperventilación y ojos cerrados y el modelo univariado para la respuesta complejidad no pudo distinguir entre las medias correspondientes a los estados de ojos abiertos y optoestimulación e hiperventilación y ojos cerrados

Se estimó el desempeño de los modelos que presentaron mejor ajuste a través de los errores de tipo I y II. Estos permitieron evaluar la confiabilidad de los modelos para contrastar las hipótesis de interés. El modelo que



presentó mayor potencia fue el modelo multivariado con una potencia del 85 % mayor que el modelos univariados del 84 % para la respuesta entropía y complejidad. Sin embargo el modelo univariado presento un error de tipo I menor. Se puede concluir que el modelo multivariado identifica mejor las diferencias de estados que los univariados. Una de las ventajas de haber tratado el problema desde la estadística, es el poder contar con resultados estadísticos sobre estas métricas en un grupo control, lo que permitirá generar protocolos de análisis, que podrán ser aplicados a cualquier paciente de interés que se pretenda analizar a futuro. Un condicionamiento que presenta el modelado en este tipo de experimento es la limitación computacional, para poder plantear diferentes escenarios de modelado de la estructura de correlación de la matriz de covarianza de los efectos aleatorios, como así también la incorporación de diferentes efectos fijos, dado por la gran cantidad de observaciones que se disponen. Como línea futura, apuntando al mejoramiento de estos modelos se podría mejorar el poder de computo del equipo. Además, se podría agregar una tercer métrica como la información de fisher para estudiar la variabilidad de la dinámica del sistema. Por otro lado, siguiendo en la misma línea, se podrían combinar estados electrofisiológicos y cognitivos para evaluar si esto genera cambios en los valores de la señales y a su vez en las medidas de información y si esa variación dentro de los estados electrofisiológicos es posible detectar la variación de los estados cognitivos previa al estímulo y post al estímulo. También otra propuesta interesante, sería mirar desde otro plano los estados electrofisiológicos, como variable respuesta en lugar de efecto fijo y abordar desde la estadística multivariada con métodos de clasificación apropiados para modelar este tipo de respuesta politómicas. Existen diversas

líneas de investigación a futuro que requieren del uso de la estadística, las cuales contribuyen a la construcción de este tipo de conocimiento.

# Capítulo 5

## 5.1. Bibliografía

Adeli, H.; Zhou, Z.; Dadmehr, N. *Analysis of EEG records in an epileptic patient using wavelet transform*. J. Neurosci. Methods, 123, 69-87 (2003).

Bandt, C.; Pompe, B. *Permutation Entropy: A Natural Complexity Measure for Time Series*. Phys. Rev. Lett.88, 174102 (2002).

Bernaola-Galván, P.; Román-Roldán, R. and Oliver, J. L. *Compositional segmentation and long-range fractal correlations in DNA sequences*. Phys.Rev. E 53, 5181-5189 (1996).

Borovkov, A.A. *Mathematical Statistics*.Mir, Moscow (1984).

Canavos, G. *Applied Probability and Statistical*.Boston: Little, Brown. Methods (1984).

Cover, T.M.; Thomas, J.A. *Elements of Information Theory, 2nd ed.*; John Wiley & Sons: Hoboken, NJ, USA (2006).

Diaz, J.M; Mateos. D.M; Boyallian, C. *Complexity-Entropy Maps as a Tool*

*for the Characterization of the Clinical Electrophysiological Evolution of Patients under Pharmacological Treatment with Psychotropic Drugs.* Entropy e-ISSN: 1099-4300 (2017)

Durban, M. *Introducción a los modelos mixtos.*

<http://halweb.uc3m.es/esp/Personal/personas/durban>.

Grosse, I.; Bernaola-Galvan, P.; Carpena, P.; Román-Roldán, R.; Oliver, J.; Stanley, H.E. *Analysis of symbolic sequences using the Jensen-Shannon divergence.* Revision Fisica E. Vol. 65, Iss. 4 (2002) .

Fisher, R.A. *The Design of Experiments.* Edinburgh and London: Oliver and Boyd (1935).

Hartigan, J.A. *Clustering Algorithms.* Jhon Wiley & Sons, Hoboken (1975)

Hao, B.L. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems* World Scientific, Singapore (1989).

Hao, B.L. *Symbolic dynamics and characterization of complexity.* Physica D 51, 161-176 (1991).

Herzel, H. and Grosse, I. *Measuring correlations in symbol sequences.* Physica A 216, 518-542 (1995).

Herzel, H. and Grosse, I. *Correlation in DNA sequences: The role of protein coding segments* Phys. Rev. E 55, 800 (1997).

Hsu, Jason C. *Multiple Comparisons. Theory and methods.* Chapman and Hall/CRC (1996)

Janackovak, S.; Boyd, S.; Yozawitz, E.; Tsuchida, T.; Lamblin, M.D., Gue-

- den, S., Pressler, R. *Electroencephalographic characteristics of epileptic seizures in preterm neonates*. Clin. Neurophysiol. 127, 2721–2727 (2016).
- Jung, T.P.; Makeig, S.; Stensmo, M.; Sejnowski, T.J. *Estimating alertness from the EEG power spectrum*. IEEE Trans. Biomed. Eng. 44, 60-69 (1997).
- Larsen, R.J.; Morris, L.M. *An introduction to Mathematical Statistic and Its Applications*. Pearson prentice Hall (1981).
- Li, W. *Mutual information functions versus correlation functions* J. Stat. Phys. 60, 823 (1990).
- Lee, Y.; Nelder, J. *Conditional and Marginal Models: Another view*. Statistical Science, 19, 219-238 (2004).
- Lin, J. *Divergence measures based on the Shannon entropy* IEEE Trans. Inf. Theory 37, 145 (1991).
- Mateos, D.M.; Zozor, S.; Olivarez, F. *On the Analysis of Signals in a Permutation Lempel-Ziv Complexity-Permutation Shannon Entropy Plane*. arXiv:1707.05164 (2017).
- Mateos, D.; Erra, R.G.; Wennberg, R.; Velazquez, J. *Measures of Entropy and Complexity in Altered States of Consciousness*. Cognitive Neurodynamics volume 12, pages73–84 (2018)
- Menendez, M.L.; Pardo, J.A.; Pardo, L.; and Pardo, M.C. *The Jensen-Shannon Divergence*. Franklin Inst. 334B, 307 (1997).
- Molenberghs, G. and Verbeke, G. *Linear Mixed Models for Longitudinal Data*. New York. Springer (2000).

- Montani, F.; Rosso, O.A. *Entropy-Complexity characterization of Brain Development in Chickens*. *Entropy* 16, 4677–4692 (2014).
- Peña, D. *Análisis de Datos Multivariantes*. McGraw-Hill Interamericana de España S.L (2002).
- Radhajrishnank, N.; Gangadhar, B.N. *Estimating Regularity in Epileptic Seizure Time-Series Data—A Complexity-Measure Approach*. *IEEE Eng. Med. Biol. Mag.* 17, 89–94 (1998).
- Ramensky, V.E.; Makeev, V.Ju.; Roytberg, M.A and Tumanyan, V.G. *DNA Segmentation Through the Bayesian Approach* *J. Comput. Biol.* 7, 1 (2000).
- Rosso, O.; Larrondo, H.; Martin, M.; Plastino, A.; Fuentes, M. *Distinguishing noise from chaos*. *Phys. Rev. Lett.* 99, 154102 (2007).
- Rosso, O.A.; Craig, H.; Moscato, P. *Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers*. *Physica A* 388, 916–926 (2009).
- Shannon, C.E. and Weaver, W.W. *The Mathematical Theory of Communication* University of Illinois Press, Urbana, IL (1949).
- Stroup, W.W. *Generalized Linear Mixed Models. Modern Concepts, Methods and Applications*. CRC Press (2012)
- Thakor, N.V.; Zhu, Y.S.; Pan, K.Y. *Ventricular Tachycardia and Fibrillation Detection by a Sequential Hypothesis Testing Algorithm*. *IEEE Trans. Biomed. Eng.* 37, 837–843 (1990).
- West, B. T.; Welch, K. B.; and Galecki, A. T. *Linear Mixed Models: A Practi-*

*cal Guide using Statistical Software, Seconde Edition.* Chapman & Hall/CRC, Boca Raton, FL (2015).

Wong, A.K.C and You M. *Entropy and distance of random graphs with application to structural pattern recognition* IEEE Trans. Pattern Anal. Mach.Intell. 7, 599-609 (1985).

Zhang, X.S.; Zhu, Y.S.; Thakor, N.V.; Wang, Z.Z. *Detecting ventricular tachycardia and fibrillation by complexity measure.* IEEE Trans. Biomed. Eng.46, 548–555 (1999).

Zozor, S.; Mateos, D.; Lamberti, P.W. *Mixing Bandt-Pompe and Lempel-Ziv approaches: Another way to analyze the complexity of continuous-states sequences.* Eur. Phys. J. B, 87, 107 (2014).

## **5.2. Apéndice A**

Tabla 5.1: Medidas estadísticas de la variable complejidad particionada por estado y canal

<b>Indice</b>	<b>Canales</b>	<b>Estado</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Complejidad	01-F1	HIPER	0,460	0,030	0,410	0,500
Complejidad	01-F1	OA	0,470	0,030	0,390	0,510
Complejidad	01-F1	OC	0,470	0,030	0,400	0,510
Complejidad	01-F1	OPTO	0,490	0,020	0,430	0,520
Complejidad	02-F3	HIPER	0,450	0,030	0,400	0,500
Complejidad	02-F3	OA	0,480	0,020	0,440	0,500
Complejidad	02-F3	OC	0,460	0,030	0,360	0,500
Complejidad	02-F3	OPTO	0,490	0,020	0,450	0,520
Complejidad	03-C3	HIPER	0,460	0,020	0,410	0,500
Complejidad	03-C3	OA	0,480	0,020	0,430	0,520
Complejidad	03-C3	OC	0,460	0,030	0,410	0,500
Complejidad	03-C3	OPTO	0,490	0,020	0,430	0,510
Complejidad	04-P3	HIPER	0,450	0,030	0,380	0,510
Complejidad	04-P3	OA	0,480	0,030	0,400	0,520
Complejidad	04-P3	OC	0,460	0,030	0,370	0,500
Complejidad	04-P3	OPTO	0,480	0,030	0,410	0,520
Complejidad	05-O1	HIPER	0,430	0,040	0,350	0,510
Complejidad	05-O1	OA	0,470	0,030	0,380	0,520
Complejidad	05-O1	OC	0,440	0,050	0,300	0,500
Complejidad	05-O1	OPTO	0,480	0,030	0,410	0,520
Complejidad	06-F7	HIPER	0,480	0,020	0,440	0,520
Complejidad	06-F7	OA	0,490	0,010	0,460	0,520
Complejidad	06-F7	OC	0,480	0,020	0,430	0,520
Complejidad	06-F7	OPTO	0,510	0,020	0,470	0,530
Complejidad	07-T3	HIPER	0,490	0,020	0,440	0,520
Complejidad	07-T3	OA	0,490	0,020	0,450	0,530
Complejidad	07-T3	OC	0,490	0,020	0,450	0,510
Complejidad	07-T3	OPTO	0,510	0,020	0,470	0,530
Complejidad	08-T5	HIPER	0,470	0,040	0,320	0,510
Complejidad	08-T5	OA	0,490	0,020	0,400	0,520
Complejidad	08-T5	OC	0,460	0,050	0,290	0,520
Complejidad	08-T5	OPTO	0,490	0,020	0,420	0,530



Tabla 5.2: Medidas estadísticas de la variable complejidad particionada por estado y canal

<b>Índice</b>	<b>Canales</b>	<b>Estado</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Complejidad	09-Fz	HIPER	0,450	0,030	0,390	0,490
Complejidad	09-Fz	OA	0,470	0,020	0,420	0,510
Complejidad	09-Fz	OC	0,460	0,030	0,370	0,490
Complejidad	09-Fz	OPTO	0,480	0,020	0,420	0,510
Complejidad	10-Cz	HIPER	0,460	0,030	0,400	0,500
Complejidad	10-Cz	OA	0,480	0,020	0,420	0,520
Complejidad	10-Cz	OC	0,460	0,030	0,370	0,500
Complejidad	10-Cz	OPTO	0,480	0,020	0,450	0,520
Complejidad	11-Pz	HIPER	0,450	0,030	0,400	0,500
Complejidad	11-Pz	OA	0,470	0,020	0,420	0,520
Complejidad	11-Pz	OC	0,450	0,030	0,380	0,500
Complejidad	11-Pz	OPTO	0,480	0,020	0,430	0,510
Complejidad	12-Oz	HIPER	0,440	0,030	0,380	0,510
Complejidad	12-Oz	OA	0,470	0,020	0,430	0,510
Complejidad	12-Oz	OC	0,450	0,030	0,370	0,500
Complejidad	12-Oz	OPTO	0,470	0,020	0,430	0,500
Complejidad	13-T6	HIPER	0,470	0,030	0,400	0,520
Complejidad	13-T6	OA	0,490	0,030	0,380	0,530
Complejidad	13-T6	OC	0,470	0,030	0,410	0,520
Complejidad	13-T6	OPTO	0,500	0,020	0,430	0,540
Complejidad	14-T4	HIPER	0,490	0,020	0,450	0,510
Complejidad	14-T4	OA	0,500	0,020	0,470	0,540
Complejidad	14-T4	OC	0,490	0,020	0,460	0,520
Complejidad	14-T4	OPTO	0,510	0,010	0,480	0,530
Complejidad	15-F8	HIPER	0,490	0,010	0,450	0,510
Complejidad	15-F8	OA	0,490	0,020	0,460	0,520
Complejidad	15-F8	OC	0,480	0,020	0,430	0,510
Complejidad	15-F8	OPTO	0,510	0,010	0,480	0,530
Complejidad	16-O2	HIPER	0,440	0,040	0,380	0,500
Complejidad	16-O2	OA	0,470	0,030	0,340	0,520
Complejidad	16-O2	OC	0,440	0,040	0,340	0,510
Complejidad	16-O2	OPTO	0,470	0,030	0,400	0,520
Complejidad	17-P4	HIPER	0,450	0,030	0,410	0,500
Complejidad	17-P4	OA	0,480	0,030	0,390	0,530
Complejidad	17-P4	OC	0,450	0,030	0,350	0,500
Complejidad	17-P4	OPTO	0,480	0,030	0,400	0,510

Tabla 5.3: Medidas estadísticas de la variable complejidad y entropía particionadas por estado y canal

<b>Indice</b>	<b>Canales</b>	<b>Estado</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Complejidad	18-C4	HIPER	0,460	0,030	0,410	0,510
Complejidad	18-C4	OA	0,480	0,020	0,420	0,520
Complejidad	18-C4	OC	0,460	0,030	0,370	0,500
Complejidad	18-C4	OPTO	0,490	0,020	0,440	0,520
Complejidad	19-F4	HIPER	0,470	0,020	0,420	0,510
Complejidad	19-F4	OA	0,480	0,020	0,430	0,520
Complejidad	19-F4	OC	0,470	0,030	0,380	0,500
Complejidad	19-F4	OPTO	0,490	0,020	0,450	0,520
Complejidad	20-F2	HIPER	0,460	0,030	0,410	0,520
Complejidad	20-F2	OA	0,470	0,030	0,380	0,530
Complejidad	20-F2	OC	0,470	0,030	0,380	0,510
Complejidad	20-F2	OPTO	0,490	0,030	0,380	0,520
Entropia	01-F1	HIPER	0,956	0,036	0,885	1,019
Entropia	01-F1	OA	0,979	0,044	0,868	1,028
Entropia	01-F1	OC	0,973	0,034	0,876	1,020
Entropia	01-F1	OPTO	0,982	0,025	0,912	1,022
Entropia	02-F3	HIPER	0,959	0,030	0,887	1,005
Entropia	02-F3	OA	0,990	0,025	0,930	1,024
Entropia	02-F3	OC	0,963	0,033	0,871	1,008
Entropia	02-F3	OPTO	0,981	0,023	0,931	1,018
Entropia	03-C3	HIPER	0,968	0,029	0,894	1,012
Entropia	03-C3	OA	0,998	0,021	0,940	1,031
Entropia	03-C3	OC	0,975	0,031	0,895	1,015
Entropia	03-C3	OPTO	0,983	0,023	0,913	1,017
Entropia	04-P3	HIPER	0,960	0,040	0,859	1,022
Entropia	04-P3	OA	0,995	0,031	0,872	1,037
Entropia	04-P3	OC	0,971	0,038	0,848	1,017
Entropia	04-P3	OPTO	0,975	0,034	0,885	1,018
Entropia	05-O1	HIPER	0,944	0,048	0,835	1,023
Entropia	05-O1	OA	0,995	0,038	0,848	1,036
Entropia	05-O1	OC	0,959	0,053	0,815	1,019
Entropia	05-O1	OPTO	0,975	0,033	0,878	1,016
Entropia	06-F7	HIPER	0,991	0,021	0,954	1,030
Entropia	06-F7	OA	1,010	0,018	0,967	1,039
Entropia	06-F7	OC	0,998	0,019	0,956	1,032
Entropia	06-F7	OPTO	1,006	0,017	0,959	1,037

Tabla 5.4: Medidas estadísticas de la variable entropía particionada por estado y canal

<b>Indice</b>	<b>Canales</b>	<b>Estado</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Entropia	07-T3	HIPER	1,001	0,022	0,928	1,033
Entropia	07-T3	OA	1,017	0,017	0,957	1,039
Entropia	07-T3	OC	1,003	0,021	0,945	1,032
Entropia	07-T3	OPTO	1,008	0,020	0,953	1,033
Entropia	08-T5	HIPER	0,985	0,040	0,844	1,033
Entropia	08-T5	OA	1,012	0,028	0,890	1,039
Entropia	08-T5	OC	0,984	0,050	0,800	1,036
Entropia	08-T5	OPTO	1,000	0,025	0,903	1,034
Entropia	09-Fz	HIPER	0,948	0,036	0,856	0,999
Entropia	09-Fz	OA	0,978	0,030	0,892	1,023
Entropia	09-Fz	OC	0,965	0,033	0,855	0,999
Entropia	09-Fz	OPTO	0,967	0,028	0,892	1,003
Entropia	10-Cz	HIPER	0,969	0,031	0,903	1,017
Entropia	10-Cz	OA	0,993	0,023	0,934	1,018
Entropia	10-Cz	OC	0,965	0,033	0,869	1,018
Entropia	10-Cz	OPTO	0,981	0,023	0,942	1,013
Entropia	11-Pz	HIPER	0,962	0,032	0,900	1,015
Entropia	11-Pz	OA	0,993	0,023	0,921	1,018
Entropia	11-Pz	OC	0,960	0,034	0,867	1,006
Entropia	11-Pz	OPTO	0,978	0,025	0,919	1,018
Entropia	12-Oz	HIPER	0,958	0,040	0,863	1,020
Entropia	12-Oz	OA	0,996	0,025	0,921	1,036
Entropia	12-Oz	OC	0,973	0,030	0,885	1,014
Entropia	12-Oz	OPTO	0,971	0,023	0,934	1,007
Entropia	13-T6	HIPER	0,990	0,030	0,910	1,036
Entropia	13-T6	OA	1,015	0,028	0,923	1,043
Entropia	13-T6	OC	0,992	0,028	0,943	1,040
Entropia	13-T6	OPTO	1,001	0,023	0,944	1,040
Entropia	14-T4	HIPER	1,005	0,021	0,955	1,033
Entropia	14-T4	OA	1,022	0,012	0,995	1,049
Entropia	14-T4	OC	1,007	0,018	0,965	1,043
Entropia	14-T4	OPTO	1,013	0,014	0,989	1,035
Entropia	15-F8	HIPER	1,008	0,016	0,971	1,036
Entropia	15-F8	OA	1,018	0,015	0,981	1,047
Entropia	15-F8	OC	1,007	0,016	0,969	1,034
Entropia	15-F8	OPTO	1,018	0,013	0,980	1,041

Tabla 5.5: Medidas estadísticas de la variable entropía particionada por estado y canal

<b>Indice</b>	<b>Canales</b>	<b>Estado</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Entropia	16-O2	HIPER	0,948	0,040	0,878	1,025
Entropia	16-O2	OA	0,988	0,037	0,849	1,022
Entropia	16-O2	OC	0,953	0,044	0,855	1,024
Entropia	16-O2	OPTO	0,971	0,033	0,885	1,019
Entropia	17-P4	HIPER	0,961	0,034	0,895	1,015
Entropia	17-P4	OA	0,992	0,031	0,892	1,024
Entropia	17-P4	OC	0,962	0,037	0,853	1,019
Entropia	17-P4	OPTO	0,974	0,030	0,879	1,023
Entropia	18-C4	HIPER	0,971	0,031	0,895	1,024
Entropia	18-C4	OA	0,997	0,023	0,932	1,025
Entropia	18-C4	OC	0,972	0,034	0,861	1,013
Entropia	18-C4	OPTO	0,982	0,025	0,923	1,023
Entropia	19-F4	HIPER	0,977	0,028	0,910	1,030
Entropia	19-F4	OA	1,000	0,020	0,936	1,030
Entropia	19-F4	OC	0,978	0,032	0,875	1,021
Entropia	19-F4	OPTO	0,989	0,022	0,946	1,021
Entropia	20-F2	HIPER	0,970	0,038	0,893	1,030
Entropia	20-F2	OA	0,989	0,036	0,866	1,031
Entropia	20-F2	OC	0,987	0,033	0,899	1,021
Entropia	20-F2	OPTO	0,987	0,033	0,872	1,028

Tabla 5.6: Medidas estadísticas de la variable complejidad por individuo

<b>Indice</b>	<b>Identidad</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Complejidad	1	0,49	0,02	0,43	0,52
Complejidad	2	0,49	0,02	0,44	0,53
Complejidad	3	0,42	0,03	0,36	0,49
Complejidad	4	0,49	0,02	0,43	0,53
Complejidad	5	0,48	0,02	0,44	0,52
Complejidad	6	0,46	0,03	0,38	0,51
Complejidad	7	0,47	0,02	0,4	0,51
Complejidad	8	0,51	0,02	0,47	0,54
Complejidad	9	0,46	0,03	0,38	0,5
Complejidad	10	0,47	0,04	0,39	0,52
Complejidad	11	0,47	0,02	0,44	0,51
Complejidad	12	0,46	0,03	0,38	0,52
Complejidad	13	0,45	0,04	0,35	0,53
Complejidad	14	0,47	0,03	0,37	0,51
Complejidad	15	0,49	0,01	0,46	0,53
Complejidad	16	0,47	0,03	0,4	0,52
Complejidad	17	0,50	0,01	0,47	0,53
Complejidad	18	0,46	0,05	0,3	0,53
Complejidad	19	0,49	0,01	0,47	0,53
Complejidad	20	0,46	0,04	0,36	0,52
Complejidad	21	0,43	0,04	0,33	0,49
Complejidad	22	0,44	0,03	0,34	0,51
Complejidad	23	0,46	0,03	0,4	0,52
Complejidad	24	0,47	0,03	0,39	0,52
Complejidad	25	0,50	0,02	0,46	0,53
Complejidad	26	0,48	0,03	0,43	0,53
Complejidad	27	0,47	0,03	0,38	0,52
Complejidad	28	0,47	0,07	0,19	0,53
Complejidad	29	0,48	0,03	0,42	0,53
Complejidad	30	0,46	0,02	0,38	0,51

Tabla 5.7: Medidas estadísticas de la variable entropía por individuo

<b>Indice</b>	<b>Identidad</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Entropia	1	1,016	0,014	0,973	1,035
Entropia	2	0,996	0,023	0,938	1,036
Entropia	3	0,913	0,039	0,830	0,993
Entropia	4	0,982	0,026	0,925	1,029
Entropia	5	0,982	0,023	0,935	1,023
Entropia	6	0,969	0,037	0,872	1,020
Entropia	7	0,993	0,015	0,958	1,019
Entropia	8	1,012	0,022	0,960	1,051
Entropia	9	0,961	0,030	0,874	0,998
Entropia	10	0,973	0,039	0,890	1,028
Entropia	11	0,979	0,022	0,932	1,018
Entropia	12	0,978	0,034	0,888	1,031
Entropia	13	0,969	0,051	0,843	1,035
Entropia	14	0,986	0,038	0,873	1,033
Entropia	15	1,011	0,012	0,982	1,036
Entropia	16	0,986	0,034	0,895	1,030
Entropia	17	1,015	0,011	0,993	1,036
Entropia	18	0,985	0,050	0,820	1,038
Entropia	19	1,006	0,012	0,978	1,028
Entropia	20	0,975	0,048	0,855	1,039
Entropia	21	0,947	0,045	0,821	1,018
Entropia	22	0,946	0,041	0,822	1,010
Entropia	23	0,979	0,037	0,884	1,036
Entropia	24	0,981	0,036	0,896	1,035
Entropia	25	1,007	0,023	0,958	1,054
Entropia	26	1,002	0,022	0,959	1,033
Entropia	27	0,983	0,034	0,886	1,026
Entropia	28	0,998	0,037	0,872	1,038
Entropia	29	1,005	0,025	0,938	1,053
Entropia	30	0,974	0,032	0,877	1,024

Tabla 5.8: Medidas estadísticas para las variables entropía y complejidad particionadas por estado y sector

<b>Indice</b>	<b>Estado</b>	<b>SectorEC</b>	<b>n</b>	<b>Media</b>	<b>D.E.</b>	<b>Mín</b>	<b>Máx</b>
Complejidad	HIPER	1	120	0,440	0,040	0,300	0,510
Complejidad	HIPER	2	90	0,450	0,040	0,360	0,510
Complejidad	HIPER	3	60	0,460	0,030	0,350	0,510
Complejidad	HIPER	4	60	0,460	0,040	0,340	0,530
Complejidad	HIPER	5	120	0,480	0,030	0,230	0,520
Complejidad	OA	1	120	0,470	0,020	0,400	0,510
Complejidad	OA	2	90	0,480	0,020	0,420	0,500
Complejidad	OA	3	60	0,480	0,020	0,420	0,510
Complejidad	OA	4	60	0,490	0,020	0,390	0,530
Complejidad	OA	5	120	0,490	0,030	0,190	0,530
Complejidad	OC	1	120	0,440	0,040	0,320	0,520
Complejidad	OC	2	90	0,450	0,030	0,350	0,510
Complejidad	OC	3	60	0,460	0,030	0,390	0,510
Complejidad	OC	4	60	0,460	0,040	0,370	0,530
Complejidad	OC	5	120	0,480	0,020	0,430	0,520
Complejidad	OPTO	1	120	0,480	0,030	0,380	0,520
Complejidad	OPTO	2	90	0,480	0,030	0,390	0,520
Complejidad	OPTO	3	60	0,490	0,020	0,410	0,530
Complejidad	OPTO	4	60	0,500	0,030	0,410	0,540
Complejidad	OPTO	5	120	0,500	0,040	0,210	0,540
Entropia	HIPER	1	120	0,944	0,050	0,820	1,023
Entropia	HIPER	2	90	0,956	0,042	0,843	1,019
Entropia	HIPER	3	60	0,964	0,037	0,845	1,019
Entropia	HIPER	4	60	0,977	0,043	0,870	1,049
Entropia	HIPER	5	120	0,996	0,029	0,881	1,037
Entropia	OA	1	120	0,993	0,027	0,903	1,036
Entropia	OA	2	90	0,992	0,024	0,933	1,031
Entropia	OA	3	60	0,996	0,020	0,943	1,030
Entropia	OA	4	60	1,015	0,023	0,926	1,054
Entropia	OA	5	120	1,013	0,022	0,880	1,053
Entropia	OC	1	120	0,954	0,046	0,821	1,033
Entropia	OC	2	90	0,963	0,037	0,844	1,026
Entropia	OC	3	60	0,973	0,031	0,884	1,025
Entropia	OC	4	60	0,980	0,041	0,874	1,048
Entropia	OC	5	120	1,002	0,021	0,939	1,036
Entropia	OPTO	1	120	0,978	0,032	0,857	1,029
Entropia	OPTO	2	90	0,981	0,032	0,869	1,031
Entropia	OPTO	3	60	0,985	0,028	0,889	1,026
Entropia	OPTO	4	60	1,002	0,030	0,900	1,051
Entropia	OPTO	5	120	1,009	0,028	0,872	1,045

## 5.3. Apéndice B

### 5.3.1. Conglomerados

#### *#Conglomerados para la variable Entropia y Complejidad*

```
mediasEntropia=apply(Entropia$Respuesta, Entropia$Canales, mean);
mediasEntropia=data.frame(mediasEntropia)
mediasComplejidad=apply(Complejidad$Respuesta, Complejidad$Canales, mean);
mediasComplejidad=data.frame(mediasComplejidad)
medias=cbind(mediasEntropia, mediasComplejidad)
HClust.1<-hclust(dist(model.matrix(~-1+mediasEntropia+mediasComplejidad, medias)),method="ward")
medias$h5 <- c(as.factor(cutree(HClust.1, k = 5)))
tabla=data.frame(medias, medias$h5)
G1=as.matrix(subset(tabla,tabla$h5==1));
G2=as.matrix(subset(tabla,tabla$h5==2));
G3=as.matrix(subset(tabla,tabla$h5==3));
G4=as.matrix(subset(tabla,tabla$h5==4))
G5=as.matrix(subset(tabla,tabla$h5==5));
```

#### *#Conglomerados para la variable Entropia*

```
medias=apply(Entropiatrain$Respuesta, Entropiatrain$Canales, mean);
medias=data.frame(medias)
HClust.1<-hclust(dist(model.matrix(~-1+medias, medias)), method="ward")
medias$h5 <- c(as.factor(cutree(HClust.1, k = 5)))
tabla=data.frame(medias, medias$h5)
G1=as.matrix(subset(tabla,tabla$h5==1))
G2=as.matrix(subset(tabla,tabla$h5==2))
G3=as.matrix(subset(tabla,tabla$h5==3))
G4=as.matrix(subset(tabla,tabla$h5==4))
G54=as.matrix(subset(tabla,tabla$h5==5))
```



### 5.3.2. Modelos ajustados

```
#Modelo multivariado (3.6)

library(nlme)

ECtrain$SectorEC_I_EC=interaction(ECtrain[,c("SectorEC", "I_EC")])

mimodelo= mimodelo=lme(Respuesta~1+CATEdad:I_EC+Sexo:I_EC+Estado:I_EC+I_EC
                        ,random=list(SectorEC_I_EC=pdIdent(~1)
                                      ,Identidad=pdIdent(~1)
                                      ,Identidad=pdIdent(~Estado:I_EC-1))
                        ,weights=varComb(varIdent(form=~1|Identidad))
                        ,method="REML"
                        ,control=lmeControl(niterEM=150
                                             ,msMaxIter=200)
                        ,na.action=na.omit
                        ,data=ECtrain
                        ,keep.data=FALSE)

#Modelo univariado (3.7) para respuesta entropia

mimodelo=lme(Entropia~1+Estado+CATEdad+Sexo
              ,random=list(SectorE_I_EC=pdIdent(~1)
                            ,Identidad=pdIdent(~1)
                            ,Identidad=pdIdent(~Estado:I_EC-1))
              ,weights=varComb(varIdent(form=~1|Identidad))
              ,method="REML"
              ,control=lmeControl(niterEM=150
                                   ,msMaxIter=200) ,na.action=na.omit
              ,data=Entropiatrain
              ,keep.data=FALSE)
```

### 5.3.3. LSD Fisher con corrección de Bonferroni

*# Las funciones descritas a continuación fueron desarrolladas por Di Rienzo, et. al. 2002 y son necesarias para poder aplicar el test LSD de Fisher con corrección de Bonferroni.*

```
t2set=function(x,split=":") strsplit(x,split)[[1]]

TheModelCoefficients=function(myModel){
  if (class(myModel)=="lme") result=myModel$coefficients$fixed
  if (class(myModel)=="gls") result=myModel$coefficients
  if (class(myModel)=="lm") result=-myModel$coefficients[complete.cases(myModel$coefficients)]
  if (class(myModel)=="glm") result=myModel$coefficients[complete.cases(myModel$coefficients)]
  if (class(myModel)=="mer") result=myModel@fixef[complete.cases(myModel@fixef)]
  if (class(myModel)=="glmerMod") result=fixef(myModel)
  if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
  result=fixef(myModel)
  result}

#-----
TheModelCoefficientsCovar<-function(myModel){
  if (class(myModel)=="lme") result=myModel$varFix
  if (class(myModel)=="gls") result=myModel$varBeta
  if (class(myModel)=="lm") result=vcov(myModel)
  if (class(myModel)=="glm") result=summary(myModel)$cov.scaled
  if (class(myModel)=="mer") result=as.matrix(summary(myModel)@vcov)
  if (class(myModel)=="glmerMod") result=as.matrix(vcov(myModel))
```

```

TheModelTerms<-function(myModel){
  if (class(myModel)=="lme") {
    result=labels(terms(myModel))
    ;if ("(Intercept)" %in% names(
TheModelCoefficients(myModel)))
    result=c("(Intercept)",result)}
  if (class(myModel)=="gls") result=names(myModel$parAssign)
  if (class(myModel)=="lm") result=labels(myModel$terms)
  if (class(myModel)=="glm") result=labels(myModel$terms)
  if (class(myModel)=="mer") result=labels(terms(myModel))
  if (class(myModel)=="glmerMod") resul=labels(terms(myModel))
  if ((class(myModel)=="lmerMod")|(
class(myModel)=="merModLmerTest"))
    result=labels(terms(myModel))
  result
}
#-----
TheModelFactors<-function(myModel,myData) {
  myModelTerms=TheModelTerms(myModel)
  myModelFactors=-unique(unlist(strsplit(myModelTerms, ":")))
  myModelFactors=intersect(myModelFactors,colnames(myData))
  myModelFactors=myModelFactors[unlist(lapply(as.data.frame(myData[,myModelFactors]),is.factor))]
  myModelFactors}
}

```

```

TheModelFactorLevels=function(myModel,myData){ myModelFactors=TheModelFactors(myModel,myData)
myModelFactorsLevels=lapply(myModelFactors,function(x) levels(myData[,x]))

names(myModelFactorsLevels)=myModelFactors

myModelFactorsLevels
}

#-----

TheModelTermsNotIncludingCovariables=function(myModel,myData){

myModelTerms=TheModelTerms(myModel)

myModelFactors=TheModelFactors(myModel,myData)

myModelTerms[sapply(myModelTerms,function(x){length(setdiff(t2set(x),myModelFactors))==0})]

}

#-----

TheModelTermsIncludingCovariables=function(myModel,myData){

myModelTerms=TheModelTerms(myModel)

myModelFactors=TheModelFactors(myModel,myData)

myModelTermsNotIncludingCovariables=myModelTerms[sapply(myModelTerms,function(x){length(setdiff(t2set(x),myModelFactors))==0})]

setdiff(myModelTerms,myModelTermsNotIncludingCovariables)

}

#-----

TheModelMatrix<-function(myModel,myData){

if (class(myModel)=="lm") result=model.matrix(myModel,data=myData)

if (class(myModel)=="glm") result=model.matrix(myModel,data=myData)

if (class(myModel)=="gls") result=model.matrix(myModel,data=myData)

if (class(myModel)=="lme") result=model.matrix(myModel,data=myData)

if (class(myModel)=="mer") result=getME(myModel,"X")

if (class(myModel)=="glmerMod") result=getME(myModel,"X")

• if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) result=getME(myModel,"X")

result

}

#-----

varianceAdjustedValues<-function(myModel,myData){

myModelMatrix=TheModelMatrix(myModel,myData)

myCovar=TheModelCoefficientsCovar(myModel)

myvar <- diag(myModelMatrix %*%tcrossprod(myCovar,myModelMatrix))

myvar

}

```

```

TheModelResidualDF<-function(myModel){
  if (class(myModel)=="lm") result=myModel$df.residual
  if (class(myModel)=="glm") result=summary(myModel)$df.residual
  if (class(myModel)=="gls") result=myModel$dim$N-myModel$dim$p
  if (class(myModel)=="lme") result=myModel$dim$N-length(myModel$coefficients$fixed)
  if (class(myModel)=="mer") result=myModel@dim["n"]-myModel@dim["p"]
  if (class(myModel)=="glmerMod") result=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))
  if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) result=nrow(getME(myModel,"X"))-
ncol(getME(myModel,"X"))

  result
}

#-----

SearchForASubstringInAStringWithSeparators=function(s1,s2,characterseparadordentrodes2){
  myResult=FALSE
  listastrings=strsplit(s2,characterseparadordentrodes2)[[1]]
  indices=grep(s1,listastrings,fixed =TRUE)
  i=0
  if (length(indices)>0){
    while ((myResult==FALSE)&(i<length(indices))) {
      i=i+1
      myResult=(myResult|(s1==listastrings[indices[i]]))
    }
  }
  myResult
}

#-----

SearchForASubstringInAListOfStringsWithSeparators=function(s1,myList,characterseparadordentrodes2)
{result=grep(s1,c())
Indices=grep(s1,myList,fixed =TRUE)
if (length(indices)>0) for (i in (1:length(indices))) if
(SearchForASubstringInAStringWithSeparators(s1,myList[indices[i]],characterseparadordentrodes2)) result=-c(result,indices[i])
result
}

#-----

SearchForASubstringInAListOfStrings=-function(s1,myList,characterseparadordentrodes2){
result=grep(s1,c())

```

```

indices=grep(s1,myList,fixed =TRUE)

if (length(indices)>0) for (i in (1:length(indices))) if (s1==myList[indices[i]])
result=c(result,indices[i])

result
}

#-----

CalculateMmatrix=function(myModel,myData,myListCovariatesValues=NULL){
myCoefficients=NULL
for (i in (1:ncol(myData))) if (is.factor(myData[,i]))
myData[,i]=as.factor(as.character(myData[,i]))
f=TheModelTerms(myModel)
indices=grep(":",f)
if (length(indices)>0)
myModelFactors=unique(ff-indices) else myModelFactors=unique(f)
myModelFactors=intersect(myModelFactors,colnames(myData))
myModelFactors=myModelFactors[unlist(sapply(as.data.frame(myData[,myModelFactors]),is.factor))]
LevelsList=lapply(myModelFactors,function(result,myData) {paste(result,levels(myData[,result]),sep="")},myData)
if (length(LevelsList)>0) {
list1=levels(interaction(LevelsList,sep=":"))
myModelFactors=-unique(unlist(strsplit(f, ":")))
myModelFactors=intersect(myModelFactors,colnames(myData))
myModelFactors=myModelFactors[unlist(lapply(as.data.frame(myData[,myModelFactors]),is.factor))]
indices=complete.cases(myData[,myModelFactors])
aa=as.data.frame(sapply(myModelFactors,function(result){paste(result,myData[indices,result],sep="")}))
list2=as.character(unique(interaction(aa[,myModelFactors],sep=":")))
myList=union(list1,list2)
myCoefficientsNames=names(TheModelCoefficients(myModel))
myCoefficients<=matrix(rep(0,length(myList)*length(myCoefficientsNames)),nrow=length(myList),ncol=length(myCoefficientsNames))
myConstantTerm=as.integer(length(grep("(Intercept)",myCoefficientsNames,fixed =TRUE))>0)
for (j in (1:length(myCoefficientsNames))) {
ss=unlist(strsplit(myCoefficientsNames[j], ":"))
indices=SearchForASubstringInAListOfStringsWithSeparators(ss,myList,':')
if (length(indices)>0) {
for (i in (1:length(indices))) {

```

```

    result=SearchForASubstringInAListOfStringsWithSeparators(ss[1],myList[indices[i]],":")
    if (length(ss)>1) for (ii in (2:length(ss)))
result=intersect(result,SearchForASubstringInAListOfStringsWithSeparators(ss[ii],myList[indices[i]],":"))
    if (length(result)>0) myCoefficients[indices[i],j]=1}
}
if (myConstantTerm>0) myCoefficients[,1]=myConstantTerm
for (i in (1:length(colnames(myData))))
if (is.numeric(myData[,i])) {
g=SearchForASubstringInAListOfStringsWithSeparators(colnames(myData)[i],myCoefficientsNames,":")
if (length(g)>0) for (j in (1:length(g))) {
    sc=colnames(myData)[i]
    ss=strsplit(myCoefficientsNames[g[j]], ":[1]")
    ss=setdiff(ss,sc)
    if (length(ss)>1) ss=paste(ss,collapse=":")
indices=which(myCoefficientsNames==ss)
mx=mean(myData[,i],na.rm = TRUE)
if (!is.null(myListCovariatesValues)) {
    if (!is.na(myListCovariatesValues[sc])) mx=myListCovariatesValues[sc]
}
if (length(indices)==0) myCoefficients[,g[j]]<-mx
    if (length(indices)>0)
        for (k in (1:length(indices))) myCoefficients[,g[j]]=myCoefficients[,indices[k]]*mx
}
}
rownames(myCoefficients)=myList
colnames(myCoefficients)=myCoefficientsNames
}
myCoefficients
}
#-----
Linear.combination.to.estimate.a.mean=function(MeanLabel,Mmatrix,myCoefficients) {
Ss=strsplit(MeanLabel, ":[1]")
result=SearchForASubstringInAListOfStringsWithSeparators(ss[1],rownames(Mmatrix),":")
if (length(ss)>1) for (i in (2:length(ss)))
result=intersect(result,SearchForASubstringInAListOfStringsWithSeparators(ss[i],rownames(Mmatrix),":"))
}

```

```

ifelse (length(result)>0, SM<-Mmatrix[result,],SM=c(1,rep(0,(length(myCoefficients)-1))))

result=as.matrix(rep((1/nrow(SM)),nrow(SM)))

ifelse ((nrow(result)==0),result=as.matrix(SM), result=-as.matrix(t(result)%*%SM))

if (nrow(result)>1) result=t(result)

colnames(result)=names(myCoefficients)

result
}

#-----

MeanAndVariance=function(MeanLabel,Mmatrix,myModel,myCoefficientsCovar=NULL) {

myCoefficients=TheModelCoefficients(myModel)

if (is.null(myCoefficientsCovar)==T) myCoefficientsCovar=TheModelCoefficientsCovar(myModel)

result=(Linear.combination.to.estimate.a.mean(MeanLabel,Mmatrix,myCoefficients))

myResult=c(result%*%myCoefficients,result%*%myCoefficientsCovar%*%t(result),result)

myResult
}

#-----

RebuildTreatmentNames=function(myData,myModelTerm,MissingCells=FALSE) {

f=strsplit(myModelTerm, ".")[[1]]

indices=complete.cases(myData[,unlist(f)])

myList=paste0(rep(f[1],nrow(as.data.frame(myData[indices,]))),myData[indices,f[1]])

if (length(f)>1) for (i in (2:length(f))) {

if (MissingCells==TRUE) myList=levels(interaction(myList,paste(rep(f[i],nrow(myData[indices,])),myData[indices,f[i]],sep=""),sep=":"))

else

myList=paste(myList,paste(rep(f[i],nrow(myData[indices,])),myData[indices,f[i]],sep=""),sep=":")

}

myList=unique(myList)

sort(myList)

}

#-----

CalculateMatrixofMeansAndStandardErrors=function(myModel,myData,myModelTerm,Mmatrix,MissingCells=FALSE,returnLinearCombinations=FALSE)

{ myList=RebuildTreatmentNames(myData,myModelTerm,MissingCells)

myTable=c()

myLinearCombination=c()

myCoefficientsCovar=TheModelCoefficientsCovar(myModel)

i=1

for (i in (1:length(myList))) {

mv=MeanAndVariance(myList[i],Mmatrix,myModel,myCoefficientsCovar)

```



```

myTable=rbind(myTable,c(mv[1],sqrt(mv[2])))

  myLinearCombination=rbind(myLinearCombination,mv[-c(1,2)])
}

rownames(myTable)=myList
colnames(myTable)=c("Media", "E.E.")

rownames(myLinearCombination)=myList
colnames(myLinearCombination)=colnames(myCoefficientsCovar)

if (!(returnLinearCombinations==T)) myTable else myLinearCombination
}

#-----
FindGreatestDFForAGivenListOfMeansForA_lme_model=function(myModel,meanslist){
  elgl=function(myMean,myTerm) {
    result=unique((attributes(gregexpr (myTerm,myMean,fixed = T)[[1]][[1]]==nchar(myTerm)))
  }
  maxgl=function(myTerm,meanslist) {
    sapply(meanslist,elgl,myTerm)
  }
  if (class(myModel)=="lme"){
    myterms=names(myModel$fixDF$terms)
    m=as.matrix(sapply(myterms,maxgl,meanslist))
    max(myModel$fixDF$terms%*%t(m))
  }
}

#-----
AvgSEMeansDiff_and_AvgDMS=function(myModel,myData,myModelTerm,Mmatrix,alfa=0.05){
  compare=function(listofmeans,myModel) {
    COVBETA=TheModelCoefficientsCovar(myModel)
    result=c()
    for (i in (1:(length(listofmeans)-1))) {
      m1=MeanAndVariance(listofmeans[i],Mmatrix,myModel)
      for (j in ((i+1):length(listofmeans))){
        m2=MeanAndVariance(listofmeans[j],Mmatrix,myModel)
        dif=m1[3:length(m1)]-m2[3:length(m2)]
        result=c(result,(dif%*%COVBETA%*%dif))
      }
    }
  }
  result}

```

```

factores=strsplit(myModelTerm, ":")

f=factores[[1]]

listofmeans=paste(rep(f[1],nrow(myData)),myData[,f[1]],sep="")

if (length(f)>1) for (i in (2:length(f))){

  listofmeans=paste(listofmeans,paste(rep(f[i],nrow(myData)),myData[,f[i]],sep="),sep=":'" ) }

listofmeans=unique(listofmeans)

if (class(myModel)=="lm") n=myModel$df.residual

if (class(myModel)=="glm") n=summary(myModel)$df.residual

if (class(myModel)=="gls") n=myModel$dim$N-myModel$dim$p

if (class(myModel)=="lme") n=FindGreatestDFForAGivenListOfMeansForA_lme_model(myModel,c(listofmeans[1],listofmeans[2]))

if (class(myModel)=="mer") n=myModel@dim["n"]-myModel@dim["p"]

if (class(myModel)=="glmerMod") n=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))

if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) n=nrow(getME(myModel,"X"))-
ncol(getME(myModel,"X"))

SEM=sqrt(mean(compare(listofmeans,myModel),na.rm=T))

result=list(AvgSEM=SEM,AvgDMS=qt((1-alfa/2),n)*SEM)

result}

#-----

MeansComparison=function(MeanLabel1,MeanLabel2,Mmatrix,myModel,myDF=0){

m1=MeanAndVariance(MeanLabel1,Mmatrix,myModel)

m2=MeanAndVariance(MeanLabel2,Mmatrix,myModel)

dif=m1[3:length(m1)]-m2[3:length(m2)]

myCoefficients=TheModelCoefficients(myModel)

COVBETA=TheModelCoefficientsCovar(myModel)

SEdiff=sqrt(dif%*%COVBETA%*%dif)

estadisticoT=-dif%*%myCoefficients/SEdiff

if (class(myModel)=="lm") myDF=myModel$df.residual

if (class(myModel)=="glm") myDF=summary(myModel)$df.residual

if ((class(myModel)=="lme")&(myDF==0))
myDF=FindGreatestDFForAGivenListOfMeansForA_lme_model(myModel,c(MeanLabel1,MeanLabel2))

if (class(myModel)=="gls") myDF=myModel$dim$N-myModel$dim$p

if (class(myModel)=="mer") myDF=myModel@dim["myDF"]-myModel@dim["p"]

if (class(myModel)=="glmerMod") myDF=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))

if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) myDF=nrow(getME(myModel,"X"))-
ncol(getME(myModel,"X"))

if (myDF==0)  ifelse((m1==m2),myResult=1,myResult=anova(myModel,L=dif)$p)

if (!(myDF==0)) ifelse((m1==m2),myResult=1,myResult=-2*(1-pt(abs(estadisticoT),df=myDF)))

myResult}

}

```

```

Test_Contrast=function(myModel,myData,myModelTerm,Mmatrix,myContrast,denDF){
  myList=RebuildTreatmentNames(myData,myModelTerm)
  myTable=c()
  for (i in (1:length(myList))) {
    a=MeanAndVariance(myList[i],Mmatrix,myModel);
    myTable=rbind(myTable,a[3:length(a)]) }
  myTable=matrix(as.numeric(as.character(myTable)),ncol=ncol(myTable))
  dif=as.matrix(t(myContrast)%*%myTable)
  myCoefficients=TheModelCoefficients(myModel)
  COVBETA=TheModelCoefficientsCovar(myModel)
  numDF=nrow(dif)
  Fstatistic<-t(dif%*%myCoefficients)%*%solve(dif%*%COVBETA%*%t(dif))%*%(dif%*%myCoefficients)
  EE=sqrt(diag(dif%*%COVBETA%*%t(dif)))
  if (class(myModel)=="glm") denDF=summary(myModel)$df.residual
  if (class(myModel)=="mer") denDF=myModel@dims["n"]-myModel@dims["p"]
  if (class(myModel)=="glmerMod") n=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))
  if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) n=nrow(getME(myModel,"X"))-
  ncol(getME(myModel,"X"))
  if (("glm" %in%
  class(myModel))|(class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="mer
  ModLmerTest")) {
    p=(1-pchisq(Fstatistic[1],numDF))
    myResult=-list("Chi"=Fstatistic[1],"numDF"=numDF,"p"=p,"contrastvalue"=dif%*%myCoefficients,"contrastSE"=EE[1])
    if (!(("glm" %in%
  class(myModel))|(class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="mer
  ModLmerTest")) {
      Fstatistic=Fstatistic/numDF# es una chi dividida sus grados de libertad
      if (!(denDF==0)) ifelse((Fstatistic[1]==0),p=1,p=(1-pf(Fstatistic[1],numDF,denDF)))
    myResult=list("F"=Fstatistic[1],"numDF"=numDF,"denDF"=denDF,"p"=p,"contrastvalue"=dif%*%myCoefficients,"contrastSE"=EE[1])
  }
  myResult
}
#-----
checklinearindependence=function(mycombination){
  mycombination=as.matrix(mycombination)
  result=FALSE
  tryCatch(error=function(e) "error", expr= {

```

```

solve(t(mycombination)%*%mycombination)
result=TRUE })

result
}

#-----

contrastLC=function(myModel,mycombination){
myModelCoefficients=TheModelCoefficients(myModel)
myModelCoefficients=myModelCoefficients[complete.cases(myModelCoefficients)]
mycombination=matrix(mycombination,ncol=length(myModelCoefficients))
if (class(myModel)=="gls") {
mycombination=as.matrix(mycombination)
myanova=anova(myModel,L=mycombination)
HB=mycombination%*%myModelCoefficients
if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%*%vcov(myModel)%*%t(mycombination))) else EE=NA
myResult=list("p"=myanova$p,"Fobs"=myanova$F,"numDF"=nrow(mycombination),"cl"=HB,"EE"=EE)
}
if (class(myModel)=="lme"){
myResult=NULL
mycombination=as.matrix(mycombination)
tryCatch(error=function(e) "error", expr={
myanova=anova(myModel,L=mycombination)
HB=mycombination%*%myModelCoefficients
if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%*%vcov(myModel)%*%t(mycombination))) else EE=NA
myResult=list("p"=myanova$p,"Fobs"=myanova$F,"numDF"=nrow(mycombination),"cl"=HB,"EE"=EE)
})
if (is.null(myResult)) {
dd=summary(myModel)$dims
denDF=as.numeric(dd$N-sum(dd$ncol))
HB=mycombination%*%myModelCoefficients
if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%*%vcov(myModel)%*%t(mycombination))) else EE=NA
Fobs=t(HB)%*%solve(mycombination%*%vcov(myModel)%*%t(mycombination))%*%HB
p=1-pf(as.numeric(Fobs),nrow(mycombination),denDF)
a=c("p"=p,"Fobs"=as.numeric(Fobs),"numDF"=nrow(mycombination))
myResult=list("p"=p,"Fobs"=as.numeric(Fobs),"numDF"=nrow(mycombination),"cl"=HB,"EE"=EE)
}
}
}

```

```

if (class(myModel)=="lm") {
  mycombination=as.matrix(mycombination)
  denDF=myModel$df
  HB=mycombination%%myModelCoefficients
  Fobs=t(HB)%%solve(mycombination%%vcov(myModel)%%t(mycombination))%%HB
  if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%%vcov(myModel)%%t(mycombination))) else EE=NA
  Fobs:=Fobs/nrow(mycombination)
  p=1-pf(as.numeric(Fobs),nrow(mycombination),denDF)
  a=c("p"=p,"Fobs"=as.numeric(Fobs),"numDF"=nrow(mycombination))
  myResult=list("p"=a[1],"Fobs"=a[2],"numDF"=a[3],"ci"=HB,"EE"=EE)
}

if (((class(myModel)=="glm")
|(class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")))
{
  mycombination=as.matrix(mycombination)
  if (class(myModel)=="glm") {
    HB=mycombination%%myModelCoefficients
    if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%%summary(myModel)$cov.scaled%%t(mycombination))) else EE=NA
    denDF=summary(myModel)$df.residual
    Fobs=t(HB)%%solve(mycombination%%summary(myModel)$cov.scaled%%t(mycombination))%%HB
  }
  if (class(myModel)=="mer") {
    HB=mycombination%%myModelCoefficients
    if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%%as.matrix(summary(myModel)@vcov)%%t(mycombination))) else EE=NA
    Fobs=t(HB)%%solve(mycombination%%as.matrix(summary(myModel)@vcov)%%t(mycombination))%%HB
  }
}

if ((class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) {
  HB=mycombination%%myModelCoefficients
  if (nrow(mycombination)==1) EE=sqrt(as.numeric(mycombination%%as.matrix(vcov(myModel))%%t(mycombination))) else EE=NA
  Fobs=t(HB)%%solve(mycombination%%as.matrix(vcov(myModel))%%t(mycombination))%%HB
}

p=1-pchisq(as.numeric(Fobs),nrow(mycombination))
a=c("p"=p,"Chi"=as.numeric(Fobs),"numDF"=nrow(mycombination))
myResult=list("p"=a[1],"Fobs"=a[2],"numDF"=a[3],"ci"=HB,"EE"=EE)
}
myResult
}

```

```

RSquare<-function(myModel){

if ((class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
y=myModel@resp$y

if
(!((class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
) {

  mydata=eval(myModel$call[[3]])

  y=mydata[,as.character(formula(myModel))[[2]]]

}

if (class(myModel)=="lme") {

  index=as.numeric(rownames(as.data.frame(myModel$fitted)))

  r=cor(myModel$fitted,y[index])

}

if (class(myModel)=="gls"){

  index=as.numeric(rownames(as.data.frame(myModel$fitted)))

  r=(cor(myModel$fitted,y[index]))

}

if (class(myModel)=="lm") {

  index=as.numeric(rownames(as.data.frame(myModel$fitted)))

  r=(cor(myModel$fitted,y[index]))

}

if (class(myModel)=="glm") {

  index=as.numeric(rownames(as.data.frame(myModel$fitted)))

  r=(cor(myModel$fitted,y[index]))

}

if
(class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
{

  index=as.numeric(rownames(as.data.frame(myModel@resp$mu)))

  r=(cor(myModel@resp$family$linkinv(getME(myModel,"X")%*%fixef(myModel)),y[index]))

}

if
((class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) {

  index=as.numeric(rownames(as.data.frame(myModel@resp$mu)))

  r=(cor(myModel@resp$family$linkinv(getME(myModel,"X")%*%fixef(myModel)),y[index]))

}

rr=as.data.frame(t(r*r))

colnames(rr)=paste("R2",seq(0,length(rr)-1,sep="_"))

rr

}

```

```

UpdateFormula=function(myModel,myModelTerm){

result=paste(as.character(formula(myModel))[[2]],as.character(formula(myModel))[[1]],as.character(formula(myModel))[[3]],"-
",myModelTerm,sep="");

result

}

#-----

ReduceModel=function(myModel,myModelTerm) {

mynewModel=update(myModel,form=UpdateFormula(myModel,myModelTerm))

result=anova(myModel,mynewModel)

if (class(myModel)=="mer") result=list("UpdatedModel"=mynewModel,"Test"=c("Term"=myModelTerm,"Chi-
square"=abs(2*result[,4]*%*%c(-1,1)), "df"=result[2,6], "p-value"=1-pchisq(abs(2*result[,4]*%*%c(-1,1)),result[2,6])))

if ((class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
result=list("UpdatedModel"=mynewModel,"Test"=c("Term"=myModelTerm,"Chi-square"=result[2,6], "df"=result[2,7], "p-
value"=result[2,8]))

result

}

#-----

dfMixedModelTypeIAnova=function(myModel,TheTitle="Building Anova Type I Table") {

f=TheModelTerms(myModel)

pb = winProgressBar(title=TheTitle, min = 0, max = length(f), width = 300)

mynewModel=myModel

result=c()

for (i in (length(f):1)) {

r=ReduceModel(mynewModel,f[i])

mynewModel=r$UpdatedModel;

result=rbind(r$Test,result)

setWinProgressBar(pb, (length(f)-i+1), title = TheTitle)

}

Sys.sleep(1)

close(pb)

result=as.data.frame(result)

result[,2]=as.numeric(as.character(result[,2]))

result[,3]=as.numeric(as.character(result[,3]))

result[,4]=as.numeric(as.character(result[,4]))

result

}

```

```

RebuildFactorLevels=function(myData,elfactor){
  unique(paste(elfactor,levels(myData[,elfactor]),sep=""))
}
#-----

ValidFactorsAndCovariables=function(myModel,myData){
  if ((class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
  cl=(myModel@call) else cl=(summary(myModel)$call)

  clf=format(cl)

  for (a in (c(" ",":","(",")","~","|"))){
    {
      l=c()

      for (i in (1:length(clf))) {
        l=c(l,split(clf[i],a,fixed=T))
      }

      clf=unique(unlist(l))
    }

    intersect(clf,colnames(myData))
  }
}
#-----

validCases=function(myModel,myData){
  seq(nrow(myData))[complete.cases(myData[,ValidFactorsAndCovariables(myModel,myData)])]
}

validCasesforAdjustedValues=function(myModel,myData){
  clf=TheModelTerms(myModel)

  for (a in (c(" ",":","(",")","~","|"))){
    {
      l=c()

      for (i in (1:length(clf))) {
        l=c(l,split(clf[i],a,fixed=T))
      }

      clf=unique(unlist(l))
    }

    for (a in (c(" ",":","(",")","~","|")) l=c(l,split(clf,a,fixed=T)); clf=unique(unlist(l))

    myvars=intersect(clf,colnames(myData))

    seq(nrow(myData))[complete.cases(myData[,myvars])]
  }
}

```



```

validCasesforPrediction=function(myModel,myData,myModelTerms=NULL){

if ((class(myModel)=="mer")|(class(myModel)=="glmerMod")|(class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
{cl=(myModel@call);

clf=as.character(unlist(formula(cl)[[3]])) else clf=as.character(unlist(formula(myModel)[[3]]))

for (a in (c(" ", ":", "(", ")", "~", "|"))) {

l=c()

for (i in (1:length(clf))) {

l=c(l, strsplit(clf[[i]], a, fixed=T))

}

clf=unique(unlist(l))

}

for (a in (c(" ", ":", "(", ")", "~", "|"))) l=c(l, strsplit(clf, a, fixed=T)); clf=unique(unlist(l))

myvars=intersect(clf, colnames(myData))

seq(nrow(myData))[complete.cases(myData[, myvars])]

}

#-----
-----

lme4VarCompExtaction=function(mimodelo){

tabla=as.data.frame(summary(mimodelo)$varcor )

d=tabla[!is.na(tabla[,3]),5]

R = diag(length(unique(tabla[!is.na(tabla[,3]),3])))

R[upper.tri(R,diag=T)] =NA

R[lower.tri(R,diag=T)] =d

rownames(R)=unique(tabla[!is.na(tabla[,3]),3])

V=tabla[is.na(tabla[,3]),c(-3)]

colnames(V)=c("RndEff", "Param", "Var", "SD")

if
((class(mimodelo)=="lmerMod")|(class(mimodelo)=="merModLmerTest")){V=cbind(V,"Var%"=V[, "Var"]/sum(V[, "Var"],na.rm=T)*100);
V=V[,c(1,2,3,5,4)]}

if (nrow(R)>0){

vacias=c()

colnames(R)=paste("Corr",seq(ncol(R)),sep=" ")

for (i in (1:(nrow(V)-nrow(R)))) vacias=rbind(vacias,rep(NA,ncol(R)))

M=cbind(V,rbind(vacias,R)) else M=V

M

}

```

```

MatrizDeDiferenciasDeMediasEstandarizadas=function(myModel,myData,myTerm,Mmatrix,alfa=0.001){
  require(Matrix)
  factores=strsplit(myTerm, ".")
  coeficientes=TheModelCoefficients(myModel)
  f=factores[[1]]
  lista=paste(rep(f[1],nrow(myData)),myData[,f[1]],sep="")
  if (length(f)>1) for (i in (2:length(f))) {
    lista=paste(lista,paste(rep(f[i],nrow(myData)),myData[,f[i]],sep=""),sep=":")
  }
  n=as.vector(table(lista))
  lista=unique(lista)
  tabla=as.vector(Linear.combination.to.estimate.a.mean(lista[1],Mmatrix,coeficientes))
  if (length(lista)>1)
    for (i in (2:length(lista))) table=rbind(tabla,as.vector(Linear.combination.to.estimate.a.mean(lista[i],Mmatrix,coeficientes)))
  rownames(tabla)=lista
  tabla=Matrix(tabla,sparse=T)
  if (class(myModel)=="lme" ) covar=tcrossprod(tabla**myModel$varFix,tabla)
  if (class(myModel)=="gls" ) covar=tcrossprod(tabla**myModel$varBeta,tabla)
  if (class(myModel)=="lm" ) covar=tcrossprod(tabla**vcov(myModel),tabla)
  if (class(myModel)=="glm" ) covar=tcrossprod(tabla**summary(myModel)$cov.scale,tabla)
  if (class(myModel)=="mer"
  {covar=tcrossprod(tabla**summary(myModel)@vcov,tabla);colnames(covar)=rownames(tabla);rownames(covar)=rownames(tabla)
  }
  if (class(myModel)=="glmerMod")
  {covar=tcrossprod(tabla**vcov(myModel),tabla);colnames(covar)=rownames(tabla);rownames(covar)=rownames(tabla)}
  if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest"))
  {covar=tcrossprod(tabla**vcov(myModel),tabla);colnames(covar)=rownames(tabla);rownames(covar)=rownames(tabla)}
  smedias=(tabla**coeficientes)
  medias=as.vector(smedias[,1]);names(medias)=rownames(smedias);rm(smedias)
  D=Matrix(as.matrix(dist(medias)))
  orden=sort(medias,decreasing=TRUE,index.return=TRUE)$ix
  medias=medias[orden]
  covar=covar[orden,orden]
  D=D[orden,orden]

  TheTitle='Calculating standardized mean differences'
  active.pb=(nrow(D)>40)

```

```

if (active.pb==TRUE) {wpb = winProgressBar(title=TheTitle, min = 0, max = nrow(D), width = 400); Sys.sleep(0.5)}

m=Matrix(diag(rep(-1,nrow(D))),sparse=TRUE)

vardiff=c()

k=1

for (k in (1:nrow(D))) {

  M=m

  M[,k]=1

  M=Matrix(M,sparse=TRUE)

  vv=diag(tcrossprod(crossprod(t(M),covar),M))

  S=1/sqrt(vv)

  vardiff=c(vardiff,vv[-k])

  D[k,]<-D[k,]*S

  if (active.pb==TRUE) setWinProgressBar(wpb,k,title=paste(TheTitle,"(",as.character(k),"/",as.character(nrow(D)),")",sep="")) }

if (active.pb==TRUE) close(wpb)

rownames(D)=c()

cn=colnames(D)

l=strsplit(myTerm,".")[1]

for (i in 1:length(l)) cn=sub(l[i],"",cn)

colnames(D)=cn;

rm(cn)

SEMD=sqrt(mean(vardiff,na.rm=T))

listofmeans=paste(rep(f[1],nrow(myData)),myData[,f[1]],sep="")

if (length(f)>1) for (i in (2:length(f))) {

  listofmeans=paste(listofmeans,paste(rep(f[i],nrow(myData)),myData[,f[i]],sep=""),sep=":")

}

listofmeans=unique(listofmeans)

if (class(myModel)=="lm") nn=myModel$df.residual

if (class(myModel)=="glm") nn=summary(myModel)$df.residual

if (class(myModel)=="gls") nn=myModel$dim$N-myModel$dim$p

if (class(myModel)=="lme") nn=FindGreatestDFForAGivenListOfMeansForA_lme_model(myModel,c(listofmeans[1],listofmeans[2]))

if (class(myModel)=="mer") nn=myModel@dim["n"]-myModel@dim["p"]

if (class(myModel)=="glmerMod") nn=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))

if ((class(myModel)=="lmerMod")|(class(myModel)=="merModLmerTest")) nn=nrow(getME(myModel,"X"))-ncol(getME(myModel,"X"))

if (nn==(-1)) nn=theModelResidualDF(myModel)

result=list(medias=medias,n=n,distancias_estandarizadas=D,covarmedias=covar,AvgSEMeansDiff_and_AvgDMS=list(AvgSEMD=SEMD,AvgDMS=qt((1-alfa/2),nn)*SEMD,DF=nn))

result}

```

```

lmeLSD=function(mean1,mean2,MDE,gle){
na=colnames(MDE$covarmedias)
d=abs(MDE$distancias_estandarizadas[na==mean1,na==mean2])
2*(1-pt(d,gle))
}

```

*#LSD Fisher con correccion de Bonferroni*

```

LSD=function(MDE, gle){
na=colnames(MDE$covarmedias)
mean1=c()
mean2=c()
resultados=data.frame(na)
resultados$indices=rep(1,nrow(resultados))
resultados$pvalor=rep(1,nrow(resultados))
resultados$medias=round(MDE$medias,2)
for (l in (1:(length(na)-1))){
resultados$pvalor[l+1]=round(lmeLSD(na[l],na[l+1],MDE,gle),3)
resultados$pvalor=p.adjust(resultados$pvalor, method="bonferroni")
if (resultados$pvalor[l+1]<0.01){
resultados$indices[l+1]=resultados$indices[l]+1
} else{resultados$indices[l+1]=resultados$indices[l]}
}
resultados
}

```

### 5.3.4. Error de Tipo I y II

```
#Calculo del error de tipo I para el modelo multivariado (4.6)
ECtest=read_excel("C:\\.\\.\\.\\ECtest.xlsx")
ECtest=data.frame(ECtest)
cols.to.factor =sapply( ECtest, function(col) length(unique(col)) < 30)
ECtest[ cols.to.factor] =lapply(ECtest[ cols.to.factor] , factor)
ECtest$Estados_num=as.numeric(ECtest$Estados_num)
ECtest$Identidad=as.numeric(ECtest$Identidad)
ECtest$SectorEC_I_EC=interaction(Ectrain[,c("SectorEC", "I_EC")])
x=c(1,2,3,4)
q =list()
for (i in (1:4)) {
  q[[i]]= subset(ECtest, ECtest$Estados_num==x[i])
  q[[i]]$Estado=as.factor(round(runif(nrow(q[[i]]), 1,4),0))
  levels(q[[i]]$Estado)=c("HIPER", "OC", "OPTO", "OA" )
  remuestreo=list()
  tabla=list()
  modelo=list()
  solucion=NULL
  count=0
  for (i in (1:4)){ remuestreo[[i]]= bootstraps(q[[i]], times = 125)
  tabla[[i]]=list()
  modelo[[i]]=list()
  for (j in (1:125)){
    try({tabla[[i]][[j]]=as.data.frame(remuestreo[[i]]$splits[[j]])
    modelo[[i]][[j]]=lme(Respuesta~1+I_EC+Estado:I_EC+CATEdad:I_EC+Sexo:I_EC
      ,random=list((SectorEC_I_EC=pdIdent(~1)
      ,Identidad=pdIdent(~1)
      ,Identidad=pdIdent(~Estado:I_EC-1))
      ,weights=varComb(varIdent(form=~1|Identidad))
    ,method="REML" ,control=lmeControl(niterEM=150, msMaxIter=200)
    ,na.action=na.omit,
    data=tabla[[i]][[j]])
  }
}
```

```

myModel=modelo[[i]][[j]]

myData=tabla[[i]][[j]]

myTerm="Estado:l_EC"

Mmatrix=CalculateMmatrix(myModel, myData);

MDE=MatrizDeDiferenciasDeMediasEstandarizadas(myModel,myData,myTerm,Mmatrix);

gle=TheModelResidualDF(myModel);

a=length(unique(LSD(MDE,gle)$indices))

if(a>2) {a=count+1} else{

a=count}

}, silent=TRUE)

solucion=rbind(solucion, a)

}

}

Error!= mean(solucion)

#-----

#Calculo del error tipo II para modelo multivariado (4.6)

q=list()

h=list()

for (i in (1:3)) {

h[[i]]=list()

q[[i]]= subset(ECtest, ECtest$Estados_num==x[i])

for (j in ((i+1): 4)){

p[[j]]=subset(ECtest, ECtest$Estados_num==x[j])

h[[i]][[j]]=rbind(q[[i]], p[[j]])

niveles=unique(as.character(h[[i]][[j]]$Estado))

h[[i]][[j]]$Estado=factor(h[[i]][[j]]$Estado, levels=niveles)}

```

```

solucion=NULL

count=0

remuestreo=list()

n=c()

tabla=list()

modelo=list()

for (i in 1:3){

  for (j in (i+1):4){

    remuestreo= bootstraps(h[[i]][[j]], times = 125)

    for (k in 1:125){

      try({tabla[[k]]=as.data.frame(remuestreo$splits[[k]])

      modelo[[k]]=lme(Respuesta~1+l_EC+Estado:l_EC+CATEdad:l_EC+Sexo:l_EC

        ,random=list((SectorEC_l_EC=pdIdent(~1)

          ,Identidad=pdIdent(~1)

          ,Identidad=pdIdent(~Estado:l_EC-1))

        ,weights=varComb(varIdent(form=~1|Identidad))

      ,method="REML",control=lmeControl(niterEM=150, msMaxIter=200)

        ,na.action=na.omit,

        data=tabla[[k]])

      myModel=modelo[[k]]

      myData=tabla[[k]]

      myTerm="Estado:l_EC"

      Mmatrix=CalculateMmatrix(myModel, myData);

      MDE=MatrizDeDiferenciasDeMediasEstandarizadas(myModel,myData,myTerm,Mmatrix);

      a=length(unique(LSD(MDE,gle)$indices))

      if(a==2) {a=count+1} else{

        a=count}

      }, silent=TRUE)

      solucion=rbind(solucion, a)}

    }

  }

ErrorII=mean(solucion)

```

### 5.3.5. Predicciones en el modelo multivariado

```
bolker_ci <- function(model, newdat, pred_int = FALSE, conf_level = 0.95) {  
  if(class(model) != "lme") {  
    stop("works for lme-models only")  
  }  
  z <- round(qnorm((1-conf_level)/2, lower.tail = FALSE), 2)  
  newdat$pred <- predict(model, newdat, level = 0)  
  Designmat <- model.matrix(formula(model)[-2], newdat)  
  predvar <- diag(Designmat %*% vcov(model) %*% t(Designmat))  
  newdat$sse <- sqrt(predvar)  
  newdat$sci_l <- newdat$pred - z*newdat$sse  
  newdat$sci_h <- newdat$pred + z*newdat$sse  
  if(pred_int == TRUE) {  
    newdat$sse2 <- sqrt(predvar+model$sigma^2)  
    newdat$predint_l <- newdat$pred - z*newdat$sse2  
    newdat$predint_h <- newdat$pred + z*newdat$sse2  
  }  
  newdat  
}
```



### 5.3.6. Divergencia de Jensen-Shanon

```
#Programa BPLZ.c escrito por Steeve Zozor, funcion de interfaz de Matlab y C
void mexFunction(int nlhs, mxArray *plhs[], int nrhs, const mxArray *prhs[]){
    void BPLZ(double*, int, int, double, int, double*, int*); /* prototype of the function called in C */
    double *signal; /* matrix of the signals */
    int d; /* embedding dimension */
    int tau; /* delay */
    int Tb; /* duration of the (bloc of) signal */
    double conf; /* confidence interval */
    double c; /* LZ complexity on BP vectors */
    int Rt; /* Sequence of ranks */
    /* matlab usage: [c Rt] = BPLZ(signal,d,tau,conf) */
    /* check for proper number of arguments */
    if((nrhs!=3)*(nrhs!=4)){mexErrMsgTxt(" not proper input arguments ");}
    if((nlhs!=1)*(nlhs!=2)){mexErrMsgTxt(" not proper output");}
    /* Create matrix for the first return argument */
    plhs[0] = mxCreateDoubleMatrix(1,1,mxREAL);
    /* Assign pointers to each input and output */
    signal = mxGetPr(prhs[0]);
    d = mxGetScalar(prhs[1]);
    tau = mxGetScalar(prhs[2]);
    c = (double *) mxGetPr(plhs[0]);
    if(nrhs==4){conf = mxGetScalar(prhs[3]);}else{conf = 0.0;}
    /* get the length of the signal */
    Tb = mxGetM(prhs[0]); /* line = times */
    if(Tb==1){mexErrMsgTxt(" the vector must be one column");}
    if(nlhs==2)
    {
        plhs[1] = mxCreateNumericMatrix(d,Tb-(d-1)*tau,mxINT32_CLASS,mxREAL);
        Rt = (int *) mxGetPr(plhs[1]);
    }
    else{Rt = (int*)malloc(d*(Tb-(d-1)*tau)*sizeof(int));}
```

```

/* call of the routine to evaluate the complexity */

/printf("conf = %f\n",conf);/

BPLZ(signal,d,tau,conf,Tb,c,Rt);

if(nlhs==1){free(Rt);}

}

/*****/

/* Function to compare test the equality between two vectors */

int EqualVectors(int *Vector1, int *Vector2, int d)

/* Inputs: Vector1, Vector2 = address of the vectors to be compared */

/*          d = dimension of the vectors */

/* Output:  eq = 1 if the vectors are equal, 0 otherwise */

/* Remind that C put the signal in line order contrary to matlab */

/* i.e. in matlab (i,j) --> (j-1) * nc + j */

/* in C (i,j) --> (i-1) * nc + j-1 */

/* !!! it supposed here that the vectors are line vectors !!! */

{

int j = 0;

int eq = 1; /* answer of the equality test : before tested they are assumed equal */

while ((j < d) && (eq)){

if ( (Vector1+j) == *(Vector2+j)) {j++;} / the current components are equal*/

else {eq = 0;} /* the current component are not equal => the vectors are not equal (unnecessary to compare the next components)*/

}

return(eq);

}

/*****/

/* Function to built the rank vector (brut force) */

void Rank(double *Vector, int d, double conf, int *Rk)

/* Inputs: Vector = address of the vector to rank */

/*          d = dimension of the vector */

/*          conf = confidence interval (test of the equality) */

/* Output: Rk = address of the vector of the ranks of the components of Vector */

{

int i, j;

double Vi;

```

```

for (i = 0; i < d; i++)/* current component, to compare to that (j) already sorted*/
{
    (Rk+i) = 0; /* initially the rank is assumed 0*/
    for (j = 0; j < i; j++)
    {
        if ((Vector+i) < *(Vector+j)-conf) {((Rk+j)) = ((Rk+j))+1;} /* lower => increase rank j*/
        else {((Rk+i)) = ((Rk+i))+1;} /* higher, increases its own rank*/
    }
}

/*printf("Y = [ "); for (i=0; i<d; i++){printf("%f ",(Vector+i));} printf("]\n");*/
/*printf("R = [ "); for (i=0; i<d; i++){printf("%d ",(Rk+i));} printf("]\n\n");*/
}

/*****

/* Function to up-to-date a phase vector and its rank vector */
void UtDRank(double *Vector, int d, int tau, double conf, double NewV, int *RkN)

/* Inputs: Vector = address of the "old" vector already ranked */
/*      d = dimension of the vector */
/*      tau = delay (needed to search for the old rank vector that is up-to-dates) */
/*      conf = confidence interval (test of the equality) */
/*      NewV = point that arrive and induces the up-to-date */
/* Output: RkN = address of the vector of the ranks up-to-dated */

{
    int i;

    int nr = d*tau; /* -nr = shift for searching the old rank vector*/
    double OldV = Vector; /* point that will quit the vector of values*/
    int pin = 0; /* rank of the new point */
    for (i = 0; i < d-1; i++)/* for component i*/
    {
        (Vector+i) = *(Vector+i+1); /* up to date of the vector of values*/
        (RkN+i) = *(RkN-nr+i+1); /* copy of the old rank*/
        if (NewV >= (Vector+i)-conf) /*if the new point is higher: with the margin due to possible numeric noise*/
        {

```

```

pin++;/* the rank increases*/

    if (OldV < (Vector+i)+conf){(RkN+i) = (RkN+i-1);}/* and if the old was lower => rank i decreases*/
    } else {if (OldV > (Vector+i)+conf){(RkN+i) = (RkN+i+1);}/* conversely : new lower, old higher, the rank increases*/
    }

(Vector+d-1) = NewV; /*up-to-date of the last point of the values*/

(RkN+d-1) = pin; /* pin is teh rank of the new point*/

/*printf("Y = [ "); for (i=0; i<d; i++){printf("%f ",(Vector+i));} printf("]\n");*/

/*printf("R = [ "); for (i=0; i<d; i++){printf("%d ",(RkN+i));} printf("]\n\n");*/

}

/*****/

/* Function to evaluate the Lempel-Ziv complexity of a sequence of a rank vector */

/*          (Bandt and Pompe construction of the rank vectors) */

void BPLZ(double *signal, int d, int tau, double conf, int Tb, double *c, int *Rt)

/* Inputs: signal = address of the signal */

/*      d = embedding dimension */

/*      tau = delay */

/*      conf = confidence interval (test of the equality) */

/*      Tb = duration of the bloc of signal */

/* Outputs: c = complexity of the sequence */

/*      Rt = sequence of rank vectors */

{int i, j, k, km, l, eq;

int t, u;

double *Y;

/int *Rt;/

int N = Tb-(d-1)*tau;

Y = (double*)malloc(d*tau*sizeof(double));

/*Rt = (int)malloc(d*N*sizeof(int)); */

/* Initializations */

for (j = 0; j < tau; j++)

{for (i = 0; i < d; i++)

    { ((Y+i+j*d) = ((signal+i*tau+j)));

    }

Rank(Y+j*d, d, conf, Rt+j*d);

}

t = tau - 1;

```

```

step0 : *c = 1.0; l = 1; if (tau > 1){goto step2;}

step1 : if ( t >= (Tb-(d-1)*tau) ){goto stop;}

else

{ t = t+1;

  u = t - (t/tau)tau; /* la division est entière*/

  UtDRank(Y+u*d, d, tau, conf, *(signal+t+(d-1)*tau), Rt+t*d);

  goto step2;

}

step2 : j = 0; k = 0; km = 1; /* we assume the production is km = 1 (max length km), j is the pointer*/

step3 : eq = EqualVectors(Rt+(l+k)d , Rt+(j+k)*d , d); /*test of the equality between vectors (-1 because the indices stars from 0 in C)*/

if (eq == 1) /* we have still a reproduction*/

{

  k++; /*still a reproduction/

  if ( (l+k) > t ) /* we ended the sequence up to the current point */

  {

    if ( (l+k) >= (Tb-(d-1)tau) ) {*c = (*c) + 1.0; goto stop;} /*addition of the last word*/

    else {

      t = t+1;

      u = t - (t/tau)tau; /* la division est entière*/

      UtDRank(Y+u*d, d, tau, conf, *(signal+t+(d-1)*tau), Rt+t*d);

      goto step3;

    } /* to make the production that is not a reproduction*/

  }

  else {goto step3;}

}

else /* we face a production that is not a reproduction*/

{

  if ( (k+1) > km ) {km = k+1;} /*the pointer is better since the production is larger -> up to date nmax */

  j++; /*test of the next point as a possible pointer/

  if(j == l) /* all the possible pointer are tested*/

  {

    c = (*c) + 1.0; /* the exhaustive production is made : c */

    l = l + km; /*we skipto the symbol following the exhaustive word*/

    if ( l > t )

```

```

{
    if ( l >= (Tb-(d-1)tau) ){goto stop;} /* the sequence ended */
    else
        { t = t+1;
          u = t - (t/tau)tau; /* la division est entière */
          UtDRank(Y+u*d, d, tau, conf, *(signal+t+(d-1)*tau), Rt+t*d);
          goto step2;
        }
    }
    else {goto step2;} /* we start for seeking the next optimal production step */
}
    else {k = 0; goto step3;} /* we start again testing the next possible pointer */
stop : free(Y); /*free(Rt)*/
}

```

### #Funcion de Matlab para calcular el radio de log-verosimilitud

Function[H, DJS]=H\_DJS(x, N1, d,  $\tau$ )

%% Para este programa es necesario compilar con un compilador de C el programa BPLZ.c

%% haciendo mexBPLZ.c

%% x=señal de concatenación entre estados A y B

%% N1 longitud de la señal para el estado A

%% d dimensión de la incrustación

%% tau el retardo

%% H entropía de permutación normalizada

N=length(x);

A=x(1:N1,:);

B=x(N1+1:N, :);

N2=length(B)

D1=int32(d^(0:d-1)\*ones(1, N1-(d-1)\* $\tau$ ));

D2=int32(d^(0:d-1)\*ones(1, N2-(d-1)\* $\tau$ ));

D=int32(d^(0:d-1)\*ones(1, N1-(d-1)\* $\tau$ \*2));

[c01, RR1]=BPLZ(A, d,  $\tau$ );

[c02, RR2]=BPLZ(B, d,  $\tau$ );

RR=[RR1, RR2];

%% Entropia de permutacion

RR1=sum(RR1\*D1,1) %calculo de vectores de rango del primer estado

RR2=sum(RR2\*D2,1) %calculo de vectores de rango del segundo estado

RR=sum(RR\*D,1) %calculo de vectores de rango del vector concatenado

P1=hist(RR1, d^d)/length(RR1); Pn1=P1(P1>0); %% aca esta la frecuencia de permutacion de los vectores de permutacion del primer estado

P2=hist(RR2, d^d)/length(RR2); Pn2=P2(P2>0); %% aca esta la frecuencia de permutacion de los vectores de permutacion del segundo estado

P=hist(RR, d^d)/length(RR); Pn=P(P>0); %% aca esta la frecuencia de permutacion del vector concatenado

H=sum(Pn\*log2(Pn));

DL=2\*(N-(d-1)\* $\tau$ \*2)\*log2\*(-sum(Pn\*log2(Pn)))+(N1-(d-1)\* $\tau$ )/(N-(d-1)\* $\tau$ \*2)\*(sum(Pn1\*log2(Pn1))+

(N2-(d-1)\* $\tau$ )/(N-(d-1)\* $\tau$ \*2)\*(sum(Pn2\*log2(Pn2))) %% radio de log-verosimilitud