DEPARTAMENTO DE INVESTIGACIÓN Y DOCTORADO

Instituto Tecnológico de Buenos Aires

Herramientas de minería de datos para estudios longitudinales con aplicaciones biomédicas

AUTOR: Lic. Lucio José Pantazis

DIRECTOR: Dr. Rafael Antonio García Galiñanes

TESIS PRESENTADA PARA OPTAR AL TÍTULO DE
DOCTOR EN INGENIERÍA

Jurado

Dra. María Paula Bonomini Dra. María Juliana Gambini Dr. Juan Miguel Medina

CIUDAD AUTÓNOMA DE BUENOS AIRES
Febrero 2024

Lic. Lucio José PANTAZIS: Herramientas de minería de datos para estudios longitudinales con aplicaciones biomédicas. *Tesis presentada como requisito parcial para acceder al grado de* ***DOCTOR EN INGENIERÍA*** *del Instituto Tecnológico de Buenos Aires.*

"Paso a Paso"
Reinaldo Carlos Merlo

Agradecimientos

Es difícil resumir a todas las personas a las que les estoy agradecida por este logro. Si bien hubo mucho trabajo individual, nunca hubiera llegado a este punto sin el enorme apoyo tanto emocional como profesional. Siempre digo que sin la gente que me rodea, no sería la persona que soy. No recomiendo a nadie encarar una tarea tan titánica en solitario. Se sufre mucho (de forma innecesaria) y se pierde la oportunidad de aprender todo lo que las personas tienen para ofrecer.

Como siempre, a mis viejos, Silvia y Ricardo, que en su incondicionalidad e infinita generosidad, han sido una referencia personal admirable. Si logro reproducir al menos en un mínimo porcentaje semejantes cualidades, estaré más que satisfecho desde lo personal.

A Sol, mi esposa, por darme fuerza cuando más me costaba. Cuando más arreciaba la frustración, siempre pude contar con su hombro, su abrazo y su contención. Y aún gestante, agradezco a Vito por venir con un paper bajo el brazo, pero sobre todo porque me inyectó energías para poder cerrar un proceso que parecía no querer terminar. Hoy puedo recibirlo con el orgullo de haber sobrepasado una etapa muy larga y difícil.

A Rafael, por animarse a hacer algo nuevo y acompañar este proceso. Sé que para él ha sido difícil sostener este proceso. A Ricardo, por acercarme al ITBA. A Carolina, por darme un marco para resurgir cuando todo parecía perdido. Hubo varios momentos en los que sin el apoyo de ellos, podría haberse cortado este trabajo antes de terminar.

A mi familia, Javu, Julia, la Yiayia, la Noni, Papú, Cele, Lau, Sofi, Carlos y tantos otros tíos y primos por estar y bancarme siempre en mis múltiples frustraciones.

A mis amigos, Santi, Manu, Barret, Pablito, Fer, Guido, Mariana, Andre, Lu, Belu, Lau, Euge, Agus, Roberto, por ofrecer su compañía, la siempre necesaria risa y la crítica cuando entro en círcuitos que no me permiten avanzar.

A Gabi y Flor por las charlas, por los almuerzos, por el apoyo, por la amistad, por ser de fierro.

A Mati, Echo, Juan y José, los de la A-2-6, por enseñarme que los que yo pensaba que eran sólo mis problemas, eran problemas comunes. Fueron un acompañamiento que apuntalaron significativamente mi día a día.

A mis compañeros de docencia, las Patris, Mari, Demian, Paco, Pablo por cubrirme siempre que hiciera falta para poder dedicarme a este trabajo. A Andrés, no sólo por compañero de docencia, sino por ser amigo de años.

A los profesores que conocí en este doctorado y me enseñaron muchísimo, Pedro, Juliana, Juan Manuel, María Laura. A Paula, Rodrigo, Giuliana, Matías y Marcela, otras grandes personas que

conocí durante estos años. A Gustavo por encargarse de innumerables gestiones.

A Salomón y Arturo, con los que pude desarrollar las herramientas para preservar mi salud mental en los momentos de mayor frustración e incertidumbre. Recomiendo a cualquier persona que requiera tanto esfuerzo mental en el día a día, que acuda a ayuda profesional.

Al ITBA y CONICET por darme la oportunidad de formarme, el lugar y el material para realizar este trabajo. A la UBA por darme la formación de base, los conocimientos adquiridos fueron clave para poder adaptarme a distintas situaciones y reformular la tesis cuando hubo que hacerlo.

Resumen

En esta tesis se describen dos desarrollos para estudios longitudinales con aplicaciones biomédicas. Estos desarrollos utilizan herramientas matemáticas muy distintas, por lo tanto, parece difícil ubicar ambos trabajos dentro de un mismo marco. Sin embargo, ambos trabajos tienen su fundamento en la interpretación de los coeficientes estadísticos y la evolución temporal de los datos, explotando las características de los datos longitudinales.

El primer desarrollo consiste de una aplicación algorítmica a una base longitudinal de pacientes diabéticos en la que se analiza secuencialmente la expresión de algunos genes estratégicos. Considerando esta base, se logró agrupar en 3 grupos a los pacientes por la variación en su expresión genética, presentando uno de los grupos características clínicas distintas a los grupos restantes. Además, en nuestro trabajo se desarrollaron algoritmos que permiten obtener nuevas características a partir de bases de datos donde las herramientas usuales de detección de patrones no son recomendables.

Este descubrimiento es relevante ya que las características de los grupos no estuvieron involucradas en el agrupamiento de pacientes y por lo tanto, presenta una potencial asociación de la cual no se tenía información.

Además, nuestra propuesta incluye un nuevo enfoque sobre el problema que permite evaluar muchos algoritmos de clustering y comparar sus resultados. Considerando estas nuevas alternativas, hemos logrado encontrar una forma de agrupar observaciones que, en las bases de datos seleccionadas, arroja mejores resultados que otros métodos preestablecidos.

El segundo desarrollo, es el de un algoritmo de detección de trayectorias atípicas para bases con alta variabilidad en las respuestas individuales, pero que siguen una cierta estructura que se puede describir a través de un modelo estadístico que se adapta a dicha variabilidad.

A partir de este modelo se detectan de forma simultánea diversas características de los individuos con evoluciones temporales que no acompañan dicho modelo estadístico. En nuestros experimentos y simulaciones, hemos obtenido mejores resultados de detección que otros algoritmos establecidos.

Contenido

Índice de figuras

Índice de tablas

Símbolos y abreviaturas

Letras

Símbolo	Descripción
I	Cantidad de individuos del estudio
J	Cantidad de mediciones
J_i	Cantidad de mediciones del individuo i
$\mathbf{Y}$	Variable de respuesta
$\mathbf{Y}_i$	Trayectoria de respuesta para el individuo i
$\mathbf{t}$	Variable temporal
$\mathbf{X}$	Matriz de covariables
P	Cantidad de covariables
ε	Errores de medición
$\mathbf{b}_i$	Vector de efectos aleatorios
$\mathbf{Z}$	Matriz de covariables asociadas a efectos aleatorios
Q	Cantidad de covariables asociadas a efectos aleatorios
$\mathbf{R}$	Matriz de covarianza de los errores
$\mathbf{G}$	Matriz de covarianza de los efectos aleatorios
N	Cantidad total de respuestas.
H	Cantidad de parámetros asociados a la covarianza
D^P	Perturbaciones puntuales en una trayectoria de respuesta
D^R	Perturbaciones en un vector de efectos aleatorios
$\mathcal{N}_J$	Distribución normal multivariada de J componentes
$\mathcal{E}$	Distribución exponencial
$\mathcal{U}$	Distribución uniforme

Letras Griegas

Símbolo	Descripción	Unidades
$\vec{\beta}$	Vector de efectos fijos	
$\vec{\theta}$	Vector de parámetros asociados a la covarianza	

Símbolo	Descripción	Unidades
$\mathbf{\Sigma}_i$	Matriz de covarianza de la trayectoria de respuesta $\mathbf{Y}_i$	

Subíndices

Subíndice	Descripción
i	Índice asignado a los individuos del estudio
j	Número de medición
p	Número de covariable
q	Número de covariable asociada a efectos aleatorios

Abreviaturas

Abreviatura	Descripción
ITBA	Instituto **T**ecnológico de **B**uenos **A**ires
FEV	**F**orced **E**xpirated **V**olume
Cap.	**C**apítulo
EF	**E**fecto/s **F**ijo/s
EA	**E**fecto/s **A**leatorio/s
EM	**E**fecto/s **M**ixto/s
MEF	**M**odelo de **E**fectos **F**ijos
MEA	**M**odelo de **E**fectos **A**leatorios
MEM	**M**odelo de **E**fectos **M**ixtos
VRE	**V**alor/es **R**esidual/es **E**xtremo/s
EAE	**E**fecto/s **A**leatorio/s **E**xtremo/s
OAE	**O**rdenada/s **A**leatoria/s **E**xtrema/s
PAE	**P**endiente/s **A**leatoria/s **E**xtrema/s

Capítulo 1
Introducción

1.1. Data Mining

En las últimas décadas, los desarrollos en materia digital han cambiado de forma sustancial la manera en la que se trabaja en diversas áreas de investigación [1], [2].

Para empezar, el aumento en las capacidades de almacenamiento en computadoras y servidores permite a los investigadores registrar una gran cantidad de variables que pueden tener una potencial influencia sobre cada fenómeno observado [3].

Por otro lado, en la actualidad, el proceso de carga de datos puede realizarse con mayor facilidad. Por ejemplo, profesionales de la salud pueden cargar de forma digital los datos en el momento de una consulta desde cualquier dispositivo [4].

Además, el incremento de la capacidad computacional de cálculo también posibilita el procesamiento y análisis de datos masivos. Por lo tanto, aumenta también la cantidad de preguntas que surgen en la investigación que se pueden responder y también la cantidad de relaciones que se pueden establecer entre variables a un costo computacional reducido [5], [6].

Sin embargo, estos desarrollos traen aparejados otros desafíos. De hecho, la alta dimensionalidad ofrece una gran cantidad de opciones y dificulta la focalización en relaciones entre variables específicas. Por otro lado, una mayor cantidad de variables también genera confusión a la hora de descartar posibles influencias sobre un proceso. Es decir, las particularidades observadas en una base de datos pueden deberse a varios factores y es difícil determinar cuál de ellos tiene una influencia real [7].

Por lo tanto, herramientas que permitan realizar nuevos descubrimientos a partir de grandes bases de datos adquieren una mayor relevancia para lidiar con estas nuevas dificultades. Esta área de la investigación, es la comúnmente llamada Data Mining, aunque también es conocida como KDD: Knowledge Discovery in Databases, definición que enuncia el principal objetivo de las aplicaciones correspondientes al área.

Una estrategia posible planteada en KDD es agrupar observaciones similares (según algún criterio pertinente) para luego detectar cambios en otras variables no involucradas en el criterio de agrupación, teniendo como objetivo encontrar en dichas similitudes posibles explicaciones a los distintos comportamientos grupales.

En algunas ocasiones, tener un conocimiento previo sobre el comportamiento de las variables permite establecer una estructura usual para los datos, aunque no perfecta, de modo que se pueden

identificar las observaciones que no cumplen el patrón común, con el objetivo de encontrar nuevas explicaciones para el mencionado comportamiento atípico.

Por otro lado, los datos longitudinales consisten de varias mediciones sobre la misma unidad observacional a lo largo de un intervalo de tiempo [8]. Por lo tanto, es importante que un desarrollo en bases longitudinales haga uso del hecho que los datos tienen un orden temporal y logre interpretar las variaciones entre mediciones consecutivas.

1.2. Objetivos

La motivación principal de esta tesis fue generar algoritmos que permitan detectar nuevos patrones en bases de datos longitudinales.

Para lograrlo, se identifican características frecuentes en bases correspondientes a aplicaciones biomédicas, además de contemplar las limitaciones de los datos longitudinales. Los algoritmos desplegados buscan adaptarse a dichas particularidades.

A lo largo del trabajo se describen dos desarrollos, que resultaron en la publicación de sendos artículos científicos [9], [10].

1.3. Agrupamiento por expresión genética

1.3.1. Descripción

El primer desarrollo [9] consiste de una aplicación algorítmica a una base longitudinal de pacientes diabéticos en la que se analiza secuencialmente la expresión de algunos genes estratégicos. La expresión genética se puede medir evaluando la concentración de cadenas de ARN mensajero presentes en el citoplasma o de sus proteínas asociadas [11]. Considerando esta base, se logró agrupar en 3 grupos a los pacientes por la variación en su expresión genética, presentando uno de los grupos características clínicas distintas a los grupos restantes.

Este descubrimiento es relevante ya que las características de los grupos no estuvieron involucradas en el agrupamiento de pacientes y por lo tanto, presenta una potencial asociación de la cual no se tenía información.

Además, nuestra propuesta incluye un nuevo enfoque sobre el problema que permite evaluar muchos algoritmos de clustering y comparar sus resultados. Considerando estas nuevas alternativas, hemos logrado encontrar una forma de agrupar observaciones que, en las bases de datos seleccionadas, arroja mejores resultados que otros métodos preestablecidos.

1.3.2. Antecedentes

Esta sección tiene como objetivo la aplicación de algoritmos de clustering a datos longitudinales de expresión genética. La capacidad de adquisición de estos datos creció enormemente en las últimas décadas [12]. Más aún, el crecimiento es mucho más vertiginoso que el conocimiento que se tiene sobre los fenómenos observados. Por lo tanto, cualquier herramienta que permita sacar nuevas

conclusiones a partir de los datos de esta área puede ayudar a cerrar la brecha mencionada [13], [14].

A la hora de analizar datos de expresión genética, muchos trabajos buscan agrupar morfológicamente las evoluciones de la expresión de una gran cantidad de genes para identificar grupos de genes que se expresan en conjunto o, en contraposición, genes que se reprimen entre sí [15], [16], [17], [18], [19].

La mayoría de los trabajos asumen que los datos son tomados en el mismo instante de tiempo, es decir, bases longitudinales balanceadas en el tiempo. Sin embargo, no siempre se puede asumir que los instantes de medición son simultáneos para todas las observaciones. Para adaptarse a esta complejidad, el trabajo de Möller et al. [20] propone clusterizar los conjuntos de expresión genética basándose en las pendientes entre observaciones, utilizando la distancia euclídea y una versión similar a K-Medias que devuelve particiones "blandas" (ver sección 5.1). Estos trabajos suelen enfocarse en el nivel celular o molecular, sin aplicaciones clínicas.

Las aplicaciones clínicas de clustering morfológico de trayectorias de expresión genética suelen tener objetivos muy específicos y basarse en el aprendizaje supervisado [21], [22], [23], [24], [25]. Es decir, se busca encontrar genes predictores de una cierta condición, pero conociendo esos datos y teniéndolos como objetivo.

Sin embargo, en muchos casos no está claro cuáles de las variables pueden llegar a estar relacionadas ya que se tiene un reducido conocimiento previo del fenómeno observado. Por lo tanto, en esta tesis se extiende el trabajo de [20], haciendo uso del espacio de las pendientes para poder aplicar distintos métodos de clustering y basándose en distancias que no sean necesariamente la euclídea.

1.3.3. Modelos mixtos de clases latentes

Otros trabajos utilizan un procedimiento que se basa en modelos de efectos mixtos (ver Capítulo 3) para identificar K clases [26], [27]. Es otro enfoque para agrupar observaciones que utilizamos para comparar con nuestra propuesta.

Este procedimiento asume que hay una cantidad fija de grupos, pero que se desconoce a qué grupo pertenece cada observación. Por eso se denominan clases latentes. Utilizando las estimaciones del modelo mixto, se determinan similitudes entre individuos para asignarle un grado de pertenencia (medida en una probabilidad) a cada grupo.

Los trabajos que utilizan este procedimiento para datos longitudinales tienen buenos resultados cuando las tendencias son relativamente estables (se asumen polinomiales), pero cuando las trayectorias de respuesta sufren cambios abruptos, los coeficientes estimados pueden ser muy sensibles a estos cambios y por lo tanto, distorsionan la capacidad de clasificación del algoritmo. Por otro lado, para aplicar este método debe conocerse de antemano el modelo que representa las trayectorias de respuesta, que como se mencionó previamente, no es el caso para la mayoría de las aplicaciones.

1.4. Detección de trayectorias atípicas

1.4.1. Descripción

El segundo desarrollo [10], es el de un algoritmo de detección de trayectorias atípicas.

En bases longitudinales donde hay una estructura en los datos explicada por un modelo estadístico, se pueden detectar como atípicas observaciones o conjuntos de observaciones que se apartan de ese modelo.

No hay mucho desarrollo sobre algoritmos que permitan esta detección simultánea de observaciones atípicas de distinta clase (estas clases serán detalladas en el Capítulo 6). Sin embargo, encontramos un trabajo que nos permite comparar con nuestra propuesta y en nuestros experimentos hemos obtenido mejores resultados.

1.4.2. Antecedentes

Una característica común de los conjuntos masivos de datos y de alta dimensionalidad es que no suele haber mucho conocimiento previo sobre los vínculos entre la mayoría de las variables observadas. Por ende, muchos de los antecedentes literarios de detección de outliers asumen una cantidad mínima de hipótesis.

Algunos métodos detectan outliers según la densidad de observaciones cercanas . Es decir, determinando qué proporción de los datos se encuentra a una distancia prefijada. Si para una observación dicha proporción es baja, se puede sospechar que ese dato se encuentra alejado del conjunto mayoritario de datos. En contraposición, a partir de una observación, se puede analizar la proporción de datos que exceden cierta distancia y si dicha proporción es alta, se puede identificar a la observación como atípica [28]. Otra alternativa es el cálculo de distancia al k-ésimo dato más cercano. Mientras mayor sea esa distancia, más alejado estará ese dato de la mayoría de las observaciones [29].

Sin embargo, en esta área la mayoría de las aplicaciones emplean el "Local Outlier Factor (LOF)" [30], que también se calcula en términos de la densidad de datos cercanos. Además, las detecciones basadas en algoritmos de clustering funcionan de manera similar, ya que los clusters minoritarios y alejados del resto de los grupos se consideran outliers, y de algún modo, se atribuye a la densidad de observaciones cercanas [31].

Estas propuestas tienen variantes tanto univariadas como multivariadas, y éstas últimas se basan generalmente en la distancia de Mahalanobis [32], [33]. Dicha distancia requiere de una estimación de un vector medio y de una matriz de covarianza en común, cuyos resultados son muy sensibles a las hipótesis distribucionales y a la presencia de datos extremos. Más aún, los datos no siempre obedecen a una media y covarianza común. Por lo tanto, los resultados de dichos algoritmos pueden ser deficitarios.

En otro enfoque, se puede utilizar el método de componentes principales, que provee combinaciones lineales ortogonales de variables que explican la mayoría de la varianza en los datos. Esto permite reducir la dimensionalidad de los datos en algunas pocas combinaciones de variables, y los datos que no se alinean con dichas combinaciones pueden ser consideradas outliers. Sin em-

bargo, las combinaciones de variables no siempre son de fácil interpretación y puede resultar difícil sacar conclusiones sobre variables específicas que puedan resultar de interés para la aplicación correspondiente.

Los métodos descriptos utilizan de algún modo la distancia entre vectores cuantitativos, identificando datos que se alejan de la mayoría de la población. Sin embargo, agregar un contexto a dichas observaciones puede identificar esos datos como verosímiles a pesar de la distancia. Por lo tanto, a mayor conocimiento previo sobre los datos se puede establecer un contexto que cambie las perspectivas sobre las detecciones positivas [34], [35].

Por otro lado, los métodos descriptos requieren que todas las observaciones tengan la misma dimensión. Por lo tanto, no se adaptan a los datos que puedan faltar en un individuo particular y excluyen todas las mediciones correspondientes al individuo.

Además, en la literatura se aborda la detección de outliers en series de tiempo, donde se pueden consultar las referencias dadas en [36], [37], [38], [39], [40] y [41]. Estos trabajos se basan en modelos estadísticos (en general regresivo o autoregresivo) para identificar valores normales y detectar cambios de tendencia o desvíos puntuales de los valores normales. Sin embargo, como se explica en la sección 2.3, las aplicaciones a series de tiempo se basan en herramientas matemáticas que no son válidas en el contexto de datos longitudinales, y no son comparables con nuestro trabajo.

La mayoría de los trabajos que abordan la detección de outliers utilizando modelos estadísticos suelen estar focalizados en los modelos de efectos fijos [42]. En estos casos, debe considerarse justamente que los outliers pueden afectar la estimación de los parámetros y que por lo tanto pueden darse dos efectos:

- Que un dato atípico afecte de tal manera los parámetros estimados que el outlier deja de ser detectado como tal. Esto se denomina “masking effect”.

- En contraposición, puede darse que por la influencia de un dato atípico, un dato que debería ser considerado como “normal” pase a ser detectado como outlier por el efecto en las estimaciones. Esto se denomina “swamping effect”.

Por lo tanto, para estos casos debería utilizarse un método robusto de estimación que no afecte de manera significativa los valores de los parámetros [43]. Sin embargo, estos algoritmos son de mayor complejidad y generalmente se requieren condiciones específicas para cada aplicación.

En cuanto a los modelos de efectos mixtos, la mayoría de los trabajos se focalizan en la detección de individuos u observaciones influyentes, en el sentido de su efecto sobre los parámetros estimados. Por lo tanto, no hay muchos antecedentes de detección de outliers en el marco de efectos mixtos. El trabajo de mayor similitud con nuestra propuesta es un trabajo de Zewotir et al. [44], en el que también se buscan distintos tipos de anomalías, aunque establecen otros umbrales para sus detecciones.

1.5. Coincidencias entre ambos trabajos

Estos desarrollos utilizan herramientas matemáticas muy distintas, por lo tanto, parece difícil ubicar ambas publicaciones en un mismo marco. Sin embargo, ambos trabajos tienen su fundamento en la interpretación de los coeficientes estadísticos y la evolución temporal de los datos, explotando las características de los datos longitudinales.

Nos parece importante recalcar la necesidad de interpretar los datos, las variables y sus correspondientes efectos en otras variables, dado que muchos desarrollos actuales de análisis de datos omiten dichas apreciaciones y pueden ser usufructuados para enriquecer las herramientas de análisis.

Además, ambas estrategias se sostienen en extensivas simulaciones donde los algoritmos son testeados en condiciones tanto favorables como desfavorables para el rendimiento de los mismos.

1.6. Aportes

Ambos desarrollos ofrecen nuevas herramientas para encontrar potenciales asociaciones en bases de datos longitudinales, motivando nuevas preguntas de investigación que no eran previamente contempladas.

Más aún, estos algoritmos contemplan las limitaciones en el contexto longitudinal de las herramientas matemáticas y estadísticas usuales. Por lo tanto, se aportan procedimientos específicos que se adaptan a una estructura de datos muy frecuente en el ambiente biomédico, en el que puede haber un tamaño muestral pequeño, pocos instantes de medición, datos faltantes y heterogeneidad en los valores.

- En el primer trabajo [9], nuestra interpretación del uso de las pendientes en los algoritmos de clustering permite mejorar las alternativas para agrupar por morfología los conjuntos de respuestas.

  Además, en la aplicación biomédica del algoritmo, se obtuvo un potencial vínculo entre patrones en la expresión genética y características clínicas observables aportando una nueva perspectiva, que motiva nuevas investigaciones sobre este vínculo.

- El segundo trabajo [10] ofrece una nueva herramienta de detección simultánea de observaciones y conjuntos de observaciones atípicas en el contexto longitudinal.

  Al basarse en medidas de dispersión y no en distribuciones específicas que pueden no cumplirse, los umbrales utilizados se adaptan bien a diversas situaciones y devienen en una cantidad razonable de detecciones, con un buen balance entre verdaderos y falsos positivos.

  Las metodologías comparables con nuestro trabajo dependen muy fuertemente de las hipótesis distribucionales y además, tienen umbrales que no se adaptan a la variablilidad de los datos y por lo tanto, puede dar demasiadas detecciones positivas (la mayoría no son relevantes) o directamente no detectan anomalías.

1.7. Organización de la tesis

Esta tesis se organiza del siguiente modo: En el capítulo 2 se hace referencia a las diversas características de los datos longitudinales, las correspondientes definiciones y notaciones. En el capítulo 3 se detallan las particularidades de los modelos de efectos mixtos y su capacidad para adaptarse a distintos tipos de variabilidad. En el capítulo 4 se describe la problemática ante datos faltantes, las definiciones y simulaciones correspondientes. En el capítulo 5 se describen los métodos tradicionales de clustering, la construcción del espacio de las pendientes, las simulaciones y resultados del primer aporte mencionado. Además, se realiza una breve recopilación de los métodos establecidos para lidiar con estas pérdidas de información. En el capítulo 6 se discute el concepto de outliers y cómo se aplica en los modelos de efectos mixtos para el segundo trabajo de esta tesis, con sus correspondientes simulaciones y resultados. Por último, el capítulo 7 contiene las conclusiones finales sobre el trabajo.

Capítulo 2
Datos Longitudinales

2.1. Definiciones

Las bases de datos longitudinales se corresponden con estudios desarrollados en un intervalo de tiempo en el que se realizan varias mediciones repetidas sobre las mismas unidades observacionales (también llamadas individuos). Son muy comunes en áreas de aplicación como la medicina en la que algunas características de los pacientes son evaluadas de manera repetida en el tiempo. Generalmente se considera un instante inicial del estudio llamado tiempo basal y establecido numéricamente como 0, aunque los estudios son muy diversos y esto no siempre ocurre.

Además, las mediciones de cada individuo se registran en distintos instantes de tiempo llamados ocasiones de medición, que puede establecerse de forma cualitativa ordinal o de forma cuantitativa. Es deseable que para todos los individuos coincidan los instantes de medición y la cantidad de observaciones. En este caso, se dice que los datos son balanceados en el tiempo.

Sin embargo, esto generalmente no ocurre en los estudios longitudinales. Por un lado, es muy difícil hacer coincidir los tiempos de medición para todos los individuos, o que las diferencias entre los instantes sea despreciable. Por otro lado, en general es difícil lograr el cumplimiento de todos los participantes a lo largo del estudio y por lo tanto, algunos individuos tendrán menor cantidad de observaciones que otros. En este caso se dice que la base de datos es desbalanceada en el tiempo.

Vale aclarar que las bases longitudinales que son balanceadas en el tiempo suelen ser ensayos diagramados con protocolos estrictos para la recopilación de datos. Y aún en esos casos el cumplimiento de los participantes no siempre se adecúa a dichos requerimientos. Por lo tanto, en general, las bases longitudinales son desbalanceadas en el tiempo.

Además, estos estudios suelen tener como objetivo el análisis del comportamiento de una variable de interés, generalmente llamada **variable de respuesta**. Sumado a esto, para un mismo individuo se tienen varias respuestas secuenciales, al conjunto de estas respuestas individuales se le llama una **trayectoria de respuesta**.

2.1.1. Notación

A continuación introducimos algunas notaciones. Los vectores y matrices son notados en formato negrita. Se indica cada individuo con la letra i ($1 \leq i \leq I$), para el cual se identifica el número de mediciones J_i y cada ocasión de medición con la letra j (para cada individuo i, $1 \leq j \leq J_i$). En el caso de bases balanceadas en el tiempo, $J_i = J$ para todo $1 \leq i \leq I$.

Los instantes de medición se representan con la letra t. Por ejemplo, $t_{i,j}$ representa la j-ésima medida repetida del i-ésimo individuo. En el caso de bases balanceadas en el tiempo, $t_{i,j} = t_j$ para todo $1 \leq i \leq I$.

Por otro lado, la variable de respuesta se nota con la letra Y. Análogamente a los instantes de medición, $Y_{i,j}$ representa la j-ésima respuesta del i-ésimo individuo. Además, se puede definir la trayectoria de respuesta $\mathbf{Y}_i$ como un vector, donde se consideran todas las respuestas del individuo i:

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,J_i} \end{pmatrix} \tag{2.1}$$

Cada componente de este vector se considera una variable aleatoria y por lo tanto, tiene su correspondiente esperanza y varianza, denotadas $E[Y_{i,j}]$ y $V[Y_{i,j}] = E[(Y_{i,j} - E[Y_{i,j}])^2]$ respectivamente. Además, para pares de medidas repetidas sobre el mismo individuo se pueden considerar las covarianzas

$$Cov[Y_{i,j_1}; Y_{i,j_2}] = E[(Y_{i,j_1} - E[Y_{i,j_1}]) \cdot (Y_{i,j_2} - E[Y_{i,j_2}])] \quad (1 \leq j_1, j_2 \leq J_i) \tag{2.2}$$

Notar que cuando $j_1 = j_2 = j$, se obtiene la fórmula para la varianza de $Y_{i,j}$. Por otro lado, todas las covarianzas de la trayectoria de respuesta Y_i se pueden agrupar en una matriz de $J_i \times J_i$, generalmente denotada Σ_i:

$$\mathbf{\Sigma}_i = Cov[\mathbf{Y}_i] = \begin{pmatrix} Cov[Y_{i,1}; Y_{i,1}] & Cov[Y_{i,1}; Y_{i,2}] & \cdots & Cov[Y_{i,1}; Y_{i,J_i}] \\ Cov[Y_{i,2}; Y_{i,1}] & Cov[Y_{i,2}; Y_{i,2}] & \cdots & Cov[Y_{i,2}; Y_{i,J_i}] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Y_{i,J_i}; Y_{i,1}] & Cov[Y_{i,J_i}; Y_{i,2}] & \cdots & Cov[Y_{i,J_i}; Y_{i,J_i}] \end{pmatrix} \tag{2.3}$$

Esta matriz es simétrica ($Cov[Y_{i,j_1}; Y_{i,j_2}] = Cov[Y_{i,j_2}; Y_{i,j_1}]$) y en su diagonal tiene las varianzas de las distintas medidas repetidas ($Cov[Y_{i,j}; Y_{i,j}] = V(Y_{i,j})$). Además, se puede demostrar que esta matriz es definida positiva.

Además de la variable de respuesta Y, puede haber otras P variables denotadas con la letra X, también llamadas covariables. Por ejemplo, $X_{i,j,p}$ representa la j ésima medición de la p-ésima covariable para el individuo i, con $1 \leq p \leq P$. También las covariables se pueden representar de forma matricial:

$$\mathbf{X}_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & \cdots & X_{i,1,P} \\ X_{i,2,1} & X_{i,2,2} & \cdots & X_{i,2,P} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i,J_i,1} & X_{i,J_i,2} & \cdots & X_{i,J_i,P} \end{pmatrix} \tag{2.4}$$

La matriz de covariables $\mathbf{X}_i$ también se suele llamar "Matriz de diseño". Vale aclarar además que como se suele analizar la evolución temporal de las variables de respuesta, una de las covariables $X_{i,j,p}$ generalmente está vinculada con el tiempo $t_{i,j}$.

2.2. Bases utilizadas

A continuación describimos las bases que serán utilizadas a lo largo de este trabajo.

2.2.1. Base de datos FEV_1:

Este estudio fue diseñado para caracterizar la función de crecimiento pulmonar en niños y adolescentes de 6 ciudades de Estados Unidos. Muchos de los participantes fueron reclutados en edades entre 6 y 7 años, y tuvieron seguimientos anuales hasta su graduación de la escuela secundaria (aproximadamente 18 años). La base disponible se focaliza en una muestra aleatoria de 300 participantes del sexo femenino que residían en la ciudad de Topeka, Kansas. A los participantes se les midió el volumen de aire exhalado durante el primer segundo de una espirometría, donde se mide el volumen espiratorio forzado luego de 1 segundo (también llamado FEV_1) en sucesivas mediciones con el objetivo de analizar el aumento de su capacidad pulmonar a lo largo del tiempo. En este caso, se puede considerar a la variable FEV_1 como la variable de respuesta.

Las medidas se dan a edades muy diversas y por lo tanto se registran las edades de forma cuantitativa (medida en años). Por lo tanto, esta base es desbalanceada en el tiempo. Por otro lado, también se registra en cada medición la altura de cada indidividuo (cada uno identificado según la variable ID).

ID	Edad	Altura	FEV_1
2	6.58	1.13	1.36
2	7.65	1.19	1.42
2	12.74	1.49	2.13
2	13.77	1.53	2.38
2	14.69	1.55	2.85
2	15.82	1.56	3.17
2	16.67	1.57	2.52
2	17.63	1.57	3.11

TABLA 2.1: Estructura de la base FEV_1, focalizado en un individuo del estudio.

La Tabla 2.1 permite observar la estructura de los datos. Esta tabla permite visualizar varias cuestiones discutidas anteriormente. Por empezar, la dificultad para mantener una constancia en la recolección anual de los datos, ya que no hay registros entre los 8 y los 12 años.

Por otro lado, se puede ver cómo crece la capacidad pulmonar de esta participante del estudio a lo largo de los años de forma casi constante, aunque se observa una respuesta atípica a los 16.67 años, donde esta tendencia creciente se interrumpe. Por lo tanto, puede ser interesante analizar los motivos de este cambio de tendencia y así adquirir mayor conocimiento sobre el crecimiento en la capacidad pulmonar. Se profundiza en este asunto en el Capítulo 6.

Las trayectorias de respuesta se detallan en la Figura 2.1. El gráfico de trayectorias de respuesta suele agruparse por individuo, uniendo las medidas sucesivas con líneas para poder observar la variación en la respuesta entre un tiempo y otro.

A pesar de la heterogeneidad en las mediciones, se puede ver que el comportamiento a lo largo de los años de los valores de la espirometría suele ser creciente, con una forma determinada.

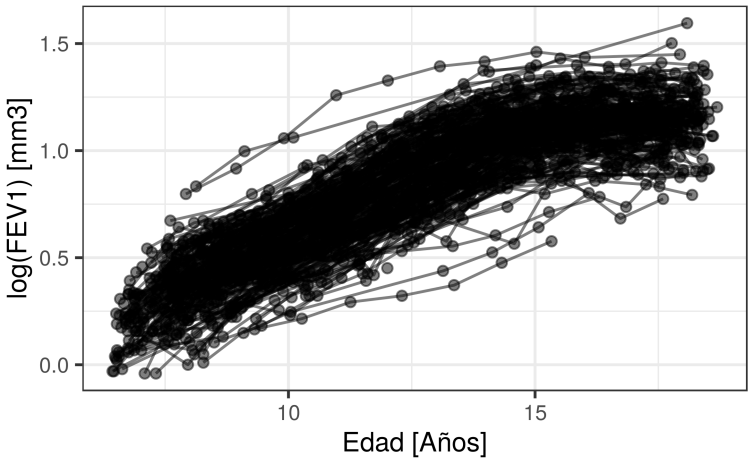


FIGURA 2.1: En esta figura se visualizan las distintas trayectorias de respuesta del logaritmo del valor de la espirometría a lo largo de los años

Puede ser interesante a partir de esta base identificar aquellos individuos cuyos curvas de crecimiento son más o menos pronunciadas respecto del común de la población. Además, se podría intentar focalizar en los individuos que tienen crecimientos similares a los de la población, pero con un punto de partida diferente, alejado de los valores grupales. Estas tareas de detección también son desarrolladas en el Capítulo 6.

2.2.2. Base de datos TLC:

Este estudio se realizó sobre 100 niños con similares niveles de plomo en sangre. El objetivo fue estudiar el efecto de un nuevo tratamiento llamado Succimer, que se suministra de forma oral. De ser efectivo el Succimer, sería menos invasivo que el tratamiento de quelación, el utilizado hasta ese momento. Se aleatorizaron los pacientes a tratamiento y placebo en las mismas proporciones. Luego, se les midió el nivel de plomo en sangre a tiempo basal; a la primer, cuarta y sexta semana, a partir de la administración de cada sustancia.

La Tabla 2.2 muestra las primeras observaciones de la base. Para cada individuo (identificado con la variable ID), se registra el grupo al que fue aleatorizado y las mediciones de plomo en las semanas correspondientes al estudio en 4 columnas distintas. En este caso, la base es balanceada en el tiempo, ya que todos los individuos tienen la misma cantidad de mediciones y en los mismos instantes de tiempo.

Se puede notar en las primeras observaciones que algunos individuos que fueron asignados al grupo de tratamiento (ID 5 y 6) tuvieron una baja drástica en el nivel de plomo en sangre en la primer semana, aunque luego vuelve a crecer de forma leve, aunque sin volver a los niveles basales. Este comportamiento es más notorio cuando se grafican las trayectorias de respuesta y más aún cuando se divide por grupo en la Figura 2.2.

ID	Grupo	Sem 0	Sem 1	Sem 4	Sem 6
1	Placebo	30.8	26.9	25.8	23.8
2	Succimer	26.5	14.8	19.5	21.0
3	Succimer	25.8	23.0	19.1	23.2
4	Placebo	24.7	24.5	22.0	22.5
5	Succimer	20.4	2.8	3.2	9.4
6	Succimer	20.4	5.4	4.5	11.9

TABLA 2.2: Estructura de la base TLC, los valores numéricos corresponden al nivel de plomo en sangre de los participantes, medido en mg/dL. Las últimas cuatro columnas corresponden a las mediciones de las semanas establecidas por el estudio.

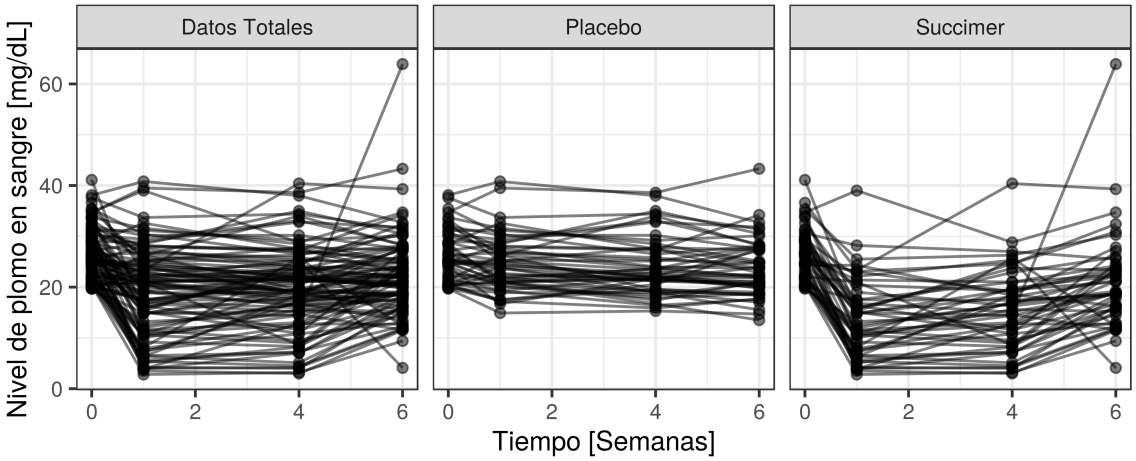


FIGURA 2.2: En esta figura se visualizan las distintas trayectorias de respuesta del nivel de plomo en sangre a lo largo del estudio. Primero se ve el conjunto total de datos y luego se separa por grupo para visualizar las distintas tendencias.

En la Figura 2.2 se ve la importancia de dividir por grupo el conjunto de datos, ya que el primer panel no parece exhibir un comportamiento uniforme, pero al dividir por grupo se aprecia la estabilidad de los valores del grupo placebo y el fuerte decrecimiento inicial en las respuestas del grupo tratamiento. Además, vale aclarar que los niveles basales de ambos grupos fueron similares, por lo que no se ve sesgo en la aleatorización de los grupos.

2.2.3. Base de datos NCGS:

El estudio denominado "National Cooperative Gallstone Study (NCGS)" tuvo como objetivo estudiar la seguridad y efecto de la droga chenodiol como tratamiento de cálculos biliares. La base de datos consiste de 103 individuos aleatorizados a placebo y tratamiento, luego se monitoreó el colesterol en sangre (variable de respuesta, medido en mg/dL) al comienzo del estudio y repetido al sexto, al duodécimo, al vigésimo y al vigésimo cuarto mes.

La Tabla 2.3 permite visualizar los primeros individuos del estudio. La variable Grupo codifica con el número 1 al tratamiento y con el número 2 al placebo, mientras que la variable ID identifica

a cada individuo y las columnas restantes corresponden a las mediciones repetidas del colesterol en sangre. Hay 68 mediciones faltantes por diversos motivos, que no pueden visualizarse a partir de la base de datos ante la falta de variables explicativas.

Grupo	ID	Mes 0	Mes 6	Mes 12	Mes 20	Mes 24
1	1	178	246	295	228	274
1	2	254	260	278	245	340
1	3	185	232	215	220	292
1	4	219	268	241	260	320
1	5	205	232	265	242	230
1	6	182	213	173	200	193

TABLA 2.3: Estructura de la base NGCS, los valores numéricos corresponden al nivel de colesterol en sangre de los participantes, medido en mg/dL. Las últimas cinco columnas corresponden a las mediciones de los meses establecidos por el estudio.

Las trayectorias de respuesta se ven en la Figura 2.3. En este caso, no se observa una diferencia por grupo en términos de crecimiento o decrecimiento del colesterol. Por otro lado, dado que gráficamente se observa una mayor variabilidad en los niveles basales, parece haber un sesgo en la asignación de tratamiento. Sin embargo, esta variabilidad se ve muy acentuada por un individuo con altos niveles basales de colesterol y fortuitamente, fue asignado a unos de los grupos y no parece haber un sesgo sistemático.

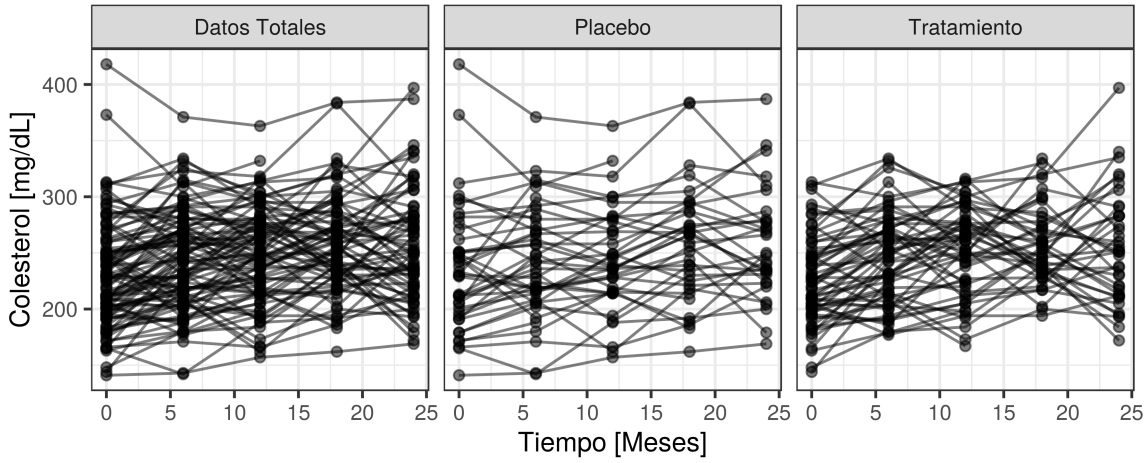


FIGURA 2.3: En esta figura se visualizan las distintas trayectorias de respuesta del nivel de colesterol en sangre a lo largo del estudio. Primero se ve el conjunto total de datos y luego se separa por grupo.

2.2.4. Base de datos DM:

Esta base de datos corresponde a un estudio del Instituto de Inmunología, Genética y Metabolismo (INIGEM), dependiente de la Universidad de Buenos Aires (UBA) y el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

En el estudio, se analizaron en tres ocasiones a pacientes adultos recientemente diagnosticados con Diabetes de tipo II (T2D). La base tiene una alta dimensionalidad, aunque nuestro trabajo se focaliza en analizar la expresión de un gen denominado Interleukina-1β (denotado IL-1β y relacionado con la inflamación sistémica) y sus relaciones con el síndrome metabólico.

Las mediciones se dieron a los 0, 6 y 12 meses desde el diagnóstico, aunque hay fluctuaciones respecto a estos valores y se puede considerar el tiempo en el estudio como una variable cuantitativa medida en meses. Las primeras observaciones de la base pueden visualizarse en la Tabla 2.4. Muchas de las variables no requieren explicación y están medidas con unidades estándard. La variable HbA_{1c} representa el porcentaje de hemoglobina glicosilada, una variable importante para el análisis diabetológico. Las variables HDL y LDL (medidas en mg/dL) representan los dos tipos de colesterol, valores altos de HDL son deseables para una buena salud mientras el caso contrario se da para el LDL. También se visualiza la cantidad de triglicéridos (TG, medidas en mg/dL) y la variable de interés IL1BΔCt que representa la expresión del gen IL-1β.

ID	Tiempo	Sexo	Edad	Talla	Cintura	Peso	IMC	HbA_{1c}	HDL	LDL	TG	IL1BΔCt
1	0	M	59	1.69	114	97.0	33.96	8.44	52	155	141	1.686
2	0	M	53	1.65	106	82.0	30.12	8.10	33	86	182	-1.356
3	0	F	41	1.51	109	85.9	37.67	9.01	48	105	134	2.366
4	0	M	38	1.75	97	76.0	24.82	13.91	29	NA	234	5.133
5	0	M	40	1.69	111	92.0	32.21	8.30	36	156	242	7.107
6	0	M	36	1.74	112	103.0	34.02	11.22	35	55	82	6.378

TABLA 2.4: Estructura de la base DM, considerando sólo algunas columnas ya que la base original contiene 109 variables. El valor NA representa uno de los datos faltantes de la base.

En la Figura 2.4 se ven las trayectorias de respuesta de la expresión genética para los distintos individuos del estudio.

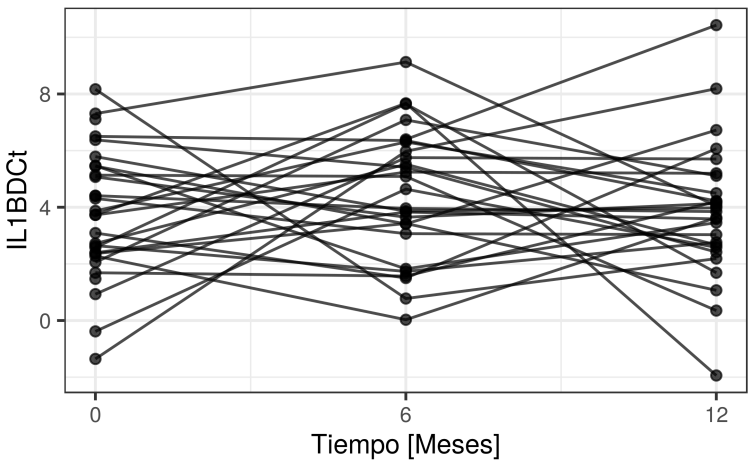


FIGURA 2.4: En esta figura se visualizan las distintas trayectorias de respuesta del nivel de expresión del gen IL-1β a lo largo del estudio.

2.2.5. Formato de la base

Hay dos formatos para las bases de datos longitudinales: formato wide y formato long. Ambas tienen distinta utilidad y el uso de un formato u otro depende del objetivo del estudio.

Formato wide

El formato wide contempla múltiples columnas para las distintas medidas repetidas de aquellas variables que cambian con el tiempo. Como ejemplos están las Tablas 2.2 y 2.3.

El formato wide es útil para realizar comparaciones numéricas entre los distintos tiempos ya que se pueden realizar operaciones entre distintas columnas manteniendo la correspondencia de cada individuo.

Formato long

El formato long considera que a pesar de tener medidas repetidas, las medidas corresponden a una misma variable. Por lo tanto, en vez de considerar varias columnas para distintas mediciones de la misma variable, se pueden fusionar esas columnas agregando una variable identificatoria de los distintos instantes de medición. Como ejemplos de formato long están las Tablas 2.1 y 2.4.

El formato long se usa generalmente para graficar evoluciones temporales de distintas variables, ya que se evalúan los valores de la misma variable a lo largo del tiempo.

Conversión de un formato a otro

Por ejemplo, la tabla 2.2 se puede convertir al formato long según la tabla 2.5:

ID	Grupo	Semana	Plomo [mg/dL]
1	Placebo	0	30.8
1	Placebo	1	26.9
1	Placebo	4	25.8
1	Placebo	6	23.8
2	Succimer	0	26.5
2	Succimer	1	14.8
2	Succimer	4	19.5
2	Succimer	6	21.0

TABLA 2.5: Estructura de la base TLC en formato long.

Notar que en este formato se reduce la cantidad de columnas y aumenta la cantidad de filas. Este formato es el utilizado para graficar las trayectorias de respuesta ya que se pueden comparar los valores de plomo en sangre con el tiempo de permanencia en el estudio.

Por otro lado, para pasar de formato long a formato wide, cada variable que varía con el tiempo resulta en una cantidad de columnas coincidente con el número de mediciones. Por ejemplo, si se quisiera pasar la Tabla 2.4 a formato wide, al haber 3 mediciones, resultaría en 3 columnas por

cada una de las siguientes variables: Cintura, Peso, IMC, HbA_{1c}, HDL, LDL, TG, IL1BΔCt, dando como resultado una gran cantidad de columnas que no parece pertinente exhibir por su extensión.

2.3. Temporalidad de los datos

Como las medidas repetidas se toman sobre el mismo individuo, las mismas suelen estar correlacionadas. Por lo tanto, las herramientas estadísticas que asumen la independencia entre observaciones no resulta adecuada en este formato.

Más aún, las medidas repetidas suelen exhibir una correlación positiva. Por ejemplo, viendo la Figura 2.5 se puede ver cómo los niveles de plomo en sangre para los mismos individuos del estudio TLC presentados en la Sección 2.2.2 (representados con puntos para distintos pares de tiempos según el panel de la figura) presentan una correlación positiva.

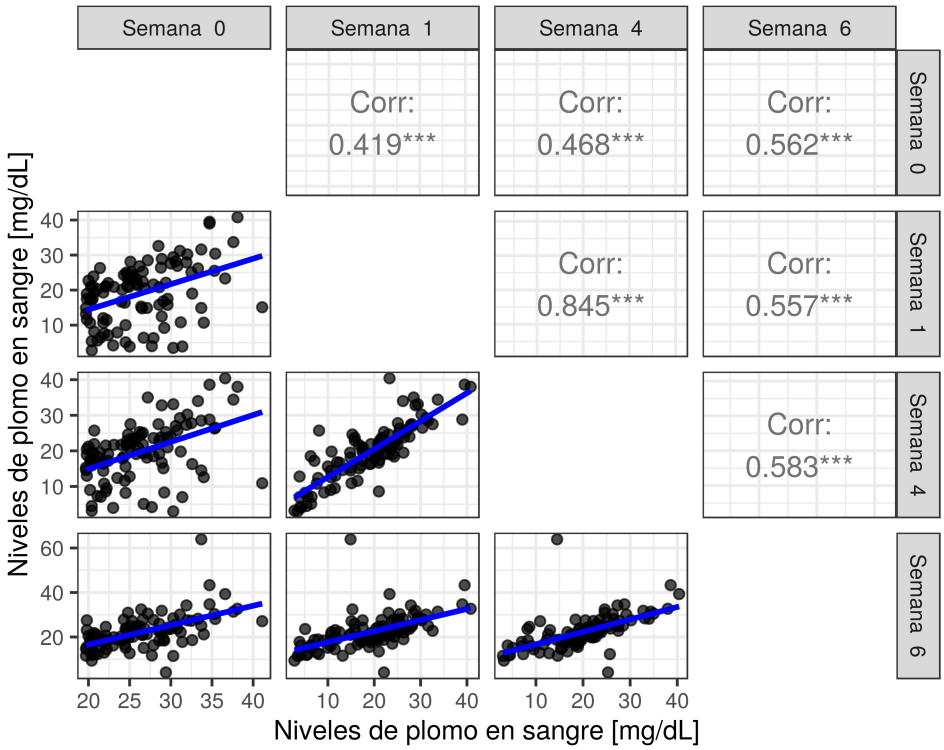


FIGURA 2.5: En esta figura se grafican con puntos los niveles de plomo en sangre en individuos de la base TLC, donde las coordenadas corresponden a los valores observados en el mismo distintas semanas del estudio.

Se puede pensar a partir de esta información que los niveles de plomo no pueden bajar con el tiempo. Sin embargo, esa interpretación es incorrecta dado que la correlación se calcula en base a todos los individuos y por lo tanto, es una característica poblacional, no individual. Para que la correlación sea negativa, debería suceder que gran parte de los de niveles de plomo mayores a la mediana en un tiempo logren un decrecimiento pronunciado y pasen a ser los de menor nivel de plomo en sangre en otro tiempo, invirtiendo la posición relativa a la mediana en casi todos los

individuos, y ese comportamiento no suele observarse en bases longitudinales.

Más aún, cuando las medidas repetidas con un orden temporal se toman sobre varios individuos, entran en juego tanto las características individuales como las características poblacionales. Esta es la mayor diferencia con el área estadística de las series de tiempo. Generalmente, las series de tiempo consisten de una gran cantidad de medidas repetidas sobre una o muy pocas unidades observacionales, mientras que los datos longitudinales suelen presentar una gran cantidad de individuos con pocas medidas repetidas. Por lo tanto, las herramientas matemáticas empleadas en cada caso son completamente distintas. Por ejemplo, el cálculo de transformadas de Fourier en un puñado de observaciones por sujetos no suele devenir en resultados confiables.

2.4. Fuentes de variabilidad

Generalmente las bases de datos biomédicas presentan una alta variabilidad entre sus observaciones. Este fenómeno se debe a la cantidad de factores que entran en juego en un proceso biológico y que pueden tener un efecto en las observaciones. Podemos dividir todos estos efectos en tres categorías: Variabilidad intersujeto, Variabilidad intrasujeto y errores de medición. En esta sección se describen las tres categorías, a modo de motivación, a partir de un ejemplo sencillo con datos simulados. En la Figura 2.6, se visualizan los resultados de una base de 2 individuos, 4 tiempos de medición y los valores de la la variable de respuesta.

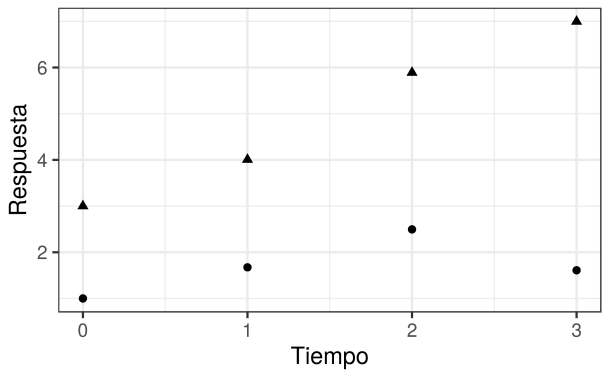


FIGURA 2.6: En esta figura se visualizan las distintas respuestas para dos individuos a lo largo de un estudio ficticio. Las respuestas de ambos individuos son identificadas con símbolos diferentes.

2.4.1. Variabilidad intersujeto

Distintos individuos pueden responder de maneras diversas ante los mismos estímulos, dando lugar a la variabilidad intersujeto. Por ejemplo, en la Figura 2.6 se observa que las tendencias de ambos individuos presentan una marcada diferencia en sus respuestas. Suponiendo que modelamos con rectas las tendencias temporales, habría una recta que representa la evolución de cada individuo y la recta que representa la evolución del conjunto de ambos (Ver Figura 2.7). Es decir, hay una estimación individual y otra estimación poblacional.

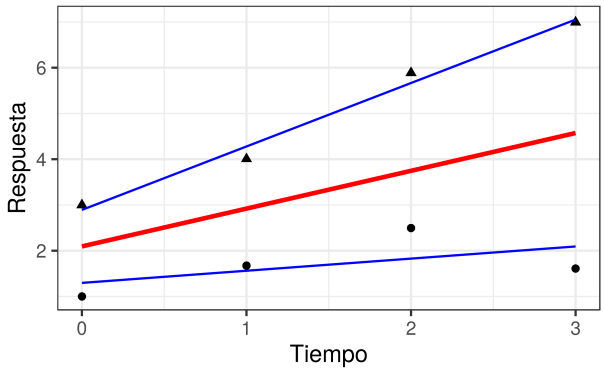


FIGURA 2.7: En esta figura se representan en rojo la tendencia temporal conjunta de ambos individuos, y con rectas azules ambas tendencias individuales.

Podemos ver que las evoluciones individuales son muy distintas entre sí y terminan afectando la estimación poblacional y de este modo, la tendencia poblacional termina sin representar con precisión a ninguno de los dos individuos. Por lo tanto, cuando la variabilidad intersujeto es alta, las estimaciones poblacionales terminan siendo menos representativas de cada individuo.

Para evitar una mayor variabilidad intersujeto, siempre que sea factible, conviene considerar una población con individuos que compartan la mayor cantidad de características posibles. En un estudio biomédico, esta diversidad se controla mediante los principios de inclusión y exclusión.

2.4.2. Variabilidad intrasujeto

A su vez, dependiendo del momento en que sea efectuada cada medición, también puede variar la respuesta de un mismo individuo. Por ejemplo, considerando las evoluciones temporales presentadas en la Figura 2.8, las rectas de tendencia individual se determinan a partir de instantes específicos de la evolución del individuo en el tiempo continuo.

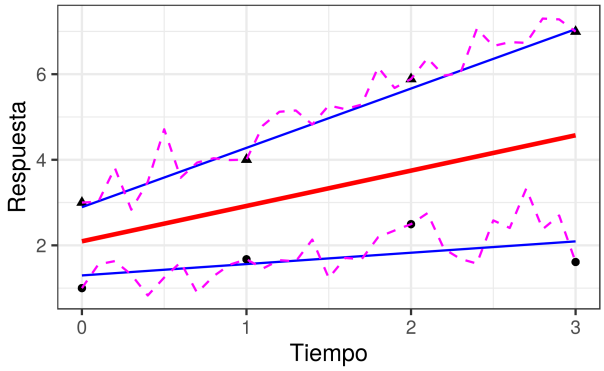


FIGURA 2.8: En esta figura se agrega la respuesta continua de cada individuo con una línea discontinua, en el que los valores medidos coinciden con los instantes de observación.

Mientras más alta sea la variabilidad intrasujeto, menos fidedignas serán las estimaciones de la tendencia individual. Para estudios longitudinales, esta diversidad se suele controlar seleccionando

cuidadosamente los instantes de medición con el objetivo de evitar que haya efectos en la respuesta propios del momento en el que se registra la observación.

2.4.3. Errores de medición

La variabilidad correspondiente a errores de medición obedece a todos los factores que puedan influir en los valores observados de un tiempo en particular. Por ejemplo, puede deberse a un instrumento de medición poco preciso o mal calibrado. Más aún, las variables cuantitativas continuas son redondeadas en una cantidad de decimales que de forma casi inevitable devienen en una diferencia mínima entre el valor real y el observado. Una ejemplificación de estos fenómenos se ve en la Figura 2.9, en algunos casos con errores exagerados para que pueda ser gráficamente visible:

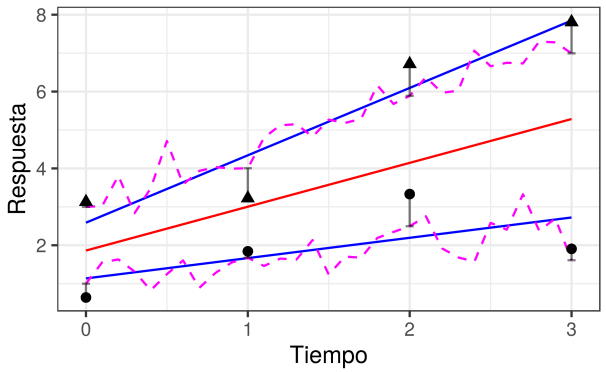


FIGURA 2.9: En esta figura se representan con puntos negros los resultados de las mediciones, que pueden diferir (marcado con segmentos verticales) de la respuesta real (siguiendo la línea discontinua) del individuo en ese instante.

Capítulo 3
Modelos de efectos mixtos (MEM)

3.1. Modelos de efectos fijos (MEF)

3.1.1. Definiciones

Uno de los principales objetivos de los estudios longitudinales es evaluar la evolución temporal de la variable de respuesta y sus vínculos con otras variables llamadas explicativas, predictivas o covariables. En algunas disciplinas, se suele considerar que las variables explicativas tienen un "efecto" sobre la variable de respuesta. Cuando una variable tiene un efecto idéntico sobre todos los individuos de una población se dice que esta variable tiene un efecto fijo (lo llamaremos EF) sobre la respuesta.

Más aún, se puede asumir que estos efectos son lineales. Es decir, para una variable de respuesta $Y_{i,j}$ y covariables $X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}$, un modelo lineal de efectos fijos (denotado MEF) asume la siguiente estructura:

$$Y_{i,j} = \beta_1 \cdot X_{i,j,1} + \beta_2 \cdot X_{i,j,2} + \cdots + \beta_P \cdot X_{i,j,P} + \varepsilon_{i,j} \quad (1 \leq i \leq I, 1 \leq j \leq J_i) \tag{3.1}$$

donde los valores de β_p $(1 \leq p \leq P)$ son los EF (constantes desconocidas, también llamados parámetros de regresión). Por otro lado, la variable $\varepsilon_{i,j}$ denota el error de medición y considera todos los factores no contemplados por las covariables que pueden influir sobre la respuesta $Y_{i,j}$. Generalmente, estos modelos (como la mayoría de los modelos de regresión) asumen que los errores tienen media nula. Por lo tanto,

$$E[Y_{i,j}] = \beta_1 \cdot X_{i,j,1} + \beta_2 \cdot X_{i,j,2} + \cdots + \beta_P \cdot X_{i,j,P}, \quad (1 \leq i \leq I, 1 \leq j \leq J_i) \tag{3.2}$$

dando así un modelo para la respuesta media a lo largo del tiempo.

Vale notar que se puede obtener una ecuación matricial que resuma todos los instantes de medición correspondientes a un mismo individuo i $(1 \leq i \leq I)$:

$$\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,J_i} \end{pmatrix} = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & \cdots & X_{i,1,P} \\ X_{i,2,1} & X_{i,2,2} & \cdots & X_{i,2,P} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i,J_i,1} & X_{i,J_i,2} & \cdots & X_{i,J_i,P} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,J_i} \end{pmatrix} \Rightarrow \mathbf{Y}_i = \mathbf{X}_i \times \vec{\beta} + \varepsilon_i \tag{3.3}$$

Más aún, se puede considerar otra ecuación matricial que concentre los vínculos entre respuestas y covariables para todos los individuos, concatenando las matrices $\mathbf{Y}_i$, ε_i y $\mathbf{X}_i$ por filas. Llamando $N = \sum_{i=1}^{I} J_i$ a la cantidad total de respuestas observadas:

$$\underbrace{\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\\ \vdots\\ \mathbf{Y}_I\end{pmatrix}}_{N\times 1} = \underbrace{\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_I\end{pmatrix}}_{N\times P} \times \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_P\end{pmatrix}}_{P\times 1} + \underbrace{\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_I\end{pmatrix}}_{N\times 1} \Rightarrow \mathbf{Y} = \mathbf{X}\times\vec{\beta} + \varepsilon \tag{3.4}$$

Por otro lado, los vectores ε_i pueden tener su propia matriz de covarianza $\mathbf{R}_i$, que en este modelo es la misma que la de $\mathbf{Y}_i$, ya que la única fuente de aleatoriedad viene dada por ε_i. Es decir,

$$\mathbf{\Sigma}_i = Cov[\mathbf{Y}_i] = Cov[\varepsilon_i] = \mathbf{R}_i \tag{3.5}$$

Por lo tanto, como los errores tienen media nula, asumir este MEF es asumir que cada trayectoria de respuesta $\mathbf{Y}_i$ sigue una distribución normal multivariada:

$$\mathbf{Y}_i \sim \mathcal{N}_{J_i}(\mathbf{X}_i\times\vec{\beta}; \mathbf{\Sigma}_i) \tag{3.6}$$

3.1.2. Estimación de los parámetros

La cantidad de variables explicativas (y sus correspondientes parámetros) debe ser suficiente para que la ecuación represente de forma fidedigna la variable de respuesta. Sin embargo, debe tenerse en cuenta que a una mayor cantidad de parámetros se reduce la precisión de cada valor estimado, ya que la misma cantidad de datos se usa para una mayor cantidad de estimaciones. Además, al sumar variables innecesarias a dicho modelo puede darse el fenómeno denominado "overfitting", en el que se da tanta prioridad a las observaciones disponibles que el modelo pierde poder predictivo para nuevas observaciones. Por lo tanto, debe apuntarse a construir un modelo parsimonioso que tenga la cantidad justa de variables que permita mantener un valor tanto explicativo como predictivo.

La matriz de covarianza de las trayectorias (dada por $\mathbf{\Sigma}_i$) tiene como máximo $\frac{J_i\cdot(J_i+1)}{2}$ parámetros relativos a la covarianza entre medidas repetidas, ya que la matriz es simétrica. De todos modos, es muy común que se tome la matriz $\mathbf{\Sigma}_i$ como diagonal con el mismo valor en todas las coordenadas de la misma. Es decir, que se tome un único parámetro (generalmente denotado σ^2) presente en todos los elementos de la diagonal y resulte $\mathbf{\Sigma}_i = \sigma^2\cdot\mathbf{I}_{Ji}$, donde $\mathbf{I}_{Ji}$ es la matriz identidad de $J_i\times J_i$. Sin embargo, una matriz de covarianza diagonal se corresponde con medidas repetidas independientes entre sí y se contradicen las características discutidas en la sección 2.3.

Más allá de estos casos extremos, se pueden asumir distintas estructuras para la matriz de covarianza $\mathbf{\Sigma}_i$ que reduzca la cantidad de parámetros relativos a la covarianza. Esta estructura resulta en un vector de parámetros de covarianza $\vec{\theta}$ con una longitud H que puede tomar cualquier valor entre 1 y $\frac{J_i\cdot(J_i+1)}{2}$. Las estructuras se asumen comunes a todos los individuos de la población

y por lo tanto, aunque el tamaño de la matriz pueda variar de individuo a individuo, los parámetros de covarianza de $\vec{\theta}$ son poblacionales y no individuales. Es decir, si hay menor cantidad de medidas repetidas en algún individuo, no se agregan parámetros nuevos sino que se asume que la matriz de covarianza respeta la estructura, sólo que potencialmente tendrá una menor cantidad de parámetros que los presentes en el vector $\vec{\theta}$, ya que se excluye como mínimo una fila y una columna de la matriz dada como estructura general para $\mathbf{\Sigma}_i = \mathbf{\Sigma}_i(\vec{\theta})$.

Una vez planteado el modelo, se deben estimar a partir de las observaciones ambos vectores de parámetros $\vec{\beta} \in \mathbb{R}^P$ (vector de EF) y $\vec{\theta} \in \mathbb{R}^H$ (vector de parámetros de covarianza). Para obtener los valores se apela a la estimación por máxima verosimilitud. Este procedimiento se basa en la distribución probabilística asumida (en este caso, normal multivariada) para hallar los valores de los parámetros que maximicen dicha probabilidad, contemplando todos los valores observados $Y_{i,j}$ y $X_{i,j,p}$. Es decir, se busca resolver un problema de optimización que consiste un sistema de $P+H$ ecuaciones (una para cada parámetro del modelo) que en general no son lineales, resolución que suele requerir algoritmos heurísticos e iterativos para poder hallar una solución.

En general los objetivos centrales de la estimación conciernen al vector de EF $\vec{\beta}$. Sin embargo, los parámetros de covarianza $\vec{\theta}$ son necesarios para realizar el cálculo. Una vez obtenidos los valores estimados para la covarianza (denotados $\widehat{\theta}$), se reemplazan los valores para estimar las matrices de covarianza obteniendo $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$, y se puede ver que los valores óptimos para estimar β vienen dados por la siguiente expresión:

$$\widehat{\beta} = \left(\sum_{i=1}^{I} \mathbf{X}_i' \times \widehat{\mathbf{\Sigma}}_i^{-1} \times \mathbf{X}_i \right)^{-1} \times \left(\sum_{i=1}^{I} \mathbf{X}_i' \times \widehat{\mathbf{\Sigma}}_i^{-1} \times \mathbf{Y}_i \right)$$

Este estimador tiene propiedades muy deseables:

- Es insesgado, por lo que la distribución del estimador está centrada en el vector de parámetros que se busca estimar. Es decir, $E[\widehat{\beta}] = \vec{\beta}$.
- Es consistente, por lo que a medida que aumenta el tamaño muestral la distribución de probabilidades del estimador se concentra alrededor del vector de parámetros $\vec{\beta}$. Es decir, $\widehat{\beta} \underset{N \to +\infty}{\longrightarrow} \vec{\beta}$.
- Se conoce su distribución asintótica, a medida que el tamaño muestral aumenta, el estimador $\widehat{\beta}$ tiene distribución multivariada: $\widehat{\beta} \underset{N \to +\infty}{\sim} \mathcal{N}_P \left(\beta, \left(\sum_{i=1}^{I} \mathbf{X}_i' \times \mathbf{\Sigma}_i^{-1} \times \mathbf{X}_i \right)^{-1} \right)$

3.1.3. Modelos para la base TLC

Todos estos modelos fueron mencionados de forma teórica, sin mencionar cómo se eligen los parámetros y las variables. En esta sección veremos cómo se ponen en práctica Con tal fin, aplicaremos distintos modelos a la base llamada TLC, introducida en la sección 2.2.2.

Modelo lineal

Cuando la variable temporal es cuantitativa, se puede modelar la respuesta media como una función de esta variable temporal. El modelo más simple, es aquel donde la respuesta (Pl: cantidad de plomo en sangre), evoluciona linealmente con el tiempo (t: en semanas, son 4 mediciones), es decir:

$$\mathrm{Pl}_{i,j} = \beta_0 + \beta_1 \cdot t_{i,j} + \varepsilon_{i,j}, \quad (1 \leq i \leq 100, 1 \leq j \leq 4) \tag{3.7}$$

Este conjunto de ecuaciones se puede representar para cada uno de los individuos de forma matricial. Además, recordando que los tiempos de medición son a las 0, 1, 4 y 6 semanas:

$$\begin{pmatrix} \mathrm{Pl}_{i,1} \\ \mathrm{Pl}_{i,2} \\ \mathrm{Pl}_{i,3} \\ \mathrm{Pl}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} \\ 1 & t_{i,2} \\ 1 & t_{i,3} \\ 1 & t_{i,4} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \Rightarrow \mathbf{Pl}_i = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 4 \\ 1 & 6 \end{pmatrix}}_{\mathbf{X}_i} \times \vec{\beta} + \varepsilon_i, \quad (1 \leq i \leq 100) \tag{3.8}$$

Notar que definiendo la matriz $\mathbf{X}_i$, se obtiene la expresión dada en 3.3. Esta matriz se denomina "matriz de diseño" y contiene toda la información sobre las variables explicativas. Por lo tanto, su composición depende del modelo asumido para un vector de respuestas y sobre todo, cuáles de las variables se consideran explicativas.

Asumiremos inicialemente que la matriz de covarianza de los errores es $\mathbf{\Sigma}_i = \sigma^2 \cdot \mathbf{I}_{Ji}$. Esto contradice el hecho de que las respuestas repetidas en un mismo individuo deben estar correlacionadas pero en la sección 3.2.1 explicaremos con mayor detalle esta decisión. Con estas hipótesis, se obtienen las siguientes estimaciones para los parámetros del modelo:

- $\widehat{\beta}_0 = 22.976$
- $\widehat{\beta}_1 = -0.401$
- $\widehat{\sigma} = 7.95618$

Con estas estimaciones, se puede superponer la estimación poblacional con las trayectorias (ver la Figura 3.1) para evaluar el ajuste del modelo a los datos. Notar que los valores estimados de la ordenada y la pendiente de la recta son consistentes con los observados en el gráfico:

Vemos que el modelo no es muy representativo de los datos (avalado por el bajo valor del R^2=0.0118, medida estándard para evaluar un modelo de regresión lineal, que debe estar cercano a 1 para dar señal de un buen ajuste). Esto se debe a que los EF son comunes a toda la población, aún cuando los individuos puedan exhibir comportamientos diversos. Por lo tanto, recurriremos a otras alternativas.

Modelo lineal por grupos

Una de las mayores falencias del modelo anterior es la imposibilidad de distinguir los modelos según el grupo de tratamiento o placebo, aún cuando la Figura 2.2 muestra que las trayectorias son distintas para individuos de distintos grupos. Por lo tanto, un modelo estadístico debería contemplar estas diferencias. Esto parece entrar en conflicto con lo mencionado anteriormente, cuando los EF se definieron como comunes a todos los individuos. Sin embargo, si se considera

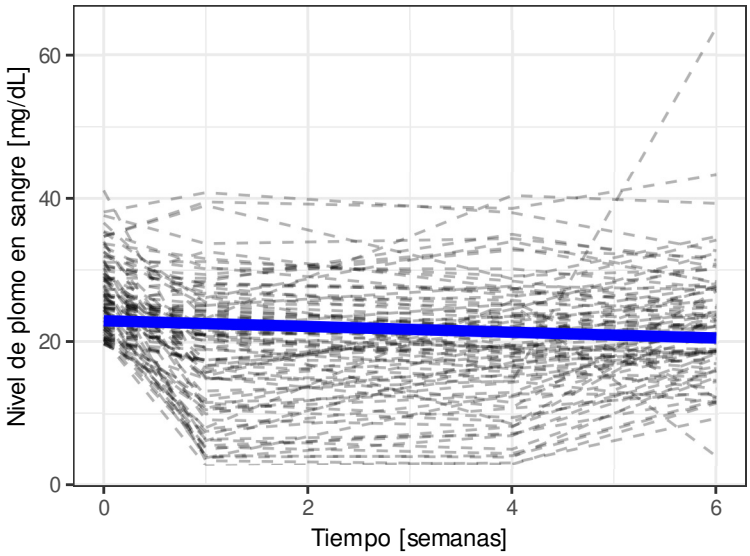


FIGURA 3.1: En esta figura se visualizan las trayectorias de respuesta (con líneas discontinuas) para la base TLC y la estimación (con una línea sólida) por el modelo lineal dado en 3.7.

una variable que sólo tiene influencia en el caso de pertenecer a uno de los grupos, el modelo de EF puede captar estas diferencias.

En este caso, consideraremos dos rectas, una por grupo, planteando el siguiente modelo:

$$\mathrm{Pl}_{i,j} = \begin{cases} \beta_0 + \beta_1 \cdot t_{i,j} + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Placebo} \\ \beta_2 + \beta_3 \cdot t_{i,j} + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \end{cases}, \begin{pmatrix} 1 \leq i \leq 100 \\ 1 \leq j \leq 4 \end{pmatrix} \tag{3.9}$$

Es decir, los coeficientes β_0 y β_1 representan la ordenada y la pendiente de la recta correspondiente al grupo placebo, respectivamente. Análogamente, β_2 y β_3 representan la ordenada y la pendiente de la recta correspondiente al grupo Tratamiento, respectivamente.

Matricialmente, se tiene lo siguiente:

$$\begin{pmatrix} \mathrm{Pl}_{i,1} \\ \mathrm{Pl}_{i,2} \\ \mathrm{Pl}_{i,3} \\ \mathrm{Pl}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} & 0 & 0 \\ 1 & t_{i,2} & 0 & 0 \\ 1 & t_{i,3} & 0 & 0 \\ 1 & t_{i,4} & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Placebo}$$

$$(1 \leq i \leq 100)$$

$$\begin{pmatrix} \mathrm{Pl}_{i,1} \\ \mathrm{Pl}_{i,2} \\ \mathrm{Pl}_{i,3} \\ \mathrm{Pl}_{i,4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & t_{i,1} \\ 0 & 0 & 1 & t_{i,2} \\ 0 & 0 & 1 & t_{i,3} \\ 0 & 0 & 1 & t_{i,4} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \tag{3.10}$$

También se puede describir el modelo de la siguiente manera:

$$\mathbf{Pl}_i = \beta_0 \cdot \mathbf{Plac}_i + \beta_1 \cdot \mathbf{t}_i \times \mathbf{Plac}_i + \beta_2 \cdot \mathbf{Trat}_i + \beta_3 \cdot \mathbf{t}_i \times \mathbf{Trat}_i + \varepsilon_i \tag{3.11}$$

donde $\mathbf{Plac}_i$ y $\mathbf{Trat}_i$ se definen del siguiente modo:

$$\begin{aligned} \mathbf{Plac}_i &= \begin{cases} 1 & \text{si el individuo } i \text{ pertenece al grupo Placebo} \\ 0 & \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \end{cases} \\ \mathbf{Trat}_i &= \begin{cases} 1 & \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \\ 0 & \text{si el individuo } i \text{ pertenece al grupo Placebo} \end{cases} \end{aligned} \tag{3.12}$$

Asumiendo que la matriz de covarianza de los errores es $\mathbf{\Sigma}_i = \sigma^2 \cdot \mathbf{I}_{J_i}$, se obtienen las siguientes estimaciones para los parámetros del modelo:

- $\widehat{\beta}_0 = 25.685$
- $\widehat{\beta}_1 = -0.372$
- $\widehat{\beta}_2 = 20.267$
- $\widehat{\beta}_3 = -0.430$
- $\widehat{\sigma} = 7.478$

Con estas estimaciones, se puede superponer la estimación poblacional con las trayectorias (ver la Figura 3.2) para verificar el ajuste del modelo a los datos. Notar que las rectas tienen diferentes ordenadas y pendientes según el grupo, consistentes con los valores estimados:

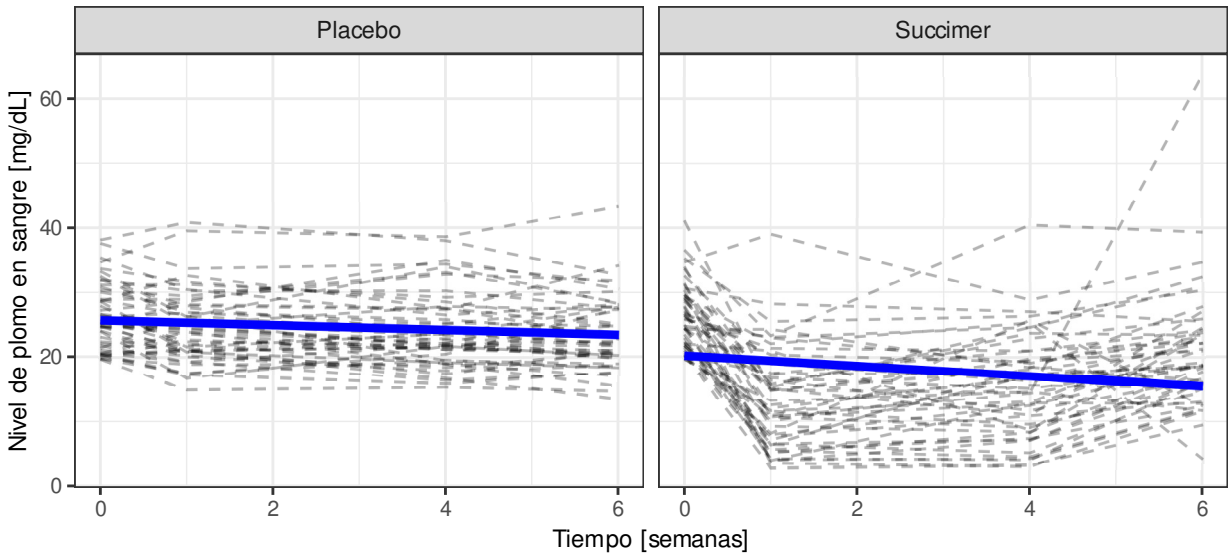


FIGURA 3.2: En esta figura se visualizan las trayectorias de respuesta (con líneas discontinuas) para la base TLC y la estimación (con una línea sólida) por el modelo lineal dado en 3.9.

Además, el R^2 incrementa su valor respecto del modelo anterior (0.129). De todas formas, el modelo sigue sin resultar satisfactorio, dado que no parece haber, al menos en el grupo Succimer, un decrecimiento constante a lo largo del tiempo.

Modelo cuadrático por grupos

Para agregarle complejidad al modelo de evolución temporal de las respuestas, se puede agregar un término cuadrático al modelo anterior. Es decir, para todo $1 \leq i \leq 100$ y $1 \leq j \leq 4$:

$$\mathrm{PI}_{i,j} = \begin{cases} \beta_0 + \beta_1 \cdot t_{i,j} + \beta_2 \cdot t_{i,j}^2 + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Placebo} \\ \beta_3 + \beta_4 \cdot t_{i,j} + \beta_5 \cdot t_{i,j}^2 + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \end{cases} \tag{3.13}$$

Matricialmente, ocurre lo siguiente, para todo $1 \leq i \leq 100$:

$$\begin{pmatrix} \mathrm{PI}_{i,1} \\ \mathrm{PI}_{i,2} \\ \mathrm{PI}_{i,3} \\ \mathrm{PI}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} & t_{i,1}^2 & 0 & 0 & 0 \\ 1 & t_{i,2} & t_{i,2}^2 & 0 & 0 & 0 \\ 1 & t_{i,3} & t_{i,3}^2 & 0 & 0 & 0 \\ 1 & t_{i,4} & t_{i,4}^2 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Placebo}$$

$$\begin{pmatrix} \mathrm{PI}_{i,1} \\ \mathrm{PI}_{i,2} \\ \mathrm{PI}_{i,3} \\ \mathrm{PI}_{i,4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & t_{i,1} & t_{i,1}^2 \\ 0 & 0 & 0 & 1 & t_{i,1} & t_{i,2}^2 \\ 0 & 0 & 0 & 1 & t_{i,1} & t_{i,3}^2 \\ 0 & 0 & 0 & 1 & t_{i,1} & t_{i,4}^2 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \tag{3.14}$$

Escrito de forma más sintética:

$$\mathbf{PI}_i = \beta_0 \cdot \mathbf{Plac}_i + \beta_1 \cdot \mathbf{t}_i \times \mathbf{Plac}_i + \beta_2 \cdot \mathbf{t}_i^2 \times \mathbf{Plac}_i + \beta_3 \cdot \mathbf{Trat}_i + \beta_4 \cdot \mathbf{t}_i \times \mathbf{Trat}_i + \beta_5 \cdot \mathbf{t}_i^2 \times \mathbf{Trat}_i + \varepsilon_i \tag{3.15}$$

donde $\mathbf{t}_i^2$ es el vector cuantitativo de tiempos con sus componentes elevadas al cuadrado, mientras que $\mathbf{Plac}_i$ y $\mathbf{Trat}_i$ se definen del mismo modo que en la ecuación (3.12).

Con las hipótesis asumidas, los coeficientes estimados son:

- $\widehat{\beta}_0 = 25.97$
- $\widehat{\beta}_1 = -0.917$
- $\widehat{\beta}_2 = 0.092$
- $\widehat{\beta}_3 = 23.973$
- $\widehat{\beta}_4 = -8.459$
- $\widehat{\beta}_5 = 1.288$
- $\widehat{\sigma} = 6.905$

Con estos valores, las estimaciones poblacionales se ven en la Figure 3.3.

Nuevamente, mejora el valor de R^2 respecto del modelo anterior (0.257). Puede notarse que dado el pequeño valor de $\widehat{\beta}_2$, en el grupo placebo la estimación es similar a la del modelo lineal. Por otro lado, en el grupo Succimer se observa una parábola más pronunciada que mejora la representación de los datos ya que se amolda al decrecimiento inicial y el posterior incremento leve. Sin embargo, no parece situarse el vértice en el lugar correcto, ya que el cambio abrupto en la trayectoria se da después de la primer semana.

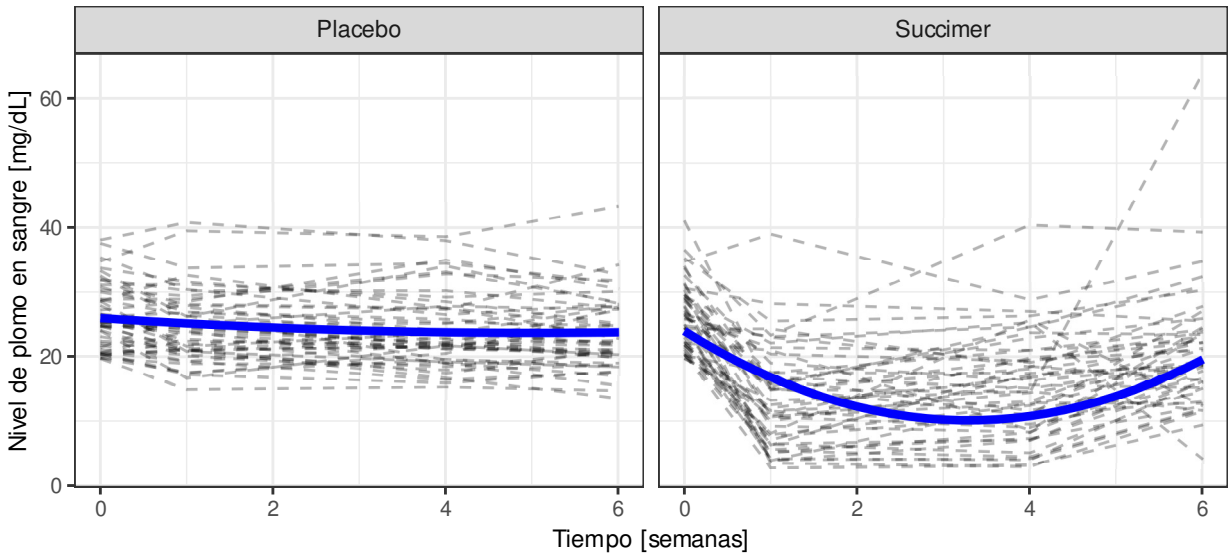


FIGURA 3.3: En esta figura se visualizan las trayectorias de respuesta (con líneas discontinuas) para la base TLC y la estimación (con una línea sólida) por el modelo lineal dado en 3.13.

Modelo lineal semiparamétrico

Como el cambio abrupto parece darse luego de la primera semana, se puede adoptar un modelo lineal a trozos (también llamado semiparamétrico), con un cambio de comportamiento en el modelo a partir de un cierto tiempo (también llamado "nudo"). Para lograr esto, se puede definir la siguiente variable:

$$t^*_{i,j} = \text{máx}\{t_{i,j} - 1, 0\} \tag{3.16}$$

Es decir, la variable tiene valor nulo para los instantes menores o iguales a 1 y empieza a tener influencia en los valores posteriores.

Con esta definición, se puede esbozar el siguiente modelo:

$$\text{Pl}_{i,j} = \begin{cases} \beta_0 + \beta_1 \cdot t_{i,j} + \beta_2 \cdot t^*_{i,j} + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Placebo} \\ \beta_3 + \beta_4 \cdot t_{i,j} + \beta_5 \cdot t^*_{i,j} + \varepsilon_{i,j} & \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \end{cases} \tag{3.17}$$

De modo similar al modelo cuadrático, esta definición se corresponde con la siguiente estructura

matricial, para todo $1 \leq i \leq 100$:

$$\begin{pmatrix} \mathrm{PI}_{i,1} \\ \mathrm{PI}_{i,2} \\ \mathrm{PI}_{i,3} \\ \mathrm{PI}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} & t^*_{i,1} & 0 & 0 & 0 \\ 1 & t_{i,2} & t^*_{i,2} & 0 & 0 & 0 \\ 1 & t_{i,3} & t^*_{i,3} & 0 & 0 & 0 \\ 1 & t_{i,4} & t^*_{i,4} & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Placebo}$$

$$\begin{pmatrix} \mathrm{PI}_{i,1} \\ \mathrm{PI}_{i,2} \\ \mathrm{PI}_{i,3} \\ \mathrm{PI}_{i,4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & t_{i,1} & t^*_{i,1} \\ 0 & 0 & 0 & 1 & t_{i,1} & t^*_{i,2} \\ 0 & 0 & 0 & 1 & t_{i,1} & t^*_{i,3} \\ 0 & 0 & 0 & 1 & t_{i,1} & t^*_{i,4} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \quad \text{si el individuo } i \text{ pertenece al grupo Tratamiento} \tag{3.18}$$

Además, se pueden resumir el modelo del siguiente modo:

$$\mathbf{PI}_i = \beta_0 \cdot \mathbf{Plac}_i + \beta_1 \cdot \mathbf{t}_i \times \mathbf{Plac}_i + \beta_2 \cdot \mathbf{t}^*_i \times \mathbf{Plac}_i + \beta_3 \cdot \mathbf{Trat}_i + \beta_4 \cdot \mathbf{t}_i \times \mathbf{Trat}_i + \beta_5 \cdot \mathbf{t}^*_i \times \mathbf{Trat}_i + \varepsilon_i \tag{3.19}$$

donde $\mathbf{Plac}_i$ y $\mathbf{Trat}_i$ se definen del mismo modo que en la ecuación (3.12).

Los parámetros estimados presentan los siguientes valores: Con las hipótesis asumidas, los coeficientes estimados son:

- $\widehat{\beta}_0 = 26.272$
- $\widehat{\beta}_1 = -1.607$
- $\widehat{\beta}_2 = 1.405$
- $\widehat{\beta}_3 = 26.54$
- $\widehat{\beta}_4 = -15.244$
- $\widehat{\beta}_5 = 16.428$
- $\widehat{\sigma} = 6.644$

Estos coeficientes deben interpretarse de la siguiente manera: la pendiente estimada para el grupo placebo previo a la primer semana es $\widehat{\beta}_1 = -1.607$, mientras que en el mismo grupo luego de la primer semana influyen ambos coeficientes $\widehat{\beta}_1$ y $\widehat{\beta}_2$. Es decir, luego de la primer semana la pendiente estimada es $\widehat{\beta}_1 + \widehat{\beta}_2 = -0.202$. Del mismo modo, las pendientes estimadas para el grupo Tratamiento son $\widehat{\beta}_4 = -15.244$ previo a la primera semana y $\widehat{\beta}_4 + \widehat{\beta}_5 = 1.184$ para los instantes posteriores. Estas características se observan en el gráfico 3.4:

Se observa en la figura que el modelo representa mejor la evolución de los datos y también se observa en un valor reducido de $\widehat{\sigma}$ con repecto a los otros modelos. Esto se sustenta aún más con un aumento del R^2 ajustado, con un valor de 0.313.

Se pueden tener además en cuenta algunas características para hacer más parsimonioso el modelo, ya que cada parámetro que se agrega a un modelo trae aparejado un incremento en la variabilidad de su estimación. Por ejemplo, no parece haber una diferencia grupal en los niveles basales, por lo tanto, podría considerarse un único parámetro β_0 en los modelos anteriores y reemplazar las ordenadas correspondientes al grupo tratamiento por este mismo valor. Además, en todas las variantes implementadas, el grupo placebo siempre se correspondió con trayectorias grupales similares a una recta, por lo que podría tener sentido eliminar los términos denotados β_2

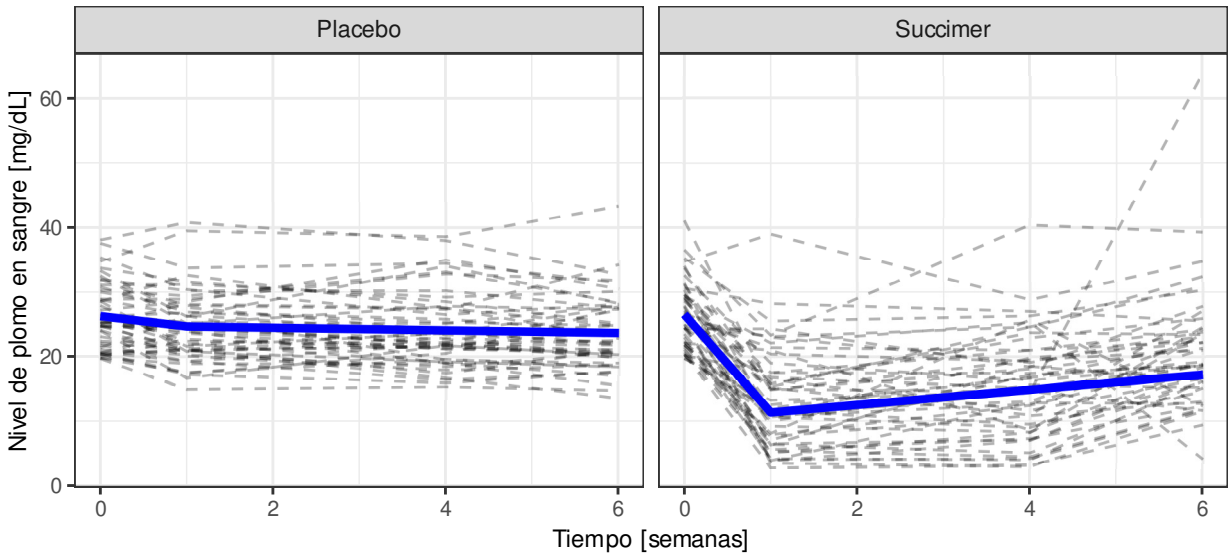


FIGURA 3.4: En esta figura se visualizan las trayectorias de respuesta (con líneas discontinuas) para la base TLC y la estimación (con una línea sólida) por el modelo lineal dado en 3.17.

en (3.13) y (3.17).

Más allá de haber obtenido un modelo que cumple algunas propiedades deseables, se observa mucha heterogeneidad y dispersión respecto de las estimaciones poblacionales. Justamente por esta característica, a pesar de que aumentó el valor de R^2 al cambiar el modelo, tiene valores muy alejados de 1, su valor óptimo.

3.2. Modelos de efectos mixtos (MEM)

Para abordar el problema de la heterogeneidad, se introducen los efectos aleatorios (lo llamaremos EA). Para motivar la comprensión de los conceptos correspondientes, apelaremos nuevamente a la base TLC en la sección 3.2.1, focalizando en el grupo Placebo.

3.2.1. Efectos aleatorios (EA)

Recordando las trayectorias empíricas de los niveles de plomo en sangre para los individuos del grupo placebo en el estudio TLC (ver Figura 3.4), las evoluciones son medianamente estables, por lo tanto, puede tener sentido pensar las trayectorias como rectas. Sin embargo, a diferencia de lo planteado en la ecuación (3.9), se considerará una recta por individuo, es decir:

$$\mathrm{PI}_{i,j} = \beta_{i,0} + \beta_{i,1} \cdot t_{i,j} + \varepsilon_{i,j}, \quad (1 \leq i \leq 50, 1 \leq j \leq 4) \tag{3.20}$$

Es decir, para cada uno de los 50 individuos (ya que se toma sólo el grupo Placebo), se considera una ordenada y una pendiente que no afecta el modelo de las otras trayectorias. Sumado al parámetro de la varianza σ, resulta en un total de 101 parámetros, lo cual parece un tanto

excesivo considerando un total de 200 datos. Una vez estimados los valores $\widehat{\beta}_{0,i}$, $\widehat{\beta}_{1,i}$ y σ, las trayectorias estimadas bajo el modelo 3.20 vienen dadas por

$$\widehat{\mathrm{PI}}_i = \widehat{\beta}_{0,i} + \widehat{\beta}_{1,i} \cdot t_{i,j} \tag{3.21}$$

En este escenario, se produce un efecto de overfitting y el modelo tiene capacidad explicativa pero poca capacidad predictiva, ya que al ser los parámetros individuales, una nueva observación no tiene en los valores estimados una ordenada ni pendiente que le corresponda y se estima una trayectoria de valor nulo.

Para poder adaptarse a la heterogeneidad, los EA se plantean de forma similar a la ecuación 3.20, pero la ordenada y pendiente no se consideran EF, sino que son realizaciones independientes de un vector aleatorio ($\vec{\mathbf{b}}_i = (b_{i,0}, b_{i,1})$) de idéntica distribución normal multivariada ($\vec{\mathbf{b}}_i \sim \mathcal{N}_2(\vec{\mathbf{0}}, \mathbf{G})$, donde $\mathbf{G}$ es la matriz de covarianza de la distribución). Notar que estas variables tienen esperanza nula, al igual que los errores de medición. Por lo tanto, estos valores sólo pueden predecir variables de media nula. De este modo, tomamos $\mathrm{PI}^*_{i,j} = \mathrm{PI}_{i,j} - E(\mathrm{PI}_{i,j})$ los valores centrados y formulamos el siguiente modelo:

$$\mathrm{PI}^*_{i,j} = b_{i,0} + b_{i,1} \cdot t_{i,j} + \varepsilon_{i,j}, \quad (1 \leq i \leq 50, 1 \leq j \leq 4) \tag{3.22}$$

Estos coeficientes se denominan efectos aleatorios (EA) y un modelo de este formato se llama un modelo de efectos aleatorios (MEA). Cada EA tiene una covariable asociada, para diferenciarla de los efectos fijos, utilizaremos la letra Z para referirnos a estas variables, por lo que para cada individuo i tendremos una matriz de diseño que denominaremos $\mathbf{Z}_i$. Con esta definición, podemos puede describir matricialmente el modelo del siguiente modo:

$$\underbrace{\begin{pmatrix} \mathrm{PI}^*_{i,1} \\ \mathrm{PI}^*_{i,2} \\ \mathrm{PI}^*_{i,3} \\ \mathrm{PI}^*_{i,4} \end{pmatrix}}_{\mathbf{PI}^*_i} = \underbrace{\begin{pmatrix} 1 & t_{i,1} \\ 1 & t_{i,2} \\ 1 & t_{i,3} \\ 1 & t_{i,4} \end{pmatrix}}_{\mathbf{Z}_i} \times \underbrace{\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}}_{\mathbf{b}_i} + \underbrace{\begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix}}_{\varepsilon_i} \Rightarrow \mathbf{PI}^*_i = \mathbf{Z}_i \times \mathbf{b}_i + \varepsilon_i \tag{3.23}$$

Es decir, la aleatoriedad de la respuesta proviene tanto de los errores ε_i como del vector de EA $\vec{\mathbf{b}}_i$. En los modelos de este tipo, estas dos fuentes de aleatoriedad se consideran independientes entre sí. Con estas hipótesis, el vector de respuestas tiene como media el vector nulo (dado que tanto $\mathbf{b}_i$ como ε_i tienen media nula) y su matriz de covarianza viene dada por:

$$\mathrm{Cov}[\mathbf{PI}^*_i] = H_i = \mathbf{Z}_i \times \mathbf{G} \times \mathbf{Z}'_i + \sigma^2 \cdot \mathbf{I}_{J_i} \tag{3.24}$$

donde $\mathbf{Z}'_i$ representa la transpuesta de la matriz $\mathbf{Z}_i$.

Notar que se asume que la matriz de covarianza de los errores ε_i se considera como antes de estructura diagonal y de idéntica varianza, es decir, $\mathbf{R}_i = \sigma^2 \cdot \mathbf{I}_{J_i}$. Además, al ser la matriz

de covarianza de $\mathbf{G} \in \mathbb{R}^{2\times 2}$ simétrica y definida positiva, $\mathbf{G} = \begin{pmatrix} g_{0,0} & g_{0,1} \\ g_{1,0} & g_{1,1} \end{pmatrix} = \begin{pmatrix} g_{0,0} & g_{0,1} \\ g_{0,1} & g_{1,1} \end{pmatrix}$. Es decir, consta de 3 parámetros que se consideran idénticos para todos los individuos. Por lo tanto, el modelo dado por 3.22 sólo contiene 4 parámetros desconocidos (3 de la matriz G y σ^2 que corresponde a la varianza de los errores $\varepsilon_{i,j}$). Entonces, si bien ambos modelos 3.20 y 3.22 consideran rectas individuales, éste último requiere de un número considerablemente menor de parámetros.

Notar que en este modelo todos los parámetros son de covarianza, por lo que ninguno de los valores estimados representa un efecto específico sobre la respuesta media. Por lo tanto, cuando se consideran EA se está haciendo una apreciación sobre cómo varían las ordenadas y pendientes individuales, en vez de asignarles un valor, agregando una capacidad predictiva al modelo respecto del dado en 3.20.

Hay otro aspecto que vale aclarar: hasta este momento, como el único componente aleatorio de los modelos provenía del vector de errores $\boldsymbol{\varepsilon}_i$, siempre coincidía su matriz de covarianza $\mathbf{R}_i$ con la del vector de respuestas $\boldsymbol{\Sigma}_i$. Además, hasta este punto siempre se consideró la matriz $\mathbf{R}_i$ con estructura diagonal que pareciera contradecirse con lo mencionado en la sección 2.3, donde se enfatiza que las medidas repetidas en un mismo individuo suelen estar correlacionadas de forma positiva. Sin embargo, al incluir EA en un modelo, éstos tienen un efecto en la covarianza de las respuestas. Por lo tanto, puede considerarse que la correlación entre medidas repetidas es abordada por los mismos EA y que los errores puedan considerarse con estructura diagonal. De este modo, si bien hay toda un área de la literatura dedicada a modelar la matriz de covarianza de los errores $\mathbf{R}_i$ asignando distintas estructuras, en este trabajo asumiremos que tienen la estructura diagonal con componentes idénticos a través de la misma.

Cualquier persona podría preguntarse a esta altura cuáles son los valores que podrían tomar las ordenadas y pendientes individuales, al no ser parámetros fijos del modelo. La respuesta es que se estiman por Máxima Verosimilitud en primera instancia los parámetros $g_{0,0}, g_{0,1}, g_{1,1}$ y σ, obteniendo sus respectivos valores estimados como $\widehat{g_{0,0}}$, $\widehat{g_{0,1}}$, $\widehat{g_{1,1}}$ y $\widehat{\sigma}$. A partir de estos valores estimados, se estima en base a la trayectoria observada $\mathbf{Pl}_i^*$ un vector $\widehat{\mathbf{b}}_i = (\widehat{b_{i,0}}, \widehat{b_{i,1}})$ del siguiente modo:

$$\widehat{\mathbf{b}}_i = \widehat{\mathbf{G}} \times \mathbf{Z}_i' \times \widehat{\mathbf{H}}_i^{-1} \times \mathbf{Pl}_i^* \tag{3.25}$$

donde $\widehat{\mathbf{G}}$ se obtiene reemplazando los valores estimados de $\widehat{g_{0,0}}$, $\widehat{g_{0,1}}$ y $\widehat{g_{1,1}}$, mientras que $\widehat{\mathbf{H}}_i^{-1}$ es similar a la inversa de la matriz $\mathbf{H}_i$ dada en 3.24, pero con los valores estimados de σ y de la matriz $\mathbf{G}$. Es decir:

$$\widehat{\mathbf{H}}_i^{-1} = \left(\mathbf{Z}_i \times \widehat{\mathbf{G}} \times \mathbf{Z}_i' + \widehat{\sigma}^2 \cdot \mathbf{I}_{J_i}\right)^{-1} \tag{3.26}$$

A partir de estos vectores y matrices, se puede conseguir una trayectoria estimada bajo el modelo 3.22 mediante la siguiente ecuación:

$$\widehat{\mathbf{Pl}^*}_i = \mathbf{Z}_i \times \widehat{\mathbf{b}}_i \tag{3.27}$$

En la Figura 3.5 se comparan los valores empíricos y las trayectorias estimadas dadas en las

ecuaciones 3.21 y 3.27: Notar que ambos modelos se adaptan correctamente a la heterogeneidad

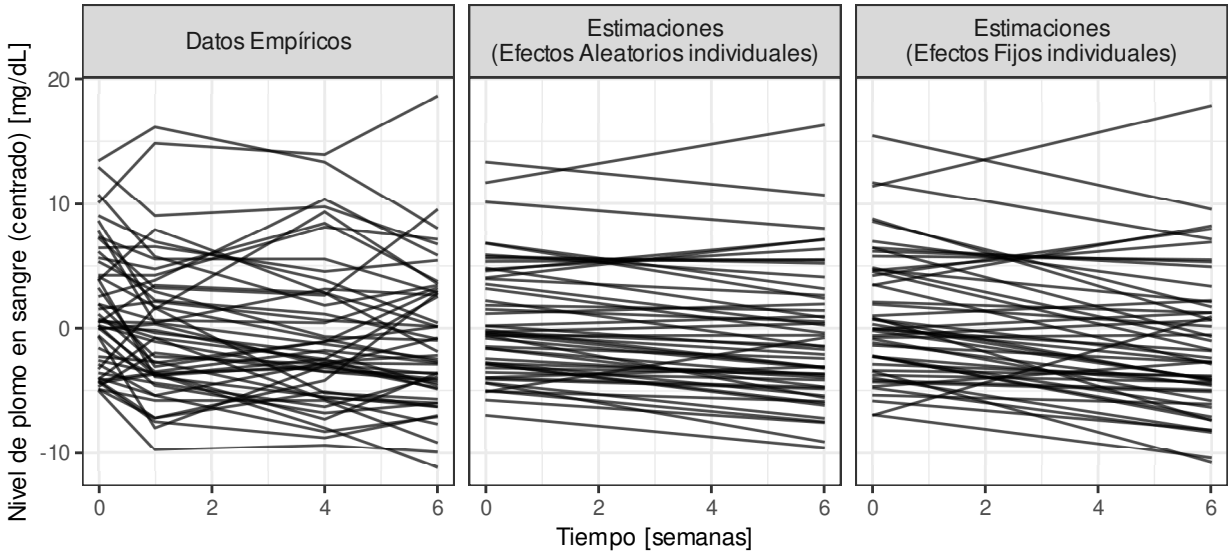


FIGURA 3.5: En esta figura se visualizan las trayectorias de respuesta para la base TLC y la estimación por los modelos dados en 3.20 y 3.22.

de los datos. De todas formas, el MEA depende de una menor cantidad de parámetros, presentando un modelo más parsimonioso.

3.2.2. Modelos de efectos mixtos (MEM)

Habiendo introducido el concepto de EA, se define un modelo de efectos mixtos (MEM) como un modelo que contiene tanto EF como EA. Es decir, los EF modelan la tendencia grupal o poblacional, mientras que los EA añaden una posibilidad de adaptarse las particularidades individuales. Por lo tanto, considerando $\beta \in \mathbb{R}^{P\times 1}$ el vector de EF (comunes a toda la población) y cada vector individual de EA $\mathbf{b}_i \in \mathbb{R}^{Q\times 1}$, se puede generalizar el MEM para cada individuo i mediante la siguiente expresión:

$$\mathbf{Y}_i = \mathbf{X}_i \times \beta + \mathbf{Z}_i \times \mathbf{b}_i + \varepsilon_i \tag{3.28}$$

Como fue mencionado en la sección 3.2.1, se asume que los vectores $\mathbf{b}_i$ y ε_i tienen distribución normal multivariada de media dada por el vector nulo y matrices de covarianzas $\mathbf{G} \in \mathbb{R}^{Q\times Q}$ y $\sigma^2 \cdot \mathbf{I}_{J_i} \in \mathbb{R}^{J_i \times J_i}$, respectivamente. Además, se considera que las distribuciones de estos vectores son independientes entre sí.

Bajo este modelo, combinando lo visto en las ecuaciones 3.6 y 3.24, la trayectoria de respuesta $\mathbf{Y}_i$ sigue la siguiente distribución:

$$\mathbf{Y}_i \sim \mathcal{N}_{J_i}\left(\mathbf{X}_i \times \beta \; ; \; \mathbf{Z}_i \times \mathbf{G} \times \mathbf{Z}_i' + \sigma^2 \cdot \mathbf{I}_{J_i}\right) \tag{3.29}$$

Respecto a la cantidad de parámetros, este modelo considera P parámetros correspondientes a los EF dados en β, la varianza de los errores $\varepsilon_{i,j}$ (es decir, σ^2) y un máximo de $\dfrac{Q\cdot(Q+1)}{2}$

parámetros correspondientes a $\mathbf{G}$,la matriz de covarianza de $\mathbf{b}_i$. Ésto último se debe a que la matriz $\mathbf{G}$ es simétrica. A pesar de que se pueden considerar estructuras en la matriz de covarianza $\mathbf{G}$ que reduzcan la cantidad de parámetros relativos a los EA, en este trabajo asumiremos esta cantidad máxima de parámetros $\frac{Q \cdot (Q+1)}{2}$. Por lo tanto, un MEM tiene $P + \frac{Q \cdot (Q+1)}{2} + 1$ parámetros. De todas formas, vale aclarar que si bien los EA permiten adaptarse a la heterogeneidad de los datos, la cantidad de parámetros asociados crece de forma cuadrática y, por lo tanto, se debe evitar la inclusión excesiva de estos términos.

Una vez estimados los parámetros del modelo, se obtienen $\widehat{\beta}$, la matriz $\widehat{\mathbf{G}}$ y $\widehat{\sigma}$. A partir de estos valores, y una trayectoria de respuesta $\mathbf{Y}_i$, se puede obtener una estimación para el vector de EA $\widehat{\mathbf{b}_i}$:

$$\widehat{\mathbf{b}_i} = \widehat{\mathbf{G}} \times \mathbf{Z}_i' \times \widehat{\mathbf{H}}_i^{-1} \times \left(\mathbf{Y}_i - \mathbf{X}_i \times \widehat{\beta}\right) \tag{3.30}$$

donde $\widehat{\mathbf{H}}_i^{-1}$ coincide con la expresión dada en (3.26). Notar además que la ecuación 3.30 es similar a la expresión obtenida en 3.25. La única diferencia es la resta a la trayectoria de respuesta respecto de su valor medio. Esto se debe a que en la ecuación 3.25 se asumía una media nula, dado que a las trayectorias se les había sustraído previamente la media, por lo que si se considera $\mathbf{Y}_i^* = \mathbf{Y}_i - \mathbf{X}_i \times \widehat{\beta}$, la expresión coincide con la dada en 3.25 (considerando que la variable **PI** representaba la variable de respuesta $\mathbf{Y}$ en ese caso).

Es decir, a partir de estos valores, se puede obtener una estimación poblacional (también llamada respuesta marginal):

$$\widehat{\mathbf{Y}_i^0} = \mathbf{X}_i \times \widehat{\beta} \tag{3.31}$$

que representa el valor esperado para una trayectoria según el comportamiento de todos los individuos de la población.

Además, se puede lograr una estimación individual de una trayectoria del siguiente modo:

$$\widehat{\mathbf{Y}_i} = \mathbf{X}_i \times \widehat{\beta} + \mathbf{Z}_i \times \widehat{\mathbf{b_i}} \tag{3.32}$$

Por lo tanto, estos modelos permiten abordar las particularidades de cada individuo, teniendo en cuenta la estructura poblacional. Es por estas características que son modelos predilectos para el análisis de datos longitudinales en la literatura, dada su capacidad de adaptarse tanto a la variabilidad intersujeto como a la variabilidad intrasujeto.

3.2.3. Aplicaciones a bases de datos

Base TLC

Ya en la sección 3.1.3 se encontró un modelo de EF adecuado para la tendencia poblacional de la base TLC. Recordar que en la Sección 3.1.3 se planteó la posibilidad de no diferenciar las ordenadas según grupo, y en el grupo Placebo se puede asumir una pendiente uniforme para la tendencia grupal a lo largo del estudio, sin necesidad de cambios posteriores a la primer semana. Esto nos permite unificar los términos β_0 y β_3 de la ecuación 3.19, al igual que se puede considerar la remoción del parámetro β_2 de dicha ecuación.

A partir de establecer estos términos como EF poblacionales, se pueden agregar EA individuales. Para cada individuo, se puede considerar un modelo lineal a trozos, es decir, considerar una ordenada, una pendiente y una pendiente para los tiempos posteriores a una semana:

$$\mathbf{Pl}_i = \overbrace{\beta_0 \cdot \mathbf{1} + \beta_1 \cdot \mathbf{t}_i \times \mathbf{Plac}_i + \beta_2 \cdot \mathbf{t}_i \times \mathbf{Trat}_i + \beta_3 \cdot \mathbf{t}_i^* \times \mathbf{Trat}_i}^{\textbf{Efectos fijos}} + \overbrace{b_{i,0} \cdot \mathbf{1} + b_{i,1} \cdot \mathbf{t}_i + b_{i,2} \cdot \mathbf{t}_i^*}^{\textbf{Efectos aleatorios}} + \varepsilon_i \tag{3.33}$$

Bajo este modelo, los parámetros estimados tienen los siguientes valores:

- $\widehat{\beta_0}$ =26.2566891
- $\widehat{\beta_1}$ =-0.3829005
- $\widehat{\beta_2}$ =-13.5450587
- $\widehat{\beta_3}$ =14.9258005
- $\widehat{\sigma}$=3.911242
- $\widehat{\mathbf{G}} = \begin{pmatrix} 11.8580 & 6.4903 & -5.9762 \\ 6.4903 & 7.8956 & -8.8783 \\ -5.9762 & -8.8783 & 10.2540 \end{pmatrix}$

La figura 3.6 permite comparar las trayectorias reales con aquellos obtenidos a través de los parámetros estimados del modelo dado en 3.33.

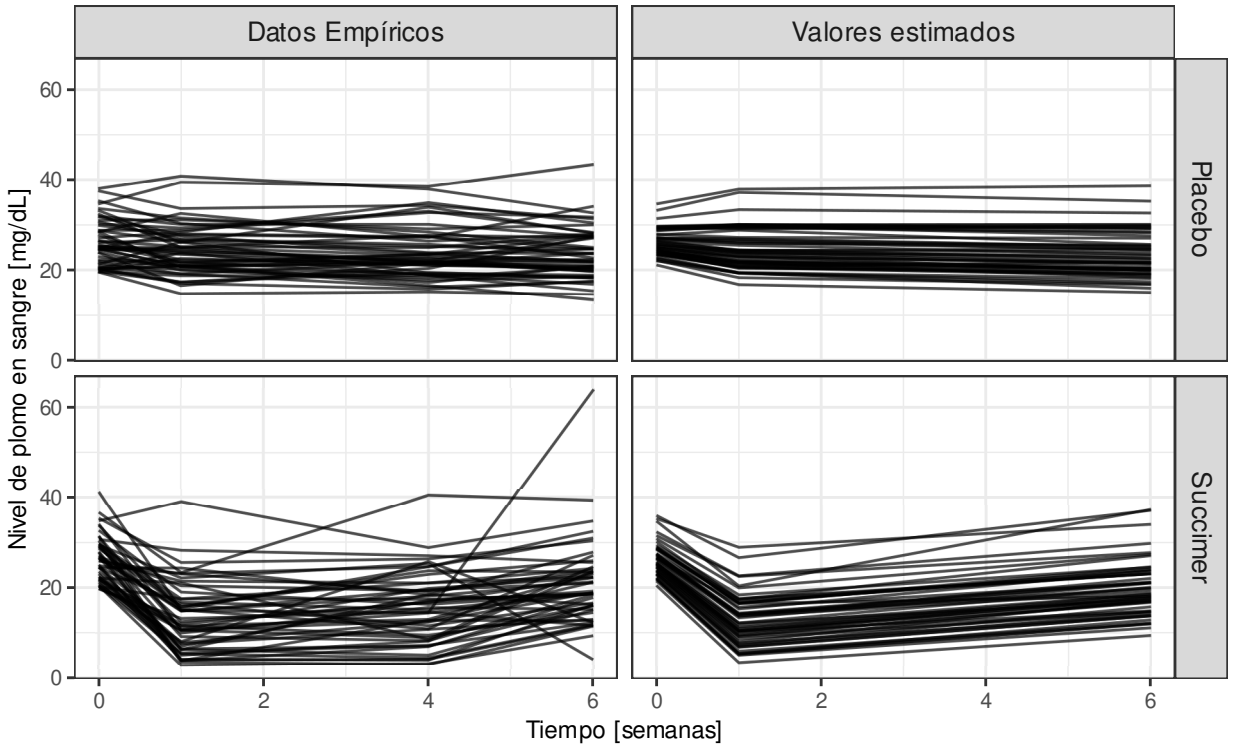


FIGURA 3.6: En esta figura se visualizan las trayectorias de respuesta para la base TLC y la estimación por el modelo dado en 3.33.

Se puede ver en la figura que las estimaciones se adaptan a las heterogeneidades en ambos grupos, salvo por algunas trayectorias atípicas en el grupo Succimer. Esto se debe a que los

parámetros de covarianza estimados para los EA no tienen valores que permitan lograr trayectorias individuales tan desviadas de los valores grupales.

Por otro lado, vale aclarar que en en el grupo Placebo si bien no se considera un modelo lineal a trozos a nivel grupal, las estimaciones individuales contemplan un posible cambio de pendiente a partir de la primera semana, proporcionado por el parámetro individual $b_{i,2}$ en el modelo 3.33.

Base FEV

En la sección 2.2.1 se introdujo la base FEV, en la que se estudia la evolución de la capacidad pulmonar en un grupo pediátrico de la ciudad de Topeka, Kansas. Se ve en la figura 2.1 que las trayectorias no describen una tendencia lineal, sino más bien cóncava. Por lo tanto, en el libro de Fitzmaurice et al. [8], se propone el siguiente MEM para la evolución temporal de la capacidad pulmonar (medida en FEV_1):

$$\log(\mathbf{FEV}_1)_i = \overbrace{\beta_0 \cdot \mathbf{1} + \beta_1 \cdot \mathbf{t}_i + \beta_2 \cdot \log(\mathbf{h})_i + \beta_3 \cdot \mathbf{t}_i^0 + \beta_4 \cdot \log(\mathbf{h}^0)_i}^{\textbf{Efectos fijos}} + \overbrace{b_{i,0} \cdot \mathbf{1} + b_{i,1} \cdot \mathbf{t}_i}^{\textbf{Efectos aleatorios}} + \varepsilon_i \tag{3.34}$$

donde $\mathbf{t}_i$ y $\mathbf{h}_i$ representan las mediciones edad (en años) y la altura (en metros) del individuo i. A su vez, $\mathbf{t}_i^0$ y $\mathbf{h}_i^0$ representan la edad y la altura del individuo i al inicio del estudio. Por otro lado, la expresión log hace referencia al logaritmo natural.

En cuanto a las características del modelo, los logaritmos permiten darle forma cóncava a las tendencias poblacionales. Además, contemplar los valores iniciales tanto de la altura como de la edad (reflejados en los parámetros β_3 y β_4) permite incluir una componente de variabilidad intersujeto que mitiga la varianza de la ordenada aleatoria $b_{i,0}$, permitiendo un modelo más preciso a nivel poblacional. Por otro lado, considerar en los EA sólo ordenadas y pendientes logra que pueda haber aún más diferencias individuales en las trayectorias individuales, sin afectar la forma cóncava poblacional, más allá de los EF β_0 y β_1.

Bajo este modelo, los parámetros estimados tienen los siguientes valores:

- $\widehat{\beta_0}$ =-0.28832334
- $\widehat{\beta_2}$ =2.23719842
- $\widehat{\beta_4}$ =0.21821482
- $\widehat{\beta_1}$ =0.02352863
- $\widehat{\beta_3}$ =-0.01650884
- $\widehat{\sigma}$=0.060237881
- $\widehat{\mathbf{G}} = \begin{pmatrix} 0.01220700 & -0.00043253 \\ -0.00043253 & 0.000050103 \end{pmatrix}$

En la figura 3.7 se pueden visualizar las trayectorias empíricas y estimadas. Se puede ver nuevamente, cómo las trayectorias estimadas terminan siendo cóncavas y adaptadas a la heterogeneidad propia de la base de datos.

Base NGCS

Recordando la base de estudio de colesterol introducido en la sección 2.2.3, se puede observar en la figura 2.3 que las trayectorias describen oscilaciones, pero siempre alrededor de una trayectoria

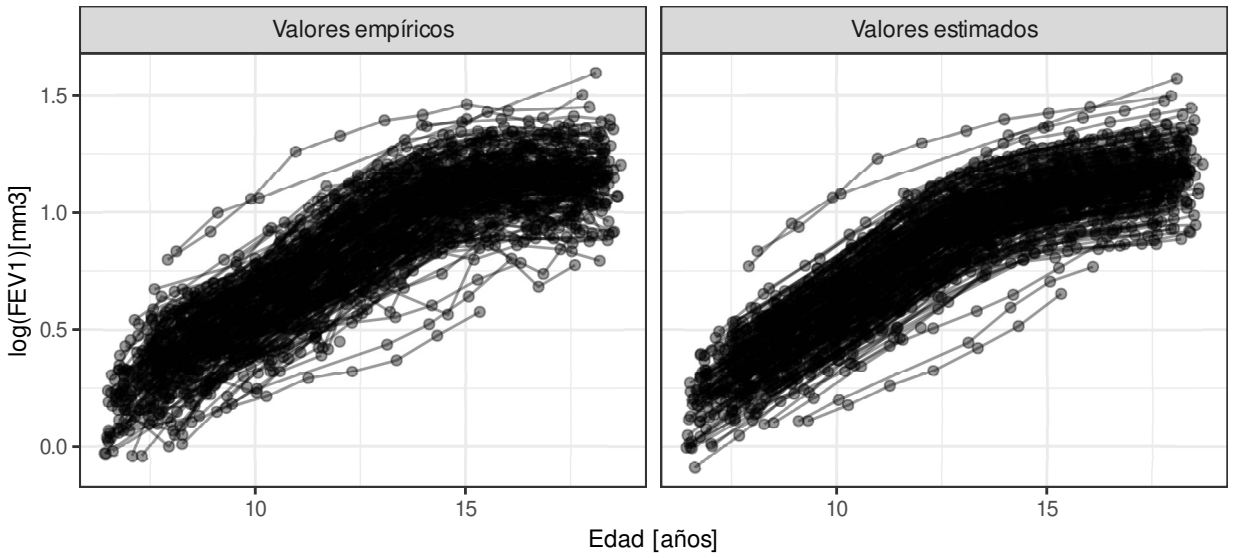


FIGURA 3.7: En esta figura se visualizan las trayectorias de respuesta para la base FEV y la estimación por el modelo dado en 3.34.

estable, por lo tanto, en este modelo podemos considerar una tendencia poblacional lineal por cada grupo y una recta individual que permita desvíos de esta estimación marginal:

$$\mathbf{Chol}_i = \overbrace{\beta_0 \cdot \mathbf{1} \times \mathbf{Plac}_i + \beta_1 \cdot \mathbf{t}_i \times \mathbf{Plac}_i + \beta_2 \cdot \mathbf{1} \times \mathbf{Trat}_i + \beta_3 \cdot \mathbf{t}_i \times \mathbf{Trat}_i}^{\textbf{Efectos fijos}} + \overbrace{b_{i,0} \cdot \mathbf{1} + b_{i,1} \cdot \mathbf{t}_i}^{\textbf{Efectos aleatorios}} + \varepsilon_i \tag{3.35}$$

Bajo este modelo, los parámetros estimados tienen los siguientes valores:

- $\widehat{\beta_0}$ =236.3723219
- $\widehat{\beta_2}$ =232.3257
- $\widehat{\sigma}$=22.6043283
- $\widehat{\beta_1}$ =0.9630882
- $\widehat{\beta_3}$ =1.220365
- $\mathbf{G} = \begin{pmatrix} \widehat{1611.900} & -13.89400 \\ -13.89400 & 0.84779 \end{pmatrix}$

En la figura 3.8 se visualizan las trayectorias empíricas y sus valores estimados. En este caso, no se ve tanta correspondencia entre los valores empíricos y los estimados, aunque puede deberse a que no haya suficientes variables en el modelo para explicar las fluctuaciones del colesterol en sangre respecto de la tendencia lineal.

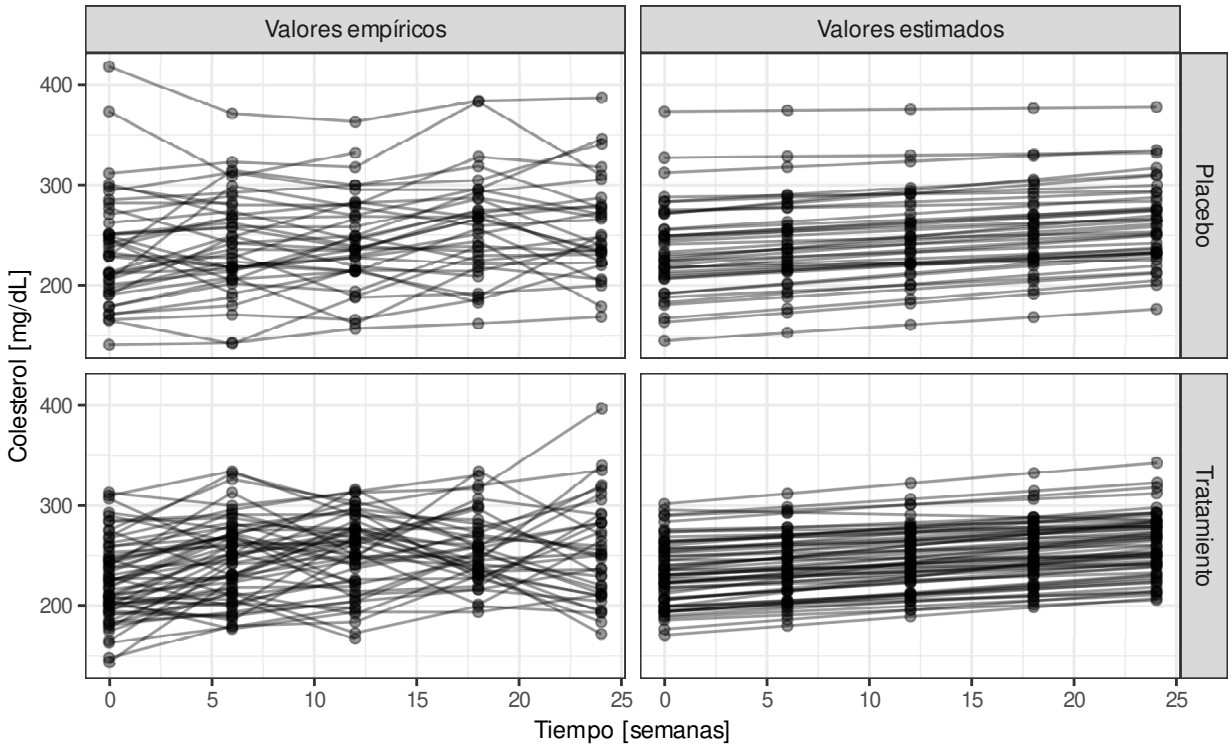


FIGURA 3.8: En esta figura se visualizan las trayectorias de respuesta para la base NGCS y la estimación por el modelo dado en 3.35.

Capítulo 4
Datos Faltantes

Es muy común que los estudios longitudinales tengan un problema de datos faltantes. Esto se puede deber a varios motivos (por ejemplo, omisiones o errores de carga, etc.), pero mayoritariamente tiene que ver con la duración de los estudios, ya que se toman muchas mediciones sobre los mismos individuos y es difícil mantener el cumplimiento en todas las instancias, aún en ensayos clínicos con estrictos protocolos. Por lo tanto, es casi una certeza la ausencia de datos en los estudios longitudinales y esta pérdida de información siempre debe ser considerada en cualquier análisis.

A veces la ausencia de datos no se debe a un problema de cumplimiento, sino que la persona que investiga no considera que la variable que se mide tenga un impacto en los valores de la variable de interés. Esto marca la diferencia entre un dato ausente y un dato faltante. El dato faltante es el dato que se considera de relevancia, pero que por algún motivo no fue registrado. Es decir, un dato faltante involucra una pérdida de información que puede resultar sensible, sobre todo porque algunas metodologías estadísticas requieren que la cantidad de observaciones por individuo sean idénticas para toda la muestra.

En esta sección asumiremos que todas las variables explicativas son observadas y nos referiremos a $\mathbf{Y}^o$ como el vector de respuestas observadas y a $\mathbf{Y}^m$ como el vector de respuestas faltantes.

4.1. Mecanismos de datos faltantes

Para analizar cuanto influyen los datos faltantes sobre las estimaciones, se definen los llamados mecanismos de datos faltantes, que determinan algunas estructuras sobre las distintas causas que ocasionan la falta de alguna respuesta en particular. Para introducir los distintos tipos de mecanismos, vamos a detallarlos usando la base TLC, realizando unos cambios hipotéticos a las condiciones del estudio.

4.1.1. Missing Completely At Random (MCAR)

Supongamos que para alivianar la obligación de los pacientes de tener que presentarse para el estudio, los investigadores sortean completamente al azar 2 individuos que puedan omitir medirse el nivel de plomo en sangre. Asumamos además, que los sorteos son en todos los instantes de medición pactados por el estudio y que las personas puedan salir sorteadas varias veces.

En este caso, la ausencia de una medición no está vinculada ni con la variable de respuesta ni con las covariables. Por lo tanto, la distribución de las respuestas observadas no se ve probabilísticamente afectada por los datos faltantes. En el caso en el que este mecanismo de datos faltantes pueda ser asumido, se dice que el mecanismo es "Missing Completely At Random" (abreviado con las siglas MCAR). Cuando se puede asumir un mecanismo MCAR de datos faltantes, las estimaciones no se ven probabilísticamente afectadas por este mecanismo ya que no se introduce ningún sesgo sistemático sobre el valor esperado de las respuestas, sino que pueden darse completamente al azar. Sin embargo, puede haber alguna pérdida de precisión debido a la reducción del tamaño muestral.

4.1.2. Missing At Random (MAR)

Supongamos otra situación en la que también se sortean individuos para ausentarse del estudio, pero, al saber los investigadores cuáles de los pacientes pertenecen al grupo placebo, realizan dos sorteos, uno por grupo donde se extraen más pacientes del grupo placebo que del grupo tratamiento. También puede darse que el sorteo se realice en una semana específica, con o sin diferencia por grupos.

En cada uno de estos casos, la ausencia de los datos depende de las covariables, no así de la variable de respuesta. En estas condiciones, se dice que el mecanismo de datos faltantes es "Missing At Random" (o MAR). En este tipo de mecanismos de datos faltantes, puede asumirse que, condicional al grupo y al instante de medición, el mecanismo de datos faltantes es MCAR.

Para los mecanismos MAR, no se introduce sesgo sistemático en la resupuesta media de $\mathbf{Y^o}$, aunque también hay pérdida de precisión. De todas formas, esta pérdida de precisión puede ser diversa según distintos niveles de las covariables. Por ejemplo, en caso de que salgan sorteadas más personas del grupo placebo que en el grupo de tratamiento, la pérdida de precisión será mayor en el primer grupo.

4.1.3. Not Missing At Random (NMAR)

Hasta ahora, las ausencias dependían en algún modo de los investigadores o al menos, que estaban bajo su control. En este caso, supongamos que los pacientes con alto nivel de plomo en sangre, no se presentan al estudio por los síntomas causados por esta circunstancia.

En este caso, la ausencia de datos se debe a los valores de la respuesta. Por lo tanto, la distribución de las respuestas observadas $\mathbf{Y^o}$ se ve alterada respecto de las respuestas totales, dado que hay una reducción de la presencia de valores altos y por lo tanto, el valor de la media observada disminuye. Por lo tanto, hay un sesgo sistemático en los valores observados y en este caso, las estimaciones son de menor confiabilidad. Estos mecanismos de datos faltantes son denominados "Not Missing At Random" (abreviado NMAR).

4.2. Simulación de datos faltantes

Para evaluar el rendimiento de algunos estimadores y su robustez ante datos faltantes, puede ser necesario partir de una base completa (con respuestas $Y_{i,j}$) como referencia, calcular las estimaciones para esta base y, luego de remover algunas respuestas (una proporción p_M de la cantidad total), volver a estimar los parámetros en base a esta nueva respuesta (denotada $\check{Y}_{i,j}$) para visualizar cuánto varían respecto de los parámetros reales. Sin embargo, las respuestas no deben removerse sin criterio, ya que según lo visto en 4.1, distintos mecanismos de datos faltantes pueden tener impactos diversos sobre las estimaciones. Más aún, como los datos se remueven bajo el control del investigador, pueden removerse de forma que pueda asumirse cada uno de los tres mecanismos presentados en 4.1.

4.2.1. Missing Completely At Random

En este escenario, la probabilidad de que una observación no esté presente no debe verse afectada por valores observados. Por lo tanto, para lograr una proporción p_M de datos faltantes, se pueden generar números $U_{i,j}$ al azar con distribución uniforme en el intervalo $(0, 1)$ y aplicar la siguiente fórmula:

$$\begin{aligned} U_{i,j} &\sim \mathcal{U}(0,1) \\ \check{Y}_{i,j} &= \begin{cases} NA & \text{si } U_{i,j} \leq p_M \\ Y_{i,j} & \text{si } U_{i,j} > p_M \end{cases} \end{aligned} \tag{4.1}$$

Esta fórmula es simple ya que la remoción de un dato debería ser lo más parecido a un sorteo para no verse influida por otras variables.

4.2.2. Missing At Random

Cuando se busca que la probabilidad de que un dato faltante esté influida por alguna variable explicativa $X_{i,j,p}$, la mayor dificultad que se presenta es que la probabilidad de remover una respuesta debe ser función de las covariables $X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}$, pero a su vez, no debe distar mucho de la proporción deseada p_M, para poder mantener la comparabilidad con otros mecanismos. Nuestra propuesta para obtener la proporción p_M de datos faltantes combina una suma ponderada de las covariables con la función logística. Es decir, consideramos una función con la siguiente estructura:

$$p(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}) = \frac{e^{K(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})}}{1 + e^{K(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})}} \tag{4.2}$$

para asegurarnos que el resultado esté entre 0 y 1 (debe ser una probabilidad). Por otro lado, la función $K(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})$ es la suma ponderada que se define del siguiente modo:

$$K(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}) = C \cdot \sum_{p=1}^{P} \frac{X_{i,j,p}}{s_p} \tag{4.3}$$

donde s_p es el desvío muestral de cada variable explicativa $X_{i,j,p}$, dando un peso inversamente proporcional a las variables de mayor desvío para que la función no presente valores tan dispersos. Por otro lado, la constante normalizadora C (asegura que en promedio la proporción de remociones ronda el valor p_M) se obtiene con la siguiente fórmula:

$$C = \frac{\log\left(\frac{p_M}{1-p_M}\right)}{\sum_{p=1}^{P}\frac{\bar{x}_p}{s_p}} \tag{4.4}$$

donde $\bar{x}_p$ es la media muestral de cada variable $X_{i,j,p}$.

A medida que el tamaño muestral aumenta, los valores de $p(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})$ se acercan a p_M y se puede generar un mecanismo de datos faltantes similar a 4.1 aunque cambiando el valor de la probabilidad p_M:

$$\begin{aligned} &U_{i,j} \sim \mathcal{U}(0,1) \\ &\breve{Y}_{i,j} = \begin{cases} NA & \text{si } U_{i,j} \leq p(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}) \\ Y_{i,j} & \text{si } U_{i,j} > p(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}) \end{cases} \end{aligned} \tag{4.5}$$

4.2.3. Not Missing At Random

Para incluir el valor de las respuestas en la probabilidad de remoción de una respuesta, se aborda de modo similar al de la sección 4.2.2, aunque agregando un término para la media y desvío muestral de la variable $Y_{i,j}$ ($\bar{y}$ y s_Y, respectivamente). Es decir, definiendo $\widetilde{p}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}, Y_{i,j})$ del siguiente modo:

$$\widetilde{p}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}, Y_{i,j}) = \frac{e^{\widetilde{K}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})}}{1 + e^{\widetilde{K}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P})}} \tag{4.6}$$

donde

$$\widetilde{K}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}, Y_{i,j}) = \widetilde{C} \cdot \left(\frac{Y_{i,j}}{s_Y} + \sum_{p=1}^{P}\frac{X_{i,j,p}}{s_p}\right) \tag{4.7}$$

y la constante normalizadora viene dada por:

$$\widetilde{C} = \frac{\log\left(\frac{p_M}{1-p_M}\right)}{\frac{\bar{y}}{s_Y} + \sum_{p=1}^{P}\frac{\bar{x}_p}{s_p}} \tag{4.8}$$

Con estas definiciones, se procede a remover las respuestas correspondiente

$$\begin{aligned} &U_{i,j} \sim \mathcal{U}(0,1) \\ &\check{Y}_{i,j} = \begin{cases} NA & \text{si } U_{i,j} \leq \widetilde{p}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}, Y_{i,j}) \\ Y_{i,j} & \text{si } U_{i,j} > \widetilde{p}(X_{i,j,1}, X_{i,j,2}, \cdots, X_{i,j,P}, Y_{i,j}) \end{cases} \end{aligned} \tag{4.9}$$

4.3. Modelos mixtos y datos faltantes

Los MEM (o en general para los modelos lineales de regresión) pueden adaptarse a los datos faltantes perdiendo la menor cantidad de información posible. Más aún, en todas las notaciones anteriores presentadas en este trabajo se admite la posibilidad de que los individuos puedan presentar distintos números de mediciones, denotados J_i.

Esta adaptabilidad se debe a que en este trabajo, generalmente consideramos una variable temporal cuantitativa, y que si en un instante hay un dato faltante, la tendencia temporal está modelada y puede removerse únicamente esa medición, sin tener que perder toda la información sobre ese individuo. Vale aclarar que para que esto suceda los datos deben estar estructurados en formato long (ver Sección 2.2.5), ya que en caso contrario se remueven todas las respuestas correspondientes a dicho sujeto. Además, al considerar modelos lineales, los mismos se pueden representar a través de matrices, y de esas matrices pueden removerse sólo las filas problemáticas. Por ejemplo, para un individuo del grupo tratamiento, la ecuación 3.33 puede expresarse matricialmente del siguiente modo:

$$\begin{pmatrix} \text{Pl}_{i,1} \\ \text{Pl}_{i,2} \\ \text{Pl}_{i,3} \\ \text{Pl}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} & t^*_{i,1} & t_{i,1} \\ 1 & t_{i,2} & t^*_{i,2} & t_{i,2} \\ 1 & t_{i,3} & t^*_{i,3} & t_{i,3} \\ 1 & t_{i,4} & t^*_{i,4} & t_{i,4} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 & t_{i,1} & t^*_{i,1} \\ 1 & t_{i,2} & t^*_{i,2} \\ 1 & t_{i,3} & t^*_{i,3} \\ 1 & t_{i,4} & t^*_{i,4} \end{pmatrix} \times \begin{pmatrix} b_{i,0} \\ b_{i,1} \\ b_{i,2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix} \tag{4.10}$$

Supongamos que para este individuo la tercer medición de plomo en sangre es una respuesta faltante, es decir, $\text{Pl}_{i,3} = NA$. La información sobre ese individuo no es necesario perderla porque se puede plantear la siguiente representación matricial:

$$\begin{pmatrix} \text{Pl}_{i,1} \\ \text{Pl}_{i,2} \\ \text{Pl}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & t_{i,1} & t^*_{i,1} & t_{i,1} \\ 1 & t_{i,2} & t^*_{i,2} & t_{i,2} \\ 1 & t_{i,4} & t^*_{i,4} & t_{i,4} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 & t_{i,1} & t^*_{i,1} \\ 1 & t_{i,2} & t^*_{i,2} \\ 1 & t_{i,4} & t^*_{i,4} \end{pmatrix} \times \begin{pmatrix} b_{i,0} \\ b_{i,1} \\ b_{i,2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,4} \end{pmatrix} \tag{4.11}$$

Notar que esta adaptabilidad es posible porque las remociones son en las matrices de diseño, pero no impactan sobre los parámetros. De todas formas, si bien la pérdida de información es mínima, hay un impacto sobre la precisión de las estimaciones.

4.4. Cuestiones prácticas y antecedentes

En la práctica, los mecanismos de datos faltantes descriptos en 4.1 son hipótesis que deben asumirse, pero que no pueden testearse. Por lo tanto, la hipótesis de mecanismos MCAR o MAR sólo puede darse en estudios prospectivos en los que se mantiene un control estricto sobre todos los efectos posibles, y aún en esos casos las hipótesis pueden ser poco realistas. En estudios observacionales, en general debe asumirse un mecanismo NMAR.

Para reducir el impacto del potencial sesgo se suelen analizar los patrones de datos faltantes (llamados $\mathbf{M}$). A diferencia de los mecanismos detallados en 4.1, los patrones de datos faltantes son analizados a través del vector binario $\mathbf{M} \in \mathbb{R}^N$, donde $N = \sum_{i=1}^{I} J_i$, es decir, la cantidad de mediciones totales contemplando los individuos y sus mediciones repetidas. El valor de cada coordenada $\mathbf{M}_{i,j}$ es 1 si la respuesta $Y_{i,j}$ es observada y 0 en caso contrario. De todas formas, ambos conceptos pueden estar vinculados. Cuando se puede considerar que el patrón de datos faltantes viene de mecanismos del tipo MCAR o MAR, se dice que el patrón de datos faltantes es "ignorable", mientras que para mecanismos del tipo NMAR, se denomina "no-ignorable".

Los patrones de ceros y unos en el vector $\mathbf{M}$ pueden motivar distintas propuestas metodológicas. Por ejemplo, cuando los patrones se pueden considerar monótonos (es decir, sólo permite deserciones del estudio o "dropouts": una vez que en un individuo hay una respuesta faltante, también son faltantes todas las mediciones subsiguientes), se puede imponer una estructura probabilística sobre los vectores de respuesta que permiten mejorar las estimaciones.

Sin embargo, en general los patrones de datos faltantes no son monótonos y puede haber datos faltantes en algunos instantes pero no necesariamente en los subsiguientes. En ese caso, hay distintas herramientas (selection models, pattern-mixture models y shared parameters models) que logran estimaciones insesgadas, involucrando factorizaciones de la verosimilitud conjunta de los vectores $\mathbf{Y}^o$ y $\mathbf{M}$.

Vale aclarar nuevamente que estos modelos no están exentos de hipótesis que pueden ser poco realistas y que además no pueden verificarse (aunque en algunos casos pueden llegar a descartarse), sobre todo si se analizan los datos en etapas posteriores al estudio o si los estudios son observacionales. Por otro lado, las causas de datos faltantes pueden ser numerosas y muy vinculadas a la aplicación que se estudia. Por lo tanto, en general estas herramientas sofisticadas no son aplicables en muchos casos y el análisis no puede abordarse de forma insesgada. Es decir, con motivación puramente pragmática, se acepta un pequeño sesgo a fin de poder llevar a cabo el análisis.

Capítulo 5
Métodos de Clustering

5.1. Definiciones

Como se mencionó en el Capítulo 1, cuando se tiene poca información previa sobre una base de datos, en ocasiones se recurre a herramientas automáticas que permitan visualizar cierta estructura en los datos. Estos procedimientos se llaman de aprendizaje no supervisado, justamente por el desconocimiento sobre los patrones que se intentan buscar.

Los algoritmos de clustering permiten agrupar datos similares entre sí, o separar datos disímiles entre sí. Los criterios para determinar cuán similares o disímiles son dos observaciones son vastos ya que, por un lado, el concepto no es específico y puede interpretarse de varias maneras, y por otro lado, puede aplicarse a variables tanto cuantitativas como cualitativas. Por otro lado, según los objetivos del estudio, se puede intentar buscar un concepto de similitud (o disimilitud) en especial que no siempre es ofrecido por los criterios mayoritariamente instalados. Por lo tanto, las opciones son muy variadas y dependen mucho del área de aplicación.

Cuando las observaciones contienen únicamente variables cuantitativas, se puede hacer uso de conceptos preestablecidos de distancia entre dos vectores, como la norma euclídea, aunque pueden usarse otras funciones de distancia.

Una vez establecido el criterio de agrupamiento, los grupos (también llamados clusters) pueden determinarse:

- de manera binaria: cada elemento o bien pertenece o bien no pertenece a cierto grupo. En inglés, son denotados como “crisp partitions” o “hard clustering”.

- de forma continua: cada elemento tiene un grado de pertenencia (entre 0 y 1) a un cierto grupo y se agrupan los datos que tienen un grado mínimo de asociación. En inglés, son denotados como “soft partitions” o “fuzzy clustering”.

5.2. Métodos

En este trabajo nos focalizamos en el uso de algoritmos de hard clustering, dado que nuestro objetivo final es comparar los grupos resultantes entre sí. Por lo tanto, esta comparación pierde sentido cuando se establece que un elemento pertenece o no a cierto grupo según un grado de asociación. Por otro lado, consideraremos variables cuantitativas para las observaciones.

Entre estos algoritmos, nos focalizamos en el uso de K-medias, K-medias basado en kernels, K-medoides, métodos de clustering jerárquico y métodos de modelos mixtos de clases latentes. Este último método tiene una estructura diferente a la de los otros algoritmos de clustering y se presenta aparte en la sección 1.3.2. Para más detalles sobre los algoritmos, referirse a los libros de Hastie, Tibshirani & Friedman [45] y Wierzchoń & Kłopotek [46]

5.2.1. Notación

Para describir los algoritmos de clustering, consideraremos n vectores de observaciones $x_i \in \mathbb{R}^m$. Estos vectores se dividen en $K \leq n$ clusters $\{\mathcal{C}_k\}_{1 \leq k \leq K}$, cada uno de estos clusters con n_k observaciones de forma que $\sum_{k=1}^{K} n_k = n$.

Por otro lado, para cada cluster se puede considerar una media grupal $\mu_k = \frac{\sum_{x_i \in \mathcal{C}_k} x_i}{n_k}$ y una media poblacional notada $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$.

5.2.2. K-Medias

El algoritmo de K-medias es el algoritmo de clustering más popular, mayoritariamente debido a su bajo costo computacional y por lo tanto, es muy utilizado en bases de datos masivas. El número de clusters K se determina previo a aplicar del algoritmo.

El objetivo del algoritmo es establecer K centros μ_k que representen promedios grupales, y buscar una partición que minimice la variación intragrupo. Esta variación se basa en la distancia euclidea entre observaciones de cada grupo respecto de su centro μ_k. Es decir, se busca minimizar:

$$W = \sum_{k=1}^{K} \sum_{x_i \in \mathcal{C}_k} (x_i - \mu_k)^2 \tag{5.1}$$

Sin embargo, este problema de optimización es $\mathcal{NP}$-hard y no tiene soluciones de complejidad polinómica. Por lo tanto, se busca un mínimo local mediante una estrategia heurística iterativa: se asigna de forma aleatoria cada observación a uno de los K grupos, en cada grupo se calcula la media μ_k como centro del grupo y luego, cada observación se reasigna al grupo correspondiente al centro más cercano. Este procedimiento se repite hasta que las asignaciones de los grupos no varían luego de la iteración. Por otro lado, como en cada paso se realiza un promedio, los cálculos son muy sensibles a datos atípicos y a la asignación inicial que afecta los primeros cálculos.

Notar además que la expresión dada en 5.1 da cuenta de cómo el método de K-medias utiliza la distancia euclídea entre vectores, ya que lo que se suma es la diferencia entre vectores elevadas al cuadrado.

5.2.3. K-Medias basado en Kernels

Una variante de este algoritmo se basa en transformar previamente las observaciones mediante una función $\phi : \mathbb{R}^m \mapsto \mathbb{R}^{\tilde{m}}$, para luego aplicar el algoritmo de K-medias a este espacio transfor-

mado. La utilización de Kernels tiene como objetivo que las observaciones transformadas tengan mayor separabilidad lineal en el nuevo espacio.

Es decir, se busca minimizar la siguiente expresión:

$$W_\phi = \sum_{k=1}^{K} \sum_{x_i \in \mathcal{C}_k} (\phi(x_i) - \mu_k^\phi)^2 \tag{5.2}$$

donde μ_k^ϕ representa la media grupal de los vectores transformados.

Al igual que K-medias, el mínimo local se obtiene de forma iterativa. Además, comparte las mismas características de sensibilidad a los datos extremos y la inicialización.

5.2.4. Jerárquico

Una alternativa a K-medias es el clustering jerárquico. A diferencia de K-medias, no se necesita saber de antemano el número de grupos, ya que el algoritmo devuelve un conjunto de particiones de todos los tamaños entre 1 y la cantidad total de observaciones n.

En su versión aglomerativa, el clustering jerárquico comienza con una partición de n elementos, un cluster por cada observación, y a partir de ellos, en cada paso, aglomera los clusters más similares entre sí, hasta juntar todas las observaciones en un único cluster. Por lo tanto, para obtener una partición de K clusters, debemos remitirnos al $(n-K)$-ésimo paso de este proceso de aglomeración.

El criterio para fusionar los dos clusters más similares puede variar (a diferencia de K-medias), ya que primero debe establecerse una distancia entre observaciones y luego una distancia entre clusters. Los métodos de aglomeración más utilizados son:

- el completo ("complete" en inglés), donde se fusionan clusters según los que tengan menor distancia *máxima* entre sus elementos.
- el promedio ("average" en inglés), donde se fusionan clusters según los que tengan menor distancia *promedio* entre sus elementos.
- el individual ("single" en inglés), donde se fusionan clusters según los que tengan menor distancia *mínima* entre sus elementos.

5.2.5. Modelos Mixtos de clases latentes

Esta metodología fue descripta en la sección 1.3.3. Habiendo explicado los conceptos de efectos mixtos en el capítulo 3, podemos detallar más el procedimiento.

Los trabajos que usan esta metodología suelen asumir un modelo polinomial (supongamos que tiene orden H) para la tendencia temporal de los datos (eligiendo las potencias de la variable temporal t, es decir, t^h con $0 \leq h \leq H$), donde los coeficientes de cada potencia involucran tanto efectos fijos poblacionales como aleatorios individuales.

Es decir, la trayectoria de respuesta para un individuo de un grupo k sigue el siguiente modelo:

$$\begin{array}{c} Y_{i,j}^{(k)} = \sum_{h=0}^{H} \alpha_h^{(k)} \cdot t_{i,j}^h + \varepsilon_{i,j} \\ \alpha_h^{(k)} = \beta_h^{(k)} + \gamma_{h,i}^{(k)} \end{array} \quad , \quad (1 \leq i \leq I,\ 1 \leq j \leq J_i,\ 1 \leq k \leq K) \tag{5.3}$$

donde $\beta_h^{(k)}$ es un efecto fijo para cada coeficiente polinomial del modelo y $\gamma_{h,i}^{(k)}$ es un efecto aleatorio individual para el mismo coeficiente.

Una vez planteado el modelo, se calculan los parámetros por máxima verosimilitud (donde se incluyen los parámetros π_k como la probabilidad de que un individuo pertenezca a cada grupo k). Con los parámetros estimados, se calcula una probabilidad a posteriori de que cada individuo pertenezca a cada grupo y se clasifica a cada individuo dentro del grupo más "probable". Este agrupamiento resultante establece una partición que se toma como agrupamiento automático.

5.2.6. K-Medoides

Otra alternativa de clustering es el algoritmo de K-medoides. A diferencia de K-medias, este procedimiento toma como centros observaciones de la base de datos, reemplazando los promedios grupales que pueden ser muy sensibles a datos atípicos.

Para lograr la partición, se eligen inicialmente al azar K observaciones como centros de cada grupo. Luego, se asigna cada una de las $n-K$ observaciones restantes al grupo correspondiente al centro más cercano. En el siguiente paso, se actualizan los centros a la observación que minimice la distancia intragrupo. Este paso se repite hasta que los grupos no se ven alterados.

Nuevamente, también para este procedimiento hay variantes a la hora de elegir la distancia entre observaciones, a diferencia de K-medias.

Vale aclarar que este algoritmo tiene una mayor complejidad computacional que K-medias. Sin embargo, la versatilidad a la hora de poder elegir las distancias entre observaciones le permite ser más robusto.

5.3. Construcción del espacio de las pendientes

En esta sección asumiremos que todos los individuos tienen la misma cantidad de mediciones y que se indexa como cero su valor basal. Es decir, para cada individuo i, la trayectoria de respuesta viene dada por $\mathbf{Y}_i = (Y_{i,0}, Y_{i,1}, \cdots, Y_{i,J})$ e instantes de medición $\mathbf{t}_i = (t_{i,0}, t_{i,1}, \cdots, t_{i,J})$.

Como se menciona en el Capítulo 1, uno de los objetivos de este trabajo es asociar individuos según la variación en las trayectorias de respuesta $\mathbf{Y}_i$. Por ejemplo, en la Figura 5.1, la partición final debería considerar similares a los individuos 1 y 3, ya que las trayectorias tienen una cierta estabilidad. Del mismo modo, debería considerar a los individuos 2 y 4 como similares ya que ambos muestran un incremento inicial con un posterior decrecimiento.

Sin embargo, la mayoría de las distancias entre vectores tradicionales, considerarían cercanos las trayectorias correspondientes a los individuos 1 y 2 o al par de sujetos 3 y 4, ya que numéricamente son similares sus coordenadas. Esto va en contra de los objetivos planteados anteriormente. Por lo tanto, en vez de aplicar los algoritmos de clustering a estos vectores, construiremos un

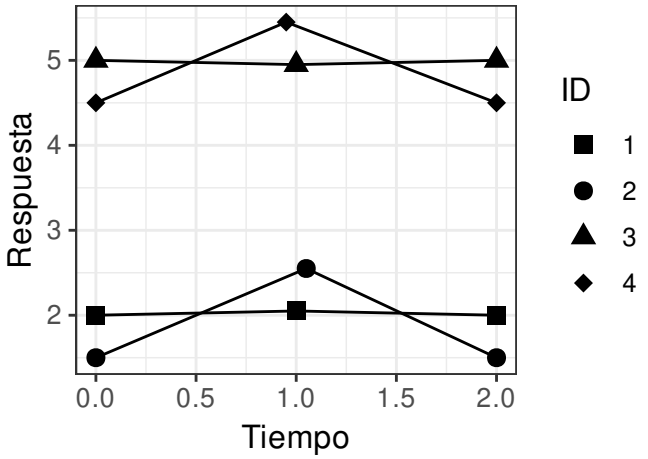


FIGURA 5.1: En esta figura se visualizan las trayectorias de respuesta para un caso hipotético. Notar que los instantes de medición no son idénticos para todos los individuos.

espacio vectorial en el que se contemplen los cambios en la respuesta respecto del tiempo t entre mediciones.

A partir de estos datos, podemos definir las pendientes $m_{i,j}$:

$$m_{i,j} = \frac{Y_{i,j} - Y_{i,j-1}}{t_{i,j} - t_{i,j-1}} \quad (1 \leq i \leq I,\ 1 \leq j \leq J) \tag{5.4}$$

Notar que el espacio de las trayectorias consiste de $J+1$ coordenadas, mientras que el espacio de las pendientes se compone de J coordenadas. Por ejemplo, en la figura 5.2 se visualizan las correspondencias entre las trayectorias de la Figura 5.1 y el espacio de las pendientes. Notar que en este espacio, los individuos 1 y 3 resultan mucho más cercanos que en el espacio de las trayectorias. Lo mismo sucede con los individuos 2 y 4. Es decir, en este espacio, las distancias entre trayectorias estables y las que presentan la misma estructura en sus variaciones se reducen notablemente.

En el espacio de las pendientes, los valores positivos corresponden a crecimientos en la variable de respuesta y en contraposición, los valores negativos corresponden a decrecimientos en la variable de respuesta. Bajo esta calificación, parecería natural agrupar individuos según el signo de cada pendiente. Sin embargo, este criterio omite las magnitudes de las pendientes. Por ejemplo, dicho criterio consideraría en grupos distintos a los individuos 1 y 3, que reflejan trayectorias estables, ya que son similares al tener pequeño valor absoluto en sus coordenadas.

Por lo tanto, en este espacio vectorial se pueden establecer distintas funciones de distancia entre sus elementos. Por ejemplo, se puede aplicar la distancia euclidea entre las pendientes de los individuos i e i':

$$||\mathbf{m}_i - \mathbf{m}_{i'}||_2 = \sqrt{\sum_{j=1}^{J}(m_{i,j} - m_{i',j})^2} = \sqrt{\sum_{j=1}^{J}\left(\frac{Y_{i,j} - Y_{i,j-1}}{t_{i,j} - t_{i,j-1}} - \frac{Y_{i',j} - Y_{i',j-1}}{t_{i',j} - t_{i',j-1}}\right)^2} \tag{5.5}$$

Esta distancia tiende a magnificar las diferencias entre pendientes y por lo tanto, es más

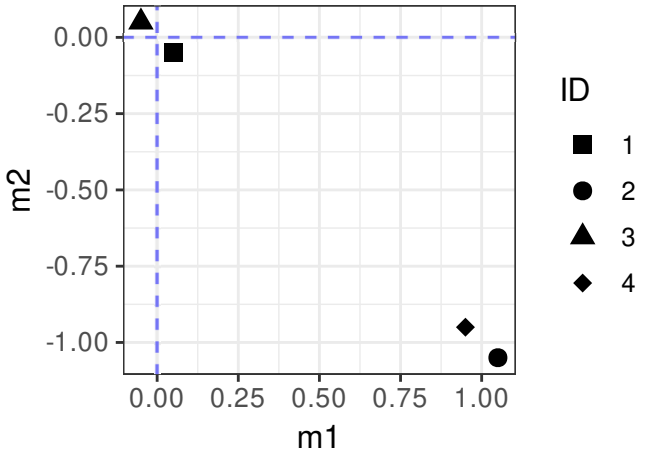


FIGURA 5.2: En esta figura se visualizan las observaciones representadas en el espacio de las pendientes para el caso hipotético de la Figura 5.1.

propensa a ser afectada por valores atípicos. Como alternativa más robusta se puede usar la distancia usualmente denominada Manhattan:

$$||\mathbf{m}_i - \mathbf{m}_{i'}||_1 = \sum_{j=1}^{J} |m_{i,j} - m_{i',j}| = \sum_{j=1}^{J} \left| \frac{Y_{i,j} - Y_{i,j-1}}{t_{i,j} - t_{i,j-1}} - \frac{Y_{i',j} - Y_{i',j-1}}{t_{i',j} - t_{i',j-1}} \right| \tag{5.6}$$

Por lo tanto, notando que distancias pequeñas entre observaciones del espacio de las pendientes, se corresponden con trayectorias similares en morfología, aplicar algoritmos tradicionales de clustering en este espacio debería agrupar individuos con los mismos patrones en la variación de sus respuestas. Esta propuesta expande las vastas opciones de algoritmos de clustering, ya que en vez de aplicarlo en el espacio de las trayectorias, se pueden aplicar en el espacio de las pendientes, aunque con mejores resultados a la hora de agrupar sujetos con trayectorias morfológicamente similares.

En cuanto a los algoritmos de clustering descriptos en la sección 5.2, los procedimientos de K-Medias y K-Medias basado en kernels utilizan la distancia Euclidea entre vectores. Por otro lado, los algoritmos de K-Medoides y clustering Jerárquicos pueden valerse de otras funciones de distancia entre vectores como la Manhattan.

5.4. Transformaciones

En la práctica, a veces se aplican transformaciones a las variables previo al algoritmo de clustering, con el objetivo de reducir el impacto de valores extremos y las diferencias entre escalas de distintas variables.

5.4.1. Datos escalados

La transformación usualmente aplicada es el escalado de cada variable, en las que los datos son restados por su media y se le divide el desvío. Por ejemplo, la variable de respuesta se puede escalar del siguiente modo:

$$\widetilde{Y}_{i,j} = \frac{Y_{i,j} - \overline{Y}_j}{s_j^Y} \tag{5.7}$$

donde $\overline{Y}_j = \dfrac{\sum_{i=1}^{I} Y_{i,j}}{I}$ y $s_j^Y = \sqrt{\dfrac{\sum_{i=1}^{I}(Y_{i,j} - \overline{Y}_j)^2}{I-1}}$. Del mismo modo, se pueden escalar los tiempos reemplazando la variable de respuesta $\mathbf{Y}$ por la variable temporal $\mathbf{t}$ en la ecuación 5.7:

$$\widetilde{t}_{i,j} = \frac{t_{i,j} - \overline{t}_j}{s_j^t} \tag{5.8}$$

5.4.2. Datos normalizados

La normalización consiste en aplicar una transformación lineal a los datos para que la imagen quede comprendida en un intervalo determinado. La normalización más utilizada toma como referencia el intervalo [0,1]. Sin embargo, como en el espacio de las pendientes los signos tienen una interpretación, fijaremos el intervalo de normalización como el [-1,1]. Por ejemplo, las respuestas pueden normalizarse mediante la siguiente fórmula:

$$\breve{Y}_{i,j} = \frac{2 \cdot Y_{i,j} - a_j^Y - b_j^Y}{b_j^Y - a_j^Y} \tag{5.9}$$

donde $a_j^Y = \text{mín}_{1 \le i \le I}\{Y_{i,j}\}$ y $b_j^Y = \text{máx}_{1 \le i \le I}\{Y_{i,j}\}$. Notar que bajo esta construcción, el valor máximo posible de $Y_{i,j}$ para distintos valores de i es b_j^Y y que el correspondiente valor transformado es 1 y de manera análoga, el mínimo valor de la transformación es -1 cuando $Y_{i,j} = a_j^Y$. Nuevamente, se puede aplicar la misma transformación a la variable $\mathbf{t}$.

5.4.3. Transformaciones previas y posteriores

La propuesta de este trabajo añade una nueva opción a la hora de realizar procesamientos de las variables. Tanto la normalización como el escalado que se propone en las secciones 5.4.1 y 5.4.2, se puede aplicar también al espacio de las pendientes. Por lo tanto, el escalado o la normalización pueden ser previas al cálculo de las pendientes (utilizando $\widetilde{\mathbf{Y}}_i$ y $\widetilde{\mathbf{t}}_i$, o $\breve{\mathbf{Y}}_i$ y $\breve{\mathbf{t}}_i$ respectivamente), o posterior al cálculo de las pendientes (utilizando $\widetilde{\mathbf{m}}_i$ o $\breve{\mathbf{m}}_i$ respectivamente).

5.5. Criterios de calidad

Una vez terminado un algoritmo de clustering, se obtiene una partición que divide a los individuos en grupos. Sin embargo, no siempre es muy claro cómo evaluar la calidad dicha partición.

Es decir, hay que establecer una medida que permita determinar cuál de las particiones es "mejor" que otra. Más aún, como se mencionó anteriormente, las distintas aplicaciones tienen distintos objetivos. Por lo tanto, la noción de "calidad" puede variar según los objetivos del estudio.

De todas formas, los criterios de calidad pueden clasificarse en dos grupos: los criterios internos y los criterios externos. Los criterios internos buscan evaluar algunas características deseables de una partición, mientras que los criterios externos buscan evaluar la concordancia entre dos particiones. Éstos últimos son de mucha utilidad cuando se tiene una partición de referencia y se busca que un agrupamiento automático se asemeje a aquellos dados por los grupos de referencia.

5.5.1. Criterios internos

Hay muchas opciones para los criterios internos. Sin embargo, puede considerarse que siempre tienen como objetivo capturar alguna de las siguientes cualidades:

- **Homogeneidad:** Las observaciones pertenecientes al mismo cluster, deberían ser similares entre sí. En el caso en que la similitud se mida con una función de distancia, esto se reduce a que la distancia entre observaciones del mismo grupo sea pequeña. Una posible medida de homogeneidad es la llamada suma de cuadrados intragrupo (WGSS: Within Group Sum of Squares), basada en los cuadrados de las distancias euclídeas entre observaciones y su correspondiente centro:

$$WGSS = \sum_{k=1}^{K} \sum_{x_i \in \mathcal{C}_k} ||x_i - \mu_k||_2^2 \tag{5.10}$$

  Mientras menor sea el valor de esta medida cuantitativa, más homogéneos serán los grupos ya que la mayoría de las observaciones $x_i \in \mathcal{C}_k$ tienen mayor cercanía a su centro μ_k (definido en la sección 5.2.1).

- **Separabilidad:** Por otro lado, las observaciones pertenecientes a distintos clusters, deberían estar más alejadas entre sí. Nuevamente considerando una distancia como medida de disimilitud, una posible medida de separabilidad es la llamada suma de cuadrados entre grupos (BGSS: Between Group Sum of Squares):

$$BGSS = \sum_{k=1}^{K} n_k \cdot ||\mu_k - \mu||_2^2 \tag{5.11}$$

  Mientras mayor sea el valor de esta medida cuantitativa, más separados serán los grupos ya que la mayoría de los centros de cada grupo μ_k tienen mayor distancia al promedio total μ (ver Sección 5.2.1).

Con estos ejemplos, un clustering que tenga un pequeño valor de WGSS y un gran valor de BGSS será deseable ya que cada cluster tendría buena concentración de observaciones y los clusters estarían bien separados entre sí. En base a este razonamiento, se construye el índice de Calinski-Harabasz:

$$CH = \frac{\frac{BGSS}{n-1}}{\frac{WGSS}{n-K}} = \frac{BGSS \cdot (n-K)}{WGSS \cdot (n-1)} \tag{5.12}$$

Para este índice, a medida que aumenta el BGSS aumenta el valor del cociente. El índice también aumenta a medida que decrece WGSS. Por lo tanto, en el caso de tener que elegir entre dos particiones, un criterio posible es eligir el que posee un mayor índice de Calinski-Harabasz. Notar además que el índice decrece cuando aumenta la cantidad de clusters, por lo que penaliza particiones en grupos excesivos.

También basados en estos principios de homogeneidad y separabilidad, hay muchos otros índices (ver [46]), aunque no siempre tan interpretables como el presentado en la ecuación 5.12. Además, para algunos criterios las mejores particiones corresponden a menores valores del índice correspondiente, por lo que antes de utilizarlo debe considerarse previamente si se trata de un índice "creciente" o "decreciente".

5.5.2. Criterios externos

Otra forma de comparar clusters no es necesariamente en función de propiedades deseables, sino evaluando cuánta concordancia tienen dos particiones en términos de si terminan agrupando a las mismas observaciones en los mismos grupos.

La dificultad de analizar la concordancia entre particiones reside en las distintas nomenclaturas que pueden tener los grupos, ya que casi todo algoritmo tiene algún componente de aleatoriedad. Por ejemplo, al correr dos veces el mismo algoritmo, la misma observación puede estar en grupos con etiquetas distintas al finalizar cada corrida. Por ejemplo, en el siguiente caso hipotético, ambas particiones son exactamente iguales, pero las observaciones tienen asignaciones distintas:

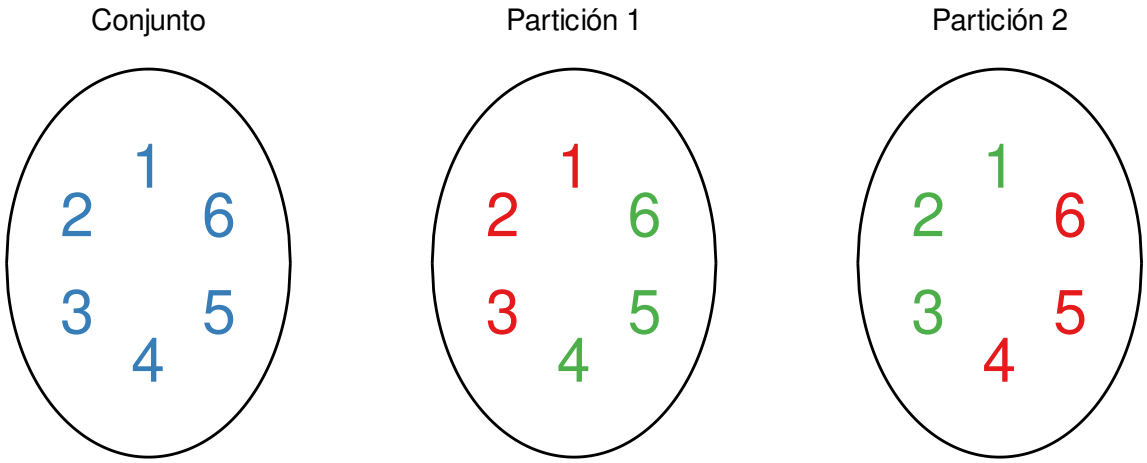


FIGURA 5.3: Ejemplo de concordancia perfecta entre dos particiones, a pesar de tener etiquetas distintas.

Se puede ver que las particiones 1 y 2 son iguales, ya que tienen a los elementos 1, 2 y 3 en el mismo grupo y el resto de los elementos en el mismo grupo también. Por lo tanto, cualquier medida de concordancia debe independizarse de los rótulos asignados a los grupos de la partición.

Con tal fin, la mayoría de los criterios externos se basan en evaluar los $n_P = \frac{n \cdot (n-1)}{2}$ **pares** de observaciones distintas para comparar dos particiones $\mathcal{P}_1$ y $\mathcal{P}_2$. Consideremos como ejemplo las

siguientes particiones, cuya concordancia es más difícil de medir:

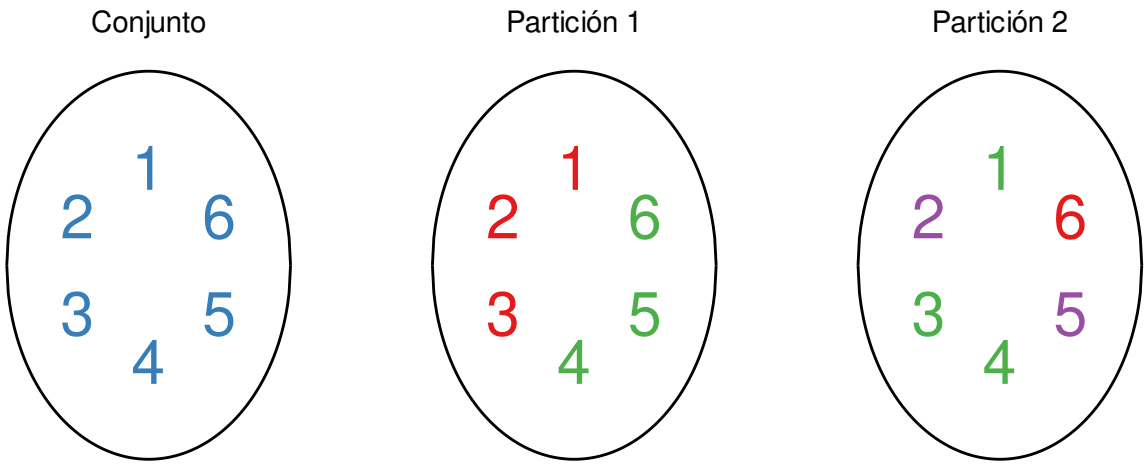


FIGURA 5.4: Ejemplo de dos particiones distintas, notar que las particiones tienen distinta cantidad de grupos.

Para cada par x_i y $x_{i'}$, podemos evaluar los siguientes indicadores:

- **Concordancia:**
  - Si las observaciones x_i y $x_{i'}$ están en el mismo grupo tanto en $\mathcal{P}_1$ como en $\mathcal{P}_2$, hay una concordancia. En este caso, se identifica al par como yy.
  - Si las observaciones x_i y $x_{i'}$ están en distintos grupos tanto en $\mathcal{P}_1$ como en $\mathcal{P}_2$, hay una concordancia. En este caso, se identifica al par como nn.
- **Discordancia:**
  - Si las observaciones x_i y $x_{i'}$ están en el mismo grupo tanto en $\mathcal{P}_1$, pero no en $\mathcal{P}_2$, hay una discordancia. En este caso, se identifica al par como yn.
  - Si las observaciones x_i y $x_{i'}$ están en distintos grupos en $\mathcal{P}_1$, pero no en $\mathcal{P}_2$, hay una discordancia. En este caso, se identifica al par como ny.

Con esta notación, evaluemos cómo se califican los $n_P = \frac{6\cdot(6-1)}{2} = 15$ pares para el ejemplo presentado en la Figura 5.4:

	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{1,6\}$	$\{2,3\}$	$\{2,4\}$	$\{2,5\}$
Partición 1	y	y	n	n	n	y	n	n
Partición 2	n	y	y	n	n	n	n	y

	$\{2,6\}$	$\{3,4\}$	$\{3,5\}$	$\{3,6\}$	$\{4,5\}$	$\{4,6\}$	$\{5,6\}$
Partición 1	n	n	n	n	y	y	y
Partición 2	n	y	n	n	n	n	n

(5.13)

Por lo tanto, totalizando los pares de cada clase en 5.13, obtenemos:

yy	yn	ny	nn
1	5	3	6

(5.14)

Esta notación puede resultar confusa, pero es la que se utiliza en la literatura correspondiente. De todas formas, podemos notar las siguientes características que pueden servir como guía para entender los resultados:

- La suma de todas las magnitudes da como resultado el número total de pares. Es decir, $n_P = \frac{n \cdot (n-1)}{2} = yy + yn + ny + nn$.
- Mientras mayores sean las concordancias (yy, nn) y menores las discordancias (yn, ny), mayor similitud habrá entre las particiones $\mathcal{P}_1$ y $\mathcal{P}_2$.

En base a estas medidas, se pueden calcular los siguientes índices:

- **Precisión (P):** $C_P = \frac{yy}{yy + ny}$
- **Recall (RC):** $C_{Re} = \frac{yy}{yy + yn}$
- **Rand (RN):** $C_{Ra} = \frac{yy + nn}{n_P}$
- **Czekanowski-Dice (CD):** $C_{CD} = \frac{2yy}{2yy + yn + ny}$
- **Folkes-Mallows (FM):** $C_{FM} = \frac{yy}{\sqrt{(yy + yn) \cdot (yy + ny)}}$
- **Jaccard (J):** $C_J = \frac{yy}{yy + yn + ny}$
- **Kulczynski (K):** $C_K = \frac{1}{2} \cdot \left(\frac{yy}{yy + ny} + \frac{yy}{yy + yn} \right)$
- **Rogers-Tanimoto (RGT):** $C_{RT} = \frac{yy + nn}{yy + nn + 2 \cdot (yn + ny)}$

Para todos estos índices, valores mayores corresponden a mejores concordancias entre particiones. Más aún, en el caso de concordancia perfecta, se tiene que tanto yn como ny son nulos, y todos esos índices tienen el número 1 como resultado máximo.

Para el ejemplo presentado en la Figura 5.4, podemos usar los datos de la Tabla 5.14 para obtener los siguientes índices:

P	RC	RN	CD	FM	J	K	RGT
$\frac{1}{4} = 0.25$	$\frac{1}{6} \approx 0.1666$	$\frac{7}{15} \approx 0.4666$	$\frac{2}{10} = 0.2$	$\frac{1}{\sqrt{24}} \approx 0.204$	$\frac{1}{9} = 0.111$	$\frac{5}{24} = 0.208$	$\frac{7}{23} = 0.304$

(5.15)

Notemos que estos índices están muy lejos de 1, lo que denota que ambas particiones son muy disímiles. Esto era esperable por la cantidad reducida de elementos, sumado a que ambas particiones tenían una distinta cantidad de grupos, lo que acentúa mucho la influencia de las discordancias.

Por otro lado, se puede probar que en el caso de concordancia perfecta presentado en 5.3, todos estos índices dan un valor de 1, ya que $ny = yn = 0$.

Estos criterios son de mucha utilidad cuando se tiene una partición de referencia y se quiere evaluar en qué medida las particiones provistas por un algoritmo coinciden con la de referencia. En nuestro caso, las utilizaremos tanto en las simulaciones como en bases controladas, ya que en ese caso se quiere evaluar cuáles de los métodos automáticos tiene una mayor concordancia entre los grupos resultantes y los de referencia.

5.6. Algoritmo

A continuación, describimos en pseudo-código la estructura del algoritmo propuesto para lograr un agrupamiento automático de individuos según las variaciones en sus respuestas.

- **Datos de entrada:**
  - Base de datos
  - Nombre de la variable de respuesta.
  - Nombre de la variable temporal.
  - Nombre de la variable identificatoria de individuos.
  - Cantidad de pendientes deseadas J (la cantidad de respuestas y tiempos debe ser $J+1$).
- **Filtrado de datos faltantes:**
  - Se recorren todos los individuos de la base.
  - Se identifican los individuos con $J+1$ respuestas y tiempos de medición como los que tienen trayectorias completas.
- **Vectores de pendientes:**
  - Para cada individuo de trayectoria completa, se calcula el vector de J pendientes asociadas (Ver Ecuación 5.4).
- **Algoritmo de clustering:**
  - Se aplica el algoritmo de clustering deseado sobre los vectores de pendientes.
  - Para cada individuo, el algoritmo de clustering devuelve un rótulo para asignarle un grupo.
- **Datos de salida:**
  - Una tabla (puede ser una matriz, una base de datos o una lista) donde para cada individuo se pueda obtener su grupo correspondiente.

Otra forma de ver este agrupamiento es que al ser similares las variaciones, la morfología de las trayectorias de respuesta tienen la misma estructura. Además, como la mayoría de los algoritmos de clustering se basan en distancias entre vectores, cualquier algoritmo de clustering puede servir para capturar similitudes entre las variaciones de las respuestas.

Sobre la partición resultante se puede realizar un análisis entre grupos con el objetivo de visualizar diferencias en otras variables de la base de datos. Cuando las bases de datos tienen una alta dimensionalidad (es decir, una gran cantidad de variables) las variables con altas diferencias entre grupos puede sugerir una potencial asociación que debe ser investigada con mayor detalle y mejores condiciones.

5.7. Simulaciones

Para testear los rendimientos de los distintos algoritmos de clustering aplicados al espacio de las pendientes, se repitieron varias iteraciones de cada algoritmo a algunas bases longitudinales. Recordando que el objetivo principal de esta sección es aplicar estos algoritmos a la base de expresión genética presentada en la sección 2.2.4, se busca además que las bases utilizadas posean características similares a las de la base mencionada: pocos instantes de medición y la posibilidad de observar cambios abruptos en la variable de respuesta.

5.7.1. Bases simuladas

Una de las instancias de evaluación se realiza sobre bases simuladas. En estas bases hay algunos atributos que permiten definirlas: los instantes de medición, los grupos de individuos y las variables de respuesta.

Simulación de los tiempos

Como se mencionó previamente, una de las características frecuentes de los datos longitudinales es que los instantes de medición no siempre coinciden para distintos individuos. Por lo tanto, para cualquier simulación, debe considerarse que los tiempos puedan ser distintos.

En esta sección consideramos 3 metodologías distintas para generar los tiempos de medición $t_{i,j}$ para un individuo i. Sin embargo, en todos los casos, se considera el tiempo inicial $t_{i,0} = 0$ y se busca generar una variable no negativa $\tau_{i,j}$ que represente el tiempo entre mediciones consecutivas. Es decir, los incrementos cumplen $\tau_{i,j} = t_{i,j} - t_{i,j-1}$, luego los tiempos $t_{i,j}$ se obtienen sumando estos incrementos. Para generar los incrementos, se utilizaron las distribuciones de distintas variables aleatorias:

- **Uniformes:** $\tau_{i,j} \sim \mathcal{U}(1 - dt, 1 + dt)$
- **Normales:** $\tau_{i,j} \sim \mathcal{N}(1, dt)$
- **Exponenciales:** $\tau_{i,j} = 1 - \sqrt{dt} + L_{i,j}, L_{i,j} \sim \mathcal{E}(\frac{1}{\sqrt{dt}})$

En estas definiciones, se elige un valor de dt que logre que las variables sean positivas en su totalidad o, en su defecto, que las probabilidades de que resulten negativas sea despreciable. Por ejemplo, en el caso de la distribución normal debe asegurarse que $1 - 3 \cdot dt$ sea positivo para que la probabilidad de obtener un resultado negativo sea ínfima. Más aún, se puede reducir aún más esa probabilidad asegurando un menor valor de dt. Además, se intenta que los incrementos tengan una media de valor 1, aunque no coincidan sus desvíos.

Nuevamente, en todos los casos se obtienen los instantes de medición sumando los incrementos, es decir, $t_{i,j} = \sum_{k=1}^{j} \tau_{i,k}$.

En la figura 5.5 se observan algunas simulaciones de los distintos mecanismos generadores de los instantes de medición. En base a esta figura, se puede notar que los instantes de medición pueden ser distintos para individuos diferentes, pero presentan una menor variabilidad en el caso uniforme, mientras que el generador exponencial es el que mayores discrepancias exhibe.

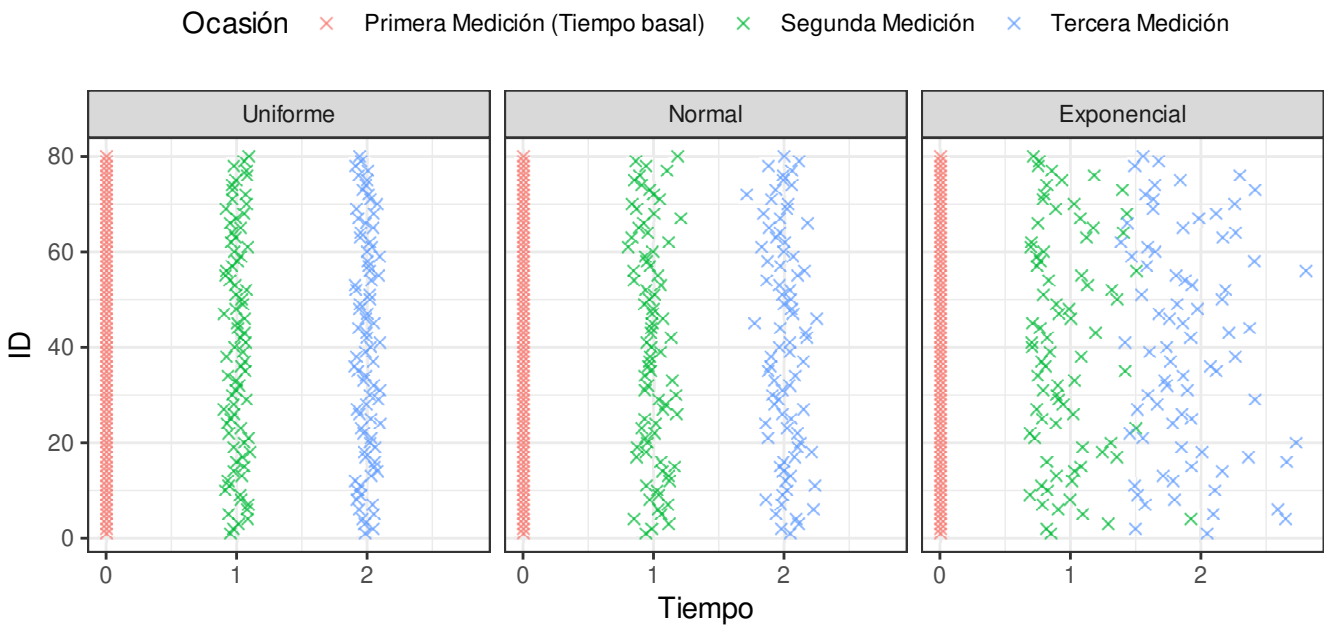


FIGURA 5.5: Instantes de medición simulados obtenidos con distintos mecanismos, fijando el valor de dt en 0.1.

Simulación de las pendientes

Recordar que el algoritmo propuesto busca agrupar individuos según similitudes en los cambios abruptos en la respuesta, que se ven reflejados en el valor de las pendientes con un valor absoluto elevado. Por lo tanto, para generar dichas trayectorias nos basamos en los valores de las pendientes.

Para lograr este objetivo la simulación utiliza un parámetro positivo D que representa el valor absoluto de una pendiente. La idea es que cada grupo tiene un punto focal o centro μ_k a partir del cual se generan los vectores de pendientes para distintos individuos en la cercanía de dicho centro.

Esta cercanía en el espacio de las pendientes aseguran similitudes morfológicas en las trayectorias.

Por otro lado, se busca que las trayectorias sean relativamente estables salvo en algún instante de tiempo en el que se genera el cambio en una respuesta. Además, podríamos asumir que el cambio en la respuesta puede ocurrir en cualquier instante de tiempo, por lo tanto, la cantidad de centros también dependerá de la cantidad de pendientes J, obteniendo $K = 2 \cdot J - 1$ centros. Por ejemplo, para $J = 3$, los $K = 2 \cdot 3 - 1 = 5$ centros se construyen de la siguiente manera:

Centro	Valores
μ_1	$(D, -D, 0)$
μ_2	$(-D, D, 0)$
μ_3	$(0, D, -D)$
μ_4	$(0, -D, D)$
μ_5	$(0, 0, 0)$

(5.16)

El número de clusters se determina del siguiente modo: cada $J - 1$ pares de mediciones consecutivas tiene asociadas dos configuraciones de pendientes, con un incremento de la respuesta en la primer medición y posterior decrecimiento, mientras que se agrega para ese mismo par de respuestas consecutivas un decrecimiento y posterior crecimiento. Estos pares dan un total de $2 \cdot (J - 1)$ posibles configuraciones, a las que además, se les agrega un vector nulo de pendientes para trayectorias que mantienen una estabilidad temporal en las respuestas, dando un resultado total de $2 \cdot (J - 1) + 1 = 2 \cdot J - 1$ configuraciones posibles.

Alrededor de cada uno de estos centros se generan $\frac{I}{K}$ (nos aseguramos que sean enteros) vectores de pendientes denotados $\mathbf{m}_i = (m_{i,1}, m_{i,2}, \cdots, m_{i,J})$, con $1 \leq i \leq I$, generando los K grupos y un número I individuos totales correspondientes a cada vector de observaciones. Para generar los vectores de pendientes se utiliza una distribución normal multivariada con vector de medias dado por cada centro μ_k y matriz de covarianza diagonal con autovalor único (es decir, $\Sigma_i = \sigma_D^2 \cdot \mathbf{I}_J$). Además, el valor de σ_D se elige relativo al parámetro D a través de un coeficiente c_V, es decir, $\sigma_D = c_V \cdot D$. Generalmente c_V se elige como un número menor que 1 para evitar que se solapen demasiado las observaciones correspondientes a distintos grupos, ya que eso dificulta naturalmente la clasificación de cada observación.

En la figura 5.6 se ve cómo influyen en las simulaciones de los vectores de pendientes los distintos parámetros, manteniendo $J = 2$ para que sea gráficamente accesible. A medida que D aumenta, más alejados están los $K = 2 \cdot 2 - 1 = 3$ centros entre sí, mientras que a medida que aumenta el valor de c_V, más dispersas son las combinaciones de pendientes alrededor de estos centros y se nota que empiezan a solaparse.

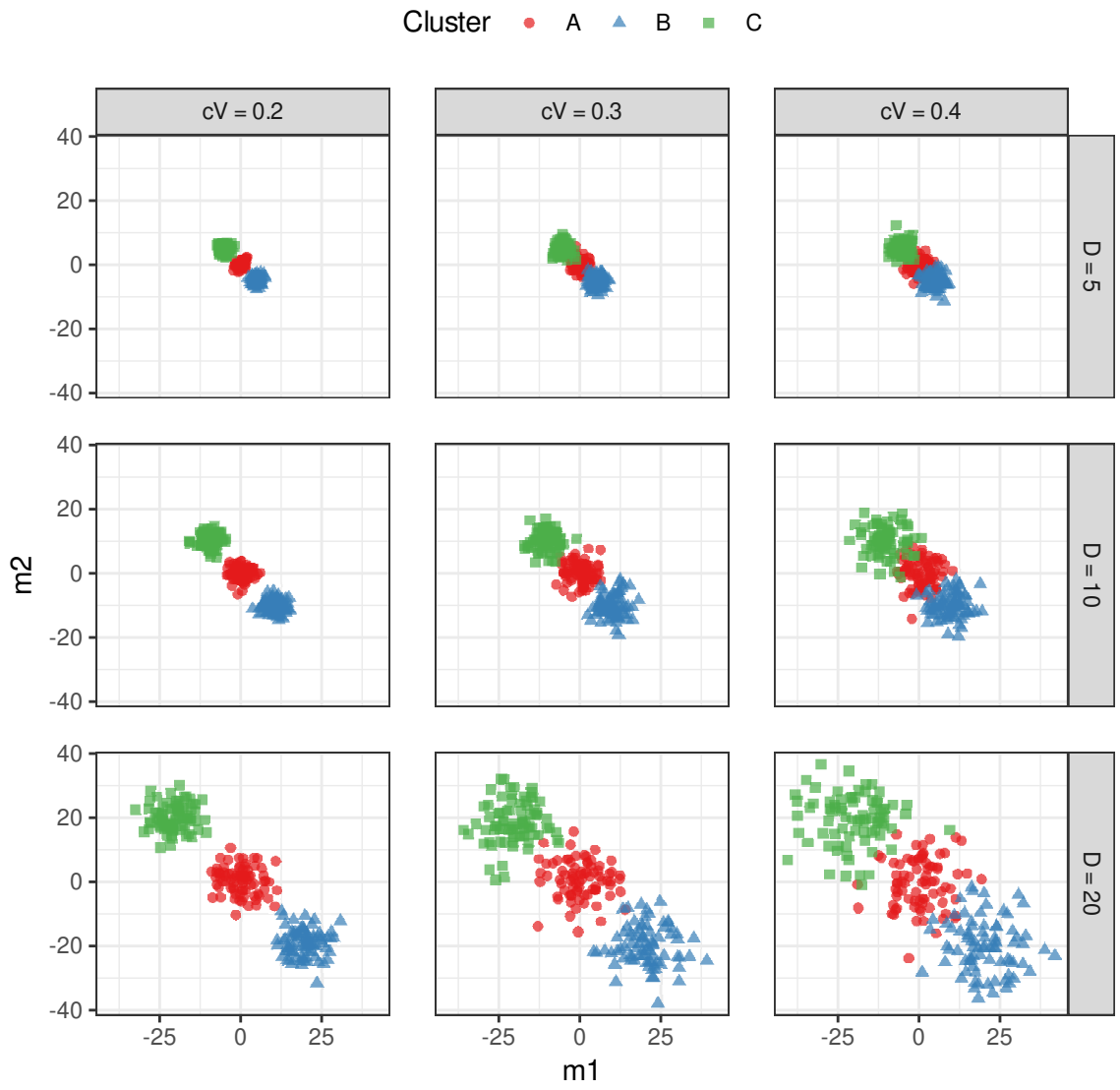


FIGURA 5.6: Combinaciones generadas de pendientes con $J = 2$ y valores variables de los parámetros D y c_V.

Simulación de las respuestas

Una vez obtenidos estos vectores de pendientes $\mathbf{m}_i$, se pueden construir las trayectorias de respuestas $\mathbf{Y}_i = (Y_{i,0}, Y_{i,1}, \cdots, Y_{i,J})$ de la siguiente manera:

$$Y_{i,j} = \begin{cases} b_{0,i} & \text{si } j = 0 \\ Y_{i,j-1} + m_{i,j} \cdot (t_{i,j} - t_{i,j-1}) + \varepsilon_{i,j} & \text{si } 1 \leq j \leq J \end{cases} \tag{5.17}$$

donde $b_{i,0} \sim \mathcal{N}(\mu_B, s_B^2)$ es una variable que genera variabilidad intersujeto en los niveles basales y $\varepsilon_{i,j} \sim \mathcal{N}(0, s_R)$ representa una perturbación en cada tiempo al valor obtenido a partir de las pendientes.

En la Figura 5.7 se ve el impacto del parámetro D sobre las trayectorias resultantes. Como era esperado, al aumentar D, mayor valor absoluto tienen las pendientes y se visualiza en las morfologías a través de cambios más abruptos en la respuesta.

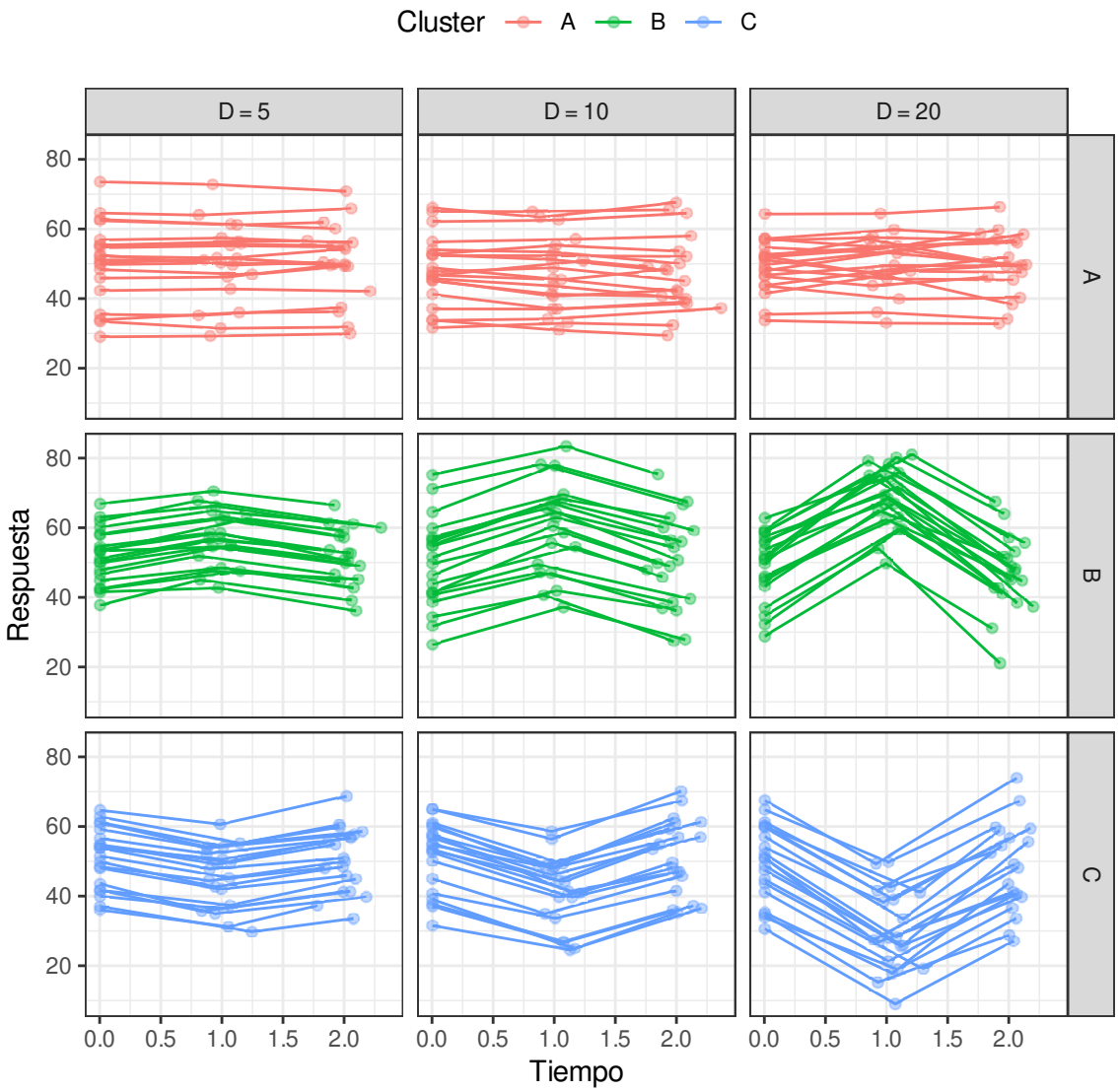


FIGURA 5.7: Trayectorias de respuesta simuladas con $J = 2$, $c_V = 0.2$, $s_R = 1$, $\mu_B = 50$, $s_B = 10$ y valores variables del parámetro D.

En la Figura 5.8 se observa la influencia del parámetro c_V sobre las trayectorias. A medida que aumenta c_V, las trayectorias dejan de tener una morfología uniforme para todos los individuos del grupo. Por ejemplo, cuando $c_V = 0.4$ el cluster A (que representa las trayectorias "estables") exhiben trayectorias con incrementos y decrecimientos más pronunciados de la respuesta, y podrían ser visualmente categorizados como trayectorias de los clusters B o C. Esto se debe al aumento de la dispersión, que aumenta la probabilidad de que un vector de pendientes que debería estar cerca de un centro μ_k, resulte cercano a otro de los centros predeterminados.

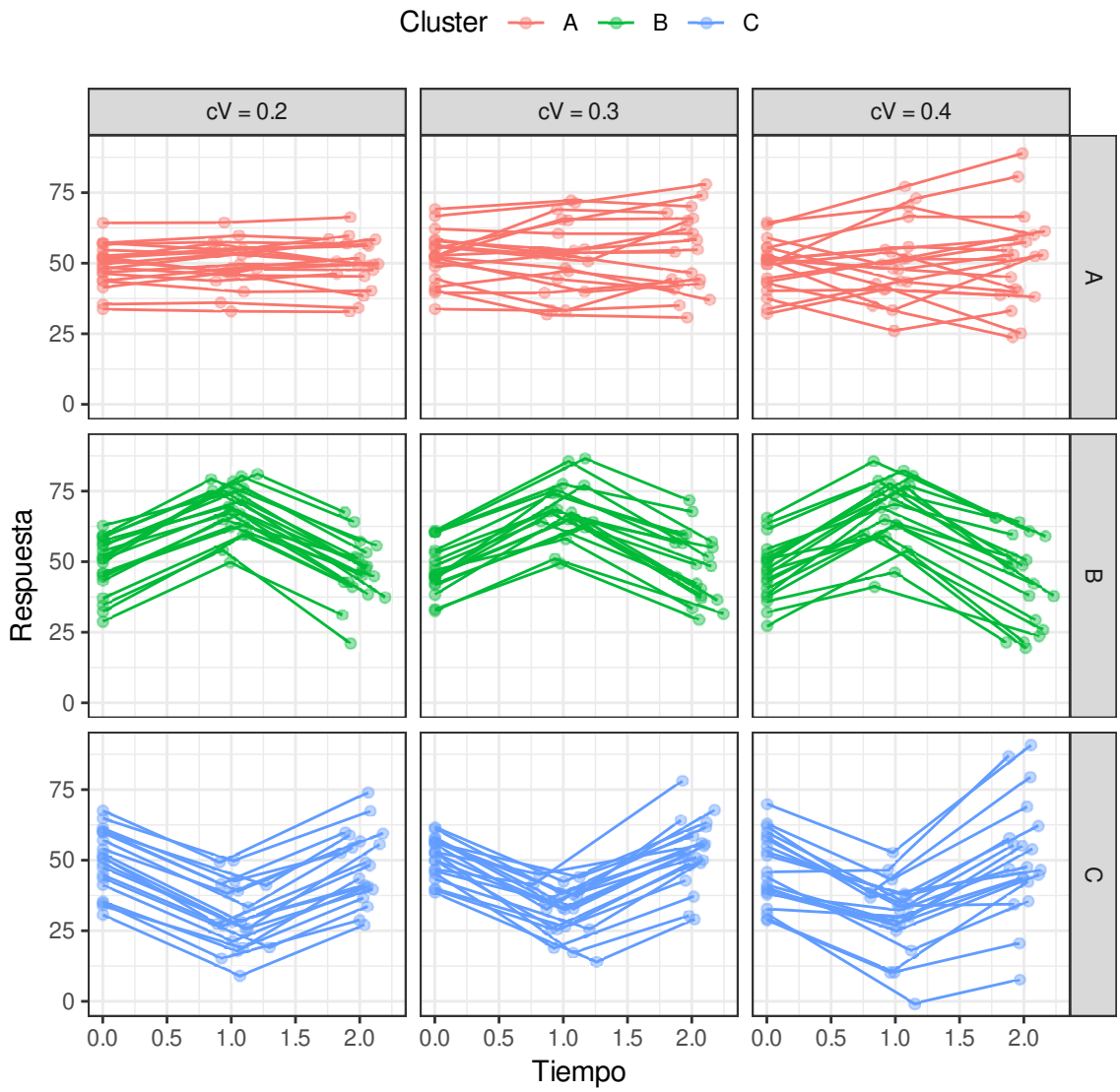


FIGURA 5.8: Trayectorias de respuesta simuladas con $J = 2$, $D = 20$, $s_R = 1$, $\mu_B = 50$, $s_B = 10$ y valores variables del parámetro c_V.

Parámetros de simulación

En la Tabla 5.1 se disponen los valores de los parámetros necesarios para generar las bases de datos simuladas. Los valores se obtuvieron de forma empírica para obtener como resultado características similares a las de las bases de datos longitudinales que se observan en la práctica. Los valores de c_V intentan que los grupos estén bien separados para algunos valores (por ejemplo, $c_V = 0.2$) y cuando c_V equivale a 0.5 los grupos empiezan a solaparse en el espacio de las pendientes, dificultando la clasificación y que permite evaluar el rendimiento y la robustez de los distintos algoritmos.

$\frac{I}{K}$	c_V	D	s_R	J	μ_B	s_B	dt
40	0.2	5	1	2			
80	0.3	10	3	3	50	10	0.1
120	0.4	20	5	4			
	0.5						

TABLA 5.1: Valores utilizados de los parámetros que generan la simulación.

Además, para estos valores de parámetros, evaluamos el rendimiento de los algoritmos de K-Medias, K-Medoides (con distancia Euclidea y Manhattan) y las distintas variantes de clustering jerárquico (completo, promedio e individual). Los otros métodos presentados fueron excluidos del análisis de bases simuladas por los siguientes motivos:

- El método de modelos mixtos de clases latentes requiere un modelo previo que se pueda adaptar a las tendencias temporales. Sin embargo, como los cambios en la respuesta son abruptos, los modelos polinomiales se ven muy afectados por estos cambios. Por otro lado, para disponer de un modelo lineal a trozos, se requiere conocer el tiempo en el que se introduce el cambio. Sin embargo, en las trayectorias dispuestas, los cambios pueden ocurrir en cualquier momento y por lo tanto, los modelos resultan demasiado complejos y además, en la práctica no se plantearía un modelo que admita cambios de tendencia en *todos* los tiempos.

- En cuanto al algoritmo de K-Medias basado en Kernels, es muy sensible a la elección del parámetro σ. En nuestras simulaciones esta sensibilidad devino en problemas de convergencia en casi todas las corridas. Este problema se dio aún apelando a una opción del algoritmo en el que se elige automáticamente el parámetro en base a los datos mediante una heurística. Por lo tanto, las iteraciones en varios casos ni siquiera arrojaban resultados, impidiendo cualquier análisis adecuado.

5.7.2. Datos TLC

Como base controlada, se toma la base TLC descripta en la sección 2.2.2. Como se ve en la figura 2.2, la diferencia en las trayectorias de respuesta de cada grupo se observa justamente en la variación inicial. Aquellos individuos que pertenecen al grupo placebo mantienen los niveles de plomo relativamente estables durante todo el estudio y aquellos individuos que recibieron tratamiento tienen un fuerte decrecimiento en la primera semana y un posterior incremento leve de los niveles en sangre. Por lo tanto, en estos casos puede ser útil el algoritmo propuesto ya que se basa en las pendientes de las trayectorias de respuesta. En cuanto a las variantes en el algoritmo de clustering, se utilizan todas las presentadas en este capítulo. Vale aclarar que para el algoritmo de modelos mixtos de clases latentes, se dispuso el modelo descripto en 3.33, ya que en este caso, mediante el análisis gráfico de las respuestas, se puede determinar que sólo luego de la primer semana las respuestas cambian abruptamente.

5.8. Resultados

Los experimentos y las simulaciones descriptas en las secciones anteriores dieron lugar a los siguientes resultados. Vale aclarar que en algunos casos, los errores estándard arrojaron valores ínfimos respecto del tamaño de los puntos y por lo tanto, no siempre son discernibles en los gráficos.

5.8.1. Bases simuladas

En las bases simuladas, se generaron $M = 100$ bases para cada combinación de los parámetros descriptos en la Tabla 5.1. En cada caso, se conoce previamente el grupo al cual pertenece cada observación y por lo tanto, se pueden comparar las particiones resultantes de cada algoritmo de clustering con los grupos de referencia utilizando los criterios externos detallados en la sección 5.5.2.

Además. salvo que se indique lo contrario, no se aplican ninguna de las transformaciones mencionadas en la sección 5.4, ya que no se observaron grandes diferencias en las distintas instancias. Por otro lado, como se explicó en la sección 5.7.1, para las bases simuladas se excluyeron las metodologías de K-medias basado en kernels (Sección 5.2.3) y el MEM de clases latentes (Sección 5.2.5) por distintos motivos.

Para el algoritmo de K-Medias basado en kernels, la selección del parámetro σ tiene mucha influencia sobre los resultados. Sin embargo, un valor óptimo de σ para una base de datos, puede arrojar resultados muy diferentes en otra. Por lo tanto, al correr nuestras simulaciones, no encontramos un valor fijo de σ que devenga en resultados apropiados para *todas* las bases simuladas, ya que siempre ocurría que en alguna iteración el algoritmo no lograba converger. Además, utilizamos una opción del algoritmo perteneciente al paquete `kernlab` del software `R` que selecciona mediante una heurística el parámetro de forma automática para la base provista como entrada. Aún con esa selección automática, los mismos problemas de convergencia se sucedieron y por lo tanto, no pudimos incluirlos en la simulación.

Respecto al MEM de clases latentes, la complejidad en la aplicación a bases simuladas reside en la necesidad de predeterminar el modelo. Dado que los cambios en la trayectoria de respuesta son abruptos, los modelos mixtos deberían considerar modelos polinomiales a trozos. Sin embargo, como los cambios abruptos pueden suceder en cualquier instante de tiempo, los modelos deberían tener nudos en todas las ocasiones de medición, lo que complejiza enormememente el modelo y puede resultar en problemas de convergencia de los métodos iterativos. Además, ante cualquier base provista como entrada, si el tiempo nodal no es conocido, no sería aconsejable considerar modelos que contemplen nudos para todos los instantes ya que aumenta considerablemente la cantidad de parámetros involucrados.

Un primer acercamiento a los resultados se ven en la Figura 5.9:

Se ve que los métodos jerárquicos presentan un menor rendimiento en casi todos los índices, mientras que el algoritmo de K-medias y el de K-medoides (utilizando tanto la distancia Euclidea como la distancia Manhattan) alcanzan valores similares en casi todos los índices. Vale aclarar que

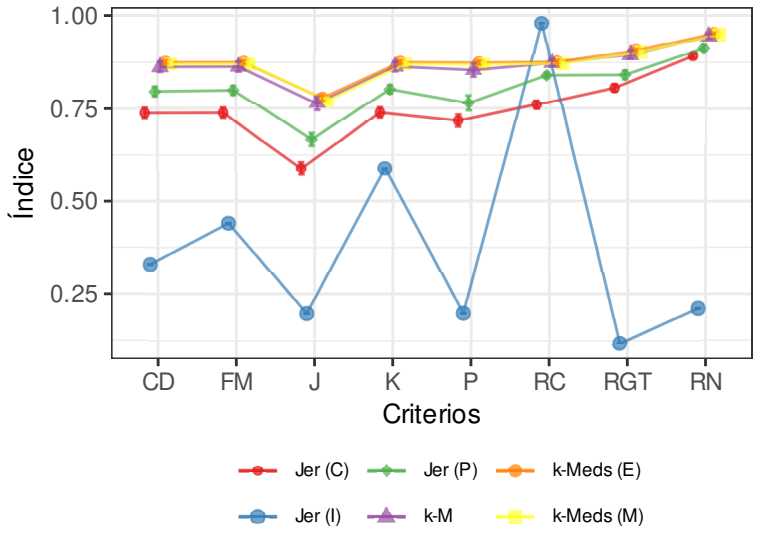


FIGURA 5.9: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $c_V = 0.2$, $D = 10$, $J = 3$ y $s_R = 3$, a las que se les aplican distintos algoritmos de clustering.

el único criterio en el que el clustering jerárquico sobrepasa el rendimiento de los otros métodos es el Recall. Esto tiene una explicación que veremos más adelante.

La Figura 5.10 muestra cómo varían los índices ante el incremento del parámetro de escala c_V.

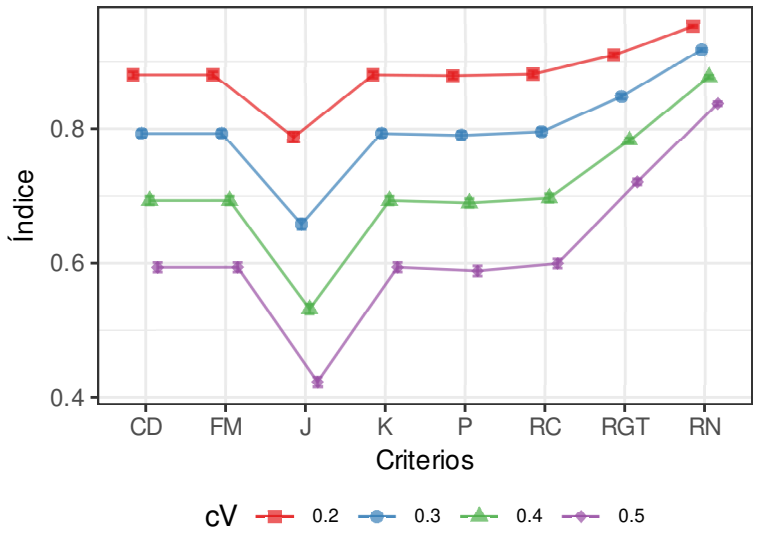


FIGURA 5.10: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $D = 10$, $J = 3$ y $s_R = 3$, utilizando el algoritmo de K-medoides con la distancia Manhattan.

Los resultados son los esperados. A medida que aumenta c_V, aumenta también el solapamiento

entre los grupos y por lo tanto, se vuelve aún más difícil para los algoritmos poder discernir los individuos que pertenecen a cada grupo. Este solapamiento deviene en menores índices de concordancia entre particiones.

Otros resultados esperados es que no se observaron diferencias sustanciales entre los rendimientos de los algoritmos en las bases cuyos instantes de medición están generados por distintos mecanismos. Del mismo modo, son similares los índices para distintas cantidades de individuos I, salvo por leves diferencias en los errores estándard. Por lo tanto, estas comparaciones no ofrecen ningún resultado de interés y las figuras correspondientes son omitidas en este trabajo.

Respecto al parámetro de distancia D y los desvíos de los errores de medición s_R, la Figura 5.11 muestra que, según lo esperado, aumentan los índices cuando aumenta D y disminuyen cuando aumenta s_R.

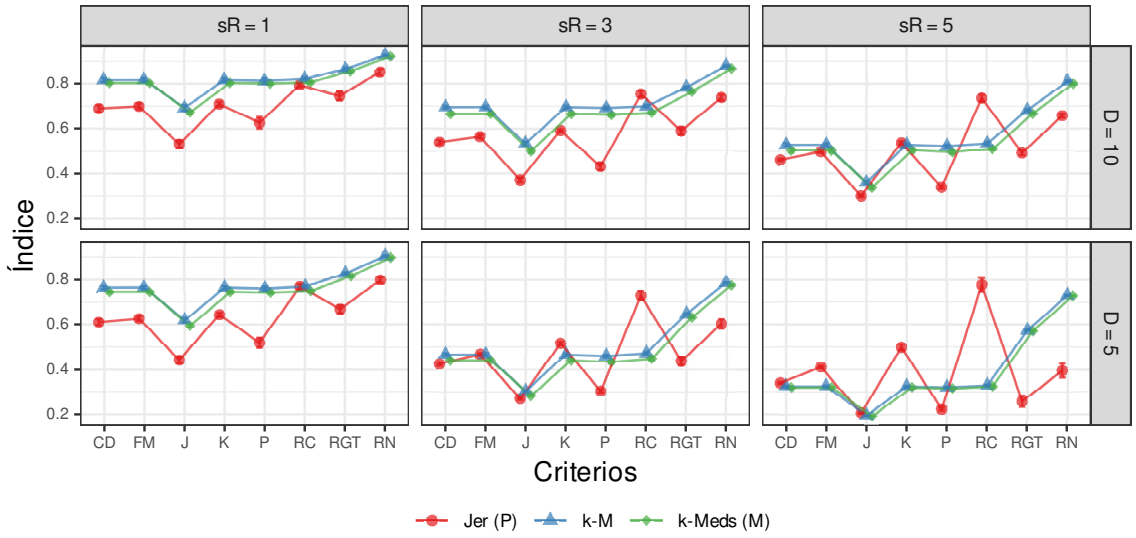


FIGURA 5.11: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $c_V = 0.4$, y $J = 3$, a las que se les aplican distintos algoritmos de clustering.

Vale notar que cuando $D = s_R = 5$, el clustering jerárquico presenta mejores índices que otros métodos para algunos criterios (el Recall y el índice de Kulczynski mayoritariamente), por lo que este método parece una mejor opción que los restantes. Sin embargo, en este caso es cuando es más natural que los algoritmos tengan problemas de clasificación ya que la distancia en el espacio de las pendientes entre los centros grupales y el vector nulo son mínimas y los errores de medición tienen más desvío. Por lo tanto, son las condiciones de mayor confusión entre grupos ya que individuos pertenecientes a distintos grupos pueden exhibir trayectorias muy similares.

En todos los casos anteriores, no se observan diferencias entre los rendimientos de los algoritmos de K-medias y de K-medoides. De todas formas, el resultado más sorprendente se da cuando $D = 20$. En este escenario, las distancias entre los centros en el espacio de las pendientes es mayor y por lo tanto, los grupos están más separados. Más aún, cuando c_V y s_R adquieren su mínimo valor, estos grupos son aún más homogéneos y casi no hay confusión gráfica entre

trayectorias de individuos pertenecientes a distintos grupos. Por lo tanto, cualquier algoritmo de clustering en estas circunstancias debería exhibir los mejores rendimientos. Sin embargo, en esta configuración, vemos en la Figura 5.12 que el algoritmo de k-Medias tiene menor rendimiento en todos los índices con un elevado error estándard, mientras que el algoritmo de k-Medoides casi no presenta errores de clasificación. Esto se debe a la sensibilidad del algoritmo de K-medias a la inicialización y outliers previamente mencionada, dando muestras de un algoritmo que en algunas circunstancias puede perder robustez. Las diferencias desaparecen a medida que aumentan c_V y s_R, pero mayoritariamente por un descenso más pronunciado en los índices de K-medoides.

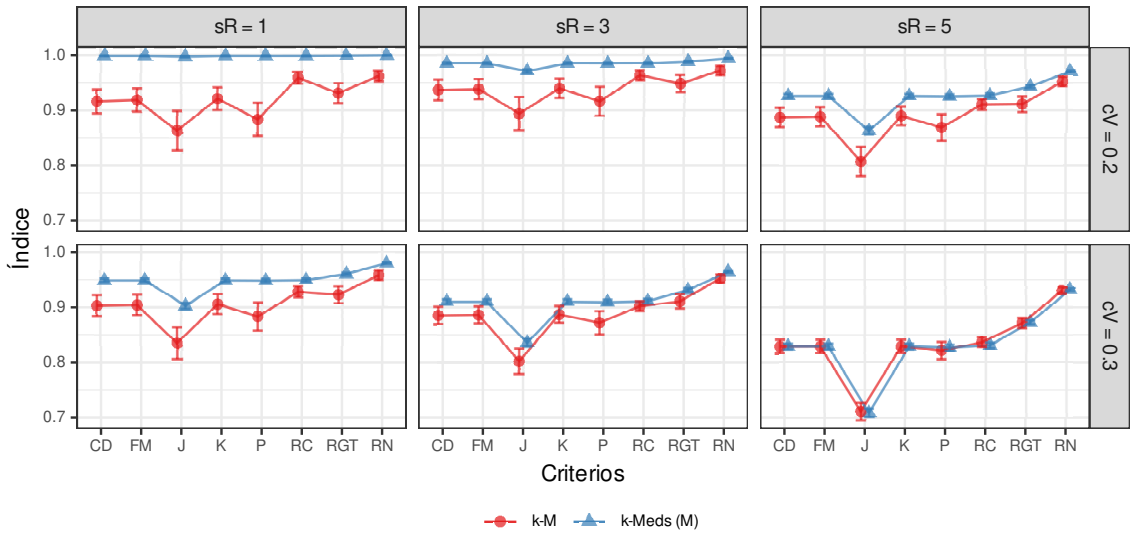


FIGURA 5.12: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $D = 20$ y $J = 3$, a las que se les aplican distintos algoritmos de clustering.

Por otro lado, la Figura 5.13 muestra cómo se ven afectados los índices por la cantidad de mediciones en cada individuo.

Por ejemplo, se ve cómo el algoritmo de K-medias tiene un rendimiento levemente superior a ambas variantes de K-Medoides cuando $J = 2$. Sin embargo, se ve que los índices de K-medias decrecen más rápidamente que los de K-medoides a medida que aumenta J, además de que presentan mayores errores estándard.

Por último, la Figura 5.14 muestra un detalle interesante respecto de las transformaciones por escala o normalización presentadas en la sección 5.4. Como fue mencionado en dicha sección, el hecho de trabajar sobre el espacio de las pendientes permite dos opciones a la hora de procesar las variables. Se pueden aplicar transformaciones previa o posteriormente al cálculo de las pendientes. En esta figura se ve que el algoritmo de K-medias presenta mejores índices cuando las transformaciones son sobre las pendientes $\mathbf{m}_i$, que cuando son sobre la variable de respuesta $\mathbf{Y}_i$. Por otro lado, el algoritmo de K-medoides parece verse perjudicado en estas circunstancias. Esto puede deberse a que las transformaciones en el espacio de las pendientes reducen la variabilidad de las mismas y por lo tanto, reduce el impacto de outliers sobre el algoritmo de K-medias. Respecto

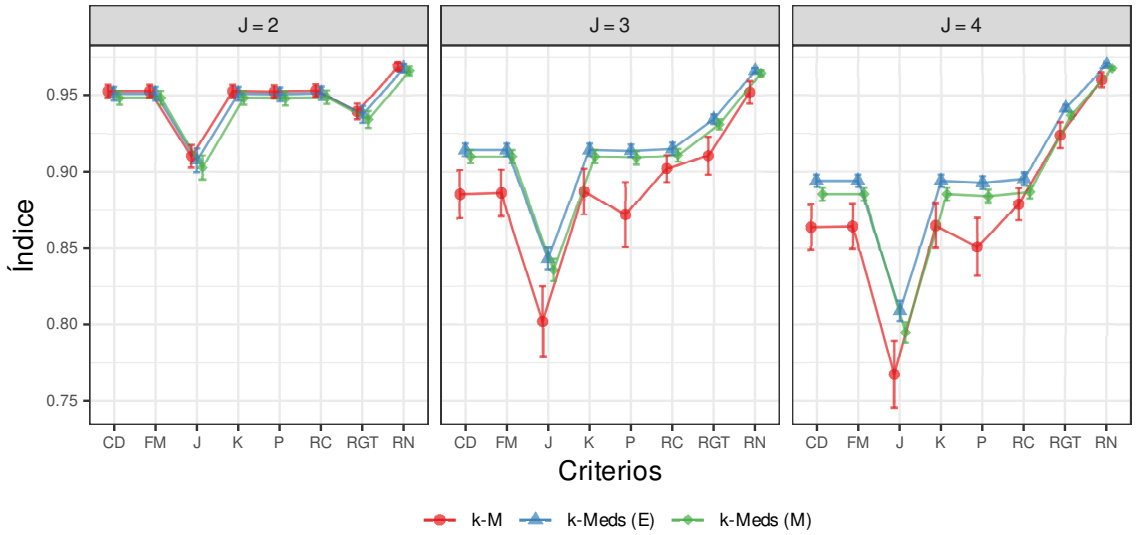


FIGURA 5.13: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $D = 20$, $c_V = 0.3$ y $s_R = 3$, a las que se les aplican distintos algoritmos de clustering.

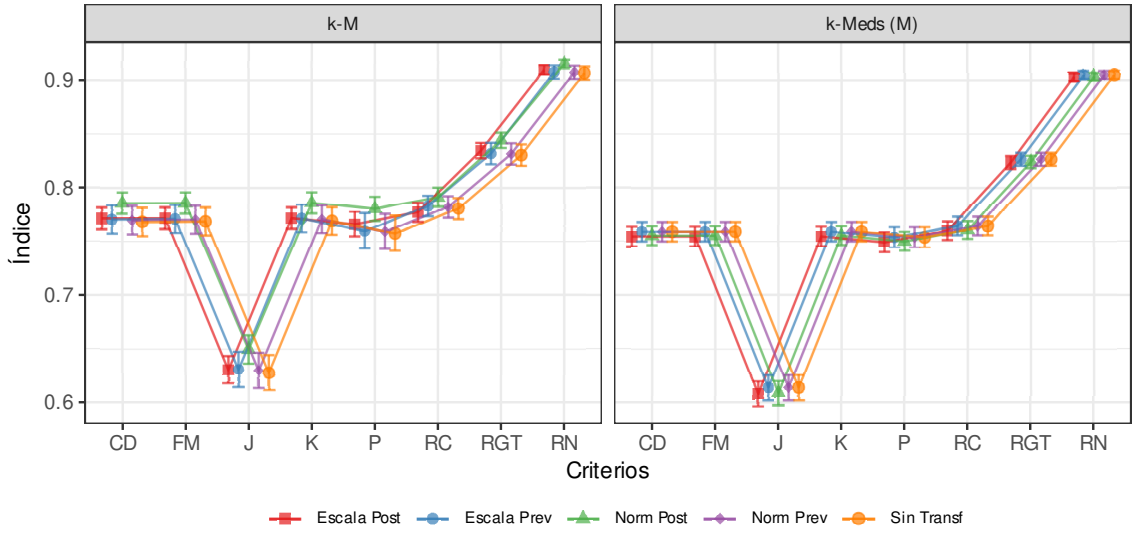


FIGURA 5.14: Valores medios de los índices de distintos criterios para M=100 bases de datos simuladas según la sección 5.7 tomando $D = 10$, $c_V = 0.3$, $s_R = 3$ y $J = 3$, a las que se les aplican distintos algoritmos de clustering.

a K-medoides, puede darse el efecto contrario, es decir, que al reducir la variabilidad, haya menor distancia entre los grupos resultantes y mayor confusión en la clasificación.

5.8.2. Datos TLC

En la Figura 5.15 se ven en qué medida los distintos algoritmos de clustering (salvo los clustering jerárquicos que se disponen en la Figura 5.16) logran identificar aquellos individuos pertenecientes tanto al grupo Placebo como Tratamiento en la base TLC. Se refuerza en este caso lo observado en las bases simuladas, los algoritmos de K-medias y K-medoides presentan similares resultados, con una leve ventaja de los algoritmos de K-medoides.

Por otro lado, se ve que los algoritmos de K-medias basado en Kernels muestra índices más bajos que éstos y con mayor error estándard. Esto se debe a que distintas aplicaciones del algoritmo da resultados muy diversos, ya que es sensible no sólo a outliers y la inicialización, sino que a la selección del parámetro σ. Además, el MEM de clases latentes arroja los menores índices.

Vale aclarar además que el desvío estándard en todos los métodos salvo el de K-Medias basado en Kernels es 0, por lo que en todos los algoritmos se obtuvo el mismo índice en las $M = 100$ repeticiones para cada criterio.

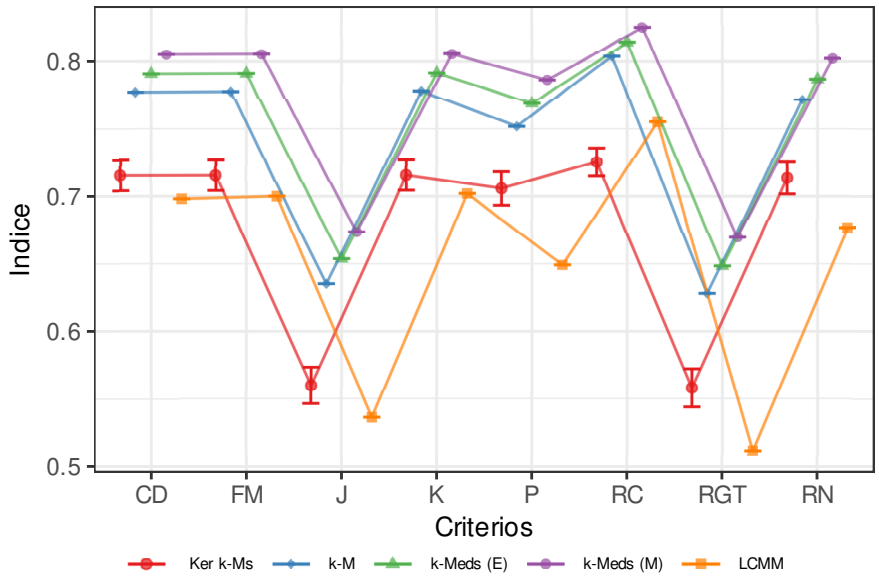


FIGURA 5.15: Valores medios de los índices de distintos criterios para M=100 iteraciones de distintos algoritmos de clustering morfológico en la base TLC.

Además, en la figura 5.16 se visualizan los rendimientos de los clustering jerárquicos.

Se puede observar en los clustering jerárquicos que en general presentan menores índices que K-medoides y K-medias en todos los criterios, salvo por el Recall. Es decir, si nos guiáramos únicamente por este índice, se eligiría un clustering jerárquico para agrupar los datos. Sin embargo, evaluando la partición resultante de aplicar el clustering jerárquico individual en el espacio de la base TLC (Figura 5.16), se ve el motivo por el cual este criterio puede ser engañoso:

En la Figura 5.17 queda claro porqué el Recall tiene valores tan altos. Dado que se calcula como $C_{Re} = \frac{yy}{yy + yn}$ (según la notación establecida en la Sección 5.5.2), resulta que en la partición obtenida por el clustering jerárquico, hay muy pocos pares de observaciones que no están agrupadas en el mismo cluster, dado que hay un cluster de un único individuo (del grupo tratamiento, no

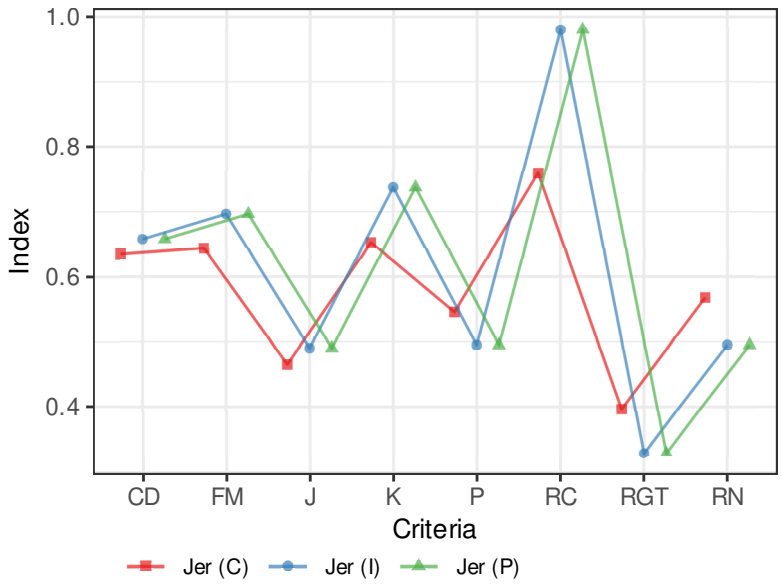


FIGURA 5.16: Valores medios de los índices de distintos criterios para M=100 iteraciones de distintos algoritmos de clustering jerárquico en la base TLC.

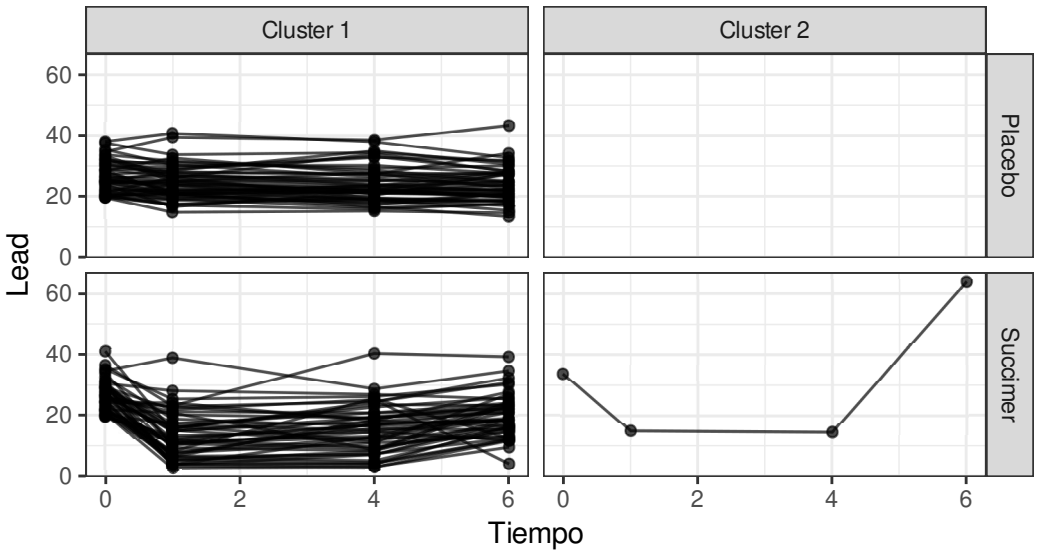


FIGURA 5.17: Trayectorias de respuesta comparando los grupos originales y los obtenidos por el clustering jerárquico individual. Notar que no hay ningún individuo del grupo Placebo agrupado en el Cluster 2, por eso el panel superior derecho no tiene observaciones.

hay ninguno del grupo placebo y por eso el panel correspondiente está vacío) y un grupo con todos los individuos restantes. Por lo tanto, el conjunto de los pares que no están agrupados en el mismo cluster son sólo aquellos que involucran al individuo del cluster que lo mantiene como único integrante. Esto explica los valores altos de los índices de Recall y Kulczynski en la figura 5.11.

Este fenómeno hace notar la importancia de observar varios índices y no dejarse llevar por los resultados de un único criterio, ya que pueden obtenerse conclusiones erróneas sobre cuál es la

mejor metodología para aplicar en una base dada.

5.8.3. Datos DM

En la base DM descripta en la Sección 2.2.4 se aplicaron los distintos algoritmos de clustering en el espacio de las pendientes correspondientes a las trayectorias de respuesta de expresión genética de IL-1β. Al filtrar los individuos con trayectorias de respuesta incompletas, quedaron 26 individuos remanentes.

En las figuras 5.18 y 5.19 se observan los resultados de los algoritmos de K-medias y K-medias basado en Kernels, respectivamente. Las particiones resultantes muestran que los distintos grupos no tienen una morfología determinada y se observan mezclas entre trayectorias estables y aquellas que presentan cambios abruptos en la respuesta. Nuevamente en base a estos resultados se refuerza lo mencionado previamente sobre la sensibilidad de estos algoritmos a datos atípicos y la inicialización de las etiquetas.

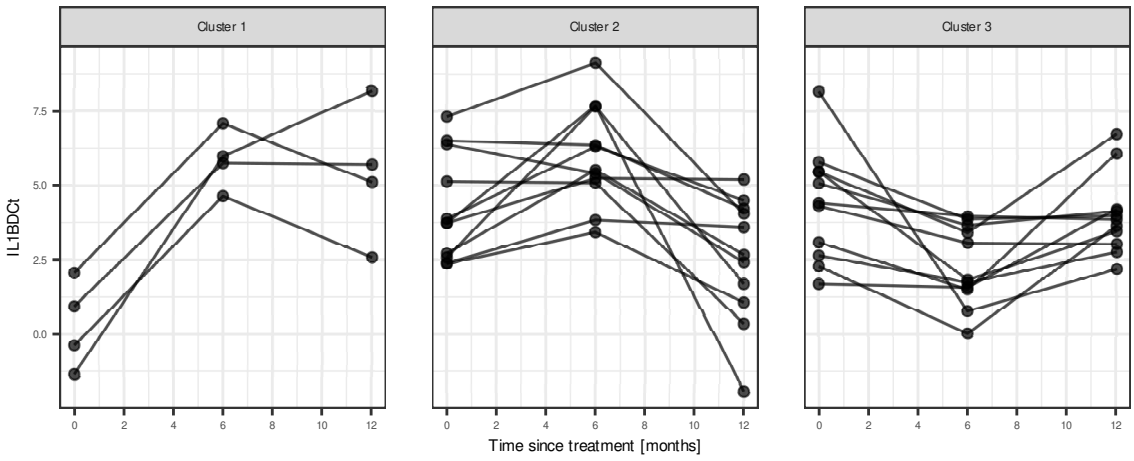


FIGURA 5.18: Resultado del algoritmo de K-medias sobre el espacio de las pendientes de la base DM.

Por otro lado, en la figura 5.20 muestra los resultados del algoritmo de K-medoides, donde se ve un primer cluster donde los individuos presentan un primer decrecimiento de la expresión genética, y un posterior incremento. Además, los individuos del cluster 2 describen un incremento en la primer instancia y un posterior decrecimiento, mientras que los individuos del cluster 3 poseen trayectorias más estables a lo largo del estudio, sin cambios muy abruptos. Vale aclarar que un individuo que fue clasificado en el cluster 2 presenta incrementos en ambas instancias, pero cuyo primer incremento lo acerca más al cluster 2 en comparación con el resto, y por eso fue automáticamente clasificado en el grupo mencionado.

Por lo tanto, todos los análisis subsiguientes se realizan sobre dicha partición. Otro motivo que sustenta el uso de esta partición resultante se debe a que la metodología correspondiente es mucho más robusta. Para verificarlo, corrimos $M = 101$ veces los algoritmos de clustering sobre el espacio de las pendientes de la base DM. En cada iteración, se comparó la similitud entre las

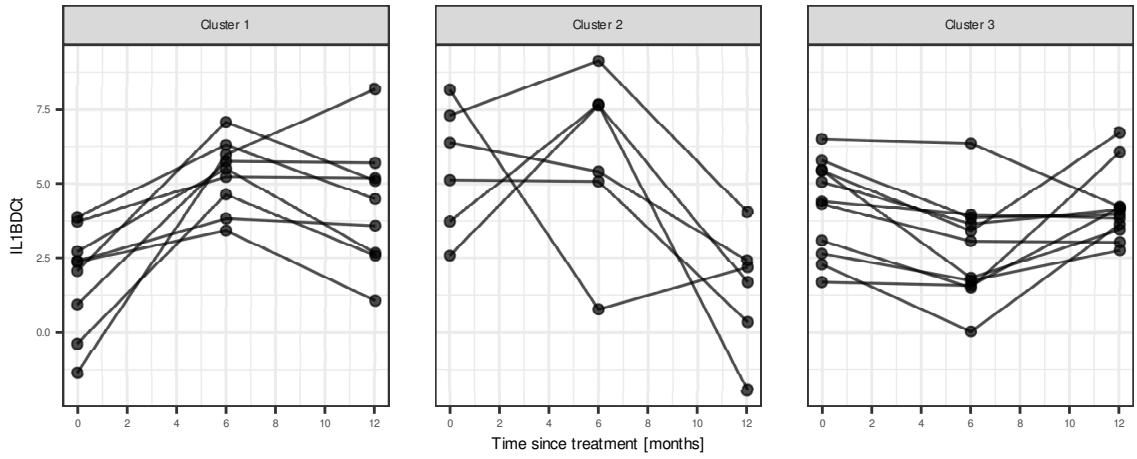


FIGURA 5.19: Resultado del algoritmo de K-medias basado en Kernels sobre el espacio de las pendientes de la base DM.

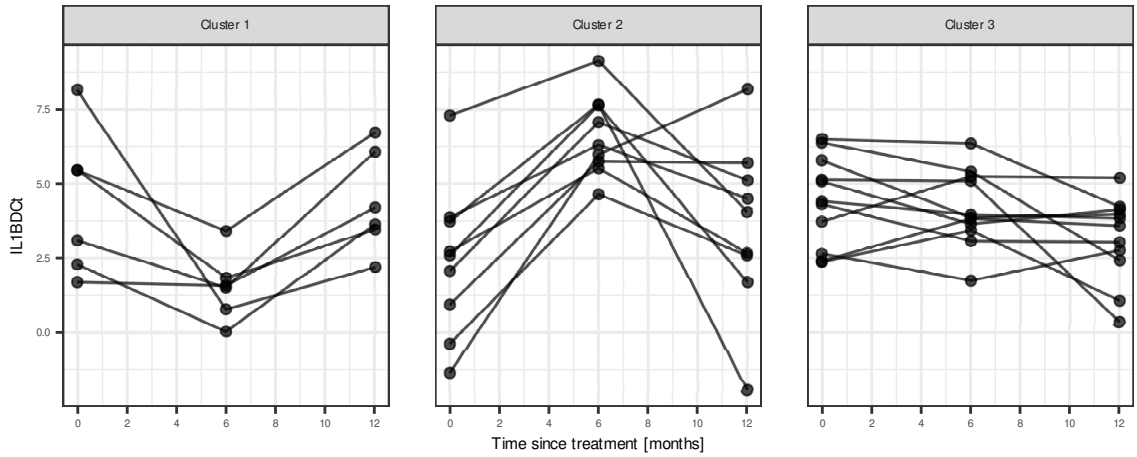


FIGURA 5.20: Resultado del algoritmo de K-medoides utilizando la distancia Manhattan sobre el espacio de las pendientes de la base DM.

particiones obtenidas con un algoritmo y partición obtenida por el mismo algoritmo en la iteración anterior, utilizando el índice de Czekanowski-Dice (abreviado CD) presentado en la sección 5.5.2. Los resultados se exhiben en la Tabla 5.2.

Algoritmo	CD
K-Medoides	1 (0)
K-Medias	0.741 (0.0189)
K-Medias basado en Kernels	0.516 (0.0122)

TABLA 5.2: Media y desvío estándard del índice de Czekanowski-Dice al comparar la partición de cada iteración con la iteración anterior del mismo algoritmo.

Los resultados son elocuentes, al correr varias veces el mismo algoritmo de K-medoides sobre esta base se obtuvo un índice constante de valor 1 (el desvío estándard es 0), que corresponden a particiones iguales en toda iteración. Por lo tanto, además de obtener una clasificación que cumple los criterios pedidos, ese agrupamiento no fue azaroso ni se ve afectado por la inicialización. Es decir, al comparar los resultados de otra variable a través de los grupos, no hay especulación sobre si el grupo resultante es fortuito ni si fue elegido por conveniencia, sino que simplemente es la partición que mejor se ajusta a las condiciones requeridas.

Por otro lado, se observa en qué medida los algoritmos de K-Medias son muy sensibles a la inicialización, ya que al correr varias veces los algoritmos sobre la misma base, se obtienen particiones distintas. De todas formas, vale aclarar que el hecho de que estas particiones sean idénticas por sí mismo no necesariamente deviene en una partición deseable. Por ejemplo, si se corre varias veces el algoritmo de clustering jerárquico, se obtiene siempre la misma partición pero la misma no es deseable porque no cumple los criterios morfológicos, como se observa en la figura 5.21.

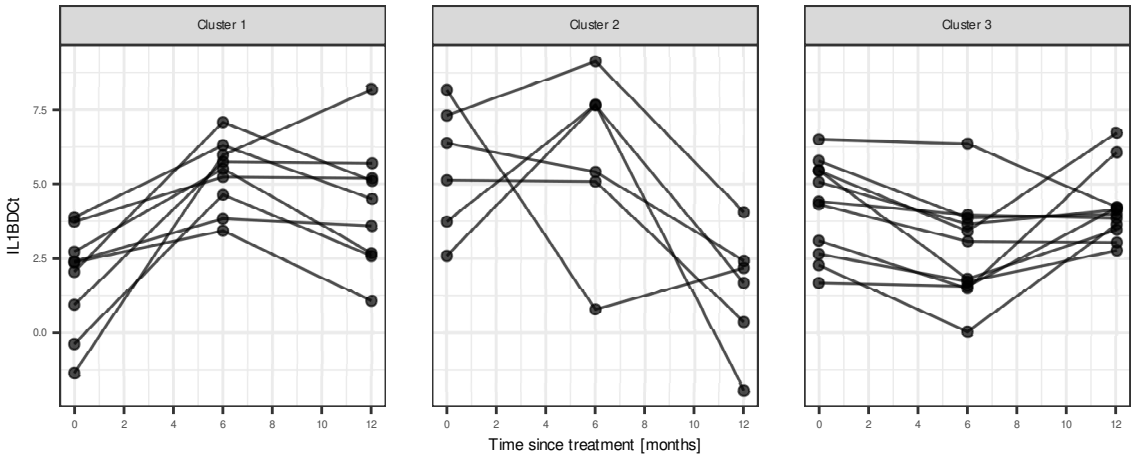


FIGURA 5.21: Resultado del algoritmo de clustering Jerárquico sobre el espacio de las pendientes de la base DM.

Sobre los grupos resultantes se analizaron las variables restantes de la base de datos:

- el peso.
- el índice de masa corporal (IMC)
- los niveles de ambos tipos de colesterol en sangre (HDL y LDL). El HDL es considerado como beneficioso para la salud mientras todo lo contrario sucede con el LDL.
- el porcentaje de hemoglobina glicosilada (notado HbA_{1c}), una variable vinculada al nivel de glucosa en sangre a lo largo de tres meses anteriores a la medición.
- Los niveles de glucosa en sangre
- La longitud de la circunferencia de cintura

- El nivel de triglicéridos en sangre
- El número de componentes del síndrome metabólico (denotado nMSC), según las guías del Panel de tratamiento para adultos dispuesto en [47].
- La edad de los participantes.

Dado el tamaño muestral pequeño, los tests estadísticos que asumen distribuciones normales no son adecuados para este conjunto de datos. Por lo tanto, se aplicaron tests no paramétricos: el test de Kruskal-Wallis fue utilizado para comparar los valores de las variables entre grupos y el test apareado de Wilcoxon para comparar los valores de las variables de cada individuo al inicio y al final del estudio. Vale aclarar que los test no paramétricos suelen ser menos potentes y que se pueden ver más afectados por tamaños muestrales pequeños. Dadas estas características, tiene sentido flexibilizar el nivel de significación del 5 % utilizado por default.

Los resultados principales están dispuestos en la Tabla 5.3. Como las últimas dos variables no se modifican a lo largo del estudio, no se realiza el test apareado de Wilcoxon. Además, por la cantidad de datos faltantes en la circunferencia de cintura en la segunda instancia, estos datos fueron removidos de la tabla.

Los tests arrojaron resultados significativos al 5 % para la mayoría de las variables para el test de Kruskal-Wallis. Además, para los niveles de LDL y el número de componentes del síndrome metabólico el test es significativo al 10 %, lo cual puede ser considerado relevante dado que los valores de la potencia pueden verse reducidos en tests no paramétricos con tamaños muestrales pequeños. Por lo tanto, el test indica que los grupos son significativamente diferentes, pero tras una inspección a posteriori, se ve que la diferencia radica en el cluster 1 ya que es donde se presenta la mayor discrepancia en los valores respecto de los otros grupos.

Las únicas variables donde las diferencias entre grupos no es sustancial es en la hemoglobina glicosilada y los niveles de glucosa en sangre. Esto se debe a que el estudio original tuvo como objetivo principal reducir la glucosa en sangre de todos los participantes y por lo tanto, todos los individuos presentaron niveles similares en estas variables.

Además, se observa que los individuos del cluster 1 presentan mejores indicadores clínicos. Por ejemplo, para dichos individuos se ve un decrecimiento en los niveles de LDL y triglicéridos, mientras que aumenta el HDL, a diferencia de lo que sucede en los otros grupos. Además, el peso, el índice de masa corporal y la longitud de la circunferencia de cintura son más bajos en el cluster 1 en comparación a los otros grupos obtenidos.

Por otra parte, los cambios a lo largo del tiempo en el cluster 1 están avalados por los bajos valores del p-valor en los tests apareados de Wilcoxon, en comparación con los otros grupos. Aún cuando los p-valores no estén por debajo de 10 %, vale recordar que estos tests se realizan de forma intragrupal y que por lo tanto, el tamaño muestral es aún menor que lo que se observa en el conjunto. Además, los p-valores de estos tests son sensiblemente menores en el cluster 1 respecto de los otros clusters, dando sustento a la teoría de que las mejoras temporales en indicadores clínicos para dicho grupo fueron más sustanciales.

Variable	Time	Cluster 1 m (IQR)	Cluster 2 m (IQR)	Cluster 3 m (IQR)	p-valor (Kruskal-Wallis)
	0 mo	81.6 (2.97)	91.5 (19.45)	96.6 (30)	
Peso (kg)	**6 mo**	78.5 (5.57)	89 (18.75)	92.7 (23.2)	***0.0068***
	12 mo	80 (4)	89.5 (13.8)	78.7 (20)	
p-valor (Wilcoxon)	*0/12 meses*	***0.1362***	*0.9055*	*0.3750*	
	0 mo	31.11 (2.2)	32.91 (6.785)	34.02 (6.32)	
IMC (kg/m^2)	**6 mo**	29.68 (0.93)	33.57 (8.66)	32.5 (4.1)	***0.0106***
	12 mo	30.48 (0.94)	33.28 (7.555)	32.8 (5.66)	
p-valor (Wilcoxon)	*0/12 meses*	***0.1250***	*0.9101*	*0.4316*	
	0 mo	42.5 (11.5)	42 (13)	39 (5)	
HDL-c (mg/dL)	**6 mo**	47 (18.5)	41 (7.5)	40.5 (7.75)	***0.0470***
	12 mo	51 (12)	43 (16)	42 (5)	
p-valor (Wilcoxon)	*0/12 meses*	***0.0544***	*0.0852*	*0.1358*	
	0 mo	124.5 (12.75)	116 (33)	127 (33)	
LDL-c (mg/dL)	**6 mo**	123 (22)	99 (23)	120 (61)	***0.0718***
	12 mo	88 (6)	105 (20)	121 (30.5)	
p-valor (Wilcoxon)	*0/12 meses*	***0.1250***	*0.9453*	*0.8125*	
	0 mo	8.605 (2.078)	9.52 (1.78)	8.15 (3.32)	
HbA_{1c} (%)	**6 mo**	6.255 (0.278)	6.37 (0.95)	6.685 (1.39)	*0.6652*
	12 mo	5.9 (0.4)	6.16 (1.19)	6.2 (1.15)	
p-valor (Wilcoxon)	*0/12 meses*	*0.0625*	*0.0039*	*0.0029*	
	0 mo	156.5 (139.75)	147 (137.5)	158 (86)	
Glucosa en sangre (mg/dL)	**6 mo**	106.5 (13.75)	107 (30.5)	114 (37.25)	*0.8086*
	12 mo	108 (7)	114 (19)	117 (21)	
p-valor (Wilcoxon)	*0/12 meses*	*0.0625*	*0.0078*	*0.0322*	
Circunferencia de cintura (cm)	**0 mo**	100.5 (9.25)	104.5 (8.5)	112 (17)	
	12 mo	100 (6)	106 (12.5)	111.5 (20.75)	***0.0149***
p-valor (Wilcoxon)	*0/12 meses*	*0.8922*	*1.0000*	*0.9056*	
	0 mo	128 (35)	182 (114.5)	134 (61)	
Triglicéridos (mg/dL)	**6 mo**	145 (80)	214 (98.5)	201 (96.25)	***0.0047***
	12 mo	82 (8)	192 (93)	130 (126)	
p-valor (Wilcoxon)	*0/12 meses*	***0.1250***	*1.0000*	*0.7597*	
nMSC	-	2.5 (1.75)	4 (1)	4 (2)	*0.05907*
Edad (Años)	-	60.5 (4.5)	42 (13)	46 (18)	*0.00423*

TABLA 5.3: Diferencias observadas en las restantes variables de la base, resumidas en mediana (m) y rango intercuartílico (IQR). Los p-valores de los tests aplicados están indicados con formato itálico.

Capítulo 6
Detección de Outliers

6.1. Introducción

Como se explicó en el capítulo 3, los MEM permiten considerar tanto tendencias individuales como estructuras poblacionales para la variable de respuesta. Esto permite la adaptación de estos modelos a la heterogeneidad que suele estar presente en los datos longitudinales, dado que suelen exhibir tanto variabilidad intersujeto como intrasujeto.

La capacidad de estimar de forma adecuada las tendencias individuales se debe a la inclusión de EA individuales, que además son parámetros que se pueden estimar e interpretar. Por ejemplo, en la sección 3.2.3, los EA $b_{0,i}$ representan la ordenada de la recta individual, mientras que los EA $b_{1,i}$ representan la pendiente de dicha recta individual, ambos siendo considerados respecto de la tendencia marginal. Por lo tanto, a partir de sus valores se pueden sacar conclusiones respecto de la respuesta basal y del crecimiento (o decrecimiento) de un individuo particular en comparación con el resto de la población. Valores positivos de $b_{0,i}$ se interpretan como respuestas basales por encima del valor poblacional estimado y valores negativos representan respuestas basales por debajo de la ordenada marginal estimada. Por otro lado, valores positivos de $b_{1,i}$ representan crecimientos más pronunciados de la respuesta respecto del crecimiento temporal poblacional, mientras que valores negativos de $b_{1,i}$ representan variaciones de menor pendiente respecto de los valores marginales.

Por los comportamientos diversos de las respuestas, puede ser muy difícil extraer conclusiones a nivel poblacional, por lo tanto, puede ser interesante detectar aquellos individuos que exhiben trayectorias con características que se diferencian de la respuesta marginal. Identificar estos individuos puede motivar nuevas preguntas sobre las explicaciones respecto de estas anomalías, cuya respuesta puede estar en variables de confusión que aún no fueron analizadas.

Dado que no hay forma de saber a priori quienes son aquellos individuos de trayectorias anómalas, es de suma dificultad la tarea de detección. Más aún, dichas trayectorias, por definición, son una pequeña proporción de los datos totales.

El trabajo de Chandola et al. [48] realiza un estudio extensivo sobre anomalías y metodologías para su detección, abordando los distintos contextos en los que se aplican estas herramientas. Para ejemplificar los conceptos, utilizaremos datos hipotéticos sobre la evolución de precios de cierto producto en distintos supermercados a lo largo de los días, disponibles en la Figura 6.1.

En esta publicación se hace una categorización de anomalías que será de utilidad para este trabajo:

- **Anomalías puntuales:** Observaciones cuyo valor se encuentran alejados del resto de la población.

  Ejemplo: Supongamos que el precio de venta de cierto producto en varios supermercados oscila entre los 509 y 511 pesos. Si en un momento o supermercado en particular, el precio de venta es de 515 pesos, muestra una diferencia sustancial respecto de los precios de referencia. (Ver Figura 6.1)

- **Anomalías contextuales:** Observaciones cuyo valor no se encuentra necesariamente alejado del resto de la población, pero discrepa del valor esperado según las condiciones en las que se toma la medición.

  Ejemplo: Consideremos el mismo ejemplo de los precios en supermercados. Si un supermercado se observa a lo largo del tiempo y viene teniendo precios similares a los del resto, pero en algún momento ofrece un descuento temporario, ese precio queda circunstancialmente lejos de la tendencia de precios usuales. (Ver Figura 6.1)

- **Anomalías colectivas:** Conjunto de observaciones que no se encuentran necesariamente alejados del resto de la población, pero exhiben un comportamiento distinto que el resto de los individuos.

  Ejemplo: Si todos los supermercados van subiendo sus precios con el tiempo, pero otro supermercado los baja levemente, es una tendencia distinta al del resto de la población (Ver Figura 6.1)

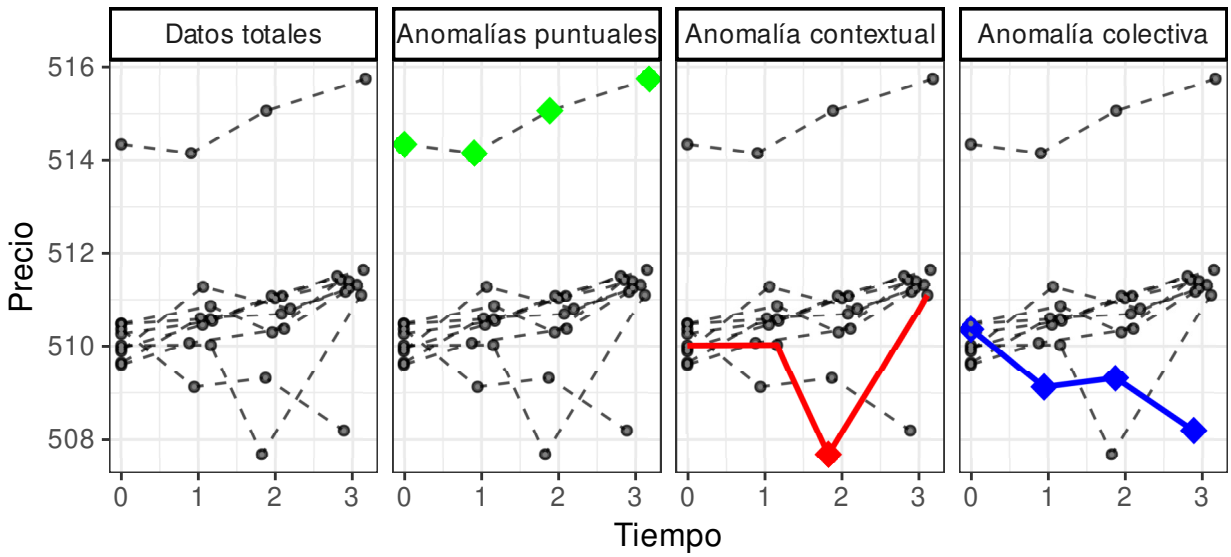


FIGURA 6.1: Representación gráfica de los distintos tipos de anomalías, representadas con rombos en el gráfico.

6.2. Outliers basados en modelos estadísticos

En algunos casos, se tiene un conocimiento previo sobre el fenómeno que se observa. Por lo tanto, se puede plantear un modelo estadístico que prediga de forma adecuada las evoluciones de las trayectorias.

Sin embargo, aún cuando dicho modelo represente correctamente la mayoría de las respuestas, algunos individuos presentan desviaciones considerables respecto de los valores predichos por el modelo, ya sea en un instante puntual o en la totalidad de su trayectoria. Por ende, cuando la detección de outliers se basa en modelos estadísticos, se asume que los datos "normales" exhiben respuestas acordes a dicho modelo, mientras que los datos atípicos o anómalos presentan algún tipo de diferencia sustancial, aunque esto requiera establecer un límite para considerar diferencias como "sustanciales".

En el caso univariado (se puede considerar un único vector x de n datos cuantitativos), hay muchos trabajos (nos basamos mayoritariamente en el trabajo de Davies et al. [49]) que asumen una cierta distribución probabilística normal para los datos y en base a la distribución se establecen límites para el rango de valores "probables". Estos límites pueden establecerse con valores fijos, aunque lo más común es que basen en medidas de dispersión de los datos, ya que mientras más dispersos resulten éstos, más probable es encontrar observaciones alejadas de algún parámetro de centralidad.

La medida de dispersión más conocida es el desvío estándard. La misma se obtiene a partir de la siguiente fórmula:

$$SD(x) = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{6.1}$$

donde $\bar{x}$ representa el promedio de los valores del vector x. En base a estos parámetros y definiendo un umbral T, se pueden definir los siguientes límites para los datos "normales" si están comprendidos en los siguientes valores de L y U:

$$\begin{aligned} L(x) &= \bar{x} - T \cdot SD(x) \\ U(x) &= \bar{x} + T \cdot SD(x) \end{aligned} \tag{6.2}$$

La utilización del desvío estándard suele estar muy vinculada a las distribuciones normales y se usa comunmente el valor de T fijado en 3. Sin embargo, las medidas utilizadas son muy sensibles a outliers y por lo tanto, se puede apelar a una medida de dispersión más robusta. Con este fin, se adopta en algunas ocasiones la mediana de la desviación absoluta que se obtiene del siguiente modo:

$$MAD(x) = \frac{1}{\Phi^{-1}(0.75)} \cdot \underset{1 \leq i \leq n}{\text{Mediana}} \{|x_i - \tilde{x}|\} \tag{6.3}$$

donde $\tilde{x}$ representa la mediana del vector x, $\Phi^{-1}(0.75)$ representa el tercer cuartil de una distribución normal estándard y la constante multiplicadora $\frac{1}{\Phi^{-1}(0.75)}$ asegura que el MAD sea un estimador insesgado del desvío estándard poblacional en el caso de que la distribución de x es normal.

En base a estos parámetros de centralidad y dispersión ($\widetilde{x}$ y $MAD(x)$, respectivamente) se pueden establecer los siguientes límites para los datos considerados "normales":

$$\begin{aligned} L(x) &= \widetilde{x} - T \cdot MAD(x) \\ U(x) &= \widetilde{x} + T \cdot MAD(x) \end{aligned} \tag{6.4}$$

Estos límites presentan mayor robustez a datos atípicos respecto de la ecuación 6.2 para distribuciones simétricas no necesariamente normales. Nuevamente, por la comparabilidad entre ambas metodologías en casos de normalidad, también se adopta generalmente $T = 3$.

Sin embargo, notar que ambas disposiciones dadas en 6.2 y 6.4 son simétricas respecto de una medida de centralidad. Por lo tanto, para distribuciones levemente asimétricas se pueden adoptar límites que se adapten a dicha condición. Por lo tanto, en ciertos casos se utiliza la regla de Tukey:

$$\begin{aligned} L(x) &= Q_1(x) - T \cdot IQR(x) \\ U(x) &= Q_3(x) + T \cdot IQR(x) \end{aligned} \tag{6.5}$$

donde $Q_1(x)$ y $Q_3(x)$ representan el primer y tercer cuartil del vector x, respectivamente, mientras que $IQR(x)$ denota el rango intercuartílico del vector y se obtiene mediante la resta entre $Q_3(x)$ y $Q_1(x)$ que resulta ser siempre positiva ya que, por definición, $Q_3(x) \geq Q_1(x)$. La regla de Tukey (fijando $T = 1.5$) es la más usada para detectar outliers en distribuciones univariadas. Más aún, fijando $T = 3$, la regla detecta outliers severos. Además, como $Q_1(x)$ y $Q_3(x)$ no son necesariamente equidistantes de la mediana del vector x, los límites pueden ser asimétricos respecto de $\widetilde{x}$ y adaptarse mejor a sesgos en el vector.

En este trabajo nos focalizaremos en estas estrategias para datos univariados, aunque las mismas técnicas se aplicarán a distintos vectores.

6.2.1. Residuos

Una vez estimados los parámetros de un MEM, se puede estimar una respuesta $\widehat{Y}_{i,j}$ que busque estimar el valor observado $Y_{i,j}$. Se denomina "residuo" a la diferencia entre estos dos valores y se denota del siguiente modo:

$$r_{i,j} = Y_{i,j} - \widehat{Y}_{i,j} \tag{6.6}$$

en base a estos valores se puede construir un vector de residuos $\mathbf{r}_i \in \mathbb{R}^{J_i}$ para el i-ésimo individuo. Más aún, concatenando todos estos vectores se obtiene un vector $\mathbf{r} \in \mathbb{R}^N$.

Cuando algún valor de $r_{i,j}$ está muy alejado del cero, se puede sospechar que para dicha medición, el modelo no está cumpliendo las expectativas de predicción, y por ende, puede investigarse la existencia de una variable aún no contemplada que pueda explicar semejante diferencia.

El vector $\mathbf{r}_i$ puede obtenerse también mediante la siguiente fórmula

$$\begin{aligned} \mathbf{r}_i &= \mathbf{A}_i \times \mathbf{Y}_i \\ \mathbf{A}_i &= \mathbf{H}_i^{-1} - \mathbf{H}_i^{-1} \times \mathbf{X}_i \times \left(\sum_{k=1}^{I} \mathbf{X}_k' \times \mathbf{H}_k^{-1} \times \mathbf{X}_k \right)^{-1} \times \mathbf{X}_i' \times \mathbf{H}_i^{-1} \end{aligned} \tag{6.7}$$

donde $\mathbf{X}_i'$ representa la transpuesta de la matriz $\mathbf{X}_i$ y‘ la matriz $\mathbf{H}_i$ está definida en la ecuación 3.24. Con estas definiciones puede deducirse la siguiente distribución multivariada para el vector $\mathbf{r}_i$:

$$\mathbf{r}_i \sim \mathcal{N}_{J_i}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{A}_i\right) \Rightarrow r_{i,j} \sim \mathcal{N}(0, \sigma^2 \cdot (\mathbf{A}_i)_{j,j}) \tag{6.8}$$

Por lo tanto, los residuos pueden estandarizarse dividiendo por su desvío, recordando que la matriz $\mathbf{A}_i$ depende de la matriz $\mathbf{H}_i$, cuyos parámetros deben estimarse. Por lo tanto, la matriz $\mathbf{A}_i$ se reemplaza por la matriz estimada $\widehat{\mathbf{A}}_i$:

$$u_{i,j} = \frac{r_{i,j}}{\widehat{\sigma} \cdot \sqrt{(\widehat{\mathbf{A}}_i)_{j,j}}} \tag{6.9}$$

Para evitar la influencia de observaciones que pueden alterar los cálculos, se puede estimar el desvío mediante la eliminación de la j-ésima medición del i-ésimo individuo. Es decir, llamando $\widehat{\sigma}_{(i,j)}$ a la estimación del desvío sin considerar la respuesta $Y_{i,j}$:

$$\begin{aligned} u^*_{i,j} &= \frac{r_{i,j}}{\widehat{\sigma}_{(i,j)} \cdot \sqrt{(\widehat{\mathbf{A}}_i)_{j,j}}} \\ \widehat{\sigma}^2_{(i,j)} &= \widehat{\sigma}^2 \cdot \left(\frac{N - u^2_{i,j}}{N-1}\right) \end{aligned} \tag{6.10}$$

6.2.2. Efectos aleatorios (EA)

Se puede demostrar (ver [50]) que cada vector estimado de EA sigue la siguiente distribución:

$$\widehat{\mathbf{b}_i} \sim \mathcal{N}_Q(\mathbf{0}; \mathbf{G} \times \mathbf{Z}_i' \times \mathbf{A}_i \times \mathbf{Z}_i \times \mathbf{G}) \tag{6.11}$$

donde la matriz $\mathbf{A}_i$ está definida en 6.7.

Por lo tanto, cada EA $\widehat{b}_{i,q}$ se puede estandarizar del siguiente modo:

$$\widehat{v}_{i,q} = \frac{\widehat{b}_{i,q}}{\sqrt{(\mathbf{G} \times \mathbf{Z}_i' \times \mathbf{A}_i \times \mathbf{Z}_i \times \mathbf{G})_{q,q}}} \tag{6.12}$$

Más aún, se puede considerar una matriz $\widehat{\mathbf{B}} \in \mathbb{R}^{I \times Q}$ concatendando por filas los I vectores estimados de EA. Análogamente, se puede construir la matriz $\widehat{\mathbf{V}} \in \mathbb{R}^{I \times Q}$ utilizando las expresiones obtenidas en 6.12.

6.3. Algoritmo

6.3.1. Objetivos

Nuestra propuesta es utilizar un algoritmo que permita detectar tanto anomalías contextuales como colectivas, considerando las distintas interpretaciones que tienen las estimaciones basadas en MEM.

Por ejemplo, como fue mencionado en la sección 6.2.1, residuos con valores alejados del cero representan diferencias grandes entre el valor esperado por el modelo y el valor empírico observado. Por lo tanto, se podría definir como una anomalía contextual. Más aún, si el valor observado termina siendo además atípico comparado con el resto de la población, hasta podría considerarse una anomalía puntual (Ver Figura 6.2). Por lo tanto, pueden detectarse estas anomalías analizando el vector de residuos $\mathbf{r}$.

Por otro lado, los EA con valores estimados alejados del cero también representan trayectorias atípicas. Para simplificar, podemos asumir que los EA consisten de una ordenada y una pendiente por individuo.

- En el caso de las ordenadas, un valor alejado del cero representan trayectorias similares a las de la población pero con niveles basales muy distintos.
- Por otro lado, las pendientes individuales que tienen valores estimados alejados del cero representan crecimientos (o decrecimientos) que difieren de los observados a nivel poblacional.

Por lo tanto, por más que los valores puntuales pueden ser numéricamente similares a los de la población, su trayectoria no evoluciona según los valores esperados. Es decir, las observaciones no necesariamente son atípicas, pero las trayectorias sí lo son (Ver Figura 6.2). Por ende, pueden considerarse estos casos como anomalías colectivas y pueden detectarse a través de las columnas de las matrices $\widehat{\mathbf{B}}$ o $\widehat{\mathbf{V}}$.

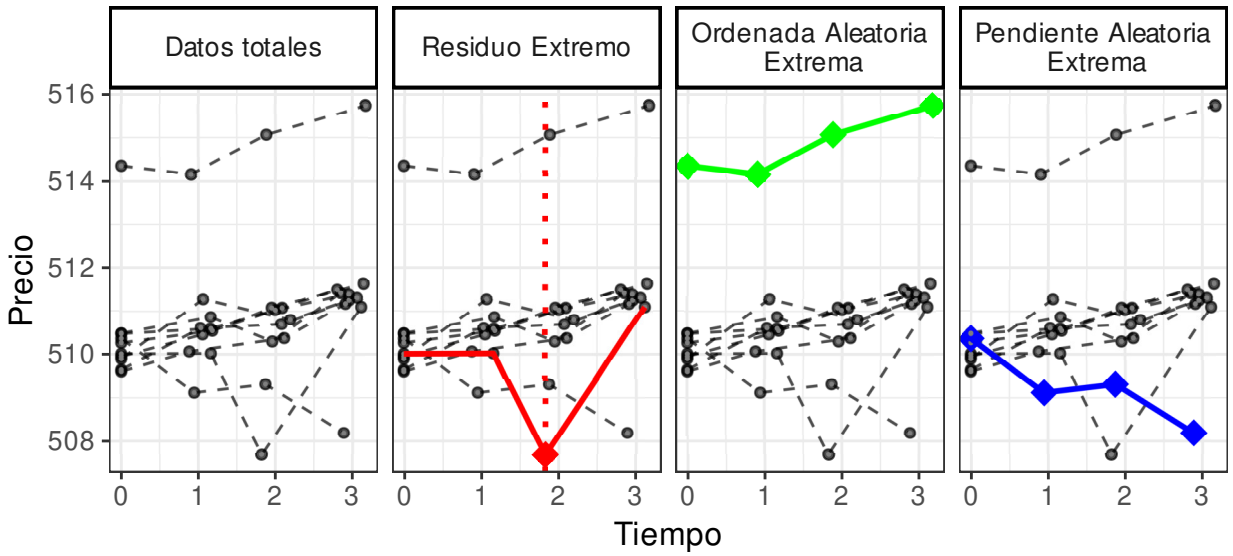


FIGURA 6.2: Representación gráfica de los distintos tipos de anomalías (representadas con rombos), según las interpretaciones de los valores estimados de un MEM.

6.3.2. Descripción del algoritmo

Para detectar los distintos tipos de anomalías, primero se requiere un MEM que represente de forma adecuada a las evoluciones de la población, y en base a los valores estimados de ese modelo,

se pueden obtener y analizar los valores del vector de residuos $\mathbf{r}$ y los EA $\widehat{\mathbf{B}}$.

A continuación disponemos el pseudo-código del algoritmo, mostrando dónde se pueden introducir variantes:

- **Datos de entrada:**
  - Base de datos
  - Nombre de la variable de respuesta.
  - Lista de nombres de las covariables asociadas a efectos fijos.
  - Lista de nombres de las covariables asociadas a efectos aleatorios.
  - Nombre de la variable identificatoria de individuos.
- **Filtrado de datos faltantes**
  - Se remueven todas las filas donde falten respuestas o alguna covariable.
- **Ajuste del modelo**
  - Se generan las matrices de diseño $\mathbf{Y}$, $\mathbf{X}$ y $\mathbf{Z}$.
  - Se estiman los parámetros del modelo y se obtienen $\widehat{\beta}$, $\widehat{\mathbf{G}}$ y $\widehat{\sigma}$.
- **Detección de residuos atípicos**
  - Calcular el vector de residuos $\mathbf{r}$ (puede reemplazarse por $\mathbf{u}$ o $\mathbf{u}^*$, ver ecuaciones 6.6, 6.9 y 6.10).
  - Calcular los límites L_P y U_P para valores normales del vector de residuos según la ecuación 6.5 (se puede reemplazar por los límites obtenidos 6.4 y 6.2).
  - Identificar los residuos fuera del intervalo $[L_P, U_P]$.
  - Guardar en una lista el individuo y el tiempo correspondiente al residuo extremo.
- **Detección de efectos aleatorios atípicos**
  - Calcular la matriz estimada de efectos aleatorios $\widehat{\mathbf{B}}$ (puede reemplazarse por $\widehat{\mathbf{V}}$ ver ecuaciones 3.30 y 6.12).
  - Para cada columna de la matriz estimada, calcular los límites L_q y U_q para valores normales del efecto aleatorio correspondiente según la ecuación 6.5 (se puede reemplazar por los límites obtenidos 6.4 y 6.2).
  - Para cada columna de la matriz estimada, identificar los efectos aleatorios fuera del intervalo $[L_q, U_q]$.
  - Guardar en una lista los individuos, la variables y los valores correspondientes a efectos aleatorios extremos.
- **Datos de salida**
  - La lista de residuos extremos con sus datos correspondientes.
  - La lista de efectos aleatorios extremos con sus datos correspondientes.

Todas estas últimas variantes serán incluidas en los experimentos y simulaciones para comparar sus resultados con los obtenidos por la propuesta principal, que involucra el uso del IQR como medida de dispersión y toma como datos de entrada el vector de residuos $\mathbf{r}$ y la matriz de EA estimados $\mathbf{B}$.

6.3.3. Comparación con otro método

Como se mencionó en la sección 1.4.2, el trabajo de mayor similitud respecto de nuestra propuesta es un trabajo de Zewotir et al. [44], donde se aplica la misma estructura que la presente en el algoritmo descripto en la Sección 6.3.2, aunque con diferentes umbrales. Por eso compararemos nuestra metodología con esta propuesta, ya que mayoritariamente, otros métodos de detección no son aplicables a nuestro contexto.

Para detectar residuos extremos, en vez de utilizar los residuos ordinarios (denotados matemáticamente como $\mathbf{r}$, ver ecuación 6.6), se detectan valores extremos de los residuos estandarizados (denotados matemáticamente como $\mathbf{u}$, ver ecuación 6.9) y se calculan distinto los límites L_P y U_P:

$$L_P = -\sqrt{\frac{4 \cdot N}{N - P + 3}} \qquad U_P = \sqrt{\frac{4 \cdot N}{N - P + 3}} \tag{6.13}$$

donde N representa la cantidad total de observaciones y P la cantidad de EF.

Por otro lado, también tiene una similitud con la estructura del algoritmo de detección de EA extremos, aunque con la diferencia de que en vez de la matriz de EA (denotados $\widehat{\mathbf{B}}$, ver ecuación 6.11), se utiliza la matriz de EA estandarizados (denotados $\widehat{\mathbf{V}}$, ver 6.12). Además, se pueden intercambiar los límites L_q y U_q por las siguientes expresiones, utilizando los cuantiles de la distribución t de student:

$$L_q = -t_{0.975,df} \qquad U_q = t_{0.975,df} \tag{6.14}$$

donde df son los grados de libertad de la distribución, y para este trabajo se asume que df es el rango de la matriz ampliada $[\mathbf{X}\ \mathbf{Z}]$, que se obtiene concatenando a las matrices $\mathbf{X} \in \mathbb{R}^{N \times P}$ (ver ecuación 3.4) y $\mathbf{Z} \in \mathbb{R}^{N \times (I \cdot Q)}$ (donde las matrices individuales $\mathbf{Z}_i$ se disponen de forma diagonal en bloques) por columnas.

Además, vale aclarar que el trabajo de Zewotir et al. [44] no aborda el problema de datos faltantes. Por lo tanto, el valor de N debe coincidir con la cantidad observada de respuestas y covariables. Teniendo en cuenta la estrategia de este trabajo, usaremos sus resultados para comparar con nuestra propuesta.

6.4. Simulaciones

Para testear el algoritmo se simularon datos que siguen un MEM sencillo, con el objetivo de obtener trayectorias de respuestas similares a las observadas en datos longitudinales biomédicos.

6.4.1. Parámetros

Según estos objetivos, se consideró un estudio hipotético en el que hay dos grupos de individuos (podrían considerarse grupo de control y de tratamiento), ambos con correspondientes ordenadas y pendientes como EF, nucleados en el vector β. Respecto de los EA, se consideran ordenadas y pendientes individuales $\mathbf{b}_i = (b_{i,0}, b_{i,1})$ con su respectiva matriz de covarianza $\mathbf{G}$. Además, se

introducen errores de medición independientes $\varepsilon_{i,j}$. Es decir,

$$\begin{gathered}\beta = (\beta_1, \beta_2, \beta_3, \beta_4) \\ \mathbf{b}_i \sim \mathcal{N}_2(0, \mathbf{G}) \\ \mathbf{G} = \begin{pmatrix} \text{Var}(b_{i,0}) & \text{Cov}(b_{i,0}, b_{i,1}) \\ \text{Cov}(b_{i,0}, b_{i,1}) & \text{Var}(b_{i,1}) \end{pmatrix} = \begin{pmatrix} g_{0,0} & g_{0,1} \\ g_{0,1} & g_{1,1} \end{pmatrix} \\ \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma)\end{gathered} \tag{6.15}$$

Por otra parte, dado que las trayectorias longitudinales suelen tener desbalance tanto en la cantidad de mediciones como los instantes de las mismas, se consideran también diferencias en estos atributos.

Para obtener distinta cantidad de mediciones, para cada individuo i se elige un número entero al azar J_i entre los valores J_{min} y J_{max}.

Por otro lado, dado el valor J_i, para obtener distintos instantes de medición en el individuo i se calculan los tiempos de medición según una suma de variables exponenciales independientes para asegurar la positividad de los intervalos temporales entre mediciones consecutivas. Es decir, el j-ésimo instante de medición para el individuo i $t_{i,j}$ se obtiene del siguiente modo:

$$t_{i,j} = \sum_{k=1}^{j} \tau_{i,k} \tag{6.16}$$

donde $1 \leq i \leq I$, $1 \leq j \leq J_i$ y $\tau_{i,k} \sim \mathcal{E}(1)$.

Una vez obtenidos los instantes de medición, se pueden construir las matrices de diseño. Al tener los distintos grupos distintas ordenadas y pendientes, se define cada matriz $\mathbf{X}_i$ del siguiente modo:

$$\begin{aligned} \mathbf{X}_i &= \begin{pmatrix} 1 & t_{i,1} & 0 & 0 \\ 1 & t_{i,2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,J_i} & 0 & 0 \end{pmatrix} && \text{si el } i\text{-ésimo individuo pertenece al grupo control} \\ \mathbf{X}_i &= \begin{pmatrix} 1 & t_{i,1} & 1 & t_{i,1} \\ 1 & t_{i,2} & 1 & t_{i,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,J_i} & 1 & t_{i,J_i} \end{pmatrix} && \text{si el } i\text{-ésimo individuo pertenece al grupo tratamiento} \end{aligned} \tag{6.17}$$

Además, las matrices $\mathbf{Z}_i$ se construyen con la siguiente estructura:

$$\mathbf{Z}_i = \begin{pmatrix} 1 & t_{i,1} \\ 1 & t_{i,2} \\ \vdots & \vdots \\ 1 & t_{i,J_i} \end{pmatrix} \tag{6.18}$$

A partir de estos datos, se puede construir una trayectoria $\mathbf{Y}_i$ que sigue el siguiente modelo:

$$\mathbf{Y}_i = \mathbf{X}_i \times \vec{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \varepsilon_i \tag{6.19}$$

6.4.2. Anomalías

En la sección anterior se describieron las trayectorias que siguen el modelo estadístico dado en 6.19. Sin embargo, el objetivo consiste en identificar trayectorias atípicas. Por lo tanto, debemos construir trayectorias que no obedezcan estrictamente dicho modelo.

Para introducir anormalidades, tratamos de seguir los conceptos de cada categorización detallada en [48]. Generar desviaciones en algunos instantes específicos tiene el objetivo de generar valores residuales extremos (lo llamaremos VRE). Por otro lado, valores extremos de la ordenada o la pendiente individual, generan anomalías colectivas. Como notación para estos últimos valores, utilizaremos las siglas OAE para ordenada aleatoria extrema, PAE para pendiente aleatoria extrema y más generalmente, EAE como efecto aleatorio extremo.

Para introducir anomalías en instantes de medición específicos, se puede construir un vector $\mathbf{D}_i^P$ que tenga valores nulos en gran parte de sus coordenadas salvo en una pequeña proporción p_P. Por otro lado, para que la anomalía sea detectable, debe tener un valor D_c considerable respecto de los errores de medición $\varepsilon_{i,j}$ (que tienen como desvío el parámetro σ), es decir:

$$\begin{gathered} U_{i,j}^1, U_{i,j}^2 \sim \mathcal{U}(0,1) \\ d_P = c_P \cdot \sigma \\ D_{i,j}^P = \begin{cases} d_P & \text{si } U_{i,j}^1 \leq p_P, U_{i,j}^2 \leq 0.5 \\ -d_P & \text{si } U_{i,j}^1 \leq p_P, U_{i,j}^2 > 0.5 \\ 0 & \text{en caso contrario.} \end{cases} \end{gathered} \tag{6.20}$$

donde además, para evitar introducir sesgo, la variable $U_{i,j}^2$ busca que las anomalías sean construidas con signos positivos y negativos en igual proporción. Juntando todas estas coordenadas para cada individuo i, se puede construir un vector $\mathbf{D}_i^P$ donde se almacenan las distintas perturbaciones puntuales al modelo para dicho individuo.

De manera similar, se pueden construir datos extremos para los EA, sólo que debe considerarse que el desvío de las ordenadas individuales viene dada por $\sqrt{g_{0,0}}$. Por lo tanto, tomando un múltiplo de este valor por una constante c_{RI}, se puede construir el valor D_i^{RI} que genera ordenadas individuales atípicas para el individuo i con una proporción de p_{RI} mediante los siguientes cálculos:

$$\begin{gathered} U_i^1, U_i^2 \sim \mathcal{U}(0,1) \\ d_{RI} = c_{RI} \cdot \sqrt{g_{0,0}} \\ D_i^{RI} = \begin{cases} d_{RI} & \text{si } U_i^1 \leq p_{RI}, U_i^2 \leq 0.5 \\ -d_{RI} & \text{si } U_i^1 \leq p_{RI}, U_i^2 > 0.5 \\ 0 & \text{en caso contrario.} \end{cases} \end{gathered} \tag{6.21}$$

Notar que como un cambio en la ordenada afecta toda la trayectoria, no hay distintas coordenadas de D_i^{RI} para cada individuo y por lo tanto, no es un vector.

Análogamente, se puede construir la perturbación a la pendiente aleatoria D_i^{RS}, con proporción p_{RS} de valores no nulos y un múltiplo c_{RS} del correspondiente desvío $\sqrt{g_{1,1}}$:

$$\begin{gathered} U_i^1, U_i^2 \sim \mathcal{U}(0,1) \\ d_{RS} = c_{RS} \cdot \sqrt{g_{1,1}} \\ D_i^{RS} = \begin{cases} d_{RS} & \text{si } U_i^1 \leq p_{RS}, U_i^2 \leq 0.5 \\ -d_{RS} & \text{si } U_i^1 \leq p_{RS}, U_i^2 > 0.5 \\ 0 & \text{en caso contrario} \end{cases} \end{gathered} \tag{6.22}$$

Para el individuo i, se puede considerar el vector $\mathbf{D}_i^R = (D_i^{RI}, D_i^{RS})$ como una perturbación general a los EA. Con estas construcciones, se puede generar una trayectoria atípica $\widetilde{\mathbf{Y}}_i$ mediante la siguiente ecuación:

$$\widetilde{\mathbf{Y}}_i = \mathbf{X}_i \times \vec{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \mathbf{D}_i^P + \mathbf{Z}_i \times \mathbf{D}_i^R + \varepsilon_i \tag{6.23}$$

6.4.3. Valores de los parámetros

Cuando las constantes c_{RI} y c_{RS} coinciden, se usa la notación c_R para simbolizar ambos valores. Del mismo modo, para apreciar la interacción entre las ordenadas y pendientes individuales, la simulación toma un único parámetro p_R para la proporción de anomalías colectivas y se toman las siguientes tres combinaciones, de forma que $p_{RI} + p_{RS} = p_R$:

- $p_{RI} = p_R, p_{RS} = 0$
- $p_{RI} = 0, p_{RS} = p_R$
- $p_{RI} = p_{RS} = \frac{p_R}{2}$

La siguiente tabla representa los valores de los parámetros que son constantes en todas las simulaciones:

β_1=1	β_2=0.1	β_3=-2	β_4=-0.2	J_{min}=2
$g_{0,0}$=0.051	$g_{0,1}$=-0.001	$g_{1,1}$=0.051	σ=0.25	J_{max}=4

TABLA 6.1: Parámetros fijos en todas las simulaciones.

Por otro lado, el resto de los parámetros tienen valores específicos según la tarea de detección que se realiza. Por ejemplo, cuando se quiere evaluar la detección de VRE, se amortigua el impacto de los EA ajustando p_R para no alterar tanto los parámetros estimados y, por ende, las detecciones. Además, el valor de T_P o T_R de los algoritmos **??** y **??**, respectivamente, puede ser incrementado para establecer límites más restrictivos para tareas de detección que no estén siendo evaluadas. Por ejemplo, si se evalúa la detección de VRE, se puede poner un valor razonable para el parámetro T_P y valores altos de T_{RI} y T_{RS}, para que no realice ninguna detección de EAE.

Parámetros para la detección de residuos extremos (VRE)

En estos casos, cuando se aplica la metodología dada en 6.5, se toma T_P como 1.5, mientras que cuando se utilizan las estrategias dadas en 6.4 y 6.2, se establece T_P=3.

p_{RI}=0, 0.025, 0.05	p_{RS}=0, 0.025, 0.05	p_P= 0.05, 0.1, 0.2, 0.25
c_{RI}=1, 2	c_{RS}=1, 2	c_P= 3, 4
T_{RI}=4	T_{RS}=4	I= 100, 200, 300

TABLA 6.2: Parámetros para la detección de VRE.

Como fue mencionado anteriormente, los valores de c_{RI} y c_{RS} son más bajos para reducir el impacto sobre los parámetros estimados y los valores de T_{RI} y T_{RS} son más altos para no consumir memoria en guardar detecciones que no serán analizadas. En cada tarea de detección subsiguiente se observan también decisiones similares en valores que no corresponden a la magnitud evaluada.

Parámetros para la detección de ordenadas individuales extremas

p_P=0, 0.1, 0.2	p_{RS}=0, 0.05, 0.1, 0.2	p_{RI}= 0.05, 0.1, 0.2
c_P=1, 2	c_{RS}=1, 2	c_{RI}= 3, 4
T_P=4	T_{RS}=4	I= 100, 200, 300

TABLA 6.3: Parámetros para la detección de OAE.

En estos casos, cuando se aplica la metodología dada en 6.5, se toma T_{RI} como 1.5, mientras que cuando se utilizan las estrategias dadas en 6.4 y 6.2, se establece T_{RI}=3.

Parámetros para la detección de pendientes individuales extremas

p_P=0, 0.1, 0.2	p_{RI}=0, 0.05, 0.1, 0.2	p_{RS}= 0.05, 0.1, 0.2
c_P=1, 2	c_{RI}=1, 2	c_{RS}= 3, 4
T_P=4	T_{RI}=4	I= 100, 200, 300

TABLA 6.4: Parámetros para la detección de pendientes individuales extremas.

En estos casos, cuando se aplica la metodología dada en 6.5, se toma T_{RS} como 1.5, mientras que cuando se utilizan las estrategias dadas en 6.4 y 6.2, se establece T_{RS}=3.

6.4.4. Evaluación

En general, para cualquier algoritmo de detección no supervisado, la dificultad a la hora de evaluar el procedimiento reside en la incapacidad de saber cuáles son las mediciones atípicas. Sin embargo, cuando los datos son simulados se sabe exactamente qué mediciones o trayectorias fueron alteradas ya que se introdujeron artificialmente.

En nuestro caso, esas perturbaciones vienen dadas por los vectores $\mathbf{D}_i^P \in \mathbb{R}^{J_i \times 1}$ y $\mathbf{D}_i^R = (D_i^{RI}, D_i^{RS}) \in \mathbb{R}^{2\times 1}$ descriptos en las ecuaciones 6.20, 6.21 y 6.22, o más aún, sus extensiones a todos los individuos $\mathbf{D}^P \in \mathbb{R}^{N\times 1}$ y $\mathbf{D}^R = (\mathbf{D}^{RI}, \mathbf{D}^{RS}) \in \mathbb{R}^{2\times I}$. Estos vectores pueden tomarse como referencia, dado que son anomalías reales y se pueden rastrear. Por lo tanto, cuando se

aplica el algoritmo sobre una base generada por las construcciones descriptas en la ecuación 6.23, se pueden comparar las detecciones obtenidas con las de referencia.

Por lo tanto, considerando una detección del algoritmo como "positiva", se pueden establecer las siguientes categorías:

- $T+$: Detecciones del algoritmo en una base simulada.
- $T-$: Observaciones no detectadas en la base simulada.
- $R+$: Valores no nulos del vector $\mathbf{D}^P$ (o la matriz $\mathbf{D}^R$, según la tarea de detección).
- $R-$: Valores nulos del vector $\mathbf{D}^P$ (o la matriz $\mathbf{D}^R$, según corresponda).

Por lo tanto, las observaciones que sean categorizadas como "+" o "-" en ambos casos son aciertos del algoritmo mientras que en caso contrario son detecciones erróneas. Por lo tanto, podemos pensar en los siguientes casos:

	$R+$	$R-$
$T+$	VP	FP
$T-$	FN	VN

(6.24)

donde VP y VN representan la cantidad de verdaderos positivos y negativos, respectivamente. Del mismo modo, FP y FN se corresponden con la cantidad de falsos positivos y negativos, respectivamente.

En base a estos valores, se pueden establecer las siguientes magnitudes que evalúan los resultados de las detecciones de cada variante del algoritmo, muy establecidas en la literatura:

- **Sensibilidad (S)**: $\frac{VP}{VP+FN}$
- **Valor predictivo positivo (VPP)**: $\frac{VP}{VP+FP}$
- **Especificidad (E)**: $\frac{VN}{FP+VN}$
- **Valor predictivo negativo (VPN)**: $\frac{VN}{FN+VN}$

Es decir,

- la sensibilidad representa qué proporción de las anomalías de referencia ($R+$) son detectadas por el algoritmo ($T+$).
- la especificidad consiste de la proporción de datos no anómalos ($R-$) que son identificados como tales ($T-$).
- el valor predictivo positivo contempla qué proporción de las detecciones positivas ($T+$) son realmente anomalías.
- el valor predictivo negativo evalúa qué proporción de los datos considerados normales ($T-$) lo son realmente ($R-$).

De algún modo, la sensibilidad y la especificidad contemplan la capacidad de detección del algoritmo y los valores predictivos permiten considerar la confiabilidad de los resultados obtenidos.

Es decir, un algoritmo de alta sensibilidad y bajo valor predictivo positivo, detecta muchas anomalías, pero también determina que varias observaciones son atípicas cuando no lo son realmente, resultando en un umbral muy laxo para considerar un dato como anómalo.

Por otro lado, un algoritmo de baja sensibilidad y alto valor predictivo positivo, detecta pocos casos anómalos pero se puede tener mayor certeza sobre el resultado de una detección positiva.

Vale aclarar además que el enfoque está puesto en las detecciones positivas, ya que el algoritmo devuelve como resultado las observaciones consideradas anómalas y por lo tanto, tanto la sensibilidad como el valor predictivo positivo permiten evaluar la calidad de dichos resultados.

Por otro lado, como las observaciones anómalas son por definición mucho menores en proporción que los datos que siguen el modelo, generalmente tanto los valores predictivos negativos como la especificidad suelen tener valores cercanos a 1 dado que el valor de verdaderos negativos (VN) suele ser mucho mayor que los otros valores de la tabla de contingencia dada en 6.24.

6.4.5. Datos faltantes

Siguiendo los métodos descriptos en 4.2 se remueven algunas respuestas previamente simuladas según la ecuación 6.23, con proporciones de remoción dadas por $p_M = 0.1$, 0.15 y 0.2. Estas remociones se realizan para evaluar el impacto de los datos faltantes sobre los resultados. Además, se ensaya con los distintos mecanismos de remoción de respuestas descriptos en 4.1 para analizar las diferencias que pueden ocasionar en las detecciones.

Vale aclarar que en estos casos, deben realizarse ajustes a los valores calculados en la tabla dada en 6.24 ante la remoción de respuestas, ya que si la respuesta removida era previamente una anomalía ($R+$), no puede ser detectada ($T-$) y por lo tanto, no puede contarse como un "falso negativo". Del mismo modo, si el dato removido era fiel al modelo dado en 6.19 (es decir, se categoriza como $R-$), entonces tampoco puede considerarse que el hecho de no detectar dicha observación ($T-$) no debería resultar en contarlo como un "verdadero negativo".

6.4.6. Resultados

En la siguiente sección mostramos los resultados de las simulaciones. Recordamos que las detecciones "positivas" son las de mayor importancia. Por lo tanto, los resultados se centran en la sensibilidad y el VPP. Además, como fue mencionado en la sección 6.4.4, los valores de la especificidad y el VPN son en general elevados y no presentan mayores diferencias. Ambas medidas son importantes pero dependiendo del área de la aplicación y de las necesidades del usuario, alguna de estas dos medidas puede adquirir mayor relevancia.

En muchos casos, sobre todo en aplicaciones biomédicas, perder una detección positiva puede resultar en un costo altísimo, por lo que el usuario puede preferir una sensibilidad más alta y un mayor número de detecciones, a partir de la cual se pueden descartar de forma manual los falsos positivos. Sin embargo, en caso de requerir mayor nivel de automatización, el usuario puede preferir una menor sensibilidad a costo de un mayor VPP, dado que el proceso de descarte de falsos positivos puede ser muy abrumador y se elige en este caso que las detecciones sean menos, pero

efectivamente positivas. De todas formas, cualquier método de detección elegido debe tener un buen balance entre sensibilidad y VPP.

Por otro lado, en todas nuestras simulaciones, la metodología de Zewotir et al. mencionada en 6.3.3 no detectó VRE y por lo tanto, no están presentes en los resultados.

Salvo que se indique lo contrario, se considera un número de $I = 100$ individuos, ya que tampoco se observaron diferencias notorias en los valores de las medidas evaluatorias al variar el tamaño muestral.

Detección de residuos extremos (VRE)

La figura 6.3 muestra la evolución de la sensibilidad media al incrementar el valor p_P aplicando el algoritmo (descripto en la Sección 6.3.2) a los residuos ordinarios **r**, con valores fijos de c_P,c_{RI} y c_{RS}, mientras que los valores de p_{RI} y p_{RS} se varían para ver cómo la proporción de datos atípicos afectan los niveles de la sensibilidad.

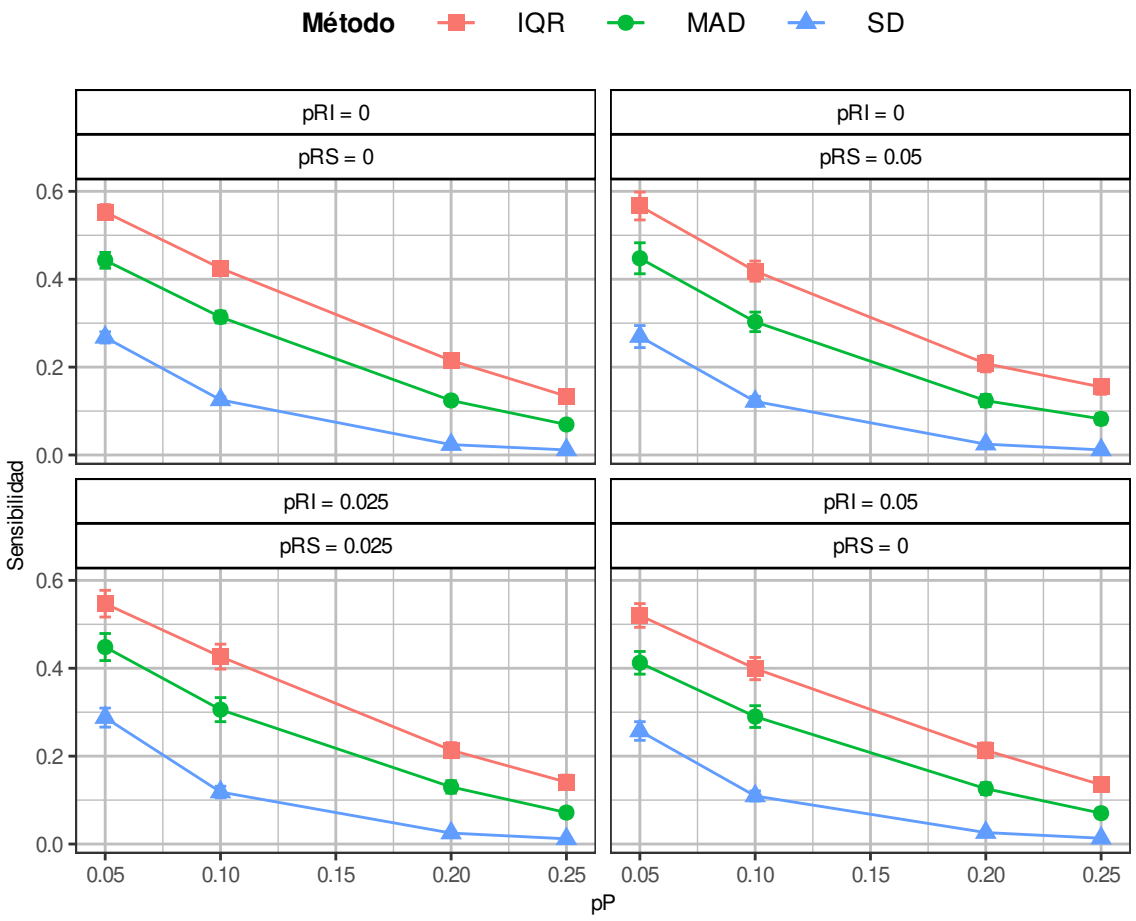


FIGURA 6.3: Sensibilidad media y error estándard de la tarea de detección de VRE con $c_P = 4$ y $c_{RI} = c_{RS} = 1$ como valores fijos.

Según el gráfico, a medida que se agregan datos atípicos, la sensibilidad exhibe una tendencia

decreciente. Este fenómeno no parece respetar la lógica, ya que al haber más datos para detectar, debería ser más fácil para el algoritmo identificar los mismos. Sin embargo, los efectos que tienen estos datos atípicos sobre los parámetros estimados, devienen en una mayor variabilidad de los residuos, generando un aumento de las medidas de dispersión y por lo tanto, mayor amplitud en los umbrales de detección.

En todos los paneles, se observa una mayor sensibilidad en el método que involucra al IQR. Esta prevalencia se observa en todas las simulaciones en las que se detectan VRE, básicamente porque los umbrales generados por el método IQR son levemente menos restrictivos que los otros métodos, dando un mayor número de detecciones positivas.

Además, vale aclarar que el descenso en la sensibilidad para el método SD es más vertiginoso que los de los otros métodos. Esto es consecuencia de la susceptibilidad del desvío estándard a los datos atípicos.

De todas formas se observa un resultado inesperado: cuando sólo se agregan valores de pendientes extremas ($p_{RI} = 0$, $p_{RS} = 0.05$), aumenta levemente la sensibilidad en la detección en comparación al caso en el que no se introducen EAE ($p_{RI} = p_{RS} = 0$). Es decir, mayor ruido aumenta la capacidad de detección del algoritmo. Sin embargo, lo contrario ocurre cuando sólo se agregan ordenadas atípicas ($p_{RI} = 0.05$, $p_{RS} = 0$). Por lo que el impacto del ruido agregado en un caso es beneficioso y en otro caso es perjudicial para la sensibilidad. Esto se debe a que al agregar variabilidad en las pendientes, las estimaciones posteriores de las trayectorias heredan dicha variabilidad y al tratarse de pendientes, tiene un impacto en todos los tiempos de medición y las trayectorias estimadas tienen mayor adaptabilidad. Por otro lado, cuando se agregan sólo ordenadas extremas, el efecto sobre las trayectorias estimadas se observa únicamente en la respuesta basal. Por lo tanto, no deviene en una mayor adaptabilidad de las trayectorias en todos los instantes de medición y en consecuencia, los residuos presentan una mayor dispersión.

La figura 6.4 muestra la evolución de la sensibilidad media al incrementar el valor p_P aplicando el algoritmo (descripto en la Sección 6.3.2) a los residuos ordinarios $\mathbf{r}$, con valores fijos de p_{RI} y p_{RS}, mientras que los valores de c_P, c_{RI} y c_{RS} varían para ver cómo la magnitud de los picos afectan los niveles de la sensibilidad.

Según lo esperado, mayores valores de c_P generan una mayor sensibilidad ya que al generar picos de mayor magnitud, son más fáciles de detectar. De todos modos, también se observa, como en el gráfico anterior, que al aumentar la magnitud de los EAE, aumenta levemente la sensibilidad en los métodos de IQR y MAD, por el efecto mencionado de la mayor adaptabilidad de las trayectorias estimadas.

La figura 6.5 analiza los resultados del VPP, con valores fijos de c_P, p_{RI} y p_{RS}, mientras que varían los valores de c_{RI} y c_{RS}.

Se ve en la figura que el método SD arroja los mejores resultados en cuanto al VPP. Esto se debe a que el método SD es el más restrictivo de los 3 considerados en este gráfico. Por lo tanto, a pesar de tener un menor número de detecciones, suelen ser en su mayoría verdaderos positivos. A pesar de esta prevalencia, los tres métodos presentan un aceptable valor de VPP y elegir el método SD viene con una pérdida notable en la sensibilidad.

Además, se observa en la figura que el método SD presenta una mejoría en su VPP cuando

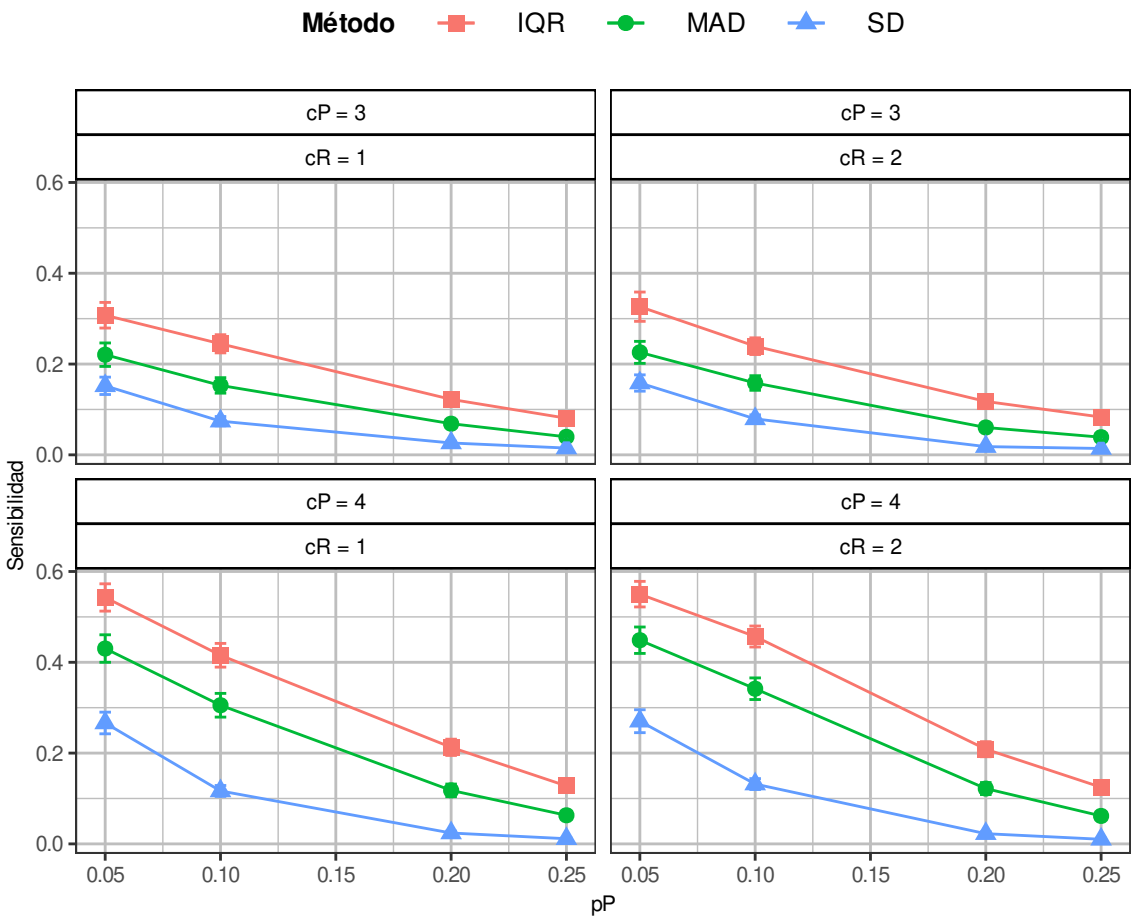


FIGURA 6.4: Sensibilidad media y error estándard de la tarea de detección de VRE con $p_{RI} = p_{RS} = 0.025$ como valores fijos.

aumenta el valor de c_R, reforzando la idea de que mayor variabilidad en los EA pueden llegar a mejorar las trayectorias estimadas.

Por otro lado, no se observaron diferencias notables al comparar las tareas de detección en las distintas clases de vectores residuales: Ordinarios (notados **r** en la ecuación 6.6), Estandarizados (notados **u** en la ecuación 6.9) y Predichos (notados **u*** en la ecuación 6.10), por lo que omitimos cualquier gráfico que involucre dicha comparación. Además, las evoluciones de las medias exhiben similares morfologías a lo largo de todas las combinaciones de las simulaciones, por lo que en las secciones subsiguientes se limitarán más los valores de los parámetros para poder describir las características principales de los resultados.

Detección de ordenadas aleatorias extremas

Aclaramos que a partir de esta sección el método de Zewotir et al. mencionado en 6.3.3 será incluido en la comparación dado que tuvo detecciones positivas de EAE. En la figura 6.6 se analiza la sensibilidad y el VPP al aplicar el algoritmo descripto en la Sección 6.3.2 a bases simuladas con

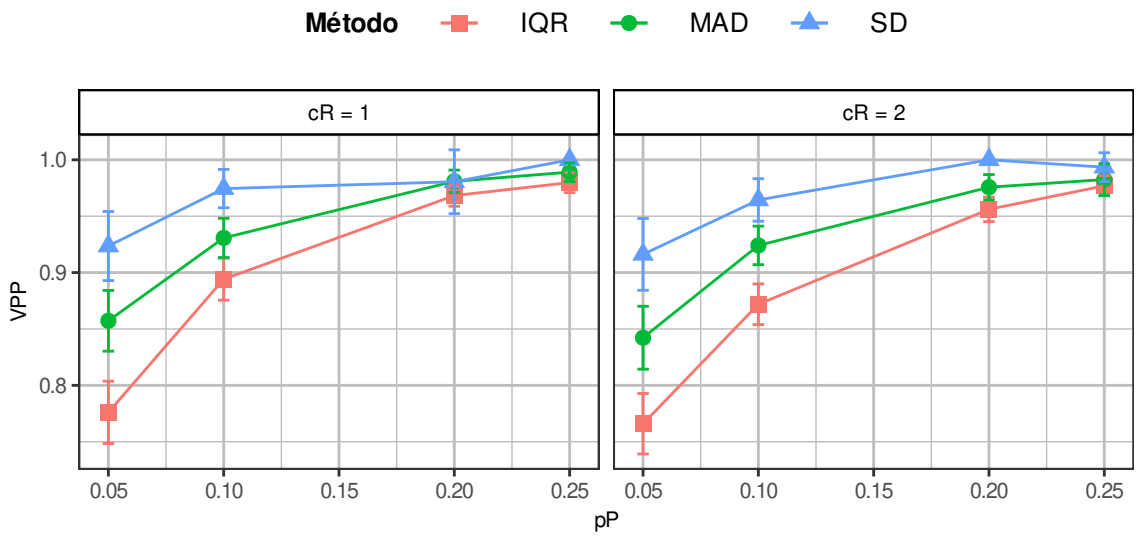
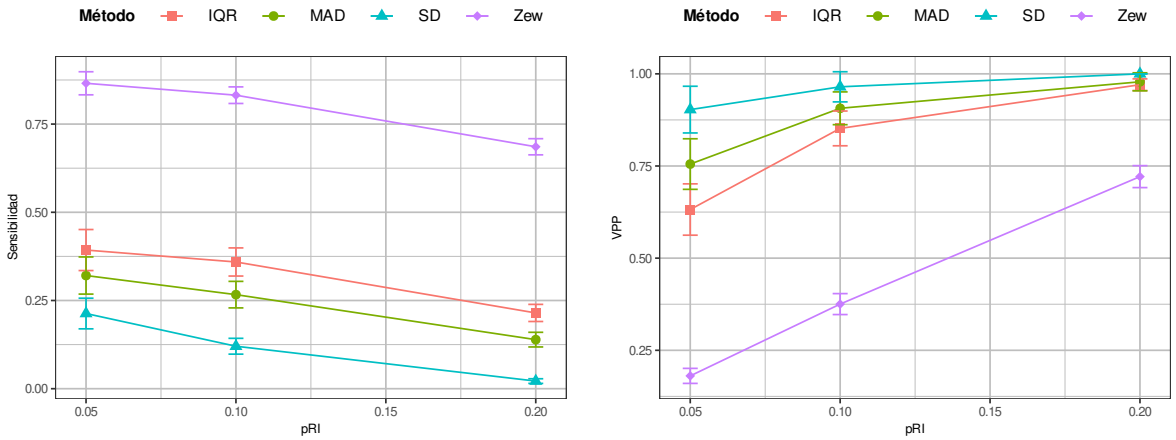


FIGURA 6.5: VPP medio y error estándard de la tarea de detección de VRE con $c_P = 4$ y $p_{RI} = p_{RS} = 0.025$ como valores fijos.

p_{RI} como valores variables y manteniendo fijos los valores de p_P, p_{RS}, c_P y c_{RS}. Para la figura las detecciones se realizan en base a la matriz de EA estimados **B**, salvo para la metodología de Zewotir et al., en la que se utiliza la matriz de EA estandarizados (notada **V**), ambas matrices descriptas en la sección 6.2.2. Vale aclarar que para los métodos IQR, MAD y SD, no se detectaron diferencias notorias entre el uso de las matrices **B** y **V** y por lo tanto, son omitidos sus resultados.



(*a*) (*b*)

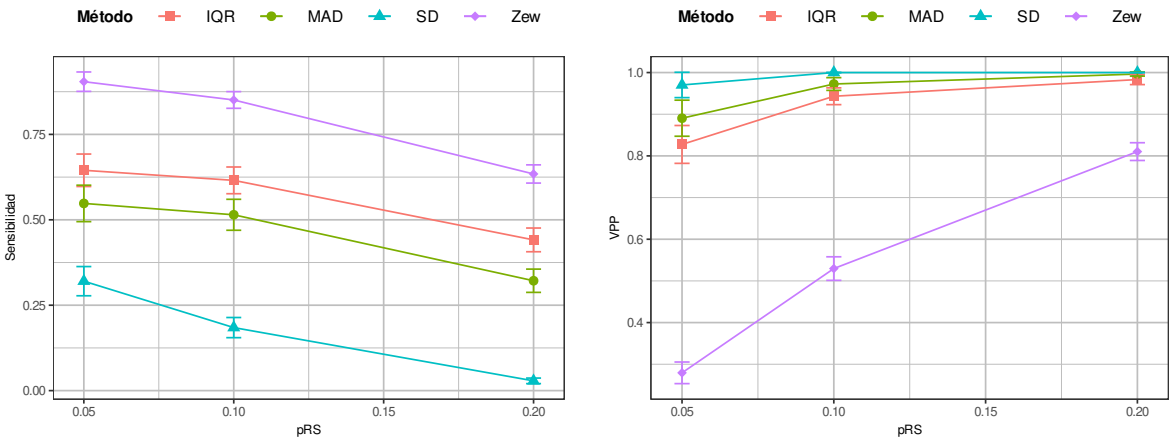
FIGURA 6.6: Análisis de (*a*) Sensibilidad y (*b*) VPP de la tarea de detección de VRE con $p_P = p_{RS} = 0$, $c_P = c_{RS} = 1$ y $c_{RI} = 4$ como valores fijos.

Claramente el método de Zewotir et al. da como resultado la mayor sensibilidad, aunque con el costo de un VPP muy reducido. Es decir, detecta un gran número de datos atípicos, pero la mayoría son falsos positivos (el 20 % de las detecciones cuando $p_{RI} = 0.05$). Salvo por este método, el resto de las evoluciones de la Sensibilidades y VPP medias mantienen comportamientos similares

a los obtenidos en la detección de VRE, con una prevalencia del método IQR en la sensibilidad por encima de los métodos de MAD y SD.

Detección de pendientes aleatorias extremas

En la figura 6.7 se visualizan los resultados de la Sensibilidad y VPP en la detección de pendientes extremas. Nuevamente, salvo para el método de Zewotir et al. (en la que se usa la matriz **V**), se utiliza la matriz de EA estimados **B**.



(*a*) (*b*)

FIGURA 6.7: Análisis de (*a*) Sensibilidad y (*b*) VPP de la tarea de detección de pendientes extremas con $p_P = p_{RI} = 0$, $c_P = c_{RI} = 1$ y $c_{RS} = 4$ como valores fijos.

En comparación con la figura 6.6, todas las metodologías aumentan notablemente su Sensibilidad y VPP, con la salvedad del método de Zewotir et al. Esta excepción tiene que ver con la naturaleza del umbral propuesto en este método. Al ser un umbral constante, además de utilizar la estandarización de los EA, se normalizan todas las distribuciones de distintos EA a una distribución única, omitiendo la interpretación de los mismos y su diferente impacto en las trayectorias estimadas. Por lo tanto, como los umbrales de los métodos IQR, MAD y SD varían según la dispersión de los EA, son más adaptables al mayor impacto en las trayectorias, producto de las pendientes aleatorias extremas.

Por otro lado, en esta tarea de detección, se observa una diferencia entre el uso de la matriz de EA estimados **B** y la estandarización correspondiente **V**. Por ejemplo, en la figura 6.8, se observa un mejor rendimiento del umbral obtenido con la matriz **B** respecto del uso de la matriz **V**, utilizando el método IQR. Esto se debe a la estandarización que mitiga el impacto de las pendientes atípicas. Por lo tanto, las detecciones basadas en la dispersión real de cada vector columna muestran mayor robustez.

Análisis de datos faltantes

Se aplicaron los métodos de remoción de datos faltantes descriptos en 4.2 a las bases simuladas, utilizando los tres mecanismos (MCAR, MAR y NMAR) y tomando como p_M los valores 0.05, 0.1

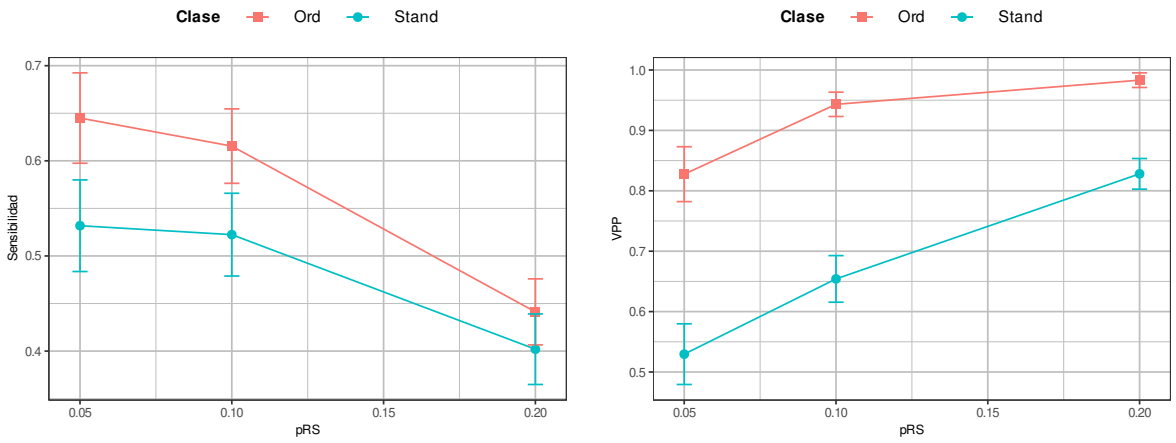


FIGURA 6.8: Análisis de (*a*) Sensibilidad y (*b*) VPP de la tarea de detección de pendientes extremas usando el IQR como medida de dispersión para los EA (ordinarios y estandarizados) con $p_P = p_{RI} = 0$, $c_P = c_{RI} = 1$ y $c_{RS} = 4$ como valores fijos.

y 0.2. Si bien hay alteraciones en las medidas de evaluación (debidos a los cambios tanto en numeradores como en denominadores), no se observaron muchas diferencias morfológicas con las tareas de detección dadas en las secciones anteriores. Por lo tanto, limitamos los análisis a unos pocos casos en los que se observan algunos resultados interesantes. Los resultados se describen en las Figuras 6.9 y 6.10, donde se analiza la sensibilidad en la detección residual manteniendo $I = 100$, $c_P = 4$, $c_{RI} = c_{RS} = 1$ y $p_{RI} = p_{RS} = 0$.

A priori, se podría pensar que al aumentar la proporción de respuestas removidas, mayor es el impacto sobre las estimaciones y que por lo tanto, sea más difícil detectar datos atípicos. Sin embargo, se produce un resultado inesperado en la Figura 6.9, ya que al aumentar la proporción de remoción de respuestas, en algunos casos aumenta levemente la sensibilidad. Además, en la Figura 6.10, se observa una menor sensibilidad de detección en el escenario más deseable, es decir, bajo el mecanismo MCAR en el que no se introduce sesgo. Estos resultados inesperados pueden deberse a que el sesgo introducido esté removiendo más detecciones positivas que negativas y que por lo tanto, reduce los valores del denominador, aumentando el cociente.

Las detecciones de EAE casi no presentaron diferencias numéricas, dado que el EA es individual y que por lo tanto, para dejar de considerar un individuo en el análisis debe removerse la correspondiente trayectoria de respuesta en su totalidad, evento que no es probable.

Complejidad computacional

La Tabla 6.5 analiza los tiempos de corrida de cada algoritmo en microsegundos. Se ve una diferencia sustancial en la complejidad al utilizar las matrices de residuos y EA ordinarios y sus respectivas estandarizaciones. Los niveles elevados de costo computacional se explican por el tamaño de las matrices (y sus respectivas matrices inversas) involucradas en los cálculos de las estandarizaciones. Aún si estos cálculos se pueden hacer de forma más eficiente por la estructura en bloques de las matrices de diseño correspondientes a cada individuo, el costo temporal es significativamente

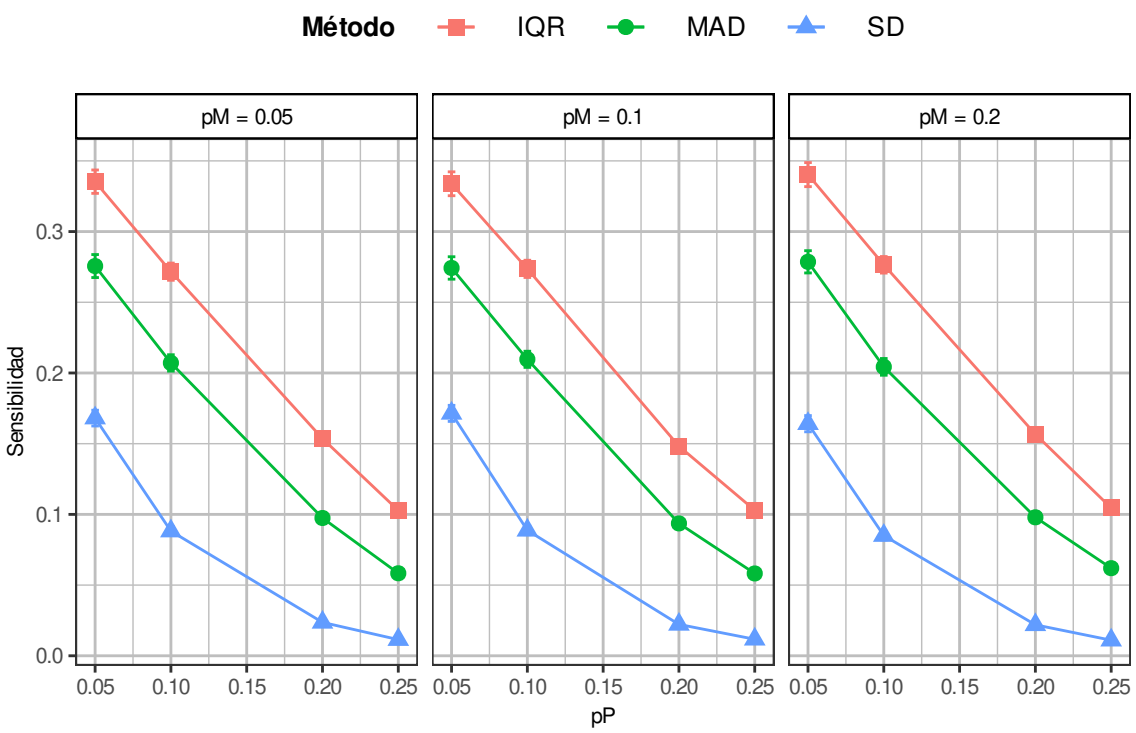


FIGURA 6.9: Comparación de sensibilidad ante distintas proporciones de remoción, utilizando el mecanismo de NMAR de remoción de datos.

mayor para las detecciones que involucran estandarizaciones.

Además, los tiempos de corrida de todos los métodos parecen aumentar de forma relativamente proporcional con la cantidad de individuos, salvo por el método de Zewotir et al., que crece de manera más pronunciada. Justamente, en este método, el cálculo del umbral para los EA requiere de matrices de diseño que no pueden reducirse por la estructura en bloques.

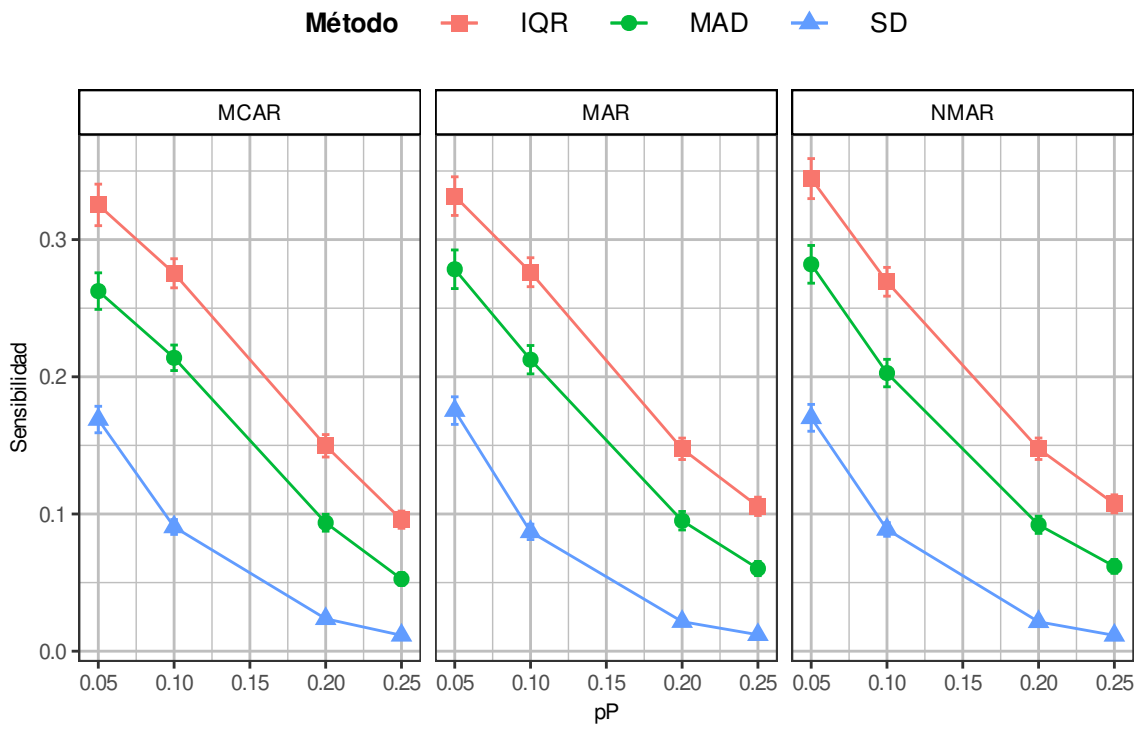


FIGURA 6.10: Comparación de sensibilidad ante distintos mecanismos de remoción, utilizando removiendo el 10 % de datos.

Método	Tarea de Detección	Clase	I=100	I=200	I=300
IQR	Residuos	Ordinarios (r)	123.68 (6.652)	201.999 (6.172)	304.836 (5.004)
		Estandarizados (u)	800.39 (6.83)	1540.423 (6.001)	2298.749 (5.982)
		Predichos (u*)	798.981 (6.914)	1542.468 (5.414)	2303.471 (5.445)
	Ordenadas aleatorias	Ordinarias (B)	117.96 (4.79)	198.392 (2.111)	290.204 (4.273)
		Estandarizadas (V)	793.246 (5.144)	1541.703 (3.665)	2285.728 (5.288)
	Pendientes aleatorias	Ordinarias (B)	130.528 (7.727)	200.202 (6.058)	286.727 (4.165)
		Estandarizadas (V)	803.665 (7.971)	1524.815 (6.079)	2284.487 (4.943)
MAD	Residuos	Ordinarios (r)	123.32 (6.688)	196.087 (5.055)	289.112 (4.381)
		Estandarizados (t)	801.081 (7.072)	1545.474 (6.253)	2313.759 (5.734)
		Predichos (t*)	800.261 (7.22)	1544.727 (5.825)	2316.206 (5.727)
	Ordenadas aleatorias	Ordinarias (B)	117.17 (4.747)	199.374 (2.385)	287.683 (3.9)
		Estandarizadas (V)	794.22 (5.773)	1548.236 (3.852)	2310.719 (5.12)
	Pendientes aleatorias	Ordinarias (B)	129.908 (7.717)	198.605 (5.902)	284.058 (3.686)
		Estandarizados (V)	805.085 (8.209)	1543.607 (6.577)	2307.901 (5.173)
SD	Residuos	Ordinarios (r)	123.324 (6.644)	195.668 (5.127)	288.259 (4.412)
		Estandarizados (u)	798.961 (7.501)	1547.536 (5.855)	2326.721 (5.995)
		Predichos (u*)	800.217 (6.836)	1549.968 (5.782)	2319.264 (6.356)
	Ordenada aleatoria	Ordinarias (B)	116.627 (4.716)	197.859 (2.079)	286.553 (3.901)
		Estandarizadas (V)	795.396 (5.858)	1542.785 (3.578)	2308.299 (5.068)
	Pendientes aleatorias	Ordinarias (B)	129.851 (7.688)	198.693 (5.879)	283.707 (3.734)
		Estandarizadas (V)	807.589 (8.283)	1548.965 (6.198)	2317.318 (5.556)
Zew	Residuos	Estandarizados (u)	824.169 (7.106)	1794.68 (7.252)	3350.533 (10.55)
	Ordenada aleatoria	Estandarizadas (V)	817.616 (5.238)	1786.657 (4.915)	3325.185 (8.747)
	Pendiente aleatoria	Estandarizadas (V)	829.924 (8.372)	1809.246 (7.23)	3345.491 (8.43)

TABLA 6.5: Análisis del tiempo medio y desvío estándard de corrida (en microsegundos) para cada metodología de elección del umbral, adoptando distintos tamaños muestrales.

6.5. Bases reales

Se aplicaron las distintas variantes de los algoritmos de detección a las bases descriptas en la sección 3.2.3, utilizando como MEM de referencia los descriptos en las ecuaciones 3.33, 3.34 y 3.35. Se calcularon las cantidades de detecciones para cada metodología y se exhiben los resultados en la Tabla 6.6. Vale aclarar que los umbrales dispuestos por Zewotir et al., se aplican sólo a vectores estandarizados y por lo tanto, el resto de los valores son omitidos.

Data	Detección	Clase	IQR	MAD	SD	Zew
FEV_1	Residuos	Ordinarios (**r**)	39	30	25	-
		Estandarizados (**u**)	41	28	25	0
		Predichos ($\mathbf{u}^*$)	41	28	25	-
	Efectos aleatorios	Ordinarios (**B**)	22	10	4	-
		Estandarizados (**V**)	6	2	2	522
NGCS	Residuos	Ordinarios (**r**)	4	0	0	-
		Estandarizados (**u**)	4	1	1	417
		Predichos ($\mathbf{u}^*$)	14	13	4	-
	Efectos aleatorios	Ordinarios (**B**)	4	3	2	-
		Estandarizados (**V**)	4	3	2	0
TLC	Residuos	Ordinarios (**r**)	19	19	5	-
		Estandarizados (**u**)	21	21	6	208
		Predichos ($\mathbf{u}^*$)	23	22	6	-
	Efectos aleatorios	Ordinarios (**B**)	4	6	0	-
		Estandarizados (**V**)	4	6	0	0

TABLA 6.6: Número de detecciones de cada algoritmo en bases reales

Para la base FEV_1, para cada método, no hay grandes distinciones entre el uso de residuos ordinarios, estandarizados y predichos. Sin embargo, la diferencia es más notoria en la detección de EAE, en la que se ve que el uso de los EA ordinarios provee mayores detecciones. Esto se puede deber a que la estandarización de los EA busca llevar todas las columnas de la matriz a una misma distribución, sin considerar las diferencias de interpretación de cada EA. Por ejemplo, se ha discutido en la sección previa los distintos impactos que tienen las ordenadas aleatorias y pendientes aleatorias sobre las trayectorias, omitir estas diferencias puede resultar en una menor capacidad de detección.

Por otro lado, el método de Zewotir et al. no detecta ningún residuo como extremo y una gran cantidad de EAE (más aún cuando se considera la cantidad total de EA dada por $I \times Q = 600$). Estos resultados son consistentes con los obtenidos en las simulaciones. Es decir, la sensibilidad es alta por el gran número de detecciones, pero en su mayoría son falsos positivos y esto deviene en que esta metodología presente un valor bajo del VPP.

El efecto inverso se observa en las bases NGCS y TLC. La estandarización de los EA no afecta la cantidad de detecciones, mientras que hay una mayor diferencia en los distintos vectores residuales utilizados, alcanzando mayor notoriedad en la base NGCS.

En este último caso, se debe a que las tendencias estimadas por el modelo no siempre se adaptan bien a las oscilaciones de las trayectorias (ver Figura 3.8). Por lo tanto, hay varios residuos alejados del cero que pueden ser muy altos y afectar el desvío estándard. Además, la cantidad de datos es menor en este caso y magnifica el impacto en las estimaciones. Cuando estos datos son removidos,

el desvío estándard se achica y por lo tanto, cambia sustancialmente los valores de los residuos predichos.

Además, vemos nuevamente en las bases TLC y NGCS que la metodología de Zewotir et al. devuelve resultados muy erráticos, ya que en este caso no se detecta ningún EAE, pero se detecta una gran cantidad de VRE en comparación con la cantidad total de datos. Por otro lado, en la base FEV_1 ocurre lo contrario: muchas detecciones en los EA pero ninguna en los residuos. En general, observamos que en esta metodología los umbrales son constantes, afectados mayoritariamente por la cantidad de datos, sin tener en cuenta cómo estos datos varían (consideración presente en los otros métodos), resultando en umbrales tanto excesivamente tolerantes como restrictivos. Por lo tanto, tienden a tener demasiadas detecciones en algunos casos, pero detecciones nulas en otros. El impacto de la cantidad de datos en esta metodología se observa al considerar la diferencia entre los tamaños muestrales de las bases: la base FEV_1 tiene $N = 1993$ observaciones, mientras que las bases NGCS y TLC tienen $N = 447$ y $N = 400$ observaciones, respectivamente.

Capítulo 7
Conclusiones

7.1. Discusión

A continuación describimos algunas de las preguntas que surgen a partir de nuestros trabajos y qué desarrollos pueden desprenderse a futuro a partir de los mismos.

7.1.1. Agrupamiento por expresión genética

En este trabajo se usan particiones que establecen de forma binaria la pertenencia de un individuo a cierto grupo. Esto se debe al hecho de que el objetivo principal es encontrar diferencias en otras variables entre los distintos grupos y esta comparación pierde relevancia cuando los grupos que se comparan no están bien definidos. Por otro lado, como los algoritmos de tipo "fuzzy clustering" agrupan observaciones dependiendo del grado de pertenencia respecto de algún umbral, si el umbral es muy restrictivo se pueden obtener varios grupos (más de los deseados) con menor cantidad de individuos. Si la cantidad de individuos por grupo es muy limitada y la cantidad de grupos es grande, la comparación entre grupos también carece de sentido. De todas formas, los algoritmos de "partición blanda" son muy utilizados para buscar asociaciones a nivel molecular y celular, ya que en esos casos la cantidad de datos suele ser inmensa y en este caso, los algoritmos mencionados pueden encontrar tendencias grupales con mayor eficiencia, ya que se ven menos influenciados por outliers y el uso del umbral puede descartar asociaciones débiles.

Nuestra propuesta obtiene buenos resultados en los casos en los que la información previa sobre el fenómeno observado es limitada y las bases de datos poseen otras características que inhabilitan algunas herramientas estadísticas establecidas. Es decir, cuando hay muchas variables a considerar, pocos individuos en la base con pocos instantes de medición (que a su vez pueden diferir entre sí). Estas circunstancias son muy comunes para profesionales de la salud que buscan analizar los datos de sus pacientes. De todas formas, vale aclarar que cualquier información previa puede aportar a la mejora de los resultados. Por ejemplo, como la propuesta tiene la versatilidad de poder utilizar cualquier algoritmo de clustering (y a su vez, basarse en cualquier función de distancia), cualquier información previa puede facilitar la selección una función de distancia que permita diferenciar con mayor eficiencia individuos distintos, según los objetivos del estudio.

Teníamos expectativas de mejores rendimientos del algoritmo de K-medias basado en el kernel radial, dado que suele diferenciar individuos según la distancia de la observación al vector nulo, que en el espacio de las pendientes representan a las trayectorias estables. Sin embargo los resultados

no fueron los esperados, ya que la distancia al vector nulo no contempla el signo de las pendientes. Por lo tanto, como trabajo futuro puede diseñarse una función de distancia híbrida que contemple tanto el kernel radial como los signos de las pendientes, que capture de manera más específicas las diferencias entre vectores de pendientes.

Por otro lado, en casos de presencia de outliers severos, el algoritmo se puede utilizar como una herramienta de detección, aplicando en el espacio de las pendientes algún algoritmo de clustering con alta sensibilidad a outliers. Por ejemplo, en las pruebas el clustering jerárquico mostró tendencia a aislar individuos con trayectorias muy atípicas. Por lo tanto, aplicando esta metodología en el espacio de las pendientes se puede obtener un grupo que sólo contenga a un outlier (o, dependiendo de la cantidad de outliers, algunos grupos reducidos) y volver a correr un algoritmo más robusto sobre los vectores de pendientes restantes.

Además, vale aclarar que las bases utilizadas no presentan un número elevado de individuos, por lo que el costo computacional de K-medoides es un contratiempo de consecuencias imperceptibles. Sin embargo, para bases de datos masivas, puede convenir el uso de K-medias ya que, como describen la mayoría de las pruebas, en muchos casos se obtienen resultados similares y el costo computacional es menor.

Por último, es importante aclarar que el algoritmo tiene fines analíticos y está pensado para estudios observacionales, en las que las condiciones de los datos obtenidos ya están establecidos. Por lo tanto, no debe usarse para obtener conclusiones determinantes. Todos los resultados obtenidos deben ser sometidos a estudios prospectivos más rigurosos y controlados.

7.1.2. Detección de trayectorias atípicas

Según nuestros experimentos, el método de detección basado en el IQR provee resultados con un buen balance entre Sensibilidad y VPP, además de tener una baja complejidad computacional. Si bien el metodo de Zewotir et al. presenta mayor sensibilidad para el caso de los EA, viene con un bajo valor de VPP, ya que la cantidad de detecciones es tan alta en comparación con la cantidad total de observaciones que la mayoría son falsos positivos.

En base a estos resultados, podemos obtener como primer conclusión que las metodologías que proponen umbrales constantes (o que se basan mayoritariamente en la cantidad de datos) tienden a ser demasiado restrictivos o demasiado tolerantes, ya que se obtiene una cantidad excesiva o nula de detecciones. Más aún, en estos casos los límites se obtienen asumiendo hipótesis distribucionales muy estrictas. Sin embargo, justamente cuando una base presenta datos atípicos, estas hipótesis se ven afectadas y los límites obtenidos pueden carecer de validez. Por otro lado, las metodologías que se basan en medidas de dispersión parecen adaptarse mejor a los efectos de los datos atípicos y proveen una cantidad razonable de detecciones.

Por otro lado, observamos que los distintos EA tienen diversos impactos sobre las trayectorias y por lo tanto, cuando se busca normalizar todas sus distribuciones, se puede perder precisión en las detecciones. Por lo tanto, considerar detecciones separadas para cada EA (con distintas metodologías que se ajusten a las necesidades del usuario) puede proveer mejores resultados que usar una misma estrategia para todos los EA.

Respecto a variantes que se pueden introducir al algoritmo descripto en la Sección 6.3.2, donde no se requiere que el valor de T sea exactamente 1.5 para el IQR ni 3 para las metodologías basadas en la MAD y SD. Estos valores de tolerancia pueden modificarse para obtener umbrales más restrictivos o tolerantes, según se requiera mejorar el rendimiento de los algoritmos de detección.

Además, los límites pueden establecerse en base a la aplicación. Por ejemplo, un usuario podría preferir detectar alguna diferencia fija o porcentual entre el valor observado y el estimado que tenga una interpretación propia del área de investigación.

Otra estrategia posible es ordenar los residuos y EA según su distancia al valor nulo e identificar un cierto porcentaje de datos que se encuentran mas alejados de dicho valor esperado.

Vale aclarar también que las detecciones pueden ser una referencia para el usuario, pero que pueda utilizar su experiencia para identificar cuáles de las detecciones son relevantes. En caso de obtener demasiados falsos positivos, se puede concluir que el umbral fue demasiado tolerante y el mismo se puede seleccionar con un criterio más restrictivo. Más aún, si las detecciones consideradas como falsos positivos comparten alguna característica respecto de otras variables, puede llegar a identificarse una variable de confusión, que pueda ser incluida para mejorar el modelo y la capacidad de detección.

Por otro lado, muchas de las estrategias descriptas en la sección 1.4.2, a pesar de haber sido descartadas para su uso en las trayectorias de respuesta, pueden ser adaptadas para ser usadas para los vectores y el contexto adecuado. Por ejemplo, los métodos basados en densidad de observaciones o algoritmos de clustering pueden aplicarse a los vectores de residuos y/o a las columnas de la matriz de EA por separado. Considerando que cada vector (o vector columna) es univariado, los datos faltantes dejan de ser una limitación ya que se aplican en un contexto unidimensional.

Además, se ha mencionado que la distancia de Mahalanobis aplicada a las trayectorias de respuesta se ve afectada por los datos faltantes, ya que no hay una longitud determinada para dichos vectores. Sin embargo, una vez que se tiene un MEM y se estiman sus coeficientes, se puede obtener una matriz de covarianza individual para los residuos según la ecuación 6.8 (tienen media nula) y puede aplicarse la distancia de Mahalanobis sin la necesidad de una media y una matriz de covarianza común a todos los individuos. Más aún, alguna estimación robusta de la matriz de covarianza puede proveer aún mejores resultados aunque debe pagarse un costo computacional por dicha mejoría.

Respecto a las simulaciones, los mecanismos utilizados para generar los instantes de medición se obtuvieron utilizando variables aleatorias exponenciales. Sin embargo, la misma estrategia puede utilizarse considerando otras variables aleatorias no negativas como la gamma, weibull o uniformes de soporte positivo.

Además, para los mecanismos de remoción de datos detallados en la sección 4.2 se utilizaron sumas ponderadas en cada caso para asegurar que las probabilidades de remoción sean comparables para distintos mecanismos. Sin embargo, estos pesos pueden alterarse o puede obtenerse una función no necesariamente lineal de las variables de forma que la proporción esperada de remoción se mantenga idéntica en distintos mecanismos.

Por otro lado, el algoritmo puede mejorar su rendimiento agregando una segunda etapa a la estimación. En la versión actual, los parámetros del modelo son estimados removiendo los datos

faltantes. Pueden utilizarse estas estimaciones para reemplazar los datos faltantes, imputando un valor que responda al modelo, en donde se reemplazan los parámetros desconocidos por los valores estimados. Una vez imputados estos datos, se puede volver a correr el algoritmo, y obtenerse mejores resultados.

Por último, se ha mencionado que los métodos para abordar los datos faltantes descriptos en la sección 4.4 requieren hipótesis sobre el mecanismo de datos faltantes, y que por lo tanto, casi nunca es aplicable a datos observacionales. Sin embargo, para los mecanismos simulados en la sección 4.2, el mecanismo es conocido y puede aplicarse alguna de las factorizaciones descriptas para estimar con mayor precisión los parámetros del modelo y en consecuencia, mejorar tanto los residuos como las estimaciones de los EA. A partir de estas mejoras se puede evaluar el impacto en cada algoritmo de detección y ver su dependencia de las factorizaciones correspondientes. Los algoritmos más robustos serán de mayor aplicabilidad a datos observacionales reales.

7.2. Conclusiones generales

En nuestro trabajo se desarrollaron algoritmos que permiten obtener nuevas características a partir de bases de datos aplicando ideas relativamente simples, pero con una fuerte base en lo conceptual.

En un mundo en creciente automatización, tendemos a delegar las conclusiones a los algoritmos y dejamos de lado nuestra capacidad analítica. Sin embargo, para que un resultado se obtenga de forma automática, deben realizarse concesiones respecto de las hipótesis y basarse en parámetros que resuman características de los datos con el fin de obtener resultados inmediatos, dando lugar a potenciales errores sistemáticos y sesgos en los resultados. Por tanto remarcamos la importancia de mantener siempre una mirada crítica sobre los resultados y que los desarrollos tecnológicos no pasen a ser un reemplazo de nuestra capacidad de pensar e interpretar, sino como una herramienta para asistir y complementar nuestras conclusiones.

En ambas propuestas es central el conocimiento previo sobre el fenómeno observado, ya que en base a esto también se ve impactado el alcance de las conclusiones. Para el algoritmo desarrollado en el capítulo 5, la estrategia es aplicable cuando el conocimiento previo es ínfimo o nulo, y se busca encontrar algún indicio de asociación que luego debe ser correctamente verificada. Por otro lado, la propuesta del capítulo 6 requiere un conocimiento previo sobre el área de aplicación suficiente para poder plantear un MEM válido.

De todas formas, también concluimos que los cálculos que requieren hipótesis distribucionales muy precisas, pueden verse muy afectados en estudios observacionales en los que no se tiene control sobre las posibles variables de confusión y los datos son tomados de forma retrospectiva. En la práctica es muy común que los datos no sigan estructuras tan precisas y suelen describir valores heterogéneos y en general, con datos faltantes que pueden introducir sesgos. En estos casos, se debe utilizar algoritmos que posean una adaptabilidad mínima para proveer resultados adecuados.

Bibliografía

[1] N. Peek, J. H. Holmes, and J. Sun, "Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 42–47, 2014.

[2] R. Bellazzi, M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler, and A. T. McCray, "Data analysis and data mining: current issues in biomedical informatics," *Methods of information in medicine*, vol. 50, no. 06, pp. 536–544, 2011.

[3] H. C. Koh, G. Tan, *et al.*, "Data mining applications in healthcare," *Journal of healthcare information management*, vol. 19, no. 2, p. 65, 2011.

[4] C. Doukas, T. Pliakas, and I. Maglogiannis, "Mobile healthcare information management utilizing cloud computing and android os," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 1037–1040, IEEE, 2010.

[5] M. Hansen, T. Miron-Shatz, A. Lau, and C. Paton, "Big data in science and healthcare: a review of recent literature and perspectives," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 21–26, 2014.

[6] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[7] F. Altiparmak, H. Ferhatosmanoglu, S. Erdal, and D. C. Trost, "Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 254–263, 2006.

[8] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*, vol. 998. John Wiley & Sons, 2012.

[9] L. J. Pantazis, G. D. Frechtel, G. E. Cerrone, R. García, and A. E. I. Molli, "Phenotype similarities in automatically grouped t2d patients by variation-based clustering of il-1β gene expression," *EJIFCC*, vol. 34, no. 3, p. 228, 2023.

[10] L. J. Pantazis and R. A. García, "Detection of atypical response trajectories in biomedical longitudinal databases," *The International Journal of Biostatistics*, vol. 19, no. 2, pp. 389–415, 2022.

[11] J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young, "Revisiting global gene expression analysis," *Cell*, vol. 151, no. 3, pp. 476–482, 2012.

[12] P. Saldi and G. W. Hatfield, "Dna m icroarrays and gene expression," 2002.

[13] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modelling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 552–564, 2012.

[14] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21, no. suppl_1, pp. i159–i168, 2005.

[15] C. Chira, J. Sedano, J. R. Villar, M. Camara, and C. Prieto, "Gene clustering for time-series microarray with production outputs," *Soft Computing*, vol. 20, no. 11, pp. 4301–4312, 2016.

[16] O. Cinar, O. Ilk, and C. Iyigun, "Clustering of short time-course gene expression data with dissimilar replicates," *Annals of Operations Research*, vol. 263, no. 1-2, pp. 405–428, 2018.

[17] N. Coffey, J. Hinde, and E. Holian, "Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data," *Computational Statistics & Data Analysis*, vol. 71, pp. 14–29, 2014.

[18] M. E. Futschik and B. Carlisle, "Noise-robust soft clustering of gene expression time-course data," *Journal of bioinformatics and computational biology*, vol. 3, no. 04, pp. 965–988, 2005.

[19] T. J. Hestilow and Y. Huang, "Clustering of gene expression data based on shape similarity," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, no. 1, p. 195712, 2009.

[20] C. S. Möller-Levet, F. Klawonn, K.-H. Cho, H. Yin, and O. Wolkenhauer, "Clustering of unevenly sampled gene expression time-series data," *Fuzzy sets and Systems*, vol. 152, no. 1, pp. 49–66, 2005.

[21] S. E. Baranzini, P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, *et al.*, "Transcription-based prediction of response to ifnβ using supervised computational methods," *PLoS Biol*, vol. 3, no. 1, p. e2, 2004.

[22] K. M. Borgwardt, S. Vishwanathan, and H.-P. Kriegel, "Class prediction from time series gene expression profiles using dynamical systems kernels," in *Biocomputing 2006*, pp. 547–558, World Scientific, 2006.

[23] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, *et al.*, "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, 2005.

[24] I. G. Costa, A. Schönhuth, C. Hafemeister, and A. Schliep, "Constrained mixture estimation for analysis and robust classification of clinical time series," *Bioinformatics*, vol. 25, no. 12, pp. i6–i14, 2009.

[25] K. H. Desai, C. S. Tan, J. T. Leek, R. V. Maier, R. G. Tompkins, J. D. Storey, *et al.*, "Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: a longitudinal clinical genomics study," *PLoS Med*, vol. 8, no. 9, p. e1001093, 2011.

[26] N. Den Teuling, S. Pauws, and E. van den Heuvel, "A comparison of methods for clustering longitudinal data with slowly changing trends," *Communications in Statistics-Simulation and Computation*, pp. 1–28, 2020.

[27] D. P. Martin and T. von Oertzen, "Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 22, no. 2, pp. 264–275, 2015.

[28] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data mining and knowledge discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[29] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438, 2000.

[30] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[31] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[32] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical association*, vol. 85, no. 411, pp. 633–639, 1990.

[33] N. Billor, A. S. Hadi, and P. F. Velleman, "Bacon: blocked adaptive computationally efficient outlier nominators," *Computational statistics & data analysis*, vol. 34, no. 3, pp. 279–298, 2000.

[34] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "A general framework for increasing the robustness of pca-based correlation clustering algorithms," in *International Conference on Scientific and Statistical Database Management*, pp. 418–435, Springer, 2008.

[35] N. Delannay, C. Archambeau, and M. Verleysen, "Improving the robustness to outliers of mixtures of probabilistic pcas," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 527–535, Springer, 2008.

[36] B. Abraham and G. E. Box, "Bayesian analysis of some outlier problems in time series," *Biometrika*, vol. 66, no. 2, pp. 229–236, 1979.

[37] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 3, pp. 350–363, 1972.

[38] A. M. Bianco, M. Garcia Ben, E. Martinez, and V. J. Yohai, "Outlier detection in regression models with arima errors using robust estimates," *Journal of Forecasting*, vol. 20, no. 8, pp. 565–579, 2001.

[39] R. S. Tsay, D. Pena, and A. E. Pankratz, "Outliers in multivariate time series," *Biometrika*, vol. 87, no. 4, pp. 789–804, 2000.

[40] S. J. Roberts, "Extreme value statistics for novelty detection in biomedical data processing," *IEE Proceedings-Science, Measurement and Technology*, vol. 147, no. 6, pp. 363–367, 2000.

[41] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: Finding unusual medical time series," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pp. 329–334, IEEE, 2005.

[42] J. Hardin and D. M. Rocke, "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator," *Computational Statistics & Data Analysis*, vol. 44, no. 4, pp. 625–638, 2004.

[43] A. M. Leroy and P. J. Rousseeuw, "Robust regression and outlier detection," *Wiley Series in Probability and Mathematical Statistics*, 1987.

[44] T. Zewotir and J. S. Galpin, "A unified approach on residuals, leverages and outliers in the linear mixed model," *Test*, vol. 16, no. 1, pp. 58–75, 2007.

[45] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[46] S. T. Wierzchoń and M. A. Kłopotek, *Modern algorithms of cluster analysis*. Springer, 2018.

[47] S. M. Grundy, J. I. Cleeman, S. R. Daniels, K. A. Donato, R. H. Eckel, B. A. Franklin, D. J. Gordon, R. M. Krauss, P. J. Savage, S. C. Smith Jr, *et al.*, "Diagnosis and management of the metabolic syndrome: an american heart association/national heart, lung, and blood institute scientific statement," *Circulation*, vol. 112, no. 17, pp. 2735–2752, 2005.

[48] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[49] L. Davies and U. Gather, "The identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782–792, 1993.

[50] G. Verbeke and G. Molenberghs, "A model for longitudinal data," *Linear mixed models for longitudinal data*, pp. 19–29, 2000.