Trabajo Fin de Máster

Comparación del análisis de datos funcionales con métodos tradicionales para analizar curvas de glucosa

Cintia Rúa Pérez

Máster en Técnicas Estadísticas

Curso 2014-2015

II

Propuesta de Trabajo Fin de Máster

Título en galego: Comparación da análise de datos funcionais con métodos tradicionais para a análise das curvas de glucosa
Título en español: Comparación del análisis de datos funcionales con métodos tradicionales para analizar curvas de glucosa
English title: Functional data analysis compared with traditional summary measures for analising glucose curves
Modalidad: Modalidad B
Autora: Cintia Rúa Pérez, Universidad de Vigo
Director: Manuel Febrero Bande, Universidad Santiago de Compostela;
Tutor: Francisco Gude Sampedro, Hospital Clínico Universitario de Santiago;
Breve resumen del trabajo: Nuestro objetivo fue conseguir información sobre las curvas de glucosa, comparándolo con la información aportada por índices resumen, y explorar su utilidad clínica a través de un marcador de inflamación. Entre la metodología empleada se encuentran el análisis de componentes principales, cluster y datos funcionales.
Recomendaciones:
Otras observaciones:

Don Manuel Febrero Bande, Catedrático de la Universidad Santiago de Compostela, don Francisco Gude Sampedro, Adjunto de la Unidad de Epidemiología Clínica de Hospital Clínico Universitario de Santiago, informan que el Trabajo Fin de Máster titulado

Comparación del análisis de datos funcionales con métodos tradicionales para analizar curvas de glucosa

fue realizado bajo su dirección por doña Cintia Rúa Pérez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 7 de julio de 2015.

El director

Don Manuel Febrero Bande

El tutor:

Don Francisco Gude Sampedro

La autora:

Doña Cintia Rúa Pérez

Agradecimientos

A mis tutores Francisco Gude y Manuel Febrero, que han dirigido este proyecto. Su apoyo y su capacidad para guiarme en este proceso han sido fundamentales.

A Manuela Alonso y Pablo Díaz, miembros de la Unidad de Epidemiología Clínica del Complejo Hospitalario Universitario, por su constante ayuda.

A todos los participantes en el proyecto A Estrada, especialmente a los que trabajaron en el centro de salud.

A mis amigos y a mi familia, que siempre han tenido una palabra de ánimo, especialmente a Juan Anido por su ayuda desinteresada en estos dos últimos años.

Índice general

Resumen

Resumen en español

Los niveles de glucosa son importantes en investigación y en clínica médica, y algunas veces se utiliza la monitorización continua de glucosa (CGM) en medidas repetidas varios días. Comúnmente, en la práctica se utilizan medidas resumen de las curvas de glucosas. Sin embargo, diferentes curvas glucémicas pueden dar medidas resumen similar, y se puede perder información con relevancia tanto clínica como fisiológica. Nuestro objetivo fue conseguir información inherente al perfil de las curvas de glucosa, comparándolo con la información aportada por los índices resumen, y explorar su utilidad clínica a través de un marcador de inflamación. Se recogieron 581 medidas de CGM en un estudio de base poblacional. Para el análisis de datos funcionales (FDA) transformamos estas medidas en curvas de glucosa suavizadas. Se ha utilizado un análisis de componentes principales para resumir los datos funcionales de la curva de glucosa. Las componentes principales de las derivadas de los datos funcionales se han comparado con las medidas tradicionales de la variabilidad glucémica. La utilidad clínica de los FDA fue estudiada a través de modelos de regresión entre las medidas de glucosa y la concentración de una citoquina pro-inflamatorioa (TNF-α).

Con el uso de las aproximaciones tradicionales no encontramos asociaciones entre la variabilidad de glucosa y el TNF-α. Sin embargo, cuando utilizamos FDA encontramos una asociación fuerte entre la variabilidad de la glucosa y en TNF-α. El análisis mediante FDA de las curvas de glucosa consiguió información que las medidas resumen tradicionales no fueron capaces de identificar.

English abstract

Glucose levels are important measures in medical care and research, and sometimes are obtained from continuous glucose monitoring (CGM) with repeated measurements over several days. It is common practice to use summary measures of glucose curves. However, different glucose curves can yield similar summary measures, and information of physiological or clinical interest may be lost. Our aim was to extract information inherent in the shape of glucose curves, compare it with the information from summary measures of glucose variability, and explore the clinical usefulness of such information by means of an inflammatory marker. In a population-based study CGM measurements were recorded for 581 individuals. For each one, these measurements were transformed into smooth glucose curves by functional data analysis. The essential modes of temporal variation between glucose curves were extracted by functional principal component analysis. The resultant functional principal component scores in its derivatives were compared with commonly used summary measures of glucose variability. Clinical usefulness of FDA was explored by regression analyses of glucose measurements on the inflammatory cytokine concentrations (TNF-α).

When using traditional approaches we did not find association between glucose variability and TNF-α levels. However, by using functional data analysis we found a strong association between glucose variability and TNF-α. FDA of glucose curves extracted information that was not identified by commonly used summary measures.

Prefacio

Desde el punto de vista clínico, la diabetes mellitus (DM) es un grupo heterogéneo de procesos cuya característica común es la hiperglucemia, como resultado de defectos en la secreción de la insulina, habitualmente por destrucción de las células beta pancreáticas de origen auto-inmunitario en la DM tipo 1, o por una progresiva resistencia a la acción periférica de la insulina, con o sin déficit asociado en la secreción, en la DM tipo 2. En ambos casos, el desarrollo de la enfermedad se atribuye a una combinación de factores genéticos predisponentes y una serie de factores ambientales que actuarían como desencadenantes.

La DM se puede considerar como uno de los principales problemas de salud mundial, entre otras razones por su elevada prevalencia, su elevado coste económico y las complicaciones generadas. Las nuevas estimaciones muestran una preocupante tendencia creciente de diabetes en personas cada vez más jóvenes, junto con el aumento de la diabetes tipo 2 en poblaciones de más edad.

Tanto en la diabetes tipo 1 como tipo 2, dos grandes estudios clínicos prospectivos, el *Diabetes Control and Complications Trial* (DCCT) y el *UK Prospective Diabetes Study* (UKPDS) han mostrado una fuerte asociación entre los niveles medios de glucemia (medidos a través de la hemoglobina glicada, A1C) y las complicaciones diabéticas. Sin embargo, en los últimos años se ha planteado la posibilidad de que la variabilidad de la glucemia, y no solamente la hiperglucemia, puede contribuir al desarrollo de complicaciones. Disponemos además, de resultados que soportan la idea de que las fluctuaciones agudas en la glucemia pueden producir alteraciones en la homeostasis, tales como la disfunción endotelial y el aumento de la actividad inflamatoria.

Sin embargo, los primeros resultados del DCCT, mostraron que la variabilidad de la glucosa no es un predictor de complicaciones. Posteriormente, los mismos autores y en base al mismo estudio, comunican que la variabilidad de la A1C más que la variabilidad de la glucosa, es predictor de complicaciones microvasculares. Debemos mencionar que la metodología de estos estudios ha sido duramente criticada; sin embargo, también se ha visto que la variabilidad de algunos índices del control de la glucemia pueden ser perjudiciales en la aparición de complicaciones en los individuos con diabetes tipo 1.

De cara a identificar una relación entre variabilidad de la glucosa y el riesgo de complicaciones vasculares, el DCCT nos provee de una base de datos muy amplia sobre la que testar estas hipótesis. Entre sus hallazgos se ha encontrado que la variabilidad de la glucosa, definida estadísticamente de diferentes formas, no añadió información alguna sobre la glucosa media para predecir el desarrollo de retinopatía o nefropatía. Tampoco ha mostrado que la variabilidad de la glucemia contribuya al desarrollo de neuropatía.

En pacientes con diabetes tipo 2, los resultados son menos consistentes. Muggeo et al. (1997) encontraron en pacientes ancianos con diabetes que la mortalidad, tanto cardiovascular como debida a cualquier causa, se asociaba fundamentalmente a la variabilidad de la glucosa en ayunas más que a sus valores absolutos.

Estudios realizados en animales de laboratorio también han mostrado un efecto deletéreo de las fluctuaciones de la glucosa sobre las células renales mesangiales del túbulo intersticial y células β-pancreáticas. Resulta interesante que cultivos de fibroblastos procedentes de la corteza renal humana han mostrado un aumento en la expresión de marcadores de fibrogénesis que es dependiente de los picos de glucosa pero independiente de la cantidad total de glucosa a la que han sido expuestas estas células.

También se ha reportado en sujetos normales, que las fluctuaciones repetidas de la glucosa producen incremento en los niveles circulantes de citoquinas inflamatorias cuando se compara con niveles elevados mantenidos de glucosa. El factor de necrosis tumoral α (TNF) es una de las principales citoquinas mediadoras en la respuesta inflamatoria e inmune. En relación con el metabolismos de la glucosa, se ha publicado que el TNF-α está sobreexpresado en tejido adiposo blanco, en estados de obesidad y resistencia insulínica.

Mientras que existen estudios que muestran que la variabilidad glucémica puede contribuir a un aumento en la producción de radicales libres, existen otros muchos que no encuentran esta asociación. Así mientras Monnier et al. (2006), encontraron una fuerte relación entre la variabilidad de la glucosa y un marcador de estrés oxidativo (8-iso PGF2α) en diabéticos tipo 2, Wentholt et al. (2008) en un estudio similar pero con más participantes, no encontraron tal asociación en diabéticos tipo 1 ni 2.

Por tanto, el papel que desempeñan la variabilidad glucémica y el estrés oxidativo en el desarrollo de complicaciones está siendo actualmente motivo de interés. De esta forma se nos presenta un nuevo desafío: cómo medir la variabilidad de la glucemia. Hasta la aparición de los dispositivos de monitorización continua de la glucosa (CGM), la variabilidad glucémica era calculada con los datos provenientes de perfiles elaborados a partir de 7 controles de glucemia capilar. A partir de la disponibilidad de los aparatos de CGM, la ingente cantidad de datos proporcionados por los nuevos sistemas de monitorización han propiciado la aparición de multitud de modelos para el análisis de la variabilidad, tanto intradía como entre días.

Molnar et al. (1970) propusieron la amplitud media de la excursiones glucémicas (MAGE) para cuantificar la VG han aparecido una serie de medidas tales como el desvío estándar (SD), el área bajo la curva (AUC), etc. Todas estas medidas tienen como característica común que se tratan de índices resumen.

Sin embargo trayectorias diferentes en el perfil de la glucosa pueden dar similares mediciones en cuanto a estas medidas resumen. De esta forma, se puede perder una información que puede ser de interés desde el punto de vista fisiopatológico y clínico. En este sentido, el análisis de datos funcionales (FDA) constituye una serie de técnicas estadísticas desarrollada para analizar curvas.

El objetivo de este trabajo consiste en la utilización de métodos de datos funcionales en comparación con las medidas resumen de variabilidad glucémica para analizar el efecto de las excursiones de la glucosa sobre factores inflamatorios relacionados con la diabetes. Como medida de inflamación y como variable respuesta se ha tomado el TNF. En las variables predictoras se han utilizado los índices de variabilidad de forma comparativa con los perfiles de glucosa como dato funcional junto a otras covariables que pueden ser potenciales confusores en esta relación.

El proyecto se estructura de la siguiente forma: en el primer capítulo se describe el proyecto A Estrada, que proporciona los datos de este estudio; en el segundo, se definen los índices de variabilidad glucémica más utilizados hasta la fecha, realizando un análisis *cluster* para escoger aquellos que van a ser evaluados de forma comparativa; en el tercer capítulo, se realiza un análisis funcional de las curvas de glucosa; y finalmente en el cuarto, se comparan los resultados de ambas técnicas.

Capítulo 1

Proyecto A Estrada

El Proyecto A Estrada se basa en un estudio de base poblacional, en una muestra representativa de la población general adulta, con un amplio tamaño muestral, extensa fenotipación y documentación individual y con almacenamiento reglado de muestras biológicas (suero, orina y sangre total).

Ideado para conocer, desde una perspectiva poblacional, cuáles son los determinantes de la inflamación y de la glicación, cuenta con ayudas de la Xunta de Galicia (10CSA918028PR: Marcadores de inflamación y su relación con enfermedades frecuentes en la población general adulta. Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica. 2010-12), del Fondo de Investigaciones Sanitarias (PI11/02219: Niveles de hemoglobina glicosilada y gap de glicación en relación con estilos de vida y las enfermedades prevalentes en la población general adulta. Fondo de Investigaciones Sanitarias. 2012 - 2014; RD12/0005/0007: Red de investigación en actividades preventivas y promoción de la salud en Atención Primaria (REDIAP). Fondo de Investigaciones Sanitarias. 2013 - 2015) y de Medtronic Inc.

Dirigido por los Dres Arturo González-Quintela (Medicina Interna) y Francisco Gude (Epidemiología Clínica), cuenta con la participación activa de un amplio grupo de profesionales pertenecientes a diferentes disciplinas: atención primaria, alergología, biología, bioquímica, enfermería, endocrinología, cardiología, odontología, psiquiatría, psicología, nutrición y bioestadística.

1.1. Localización

El municipio de A Estrada (Pontevedra) cuenta con una población adulta mayor de 18 años de 18744 habitantes, sobre una superficie de 282 km^2 de superficie. Cerca de un cuarto de la población vive en la ciudad y el resto en un entorno rural.

1.2. Selección de los participantes

A partir de la tarjeta sanitaria, que cubre más del 95 % de la población, se tomó una muestra aleatoria de la población. Esta tarjeta contiene el nombre, la fecha de nacimiento y la dirección de toda persona que es atendida en atención primaria.

La selección de la muestra fue estratificada en grupos de edad: 18 a 29 años, de 30 a 39, de 40 a 49, de 50 a 59, 60 a 69, 70 a 79, y mayores de 80 años. Se generó una muestra aleatoria de 500 personas

Grupos de edad (años)	Población base (n)	Selección aleatoria (n)	Válidas para participar (n)	Participa en el estudio (n)	Tasa de participación (%)
18-29	2483	500	315	198	62.9
30-39	3365	500	330	235	71.2
40-49	3148	500	382	275	72
50-59	2625	500	372	252	67.7
60-69	2585	500	381	279	73.2
70-79	2345	500	312	200	64.1
80+	1923	500	138	77	55.8
Total	18474	3500	2230	1516	67.9

Cuadro 1.1: Resumen de la participación en el estudio.

de cada grupo de edad, obteniendo un total de 3500 individuos. De ellos, se excluyeron 428 por finalización del estudio sin intento de reclutamiento, 84 habían fallecido, 211 no respondieron, 134 habían cambiado su domicilio fuera del municipio, y 19 no tenían asistencia sanitaria. Además, 394 personas no cumplían los criterios de inclusión por presentar demencia, retraso mental, enfermedad cerebrovascular grave, cáncer, enfermedad terminal, o incapacidad para comunicarse. De los restantes elegibles, un total de 1516 personas (55 % mujeres, 45 % hombres) accedieron a participar (tasa de participación: 67.9 %), siendo la tasa de participación inferior en hombres que en mujeres (65 % vs 71 %). La tasa de participación también fue inferior en los más mayores y en los jóvenes, no encontrándose diferencias en cuanto el lugar de residencia (rural/urbano) entre aquellos que aceptaron y no aceptaron participar.

1.3. Determinaciones

Desde Noviembre de 2012 a Marzo de 2015, todos los participantes acudieron a una consulta del centro de salud de A Estrada para la realización de una entrevista clínica y determinaciones que incluían:

(a) Cuestionario estructurado con datos demográficos y antropométricos.

(b) Estilos de vida con registro de la actividad física, ingesta dietética, consumo de tabaco y alcohol.

(c) Una batería de test psicológicos.

(d) Examen periodontal.

(e) Pruebas alérgicas

(f) Muestra sanguínea.

Además, un subgrupo participó en la parte de glicación del proyecto que incluía la monitorización continua de la glucosa (CGM). Constituyeron criterios de exclusión adicionales la incapacidad para cumplimentar debidamente el protocolo, comer fuera habitualmente, alergia a los adhesivos y cualquier condición médica que pudiera afectar al buen funcionamiento del dispositivo ($n = 451$). De las 1065 elegibles, aceptaron participar 622, los cuales completaron un período de CGM de 6 días. Una descripción más detallada de todo el proceso puede verse en la Figura 1.1 y en la Tabla 1.1.

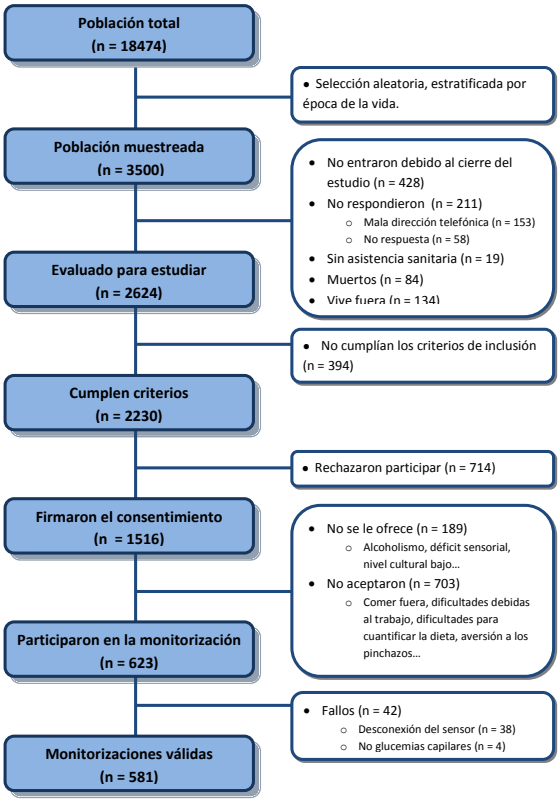

Figura 1.1: Diagrama de flujo

1.4. Características demográficas y clínicas

En el Cuadro 1.2 se muestra un resumen descriptivo de los participantes del estudio. En la primera columna aparecen los participantes en el estudio de inflamación, y en la segunda las características de los participantes del estudio de CGM.

Como puede observarse, los participantes en el estudio CGM eran ligeramente más jóvenes, hubo mayor participación de mujeres, un nivel más elevado de estudios y pertenecientes al hábitat urbano.

1.5. Monitorización continua de la glucosa

La glucosa es un carbohidrato simple o monosacárido que se ingiere en la dieta. Tras su absorción y paso al torrente sanguíneo sufre una serie de transformaciones metabólicas en las que se liberan dióxido de carbono, agua y algunos compuestos de nitrógeno. Este proceso de oxidación libera energía que puede ser utilizada por las células, lo que convierte a la glucosa en la principal fuente de energía para el ser humano.

Los niveles de glucosa no se mantienen constantes permanentemente sino que fluctúan en función de la ingesta y del gasto metabólico. La unidad de medida de la glucosa en sangre es milimoles por litro (mmol/L) o miligramos por decilitro (mg/dL); en lo que sigue, si no se especifica, utilizaremos la

	Participantes ($n = 1516$)	CGM ($n = 622$)	CGM exitoso ($n = 583$)	CGM fallido ($n = 39$)
Hombres, n (%)	678(44.7)	238(38.3)	221(37.9)	17(43.6)
Edad, años	53 ± 18	48 ± 15	48 ± 15	51 ± 15
Viven solos, n (%)	124(8.2)	39(6.3)	35(6.0)	4(10.3)
Urbano, n (%)	534(35.2)	243(39.1)	227(38.9)	16(41.0)
Animales de labor	754(49.7)	288(46.3)	267(45.8)	21(53.8)
Nivel de estudios				
0	339(22.4)	79(12.7)	73(12.5)	6(15.4)
I	525(34.6)	216(34.7)	193(33.1)	23(59.0)
II	124(8.2)	50(8.0)	48(8.2)	2(5.1)
III	145(9.6)	71(11.4)	69(11.8)	2(5.1)
IV	383(25.3)	206(33.1)	200(34.3)	6(15.4)
Estatus laboral				
Trabajador	604(39.8)	291(46.8)	277(47.5)	14(35.9)
Desempleado	237(15.6)	119(19.1)	108(18.6)	11(28.2)
Ama de casa	112(7.4)	55(8.8)	52(8.9)	3(7.7)
Jubilado	481(31.7)	118(19.1)	108(18.6)	10(25.6.)
Otros	80(5.3)4	38(6.1)	37(6.3)	1(2.6)
Auto-percepción salud				
Excelente	221(14.5)	94(15.1)	89(15.3)	5(12.8)
Buena	755(49.8)	339(54.5)	318(54.5)	21(53.8)
Mala	539(35.6)	189(30.4)	176(30.2)	13(33.3)
Tabaco				
No fumadores	884(58.3)	351(56.5)	335(57.4)	16(41.1)
Ex-fumadores	334(22.0)	147(23.6)	128(22.0)	19(48.7)
Fumadores	298(19.7)	124(19.9)	124(19.9)	19(48.7)
Alcohol				
Abstemio	264(17.4)	105(16.9)	99(17.0)	6(15.4)
Ocasional	630(41.6)	285(45.8)	271(46.5)	14(35.9)
Habitual	620(40.9)	231(37.1)	212(36.4)	19(48.7)
Actividad física				
Baja	620(41.1)	250(40.2)	235(40.7)	15(38.5)
Moderada	523(34.7)	217(34.9)	202(34.9)	15(38.5)
Alta	366(24.2)	150(24.1)	141(23.4)	9(23.0)
IMC, kg/m^2	28.2 ± 5.1	28.3 ± 5.2	28.2 ± 5.2	29.0 ± 4.9
Diabetes, n (%)	187(12.3)	71(11.4)	68(11.7)	3(7.7)
Glucosa en ayunas	95 ± 23	93 ± 22	93 ± 22	96 ± 26
A1c, %	5.7 ± 0.8	5.6 ± 0.8	5.6 ± 0.8	5.4 ± 0.7

Cuadro 1.2: Resumen descriptivo de los sujetos que participaron en el estudio. 0 ≡ no sabe leer ni escribir o no completó Estudios de Graduado Escolar (EGB). I≡ EGB y Educación Secundaria Obligatoria (ESO). II≡ Bachiller. III≡ Formación Profesional. IV≡ Estudios Universitarios.

segunda medida. Los valores óptimos oscilan entre 70-80 mg/dL y 120-130 mg/dL. Cuando los mecanismos reguladores del organismo comienzan a fallar, estos niveles de glucosa se alteran pudiendo ser más altos o más bajos de lo recomendable y dando lugar a una serie de complicaciones clínicas propias de la Diabetes Mellitus.

Las diferentes técnicas de determinación han ido evolucionando a lo largo del tiempo, desde la determinación en orina inicial hasta los novedosos sistemas de monitorización continua de la glucosa. Esta evolución perseguía tres objetivos fundamentales en cualquier técnica de medición: comodidad,

fiabilidad y precisión.

Actualmente, las técnicas predominantes para la determinación de los niveles de glucosa son: la determinación de la glucosa en sangre venosa (ya sea en ayunas o tras una sobrecarga oral de glucosa), la determinación de la glucemia capilar y la monitorización continua de la glucemia intersticial, siendo estas dos últimas las utilizadas en en este proyecto.

Los dispositivos utilizados para la medición de la **glucemia capilar** (SMBG) reciben el nombre de glucómetros. Estos dispositivos se usan conjuntamente con unas tiras reactivas, que deben ser introducidas en el glucómetro, y con unas lancetas de punción, destinadas a atravesar la piel permitiendo conseguir una pequeña muestra de sangre capilar destinada a ponerse en contacto con la tira reactiva insertada en el glucómetro.

A finales de 1980, con la introducción de la automonitorización de la glucosa en sangre, se ha realizado un gran progreso para la medida de la glucosa. Más tarde, a principios de este siglo, se introdujo el sistema de ***Continuous Glucose Monitoring*** (CGM), que proporciona una imagen completa de las fluctuaciones de glucosa y una mayor información sobre el control glucémico.

A lo largo de los últimos años aparecieron varios dispositivos de CGM. De forma breve, actualmente contamos con dos tipos diferentes de sistemas de monitorización: en tiempo real y retrospectivos. En nuestro estudio hemos utilizado el iPro®2, diseñado por Medtronic, el cual es considerado como un sistema de monitorización retrospectiva (profesional). A diferencia de los sistemas en tiempo real, el que nosotros hemos utilizado presenta algunas ventajas para su uso en estudios clínicos y epidemiológicos:

(1) Uso más fácil para los pacientes, ya que no requiere entrenamiento previo ni puesta a punto horaria.

(2) Dado que el paciente no recibe información de sus niveles de glucosa, éste no modifica su conducta en cuanto a introducir cambios en la dieta, actividad física o medicación.

(3) Desde un punto de vista técnico los sistemas retrospectivos son más fiables que los de tiempo real.

Los avances en los sistemas para la medición de la glucosa han ayudado a conocer mejor el comportamiento de la glucosa, no solamente para establecer sus niveles medios en ayunas sino también para conocer sus fluctuaciones a lo largo del día (**variabilidad glucémica** (VG)). Aspecto que abordaremos con mayor profundidad en el segundo capítulo del presente proyecto.

El sensor de la CGM se inserta en la región abdominal de cada paciente durante su visita inicial al centro de salud y, con el objetivo de calibrar el dispositivo CGM, se solicita a los participantes realizar al menos tres determinaciones diarias de glucemia capilar. En la Figura 1.3 se muestra la imagen del glucómetro utilizado para determinar las glucemias capilares y en la Figura 1.2 una imagen de uno de los dispostivos de CGM utilizados en el proyecto.

El dispositivo de CGM almacena los niveles de glucosa que se encuentran en un rango de 40 y 400 mg/dL durante las 24 horas de los 6 días cada 5 minutos. Durante la monitorización, los participantes debían seguir con su ritmo de vida normal, anotando la hora y la ingesta de alimentos, y la actividad física que realizaban cada día.

Figura 1.2: Dispositivo de monitorización continua de la glucosa: iPro Medtronic.

Figura 1.3: Medidor de la glucosa capilar: One TouchVerio Pro; LifeScan, Milpitas, CA, USA.

1.6. Factor de necrosis tumoral α

El TNF-α es una citoquina pro-inflamatoria que es producida por varios tipos de células, mayoritariamente macrófagos y linfocitos. También, puede ser producida en menor medida por los adipocitos del tejido adiposo.

Distintos estudios han encontrado un aumento en los niveles de TNF en pacientes con diabetes tipo 1, en individuos con resistencia a la insulina y en obsesos. Ceriello et al. (2008) encontraron que el aumento de la glucosa produce un aumento en los niveles séricos del TNF-α tanto en individuos sanos como en individuos con el síndrome de la resistencia a la insulina.

1.7. Consideraciones éticas

El estudio fue llevado a cabo de acuerdo con los principios de la Declaración de Helsinki y con la legislación vigente. Fue aprobado por el Comité Ético de Investigación Clínica de Galicia, Santiago de Compostela, España (referencias 2010-315 y 2012-025).

Capítulo 2

Evaluación de la *performance* de la CGM

En cualquier tipo de medición clínica y de laboratorio resulta crucial la obtención de mediciones precisas y fiables. Solamente aquellas lecturas del sensor que reflejen de forma precisa los niveles de glucosa pueden resultar de interés tanto para actividades de investigación como para que los pacientes con diabetes tengan un mejor control de su enfermedad.

El objetivo de este capítulo es el de evaluar la *performance* de los dispositivos de monitorización para obtener niveles fiables de glucosa cuando lo comparamos con las medidas de glucemia capilar.

Ocasionalmente para una resolución aproximada al problema de comparación de dos métodos de medición se calculaban los coeficientes de correlación, pero esta solución podría ser engañosa. Por ello, se ha preferido utilizar las siguientes aproximaciones: método de Bland and Altman (1986), el porcentaje de la media absoluta relativa de las diferencias Castle et al. (2010), y el gráfico de error de Clarke et al. (1987).

2.1. Método Bland - Altman

El método de ***Bland-Altman*** considera las diferencias entre las mediciones de los dos métodos y calcula los conocidos **límites de concordancia**; límites de predicción para presentar la diferencia entre pares de futuras mediciones.

Trata de graficar las diferencias entre cada par de de valores de los dos métodos de medida contra la media de cada par. Los datos se distribuirán sobre la recta correspondiente a la diferencia cero de las medidas. Estos gráficos proporcionan información sobre la precisión del sensor en el rango de valores absolutos de glucosa.

Bland-Altman propone el modelo matemático siguiente:

$$Y_{mi} = \alpha_m + \varepsilon_{mi}, \quad \varepsilon_{mi} \sim N\left(0, \sigma_m^2\right),$$

dónde Y_{mi} es la variable respuesta que representa el valor de la respuesta en la i-ésima observación del m tratamiento. En lo que sigue denotaremos Y_{mi} para referirnos a la variable y la notación y_{mi} para referirnos a una observación concreta. α_m es la respuesta real en el m tratamiento, es decir, la que se obtendría siempre con el i-ésimo tratamiento si se ejecutase el experimento en, exactamente, las mismas condiciones. Por último, ε_{mi} es el error que agrupa la contribución de las fuentes de variaciones

menores y no planificadas.

Para examinar las diferencias entre tratamientos equivaldría a examinar las diferencias entre los parámetros α_m. Las diferencias $d = y_{1i} - y_{2i}$ tienen una varianza de $\sigma_1 + \sigma_2$ y el intervalo de predicción para la diferencia entre los dos métodos es:

$$\alpha_1 - \alpha_2 \pm 1.96\sqrt{\sigma_1 + \sigma_2}.$$

En la práctica, el término $\alpha_1 - \alpha_2$ es estimado por la media de las diferencias y el último término se calcula como la desviación estándar de las diferencias, por tanto:

$$\bar{d} \pm 1.96 sd(d).$$

Sólo quedaría estimar los componentes de la varianza del modelo mixto lineal, que se realizará con la ayuda del paquete `MethComp` del *software* estadístico R.

Una vez calculados los límites de concordancia, se pueden presentar los gráficos de Bland-Altman. Estos gráficos son una técnica muy útil para el estudio de dos métodos de medidas repetidas. Simplemente es un gráfico de dispersión donde la diferencia entre las medidas emparejadas se asocia a su valor medio. A partir de éste, se podrán detectar valores extremos.

En vista de la Figura 2.1, el primer día tiene mayor dispersión en los pares de puntos, la media de las diferencias es mayor y por tanto, indica menor precisión frente al resto de los días, posiblemente por efectos relacionados con el arranque del sensor. Según este análisis puede ser conveniente que, en lo que resta de trabajo, eliminaremos el primer día.

2.2. Media absoluta y relativa de las diferencias

La siguiente medida que utilizaremos para evaluar la *performance* será la ***Mean Absolute Relative Difference*** (MARD) utilizando la glucosa capilar como valor de referencia. El MARD es el porcentaje de la diferencia media absoluta de las lecturas de la monitorización y los valores de la glucosa de referencia, SMBG, para todos los pares de puntos. Los valores del MARD deben ser inferiores al 14 % para que el estudio sea válido y mayores de 18 % son considerados como una mala aproximación.

En este estudio el MARD resultó del 7.95 % por tanto, se podría asegurar que la monitorización es buena. Para un estudio más exhaustivo calcularemos el MARD día a día.

%	Día 1	Día 2	Día 3	Día 4	Día 5	Día 6
Desayuno	12.90	8.07	7.05	7.06	7.16	7.36
Comida	11.19	7.49	7.44	7.18	7.68	6.53
Cena	12.19	7.20	6.63	7.15	7.21	6.02
Total	12.07	7.59	7.04	7.13	7.34	6.64

Cuadro 2.1: Porcentaje de MARD para las diferentes comidas en 6 días.

En vista del Cuadro 2.1, los valores del MARD son siempre inferiores al 14 % por lo que, como se ha comentado antes, podremos suponer que los valores de la glucosa dados por el GCM son buenos. Cabe destacar que el primer día tiene un porcentaje ligeramente mayor que el resto, al igual que en el análisis de Bland y Altman, se asocia a la calibración del sensor.

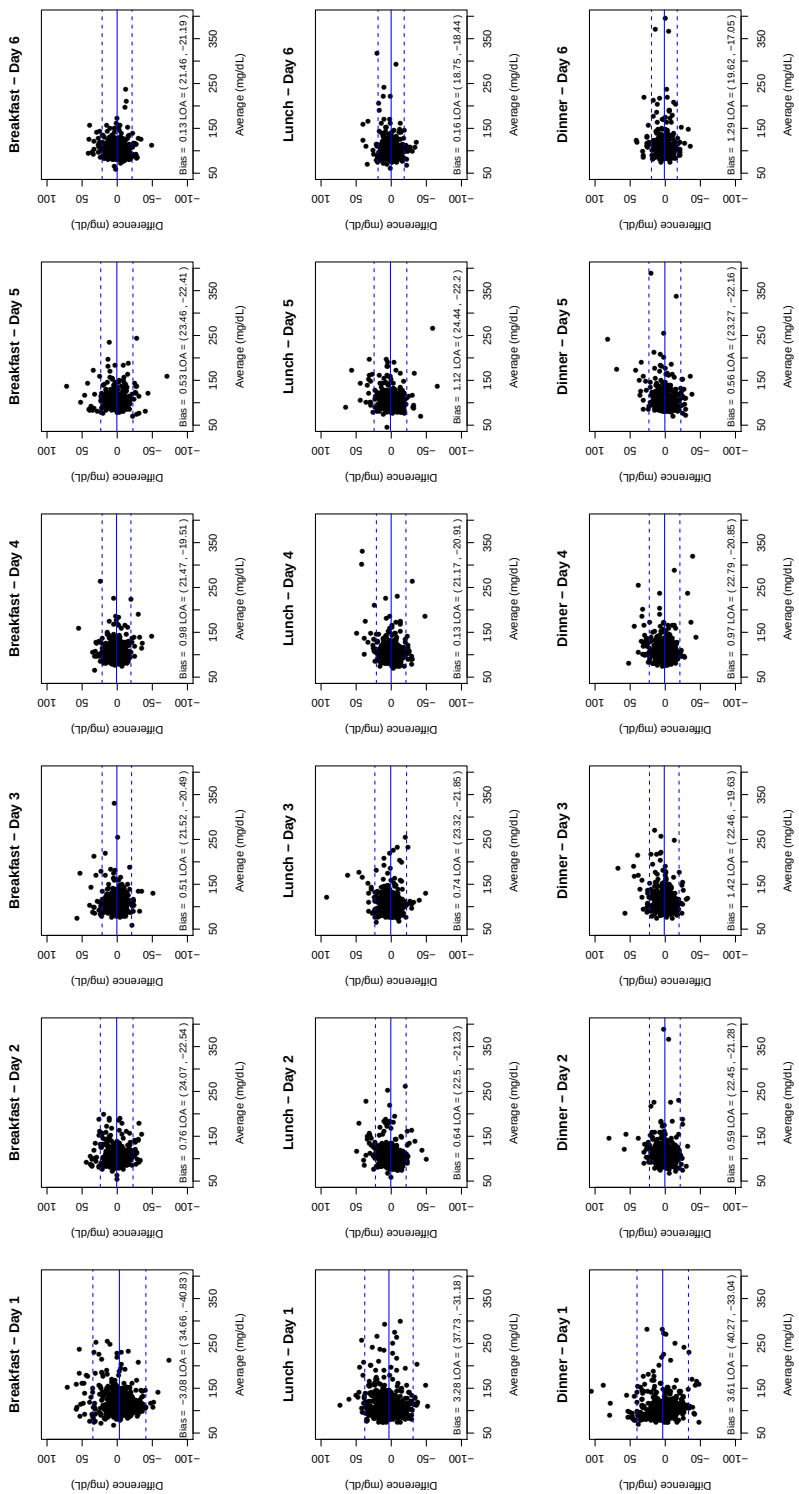

Figura 2.1: Gráfico de Bland y Altman para los diferentes días de monitorización. El eje x representa la media de las medidas de glucosa por ambos métodos, el eje y la diferencia de los dos métodos de medición, expresado en porcentaje. La línea azul más gruesa es el sesgo (media de las diferencias de las glucosas) y las líneas azules más finas representan los límites de concordancia.

2.3. Método de Clarke

El ***Clarke Error Grid*** (EGA) define como el eje X el valor de la glucosa de referencia y el eje Y como el valor de la glucosa generado por el sistema de monitorización. La diagonal representa el perfecto acuerdo entre los dos métodos, los pares de puntos que se encuentran por encima y por debajo de la diagonal son las sobreestimaciones y subestimaciones, respectivamente.

Divide el área de la gráfica en cinco zonas, el tamaño y forma de cada una de ellas depende de la relevancia clínica de los pares de puntos. Para cada par de lecturas, $(\text{SMBG}(t_1), \text{SMBG}(t_2))$, tomadas en los tiempos t_1 y t_2, la tasa de cambio (mg/(dl.min)) se calcula como la diferencia entre los pares de puntos divida por el tiempo transcurrido, es decir, $(\text{SMBG}(t_2) - \text{SMBG}(t_1))/(t_2 - t_1)$, análogo para los datos de CGM.

Las zonas del gráfico se definen en función de la tasa de referencia del BG. A continuación presentamos el procedimiento de la división de las diferentes zonas.

1. La diagonal principal significa un ajuste perfecto y si la tasa de SMBG está dentro de -1 a 1 mg/(dl.min) (no hay cambios significativos), se considera los límites de la zona A.

2. Si falla en una tasa de -2 a -1 mg/(dl.min), los límites superiores de las zonas A, B y D son expandidas en 20 mg/dl.

3. Si la tasa aumenta de $1-2$ mg/(dl.min), los límites inferiores de las zonas A, B y D se extienden por 10 mg/dl. Si aumenta más rápido que 2 mg/(dl.min) los límites inferiores de las zonas A, B y D se expanden por 20 mg/dl.

4. La zona de C se divide en sobreestimación (superior) y la subestimación (menor) del tipo de referencia. La tasa de referencia es de -1 a 1 mg/(dl.min), sin mostrar fluctuación de BG significativo. Sin embargo, el sensor muestra una fluctuación significativa BG, lo que podría conducir a un tratamiento excesivo.

5. En la zona de E, las lecturas de la pantalla del sensor son opuestas a la tasa de referencia. La parte alta de esta zona es una disminución real BG, se estima como aumento BG, mientras en la parte más baja, que es un aumento real BG, se interpreta como BG baja.

Según la zona en la que se encuentren los pares de puntos, la relevancia clínica es diferente:

- Zona A, mediciones de glucosa que se desvían de la referencia no más del 20 % o los niveles de glucosa menores a 70 mg/dl.
- Zona B, medidas que difieren de la de referencia por más del 20 %, pero clínicamente aceptable.
- Zona C, son medidas que pueden ocasionar sobre-correcciones de los valores de glucosa aceptables, la medida puede conllevar a cambios innecesarios en el tratamiento con insulina.
- Zona D, potencialmente peligrosa, eventos de hipo- o hiperglucemias podrían perderse.
- Zona E, niveles que seguramente conduzcan a una decisión contraria a la que se requiere, por ejemplo, tratamiento de hipoglucemias en lugar de hiperglucemias.

Cuando una cantidad superior al 95 % de las mediciones apareadas se trazan en las zonas A y B se considera, generalmente, un sistema preciso. Como se trata de una gráfica que tiene relevancia clínica, para el estudio, diferenciaremos pacientes con diabetes y sin diabetes.

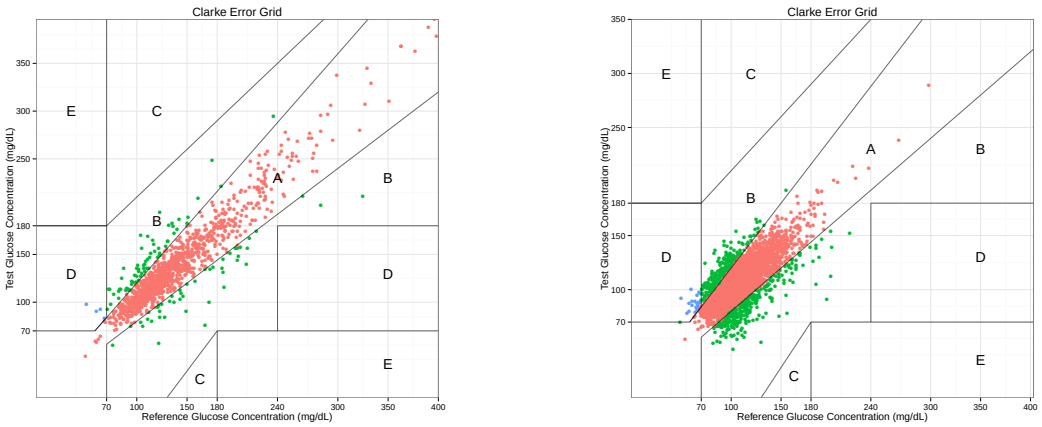

Figura 2.2: Gráfico EGA, a la izquierda para los pacientes diabéticos y a la derecha para pacientes sin diabetes. Ambos divididos en diferentes zonas (A, B, C, D y E) según la relevancia clínica de cada par de puntos.

%	Zona A	Zona B	Zona C	Zona D	Zona E
Diabéticos	85.03	14.65	0	0.32	0
No diabéticos	87.96	11.83	0	0.21	0

Cuadro 2.2: Porcentaje de pares de puntos que se encuentran en las diferentes zonas según el gráfico de EGA.

Para la obtención de los gráficos Clarke utilizaremos funciones implantadas en el paquete `ega` de R. En nuestro caso, obtenemos que la mayor parte de pares de puntos se encuentran en las zonas A y B (Figura 2.2), exactamente un 99.68 % para diabéticos y 99.79 % para pacientes sin diabetes (Cuadro 2.2). Por tanto, podemos considerar que nuestro sistema es muy preciso.

2.4. Conclusión de la evaluación

En este estudio poblacional cabe destacar el alto grado de aproximación entre las mediciones del glucómetro y las del sensor de monitorización. Los resultados generales y, en concreto, los del día a día, los MARD obtenidos son comparables, e incluso mejores, que los que informan en otros estudios de CGM (Clarke et al. (1987) y Kovatchev et al. (2008)).

Pero, el sensor fue menos preciso durante el primer día de implantación del dispositivo, probablemente debido a la inicialización del sensor y efectos relacionados con el arranque del dispositivo por lo que no se tendrá en consideración en futuros análisis. Esto se puede visualizar en los gráficos de Bland-Altman (véase Figura 2.1) que indican menor precisión durante el primer día frente al resto de días.

Además, el 99.2 % de las mediciones recogidas fueron clínicamente aceptables (zonas A y B de la cuadrícula de Clarke, Figura 2.2) y la mayoría de estos valores (87.6 %) se consideraron clínicamente exactos (dentro de la zona A).

Capítulo 3

Variabilidad glucémica

3.1. Preliminares

La oscilación de los niveles de glucosa en sangre se conoce como variabilidad glucémica (VG). Hay muchos factores que contribuyen al desarrollo de la VG, como el momento del día, el estado de salud y el grado de estrés.

La monitorización ha demostrado la inestabilidad en el control glucémico, de hecho el estado de las terapias para pacientes con diabetes se alejan bastante del ideal. Pero por otro lado, el *Diabetes Control and Complications Trial* (DCCT) aseguró que la hemoglobina glucosilada (HbA1c) no es la medida más completa para la glucosa, ya que el riesgo puede ser únicamente debido a las fluctuaciones glucémicas. Cabe destacar que, los pacientes con mayor variabilidad sufren mayores hiperglucemias, pero también hipoglucemias.

La monitorización continua presenta la glucosa en sangre (BG) como una serie temporal

$$\{BG(t_i) \mid i = 1, 2, \ldots, N\},$$

donde t es el momento de recogida y N el número total de mediciones.

En este estudio, se coloca el sensor durante seis días obteniendo 288 mediciones de glucosa por día. Como se ha visto en el capítulo anterior, desecharemos el primer día obteniendo, finalmente, 1440 mediciones de glucosa.

La ingente cantidad de datos proporcionados por los nuevos sistemas de monitorización (frente a los otros métodos de medida), control de los niveles de glucemia y la falta de un consenso sobre cuál es la manera adecuada de tratar esos datos, han propiciado la aparición de una multitud de índices para el análisis de la variabilidad, tanto intradía como entre días.

3.2. Índices de variabilidad

Como se ha mencionado anteriormente, existen una multitud de índices que intentan cuantificar la variabilidad glucémica. Sin embargo, deberíamos tener en cuenta algunas consideraciones antes de comenzar a enumerar estos índices. La primera es que se trata de un campo relativamente nuevo, y por tanto, sujeto a problemas de imprecisiones en los datos, costes y disponibilidad. Una segunda consideración es la existencia de una alta correlación entre los distintos índices existentes, incluso con

duplicaciones y reinvención en los métodos. Tercero, existe una gran confusión entre las medidas de variabilidad y las medidas de control glucémico (en diabéticos). Finalmente, deberíamos tener en cuenta si necesitamos estos índices para fines clínicos o de investigación.

Antes de desarrollar estas medidas de variabilidad, no debemos olvidar los tradicionales porcentajes de recuento de hipo- e hiperglucemias. La hipoglucemia se presenta más frecuentemente, en pacientes con diabetes en tratamiento con insulina, se corresponden a los niveles de glucosa menores a 70 mg/dl. Por otro lado, la hiperglucemia es el hallazgo básico en todos los tipos de diabetes, cuando el individuo supera niveles de glucosa de 180 mg/dl. En este nuestro estudio, la media de los porcentajes de hipoglucemia es de 1.45 % y para la hiperglucemia es del 8.7 %.

La medida de variabilidad más lógica y sencilla sería la **desviación estándar** de la glucosa (SD), ya que mide las fluctuaciones de los niveles de glucosa. Se puede considerar que la VG es baja si el valor de la SD se encuentra tres veces por debajo del promedio normal de la glucosa. Una de las desventajas de utilizar esta medida es que está influenciada por las excursiones de hiperglucemia y no sería sensible a la hipoglucemia, ya que el rango del último es menor al de hiperglucemia. Destacar que la medición de la glucosa es altamente asimétrica.

Schlichtkrull et al. (1965) introdujeron un índice con el fin de evaluar la eficacia del tratamiento con insulina en pacientes con diabetes, el ***M-Value*** (M). Trata de medir la distancia a la que se encuentra el valor de la glucosa en relación a un valor arbitrario pero, tiene la desventaja de no distinguir entre las hipo- e hiperglucemias. Sigue la expresión,

$$M = \frac{1}{N}\sum_{i=1}^{N}\left|10\log_{10}\left(\frac{BG}{IGV}\right)\right|^3 + \frac{\text{máx}(BG) - \text{mín}(BG)}{20},$$

siendo IGV el valor ideal de glucosa y N el número total de lecturas. La elección del primer valor es arbitrario; en personas no diabéticas una buena opción sería 100, mientras que para diabéticos ascendería a 120.

Para el ***Lability Index*** (LI), Ryan et al. (2004), se basaron en los cambios de los niveles de glucosa a lo largo del tiempo, comparándolos con una evaluación clínica de la inestabilidad de la glucemia. Se compone de la suma de todos los cuadrados de las diferencias de glucosa consecutivas dividas por el intervalo de tiempo de las mediciones, es decir,

$$LI = \sum_{i=1}^{N-1}\frac{(BG_i - BG_{i+1})^2}{(t_{i+1} - t_i)},$$

donde t el minuto de la medición y con la BG medida en mmol/L.

Molnar et al. (1970), propusieron la ***Mean Amplitude of Glycemic Excursions*** (MAGE), que contabiliza los cambios de la glucosa para un perfil glucémico. Se trata de la media aritmética de los picos de los niveles de glucosa hasta los puntos más bajos, cuando los segmentos descendentes exceden de una SD, para el mismo periodo de 24 horas. Ya que las fluctuaciones glucémicas se cuentan sólo si exceden de la SD, por lo cual los pacientes con alta SD, presentan una puntuación alta de MAGE. En resumen,

$$MAGE = \sum\frac{\lambda}{x} \text{ con } \lambda > SD,$$

siendo λ los segmentos descendentes válidos y x el número total de λ. En la Figura 3.1 se muestran los niveles de glucosa para un sujeto. Los puntos de color rojo corresponden a los puntos de inflexión de la curva de glucosa. Los puntos verdes son los puntos de inflexión que tienen una diferencia superior

a 8 con el anterior. Además, son los que se han utilizado para el cálculo del MAGE, para calcular las bajadas.

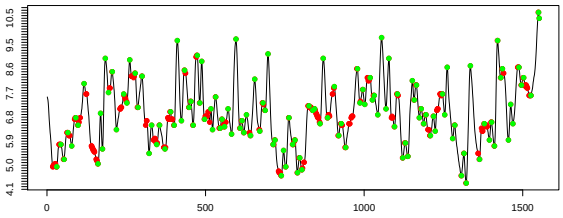

Figura 3.1: Niveles de glucosa para un individuo, los puntos rojos son todos los puntos de inflexión y los verdes los que se utilizarán para el cálculo del MAGE.

La ***Mean Of Daily Differences*** (MODD) desarrollado por Molnar et al. (1972), mide los cambios de los niveles de glucosa resultante de la variación en la respuesta de la terapia día a día. Se calcula en base a los promedios de los valores de glucosa tomados en dos días consecutivos en el mismo tiempo, es decir,

$$MODD = \frac{1}{(K(m-1))} \sum_{i=1}^{N-K} |BG_i - BG_{K+i}|,$$

donde m es el número de días, K el número de mediciones en 24 horas y con BG medida en mmol/L. Es interesante destacar que altos valores de MODD indican hábitos de irregularidad en estilo de vida de las personas.

Más tarde, Wojcicki (1995) diseñó un nuevo índice para eliminar el problema del *M-value* de utilizar diferentes niveles de referencia, con el objetivo de comparar datos de control de la glucosa desde diferentes centros clínicos. Este nuevo índice lo bautizó como ***J-Index*** (J), que es válido para todos los perfiles diarios de glucosa excluyendo aquellos en los que sufren episodios de hipoglucemias. Esta medida intenta resaltar dos componentes importantes de la glucemia; los niveles medios y la variabilidad glucémica. Por ello, el *J-Index* sigue la fórmula

$$J = 0.001(MBG + SD)^2,$$

siendo MBG la media y SD la desvío estándar de los niveles de glucosa.

La ***Continuous Overlapping Net Glycemic Action*** (CONGA), propuesta por McDonnell et al. (2005), describe la VG intradía. Para cada observación de las n primeras horas, se calcula la diferencia entre la observación actual y la n observaciones después. La CONGA se define como la desviación estándar de estas diferencias,

$$CONGA_n = \sqrt{\frac{\sum_{i=n+1}^{N}(D_i - \overline{D})^2}{N-1}},$$

siendo $D_t = BG_i - BG_{i-n}$ y con la BG medida en mmol/L. Destacar que ante valores altos de CONGA la variabilidad glucémica es mayor. El valor por defecto que se utiliza son 60 minutos, en nuestro caso $n = 12$, ya que las lecturas fueron recogidas cada 5 minutos.

El ***Coefficient of Variation*** (CV) propuesto recientemente por Bergenstal et al. (2013) deriva de la desviación estándar. Se define como el mejor parámetro a utilizar en la caracterización de la VG

porque es, relativamente, más constante que otras medidas. Además es la única medida que no asume una distribución normal.

$$\%CV = \frac{100SD}{MBG}.$$

Como índice de variabilidad de la glucosa se podría utilizar el ***Area Under the Curve*** (AUC) de la glucosa. Para ello, utilizaremos la fórmula de Tai (1994) que se calcula dividiendo el área bajo la curva entre dos valores designados en el eje de abscisas en segmentos pequeños, cuyas áreas se pueden calcular con precisión desde sus respectivas fórmulas geométricas. La suma total de las áreas será el área total bajo la curva. Además, las muestras se pueden tomar con intervalos de tiempo diferentes.

$$AUC = \frac{1}{2}\sum_{i=1}^{n} t_{i-1}\left(BG_{i-1} + BG_i\right),$$

donde BG es la glucosa en sangre y t es la diferencia del tiempo de una medición a otra que, en nuestro caso, es de 5 minutos.

En el Cuadro 3.1 mostramos un resumen descriptivo de los índices descritos con anterioridad. Es de interés resaltar que los valores máximos de los índices los toma el mismo sujeto, con ID 0080. Una de las razones de que los índices tengan valores tan altos es que este individuo padece diabetes y no está controlado.

	MG	SD	M	LI	MAGE	MODD	CONGA	CV	AUC
Mínimo	80.67	6.22	2.471	0.65	10.57	0.28	0.29	6.39	66.48
Máximo	209.13	118.71	63.34	36.78	168.09	5.12	2.86	62.40	188.60
Media	109.94	17.79	7.19	4.59	31.96	0.79	0.87	15.68	98.09

Cuadro 3.1: Resumen descriptivo de los índices de variablidad. MG ≡ media de la glucosa. SD ≡ desviación estándar. M≡ M-value. LI ≡ *lability index*. MAGE ≡*Mean Amplitude of Glycemic Excursions*. MODD ≡ *Mean Of Daily Differences*. CONGA ≡ *Continuous Overlapping Net Glycemic Action*. CV ≡ coeficiente de variación. AUC ≡ área bajo la curva.

3.2.1. Medidas de variabilidad y riesgo

Existen otros índices que además de estudiar la variabilidad se centran también, en medir el riesgo de hipo- e hiperglucemia de valores extremos.

Así pues, se propuso el ***Low Blood Glucose Index*** (LBGI) y el ***High Blood Glucose Index*** (HBGI), Kovatchev et al. (2003), que mezclan tanto la variabilidad como el riesgo de hipo- e hiperglucemias. El primero es una medida de frecuencia de niveles bajos de glucosa, basado en episodios de hipoglucemia y por otro lado, el HBGI calcula la frecuencia de niveles altos asociados a las hiperglucemias. Por tanto, estos índices valoran el riesgo de hipo- e hiperglucemias y, en consecuencia pueden ayudar a predecir y prevenir estos episodios.

El primer paso a seguir, para el cálculo de estos índices, será realizar una transformación a los niveles de glucosa en sangre para normalizar su escala

$$f(BG) = 1.509\left(\ln(BG)^{1.084} - 5.381\right),$$

después convertimos los valores de la glucosa en valores de riesgo de tal forma que $rl(BG) = 10f(BG)^2$ si $f(BG) < 0$ (0 en otro caso) y $rh(BG) = 10f(BG)^2$ si $f(BG) > 0$ (0 en otro caso). Finalmente, los

índices son el promedio de los riesgos calculados

$$LBGI = \frac{1}{N}\sum_{i=1}^{N} rl(BG_i) \text{ y } HBGI = \frac{1}{N}\sum_{i=1}^{N} rh(BG_i).$$

Más tarde se formuló un índice que combina los dos anteriores, el ***Average Daily Risk Ratio*** (ADRR), de Kovatchev et al. (2006). Fue diseñado para ser sensible tanto para hipoglucemias como para hiperglucemias. Tiene la ventaja, sobre el LBGI y HBGI, de que sólo es un índice de riesgo y que es predictivo para hipo- e hiperglucemias. Pero, una gran desventaja es que no se tiene en cuenta la velocidad de la variabilidad. Finalmente se calcula siguiendo la expresión

$$ADRR = \frac{1}{m}\sum_{i=1}^{m}(LR^i + HR^i),$$

con $LR^i = \text{máx}\{rl(BG_j^i)\}$ y $HR^i = \text{máx}\{rh(BG_j^i)\}$, con $j = 1, \ldots, N$ y $i = 1, \ldots, m$, siendo m el número de días.

Hill et al. (2007) propusieron la ***Glycemic Risk Assessment in Diabetes Equation*** (GRADE) que, resume los datos de CGM a una única evaluación de riesgo. Como las anteriores, fue diseñada para advertir de posibles hipo- e hiperglucemias. Tiene como limitación que opera con niveles de glucosa dentro de un rango de 37-630 mg/dl. El GRADE se calcula mediante una aproximación de la mediana sobre valores de riesgo

$$GRADE = \text{median}(425(\log_{10}(\log_{10}(BG) + 0.16)^2),$$

con BG medida en mmol/L.

En el Cuadro 3.2 se muestra un resumen descriptivo de los índices de variabilidad y riesgo. En estos índices además de la variabilidad también tienen en cuenta el riesgo. El sujeto que más en riesgo está es el que tiene el ID 1231 ya que tiene el ADRR y el HBGI más alto. El 0080 es el que mayor GRADE tiene. Por último el individuo 0637 es el individuo que más riesgo tiene de hipoglucemia, pues tiene el mayor valor de LBGI.

	LBGI	HBGI	ADRR	GRADE
Mínimo	0	0	1.47	0.07
Máximo	6.14	17.29	66.53	17.42
Media	0.93	0.73	8.52	1.48

Cuadro 3.2: Resumen descriptivo de los índices de riesgo y variablidad. LBGI ≡ *Low Blood Glucose Index* . HBGI ≡ *High Blood Glucose Index* . ADRR ≡ *Average Daily Risk Ratio*. GRADE ≡ *Glycemic Risk Assessment in Diabetes Equation*.

En el Cuadro 3.3 se recogen un resumen de las expresiones que sigue cada índice según el autor y los rangos de referencia que en la literatura se propuso después de diferentes estudios.

3.3. Selección de los índices: Análisis Cluster

En esta sección elegiremos un par de índices para el estudio, ya que son muy parecidos unos con los otros. Para la elección, realizaremos un análisis de conglomerados (en terminología inglesa *cluster*)

Índice	Fórmula	Rangos
%CV	$\frac{100SD}{MBG}$	
AUC	$\frac{1}{2}\sum_{i=1}^{N}(t_{i+1}-t_i)(BG_{i-1}+BG_i)$	
GRADE	$median(425(\log_{10}(\log_{10}(BG/18)+0.16)^2)$	$GRADE < 5$ buen control.
J	$0.001(MBG+SD)^2$	$10 \leq J \leq 20$ control ideal, $20 < J \leq 30$ control bueno, $30 < J \leq 40$ control malo, $J > 40$ falta de control.
M	$\sum_{i=1}^{N}\frac{\left\|10\log_{10}\left(\frac{BG}{TGV}\right)\right\|^3}{N}+\frac{\text{máx}(BG)-\text{mín}(BG)}{20}$	$0 \leq M < 19$ control bueno, $19 \leq M < 32$ control medio, $M \geq 32$ control malo.
LBGI	$\frac{1}{N}\sum_{i=1}^{N} rl(BG_i)$	$LBGI \leq 1.1$ riesgo mínimo, $1.1 < LBGI \leq 2.5$ riesgo bajo, $2.5 < LBGI \leq 5$ riesgo medio, $LBGI > 5$ riesgo alto.
HBGI	$\frac{1}{N}\sum_{i=1}^{N} rh(BG_i)$	$HBGI \leq 4.5$ riesgo bajo, $4.5 < HBGI \leq 9$ riesgo medio, $HBGI > 9$ riesgo alto.
ADRR	$\frac{1}{m}\sum_{i=1}^{m}(LR+HR)$	$ADRR < 10$ riesgo mínimo, $10 \leq ADRR < 20$ riesgo bajo, $20 \leq ADRR < 40$ riesgo medio, $ADRR \geq 40$ riesgo alto.
MAGE	$\sum \frac{\lambda}{x}$ si $\lambda > SD$	$22 \leq MAGE \leq 60$ sanos, $60 < MAGE \leq 90$ diab. estables, $90 < MAGE \leq 200$ diab. inestables.
LI	$\sum_{i=1}^{N-1}\frac{(BG_i-BG_{i+1})^2}{(t_{i+1}-t_i)}$	
MODD	$\frac{1}{(K(m-1))}\sum_{i=1}^{N-K}\lvert BG_i - BG_{K+i}\rvert$	$0.3 < MODD \leq 0.5$ sanos, $0.5 < MODD \leq 2$ diab. estables, $MODD > 2$ diab. inestables.
CONGA	$\sqrt{\frac{\sum_{i=n+1}^{N}(D_i-\overline{D})^2}{N-1}}$ con $D_t = BG_i - BG_{i-n}$	

Cuadro 3.3: Resumen de los índices de variabilidad de la glucosa. En la tercera columna se muestran los rangos de referencia. LBGI ≡ *Low Blood Glucose Index* . HBGI ≡ *High Blood Glucose Index* . ADRR ≡ *Average Daily Risk Ratio*. GRADE ≡ *Glycemic Risk Assessment in Diabetes Equation*. M≡ M-value. LI ≡ *lability index*. MAGE ≡*Mean Amplitude of Glycemic Excursions*. MODD ≡ *Mean Of Daily Differences*. CONGA ≡ *Continuous Overlapping Net Glycemic Action*. CV ≡ coeficiente de variación. AUC ≡ área bajo la curva.

de todos los índices introducidos en la sección anterior.

El principal objetivo del análisis *cluster* es agrupar el conjunto datos de forma que cada grupo de observaciones sea lo más distinto posible uno de los otros, y a su vez, dentro de cada grupo, similares.

Para estudiar el análisis *cluster*, partiremos de un análisis de componentes principales para evitar introducir variables no relevantes. Esta técnica consiste en reducir la dimensión, ya que permite pasar de una gran cantidad de variables interrelacionadas a unas pocas componentes principales. Se intentará buscar combinaciones lineales de variables originales que representen lo mejor posible a la variabilidad presente en los datos, es decir, es una herramienta básica para explicar un conjunto de datos mediante variables ortonormales que cumplen la propiedad de maximizar la varianza. Las componentes principales, que formen un vector aleatorio de dimensión menor, serán empleadas para el análisis *cluster* posterior.

Definición 3.1. Sea $\boldsymbol{x} = (x_1, \ldots, x_d)'$, un vector aleatorio d-dimensional con vector de medias $\boldsymbol{\mu} = E(\boldsymbol{x})$ y matriz de covarianzas $\boldsymbol{\Sigma} = E((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})')$. Se definen las d componentes principales de $\boldsymbol{x}$ como las variables aleatorias $(z_1, \ldots, z_d)$ tales que,

$$z_1 = \boldsymbol{v_1'x}, \ldots, z_d = \boldsymbol{v_d'x}, \quad \boldsymbol{v_1}, \ldots, \boldsymbol{v_d} \in \mathbb{R}^d$$

$$\begin{aligned}
Var(z_1) = &\quad \text{máx}\{Var(\boldsymbol{v'x}) \mid \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v'v} = \mathbf{1}\},\\
Var(z_2) = &\quad \text{máx}\{Var(\boldsymbol{v'x}) \mid \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v'v} = \mathbf{1}, \boldsymbol{v_1'v} = 0\},\\
\vdots &\quad \vdots\\
Var(z_j) = &\quad \text{máx}\{Var(\boldsymbol{v'x}) \mid \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v'v} = \mathbf{1}, \boldsymbol{v_1'v} = 0, \ldots, \boldsymbol{v_{j-1}'v} = \mathbf{0}\},\\
\vdots &\quad \vdots\\
Var(z_d) = &\quad \text{máx}\{Var(\boldsymbol{v'x}) \mid \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v'v} = \mathbf{1}, \boldsymbol{v_1'v} = 0, \ldots, \boldsymbol{v_{d-1}'v} = \mathbf{0}\}.
\end{aligned}$$

En nuestro conjunto tenemos 580 individuos con 18 índices de variabilidad para cada uno. Obtenemos las componentes principales a partir de los autovectores de la matriz de covarianzas. Si escogemos la primera y la segunda componente principal, explicamos un 99 %, con lo que es más que suficiente.

Utilizaremos el algoritmo K-medias para realizar el análisis *cluster* y escoger un índice de cada *cluster*. Este método trata de buscar la partición de los n individuos en k grupos de manera que minimice la suma de cuadrados dentro de los grupos sobre todo el conjunto de variables. Vamos a aplicar el método suponiendo 4 *cluster*, pues queremos seleccionar 4 índices uno de cada *cluster*.

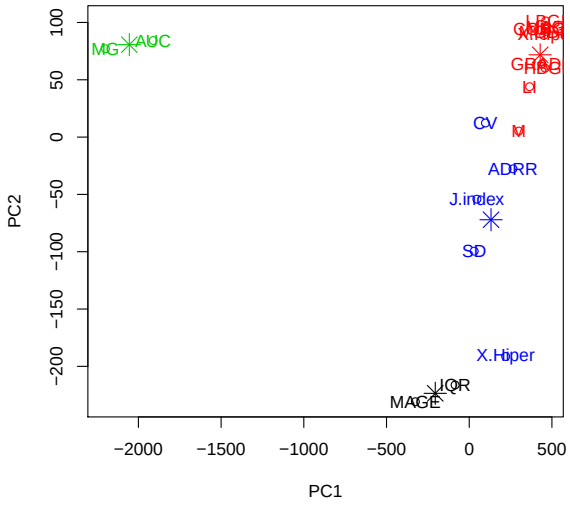

Figura 3.2: Cada conjunto de colores representa un *cluster*.

En vista de la Figura 3.2, escogeremos un índice de cada *cluster*. Nos quedaremos entonces, con la AUC, MAGE, SD y la CONGA, por ser los más utilizados en la literatura y pertenecer a cada uno de los cuatro *cluster* diferenciados.

El primer paso será construir la matriz de distancias entre ellos, que nos ayudará a saber que índices están más próximos. Empleamos la distancia euclídea para calcularla,

$$d^2(i,j) = \|x_i - x_j\|^2 = (x_i - x_j)^t(x_i - x_j) = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2.$$

Una vez, calculada la distancia escogida, podemos obtener la matriz de distancias que tendrá la forma siguiente

$$D = \begin{pmatrix} 0 & D_{12} & \dots & D_{1n} \\ D_{21} & 0 & \dots & D_{2n} \\ \dots & \dots & \dots & \dots \\ D_{n1} & D_{n2} & \dots & 0 \end{pmatrix},$$

es decir, $D_{ii} = 0$ la distancia entre una observación y si misma es siempre 0 y además $D_{ij} = D_{ji}$. Obtendremos así que índices están más próximos entre sí.

		No diabéticos	Diabéticos
SD	Mínimo	6.22	9.88
	Media	15.55	35.27
	Máximo	38.79	118.71
AUC	Mínimo	66.48	80.47
	Media	94.17	127.74
	Máximo	125.60	188.60
MAGE	Mínimo	10.57	21.57
	Media	28.22	60.90
	Máximo	82.24	168.09
CONGA	Mínimo	0.30	0.45
	Media	0.79	1.44
	Máximo	1.97	2.86

Cuadro 3.4: Mínimo, máximo y media de los cuatro índices seleccionados (SD, AUC, MAGE, CONGA) diferenciando entre diabéticos y no diabéticos.

En el Cuadro 3.4 se aprecia que los valores de los índices para sujetos que padecen diabetes es más elevado que para personas no diabéticas pues, o bien, porque existe más variabilidad en los valores de glucosa (como en el caso de la SD) o porque los valores de la glucosa son más elevados (como en el caso de la AUC).

A la luz de estos resultados, será conveniente realizar los futuros análisis sin los sujetos que padecen diabetes ya que sus niveles de glucosa son anómalos en comparación al resto. En lo que sigue, analizaremos como estas medidas de variabilidad nos ayudarían a explicar el factor de inflamación como es el TNF.

3.4. Modelos de regresión para los índices de variabilidad

Para el análisis de los índices de variabilidad al TNF, utilizaremos modelos de regresión. Se tomará como variable respuesta el TNF y como variables explicativas los índices y alguna otra covariable como puede ser el índice de masa corporal (BMI) y la edad. Cabe la posibilidad de que el efecto de las covariables de la variable explicativa no tenga efecto lineal sobre el TNF por lo que se ajustará un modelo aditivo (Wood (2006)).

Los modelos aditivos generalizados (GAM) es un modelo generalizado lineal donde el predictor lineal está dado por la suma de funciones suavizadoras de las covariables.

$$\mathbb{E}[Y] = g^{-1}\left(\alpha + \sum_{j=1}^{p} f_j(x_j)\right),$$

donde Y es la variable respuesta, f_j el efecto parcial suave (desconocido) de x_j en el predictor y α corresponde a la parte no paramétrica del modelo.

Los modelos GAM se ajustarán con la función `gam` del paquete `mgcv` del *software* estadístico `R`. Se estima el modelo GAM como un GLM penalizado cuadráticamente y el grado de suavización óptimo se obtiene como parte del ajuste. Los covariables suavizadas se representan utilizando bases de splines penalizadas con parámetros suavizadores.

En un principio, el modelo de interés a estimar es el siguiente:

$$\text{Modelo } 1 \equiv TNF \sim f_1(Edad) + f_2(AUC) + f_3(SD) + f_4(CONGA) + f_5(MAGE) + f_6(BMI)$$

donde f_i para $i = 1, \ldots, 6$ son las funciones suavizadoras para las covariables de la edad, BMI y los cuatro índices seleccionados.

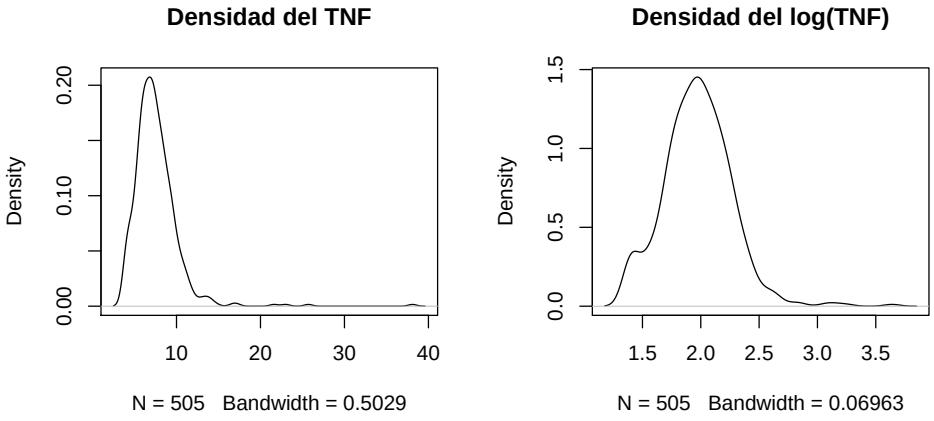

Figura 3.3: Función densidad del TNF y la del logaritmo del TNF.

Es de interés, saber como es la función densidad del TNF, saber si sigue una normal o no. En el gráfico de la izquierda de la Figura 3.3 se muestra la función densidad de este factor y la derecha, se ha representando la función densidad del logaritmo para comprobar si nos bastaría esta transformación. Se puede constatar que existen una serie de sujetos que tienen un TNF elevado con lo que, la función densidad se vuelve asimétrica. Por tanto, para tener una buena estimación del modelo es recomendable eliminar estos individuos.

En la Figura 3.4 se visualiza, de nuevo, las función de densidad del TNF y del logaritmo del TNF, pero sin contar con 6 individuos que su TNF es superior a 15. A la vista de la figura, se puede apreciar que el logaritmo del TNF se ajusta a una función de densidad normal. En lo que sigue estimaremos los modelos tomando como variable respuesta el logaritmo y por tanto, usaremos como función *link* la

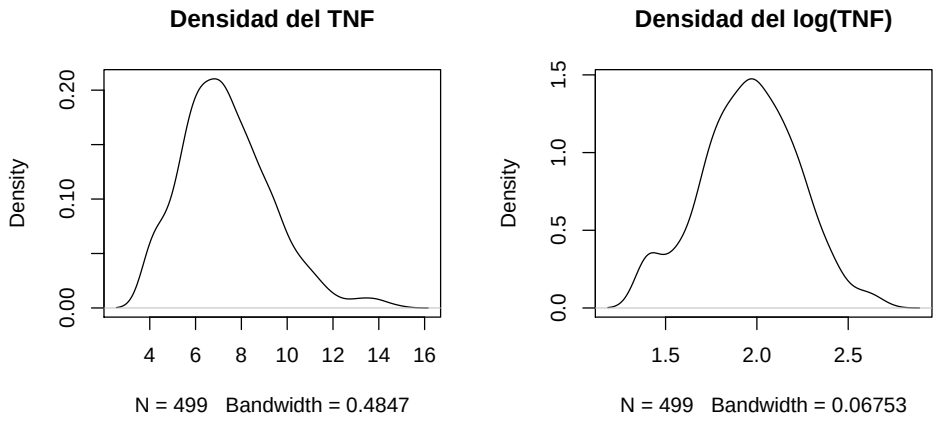

Figura 3.4: Función densidad del TNF y la del logaritmo del TNF, eliminando 6 sujetos que superan un valor 15 de TNF.

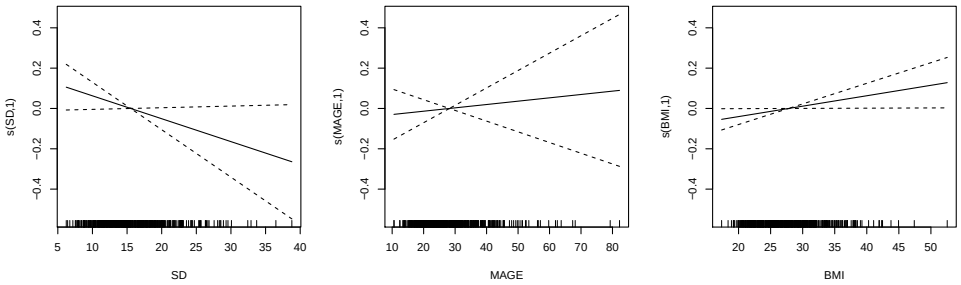

Figura 3.5: Efecto del BMI y los índices SD y MAGE sobre el factor de inflamación del TNF.

identidad.

Una vez ajustado el Modelo 1, con la función `gam` se aprecia que las medidas de variabilidad SD y MAGE y al BMI no habría que aplicarles una función suavizadora pues, como se aprecia en la Figura 3.5, el efecto que tienen sobre el TNF es lineal. Por tanto, el modelo a estimar sería el siguiente:

$$\text{Modelo } 2 \equiv TNF \sim f_1(Edad) + f_2(AUC) + SD + f_3(CONGA) + MAGE + BMI$$

donde f_i $(i = 1, \ldots, 3)$ son las funciones suavizadoras. El resultado de estimar el Modelo 2 es el siguiente:

```
Family: gaussian
Link function: identity

Formula:
logt ~ s(Edad) + s(AUC) + s(CONGA) + SD + MAGE + BMI

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1.945376   0.099805  19.492   <2e-16 ***
SD          -0.011358   0.006087  -1.866   0.0626 .
MAGE         0.001658   0.003485   0.476   0.6345
BMI          0.005171   0.002527   2.047   0.0412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
edf Ref.df     F  p-value
s(Edad)  2.983  3.764 6.537 7.17e-05 ***
s(AUC)   2.033  2.623 1.509    0.211
s(CONGA) 2.286  2.906 1.513    0.210
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0733   Deviance explained = 9.25%
GCV = 0.067335  Scale est. = 0.06581   n = 499
```

En vista de los resultados obtenidos, las únicas covariables que son significativamente diferente a cero son la edad y el BMI, es decir, podría decirse que la *deviance* explicada, 9.25 %, seguramente sea totalmente achacable a estas dos covariables, lo comprobamos ajustando el siguiente modelo,

$$\text{Modelo } 3 \equiv TNF \sim f(Edad) + BMI$$

donde f es la función suavizadora de la edad. La estimación del Modelo 3 ajustado es:

```
Family: gaussian
Link function: identity

Formula:
logt ~ s(Edad) + BMI

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.852831   0.069301  26.736   <2e-16 ***
BMI         0.003839   0.002458   1.562    0.119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
edf Ref.df     F  p-value
s(Edad) 7.301  8.267 4.194 5.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0693   Deviance explained = 8.48%
GCV = 0.067351  Scale est. = 0.066096  n = 499
```

La *deviance* explicada del nuevo modelo sin los índices de variabilidad es de 8.48 % contra el 9.25 % del Modelo 2, además el BMI no es significativamente distinto a cero. En la Figura 3.6 podemos ver el efecto suavizado de la edad sobre el TNF, a partir de los 45 años se aprecia una gran subida de los valores del TNF. Notar que a partir de los 70 el intervalo de confianza se agranda, con lo que a la hora

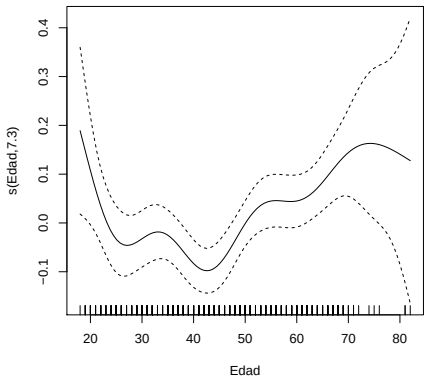

Figura 3.6: Efecto de la covariables edad sobre el factor de inflamación del TNF.

de realizar predicciones podría conllevar a errores.

A la luz de los resultados obtenidos con la estimación de los modelos GAM, los índices de variabilidad no influyen en los valores de TNF que puede tener un individuo, al igual que el BMI. Por otro lado, la edad sí que influye en los valores del TNF, los sujetos cuánto mayor sean, más elevados serán los valores de TNF.

Se puede concluir que, según el análisis realizado, las excursiones de glucosa no tienen ningún efecto sobre el TNF pues, los índices de variabilidad de la misma no lo tienen y se crearon con el fin de medir dichas excursiones.

En el siguiente capítulo se utilizará otra metodología para medir la variabilidad de la glucosa para poder asegurar, o no, si la afirmación de que las fluctuaciones de la glucosa no tienen ningún efecto sobre el factor inflamatorio TNF.

Capítulo 4

Análisis de datos funcionales

En este capítulo se estudiará la metodología de tratar los niveles de glucosa como datos funcionales, para saber que efecto tiene la glucosa sobre el factor de inflamación TNF y la predicción del mismo. Comenzaremos introduciendo el concepto de datos funcionales y el análisis exploratorio de los mismos. Se proseguirá comparando modelos de regresión funcionales y la elección de uno de ellos.

El procedimiento anteriormente descrito se realizará en *software* estadístico `R`. Para trabajar con este tipo de datos, entre los paquetes disponibles, se empleará el paquete `fda.usc` (Febrero-Bande and Oviedo de la Fuente (2012)). Existen otros paquetes como el bien conocido `fda`, pero el anterior trabaja en un campo más amplio. El objetivo de esta sección será estudiar si las fluctuaciones de glucosa, tomando la curva de glucosa como dato funcional, tienen efecto sobre el factor TNF.

4.1. Preliminares

Durante el análisis de los datos funcionales se restringirá al espacio de Hilbert $\mathcal{L}_2$,

$$\mathcal{L}_2 = \left\{ f : \mathbb{R} \to \mathbb{R} \mid \int_{\mathbb{R}} f^2(t)dt < \infty \right\}.$$

A continuación se introducirá el concepto de dato funcional.

Definición 4.1. Una variable aleatoria $\mathcal{X}$ se llama variable funcional si toma valores en un espacio métrico completo o un semimétrico funcional. Una observación x de $\mathcal{X}$ se llama dato funcional.

Definición 4.2. Un conjunto de datos funcionales $\mathcal{X}_1, \ldots, \mathcal{X}_n$ son las observaciones de n variables funcionales, idénticamente distribuidas como $\mathcal{X}$.

En la Figura 4.1 presentamos para cada individuo la curva de glucosa media de los 5 días para las diferentes comidas del día. La primera gráfica corresponde al momento en que el individuo comienza a cenar hasta tres horas después; la segunda, desde que empieza a comer hasta cuatro después; y la tercera, desde que cena hasta 8 horas pasadas. Ya que cada paciente ha apuntando en una hoja de registro todas las horas de ingesta.

4.1.1. Representación de datos funcionales

Representación con bases

Una curva puede ser representada en una base cuando se supone que los datos pertenecen al espacio $\mathcal{L}_2$. Una base es un conjunto de funciones conocidas $\{\phi_k\}_{k\in\mathbb{N}}$, tales que cualquier función puede ser

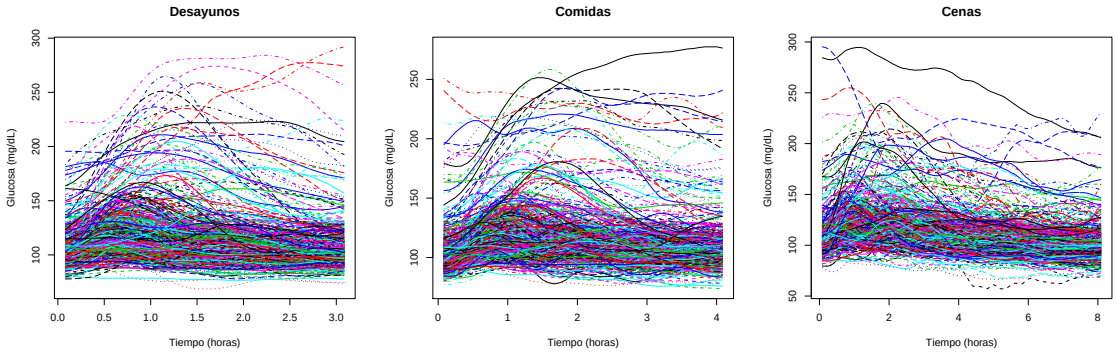

Figura 4.1: Curvas de glucosa de los sujetos para los desayunos, comidas y cenas, tres, cuatro y ocho horas después de la ingesta.

aproximada mediante una combinación lineal de k_n funciones.

Podemos aproximar la observación funcional de la forma,

$$\mathcal{X}(t) = \sum_{k\in\mathbb{N}} c_k\phi_k(t) \approx \sum_{k=1}^{k_n} c_k\phi_k(t) = \boldsymbol{c^T\Phi}(t).$$

Las bases más comunes que se suelen utilizar son las bases de Fourier, bases B-splines, bases Wavelets, ... La elección de la base depende de los análisis y de los datos. Por ejemplo, si los datos son periódicos, se suelen utilizar series de Fourier. Para cálculos sencillos y rápidos, es común el uso de B-splines.

Por otro lado, la metodología no paramétrica, en particular la suavización tipo núcleo, puede usarse para la representación de datos funcionales. Por lo cual, supongamos que observamos $\mathcal{Y}(tj) = \mathcal{X}(t_j) + \varepsilon(t_j)$ donde $\varepsilon(t_j)$ representa el ruido originario al medir los datos con matriz de covarianzas $\sum_\varepsilon = \boldsymbol{W}^{-1}$.

Para recuperar la señal original podemos utilizar un suavizador lineal, es decir,

$$\hat{\mathcal{X}}(t_j) = \sum_{i=1}^{T} s_j(t_i)\mathcal{Y}(t_i) \Rightarrow \boldsymbol{\hat{\mathcal{X}}} = \boldsymbol{S\mathcal{Y}}$$

Por ejemplo, para el estimador de Nadaraya–Watson:

$$s_j(t_i) = \frac{K\left(\frac{t_i - t_j}{h}\right)}{\sum_{k=1}^{m} K\left(\frac{t_i - t_k}{h}\right)}$$

donde $K(\cdot)$ es la función tipo núcleo y h el parámetro ventana. Otras posibilidades para la elección de S puede ser la representación finita en base ($\boldsymbol{S}(K) = \boldsymbol{\Psi}(\boldsymbol{\Psi'W\Psi})^{-1}\boldsymbol{\Psi'W}$) donde $\boldsymbol{\Psi} = (\psi_1(t_i) \mid \cdots \mid \psi_j(t_i) \mid \cdots \mid \psi_k(t_i))$ construido por columna), suavización penalizada ($\boldsymbol{S}(\boldsymbol{K}, \lambda) = \boldsymbol{\Psi}(\boldsymbol{\Psi'W\Psi} + \lambda\boldsymbol{R})^{-1}\boldsymbol{\Psi'W}$ donde $\boldsymbol{R}$ es la matriz de penalización).

La elección del parámetro de suavización (h, K o λ) es crucial, y en principio, no existen ninguna regla universal que permita la elección óptima. Pero, en general se pretende minimizar el error cuadrático medio,

$$\hat{\mathcal{X}} = \arg\min_\theta \mathbb{E}\left[\int (\mathcal{X}(t) - \hat{\mathcal{X}}_\theta(t))^2 dt\right] = \arg\min_\theta \int Sesgo^2(t) + Var(t)dt.$$

Por ello podemos utilizar el criterio *Genereralized Cross-Validation* (GCV)

$$\mathrm{GCV}(\theta) = \frac{(\boldsymbol{\mathcal{Y}} - \hat{\boldsymbol{\mathcal{X}}}_\theta)'\boldsymbol{W}(\boldsymbol{\mathcal{Y}} - \hat{\boldsymbol{X}}_\theta)}{\left(1 - \frac{tr(\boldsymbol{S})}{n}\right)^2}.$$

Si el proceso $\boldsymbol{\mathcal{Y}}$ es gaussiano, la varianza de predicción viene dado por $Var[\hat{\boldsymbol{\mathcal{Y}}}] = \boldsymbol{S}\boldsymbol{\Sigma}_\varepsilon\boldsymbol{S}'$.

Representación con las componentes principales

Sea $\mathcal{X}(t) \in \mathcal{L}_2(T)$ y $\sum(s,t) = \mathbb{E}\left[(\mathcal{X}(s) - \overline{\mathcal{X}}) - (\mathcal{X}(s) - \overline{\mathcal{X}})\right]$ y el operador

$$\begin{array}{rccc} T_\Sigma: & \mathcal{L}_2(T) & \to & \mathcal{L}_2(T) \\ & f(t) & \to & \int_T \Sigma(s,t)f(s)ds \end{array}$$

donde T_Σ es el operador lineal, esto hace que tenga sentido hablar sobre autovalores λ_k y autovectores v_k que resuelven,

$$\int_T \Sigma(s,t)v_k(s)ds = \lambda_k v_k(t).$$

Los autovectores maximizan la varianza y son ortogonales con cada uno de ellos, $\{v_i\}_{i\in\mathbb{N}}$ forman una base ortogonal de $\mathcal{L}_2(T)$, es decir, $\mathcal{X} = \sum_{i=1}^{\infty}\langle\mathcal{X}, v_i\rangle v_i$. Además, $Z_i = \langle\mathcal{X}, v_i\rangle$ verifica que $\mathbb{E}[Z_i] = 0$ para todo $i \in \mathbb{N}$ y $E[Z_iZ_j] = \delta_{ij}\lambda_k$ para todo $i, j \in \mathbb{N}$.

Las bases de componentes principales son las más efectivas para resumir la información de $\mathcal{X}$.

4.2. Análisis exploratorio

El análisis exploratorio de los datos funcionales podrá ser útil para poder resumir las características de los datos, detectar posibles errores o datos *outliers*. En nuestro caso, nos puede ayudar para saber si incluimos a los pacientes diabéticos en nuestro análisis o no, o bien, si es necesario excluir a algún paciente con curvas de glucosa fuera de lo común.

Las medidas de profundidad nos permiten ordenar los datos desde el más interno hasta el más externo. Los más profundos pueden ser empleados para definir medidas de profundidad, así mismo lo más exteriores serán los candidatos a ser *outliers*.

La interpretación cara la medida de centralidad utilizando profundidades viene dada por la función de profundidad empleada, por tanto puede considerarse una extensión de las medidas clásicas univariantes de datos funcionales (media, moda, mediana...). En lo que sigue, se calcularán las profundidades de toda la muestra (para desayuno, comida y cena), y se ordenarán de forma decreciente obteniendo un rango del punto más central $\left(x_{[1]}\right)$ al punto más lejano$\left(x_{[n]}\right)$. Además, solamente se empleará los $(1-\alpha)$ puntos más profundos, de esta manera las medidas de localización serán más robustas.

Sea $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ una muestra de variables funcionales aleatorias e independientes e idénticamente distribuidos con dominio $\mathcal{T} = [a,b]$ y sea D una medida de profundidad en $\mathbb{R}$, podemos definir los diferentes tipos de profundidades para datos funcionales:

- **Profundidad de Fraiman-Muniz** (FMD)

  Para todo $t \in \mathcal{T}$, se considera $z_i(t) = D(\mathcal{X}_i(t))$ con $\mathcal{X}_i \in S_n$ y siendo D una profundidad univariante del dato i evaluado en t . Se define la profundidad de Fraiman-Muniz como la integración univariante a lo largo de $\mathcal{T}$:

  $$FMD(\mathcal{X}_0) = \int_{\mathcal{T}} z_0(t)dt,$$

Dependiendo de la profundidad univariante elegida, FMD puede tener interpretaciones de media. mediana o moda.

- **Profundidad Modal** (MD)

  Sea $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ una muestra de variables funcionales aleatorias e independientes e idénticamente distribuidas. Sea $K : \mathbb{R}^+ \to \mathbb{R}^+$ una función tipo núcleo asimétrica y h el parámetro ventana. Se define profundidad modal como

  $$MD(\mathcal{X}_0) = \sum_{j=1}^{n} K\left(\frac{d(\mathcal{X}_0, \mathcal{X}_j)}{h}\right),$$

  donde d es una distancia funcional. Esta profundidad como cuántos datos hay alrededor del dato $\mathcal{X}_0$, por tanto su interpretación se asemejaría a una moda.

- **Profundidad basada en proyecciones aleatorias** (RPD)

  Sea $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ una muestra de variables funcionales aleatorias e independientes e idénticamente distribuidas, se toma una dirección aleatoria α (independiente de $\mathcal{X}_0$) y se proyectan los datos a lo largo de esa dirección. Por tanto, la profundidad muestral de $\mathcal{X}_0$ se define como la profundidad univariante de la proyección unidimensional correspondiente. Como se supone que $\mathcal{X}_0$ recorre el espacio de Hilbert $\mathcal{L}_2$, la proyección está dada por el producto interno, es decir, $P_0^\alpha = \langle \alpha, \mathcal{X}_0 \rangle$. Con lo cuál, se define RPD como,

  $$RPD(\mathcal{X}_0, \alpha) = D\left(P_0^\alpha\right),$$

  siendo D una medida de profundidad univariante.

  Teóricamente, una proyección es suficiente pero, sería mejor generar una colección de proyecciones aleatorias $\left(\{\alpha_I\}_{I=1}^M\right)$ y calcular la profundidad basándose en esas proyecciones, es decir,

  $$RPD\left(\mathcal{X}_0, \{\alpha_I\}_{I=1}^M\right) = \frac{1}{M}\sum_{I=1}^{M} D\left(P_0^{\alpha_I}\right).$$

  También utilizaremos la RPD basada en la distancia de Tukey,

  $$RTD\left(\mathcal{X}_0, \{\alpha_I\}_{I=1}^M\right) = \min_M D\left(P_0^{\alpha_I}\right).$$

En las cuatro primeros gráficos de las Figuras 4.2, 4.3 y 4.4, representamos el nivel de profundidad de las curvas del desayuno, comida y cena, respectivamente. Se han aplicado los cuatro métodos expuestos con anterioridad. Las curvas más oscuras corresponden a los que poseen mayor profundidad y las grises las que menos, es decir, la profundidad se rige por la escala de grises. Además se han tomando para la media recortada el 10 % de las curvas menos profundas para que la medida sea más robusta. La línea roja corresponde al sujeto con la curva media de glucosa más profunda, mientras que la línea amarilla es la media recortada del 10 % de las menos profundas.

Los gráficos inferiores de las tres figuras mencionadas, representan un análogo al *boxplot*. Las líneas rojas son los datos que se encuentran en el núcleo, es decir, el 50 % de las más profundas, las grises son los datos que se encuentran en el 5 % de las menos profundas (posibles candidatos a ser *outliers*) y las azules son el 95 % de las curvas que su profundidad está entre las dos anteriores.

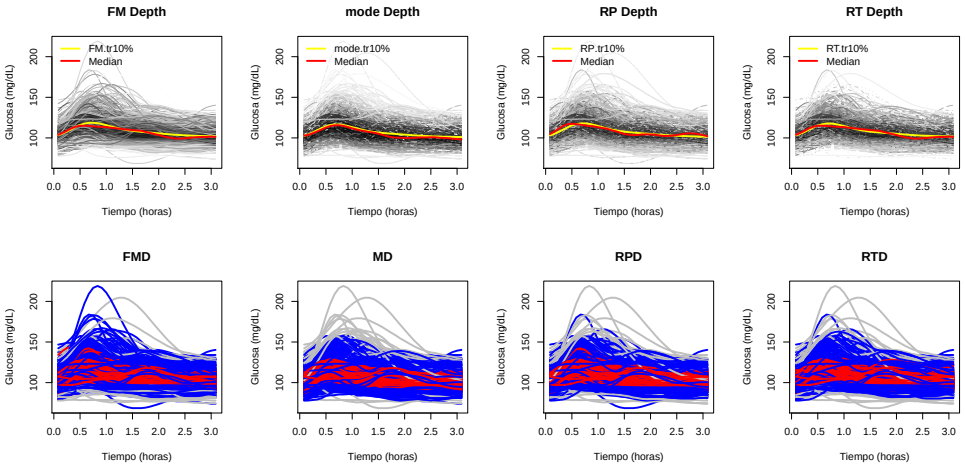

Figura 4.2: De izquierda a derecha: Cálculo de la profundidad de las curvas del desayuno por los métodos de Fraiman-Muniz, modal, proyecciones aleatorias y Tukey. Cuatro superiores: nivel de profundidad de las curvas según la escala de grises. Cuatro inferiores: las líneas rojas son las curvas del 50 % más profundas, las grises son el 5 % de las menos profundas y las azules son el 95 % de las del medio.

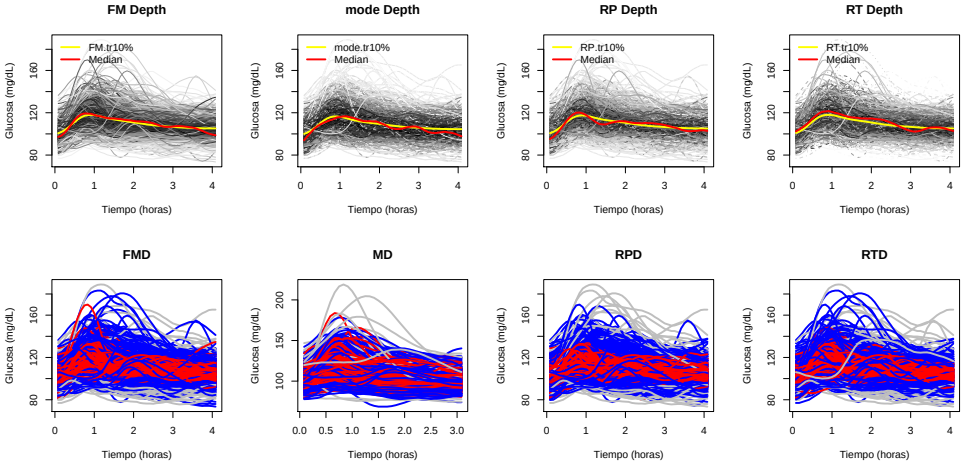

Figura 4.3: De izquierda a derecha: Cálculo de la profundidad de las curvas de las comidas por los métodos de Fraiman-Muniz, modal, proyecciones aleatorias y Tukey. Cuatro superiores: nivel de profundidad de las curvas según la escala de grises. Cuatro inferiores: las líneas rojas son las curvas del 50 % más profundas, las grises son el 5 % de las menos profundas y las azules son el 95 % de las del medio.

Es de interés fijarnos en la Figura 4.5, dónde se dibujan las curvas más profundas según los diferentes métodos. En los desayunos, la curva más profunda calculada por el método de Fraiman-Muniz coincide con el de proyecciones aleatorias. Para los cuatros métodos la glucosa en el desayuno, llega a niveles más altos la primera media hora y después decrece, las dos últimas hora los niveles se mantienen.

Por otro lado, si calculamos la profundidad de las curvas, utilizando el 20 % de las más profundas nos quedaríamos con las curvas que se encuentran en el núcleo. Usando las profundidades por los cuatro métodos ya comentados con anterioridad, con Fraiman-Muniz tenemos 17 individuos que se encuentran en el núcleo de las curvas de glucosa para los desayunos, comidas y cenas; con la profundidad modal 16; con proyecciones aleatorias basadas en Tukey 11; y por último 12 para proyecciones aleatorias.

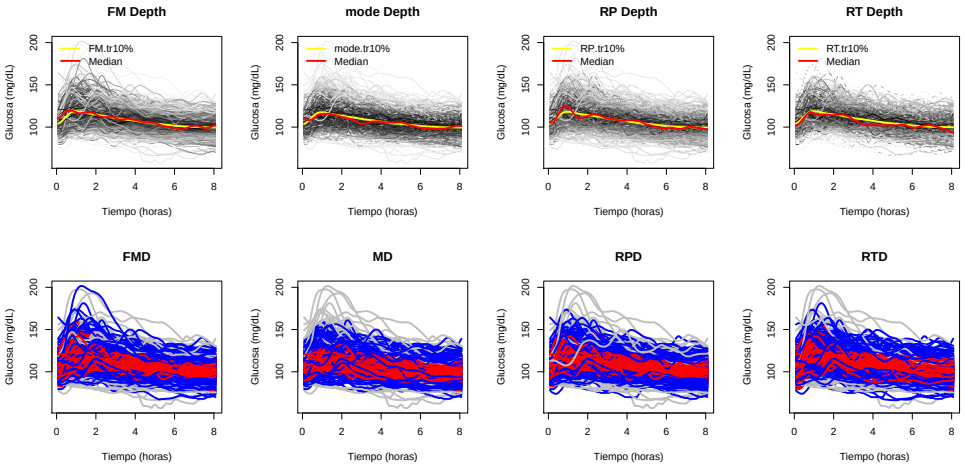

Figura 4.4: De izquierda a derecha: Cálculo de la profundidad de las curvas de las cenas por los métodos de Fraiman-Muniz, modal, proyecciones aleatorias y Tukey. Cuatro superiores: nivel de profundidad de las curvas según la escala de grises.Cuatro inferiores: las líneas rojas son las curvas del 50 % más profundas, las grises son el 5 % de las menos profundas y las azules son el 95 % de las del medio.

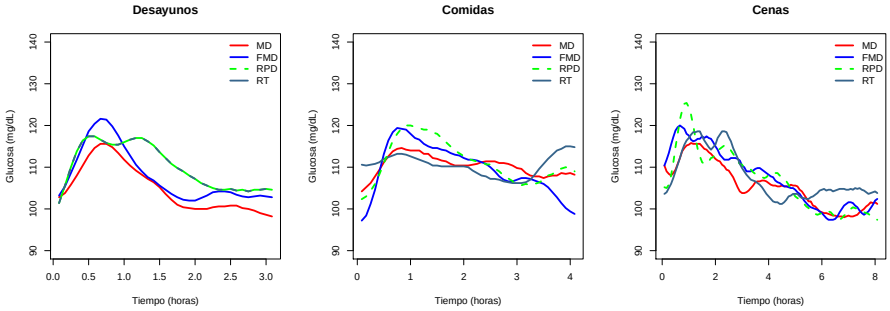

Figura 4.5: De izquierda a derecha: Curvas más profundas para desayunos, comidas y cenas. La línea roja, profundidad modal (MD); la azul, profundidad de Fraiman-Muniz, (FMD); la línea punteada verde, profundidad por proyecciones aleatorias (RPD); y la línea azul oscura, RPD basada en Tukey (RT).

En la Figura 4.6 representamos las curvas de glucosa para los 6 individuos que coinciden utilizando cualquier de los cuatro métodos, éstos individuos serían los mejores representantes de curvas de glucosa. Se correspondería con los individuos con ID: 0070, 0878, 1123, 1192 y 1229. Cabe destacar que ninguno de los 5 padece diabetes y que tienen un índice de masa corporal entre 22.91 y 35.30.

En vista de los resultados ofrecidos por las medidas de profundidad, podemos concluir que para cualquier momento del día, ya sea desayuno, comida y cena, los sujetos candidatos a ser *outliers* son los que tienen niveles de glucosas muy altos o muy bajos. Ya que los individuos que siguen este patrón suelen ser los diabéticos, para los análisis que restan en este proyecto trabajaremos solamente con los sujetos no diabéticos.

Cabe la posibilidad de que en el grupo que se ha identificados como no diabéticos existan sujetos diabéticos no diagnosticados o simplemente que presenten anomalías en las curvas de glucosa por tanto, identificaremos valores atípicos en el conjunto de los datos funcionales del grupo identificado como no diabético. Con el fin de identificar los *outliers* se hará uso de las medidas de profundidad, por lo que

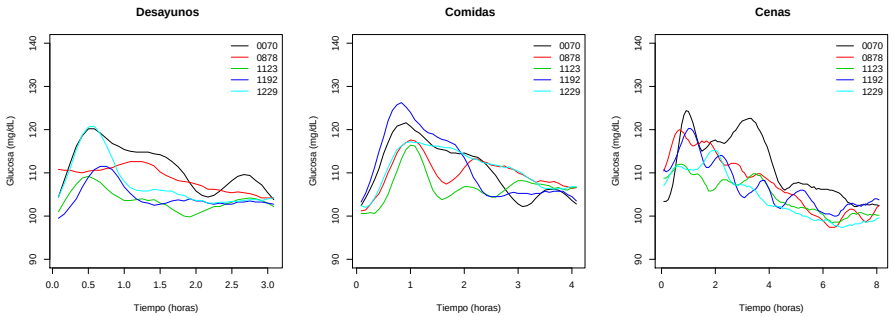

Figura 4.6: Curvas de glucosa para desayunos, comidas y cenas de los individuos con más profundidad en los datos funcionales, dónde coinciden los cuatro métodos expuestos en el este proyecto.

si un dato es atípico, la curva correspondiente tendrá una profundidad significativamente baja. Por tanto, para la detección de *outliers* nos tendremos que centrar en las curvas con profundidad baja.

Se estudiarán los candidatos a ser *outliers* en las derivadas de las curvas de glucosa, pues recordemos que nuestro objetivo es estudiar la variabilidad de las mismas. Para encontrarlos utilizaremos, tanto para desayunos, comidas y cenas, la profundidad de Fraiman-Muniz, por rapidez computacional, con 1000 réplicas *bootstrap*, con 0.1 de parámetro de suavizado para las muestras bootstrap y retirando el 1 % de las curvas con menor profundidad.

Una vez aplicado el método, obtenemos que existen 9 sujetos candidatos para ser atípicos en el desayuno, 8 en las comidas y 9 en las cenas. En la Figura 4.7, se presenta el gráfico de las derivadas de las curvas de glucosa resaltando en color rojo los datos atípicos para los desayunos, comidas y cenas.

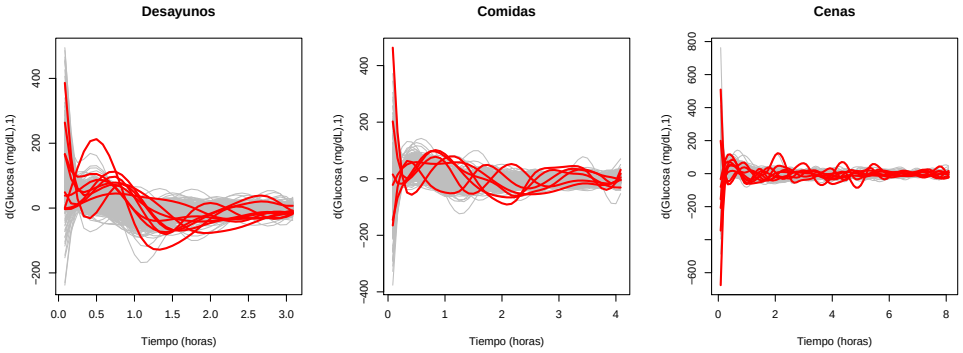

Figura 4.7: Derivadas de las curvas de glucosa para los sujetos, en color rojo los candidatos a ser datos atípicos, en gris el resto, para desayunos, comidas y cenas.

Destacar que, en vista del Cuadro 4.1, podemos ver que el sujeto con código 1110 es candidato a ser *outlier* en las curvas de glucosa de los desayunos y las comidas; y el 1484 en las comidas y en las cenas. En dicho cuadro se recogen los valores de las profundidades de los candidatos a ser *outliers*. Para cada momento del día, el valor q marca el límite para detectar los datos atípicos, si el valor profundidad del sujeto se aleja de q mayor razón para eliminar a ese individuo.

Desayunos ($q \equiv 0.6019$)		**Comida** ($q \equiv 0.6112$)		**Cena** ($q \equiv 0.6247$)	
Código	Profund.	Código	Profund.	Código	Profund.
0041	0.5708	0058	0.5847	0007	0.5963
0248	0.6011	0060	0.6032	0042	0.6132
0525	0.5730	0208	0.5517	0102	0.6178
0847	0.6006	0228	0.6086	0358	0.6231
1020	0.6010	0988	0.6044	0391	0.5954
1110	0.5738	1110	0.6087	0848	0.6096
1241	0.6015	1441	0.5693	1383	0.6223
1387	0.5926	1484	0.6089	1444	0.6234
1456	0.5838			1484	0.6201

Cuadro 4.1: Valores de las profundidades de los candidatos a ser *outliers* en los desayunos, comidas y cenas. q es el límite para la detección de los *outliers*.

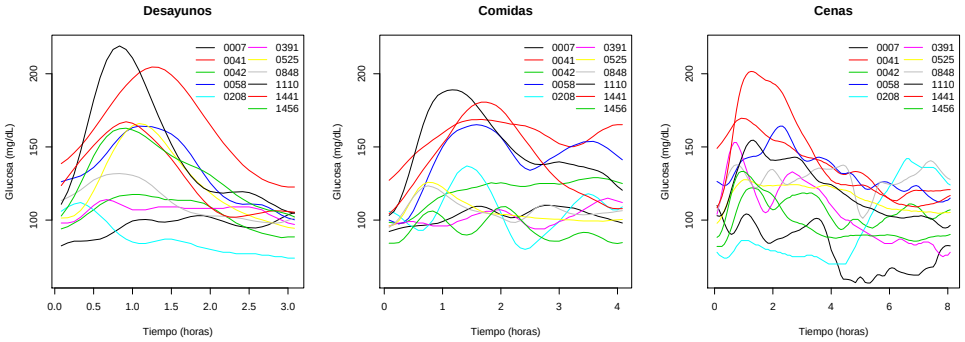

Figura 4.8: Curvas de glucosa del desayunos, comidas y cenas en los sujetos *outliers* que se eliminarán al realizar el análisis posterior.

Por tanto, si nos fijamos en los valores de profundidades recogidos en el cuadro mencionado, parece razonable eliminar 11 de los 24 candidatos pues, presentan una diferencia mayor a 0.01. En la Figura 4.8 se representan las curvas de glucosa de los sujetos a descartar. Por ejemplo, el individuo con el código 0007 presenta, en los niveles de glucosa de las cenas, peligro de sufrir una hipoglucemia. El 0208, en la cena mantiene valores muy bajos y a las 4 de la mañana el nivel de glucosa se dispara.

4.3. Regresión con datos funcionales

Nuestro objetivo es saber cómo explica la variabilidad de las curvas de glucosa el factor de inflamación TNF, ajustando por covariables como puede ser la edad. Lo descrito es generalmente un modelo de regresión y puesto que vamos a utilizar alguna variable funcional, necesitaremos modelos de regresión funcionales. Por tanto, debemos utilizar **modelos funcionales de regresión lineal generalizados** (GLM).

Para definir los modelos GLM es necesario saber que es un **modelo regresión lineal funcional**. Se dice que un modelo de regresión es funcional cuando una de las variables involucradas en el modelo (bien sea la variable explicativa o la predictora) es funcional. En nuestro caso, se van a desarrollar modelos de regresión funcional con variables explicativas las tres covariables funcionales (los niveles de glucosa en los desayunos, comidas y cenas).

Para definir un modelo de regresión lineal funcional se supone que $\mathcal{X} \in \mathcal{L}_2(T)$ e $y \in \mathbb{R}$. Asumiendo que ambas variables están centradas, es decir, $\mathbb{E}[\mathcal{X}(t)] = 0$, para todo $t \in [0,T]$ y $\mathbb{E}[y] = 0$, se establece la siguiente relación entre $\mathcal{X}$ e y:

$$y = \langle \mathcal{X}, \beta \rangle + \varepsilon = \int_T \mathcal{X}(t)\beta(t)dt + \varepsilon, \tag{4.1}$$

donde $\langle , \rangle$ denota el producto escalar, $\beta \in \mathcal{L}_2(T)$ y ε el error (independiente de $\mathcal{X}$).

En nuestro caso, como variable respuesta usaremos el logaritmo del TNF ya que, como vimos en el capítulo anterior, podemos asumir que tiene una distribución normal. Como variables explicativas funcionales escogeremos las curvas de desayunos, comidas y cenas; y alguna otra covariable escalar como puede ser la edad o el índice de masa corporal (BMI).

Dado $\mathbf{X} = \{\mathcal{X}^j\}_{j=1}^p$ un conjunto de covariables funcionales con valores en el espacio de Hilbert $\mathbf{E} = E^1 \times \cdots \times E^p$, supondremos que $\mathbb{E}\left[Y \mid \mathcal{X}\right]$ está relacionada con $\beta_0 + \langle \mathbf{X}, \beta \rangle$ utilizando una función *link*,

$$\mathbb{E}[Y \mid \mathbf{X}] = \eta = g^{-1}\left(\beta_0 + \langle \mathbf{X}, \beta \rangle\right), \tag{4.2}$$

donde $\beta = (\beta_1, \ldots, \beta_p)$ toma valores en el espacio de Hilbert, $\langle \mathbf{X}, \beta \rangle = \sum_{j=1}^p \langle \mathcal{X}^j, \beta_j \rangle$ y g es una función *link*, describiendo la relación funcional entre el valor esperado η al dato y y la componente lineal. Como en nuestro caso, la variable respuesta sigue una distribución normal, no hará falta aplicar la función *link*, es decir, usaremos una función *link* identidad.

Functional Spectral Additive Models (FSAM)

Como la relación de las derivadas de las curvas de la glucosa con el factor de inflamación cabe la posibilidad de que sea no lineal, debemos utilizar modelos aditivos. Los modelos aditivos son una extensión de los modelos lineales, en los cuales el predictor lineal no está restringido a ser lineal, pero es la suma de funciones suavizadoras aplicadas a las covariables.

Dado $\mathbf{X} = \{\mathcal{X}^j\}_{j=1}^p$, conjunto de variables funciones, un modelo aditivo funcional se puede expresar de la manera siguiente:

$$\mathbb{E}[Y|\mathbf{X}] = \eta = \beta_0 + \sum_{j=1}^{p} f_j\left(\mathcal{X}^j\right), \tag{4.3}$$

donde f_j son las funciones suavizadores que usan la información de la covariable funcional.

Siendo x_j^k la puntuación de la componente principal funcional k de X^l, la representación de la variable funcional será de la forma siguiente:

$$\mathcal{X}(t) = \mu(t) + \sum_k x_k v_k(t),$$

donde $v_k(t)$ es el k autovector.

Por tanto, se puede representar cada función suavizadora de la forma $f_j(\mathcal{X}^j) = \sum_{k=1}^{K_j} f_j^k(x_j^k)$ funciones suavizadoras de las puntuaciones x_j^k, de la componente principal k de la variable j. Esta aproximación se conoce como **modelos espectrales aditivos funcionales** (FSAM). Notar que, la descomposición en componentes principales es sólo válida en los espacios de Hilbert.

Los modelos FSAM añadiendo covariables escalares (edad, BMI...) seguirán la expresión:

$$\mathbb{E}[Y \mid \mathbf{X}, Z] = \eta = \alpha + \sum_{s=1}^{q} f_s(Z_s) + \sum_{j=1}^{p} \sum_{k=1}^{k_j} f_j^k(x_j^k), \tag{4.4}$$

donde $Z = (Z_1, \ldots, Z_q)$ son las covariables no funcionales. Para la estimación del modelo representaremos los datos funcionales utilizando componentes principales por tanto, x_j^k la puntuación de la componente principal funcional k de X^l.

En nuestro caso, hemos ajustado un modelo FSAM tomando como variable respuesta el logaritmo del factor de inflamación TNF y como variables explicativas las derivadas de los momentos del día y la edad. Notar que se han probado con otras variables escalares como puede ser BMI, pero no hay resultados significativos para el modelo de regresión ajustado.

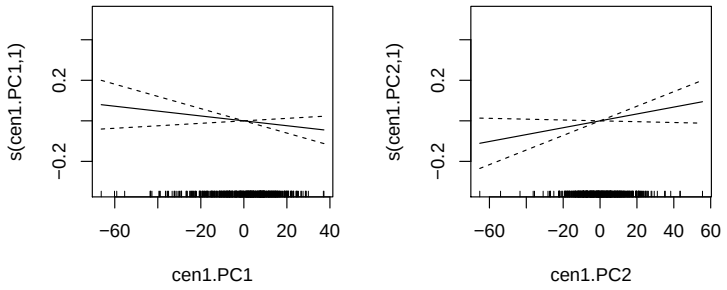

Figura 4.9: Función de regresión de la primera y segunda componente principal de la cena sobre el $\log(TNF)$.

El camino que se ha tomado para estimar el modelo fue representar los datos funcionales utilizando sus dos primeras componentes principales para las derivadas del desayuno, comida y la cena, con las cuales bastarían para resumir las derivadas de los datos funcionales. Además, la única derivada que tiene efecto lineal sobre el logaritmo del TNF es la cena (Figura 4.9), pues los desayunos y comidas presentan un efecto no lineal (Figura 4.10), por tanto sería la única variable funcional que no precisa de una función suavizadora. A la edad también se le aplicará una función suavizadora pues, el efecto que tiene sobre el logaritmo del TNF es no lineal, es decir, el modelo que habrá que estimar es:

$$\text{Modelo } 4 \equiv \log(TNF) \sim f_1(Edad) + \sum_{k=1}^{2} f_2^k\left(\partial(des)^k\right) + \sum_{k=1}^{2} f_3^k\left(\partial(com)^k\right) + \langle \partial(cen), \beta \rangle,$$

donde, f_i para $i = 1, 2, 3$ serían las funciones suavizadoras de la edad y de las derivadas del desayuno y comida, respectivamente. Los resultados de la estimación del modelo descrito serían:

```
Family: gaussian
Link function: identity

Formula:
[1] "logt~+s(Edad,k=-1)+cen1.PC1+cen1.PC2+s(des1.PC1,k=-1)+s(des1.PC2,k=-1)
```

```
     +s(com1.PC1,k=-1)+s(com1.PC2,k=-1)"

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.9601424  0.0116301 168.540   <2e-16 ***
cen1.PC1    -0.0011925  0.0009071  -1.315   0.1893
cen1.PC2     0.0016250  0.0009500   1.711   0.0878 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
edf Ref.df     F  p-value
s(Edad)     3.004  3.787 7.128 2.61e-05 ***
s(des1.PC1) 1.000  1.000 4.518   0.0341 *
s(des1.PC2) 2.454  3.151 0.913   0.4357
s(com1.PC1) 4.746  5.894 1.899   0.0807 .
s(com1.PC2) 1.000  1.000 4.510   0.0342 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.086   Deviance explained = 11.3%
GCV = 0.06813  Scale est. = 0.066007  n = 488
```

En vista de este resultado, la cena no es significativamente distinta a cero, con lo se puede decir que los niveles de glucosa en la cena no tienen efecto sobre el factor inflamatorio TNF. Además, se obtiene una *deviance* explicativa del 11.3 % y un R ajustado de 0.086. A partir de la función de regresión de la edad (Figura 4.10) se aprecia una tendencia creciente respecto la edad del individuo, por lo que podemos asegurar que si los sujetos tienen una edad mayor el TNF es mayor.

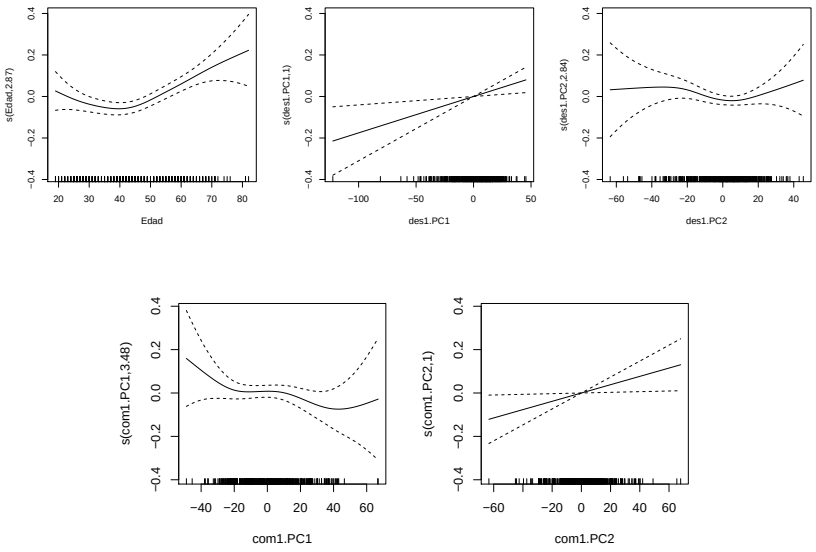

Figura 4.10: Funciones de regresión para las componentes de la edad y de la derivada del desayuno y la comida, con intervalos de confianza al 95 % para el Modelo 4.

En la Figura 4.10 se muestran las funciones de regresión de la edad y de las dos primeras componentes principales de la derivada del desayuno y comida, con su intervalo de confianza al 95 %. Recordemos que las únicas variables significativamente distintas a cero son la edad, la primera componente principal del desayuno y la segunda de la cena. En las tres gráficas, se observa que los intervalos de confianza en los extremos son más amplios, lo que podría conllevar problemas en la predicción.

Por otro lado, se representa la segunda componente principal frente la primera de las derivadas del desayuno y la comida en la Figura 4.11. Bajo normalidad del proceso $\mathcal{X}$, los *scores* de la primera componente principal tendrían media cero y varianza proporcional al primer autovalor; la segunda componente principal análogo al segundo autovalor. Bajo esta suposición, los puntos alejados a la nube de puntos serían los sujetos candidatos a ser *outliers*.

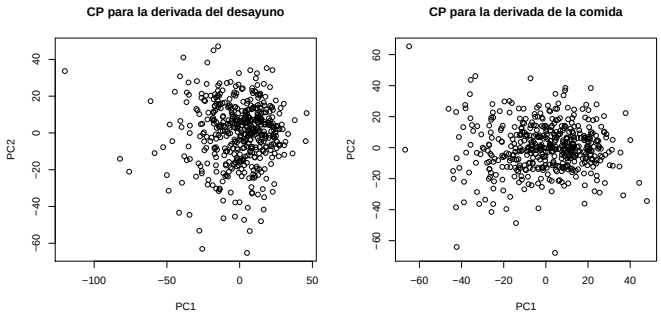

Figura 4.11: Puntuaciones de las dos primeras componentes principales para el desayuno y la comida.

Ahora bien, para saber si el Modelo 4 predice bien o no lo ajustaremos para 450 sujetos, seleccionados aleatoriamente, dejando el resto para realizar predicciones del TNF y compararlas con los valores reales.

Los valores de predicción oscilan alrededor de la media de los valores del TNF (7.36), ya que ajustando 50 modelos diferentes el mínimo valor es de 6.118 y el máximo 8.709, por esta razón para valores extremos (altos y bajos) existe un mayor error de predicción, siendo 4.3 la media de todos los 50 errores cuadráticos. En consecuencia, podemos asegurar que el Modelo 4 no es apropiado para predecir futuros valores de TNF a partir de las curvas de glucosa precisamente por el comportamiento de los extremos del modelo de regresión.

Además, en la Figura 4.12 se muestra un gráfico que aporta información sobre las predicciones del modelo aditivo según las componentes principales de las derivadas de los desayunos y comidas. Los gráficos se componen de parcelas de color y superpuestos, un gráfico de contorno en la escala del predictor lineal, el logaritmo del TNF en nuestro caso.

Para su interpretación nos apoyaremos en la Figura 4.13, en la cual se dibujan las componentes principales para el desayuno y la comida. Por ejemplo, para los desayunos: una curva funcional que tenga valores positivos en la primera componente correspondería a la que más se asemeje con la línea negra del primer gráfico de la Figura 4.13, es decir, que esté por debajo de la media la primera hora y que suba a partir de ésta, el TNF resultante sería bajo (Figura 4.12); para valores negativos de la primera componente sería lo contrario a lo descrito.

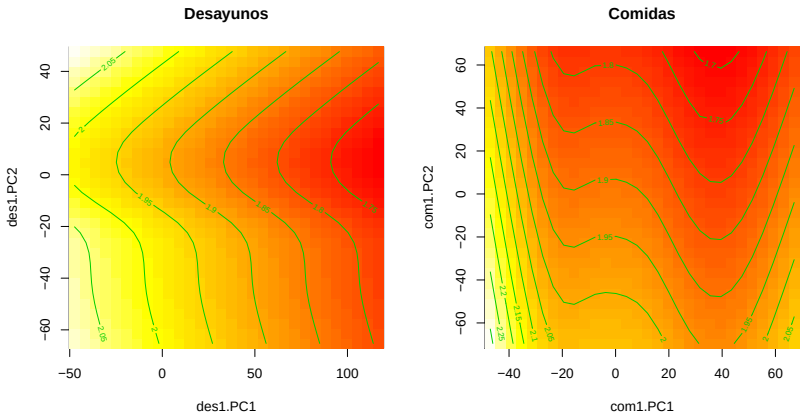

Figura 4.12: Trazado de predicciones del ajuste del Modelo 4 contra los pares de componentes de principales de las derivadas de los desayunos y comidas.

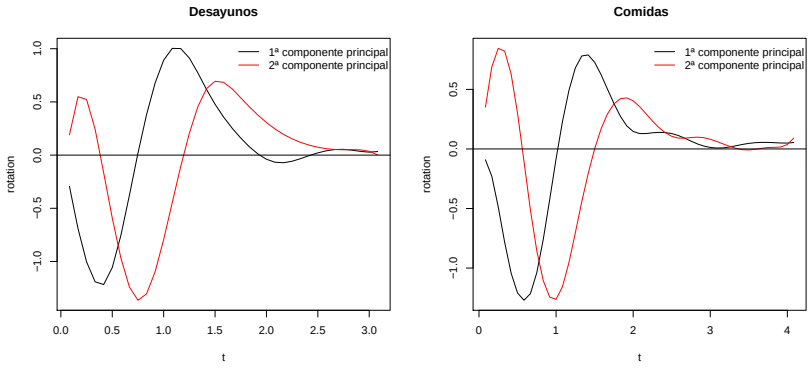

Figura 4.13: Primera y segunda componentes principales del desayuno y comida.

Functional Generalized Kernel Additive Models(GKAM)

Otro camino a tomar, será la aproximación con los **modelos aditivos generalizados tipo núcleo funcionales** (GKAM) (Febrero-Bande and González-Manteiga (2013)). A diferencia de los anteriores modelos, en éstos no se representa la contribución de las componentes principales sino que se estima la contribución directamente a partir de las distancias estimadas por Nadaraya–Watson.

Para estimar conjuntamente las funciones se utiliza el algoritmo de *backfitting*, que consiste en estimar cada componente de la forma siguiente:

$$\hat{f}_j^l(\mathcal{X}^j) = \frac{\sum_{i=1}^n \left(Y_i - \hat{Y}_i^{-j,l}) K_j \left(\frac{d_j(\mathcal{X}^j, \mathcal{X}_i^j)}{h_j}\right)\right)}{\sum_{i=1}^n K_j \left(\frac{d_j(\mathcal{X}^j, \mathcal{X}_i^j)}{h_j}\right)}, \tag{4.5}$$

dónde $\hat{Y}_i^{-j,l} = \sum_{i=1}^{j-1} \hat{f}_i^l(\mathcal{X}^i) + \sum_{i=j+1}^{p} \hat{f}_i^{(l-1)}(\mathcal{X}^i)$ es la predicción de la variable j, d_j es la distancia en el espacio ε_j, K_j es una función asimétrica tipo núcleo y h_j el parámetro ventana.

Con lo cual, la expresión que sigue los modelos GKAM es la siguiente:

$$\mathbb{E}[Y|\mathbf{X}] = \eta = \alpha + \sum_{j=1}^{p} f_j\left(\mathcal{X}^j\right),$$

donde f_j se estimará empleando la expresión (4.5).

Como el estimador sólo calcula las distancias entre datos para cada covariable, esta estimación es aplicable en el espacio funcional métrico. Dónde además, se pueden incluir covariables escalares. El uso de este tipo de estimador asegura la convergencia global del algoritmo y una única solución global bajo las siguientes condiciones de identificación:

- $\mathbb{E}[f_j] = 0$ para $j = 1, \ldots, p$,
- $\beta_0 = 0$ y
- $\mathbb{E}\left[\left(\sum_{j=1}^{p} f_j\right)^2\right] = 1$.

En nuestro caso, se tratará de ajustar el siguiente modelo:

$$\text{Modelo } 5 \equiv \log(TNF) \sim f_1(Edad) + f_2(\partial(des)) + f_3(\partial(com)) + f_4(\partial(cen)),$$

con f_i para $i = 1, \ldots, 4$, las funciones suavizadoras de las covariables, edad y las derivadas del desayuno, comida y cena, que se estimarán por el algoritmo *backfitting*.

Una vez ajustado el modelo obtenemos los resultados siguientes,

```
*** Summary Functional Data Regression with backfiting algorithm ***

Family: gaussian
Link function: identity

alpha= 1.96   n=  488
Algorithm converged? Yes  Number  of iterations  1

****    ****    ****    ****    ****    ****
           h cor(f(X),eta) edf
f(des1) 43.9         0.532 1.7
f(com1) 33.6         0.906 2.5
f(cen1) 47.6         0.039 1.6
****    ****    ****    ****    ****    ****

edf: Equivalent degrees of freedom
Residual deviance= 34.694  Null deviance= 35.171
AIC=  110.365 Deviance explained= 1.4 %
R-sq.= 0.014  R-sq.(adj)= 0.002
Names of possible influence curves: 30 45 103 105 180 281 333 387 416 418
It prints only the 10 most influence curves
```

La *deviance* explicada de la estimación de este modelo es de 1.4 %, más bajo que el modelo anteriormente estimado. En la Figura 4.14 se presentan una serie de gráficos para la diagnosis del modelo:

- En el primer gráfico están representados los valores reales frente a sus predictores lineales, además el R ajustado es de 0.01.

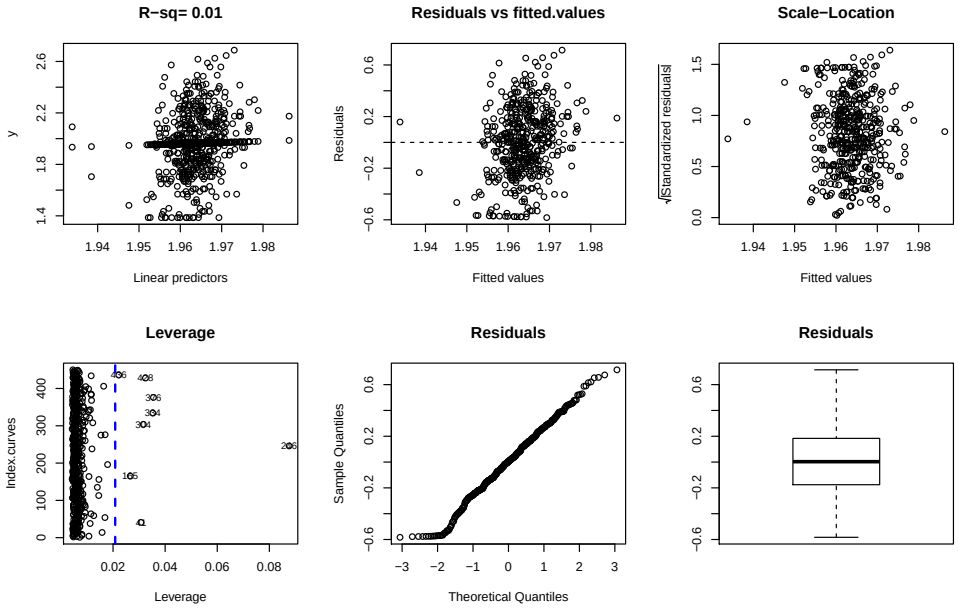

Figura 4.14: Diagnosis del Modelo 5.

- El segundo representa los residuos frente a los valores ajustados, para la validación del modelo, destacar que existen algunos datos que se alejan del resto del modelo.
- El tercero, es el gráfico que se denomina de Localización-Escala pues se representan las raíces cuadradas de los residuos estandarizados (indican la dispersión) frente a los valores ajustados (representan la regresión) por lo que, como en nuestro caso como no se observa una evolución indica la no heterocedasticidad en el modelo.
- Al cuarto se le denomina gráfico *Leverage*, que permite constatar la presencia de datos atípicos, que sería los que se encuentran a la derecha de la línea vertical discontinua.
- El quinto, gráfico QQ-plot, nos aporta un test de normalidad de los residuos estandarizados. En el caso de que se cumpla, los puntos tendría que situarse formando una diagonal.
- El último gráfico, es el *Boxplot* de los residuos.

Ahora bien, realizamos predicciones de un modelo ajustando con 450 sujetos seleccionados aleatoriamente. Con 50 modelos ajustado, el Modelo 5 predice valores cercanos a la media, 7.36 (entre 7.03 y 7.18). Por tanto, la estimación de este tipo de modelos, no es capaz de extraer información relevante.

Modelos tipo *kernel*

Sea Y, una variable respuesta escalar y $\mathcal{X}$ la variable explicativa funcional, los modelos no lineales siguen la expresión siguiente,

$$\mathbb{E}[Y|\mathcal{X}] = m\left(\mathcal{X}\right) + \varepsilon.$$

La función desconocida m se estima de la manera que sigue utilizando una estimación tipo núcleo (Nadaraya-Watson),

$$\hat{m}(\mathcal{X}) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(\mathcal{X},X_i)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{d(\mathcal{X},X_i)}{h}\right)},$$

donde K es una función asimétrica tipo núcleo, h el parámetro ventana y d una distancia en el espacio funcional.

Modelo	R^2-ajustado
$\partial(desayuno)$	0.1566
$\partial(comida)$	0.0968
$\partial(cena)$	0.0355

Cuadro 4.2: R^2-ajustado para tres modelos no lineales tomando como variable explicativa el logaritmo del TNF.

Las fluctuaciones de las curvas de glucosa del desayuno, comida y cena no aportan información de los valores del TNF ya que, en vista del Cuadro 4.2, los modelos no lineales ajustados (tomando las derivadas de las curvas como variable explicativa y variable respuesta el logaritmo del TNF) presentan un R^2-ajustado muy bajo.

***Modelo Parcialmente Lineal* (PLM)**

Los modelos PLM son una extensión de los modelos de regresión tipo *kernel*, en este tipo de modelos también se utiliza un procedimiento tipo núcleo. Sean $Y \in \mathbb{R}$, $\mathcal{X}$ una covariable funcional y $Z = (Z_1, \ldots, Z_q)$ covariables escalares, los modelos PLM siguen la expresión:

$$Y = m(\mathcal{X}) + \sum_{s=1}^{q} Z_j \beta_j + \varepsilon.$$

La función suavizadora desconocida m se estima de la manera siguiente:

$$\hat{m}(\mathcal{X}) = \sum_{i=1}^{n} w_{n,h}\left(\mathcal{X}, X_i\right)\left(Y_i - (Z^i)^\top \hat{\beta}_h\right),$$

donde $w_{n,h}$ se estima utilizando una versión funcional del estimador de Nadaraya-Watson,

$$w_{n,h}(\mathcal{X}, X_i) = \frac{K\left(\frac{d(\mathcal{X}, X_i)}{h}\right)}{\sum_{l=1}^{n} K\left(\frac{d(\mathcal{X}, X_l)}{h}\right)},$$

con parámetro ventana h, K una función tipo núcleo asimétrica y d una distancia en el espacio funcional.

Modelo	R^2-ajustado	p-valor (Edad)
Edad+$\partial(desayuno)$	0.1809	0.0007
Edad+$\partial(comida)$	0.1188	0.0019
Edad+$\partial(cena)$	0.0605	0.0027

Cuadro 4.3: R^2-ajustado para tres modelos semi lineales tomando como variable explicativa el logaritmo del TNF.

A diferencia de los modelos tipo *kernel*, los modelos PLM aumentan ligeramente el R^2-ajustado (Cuadro 4.3). Además la edad en los tres modelos de regresión es significativa, (tercera columna, Cuadro 4.3). En consecuencia, se podría afirmar que la gente con más edad tiene valores de TNF más elevado, por otro lado la influencia de la variabilidad de la glucosa pasa desapercibida.

Capítulo 5

Conclusiones

El objetivo principal del presente proyecto era estudiar la comparación de, realizar un análisis utilizando los índices de variabilidad de la glucosa frente al uso de las curvas completas de la glucosa para el mismo análisis. El estudio se ha basado en como afecta los niveles de glucosa a las citoquinas inflamatorias, como puede ser el factor de necrosis tumoral (TNF), que está relacionada con la diabetes.

En vista, de las dos secciones anteriores podemos concluir que los niveles de glucosa no están relacionados fuertemente con el factor inflamatorio, relacionado con la diabetes, como es el TNF. En el estudio de las fluctuaciones en las curvas de glucosa del desayuno y la comida podría decirse que afectan ligeramente a los valores de TNF. Por otro lado, después del análisis de las medidas resumen de las curvas de glucosa se puede concluir que los estos índices no serían válidos para estudiar el efecto de la glucosa sobre el TNF.

En la Figura 5.1 se presentan los *boxplot* de los errores cuadráticos medio de las predicciones de los diferentes modelos funcionales ajustados para 50 muestras de 450 elegidas aleatoriamente. El modelo que tiene menor intervalo de error en la predicción es el semilineal con la derivada del desayuno. Notar, que los errores de predicción en los modelos FSAM son muy semejantes a los obtenidos con los FKAM.

Alejándonos de nuestro tema, podemos ver que la correlación entre los niveles y fluctuaciones de la glucosa (para el desayuno, comida y cena) y el logaritmo del TNF es muy baja (véase Cuadro 5.1). Para ello, se ha utilizado la correlación de distancia introducida por Székely et al. (2007). Esta correlación es una medida que estudia la dependencia entre los vectores aleatorios de cualquier dimensión y también sirve para variables funcionales. Esta medida toma valores entre 0 y 1 aunque en este caso se hace una corrección por sesgo que puede ofrecer datos negativos.

Variable	Correlación	p-valor
Desayuno	0.0009	0.3796
Comida	-0.0013	0.6782
Cena	-0.0014	0.6840
$\partial(desayuno)$	0.0058	0.0225
$\partial(comida)$	0.0057	0.0253
$\partial(cena)$	-0.0050	0.9574

Cuadro 5.1: Correlación de las variables con el logaritmo del TNF y test de independencia.

Además, en la tercera columna del Cuadro 5.1 se muestra el p-valor del t-test no paramétrico de

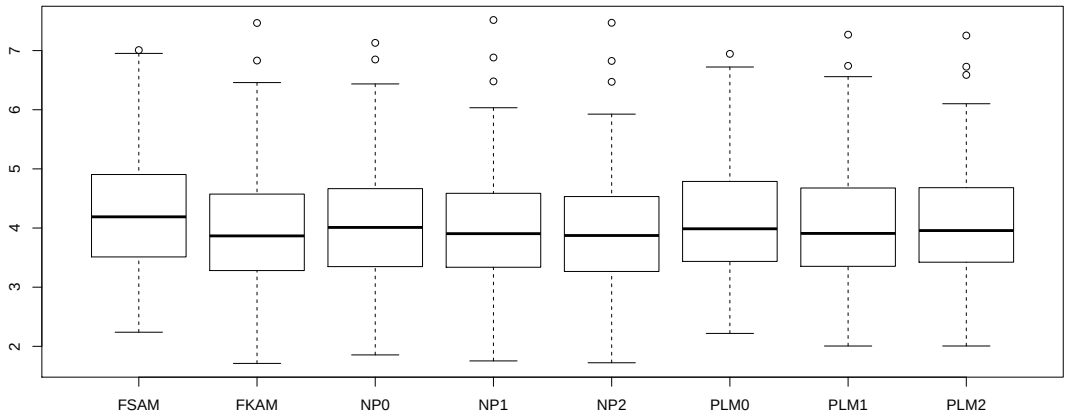

Figura 5.1: *Boxplot* de los errores cuadráticos medios de las predicciones para 50 modelos 5 y 6 con 450 individuos. FSAM $\equiv$ Modelo 4, FKAM $\equiv$ Modelo 5, NP0 $\equiv \partial(des)$, NP1 $\equiv \partial(com)$, NP3 $\equiv \partial(cen)$, PLM0 $\equiv edad + \partial(des)$, PLM1 $\equiv edad + \partial(com)$ y PLM2 $\equiv edad + \partial(cen)$.

la independencia funcional. Con un nivel de significación del 5 % para las derivadas del desayuno y la cena , son las únicas variables funcionales para las que no existen evidencias para rechazar la hipótesis nula de independencia pero, con una correlación muy baja. Este resultado es análogo al obtenido por los modelos de regresión funcionales.

En conclusión, para el estudiar el efecto de las curvas de glucosas sobre un factor, utilizar la metodología de datos funcionales puede aportar mayor información que los índices de variabilidad. Pero, en el caso del TNF, tanto los perfiles de glucosa como sus derivadas, no aportan ninguna información relevante sobre él, en la misma línea de las conclusiones obtenidas por las medidas resumen.

Apéndice A

Hoja de registro para el paciente

Al paciente se le dan las siguientes instrucciones para cubrir la hoja de registro de la monitorización de glucemia intersticial y registro dietético:

El sistema que se le acaba de implantar es un Sistema de monitorización continua de la glucosa (iPro®2). Este sistema permite controlar sus niveles de glucosa durante 24 horas. El sistema está formado un sensor y un receptor que están conectados a su cuerpo y que registra los datos de glucemia.

El equipo que se le ha suministrado es un sistema...SEA CUIDADOSO.

MEDICIÓN DEL NIVEL DE GLUCOSA EN SANGRE:

- Anote los controles de glucemia capilar y la hora (en formato 24h, 00 : 00) de los mismos en esta libreta en lugar donde le han indicado.
- Mida los valores de glucosa en sangre 3 veces al día, por ejemplo: antes del desayuno, del almuerzo y de la cena.
- Utilice siempre el mismo medidor de glucemia capilar.

CUIDADOS Y UTILIZACIÓN:

- Mantega sus hábitos cotidianos.
- Compruebe la zona de inserción 4 veces al día para asegurarse de que el sensor y el receptor estén correctamente colocados (si el sensor y/o el receptor se descoloca por completo métalo en una bolsa de plástico con autocierre y póngase en contacto con el equipo sanitario).
- Retire el sensor si tiene enrojecimiento, dolor, hipersensibilidad o inflamación en la zona de inserción y póngase en contacto con el equipo sanitario.
- Puede ducharse y andar con el sensor y el receptor colocados. El iPro®2 es resistente al agua hasta una profundidad de 2.4 metros durante 30 minutos, no hay límite de tiempo si nada en superficie en una piscina o en la ducha.

Apéndice B

Código R

1. Evaluación de la *performance*

Detalles:
Se evalúa si la monitorización es buena o no, utilizando tres métodos diferentes: la media absoluta y relativa entre las diferencias de los pares de glucosa (capilar y monitorización); el gráfico de Clarke y el gráfico de Bland-Altman.

Código:

```
## Librerías previas
library(MethComp)

## Lectura datos
data <- read.csv2("calibracion.csv", header = T)

## Separando las medidas de capilar y monitorización
desayuno <- data.frame(data$ID, data$Día, data$DNO.SMBG, data$DNO.CGM)
colnames(desayuno) <- c("id", "dia", "smbg", "cgm")
comida <- data.frame(data$ID, data$Día, data$CDA.SMBG, data$CDA.CGM)
colnames(comida) <- c("id", "dia", "smbg", "cgm")
cena <- data.frame(data$ID, data$Día, data$CNA.SMBG, data$CNA.CGM)
colnames(cena) <- c("id", "dia", "smbg", "cgm")

smbg <- c(data$DNO.SMBG, data$CDA.SMBG, data$CNA.SMBG)
cgm <- c(data$DNO.CGM, data$CDA.CGM, data$CNA.CGM)
dat <- data.frame(smbg, cgm, data$Día)

################################### MARD ###################################

## MARD día a día
n <- length(unique(data$Día))
MARD.dnod = MARD.cdad = MARD.cnad = NULL
for(i in 1:n){
    MARD.dnod <-c(MARD.dnod, mean(na.omit(abs(data[data$Día == i,]$DNO.SMBG -
    data[data$Día == i,]$DNO.CGM)/ data[data$Día == i,]$DNO.SMBG))*100)
```

```
    MARD.cdad <-c(MARD.cdad, mean(na.omit(abs(data[data$Día == i,]$CDA.SMBG -
                  data[data$Día == i,]$CDA.CGM)/ data[data$Día == i,]$CDA.SMBG))*100)

    MARD.cnad <-c(MARD.cnad, mean(na.omit(abs(data[data$Día == i,]$CNA.SMBG -
                  data[data$Día == i,]$CNA.CGM)/ data[data$Día == i,]$CNA.SMBG))*100)
}
MARD.dnod;MARD.cdad;MARD.cnad

## MARD de los 6 días juntando las comidas.
MARD.dias = NULL
for(i in 1:n){
    MARD.dias <- c(MARD.dias, mean(na.omit(abs(dat[dat$data.Día == i,]$smbg -
                   dat[dat$data.Día == i,]$cgm)/ dat[dat$data.Día == i,]$smbg))*100)
}
MARD.dias
## MARD total
MARD.total <- mean(na.omit(abs((smbg - cgm))/smbg))*100
MARD.total



################################ CLARKE ERROR GRID ############################
##Librería previa
library(ega)

data <- read.csv2("calibracion.csv", header = T)
base <- read.csv2("Base.csv", header = T)
base <- data.frame(base$Código, base$Sexo, base$DM)
colnames(base) <- c("ID", "Sexo", "DM")

## Juntamos la base con información de DM
m1 <- merge(data, base, by = "ID")

smbg <- c(m1$DNO.SMBG, m1$CDA.SMBG, m1$CNA.SMBG)
cgm <- c(m1$DNO.CGM, m1$CDA.CGM, m1$CNA.CGM)
dat <- data.frame(smbg, cgm, m1$Día, m1$DM )
colnames(dat)<-c( "smbg", "cgm", "dia", "dm")

# Construimos el gráfico de Clarke
plotClarkeGrid(dat$smbg[dat$dm == 1], dat$cgm[dat$dm ==1])
plotClarkeGrid(dat$smbg[dat$dm == 0], dat$cgm[dat$dm ==0])
clard <- getClarkeZones(dat$smbg[dat$dm == 1], dat$cgm[dat$dm ==1])
tclad <- table(clard)
# Diabéticos
prop.table(tclad)*100
clarnd <- getClarkeZones(dat$smbg[dat$dm == 0], dat$cgm[dat$dm ==0])
# No diabéticos
tcland <- table(clarnd)
prop.table(tcland)*100
```

```
################################## BLAND-ALTMAN ##################################
## Se preparan los datos de manera que aparezcan por columnas
## el sujeto, el día, el método y los valores de glucosa.
met <- function(data,momento1, momento2){
    data2 <- data.frame(data$ID,data$Día,momento1,momento2)
    colnames(data2) <- c("id", "dia", "smbg", "cgm")

    met.long <- reshape(data2, varying = c("smbg", "cgm"),
                        idvar = c("id","dia"), times = c("smbg", "cgm"),
                        v.names = "y", direction = "long")

    colnames(met.long) <- c("item", "repl", "meth", "y")
    met.long$repl <- factor(met.long$repl)
    met.long$y <- as.numeric(met.long$y)
    return(met.long)
}

des.long <- met(data, data$DNO.SMBG, data$DNO.CGM)
cda.long <- met(data, data$CDA.SMBG, data$CDA.CGM)
cna.long <- met(data, data$CNA.SMBG, data$CNA.CGM)

## Gráfico de Bland-Altman para los 6 días
## en desayuno, comidas y cenas.

des.long <- na.omit(des.long)
par(mfrow = c(2, 3), mar = c(5,4,4,4)+0.1)
BA.plot(des.long[des.long$repl == 1,])
BA.plot(des.long[des.long$repl == 2,])
BA.plot(des.long[des.long$repl == 3,])
BA.plot(des.long[des.long$repl == 4,])
BA.plot(des.long[des.long$repl == 5,])
BA.plot(des.long[des.long$repl == 6,])

cda.long <- na.omit(cda.long)
BA.plot(cda.long[cda.long$repl == 1,])
BA.plot(cda.long[cda.long$repl == 2,])
BA.plot(cda.long[cda.long$repl == 3,])
BA.plot(cda.long[cda.long$repl == 4,])
BA.plot(cda.long[cda.long$repl == 5,])
BA.plot(cda.long[cda.long$repl == 6,])

cna.long <- na.omit(cna.long)
BA.plot(cna.long[cna.long$repl == 1,])
BA.plot(cna.long[cna.long$repl == 2,])
BA.plot(cna.long[cna.long$repl == 3,])
BA.plot(cna.long[cna.long$repl == 4,])
BA.plot(cna.long[cna.long$repl == 5,])
BA.plot(cna.long[cna.long$repl == 6,])
```

2. Función para el cálculo de los índices de variabilidad

Detalles:
La función `index` da el valor de las medidas resumen de la variabilidad de la glucosa a partir de los datos.

Uso:
`index(data,d,n)`

Argumentos:

`data`	Base de datos con las columnas: código, día y glucosa.
`d`	Tiempo, en minutos, entre una medida de glucosa y la siguiente.
`n`	Para el cálculo de CONGA, la observación realizada `n` horas antes.

Código:

```
index <- function(data, d, n){
     glucosemg <- data$Glucosa
     day <- data$Día
     id <- unique(data$ID)
     glucose <- glucosemg/18
     N <- length(glucose)
     m <- length(table(day)) # número de días
     Th <- 24*m              # número total de horas
     # número de observaciones en una hora
     K <-  dim(data[day == 1, ])[1]
     dat <- cbind(c(1:N), data)
     colnames(dat)[1] <- "num"; orden <- dat$num

     sd <- tapply(dat$Glucosa, dat$Día, sd) # sd por cada día
     IGV <- 120              # valor de glucosa ideal
     MG <- mean(glucosemg)
     SD <- sd(glucosemg)
     CV <- 100*SD/MG
     IQR <- IQR(glucosemg)

     # M-Value
     M <- mean(abs(10*log10(glucosemg/IGV))^3) +
         (max(glucosemg)-min(glucosemg))/20


     J <- 0.001*(MG + SD)^2  # J-Index (mg/dl)


     FG <- 1.509*((log(glucosemg))^1.084 - 5.381)
     rl <- ifelse(FG < 0, 10*FG^2, 0)
     rh <- ifelse(FG > 0, 10*FG^2, 0)
     # Low Blood Glucose Index
     LBGI <- 1/N*sum(rl)
     # High Blood Glucose Index
     HBGI <- 1/N*sum(rh)
```

```
LR = HR = NULL
for(i in 1:m){
     LR <- c(LR, max(rl[day == i]))
     HR <- c(HR, max(rh[day == i]))
}
# Average Daily Risk Ratio
ADRR <- 1/m*sum(LR + HR)

HIPO <- 100*mean((glucosemg < 70))
HIPER <- 100*mean((glucosemg > 140))

# Lability Index
LI <- sum((glucose[1:N-1] - glucose[2:N])^2)/d


# Glycemic risk assessment diabetes equation score
GRADE <- median(425*(log10(log10(glucose)) + 0.16)^2)

# Continous overall glycemic action
CONGA <- sd(glucose[(n+1):N] - glucose[1:(N-n)])

#Mean of Daily Differences
MODD <- sum(abs(glucose[1:(N-K)] - glucose[(1+K):N]))/
       (K*(m-1))

# Área bajo la curva
AREA <- (1/2*5*sum(glucosemg[1:N-1] + glucosemg[2:N]))/
       (m*24*60)

# Cálculos previos
infl = maxi = mini  = diff =  downs = NULL
for (i in 2:(N - 1)){
     if ((glucosemg[i] - glucosemg[i - 1]) *
         (glucosemg[i + 1] - glucosemg[i]) <= 0){
    infl <- c(infl, i)}
}
infl <- c(infl,N)
eps <- 8
n_infl <- length(infl)

for (j in 1:(n_infl)) {
     I1 <- (infl[j] - eps):(infl[j] + eps)
     I <- subset(I1, I1 <= max(infl) & 0 < I1)
     if (max(glucosemg[I]) == glucosemg[infl[j]] |
         min(glucosemg[I]) == glucosemg[infl[j]]) {
          maxi <- c(maxi, infl[j])
          mini <- c(mini, infl[j])
     }
}
mm <- c(sort(c(maxi, mini)), infl[n_infl])
```

```
     def <- ifelse(glucosemg[mm[1:(length(mm)-1)]] ==
                   glucosemg[mm[2:length(mm)]], NA,
                   mm[1:(length(mm)-1)])
     def <- c(subset(def, def != "NA"), infl[n_infl])

     for(k in 1:m){
          ii <- day[def] == k
          deff <- def[ii]
          for (j in 1:length(deff)) {
               diff <- c(diff, glucosemg[deff[j]] -
                         glucosemg[def[j+1]])
          }
     downs <- c(downs, (subset(diff,diff > 0 & diff > sd[k])))
     }

     MAGE <- sum(downs)/length(downs)

     values <- round(c(id,MG, SD, CV, IQR, M, J, LBGI, HBGI,
           ADRR, HIPO, HIPER, LI, MAG, GRADE, MAGE,  CONGA,
           MODD, AREA), 6)
     return(values)
}
```

Observaciones:

Utilizamos el siguiente código para obtener una matriz con los valores de los índices para todos los sujetos con monitorización.

```
lista <- list.files(pattern = ".csv")
ind <- NULL
for(i in lista){
     ind <- c(ind, index(read.csv2(i), 5, 12))
}

indi <- t(matrix(ind, nrow = 19))
colnames(indi) <- c("id","MG", "SD", "CV", "IQR", "M",
                    "J-index", "LBGI", "HBGI", "ADRR", "%Hipo",
                    "%Hiper", "LI", "MAG", "GRADE", "MAGE",
                    "CONGA", "MODD", "AUC")
indi <- data.frame(indi)
```

3. Función para separar los desayunos, comidas y cenas

Detalles:
La función `particion` separa los niveles de glucosa en los momentos del día que se quiera y las horas que se deseen.

Uso:
`particion(data,horas,codigo)`

Argumentos:

`data`	Base de datos con las columnas: código (ID), día, glucosa y el código del momento de la ingesta (por ejemplo, la cena).
`horas`	Número de horas que se desean tomar después de la ingesta.
`codigo`	Momento de la ingesta de comida:1 ≡ desayuno, 2 ≡ media mañana, 3 ≡ almuerzo, 4 ≡ merienda y 5 ≡ cena.

Código:

```
particion <- function(data, horas, codigo) {
     data <- as.data.frame(data)
     h <- (horas*60)/5
     data$come[is.na(data$come)] =  0
     n <- which(data$come == codigo)
     N <- length(table(data$Día))

     da <- NULL
     for(i in 1:(N-1)) {
          cc <- n[i]:(n[i] + h)
          data[cc, ]$Día[data[cc, ]$Día == i+1] = i
          da <- rbind(da,data[cc, ])
     }
     datat <- data.frame(da$ID, da$Día, da$Glucosa)
     datat$time <- rep(1:(h + 1), N-1)
     colnames(datat) <- c("id", "dia", "glucosa", "tiempo")
     datafinal <- reshape(datat, idvar = "dia",  v.names = "glucosa",
     timevar <- "tiempo", direction = "wide")

     return(datafinal)
}
```

Observaciones:

Utilizamos el siguiente código para obtener tres archivos, los niveles de glucosa en los desayunos, comidas y cenas.

```
lista <- list.files(pattern = ".csv")
desayunos = comidas = cenas = NULL
for(i in 1:length(lista)){
    desayunos <- rbind(desayunos, particion(read.csv2(lista[i]), 3, 1))
    comidas <- rbind(comidas, particion(read.csv2(lista[i]), 4, 3))
    cenas <- rbind(cenas, particion(read.csv2(lista[i]), 8, 5))
}
write.table(desayunos, file = "desayunos.txt")
write.table(comidas, file = "comidas.txt")
write.table(cenas, file = "cenas.txt")
```

4. Análisis Cluster

Detalles:
En el siguiente código se realiza un análisis *cluster* de los índices de variabilidad, escogiendo 4 *clusters*

Código:

```
# Seleccionamos las columnas que ocupan los índices
indext <- t(na.omit(base[,32:49]))
# Análisis de componentes principales
acpt <- prcomp(indext)
compt <- predict(acpt)[,1:2]
# Análisis para 4 cluster
kmt <- kmeans(compt, 4 )
plot(compt, col = kmt$cluster)
points(kmt$centers, col = 1:4, pch = 8, cex = 2)
text(compt[,1], compt[,2], labels = rownames(indext), col = kmt$cluster)
```

5. Modelos Aditivos

Detalles:
Se ajustarán diferentes modelos aditivos para estudiar el efecto que tienen los índices de variabilidad sobre el factor inflamatorio TNF.

Código:

```
# Seleccionamos a los no diabéticos
base0 <- base[base$DM == 0,]
# Elimanos a los individuos con TNF alto para conseguir normalidad
base01 <- base0[base0$TNF < 15,]
base01 <- base01[!is.na(base01$TNF),]
# Densidad del TNF
plot(density(base01$TNF), main = "Densidad del TNF")
plot(density(log(base01$TNF)), main = "Densidad del log(TNF)")
base01$logt <- log(base01$TNF)
# Ajuste de los modelo GAM
fit1 <- gam(logt ~ s(Edad) +s(Glu)+ s(BMI)+s(CONGA)  , data = base01)
summary(fit1)
plot(fit1)

fit2 <- gam(logt ~ s(Edad) + s(AUC) + s(CONGA)+SD+MAGE +BMI , data = base01)
summary(fit2)
plot(fit2)

fit3 <- gam(logt~s(Edad) + BMI, data = base01)
summary(fit3)
plot(fit3)
```

5. Análisis Exploratorio de los datos funcionales

Detalles:
Utilizando técnicas estadísticas, como el cálculo de profundidades, se realiza un análisis exploratorio de los datos funcionales (desayunos, comidas y cenas). Así como, la detección de las curvas atípicas.

Código:

```
# Cargamos la librería
library(fda.usc)
datos <- read.table("data.txt", header = T)
datos$BMI <- datos$Peso/(datos$Talla/100)^2
datos <- datos[datos$DM == 0,]
attach(datos)
# Se crean los objetos funcionales
desa <- datos[, 50:86]
des_fda <- fdata(desa)
des_fda$names$ylab <- "Glucosa (mg/dL)"
des_fda$names$xlab <- "Tiempo (horas)"
des_fda$argvals <- ((des_fda$argvals * 5) / 60 )
des_fda$rangeval <- range(des_fda$argvals)


comi <-datos[, 87:135]
com_fda <- fdata(comi)
com_fda$names$ylab <- "Glucosa (mg/dL)"
com_fda$names$xlab <- "Tiempo (horas)"
com_fda$argvals <- ((com_fda$argvals * 5) / 60 )
com_fda$rangeval <- range(com_fda$argvals)

cenn <- datos[, 136:232]
cen_fda <- fdata(cenn)
cen_fda$names$ylab <- "Glucosa (mg/dL)"
cen_fda$names$xlab <- "Tiempo (horas)"
cen_fda$argvals <- ((cen_fda$argvals * 5) / 60 )
cen_fda$rangeval <- range(cen_fda$argvals)

# Curvas de glucosa
par(mfrow = c(1, 3))
plot(des_fda, main = "Desayunos")
plot(com_fda, main = "Comidas")
plot(cen_fda, main = "Cenas")

#################################################
## Profundidades
par(mfrow = c(2,4))

#### DESAYUNO
# Profundidad de Fraiman-Muniz
des.fm <- depth.FM(des_fda,trim = 0.1, draw = TRUE)
# Profundidad modal
```

```
des.m <- depth.mode(des_fda,trim =0.1, draw = TRUE)
# Profundidad por proyecciones aleatorias
des.RP <- depth.RP(des_fda, trim =0.1,draw = TRUE)
# Profundidad por proyecciones aleatorias por método de Tukey
des.RT <- depth.RT(des_fda,trim =0.1, draw = TRUE)

# Construimos la media recortada a mano.
c1 <- cut(des.fm$dep, quantile(des.fm$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(des_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c1], main = "FMD")
c2 <- cut(des.m$dep, quantile(des.m$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(des_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c2], main ="MD")
c3 <- cut(des.RP$dep, quantile(des.RP$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(des_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c3], main = "RPD")
c4 <- cut(des.RT$dep, quantile(des.RT$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(des_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c4], main = "RTD")

#### COMIDA
com.fm <- depth.FM(com_fda, trim = 0.1, draw = TRUE)
com.m <- depth.mode(com_fda,trim = 0.1, draw = TRUE)
com.RP <- depth.RP(com_fda,trim = 0.1, draw = TRUE)
com.RT <- depth.RT(com_fda,trim = 0.1, draw = TRUE)

# Construimos la media recortada a mano.
c1 <- cut(com.fm$dep, quantile(com.fm$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(com_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c1], main = "FMD")
c2 <- cut(com.m$dep, quantile(com.m$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(des_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c2], main = "MD")
c3 <- cut(com.RP$dep, quantile(com.RP$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(com_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c3], main = "RPD")
c4 <- cut(com.RT$dep, quantile(com.RT$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(com_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c4], main = "RTD")

#### CENA
cen.fm <- depth.FM(cen_fda, trim = 0.1,  draw = TRUE)
cen.m <- depth.mode(cen_fda,trim= 0.1,draw = TRUE)
cen.RP <- depth.RP(cen_fda,trim = 0.1, draw = TRUE)
cen.RT <- depth.RT(cen_fda, trim = 0.1,draw = TRUE)

# Construimos la media recortada a mano.
c1 <- cut(cen.fm$dep, quantile(cen.fm$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(cen_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c1], main ="FMD")
c2 <- cut(cen.m$dep, quantile(cen.m$dep, c(0, 0.05, 0.50, 1)),
```

```
          include.lowest = TRUE)
plot(cen_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c2], main = "MD")
c3 <- cut(cen.RP$dep, quantile(cen.RP$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(cen_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c3], main = "RPD")
c4 <- cut(cen.RT$dep, quantile(cen.RT$dep, c(0, 0.05, 0.50, 1)),
          include.lowest = TRUE)
plot(cen_fda, lwd = 2, lty = 1, col = c("gray", "blue", "red")[c4], main ="RTD")


#### Cálculamos las curvas con mayor profundidad
par(mfrow = c(1,3))
plot(des.m$median, col = "red", lty = 1, lwd = 2,ylim = c(90, 140),
     main ="Desayunos")
lines(des.fm$median, col = "blue", lty = 1, lwd = 2)
lines(des.RP$median, col= "green", lty = 1,lwd = 2)
lines(des.RT$median, col = "steelblue4", lty = 2, lwd =2)
legend("topright", legend = c("MD", "FMD", "RPD","RT"),
       col = c("red", "blue", "green","steelblue4"),
       lty = c(1, 1, 2,1), lwd = 2, bty ="n")

plot(com.m$median, col = "red", lty = 1, lwd = 2,ylim = c(90, 140),
     main ="Comidas")
lines(com.fm$median, col = "blue", lty = 1, lwd = 2)
lines(com.RT$median, col = "steelblue4", lty = 1, lwd =2)
lines(com.RP$median, col= "green", lty = 2,lwd = 2)
legend("topright", legend = c("MD", "FMD", "RPD","RT"),
       col = c("red", "blue", "green","steelblue4"),
       lty = c(1, 1, 2,1), lwd = 2, bty ="n")

plot(cen.m$median, col = "red", lty = 1, lwd = 2,ylim = c(90, 140),
     main ="Cenas")
lines(cen.fm$median, col = "blue", lty = 1, lwd = 2)
lines(cen.RP$median, col= "green", lty = 2,lwd = 2)
lines(cen.RT$median, col = "steelblue4", lty = 1, lwd =2)
      legend("topright", legend = c("MD", "FMD", "RPD","RT"),
      col = c("red", "blue", "green","steelblue4"),
      lty = c(1, 1, 2,1), lwd = 2, bty ="n")

#### Curvas con el patrón ideal
# Se calculan el 20% de las curvas más profundas
des.fm2<-depth.FM(des_fda,trim=.8)
com.fm2<-depth.FM(com_fda,trim=.8)
cen.fm2<-depth.FM(cen_fda,trim=.8)
ldes<-which(des.fm2$dep>quantile(des.fm2$dep,.8))
lcom<-which(com.fm2$dep>quantile(com.fm2$dep,.8))
lcen<-which(cen.fm2$dep>quantile(cen.fm2$dep,.8))
# Coinciden los tres métodos
l1 <- intersect(intersect(ldes,lcom),lcen)

des.m2<-depth.mode(des_fda,trim=.8)
```

```
com.m2<-depth.mode(com_fda,trim=.8)
cen.m2<-depth.mode(cen_fda,trim=.8)
l2 <-intersect(intersect(des.m2$ltrim,com.m2$ltrim),cen.m2$ltrim)

des.RT2<-depth.RT(des_fda,trim=.8)
com.RT2<-depth.RT(com_fda,trim=.8)
cen.RT2<-depth.RT(cen_fda,trim=.8)
l3 <-intersect(intersect(des.RT2$ltrim,com.RT2$ltrim),cen.RT2$ltrim)

des.RP2<-depth.RP(des_fda,trim=.8)
com.RP2<-depth.RP(com_fda,trim=.8)
cen.RP2<-depth.RP(cen_fda,trim=.8)
l4 <-intersect(intersect(des.RP2$ltrim,com.RP2$ltrim),cen.RP2$ltrim)

aa <- intersect(intersect(intersect(l1,l2), l3), l4)
par(mfrow = c(1,3))
plot(des_fda[aa], main = " Desayunos", ylim = c(90,140), lty = 1)
legend("topright", legend = c("0070", "0878", "1123", "1192", "1229"),
       lty = 1, col = 1:5, bty = "n")
plot(com_fda[aa], main = " Comidas", ylim = c(90,140), lty = 1)
legend("topright", legend = c("0070", "0878", "1123", "1192", "1229"),
       lty = 1, col = 1:5, bty = "n")
plot(cen_fda[aa], main = " Cenas", ylim = c(90,140), lty = 1)
legend("topright", legend = c("0070", "0878", "1123", "1192", "1229"),
       lty = 1, col = 1:5, bty = "n")


################################################
# Eliminamos los diabéticos
datos <- datos[datos$DM == 0,]
desa <- datos[, 50:86]
des_fda <- fdata(desa)
des_fda$names$ylab <- "Glucosa (mg/dL)"
des_fda$names$xlab <- "Tiempo (horas)"
des_fda$argvals <- ((des_fda$argvals * 5) / 60 )
des_fda$rangeval <- range(des_fda$argvals)

comi <-datos[, 87:135]
com_fda <- fdata(comi)
com_fda$names$ylab <- "Glucosa (mg/dL)"
com_fda$names$xlab <- "Tiempo (horas)"
com_fda$argvals <- ((com_fda$argvals * 5) / 60 )
com_fda$rangeval <- range(com_fda$argvals)

cenn <- datos[, 136:232]
cen_fda <- fdata(cenn)
cen_fda$names$ylab <- "Glucosa (mg/dL)"
cen_fda$names$xlab <- "Tiempo (horas)"
cen_fda$argvals <- ((cen_fda$argvals * 5) / 60 )
cen_fda$rangeval <- range(cen_fda$argvals)
```

```r
# Primeras derivadas
des1 <- fdata.deriv(des_fda,1)
com1 <- fdata.deriv(com_fda, 1)
cen1 <- fdata.deriv(cen_fda, 1)

##################################################
# Estudio de los outliers
par(mfrow = c(1,3))
out1 <- outliers.depth.pond(des1,  nb = 1000, smo = 0.1, trim = 0.01, dfunc=depth.FM)
out_de <- out1$outliers
plot(des1, col = "gray", lty = 1, main = "Desayunos")
lines(des1[out_de], col = "red", lwd = 2, lty = 1)
out1$quantile # cuantil
out1$dep.out  # profundidad de los outliers
out1$quantile - out1$dep.out > 0.01

out2 <- outliers.depth.pond(com1,  nb = 1000, smo = 0.1, trim = 0.01, dfunc= depth.FM)
out_co <- out2$outliers
plot(com1, col = "gray", lty = 1, main = "Comidas")
lines(com1[out_co], col = "red", lwd = 2, lty = 1)
out2$quantile
out2$dep.out
out2$quantile - out2$dep.out > 0.01

out3 <- outliers.depth.pond(cen1,  nb = 1000, smo = 0.1, trim = 0.01, dfunc=depth.FM)
out_ce <- out3$outliers
plot(cen1, col = "gray", lty = 1, main = "Cenas")
lines(cen1[out_ce], col = "red", lwd = 2, lty = 1)
out3$quantile
out3$dep.out
out3$quantile - out3$dep.out > 0.01

# Detectamos los outliers con diferencia mayor a 0.01
outliers <- c("5", "30", "31", "43", "135", "211", "259", "364", "447", "565", "569" )
datos[outliers,]
par(mfrow = c(1,3))
plot(des_fda[outliers], col = 1:11, lty = 1, ylim = c(60,220), main = "Desayunos")
legend(1.7,226.2, legend = c("0007", "0041", "0042", "0058", "0208"), col = 1:5,
       lty = 1, bty = "n")
legend("topright", legend = c("0391", "0525", "0848", "1110","1441","1456"),
       col = 6:11, lty = 1, bty = "n")
plot(com_fda[outliers], col = 1:11, lty = 1, ylim = c(60,220), main = "Comidas")
legend(2.25,226.2, legend = c("0007", "0041", "0042", "0058", "0208"), col = 1:5,
       lty = 1, bty = "n")
legend("topright", legend = c("0391", "0525", "0848", "1110","1441","1456"),
       col = 6:11, lty = 1, bty = "n")
plot(cen_fda[outliers], col = 1:11, lty = 1, ylim = c(60,220), main ="Cenas")
legend(4.3,226.2, legend = c("0007", "0041", "0042", "0058", "0208"), col = 1:5,
       lty = 1, bty = "n")
legend("topright", legend = c("0391", "0525", "0848", "1110","1441","1456"),
       col = 6:11, lty = 1, bty = "n")
```

6. Regresión de los datos funcionales

Detalles:
Se ajustan diferentes modelos de regresión funcionales tomando para todos ellos, como variable, el logaritmo del TNF.

Código:

```
# Elimmamos los outliers
outliers <- c(5, 30, 31, 43, 135, 211, 259, 364, 447, 565, 569 )
datos <- datos[-outliers,]
datos<- datos[datos$DM == 0,]
datos <- datos[datos$TNF < 15,]
datos <- datos[!is.na(datos$TNF),]
attach(datos)

desa <- datos[, 50:86]
des_fda <- fdata(desa)
des_fda$names$ylab <- "Glucosa (mg/dL)"
des_fda$names$xlab <- "Tiempo (horas)"
des_fda$argvals <- ((des_fda$argvals * 5) / 60 )
des_fda$rangeval <- range(des_fda$argvals)

comi <-datos[, 87:135]
com_fda <- fdata(comi)
com_fda$names$ylab <- "Glucosa (mg/dL)"
com_fda$names$xlab <- "Tiempo (horas)"
com_fda$argvals <- ((com_fda$argvals * 5) / 60 )
com_fda$rangeval <- range(com_fda$argvals)

cenn <- datos[, 136:232]
cen_fda <- fdata(cenn)
cen_fda$names$ylab <- "Glucosa (mg/dL)"
cen_fda$names$xlab <- "Tiempo (horas)"
cen_fda$argvals <- ((cen_fda$argvals * 5) / 60 )
cen_fda$rangeval <- range(cen_fda$argvals)

dataf <- datos[,2:30]
dataf <- data.frame(dataf, BMI)
dataf$logt <- log(dataf$TNF)

################################
## Utilizando derivadas

des1 <- fdata.deriv(des_fda,1)
com1 <- fdata.deriv(com_fda, 1)
cen1 <- fdata.deriv(cen_fda, 1)

ldata <- list(df = dataf,   des1 = des1, com1 = com1, cen1 = cen1)

b1.pc0<-create.pc.basis(des1,1:2)
```

```
b1.pc1<-create.pc.basis(com1,1:2)
b1.pc2<-create.pc.basis(cen1,1:2)
basis.x<-list(des1 = b1.pc0, com1 = b1.pc1, cen1 = b1.pc2)

#### MODELOS FSAM
res.gsam<-fregre.gsam(logt ~ s(Edad)+ s(BMI) +s(des1) + s(com1) + s(cen1),data=ldata,
                      basis.x=basis.x)
summary(res.gsam)
res.gsam<-fregre.gsam(logt ~ s(Edad)+s(BMI) +s(des1)+cen1+s(com1) ,data=ldata,
                      basis.x=basis.x)
plot(res.gsam)
summary(res.gsam)

par(mfrow = c(1,2))
pc_com <- fdata2pc(com1, ncomp=2)
pc_des <- fdata2pc(des1, ncomp=2)
plot(pc_des$x[,1:2], main = "CP para la derivada del desayuno")
plot(pc_com$x[,1:2], main ="CP para la derivada de la comida")

## Predicciones
vis.gam(res.gsam, view = c("des1.PC1", "des1.PC2"), plot.type ="contour",
        main ="Desayunos")
vis.gam(res.gsam, view = c("com1.PC1", "com1.PC2"), plot.type ="contour",
        main ="Comidas")

## Puntuaciones de las componentes principales
plot(b1.pc0$basis[1:2], main = "Desayunos", lty = 1)
abline(h=0)
legend("topright", legend  = c("1ª componente principal", "2ª componente principal"),
       col = 1:2, lty = 1, bty = "n")
plot(b1.pc1$basis[1:2], main =" Comidas", lty =1)
abline(h=0)
legend("topright", legend  = c("1ª componente principal", "2ª componente principal"),
       col = 1:2, lty = 1, bty = "n")

#### MODELOS FKAM
res.gkam <- fregre.gkam(logt ~ Edad + des1 + com1 + cen1, data = ldata)
summary(res.gkam)

#### MODELOS TIPO KERNEL
res.np0 <- fregre.np(des1, dataf$logt)
summary(res.np0)

res.np1 <- fregre.np(com1, dataf$logt)
summary(res.np1)

res.np2 <- fregre.np(cen1, dataf$logt)
summary(res.np2)

#### MODELOS PARCIALMENTE LINEALES
res.plm0 <- fregre.plm(logt ~ Edad + des1, data = ldata)
```

```r
summary(res.plm0)

res.plm1 <- fregre.plm(logt ~ Edad + com1, data = ldata)
summary(res.plm1)

res.plm2 <- fregre.plm(logt ~ Edad + cen1, data = ldata)
summary(res.plm2)
```

5. Comparación de los modelos de regresión funcionales

Detalles:
El objetivo es construir un *boxplot* con los errores de predicción cuadráticos para poder comparar los diferentes modelos ajustados.

Código:

```r
err = pred = errk = predk =NULL
errn0 = predn0 = errn1 = predn1 = errn2 = predn2 = NULL
errp0 = predp0 = errp1 = predp1 = errp2 = predp2 = NULL
for(i in 1:50){
    # Ajustamos los modelos para 450 individuos
    l <- sample(1:(dim(dataf)[1]), size = 450)
    ldata <- list(df = dataf[l,], des1 = des1[l], com1 = com1[l], cen1 = cen1[l])
    b1.pc0<-create.pc.basis(des1[l],1:2)
    b1.pc1<-create.pc.basis(com1[l],1:2)
    b1.pc2<-create.pc.basis(cen1[l],1:2)
    basis.x <- list(des1 = b1.pc0, com1 = b1.pc1, cen1 = b1.pc2)
    # Predecimos el resto
    newl <- list(df = dataf[-l,], des1 = des1[-l], com1 = com1[-l], cen1 = cen1[-l])

    res.gsam<-fregre.gsam(logt ~ s(Edad)+ s(des1) +s(com1)  + cen1,data=ldata,
                          basis.x=basis.x)

    pred[i] <- predict.fregre.gsam(res.gsam,newl)
    err[i] <-mean((dataf$TNF[-l]-exp(pred))^2)

    res.gkamp <- fregre.gkam(logt ~ Edad + cen1, data = ldata)
    predk[i]<-predict.fregre.gkam(res.gkamp, newl)
    errk[i] <-mean((dataf$TNF[-l]-exp(predk))^2)

    res.np0p <- fregre.np(des1[l], dataf$logt[l])
    predn0[i] <- predict.fregre.fd(res.np0p, des1[-l])
    errn0 <-mean((dataf$TNF[-l]-exp(predn0))^2)


    res.np1p <- fregre.np(com1[l], dataf$logt[l])
    predn1[i] <- predict.fregre.fd(res.np1p, com1[-l])
    errn1[i] <- mean((dataf$TNF[-l]-exp(predn1))^2)
```

```
    res.np2p <- fregre.np(cen1[l], dataf$logt[l])
    predn2[i] <- predict.fregre.fd(res.np2p, cen1[-l])
    errn2[i] <- mean((dataf$TNF[-l]-exp(predn2))^2)

    res.plm0p <- fregre.plm(logt~ Edad + des1, data = ldata)
    predp0[i] <- predict.fregre.plm(res.plm0p, newl)
    errp0[i] <- mean((dataf$TNF[-l]-exp(predp0))^2)

    res.plm1p <- fregre.plm(logt~ Edad + com1, data = ldata)
    predp1[i] <-predict.fregre.plm(res.plm1p,newl)
    errp1[i] <-mean((dataf$TNF[-l]-exp(predp1))^2)

    res.plm2p <- fregre.plm(logt~ Edad + cen1, data = ldata)
    predp2[i] <-predict.fregre.plm(res.plm2p, newl)
    errp2[i] <-mean((dataf$TNF[-l]-exp(predp2))^2)


    error <- cbind(err,errk,errn0,errn1,errn2,errp0,errp1,errp2)
    colnames(error) <-c("FSAM", "FKAM", "NP0", "NP1", "NP2", "PLM0","PLM1", "PLM2")
}

boxplot(error)
```

6. Correlación

Detalles:
A mayores, calculamos la correlación y el t-test de independencia entre el logaritmo del TNF y las diferentes curvas de glucosa (y sus derivadas).

Código:

```
dcor.xy(des1, dataf$logt)
dcor.xy(com1, dataf$logt)
dcor.xy(cen1, dataf$logt)
dcor.xy(des_fda, dataf$logt)
dcor.xy(com_fda, dataf$logt)
dcor.xy(cen_fda, dataf$logt)
```

Bibliografía

Bergenstal, R. M., Ahmann, A. J., Bailey, T., Beck, R. W., Bissen, J., Buckingham, B., Deeb, L., Dolin, R. H., Garg, S. K., Goland, R., et al. (2013). Recommendations for standardizing glucose reporting and analysis to optimize clinical decision making in diabetes: the ambulatory glucose profile (AGP). *Diabetes Technology & Therapeutics*, 15(3):198–211.

Bland, J. M. and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Castle, J. R., Engle, J. M., El Youssef, J., Massoud, R. G., Yuen, K. C., Kagan, R., and Ward, W. K. (2010). Novel use of glucagon in a closed-loop system for prevention of hypoglycemia in type 1 diabetes. *Diabetes Care*, 33(6):1282–1287.

Ceriello, A., Esposito, K., Piconi, L., Ihnat, M. A., Thorpe, J. E., Testa, R., Boemi, M., and Giugliano, D. (2008). Oscillating glucose is more deleterious to endothelial function and oxidative stress than mean glucose in normal and type 2 diabetic patients. *Diabetes*, 57(5):1349–1354.

Clarke, W. L., Cox, D., Gonder-Frederick, L. A., Carter, W., and Pohl, S. L. (1987). Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10(5):622–628.

Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *TEST*, 22(2):278–292.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package `fda.usc`. *Journal of Statistical Software*, 51(4):1–28.

Hill, N., Hindmarsh, P., Stevens, R., Stratton, I., Levy, J., and Matthews, D. (2007). A method for assessing quality of control from glucose profiles. *Diabetic Medicine*, 24(7):753–758.

Kovatchev, B., Anderson, S., Heinemann, L., and Clarke, W. (2008). Comparison of the numerical and clinical accuracy of four continuous glucose monitors. *Diabetes Care*, 31(6):1160–1164.

Kovatchev, B. P., Cox, D. J., Kumar, A., Gonder-Frederick, L., and Clarke, W. L. (2003). Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data. *DiabetesTechnology & Therapeutics*, 5(5):817–828.

Kovatchev, B. P., Otto, E., Cox, D., Gonder-Frederick, L., and Clarke, W. (2006). Evaluation of a new measure of blood glucose variability in diabetes. *Diabetes Care*, 29(11):2433–2438.

McDonnell, C., Donath, S., Vidmar, S., Werther, G., and Cameron, F. (2005). A novel approach to continuous glucose analysis utilizing glycemic variation. *Diabetes Technology & Therapeutics*, 7(2):253–263.

Molnar, G., Taylor, W., and Ho, M. (1972). Day-to-day variation of continuously monitored glycaemia: A further measure of diabetic instability. *Diabetologia*, 8(5):342–348.

Molnar, G. D., Rosevear, J. W., Ackerman, E., Gatewood, L. C., Taylor, W. F., et al. (1970). Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes*, 19(9):644–655.

Monnier, L., Mas, E., Ginet, C., Michel, F., Villon, L., Cristol, J.-P., and Colette, C. (2006). Activation of oxidative stress by acute glucose fluctuations compared with sustained chronic hyperglycemia in patients with type 2 diabetes. *Jama*, 295(14):1681–1687.

Muggeo, M., Verlato, G., Bonora, E., Zoppini, G., Corbellini, M., and De Marco, R. (1997). Long-term instability of fasting plasma glucose, a novel predictor of cardiovascular mortality in elderly patients with non–insulin-dependent diabetes mellitus the verona diabetes study. *Circulation*, 96(6):1750–1754.

Ryan, E. A., Shandro, T., Green, K., Paty, B. W., Senior, P. A., Bigam, D., Shapiro, A. J., and Vantyghem, M.-C. (2004). Assessment of the severity of hypoglycemia and glycemic lability in type 1 diabetic subjects undergoing islet transplantation. *Diabetes*, 53(4):955–962.

Schlichtkrull, J., Munck, O., and Jersild, M. (1965). The M–value, an index of blood-sugar control in diabetics. *Acta Medica Scandinavica*, 177(1):95–102.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Tai, M. M. (1994). A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. *Diabetes Care*, 17(2):152–154.

Wentholt, I., Kulik, W., Michels, R., Hoekstra, J. L., and DeVries, J. (2008). Glucose fluctuations and activation of oxidative stress in patients with type 1 diabetes. *Diabetologia*, 51(1):183–190.

Wojcicki, J. (1995). J–index. A new proposition of the assessment of current glucose control in diabetic patients. *Hormone and Metabolic Research*, 27(1):41–42.

Wood, S. (2006). *Generalized Additive Models: An introduction with R*. CRC press.