Traballo Fin de Máster

Caracterización de perfís de glicosa en poboación non diabética

Daniel Mato Regueira

Máster en Técnicas Estatísticas

Curso 2015-2016

Proposta de Traballo Fin de Máster

Título en galego: Caracterización de perfís de glicosa en poboación non diabética
Título en español: Caracterización de perfiles de glucosa en población no diabética
English title: Characterization of glucose profiles in a nondiabetic population
Modalidade: Modalidade B
Autor/a: Daniel Mato Regueira, Universidade de Vigo;
Director: Manuel Febrero Bande, Universidade de Santiago de Compostela;
Titor: Francisco Gude Sampedro, Hospital Clínico Universitario de Santiago;
Breve resumo do traballo: A Unidade de Epidemioloxía Clínica do Hospital Universitario de Santiago de Compostela mantén un proxecto activo no que se rexistran os niveis de glicosa de forma continua para un marco de poboación de referencia conxuntamente con variables obtidas a través da análise da sangue e dunha sondaxe médica exhaustiva que inclúe hábitos de vida. O obxectivo deste traballo é relacionar estes perfís coas variables obtidas na sondaxe para determinar diferenzas entre os grupos (se os houbera) así como establecer pautas de control para poboación saudable.
Recomendacións: Ter cursado Datos Funcionais, Regresión Non Paramétrica e Análise Multivariante.
Outras observacións:

Don Manuel Febrero Bande, Catedrático da Universidade de Santiago de Compostela; e Don Francisco Gude Sampedro, Adxunto da Unidade de Epidemioloxía Clínica de Hospital Clínico Universitario de Santiago; informan que o Traballo Fin de Máster titulado

Caracterización de perfís de glicosa en poboación non diabética

foi realizado baixo a súa dirección por Don Daniel Mato Regueira para o Máster en Técnicas Estatísticas. Estimando que o traballo está terminado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 1 de Xullo do 2016.

O director:

Don Manuel Febrero Bande

O titor:

Don Francisco Gude Sampedro

O autor:

Don Daniel Mato Regueira

Agradecementos

Ós meus titores Francisco Gude e Manuel Febrero polo seu apoio na realización deste traballo.

A todos os pacientes que participaron no proxecto de A Estrada, que fixeron posible que se levara a cabo este estudo.

Ás miñas compañeiras María Jesús Pérez e Carla Díaz que realizaron xunto a min as prácticas no Hospital Clínico de Santiago que fixeron o paso por estas moito máis amenas.

E por último, pero non menos importante, á miña familia e amigos polo seu constante ánimo e axuda.

Índice xeral

Resumo

Resumo en galego

A Diabetes Mellitus (DM) é unha enfermidade metabólica producida por unha secreción deficiente de insulina, o que produce un exceso de glicosa en sangue. A súa prevalencia depende tanto de factores propios do individuo como de estilos de vida. Neste traballo utilizarase análise de datos funcionais (FDA) para tratar curvas de glicosa provenientes de individuos do estudo AEGIS.

A FDA é unha parte da estatística que traballa con mostras de funcións aleatorias. Nesta podemos empregar moitas das técnicas máis coñecidas para a análise univariante, cunha adaptación dos conceptos ó novo espazo que se considera (neste caso, o espazo de Hilbert $\mathcal{L}_2$).

O obxectivo deste traballo será dar unha visión exploratoria das curvas de glicosa extraidas do estudo AEGIS, así como atopar individuos pre-diabéticos. Para isto, realizaremos unha análise de outliers e definiremos regras de clasificación adecuadas para poder coñecer que tipo de comportamento teñen os individuos atípicos: diabéticos ou non diabéticos. Por último, realizarase unha versión funcional do test ANOVA para comprobar se existen diferenzas entre pacientes con distintas características.

English abstract

Diabetes Mellitus (DM) is a metabolic disorder caused by a deficient secretion of insulin, which produces an excess of glucose in the blood. Its prevalence depends on factors such as the individual's own lifestyle. This paper used functional data analysis (FDA) to treat glucose curves coming from individuals of the AEGIS study.

The FDA is a part of the statistic that works with samples of random funcions. In this we can use many of the techniques known for the univariate analysis, an adaption of the concepts that the new space is considered (in this case, the Hilbert space $\mathcal{L}_2$).

The purpose of this project is to give an insight exploratory of the glucose curves extracted from the AEGIS study, as well as finding individuals with pre-diabetes. For this, we will conduct an analysis of outliers and define appropriate classification rules to know what kind of behaviour individuals have atypical: diabetics or non-diabetics. Finally, there will be a functional version of the ANOVA test to check whether there are differences between patients with different characteristics.

Prefacio

A diabetes é unha enfermidade crónica e irreversible do metabolismo na que se produce un exceso de glicosa ou azucre na sangue e en ouriños, xeralmente debida a unha diminución da secreción da hormona insulina ou a unha deficiencia da súa acción.

A glicosa[1] é o principal factor nutritivo para os músculos e é a única fonte de enerxía para o cerebro. Esta entra no corpo a través da dixestión, pero tamén pode ser liberada polo fígado e músculos, onde se almacena glicóxeno; e, en menor parte, polos riles.

Xeralmente, os niveis de glicosa en xaxún están comprendidos entre 70 mg/dl e 110 mg/dl pero estes aumentan ata os 120-140 mg/dl tras a inxesta de glicosa, 1 gr por cada kg de peso do individuo. Ademais, outras características como dieta, idade ou outros hábitos de vida poden influír nos niveles de glicosa dun individuo.

Para regular o metabolismo deste carbohidrato, o corpo emprega dúas hormonas: a insulina, que é segregada ante niveis altos de glicosa facilitando a absorción e a utilización desta polos tecidos; e o glicagón, segregada ante niveis baixos de glicosa estimulando así a conversión de glicóxeno en glicosa e permitindo manter os niveis desta nun rango saudable. Cando esta regulación falla, o individuo entrará en estado de hipoglucemia ou hiperglucemia, caracterizados por niveis baixos e altos de glicosa en sangue respectivamente. É neste tipo de situacións onde aparece o trastorno do que estamos a falar: a Diabetes Mellitus.

O aumento na prevalencia de diabetes ó que estamos asistindo nos últimos anos foi erixido nun dos maiores desafíos que debemos afrontar tanto desde un punto de vista clínico como en termos de saúde pública. As estimacións da Federación Internacional de Diabetes cifran en 366 millóns as persoas que actualmente teñen diabetes no mundo e estímase que para o ano 2030 haberá 552 millóns, presentando a maioría deles diabetes mellitus tipo 2 (DM2) (90-95 %) (Whiting D.R. 2011).

A diabetes pódese clasificar en catro categorías clínicas:

- Diabetes tipo 1. Debido á destrución das células β no páncreas que conduce a unha deficiencia de insulina
- Diabetes tipo 2. Debido a un defecto da secreción da insulina nun contexto de resistencia periférica ó efecto da mesma.
- Diabetes debido a outras causas: defectos xenéticos, enfermidades do páncreas, tóxicos e fármacos.
- Diabetes xestacional, aquela que se produce durante o embarazo.

A hiperglucemia crónica conleva lesións en múltiples tecidos, con danos especialmente sensibles nos pequenos vasos (microanxiopatía) da retina, os riles e os nervios periféricos. Por isto, a diabetes é unha das principais causas de cegueira, amputacións e enfermidade renal terminal nas sociedades desenroladas.

Adicionalmente, a diabetes conleva un importante risco de enfermidades cardiovasculares (ECV) (macroanxiopatía), tanto por sí mesma como pola súa asociación a outros factores de risco, como hipertensión arterial e dislipemia.

[1] Tamén coñecida como dextrosa, é un carbohidrato pertencente ó subgrupo dos monosacáridos ou azucres simples.

Tendo en conta que algunhas modificacións nos estilos de vida así como algúns fármacos poden previr ou atrasar a aparición de DM2 nos individuos con risco elevado, resulta crucial desenrolar ferramentas de predición de risco para o seu uso en programas de prevención e screening baseado en poboacións (Gillies C.L. 2007).

Como quedou demostrado nos ensaios clínicos, as complicacións micro- e macro-vasculares da diabetes son debidas principalmente á disglicemia, a cal a súa vez ten dous compoñentes: a hiperglucemia crónica sostida e as flutuacións glicémicas agudas. Ambos compoñentes conducen ás complicacións da diabetes a través de dous mecanismos principais: glicación excesiva das proteínas e activación do estrés oxidativo (Brownlee M. et al 2005). A variabilidade glicémica tamén a un aumento de mortalidade nas unidades de críticos en non diabéticos con hiperglicemia de estrés (Eslami S. et al 2011). A variabilidade glicémica é un fenómeno fisiolóxico que se refire ás flutuacións da glicosa ó longo do tempo, e que pode describirse mediante variabilidade intra-día, con diferencias entre os valores de glicemia en xaxún; ou pos-prandiales ou mediante variabilidade entre-días.

Para cuantificar a variabilidade glicémica propuxéronse distintos índices, aínda que non existe un medida estrela universalmente aceptada. A forma máis sinxela consiste en calcular a desviación estándar das medicións de glicosa e/ou o seu coeficiente de variación. Moitos estudos sobre a variabilidade glicémica tamén utilizan o MAGE (mean amplitude of glycemic excursions) introducido por Service en 1970, que ademais se considera a métrica estándar na medición da variabilidade glicémica (Service F.J. et al. 1970). Este índice incorpora soamente datos das excursións maiores. É posible obter estes índices a partir da curva de glicosa en 7 puntos. Non obstante, perderíanse moitos picos e nadires das excursións glicémicas simplemente porque estes pasasen entre dúas medicións.

A aparición de novas tecnoloxías na monitorización continua de glicosa (CGM) permítenos coleccionar gran cantidade de datos fiables de glicosa en persoas sometidas a actividades habituais da súa vida diaria, coa vantaxe de obter información clínica adicional importante, que non sería posible obter cos sistemas de monitorización intermitente da glicosa. A pesar diso, o índice MAGE non puido predicir o desenrolo de retinopatía ou nefropatía nunha cohorte de diabéticos tipo 1 (Kilpatrick E.S. et al. (2009)). Resulta razoable especular que os achados contraditorios con respecto á relación entre a variabilidade da glicosa e as complicacións da diabetes poderían deberse en parte ás limitacións derivadas ó uso destes índices.

As curvas de glicosa, ó igual que outras sinais biolóxicas, teñen propiedades lineais e non lineais que poden analizarse por métodos estatísticos, pero que dificilmente se poden resumir nun só índice. A natureza oscilatoria dos perfís glicémicos relaciónase coas diferentes condicións clínicas (inxesta, exercicio físico) e bioquímicas (radicais libres, sensibilidade á insulina, marcadores inflamatorios) que modulan as súas traxectorias.

É práctica común empregar medidas resumo para describir os niveis de glicosa, incluso en estudos nos que se utiliza monitorización continua. Non obstante, diferentes curvas poden presentar medidas resumo similares, con perda de información clínica ou fisiolóxica de interese. A análise de datos funcionais (FDA) usa unha serie de técnicas estatísticas desenroladas especificamente para analizar curvas, que nos proveña dunha aproximación relativamente novidosa e extremadamente útil para a construción de modelos de predición procedentes da CGM.

Neste sentido, este traballo centrarase no aproveitamento da análise de datos funcionais poder extraer información poboacional a través da monitorización continua da glicosa provinte do estudo AEGIS.

A Estrada Glycation and Inflammation Study (AEGIS) é un estudo de corte transversal e base poboacional realizado no municipio de A Estrada (Pontevedra), cuxa fase de recollida de datos e mostras concluíu en Xuño de 2015, e na que participaron 1516 persoas maiores de 18 anos, elixidas mediante mostreo aleatorio. Financiado polo Instituto de Salud Carlos III baixo o título: “Niveis de hemoglobina glicosilada e gap de glicación en relación con estilos de vida e as enfermidades prevalentes na poboación xeral adulta”. Desde Novembro do 2012 a Marzo do 2015, todos os participantes acudiron ó Centro de Saúde onde foron entrevistados mediante un cuestionario estruturado que incluía: datos demográficos, antropométricos e enfermidades crónico-prevalentes, estilos de vida, unha batería de probas psicolóxicas, presión arterial, etc. Ademais, os participantes foron convidados a participar no proxecto glicación,

que incluíu a monitorización continua da glicosa e o rexistro exhaustivo da dieta durante 6 días máis. Finalmente participaron neste subproxecto 622 individuos, dos que 581 completaron polo menos dous días de monitorización.

Para poder realizar esta análise, empregaremos o programa estatístico de libre disposición R. Ademais, co gallo de aplicar a análise de datos funcionais, debemos dispoñer de paquetes como o paquete **fda** (Wickham H. 2015), que é o paquete básico e o máis empregado para o tratamento de datos funcionais; o paquete **rainbow** (Shan H.L. 2016), sobre todo para a representación de datos funcionais (aínda que, como veremos, tamén se pode usar para encontrar certas profundidades e outliers) e o **fds** (Hydman J.R. 2015), para a análise de series temporais funcionais . Por último, o paquete que máis imos empregar é o **fda.usc** (Febrero Bande M. et al 2016), que nos proporciona un marco máis amplo para este tipo de análise. Este último paquete foi implementado polo grupo da Universidade de Santiago de Compostela, complementando e estendendo algunhas funcións do paquete **fda**.

Os obxectivos que se perseguirán con este traballo será realizar unha análise exploratoria das curvas de glicosa extraidas da monitorización continua onde se conseguirá un entendemento básico dos datos e as relacións existentes entre elas. Ademais explicarase e aplicarase técnicas para atopar datos outliers, onde se intentará identificar curvas pertencentes a individuos que non están diagnosticados como diabéticos pero sí que posúen comportamentos distintos ós da poboación xeral. Logo, axustarase unha regra de clasificación para estes datos e, por último, realizaranse tests ANOVA para buscar diferencias nas distribucións de ditas curvas.

Este traballo estruturarase da seguinte maneira: comezarase explicando as técnicas para unha análise exploratoria en datos funcionais, introducindo pequenos exemplos para unha mellor compresión; na seguinte sección tratarase da busca de outliers, polo que se terá que explicar antes a profundidade. Logo, propoñerase unha revisión da teoría de clasificación en datos funcionais (tanto supervisada como non supervisada) así como a versión funcional do test Anova; e por último, aplicaremos todo o anteriormente exposto ó estudo AEGIS.

Capítulo 1

Análise exploratorio en datos funcionais

A maneira máis tradicional de realizar un estudo estatístico é mediante observacións univariantes ou multivariantes dos individuos. Non obstante, grazas ás últimas melloras tecnolóxicas que se están producindo tanto a nivel de precisión como de rapidez informática; están aparecendo novas formas de estudar os datos. Unha delas é a análise de datos funcionais.

Neste capítulo, introduciranse dito tipo de datos, así como se explicará a súa motivación e como se pode realizar unha análise exploratoria destes mesmos. Ademais, acompañarase o estudo teórico con exemplos sinxelos sempre que se considere oportuno.

1.1. Datos funcionais? Que son e cal é a súa motivación

Neste capítulo realizarase unha análise exploratoria, onde se intenta relevar características coñecidas e evidentes do estudo. Este tipo de análise non se preocupa de cuestións globais para a poboación nin de eventos non observados dos datos.

Como se sabe, na análise multivariante clásica tómanse como variables vectores que se miden en cada individuo. Non obstante, os datos funcionais caracterízanse pola evolución dunha variable medida ó longo do tempo de xeito que os valores de cada individuo van ser unha función.

Os obxectivos da análise de datos funcionais son basicamente os mesmos que os de outra forma de análise estatística que son, por exemplo:

- Representar os datos de maneira correcta para realizar unha análise detallada.
- Mostrar os datos co gallo de describir varias características.
- Buscar fontes de patróns e variación nos datos.
- Explicar a variación dunha variable dependente mediante o uso de información doutras variables independentes.

Para tratar de chegar a estes obxectivos débese presentar en primeiro lugar as seguintes definicións de variable funcional e conxunto de datos funcionais.

Definición 1.1.1. (Ferraty F. e Vieu P. 2006) Unha variable aleatoria $\mathcal{X}$ é unha variable funcional se toma valores nun espazo completo e normado (ou seminormado) $\mathcal{E}$.

Definición 1.1.2. (Ferraty F. e Vieu P. 2006) Un conxunto de datos funcionais $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ son as n observacións para as $\mathcal{X}_1, \ldots, \mathcal{X}_n$ variables funcionais identicamente distribuídas como $\mathcal{X}$.

Para facernos unha idea de como serían uns datos funcionais, recurriremos a monitorización do estudo AEGIS. Na Figura 1.1, preséntase as curvas do almorzo de toda a poboación do estudo de AEGIS con monitorización. En azul están os non diabéticos e en vermello os diabéticos.

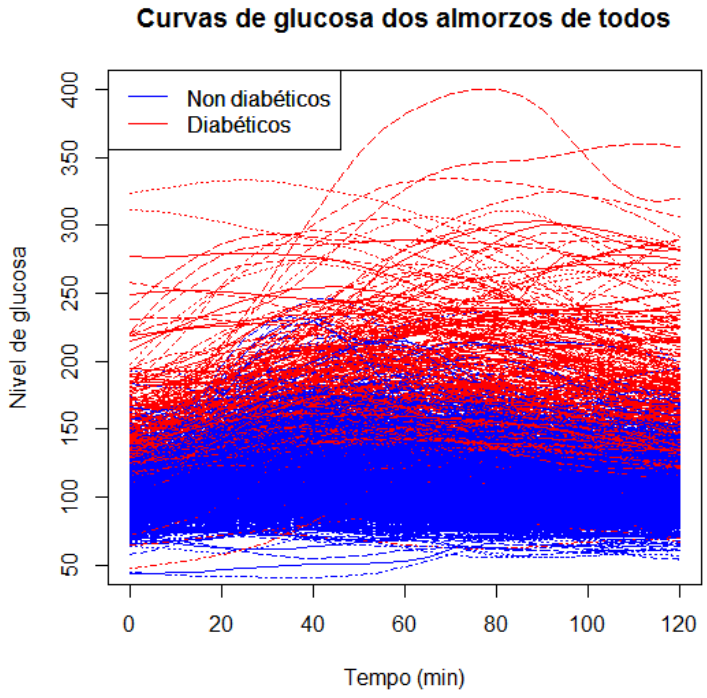

Figura 1.1: Representación das curvas de glicosa para a poboación do estudo de AEGIS. Os non diabéticos están en azul e os diabéticos en vermello.

Como podemos observar, a determinación das características dunha serie de datos funcionais non é tarefa sinxela cun simple gráfico de dispersión das curvas. Isto é debido a que a forma da curva depende da medida de proximidade entre elas.

Polo tanto, é importante dar outras medidas para atopar características xerais dos datos. Estas poden ser a media e a mediana como medida de localización; e a varianza e covarianza como medida de dispersión. Non obstante, a definición destas veñen dadas a través dunha función *distancia*, polo que debemos escoller antes un xeito adecuado de calculala.

A maioría dos espazos para os datos funcionais son espazos métricos completos onde só existe a noción de distancia entre elementos do espazo. Se a métrica pode ser expresada como $d(X(t), Y(t)) = ||X(t) - Y(t)||$ con unha norma $||.||$ verificando a desigualdade triangular, temos un espazo normado ou espazo de Banach[1]. Nestes espazos hai tamén unha noción de tamaño dos elementos no espazo. Se

[1]Nótese que non todos os espazos vectoriais son espazos de Banach. Para que o sexan, necesitan definirse sobre eles unha relación de equivalencia de tal maneira que as clases de equivalencia (formadas por funcións iguais en case todas as partes (é dicir, en todo menos nun conxunto de medida nula) si constitúan un espazo vectorial normado.

a norma verifica a lei do paralelogramo[2], o produto interior pode ser definido no espazo da seguinte maneira:

$$< x, y >= \frac{1}{4}(||x+y||^2 - ||x-y||^2)$$

Entón estaríamos nun espazo de Hilbert xa que é un espazo completo e cun produto interior asociado. O exemplo máis coñecido deste tipo de espazos é o espazo $\mathcal{L}_2[a,b]$ de funcións cadrado integrables definido en $[a,b]$ con

$$< f, g >= \int_a^b fg$$

Polo tanto, dependendo do espazo no que se estea traballando, usaremos distintas distancias convertendo así o problema de elixir un espazo en fundamental para o estudo.

Se nos centramos en espazos $\mathcal{L}_p$[3], usarase a regra de Simpson[4] para medir a distancia entre os elementos. É dicir, se $f(t) = X_1(t) - X_2(t)$,

$$||f||^p = \left(\frac{1}{\int_a^b w(t)dt} \int_a^b |f(t)|^p w(t)dt \right)^{\frac{1}{p}}$$

onde w son os pesos.

Non obstante, tamén se pode considerar as distancias entre curvas baixo a suposición de que pertenzan a espazos métricos ou semimétricos. En tal caso, podemos computar a distancia para o caso semimétrico en $\mathcal{L}_2$ das derivadas de orde q, por exemplo, do seguinte xeito:

$$d_q(f,g) = \sqrt{\frac{1}{\sqrt{T}} \int_T (f^{(q)} - g^{(q)}(g))^2}$$

Outros exemplos están expostos en Ferraty F. e Vieu P. (2006).

Por último, destacar que tanto os casos comentados coma os da referencia están implementados no paquete **fda.usc**.

Non obstante, o espazo $\mathcal{E}$ pode non ser un espazo de Hilbert, polo que se necesitaría unha representación máis flexible. Por esta razón, de agora en diante imos empregar na medida do posible a clase $fdata$ do paquete **fda.usc**, que usa simplemente os valores avaliados nunha malla de discretización de puntos $\{t_1, ..., t_n\}$ que poden ser non equidistantes.

Aínda así, non estamos exentos de problemas, xa que deste xeito debemos realizar todos os cálculos (como cálculo de distancias, por exemplo) mediante aproximacións numéricas. Ademais, a densidade da malla pode ser demasiado alta, afectando así á precisión de ditos cálculos.

Por todo isto, para definir unha variable funcional no programa estatístico R utilizamos o comando $fdata$ do paquete **fda.usc**. Para isto, necesitamos especificarlle, como mínimo, as seguintes opcións:

---

[2] A norma do paralelogramo é a que se cumpre cando:

$$||x+y||^2 + ||x-y||^2 = 2(||x||^2 + ||y||^2)$$

para x, y elementos do espazo.

[3] Conxunto de funcións cuxo valor absoluto se eleva a p-ésima potencia ten integral finita.

[4] A regra de Simpson ou regra de Kepler é un método de integración numérica que se utiliza para obter a aproximación da seguinte integral:

$$\int_a^b f(x)dx \approx \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

data: Matriz de dimensións $n \times m$ contendo o conxunto de n curvas discretizadas en m puntos.

argvals: Localización dos puntos de discretización (por defecto $t_1 = 1$ e $t_m = m$).

rangevals: Rango de puntos de discretización (por defecto tómase o rango de *argvals*).

Ademais de poder definir datos funcionais, despois de ter esta clasificada como *fdata*, pódese calcular a derivada do dato funcional co comando *fdata.deriv* con distintos métodos (tanto de maneira numérica ou a través da representación por bases). A elección do método dependerá do caso no que se estea e da malla de discretización.

Visto isto, vólvese ás curvas de glicosa para todos o individuos da mostra nos almorzos. Estas están medidas dúas horas despois de dita comida cada 5 minutos. Polo tanto, como se xustificou, crearemos un obxecto da clase *fdata* do seguinte xeito:

```
> funcional<-fdata(base1[,3:27],argvals=seq(0,120,5))
```

A partir disto, pódese realizar unha pequena ollada os datos con cálculos simples, como a representación dos datos funcionais separando a poboación segundo a cantidade de hemoglobina glicada (pigmento vermello contido nos hematíes da sangue que está alterada pola presenza de cantidades altas de glicosa) en sangue de cada paciente, Figura 1.2; ou segundo sexa diabético ou non, Figura 1.1.

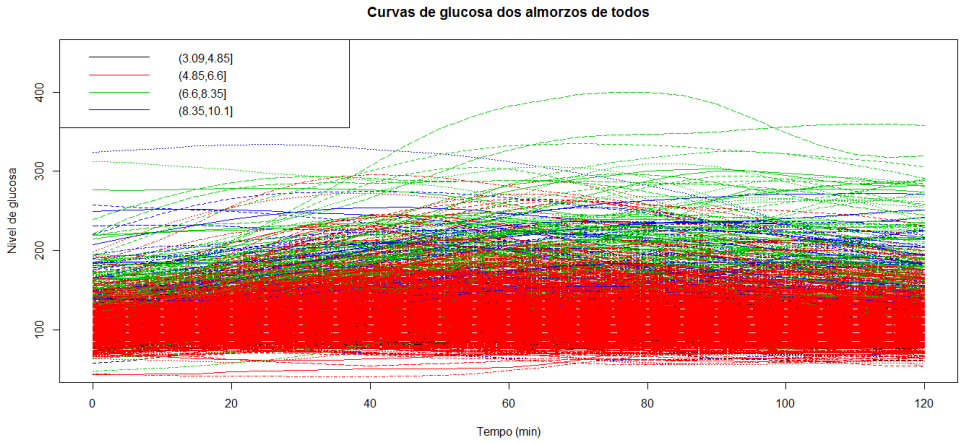

Figura 1.2: Representación das curvas de glicosa para a poboación do estudo de AEGIS separando a cantidade de hemoglobina glicada do paciente.

Como mencionamos xa, a escolla dun espazo apropiado para traballar é un aspecto importante. Neste caso, polo feito de que estamos a tratar de curvas de glicosa, supoñerase que o espazo no que realizamos estes cálculos é o $\mathcal{L}_2$, considerando así a distancia como a norma definida neste espazo.

Outra opción considerada nesta sección é o cálculo da derivada. Aínda que para a realización desta precísanse empregar unha base b-spline, que se explicarán máis adiante, pódese calcular e representar esta na Figura 1.3. Nótese que neste caso, aínda que menos evidente que nas curvas orixinais, os diabéticos (vermello) son os que presentan máis flutuacións, aparentemente.

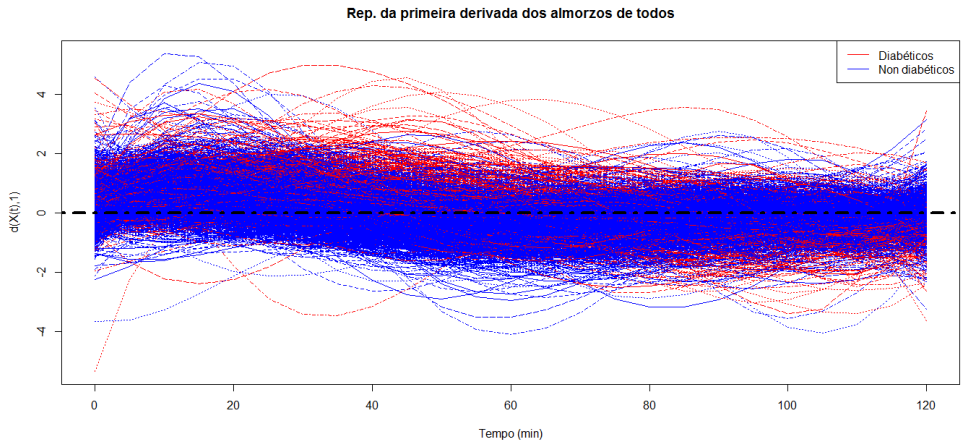

Figura 1.3: Representación da derivada das curvas de glicosa para a poboación do estudo de AEGIS. Os diabéticos están en vermello e os non diabéticos en azul.

1.2. Representación

O seguinte paso que se debe realizar cando traballamos con datos funcionais é a súa representación. En xeral, terase un dato funcional observado nun conxunto discreto de puntos $\{t_j\}_{j=1}^{T} \in [a,b]$ e trataremos de realizar a representación dos datos en base $\mathcal{L}_2$ (ou en bases penalizadas), baseado en compoñentes principais funcionais, baseado en compoñentes funcionais parciais por mínimos cuadrados e baseado en métodos de suavizado kernel.

1.2.1. Representación por bases

Definición 1.2.1. (Ramsay J.O. e Silverman B.W. 2005) Unha base é un conxunto de funcións $\{\phi_k\}_{k\in\mathbb{N}}$ tales que calquera función pode ser representada como unha ponderación dun número o suficientemente grande k_n destas funcións.

Unha curva pode ser representada por unha base cando asumimos que os datos pertencen a un espazo $\mathcal{L}_2$. Sexa entón $\mathcal{X} \in \mathcal{L}_2(T)$ con $\mathbb{E}(\mathcal{X}(t)) = 0$, $t \in [0,T]$ e $y \in \mathcal{R}$ con $\mathbb{E}(y) = 0$. O modelo de regresión linear funcional pode escribirse deste xeito:

$$y =< \mathcal{X}, \beta > +\epsilon \overset{(a)}{=} \int_T \mathcal{X}(t)\beta(t) + \epsilon$$

sendo β un elemento de $\mathcal{L}_2(T)$ e en (a) consideramos que estamos en dito espazo. Unha maneira de estimar os parámetros é mediante unha base en $\mathcal{L}_2$ do seguinte xeito:

$$\beta(t) = \sum_{k=1}^{K_\beta} \theta_k(t) \Rightarrow \beta = \theta' b$$

e

$$\mathcal{X}_i(t) = \sum_{k=1}^{K_x} c_{ik}\psi_k \Rightarrow \mathcal{X} = C\psi(t)$$

Polo que se pode sobrescribir o modelo do seguinte xeito:

$$y = C\psi\theta' b + \epsilon = Zb + \epsilon, \text{ con } b = (Z'Z)^{-1}Z'y$$

Se ademais reescribimos o produto entre ψ e θ tal que $J_{\psi\theta} = (< \psi_i, \theta_j >)_{ij}$, obtemos a seguinte expresión:

$$\hat{y} = CJ_{\psi\theta}by = Zb = Z(Z'Z)^{-1}Z'y = Hy$$

Véxase que unha vez chegado a este punto, pódese concluír que a escolla dunha base apropiada para os nosos datos é imprescindible. As bases máis comúns son as de Fourier, B-splines, Wavelets, etc. En xeral, non hai ningunha regra universal que decida que base é mellor escoller para uns datos en concreto, aínda que sí que hai "pistas". Por exemplo, se os datos son periódicos, o mellor é escoller unha base de Fourier mentras que se buscamos rapidez no cálculo deberiamos encamiñarnos cara os B-splines. De seguido, imos explicar as bases máis coñecidas.

Base poligonal

Suavizar os datos observados non sempre é necesario, e especialmente se non temos interese en axustar os datos en sí, senón que estamos interesados nalgún parámetro funcional que non están ligados ós datos directamente. Nos capítulos do libro Ramsay J.O e Silverman B.W. (2005) nos que explica o modelo lineal funcional vese que se poden interpolar os datos cunha base simple e levar o tema de suavizado á estimación do parámetro funcional desexado. De feito, datos lineais a trozos ou poligonais son moi recomendables e poden ofrecer unha estimación aproximada da primeira derivada.

Aínda así, esta opción non é tan usada como a base B-spline ou a base de Fourier.

Base de B-splines

A base máis común para datos non periódicos son as funcións spline.

Definición 1.2.2. (de Boor C. 1978) Unha función f dise spline polinómico de grao m se satisfai:

1. $f(x)$ é $(m-1)$ veces continuamente diferenciable
2. $f(x)$ é un polinomio de grado m para $x \in [k_j, k_{j+1})$ con $j = 1, \ldots, m-1$.

Por tanto, cada spline polinómico pode ser representado por unha base de $d = (m+l-1)$ funcións, da seguinte maneira:

$$f(x) = \sum_{j=1}^{d} \beta_j B_j(x)$$

O punto de corte dos subintervalos son os chamados "nodos".

Entón unha posible base flexible local é a formada por Basic-splines (B-splines) (de Boor C. 1978). Os B-splines de grado m obtéñense fusionando $(m+1)$ polinomios de grao M suavemente nos $(m-1)$ nodos interiores. Matematicamente pódense expresar do seguinte xeito:

- **B-spline de grao** $m = 0$

$$B_j^0(x) = I_{[k_j, k_{j+1})}(x) = \begin{cases} 1 & se \quad k_j \leqslant x < k_{j+1} \\ \\ 0 & noutro \quad caso \end{cases}$$

- **B-spline de orde superior**: Calcúlanse de maneira recursiva:

$$B_j^m = \frac{x - k_j}{k_{j+m} - k_j} B_j^{m-1}(x) + \frac{k_{j+m+1} - x}{k_{j+m+1} - k_{j+1}} B_{j+1}^{m-1}(x)$$

Por último, destacar que unha das mellores características deste método é o seu rápido funcionamento e o sinxelo cálculo das súas derivadas.

Base de Fourier

A expansión básica máis coñecida é a proporcionada polas series de Fourier:

$$x_i(t) = c_0 + c_1 \operatorname{sen}(wt) + c_2 \cos(wt) + c_3 \operatorname{sen}(2wt) + ...$$

definida pola base

- **De grao 0**:
$$\phi_0(t) = 1$$
- **De grao par**:
$$\phi_{2r}(t) = \cos(rwt)$$
- **De grao impar**:
$$\phi_{2r-1}(t) = \operatorname{sen}(rwt)$$

con $r = 1, 2, 3, ...$.

Esta base é periódica (con período $2\pi/w$). Se os valores de t se escollen equiespaciados no intervalo T e o período é igual á lonxitude de T, entón a base é ortogonal no sentido de que a matriz do produto cruzado $\phi_i'\phi_i$ é diagonal e pode ser igual á identidade dividindo as funcións base polas constantes adecuadas, $\sqrt{m}$ para grao 0 e $\sqrt{m/2}$ para o resto de graos; sendo m o número de observacións.

A versión ortonormal da base de Fourier é coñecida como base ortonormal[5] de funcións trigonométricas en $\mathcal{L}_2$ e ven dada pola seguinte expresión:

- **De grao 0**:
$$\phi_0(t) = 1/\sqrt{T}$$
- **De grao par**:
$$\phi_{2r}(t) = \frac{\cos(rwt)}{\sqrt{T/2}}$$
- **De grao impar**:
$$\phi_{2r-1}(t) = \frac{\operatorname{sen}(rwt)}{\sqrt{T/2}}$$

Por último, destacar que a Transformada Rápida de Fourier permite atopar eficientemente todos os coeficientes cando m é potencia de 2 e os argumentos son equiespaciados. Neste caso, podemos atopar os coeficientes c_j e os m valores suavizados de $x(t)$ en $\mathcal{O}(m \log m)$ operacións. Pola contra, os B-splines e as wavelets poden igualar e incluso superar a súa eficiencia computacional.

[5]Unha familia $\{v_i\}_{i \in I}$ de elementos dun espazo $\mathcal{E}$ é un sistema ortogonal se $< v_i, v_j >= 0$. Se ademais $||v_i|| = 1$, $\forall i \in I$, dise que é un sistema ortonormal.(Fernandez F.J. 2012)

Base de Wavelets

A diferencia coa transformada de Fourier, o representación por bases de wavelets non asume que os datos sexan periódicos, polo que se utilizarán moitas menos funcións que as que se precisarían se se utilizaran funcións seno e coseno para alcanzar unha aproximación adecuada dos datos funcionais.

As wavelets úsanse como funcións básicas para representar outras como se fai coas función seno e coseno na base de Fourier.

Podemos construír unha base para todas as funcións en $(-\infty, \infty)$ que sexan cadrado integrables escollendo unha adecuada wavelet nai ψ e considerando todas as dilatacións e translacións da forma:

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$$

para uns enteiros j e k. Constrúese a wavelet nai para asegurar que a base é ortogonal, no sentido de que a integral do produto de calquera dúas funcións base distintas é cero.

Entón, a idea de base de wavelet é facilmente adaptable para tratar funcións definidas nun intervalo limitado, dunha maneira máis fácil que se houbese imposicións sobre a periodicidade da fronteira.

A expansión wavelet dunha función f proporciona unha análise de multiresolución no sentido de que os coeficientes de ψ_{jk} dan información sobre f próxima á posición $2^{-j}k$ sobre a escala 2^{-j}, isto é, en frecuencias próximas a $c2^k$ para algunha constante c.

En consecuencia, wavelets da unha secuencia sistemática de grados de localización. Ó contrario das series de Fourier, as expansións wavelet traballan ben con descontinuidades e con rápidos cambios dos datos.

Supoñamos unha función x observada sen erro en nodos equiespaciados no intervalo T. Entón hai unha transformación wavelet discreta (DWT) de xeito que proporcione coeficientes m_i relacionados con coeficientes da función x. Así, podemos calcular a DWT e a súa inversa en $\mathcal{O}(m_i)$ operacións. Se supoñemos que as observacións de x teñen ruído, o feito de que moitas clases de funcións teñan expansións wavelet conduce a unha simple aproximación suave non linear.

1.2.2. Representación por compoñentes principais

Outra ferramenta para a representación de datos funcionais son as compoñentes principais funcionais (FPCA)[6](Ramsay J.O. e Silverman B.W. 2005). Estas intentan explicar os datos funcionais a través da combinación ortonormal das variables, intentando maximizar a varianza.

Usando este método, unha dato funcional calquera virá representado na "autobase", que é unha base ortonormal do espacio de Hilbert $\mathcal{L}_2$. Sexa entón $\mathcal{X}(t) \in \mathcal{L}_2(T)$ e $\Sigma(s,t) = \mathbb{E}[(\mathcal{X}(s) - \bar{\mathcal{X}}) - (\mathcal{X}(s) - \bar{\mathcal{X}})]$ e o operador linear $T_\Sigma : \ f(t) \to \int_T \Sigma(s,t) f(s) ds$. Entón, podemos falar de autovalores λ e autovectores v_k que resolven a ecuación:

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

Os autovectores maximizan a varianza e son ortogonais con cada un deles, $\{v_i\}_{i\in\mathbb{N}}$ forman unha base ortogonal de $\mathcal{L}_2(T)$, ou sexa, $\mathcal{X} = \sum_{i=1}^{\infty} < \mathcal{X}, v_i > v_i$. Ademais, $Z_i =< \mathcal{X}, v_i >$ verifica que $\mathbb{E}[Z_i] = 0$, $\forall i \in \mathbb{N}$ e $\mathbb{E}[Z_i, Z_j] = \delta_{ij}\lambda_k$ para todo $i, j \in \mathbb{N}$.

Nótese que as bases de compoñentes principais son as máis efectivas para resumir a información de $\mathcal{X}$.

1.2.3. Representación por modelos PLS

Na anterior subsección viuse que as compoñentes principais funcionais poden ser unha boa elección para poder representar os datos nun número reducido de dimensións. Pero ademais da variable funcional, nos estudos estatísticos xeralmente dispoñen de máis variables, como variables escalares. Neste

[6]Do inglés *Functional principal component analysis*

caso, pódese usar directamente a información adicional aplicando mínimos cadrados parciais funcionais (FPLS).

No artigo Preda C. e Saporta G. (2005) tratan dunha maneira amena os PLS. A idea na que se basean para o aproveitamento dos PLS é a construción dun conxunto de variables aleatorias $\{\nu_i\}_{i\geqslant 1}$ no espazo linear estendido por $\mathcal{X}$ tendo en conta a covarianza existente entre a variable funcional e a escalar.

Os compoñentes PLS serán obtidos da seguinte maneira:

1. Defínese $y_0 = y - \bar{y}$ e $\mathcal{X}_0 = \mathcal{X} - \bar{\mathcal{X}}$. Sexa ademais l=0.
2. Sexa $t_{l+1} =< \mathcal{X}_l, w_{l+1} >$, onde $w_{l+1} \in \mathcal{L}_2$ de tal maneira que $Cov(y_l, t_{l+1})^2$ é maximal. Entón:
$$w_{l+1} = \frac{Cov(y_l, \mathcal{X}_l)}{||Cov(y_l, \mathcal{X}_l)||}$$
3. Sexa $y_{l+1} = y_l - u_{l+1}t_{l+1}$ e $\mathcal{X}_{l+1} = \mathcal{X}_l - \nu_{l+1}t_{l+1}$ onde:
$$u_{l+1} = \frac{Cov(y_l,t_{l+1})}{Var[t_{l+1}]} \qquad \nu_{l+1} = \frac{Cov(\mathcal{X}_l,t_{l+1})}{Var[t_{l+1}}$$
4. Sexa $l = l + 1$ e volvemos ó paso 2.

Mediante este proceso iterativo chegamos a:

$$\mathcal{X} = \bar{\mathcal{X}} + \sum_l t_l\nu_l \qquad y = \bar{y} + \sum_l u_l t_l + e$$

Ó igual que pasou na representación por compoñentes principais, necesitamos un método para realización da estimación e poder así aplicalo a casos reais.

Sexa entón $X = \mathcal{X}_i(\mathcal{T}_j))$ a matriz de dimensión $n \times T$ na que están as avaliacións dos datos funcionais na malla de discretización $\{\mathcal{T}_j\}_{j=1}^{T}$ e sexa tamén o vector resposta y de tamaño $n \times p$. Entón, a estimación realizarase seguindo o seguinte esquema:

1. Seleccionar un vector de pesos w distinto de 0 de lonxitude T (a primeira compoñente principal ou unha fila de X son exemplos válidos) e normalizalo.
2. Calcular o vector de puntuacións $t = Xw$. Conseguiremos un vector de lonxitude n.
3. Calcular o vector de y-cargas denotado por $q = y't$. Dito vector terá dimensión $p \times 1$.
4. Calcular o vector de y-puntuacións $u = yq$ onde u terá dimensión $n \times 1$.
5. Calcular o novo vector de pesos $w_1 = X'u$ e normalízase.
6. Se $||w - w_1|| < \epsilon$, conseguiuse a converxencia do método. No caso contrario, tomarase $w = w_1$ e debemos volver ó paso 2.

Así conseguimos un par (t, u) de puntuacións para X e y respectivamente. De todos xeitos, estes pasos poderían ser resumidos tendo en conta que o que se conseguiu non é máis que o primeiro autovector das matrices $X'YY'X$ e $XX'YY'$.

Por último, deberíamos estimar os compoñentes (p, b) de X e y. Isto faise do seguinte xeito:

1. Calcular o vector de cargas $p = \frac{X't}{t't}$.
2. Cambiamos X calculando $X_1 = X - tp'$
3. Realizar a regresión de Y sobre t: $b = \frac{y't}{t't}$
4. Axustar y usando b: $y_1 = y - tb'$
5. No caso de necesitar máis, entón debemos tomar $X = X_1$ e $y = y_1$ e volver ó primeiro paso.

1.2.4. Representación por suavización

De igual xeito que podemos aplicar ideas de análise multivariante clásico como compoñentes principais para representar os datos funcionais, tamén podemos aproveitar a metodoloxía non paramétrica, en concreto o método de suavización tipo kernel. Pero ó igual que ocorría en multivariante, aparece o problema de escolla correcta dun estimador e dun parámetro ventá.

Supoñamos que temos unha observación

$$y(t_j) = \mathcal{X}(t_j) + \epsilon(t_j)$$

onde $\epsilon(t_j)$ representa o ruído orixinario ó medir os datos con matriz de covarianzas $\Sigma_\epsilon = W^{-1}$.

Podemos volver ós datos orixinais cun suavizador linear:

$$\bar{\mathcal{X}}(t_j) = \sum_{i=1}^{T} s_{ij}(t_i) y(t_i)$$

con s_{ij} son os pesos dos puntos t_j e $y(t_i)$ son os valores observados de y no punto t_i.

Entón, tendo en conta que imos realizar unha suavización tipo kernel, a suavización non paramétrica de datos funcionais ven dada pola matriz S tal que:

$$s_{ij} = \frac{1}{h} K\left(\frac{t_i - t_j}{h}\right)$$

Como imos traballar co paquete **fda.usc**, imos fixarnos nas opcións máis importantes que trae. Con este paquete, podemos calcular a matriz de suavización S mediante:

- O método de Nadaraya Watson no que:

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

  onde K é a función kernel e h o parámetro ventá.

- O k-ésimo veciño máis próximo no que:

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

  onde K é a función kernel uniforme e h_l é o parámetro ventá dependente do punto de onde se estima.

Unha vez visto isto, salta á vista que nos queda por mencionar os tipos de kernels. No paquete **fda.usc** pódense aplicar os seguintes:

- Gausiano: $k(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
- Epanechnikov: $K(u) = \frac{3}{4} 1_{[-1,1]}(1 - u^2)$
- Triweigth: $K(u) = \frac{35}{32} 1_{[-1,1]}(1 - u^2)^3$
- Uniforme: $K(u) = \frac{1}{2} 1_{[-1,1]}(u)$
- Coseno: $K(u) = \frac{\pi}{4} 1_{[-1,1]} \cos(\frac{\pi u}{2})$
- Cuadrático: $K(u) = \frac{15}{16} 1_{[-1,1]}(1 - u^2)^2$

1.2.5. Validación cruzada e validación cruzada xeneralizada

The choice of the parameter number of basis and the most appropriate basis for the observed data is also vital and, in principle, there is no universal rule that would enable an optimal choice. The decision on what basis to choose should be based on the objective of the study and on the data.

A elección do número de parámetros da base e a escolla da base máis apropiada para os datos observados, xunto coa busca dun parámetro de suavización adecuado (segundo o caso que esteamos a considerar) é crucial. Para resolver estes problemas contamos cunha múltiple variedade de criterios de selección de parámetros. Non obstante, no paquete **fda.usc** están implementadas dúas: a validación cruzada (CV) e a validación cruzada xeneralizada (GCV). Estas teñen as seguintes expresións:

$$CV(\nu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i}^{\nu})^2 w_i$$

$$GCV(\nu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i}^{\nu})^2 w_i \Xi(\nu)$$

onde $\hat{y}_{(-i)}^{\nu}$ indica o estimador que se conseguiu ó extraer o par (t_i, y_i), w_i é o peso no punto t_i e $\Xi(\nu)$ denota o tipo de función de penalización. Estas funcións de penalización poden ser: validación cruzada xeneralizada, criterio de información de Akaike, erro de predición finito, modelo selector de Shibata e selector de parámetro ventá de Rice. A expresión de todas estas funcións de penalización pódense ver en Febrero Bande M. et al (2016).

1.2.6. Exemplo práctico

Volvamos entón ás curvas dos almorzos da poboación estudada polo estudo AEGIS. Analizando as características destes datos, chegamos a conclusión de que a base que mellor se adecúa a eles tanto pola súa precisión e rapidez computacional é a base de B-splines. Escolleremos ditas bases co seguinte comando:

```
> create.bspline.basis(rangeval=funcional$rangeval,nbasis=10)
```

onde se ten en conta o rango dos valores dos datos funcionais. Representando estas bases obtemos a Figura 1.4, B-splines para ditos datos con 10 bases.

Como vimos, tamén se pode intentar explicar os datos funcionais a través da combinación ortonormal das variables, intentando maximizar a varianza. É dicir, aplicando unha representación mediante compoñentes principais.

Aplicamos entón a función *fdata2pc* obtendo a seguinte saída de R:

```
> princomfun1<-fdata2pc(funcional,ncomp=3)
> summary(princomfun1)

     - SUMMARY:   fdata2pc  object   -

-With 3  components are explained  98.06 %
 of the variability of explicative variables.

-Variability for each component (%):
  PC1   PC2   PC3
83.98 10.16  3.92
```

Pola cantidade de variabilidade explicada polas compoñentes principais, o lóxico será que traballemos coas dúas primeiras pois con elas supérase o 90 % da variabilidade explicada. Ademais de obter

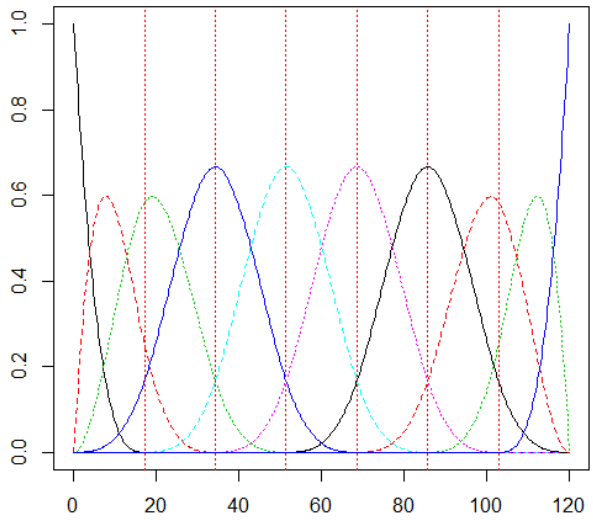

Figura 1.4: B-splines para os almorzos con 10 bases.

a saída anterior na consola, cando facemos un summary, aparece a Figura 1.5, onde se mostran as compoñentes principais (na diagonal) e o biplot asociado ós scores de cada compoñente.

De igual xeito, podemos calcular as compoñentes principais da derivada dos datos funcionais, anteriormente calculada. Non obstante, o número de compoñentes principais que son necesarias para a representación de ditos datos elévase a 5 para conseguir acadar o 90 % da variabilidade explicada:

```
> princomder1<-fdata2pc(derivada1,ncomp=5)
> summary(princomder1)

     - SUMMARY:   fdata2pc  object   -

-With 5  components are explained  94.26 %
 of the variability of explicative variables.

-Variability for each component (%):
  PC1   PC2   PC3   PC4   PC5
34.62 31.97 11.05 10.28  6.33
```

Realizando unha pequena revisión ós datos que temos, podemos usar a hemoglobina glicada, medida en cada individuo, para representar os datos mediante modelos PLS. A idea básica é construír un conxunto de compoñentes PLS nun espazo linear tendo en conta a correlación que hai entre estas dúas variables. Isto podémolo facer en R coa seguinte liña de comando:

```
> pls2<-fdata2pls(funcional,bas$a1c1,ncomp=3)
```

Como podemos ver na Figura 1.6, a porcentaxe de variabilidade explicada polas dúas primeiras compoñentes diminuíu con respecto a considerar só ó dato funcional. Non obstante, a variabilidade da terceira compoñente aumentou considerablemente.

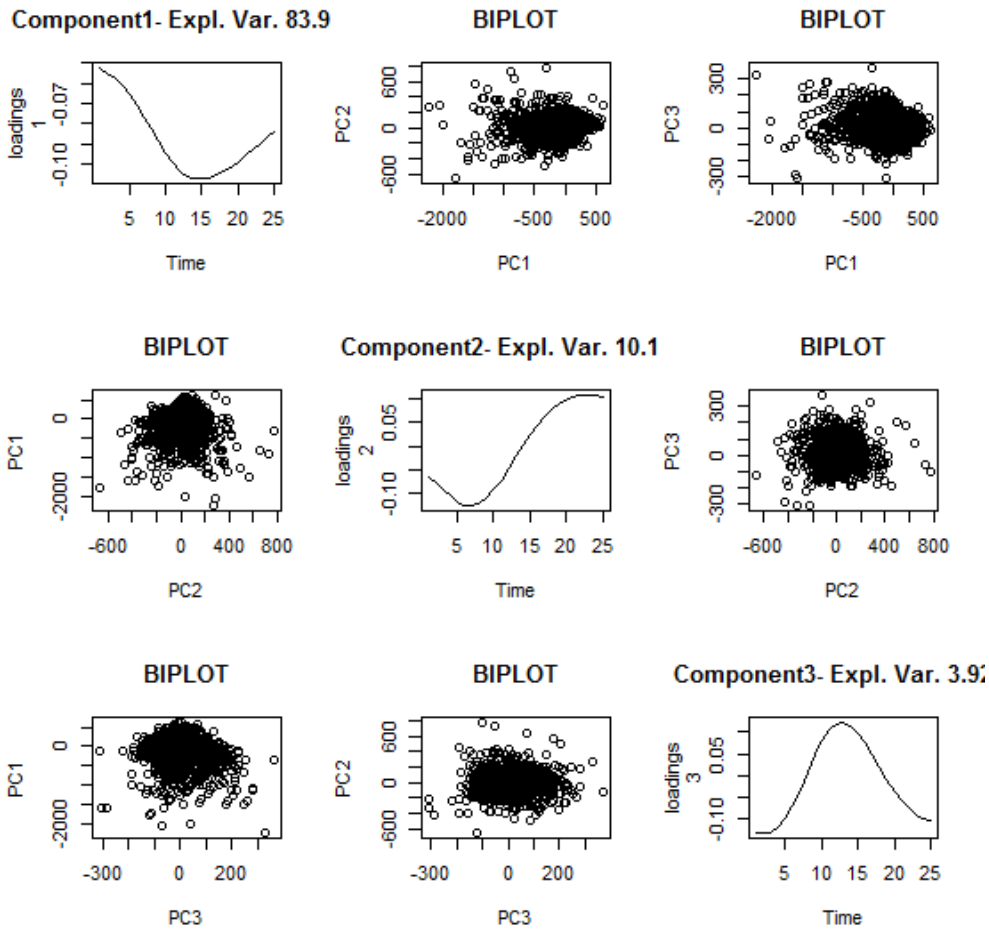

Figura 1.5: Resumo das compoñentes principais para os datos funcionais orixinais.

Por outra banda, pódese considerar a representación tanto en bases B-spline, Fourier ou estimacións tipo kernel (tanto co estimador Nadaraya-Watson como con k veciños máis próximos). Polo tanto, escollemos unha curva das curvas de glicosa e, para ela, representamos ditas representacións. Como resultado obtemos a Figura 1.7.

Claramente, a aproximación que máis se lle asemella é a estimación kernel con estimador de Nadaraya-Watson. Por isto, realizamos dita aproximación mediante a liña de comando:

```
> primeiro<-min.np(funcional,h=seq(1,7,length=15),type.S=S.NW)
```

E se expoñemos os datos funcionais fronte a base que utilizamos obtemos a Figura 1.8.

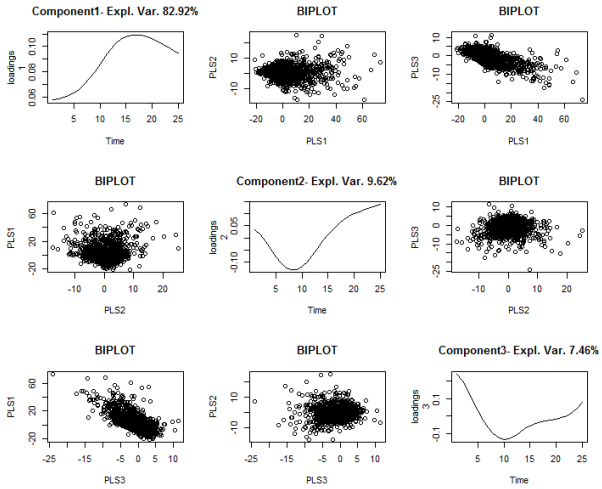

Figura 1.6: Resultado de aplicar un summary ó modelo $pls1$.

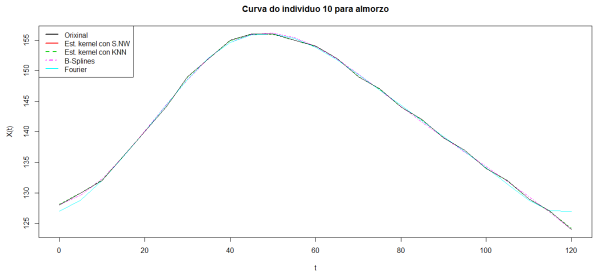

Figura 1.7: Curva de glicosa do individuo 10 (en negro), xunto coa aproximación kernel (con estimador Nadaraya-Watson en vermello e con k veciños máis próximos en verde), coa estimación con bases de Fourier (en azul) e mediante B-splines (rosa).

1.3. Medidas de dispersión e localización

Nesta sección trataremos as medidas de localización e dispersión, así como aplicar estas a un exemplo práctico. Recordemos que trataremos os datos ó espazo de Hilbert $\mathcal{L}_2$,

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

1.3.1. Medidas de localización

Agora que xa se ten establecido o espazo de funcións, procederase ó cálculo de medidas de localización. En primeiro lugar, calcularase a media, xa que é a medida máis popular de localización. Para o caso de datos funcionais, a media ou centro de gravidade dos datos ten a seguinte expresión:

$$\min_{a \in \mathbb{F}} \sum_{\mathbb{F} \in S} d(\mathcal{X}, a)^2$$

En espazos $\mathcal{L}_2$, a media mostral defínese analogamente á anterior:

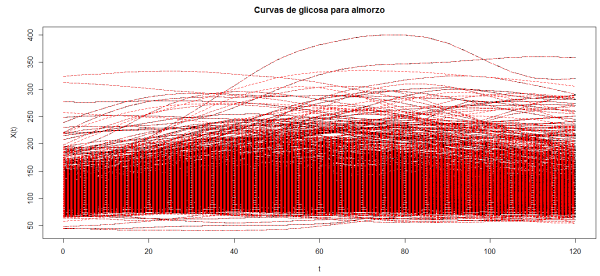

Figura 1.8: Representación da base orixinal (en negro) e a base suavizada con estimador de Nadaraya-Watson en vermello.

$$\min_{a \in S_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a)^2$$

De igual forma se poden definir medias en funcións das diferentes métricas que poden ser usadas. Para explicalas, estudaranse en dúas situacións, sexan ou non considerados espazos de Hilbert.

- **Espazo de Hilbert**: Supoñamos que $\{\mathcal{X}_i\}_{i=1}^n$ é o noso conxunto de datos funcionais onde cada elemento se pode representar mediante a base $\{\psi_j\}_{j \in \mathbb{N}}$. Entón temos que

$$\mathcal{X}_i = \sum_{j \in \mathbb{N}} c_{ij} \psi_j$$

Sexa entón $\bar{\mathcal{X}}_i = \sum_{j \in \mathbb{N}} \bar{c}_{ij} \psi_j$ a media. En consecuencia:

$$\min_{\bar{\mathcal{X}}} \sum_{i=1}^{n} d(\mathcal{X}_i, \bar{\mathcal{X}})^2 = min_{\bar{\mathcal{X}}} \sum_{i=1}^{n} < \mathcal{X}_i - \bar{\mathcal{X}}, \mathcal{X}_i - \bar{\mathcal{X}} >= min_{\bar{c}} \sum_{i=1}^{n} < \sum_j (c_{ij} - \bar{c}_j)\psi_j, \sum_l (c_{il} - \bar{c}_l)\psi_l >$$
$$= \min_{\bar{c}} \sum_i (\vec{c}_i - \vec{\bar{c}}) J_\psi (\vec{c}_i - \vec{\bar{c}})$$

sendo $(J_\psi)_{ij} = (< \psi_i, \psi_j >)$. Esta última expresión é unha forma cadrática con matriz definida positiva e polo tanto o mínimo é obtido con:

$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} c_{ij}$$

- **En espazos non Hilbertianos**: En espazos métricos ou espazos de Banach, non hai maneira de buscar a media ó longo do espazo así que non hai forma pechada para a media. Non obstante, podemos utilizar unha regra empírica para encontrar o elemento da mostra que minimiza un certo criterio. Por exemplo, se escollemos $\mathcal{X}_i \in S_n$ tal que:

$$\sum_{j=1}^{n} d(\mathcal{X}_j, \mathcal{X}_i)^2 \leqslant \sum_{j=1}^{n} d(\mathcal{X}_j, \mathcal{X}_l)^2$$

para $l = 1, \ldots, n$.

Esta estratexia pode ser aproveitada para calcular a mediana, outra das medidas máis coñecidas de localización. Aínda que se omitirán os pasos anteriores por considerarse un procedemento análogo, expresaremos a súa fórmula:

$$\min_{a\in\mathbb{F}} \sum_{x\in S} d(\mathcal{X}, a)$$

E en espazos $\mathcal{L}_2$, a mediana mostral defínese analogamente á anterior:

$$\min_{a\in S_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a)$$

Por último, cabe destacar que existe unha medida de localización deseñada expresamente para este tipo de datos: a profundidade. Non obstante, por ser tan importante tanto na análise exploratoria como na detección de outliers, explicarase no seguinte capítulo.

1.3.2. Medidas de dispersión

Ó igual que se realizou no caso de medidas de localización, estudaremos as medidades de dispersión máis coñecidas que son a varianza e a covarianza.

A varianza enténdese dun xeito análogo ó caso univariante, é dicir, o promedio das desviacións cadráticas con respecto a media. Esta, na análise de datos funcionais, ten a seguinte expresión:

$$Var[S_n] = \frac{1}{n}\sum_{i=1}^{n} d(\mathcal{X}_i, \bar{\mathcal{X}})^2$$

Esta definición pode ser aplicada a todo espazo métrico.

Outra medida que debemos estudar é a matriz de covarianzas. A covarianza na análise estatística clásica é o valor que indica o grao de variación conxunta de dúas variables (ou máis) aleatorias. En datos funcionais, as variables que se teñen en conta son as distintas medidas que toma cada curva, polo que se se trata dunha variable funcional que está medindo o mesmo parámetro ó longo do tempo, o razoable é que haxa moita relación entre unha medición e a seguinte ou a anterior.

Esta medida ten a seguinte expresión:

$$\Sigma = \mathbb{E}\left[(\mathcal{X}_i - \bar{\mathcal{X}})(\mathcal{X}_i - \bar{\mathcal{X}})^t\right]$$

Pero esta definición só é válida en espazos $\mathcal{L}_2$.

Por último, tamén podemos comprobar a varianza marxinal que, en datos funcionais, é a diagonal da matriz de covarianzas. Esta medida soe dar interesante información dos datos.

1.3.3. Exemplo práctico

Agora aplicaremos os coñecementos obtidos ós datos cos que estabamos a traballar nas anteriores seccións. De igual xeito, séguese utilizando o paquete **fda.usc**. Volvemos a recurrir as curvas de glicosa dos almorzos da poboación do estudo de AEGIS anteriormente tratadas.

Comezaremos coa media. Para que sirva de exemplo, usaremos a distancia en $\mathcal{L}_2$ e a distancia do supremo para calcular dita medida e representarémola xunto coa media teórica na Figura 1.9. Os comandos que se precisaron e as súas saídas son as seguintes:

```
> ### Cálculo da L2 media (media empírica)
> D2 = metric.lp(funcional) #Distancia L2 entre as curvas
> crit2 = apply(D2^2, 1, sum)
> which.min(crit2)
```

```
[1] 4
```

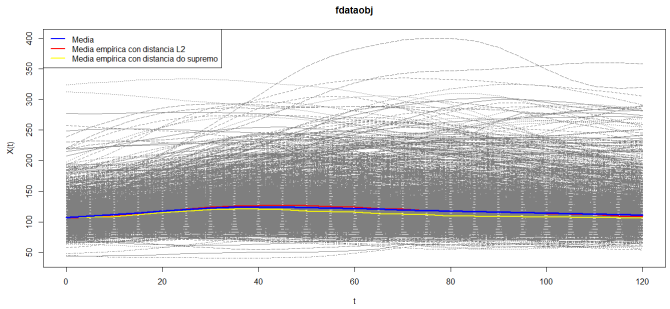

Figura 1.9: Representación gráfica dos datos de glicosa coa media teórica $\mathcal{L}_2$ (vermello) e a media mostral coa distancia do supremo (amarelo).

```
> ### Cálculo da distancia do supremo
> D0=metric.lp(funcional,lp=0)
> crit0=apply(D0,1,sum)
> which.min(crit0)

[1] 2677
```

```
> plot(funcional,col="gray50")
> lines(funcional[4,],col="red",lwd=2) #curva "media" pola regra empírica
> lines(funcional[2677,],col="yellow",lwd=2) #curva "media" pola regra do
> lines(func.mean(funcional),col="blue",lwd=3) # Media dos datos
```

En canto a varianza, como pode ser aplicado a calquera espazo métrico, usamos os seguintes comandos de R para poder calculalo:

```
> ## Varianzas
> barx<-func.mean(funcional)
> vsn<-mean(metric.lp(funcional,barx)^2)
> vsn

[1] 105930.7
```

```
 > sqrt(vsn)

 [1] 325.47
```

Obténdose, como se poderá observar, unha varianza de 105930.7 cunha desviación típica de 325.47.

Por último, móstranse os comandos necesarios para calcular a matriz de covarianzas. Ademais, estes sacan a Figura 1.10. Nela vese que hai unha gran correlación entre os datos recollidos a minutos próximos, como parece obvio.

```
> mcor=cor(funcional$data)
> image(funcional$argvals,funcional$argvals,mcor)
> contour(funcional$argvals,funcional$argvals,mcor,add=T)
```

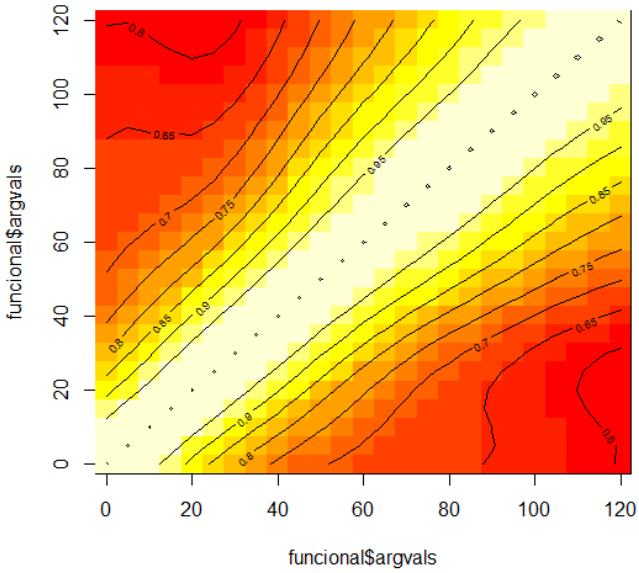

Figura 1.10: Representación gráfica da matriz de covarianzas dos datos de glicosa da poboación do estudo AEGIS.

Capítulo 2

Medidas de profundidade e busca de outliers en datos funcionais

A busca de outliers é unha parte importante para a boa realización dun estudo estatístico. No caso multivariante clásico, contamos con ferramentas como a distancia de Cook para atopalos. Non obstante, no caso de datos funcionais hai unha medida de localización estrela que se usa para atopar os datos atípicos: a profundidade.

A profundidade, ó igual que a media ou a mediana, pódese usar para construír unha medida de localización (como se usa a distribución ou densidade en datos multivariantes). Tratarase neste capítulo debido a que o seu cálculo é un tema fundamental para a busca de datos atípicos ou outliers. Logo veremos como se relacionarán estes dous conceptos e aplicaremos a un exemplo práctico sinxelo.

2.1. Medidas de profundidade

Como xa se comentou, a profundidade é unha medida de localización moi estudada. Na literatura, foron propostos varios criterios co obxectivo de cuantificar que profundo está un punto na mostra.

Normalmente, no caso univariante, a mediana sería o dato máis profundo dun conxunto de nube de puntos. Neste traballo, imos traballar coas profundidades que están recollidas no traballo de Cuevas A. et al (2007): profundidade Fraiman e Muñiz (FMD), profundidade modal (MD) e profundidade por proxección aleatorias (RPD):

- **Profundidade de Fraiman-Muniz (FMD)**: Sexa $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ iid (independentes e identicamente distribuídas) realizacións dunha variable funcional aleatoria con dominio $\mathbb{T} = [a, b]$, sexa D unha medida de profundidade en $\mathbb{R}$ e sexa $F_{n,t}$ unha distribución empírica de $X_1, \dots, X_n$. Por exemplo poderíase considerar, para todo $t_0 \in \mathbb{T}$, $z_i(t_0) := D(\mathcal{X}_i(t_0)) = 1 - |\frac{1}{2} - F_{n,t}(\mathcal{X}_i(t))|$ como a profundidade univariante do dato i en t_0.

  A profundidade de Fraiman-Muniz é:

$$FMD(\mathcal{X}_i) = \int_{\mathbb{T}} z_i(t)dt$$

- **Profundidade modal (MD)**: Sexa $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ observacións independentes e identicamente distribuídas dunha variable funcional aleatoria e sexa $K : \mathbb{R}^+ \to \mathbb{R}^+$ unha función kernel

asimétrica con parámetro ventá h. Entón a profundidade modal defínese como:

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

Debido á súa complexidade computacional, esta profundidade non é aconsellable para un conxunto de datos cun elevado número de curvas.

- **Profundidade de proxeccións aleatorias (RPD)**: Sexa $S_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ observacións iid dunha variable funcional aleatoria, $h \in \mathbb{H}$ unha realización de dirección independente do proceso $\mathbb{H}$ e $P_i^h =< h, \mathcal{X}_i >\in \mathbb{R} >$ a proxección de $\mathcal{X}_i$ na dirección de h. Entón a profundidade de proxeccións aleatorias consiste en:

$$RPD(\mathcal{X}_i, h) := D(P_i^h)$$

Na práctica, poderase escoller dúas variantes: unha na que se utiliza un número M de direccións, sendo RPD a media aritmética de todas as profundidades (RPD); ou outra na que se considera o mínimo de todas esas profundidades (variante de Tukey, RTD) Cuesta-Albertos J.A. e Nieto-Reyes A. (2008). Estas terán a seguinte expresión:

$$RPD\left(\mathcal{X}_i, \{h_l\}_{l=1}^M\right) = \frac{1}{M}\sum_{l=1}^{M} D(P_i^{h_l})$$

e

$$RTD\left(\mathcal{X}_i, \{h_l\}_{l=1}^M\right) = \min_M D(P_i^{h_l})$$

Unha das características máis agradecidas das medidas de profundidade é que todas elas poden ser adaptadas ó traballo con moitas características das funcións ó mesmo tempo, é dicir, con derivadas ou outras transformacións. Por exemplo, se supoñemos que ós nosos datos $\mathcal{X}_i$ aplicamos unha transformación K, $\{\vec{T}(\mathcal{X}_i\}_{i=1}^n := \{T^1(\mathcal{X}_i), \dots, T^K(\mathcal{X}_i)\}_{i=1}^n$; para modificar a profundidade podemos:

- Calcular as profundidades e ponderalas, ou sexa:

$$D(\vec{T}(\mathcal{X})) = \sum_{k=1}^{K} w_k D^k(T^k(\mathcal{X}))$$

sendo $w = (w_1, \dots, w_k)$ un vector de pesos e D^k unha medida de profundidade usada na transformación k.

- Modificar o procedemento para incorporar a información adicional. Isto habería que facelo en cada tipo de profundidade dun xeito distinto:

  - Para a profundidade de Fraiman-Muniz, calcularase a profundidade marxinal multivariante.

- Para a profundidade modal usaremos unha nova distancia entre os datos (se a transformación é a derivada usaremos a métrica de Sobolev[1], por exemplo)
- Para as proxeccións aleatorias considerarase unha profundidade multivariante para ser aplicada ás diferentes proxeccións.

2.2. Bandas de confianza bootstrap

Se o que se quere é medir a dispersión dos estimadores de localización, o mellor é facelo mediante a técnica bootstrap. Para isto, sexa $S_n = \{\mathcal{X}_i\}_{i=1}^n$ a mostra dispoñible e $\hat{\theta}(S_n)$ o estimador de localización de $\theta(S)$. Entón, as bandas de confianza bootstrap $(1-\alpha)$ centradas en $\hat{\theta}(S_n)$ defínense como o cuantil $q_{1-\alpha}$ das distancias $d(\hat{\theta}(S_n), \hat{\theta}(S_n*))$ obtidas mediante remostraxe.

Entón, dados os datos orixinais, as bandas[2] de confianza bootstrap son construídas do seguinte modo:

- Obtense unha remostraxe $S_n^j* = \{\mathcal{X}_i*\}_{i=1}^n$ onde $\mathcal{X}_i* = \mathcal{X}_{i*} + Z_i$ sendo Z_i un proceso independente de $\mathcal{X}_i*$ con $\mathbb{E}[Z] = 0$ e $\Sigma_Z \propto \Sigma_{\mathcal{X}}$, por exemplo $Z \sim N(0, h\Sigma_{\mathcal{X}})$
- Calculamos $\hat{\theta}(S_n^j*)$.
- Repetimos os anteriores pasos B veces para obter o cuantil $1-\alpha$ de $\left\{d(\hat{\theta}(S_n^j), \hat{\theta}(S_n^j*))\right\}_{j=1}^B$

2.3. Busca de outliers ou datos atípicos

Despois de dedicar dúas seccións deste capítulo para buscar os datos máis profundos, nesta intentarase facer ó contrario: buscar os datos con menos profundidade para atopar curvas atípicas ou outliers.

[1] O espazo Sobolev Deza M. (2014) $W^{k,p}$ é un subconxunto dun espazo $\mathcal{L}_p$ tal que f e as súas derivadas ata a orde k ten unha finita $\mathcal{L}_p$ norma. Formalmente, dado un subconxunto $G \subset \mathbb{R}^n$, definimos:

$$W^{k,p} = W^{k,p}(G) = \{f \in \mathcal{L}_p(G) : f^{(i)} \in \mathcal{L}_p(G), 1 \leqslant i \leqslant k\}$$

onde $f^{(i)} = \delta_{x_1}^{\alpha_1} ... \delta_{x_n}^{\alpha_n} f$ e $\alpha_1 + \cdots + \alpha_n = i$ e as derivadas están consideradas no sentido débil. Entón a norma Sobolev en $W^{k,p}$ ven definida por:

$$||f||_{k,p} = \sum_{i=0}^{k} ||f^{(i)}||_p$$

Polo tanto, a métrica Sobolev é a norma métrica $||f-g||_{k,p}$ en $W^{k,p}$ que fai que $W^{k,p}$ sexa un espazo de Banach. Este espazo é Hilbertiano co produto interior:

$$< f, g >_k = \sum_{i=1}^{k} < f^{(i)}, g^{(i)} >_{\mathcal{L}_2} = \sum_{i=1}^{k} \int_G f^{(i)} \bar{g}^{(i)} \mu(d\omega)$$

[2] Aínda que se chama banda, é a bola centrada no estimador de radio $q_{1-\alpha}$ do espazo funcional considerado. Sería unha banda só se estamos considerando a distancia do máximo. Noutro caso, a bola depende da métrica e sería case imposible representala graficamente.

Neste tema, non hai unha definición xeral aceptada de outlier en datos funcionais. Polo tanto, neste traballo definirase outlier como un dato xerado por un proceso distinto do resto da mostra coas seguintes características:

- O número de datos atípicos nunha mostra non é coñecido, pero é probablemente baixo.
- Os outliers, de habelos, terán unha profundidade significativamente baixa.

En Febrero M. et al (2007a), explícase o procedemento para obter os outliers. Os pasos son os seguintes:

- Obtemos as profundidades do conxunto de datos: $\{D(\mathcal{X}_i)\}_{i=1}^{N}$
- Sexa $\mathcal{X}_{i_1}, \ldots, \mathcal{X}_{i_k}$ a k-ésima curva tal que $D(\mathcal{X}_{i_j}) \leqslant C$, para un C dado. Entón, marcamos $\mathcal{X}_{i_1}, \ldots, \mathcal{X}_{i_k}$ como o conxunto de curvas de outliers na mostra.
- Elixir C de tal maneira que en ausencia dos valores atípicos, a porcentaxe de observacións correctas mal etiquetadas como valores atípicos é aproximadamente igual a unha pequena proporción (digamos entre o 1 e o 2 %).

O punto de corte C débese establecer a través de técnicas de remostraxe xa que non hai distribución teórica dos datos. Ademais, a determinación de C non debe estar afectada pola presenza de atípicos.

En xeral, hai dous tipos de procedementos para buscar outliers: as recortadas e as ponderadas. A diferencia entre elas é que unha realiza unha mostra bootstrap despois de rexeitar unha certa porcentaxe de datos menos profundos (a recortada) mentres que a outra da unha certa probabilidade a cada curva proporcional a súa profundidade (ponderada).

- **Detección de valores atípicos baseado en un recorte:**
  1. Obter a profundidade funcional $\{D_n(\mathcal{X}_i)\}$ para unha profundidade funcional.
  2. Obter B mostras bootstrap X_i^b* das curvas do conxunto de datos obtidos despois de borrar $\alpha\,\%$ curvas menos profundas, para cada $i = 1, \ldots, n$ e $b = 1, \ldots, B$.
  3. Obter mostras bootstrap $Y^b = X_i^b + Z_i^b$, onde Z_i^b é tal que $Z_i^b(t_l)$ para $l = 1, \ldots, m$ é normalmente distribuído con media 0 e matriz de covarianzas $\gamma\Sigma_x$, onde Σ_x é a matriz de covarianzas de $X(t_l)$ e γ é o parámetro de suavizado de bootstrap.
  4. Para cada conxunto bootstrap $b = 1, \ldots, B$, obtemos C^b como o percentil empírico $c\,\%$ da distribución das profundidades $D(Y_i^b)$.
  5. Coller C como a mediana dos valores de C^b con $b = 1, \ldots, B$, onde a estimación do punto de corte C está baseada no remostraxe das curvas orixinais con probabilidade proporcional á súa profundidade.
- **Detección de outliers baseado en ponderación**
  1. Obter a profundidade funcional $\{D_n(\mathcal{X}_i)\}$ para unha profundidade funcional.
  2. Obter B mostras bootstrap X_i^b de curvas nas cales cada curva orixinal é aleatoria con probabilidade proporcional á súa profundidade.
  3. Obter mostras bootstrap $Y^b = X_i^b + Z_i^b$, onde Z_i^b é tal que $Z_i^b(t_l)$ para $l = 1, \ldots, m$ é normalmente distribuído con media 0 e matriz de covarianzas $\gamma\Sigma_x$, onde Σ_x é a matriz de covarianzas de $X(t_l)$ e γ é o parámetro de suavizado de bootstrap.
  4. Para cada conxunto bootstrap $b = 1, \ldots, B$, obtemos C^b como o percentil empírico $c\,\%$ da distribución das profundidades $D(Y_i^b)$.

5. Coller C como a mediana dos valores de C^b con $b = 1, \dots, B$, onde a estimación do punto de corte C está baseada no remostraxe das curvas orixinais con probabilidade proporcional á súa profundidade.

Por último, describirase outra medida de profundidade alternativa a todas estas na que se emprega as compoñentes descritas no anterior capítulo: o "High density region"e o "Half-space". A utilización destes dous métodos da, por defecto, un HDR plot e un bagplot respectivamente.

Un bagplot (Rousseauw P.J. et al 2012) é unha extensión bivariante do concepto de boxplot. Para comprendelo, necesítase unha xeneralización dos rangos dos datos univariantes a multivariantes que foi estudado por (Rousseauw P.J. et al 2012) que introduciu a profundidade "half-space". A localización da profundidade do semiespazo ("half-space"), $ldepth(\theta, Z)$, para un punto $\theta \in \mathbb{R}^2$ relativo a unha nube de puntos bivariantes $Z = \{z_1, \dots, z_n\}$ é o menor número de z_i contidos en calquera semiespazo pechado cuxa fronteira pasa por θ. Usando este concepto, proponse unha versión bivariante do boxplot onde destacan: unha bolsa (bag) que contén ó 50 %, unha valla que separa os outliers dos inliers e un circuíto indicando os datos fóra da bolsa pero dentro da valla. Como se dixo, pódese consultar o procedemento a seguir para calcular cada unha das compoñentes citadas en (Rousseauw P.J. et al 2012). Ademais, podemos aplicar rexións de confianza para a mediana no bagplot. Para iso, o artigo citado explica como conseguir o maior valor k para o cal

$$\mathbb{P}(ldepth(\hat{\theta}, X_n) \geqslant k) \geqslant 0.95$$

onde X_n provén dunha distribución con mediana poboacional $\hat{\theta}$. Entón constrúese a correspondente rexión D_k:

$$D_k = \{\theta \in \mathbb{R}^2 : ldepth(\theta, Z) \geqslant k\}$$

que lle chamaremos mancha.

Un HDR plot (gráfico de rexións de alta densidade) foi proposto por Hydman (1996) (Rousseauw P.J. et al 2012). Para construílo, primeiro estímase a densidade dos datos cun método kernel, por exemplo e logo o 50 % do HDR calcúlase pola densidade que rodea o 50 % da masa.

Aínda que os métodos se parecen, isto non é certo. Algunhas diferencias son:

- O HDR non necesita ser convexo nin conexo.
- O HDR non é unha xeneralización do diagrama de caixas ou boxplot, xa que a súa versión univariante contería moitas caixas.
- O HDR está baseado na idea de densidade mentres que o bagplot nos rangos.
- O HDR depende do estimador da densidade e do parámetro ventá mentres que o bagplot é invariante ante transformacións.
- Interprétanse de xeito distinto para buscar outliers: un atípico no bagplot identifícase como un punto lonxe da masa dos datos, mentres que no HDR é un punto nunha zona baleira.

2.4. Exemplo práctico

Como xa é costume no traballo, aplicaremos o exposto no capítulo a un exemplo sinxelo para poder comprendelo todo mellor. Volvemos a rescatar os datos do estudo poboacional de AEGIS, de curvas de glicosa para os almorzos. Comezarase estudando os distintos tipos de profundidade. Isto implementarase en R coas seguintes liñas de comando:

```
> prof1<-depth.FM(funcional,draw=T)          #Fraiman-Muniz
> prof2<-depth.mode(funcional,draw=T)        #Modal depth
> prof3<-depth.RP(funcional,draw=T)          #Random Projection Depth
> prof4<-depth.RT(funcional,draw=T)          #Tuckey ramdom projection Depth
```

que nos devolve a Figura 2.1. Nela podemos comprobar cal é a profundidade de cada curva cunha simple ollada á súa tonalidade, pois canto máis clara é, menos valor vai ter a súa profundidade (se necesitásemos o valor exacto de cada unha delas, a función devolve o seu valor, entre outras, na variable *dep*).

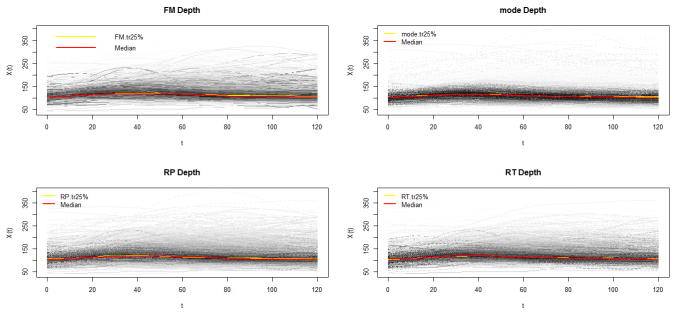

Figura 2.1: Representación gráfica das curvas de glicosa para os almorzos. A representación da profundidade ven dada pola tonalidade de cor desta: canto máis escura máis profunda. Ademais aparece a mediana (vermello) e a curva máis profunda segundo a profundidade.

En canto a profundidade modal, pode ser interesante saber cal é a reacción desta ó aumentar o parámetro ventá (pois, como se pode recordar, esta profundidade depende dun parámetro ventá). En concreto, imos doblar o valor do parámetro ventá que escolle por defecto R (124.7), é dicir, consideraremos un parámetro de 249.4. Na Figura 2.2, podemos ver que aumenta a profundidade das curvas máis "centrais".

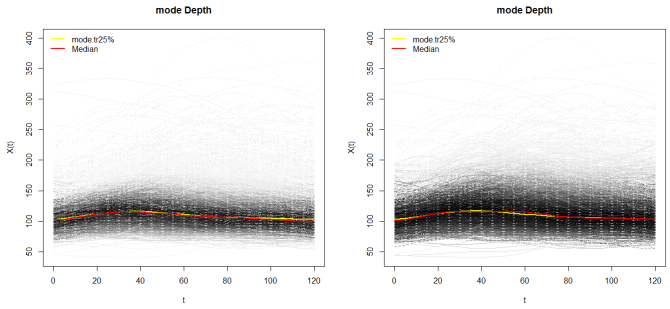

Figura 2.2: Representación gráfica do dato funcional con tonalidade distinta segundo a súa profundidade. Á esquerda co parámetro ventá por defecto e a dereito co doble da primeira.

Ademais, coa seguinte liña de comando:

```
> cur<-c(prof1$lmed,prof2$lmed,prof3$lmed,prof4$lmed)
> plot(funcional,type="n",main="")
>lines(funcional[cur],lwd=c(4,1,2,1),lty=c(1,2,1,1),col=c("blue","green","red","yellow"))
```

pódese comparar o dato máis profundo segundo as profundidades que estamos a considerar. O resultado pódese ver na Figura 2.3. Á vista de dita gráfica, parece que os datos máis profundos están considerablemente cerca.

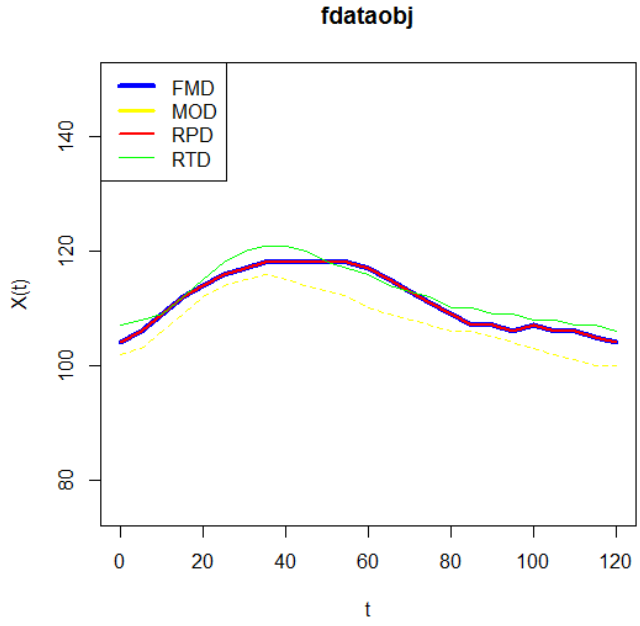

Figura 2.3: Representación gráfica do dato máis profundo das curvas segundo a profundidade considerada.

Ademais, mediuse a dispersión dos estimadores de localización (media e profundidades FM, RP e modal) cunha remostraxe de 100 mostras bootstrap con bandas de confianza ó 95 % co comando de R:

```
> out.boot1=fdata.bootstrap(funcional,statistic=func.mean,nb=100,draw=TRUE)
> out.boot2=fdata.bootstrap(funcional,statistic=func.med.FM,nb=100,draw=TRUE)
> out.boot2=fdata.bootstrap(funcional,statistic=func.med.mode,nb=100,draw=TRUE)
> out.boot2=fdata.bootstrap(funcional,statistic=func.med.RP,nb=100,draw=TRUE)
```

O resultados desta liña de código móstranse na Figura 2.4.

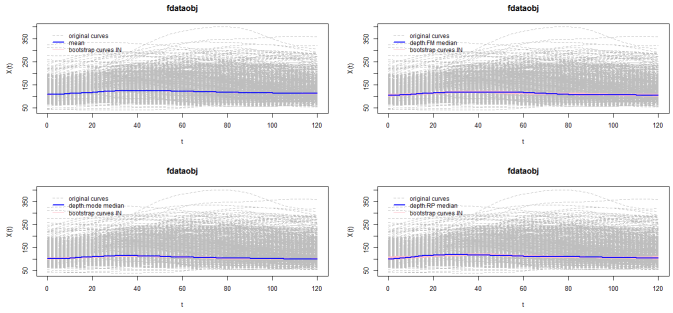

Figura 2.4: Representación das bandas de confianza bootstrap (mediante bootstrap suavizado) para $\alpha = 0.05$ para a media e profundidades FM, RP e modal para os datos de espectro espectrométrico.

De seguindo, calcúlanse os outliers polos métodos explicados. Primeiro buscáronse atípicos coa profundidade de Fraiman-Muniz, tanto polo procedemento recortado como ponderado. Por outra banda, calcúlase os outliers con respecto á profundidade por proxeccións aleatorias. Logo representámolo todo na Figura 2.5, onde á dereita están os outliers coa profundidade FM e á esquerda coa RP. Se se prefire identificar as curvas, mirar o punto de corte do método ou a profundidade de cada dato outlier deberase profundizar na función *outlier.depth*. Por exemplo, buscaremos ditos datos para, por exemplo, os outliers atopados mediante o método ponderado e utilizando a profundidade FM. Para atopar que número ocupa o dato atípico procederemos do seguinte xeito:

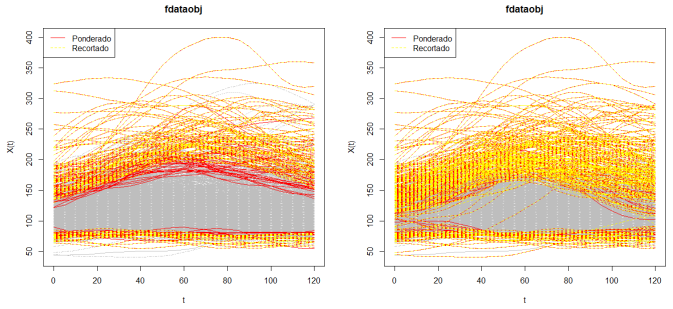

Figura 2.5: Representación dos outliers (en cores) co método recortado e ponderado utilizando a profundidade de Fraiman-Muniz á esquerda e a profundidade por proxeccións aleatorias á dereita.

```
> outFM1$outliers

  [1] "7"    "9"    "64"   "81"   "96"   "105"  "131"  "195"  "247"  "502"
 [11] "519"  "521"  "562"  "585"  "616"  "620"  "660"  "673"  "687"  "820"
 [21] "1186" "1187" "1188" "1290" "1298" "1301" "1350" "1370" "1372" "1394"
```

```
 [31] "1408" "1410" "1411" "1413" "1627" "1631" "1648" "1650" "1711" "1712"
 [41] "1788" "1808" "1847" "1850" "1851" "1889" "1918" "1949" "1979" "1980"
 [51] "1981" "2038" "2039" "2041" "2100" "2182" "2281" "2319" "2369" "2391"
 [61] "2394" "2426" "2502" "2665" "2667" "2668" "10"   "62"   "63"   "83"
 [71] "98"   "522"  "684"  "685"  "686"  "688"  "835"  "1100" "1101" "1103"
 [81] "1185" "1189" "1270" "1341" "1343" "1409" "1426" "1427" "1608" "1769"
 [91] "1830" "1848" "1867" "1870" "1978" "2037" "2180" "2199" "2200" "2201"
[101] "2318" "2321" "2501" "2666" "2"    "61"   "85"   "641"  "642"  "1099"
[111] "1102" "1150" "1267" "1297" "1299" "1342" "1371" "1594" "1864" "2076"
[121] "2320" "2428" "2669" "84"   "564"  "1300" "1373" "1504" "1606" "1633"
[131] "99"   "1609" "2306" "1966" "1863" "2472" "1269" "2302"
```

Mentres que se se quere obter a profundidade de cada dato e o punto C de corte, introduciremos a seguinte liña de comando:

```
> outFM1$quantile
```

```
0.04337805
```

```
> outFM1$dep.out
```

```
  [1] 0.038189781 0.015007299 0.042306569 0.040963504 0.032087591 0.039328467
  [7] 0.037197080 0.035036496 0.039708029 0.037197080 0.020817518 0.042540146
 [13] 0.042919708 0.028875912 0.029372263 0.024204380 0.018627737 0.004583942
 [19] 0.037722628 0.043357664 0.035386861 0.028467153 0.038540146 0.011795620
 [25] 0.024875912 0.027678832 0.004788321 0.040000000 0.040700730 0.011854015
 [31] 0.009664234 0.003649635 0.014919708 0.028525547 0.042540146 0.023416058
 [37] 0.013284672 0.020613139 0.022481752 0.019912409 0.039416058 0.021372263
 [43] 0.003124088 0.009897810 0.029664234 0.012992701 0.023416058 0.018540146
 [49] 0.021605839 0.005985401 0.018686131 0.027124088 0.028175182 0.009751825
 [55] 0.006861314 0.031941606 0.022948905 0.030627737 0.043270073 0.029313869
 [61] 0.040817518 0.009197080 0.022540146 0.027007299 0.017576642 0.016000000
 [67] 0.042303665 0.034644727 0.036290202 0.034854151 0.021301421 0.031772625
 [73] 0.028481675 0.025071055 0.023216156 0.039820494 0.041106956 0.033298429
 [79] 0.042872102 0.037516829 0.031473448 0.026925954 0.036798803 0.033178758
 [85] 0.042393418 0.023784592 0.043261032 0.026387435 0.037038145 0.033896784
 [91] 0.040837696 0.030665669 0.032191473 0.035512341 0.042543007 0.038534031
 [97] 0.034345550 0.024652206 0.038175019 0.028840688 0.030426328 0.025759162
[103] 0.035721765 0.026716530 0.028588771 0.030561457 0.039848255 0.032655539
[109] 0.034871017 0.036024279 0.036965099 0.042063733 0.033141123 0.034658574
[115] 0.035235205 0.035174507 0.038634294 0.043308042 0.042852807 0.033990895
[121] 0.037845220 0.040576631 0.029833080 0.037172335 0.040229270 0.040718380
[127] 0.042766527 0.040106993 0.043286206 0.042766527 0.042727969 0.043357608
[133] 0.043312883 0.043268124 0.043100537 0.043346124 0.043072197 0.042519201
```

Por último, calculamos os outliers a través da profundidade HDR e HS. Isto aínda non está implementado no paquete **fda.usc**, polo que deberase instalar o paquete **rainbow** e teremos que converter os datos funcionais, que os tiñamos na clase $fdata$, á clase fs coa función fds. Logo destes pasos, obtense as Figuras 2.6 e 2.7 onde se poden ver os outliers a través da profundidade HDR e "Half-space", respectivamente.

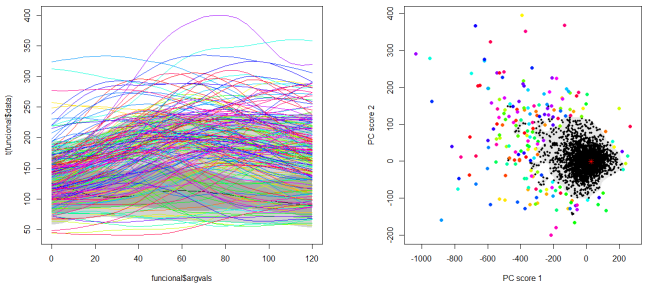

Figura 2.6: Representación dos outliers (en cores) co método HDR, tanto coa gráfica do tipo funcional como coa gráfica do tipo bivariante.

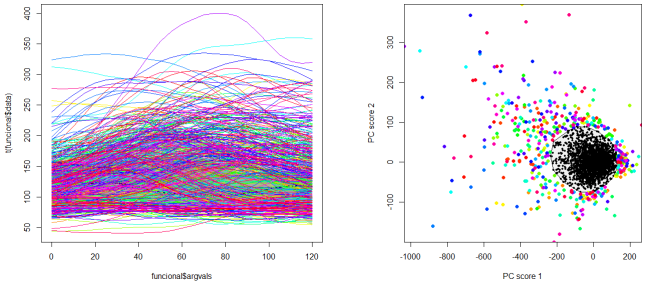

Figura 2.7: Representación dos outliers (en cores) co método HS, tanto coa gráfica do tipo funcional como coa gráfica do tipo bivariante.

Nótese que como se comentou na parte teórica, o bagplot e o HDR plot son parecidos pero non iguais xa que, por exemplo, neste caso o HDR plot non é convexo.

Capítulo 3

Clasificación en datos funcionais

Os métodos de clasificación baséase na idea de que as medidas dos individuos baixo estudo forman grupos ou nubes de datos máis ou menos ben separados no espazo de características e, neses caso, é posible construír unha función que permita separar os datos dun grupo de datos dos demais grupos. Estes datos poden pertencer a un ou varios grupos máis ou menos amplos. Estas técnicas son bastante útiles para, por exemplo, o recoñecemento de patróns.

Dentro da clasificación débese distinguir entre clasificación supervisada e non supervisada. Na clasificación supervisada xa se ten unha mostra distribuída en grupos e o obxectivo é establecer unha regra para estimar a probabilidade a posteriori dunha nova observación de pertencer a un grupo ou a outro. A regra óptima consiste en asignar a nova observación ó grupo que maximice a súa probabilidade a posteriori. Polo contrario, na clasificación non supervisada non contamos con coñecemento de cales son os grupos a priori, polo que o seu obxectivo é a distinción dos grupos polos que están formados.

3.1. Clasificación non supervisada

Na clasificación non supervisada temos unha mostra $\{\mathcal{X}_i\}_{i=1}^n$ cun conxunto de características e o obxectivo é analizar a que grupos pertencen cada dato apoiándonos nas características para separar ditas clases.

Para unha escolla correcta do algoritmo a empregar, hai que fixar antes o obxectivo ó que se quere chegar:

- Se o obxectivo é construír k grupos onde k é un número dado deles e se quere prover da mellor partición para estes grupos, debemos optar polo método de partición.
- Non obstante, se non coñecemos o número de grupos, deberase empregar $k \in \{1, \dots, n\}$ usando certas regras para separar ou agregar os datos. Estes métodos son os métodos xerárquicos.

Neste traballo o interesante será cumprir o primeiro obxectivo, é dicir, separar os datos en k grupos. Un exemplo para cumprir este obxectivo é o algoritmo de k-medias. Este é un método de agrupamento que ten como obxectivo a partición dun conxunto de n observacións en k grupos de maneira que minimice a suma de cadrados dentro dos grupos sobre todo o conxunto de variables.

O método intenta atopar k grupos ó redor dos centros iniciais $\{m_1^{(1)}, \dots, m_k^{(1)}\}$. Entón, o algoritmo alterna os dous seguintes pasos:

1. Etapa de asignación: cada observación asígnase a un grupo con quen teña a media máis cerca:

$$S_i^{(t)} = \left\{ \mathcal{X}_p : ||\mathcal{X}_p - m_i^{(t)}|| \leqslant ||\mathcal{X}_p - m_j^{(t)}||^2, \forall 1 \leqslant j \leqslant k \right\}$$

onde $\mathcal{X}_p$ é asignado a exactamente un grupo.

2. Actualización das medias: calcula as novas medias para que van ser os novos centroides das observacións nos novos grupos con $m_i^{(t+1)}$:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathcal{X}_j \in S_i^{(t)}} \mathcal{X}_j$$

Nótese que este algoritmo pode producir problemas pola utilización da media ou pola presenza de outliers e para solucionalo proponse empregar outro centroide en vez da media e usar o algoritmo no $(1-\alpha)\,\%$ dos datos máis profundos, respectivamente.

3.2. Clasificación supervisada

Na clasificación supervisada, temos unha mostra totalmente clasificada por grupos que será a chamada mostra de adestramento, é dicir, terase:

$$\{\mathcal{X}_i, G_i\}_{i=1}^n \in \mathbb{F} \times \mathbb{G} = \{1, \ldots, G\}$$

onde G é a variable que indica a pertenencia a un grupo determinado.

O obxectivo é estimar a probabilidade a posteriori para unha nova observación X de pertencer a cada grupo, ou sexa:

$$p_g(X) = \mathbb{P}(G = g / \mathcal{X} = X) = \mathbb{E}[1_{\{G=g\}} \backslash \mathcal{X} = X]$$

A regra de clasificación óptima é asignar á nova observación ó grupo que maximiza a probabilidade a posteriori:

$$\hat{G}_x = arg \max_{g \in \mathbb{B}} \hat{p}_g(X)$$

Agora ben, hai varias maneiras de calcular dita probabilidade. Por unha banda, a probabilidade pode ser escrita como unha esperanza, polo que a probabilidade pode ser estimada desta maneira:

$$\hat{p}_{g,h}(X) = \frac{\sum_{i=1}^n 1_{\{G_i=g\}} K(h^{-1}) d(X, \mathcal{X}_i)}{K(h^{-1}) d(X, \mathcal{X}_i)}$$

onde K é unha función kernel asimétrica e h o parámetro ventá[1]. Este estimador cumpre:

- $0 \leqslant \hat{p}_{g,h}(X) \leqslant 1$
- $\sum_{g \in \mathbb{G}} \hat{p}_{g,h}(X) = 1$

[1]Normalmente escollido minimizando unha función de perda, por exemplo a validación cruzada, $h_{opt} = arg\,\text{mín}\, LCV(h)$ con

$$LCV(h) = \sum_{g=1}^{G} (1_{\{G_i=g\}} - \hat{p}_{g,h}^{(-i)}(\mathcal{X}_i))$$

Por outra banda, outra posibilidade é considerar o problema de clasificación como un problema de regresión loxística:

$$\pi_{i,g} = \mathbb{P}(G = g/\mathcal{X} = \mathcal{X}_i) = \frac{exp\{\alpha_g + < \mathcal{X}_i, \beta_g >\}}{1 + exp\{\alpha_g + < \mathcal{X}_i, \beta_g >\}}$$

ou equivalentemente

$$l_{i,g} = \alpha_g + < \mathcal{X}_i, \beta_g > , \; l_{i,g} = \ln[\pi_{i,g}/(1 - \pi_{i,g})]$$

Seguindo esta estratexia, poderase realizar clasificación cos modelos FGLM, GSAM e GKAM. (Febrero M. e Gonzalez Manteiga 2013)

No caso de que se conte unicamente con dous grupos, existe outro método importante de clasificación: o DD-plot. Este método está definido de forma que compara dúas distribucións ou grupos como o gráfico dimensional $(D_1(x), D_2(x))$, onde $D_i(x)$ é a profundidade do punto x respecto ós datos do grupo i-ésimo. É dicir, consiste na representación gráfica da función:

$$\begin{aligned} \mathcal{X} &\longrightarrow \mathbb{R}^2 \\ x &\rightsquigarrow (D_1(x), D_2(x)) \end{aligned}$$

Se os dous grupos son o mesmo, os puntos do gráfico estarán agrupados sobre a diagonal. Non obstante, se os dous grupos están claramente separados, o DD-plot terá forma de L.

A pesar de que a clasificación con este método ten importantes vantaxes, como ser capaz de identificar patróns complexos, tamén ten unha serie de desvantaxes:

- Situacións nas que aparecen illas non se poden resolver.
- A complexidade computacional aumenta a un paso de N^k con N o tamaño da mostra e k os graos do polinomio que se está considerando.
- Non se pode usar con máis de dous grupos.

Por todo isto, preséntase o clasificador DD^G. Supoñamos que temos g grupos (clases ou distribucións) para ser separados usando a profundidade dos datos. O clasificador DD^G empeza seleccionando a profundidade D e calculando o seguinte:

$$\begin{aligned} \mathcal{X} &\longrightarrow \mathbb{R}^g \\ x &\rightsquigarrow d = (D_1(x), D_2(x), \dots, D_g(x)) \end{aligned}$$

onde D_k é a profundidade de x con respecto a $k = 1, \dots, g$. Entón, o clasificador DD^G comprime a información de $\{y_i, x_i\}$ nun espazo real de dimensión $(g+1)$ coa forma $\{y_i, D_1(x_i), \dots, D_g(x_i)\}$. Neste espazo, $\mathbb{R}^g$, pódense ter algunhas técnicas de clasificación como o Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Generalized Linear Models (GLM) e Generalized Additive Models (GAM), entre outros.

Por último, cabe destacar as propiedades do clasificador DD^G:

- Moitos métodos dispoñibles para a clasificación.
- Posible redución na dimensión do problema de clasificación, vantaxe importante en problemas de alta dimensión.
- O uso de métodos de clasificación clásicos no DD-plot pode proporcionar información útil acerca do que está pasando (cales son profundidades influentes ou as probabilidades de pertenza a un certo grupo determinado).

- Non importa que complexo é o espazo a analizar, só que función de profundidade pode ser definido.

3.3. Exemplo práctico

Logo de repasar teoricamente as técnicas de clasificación, volvemos a empregar os datos do estudo de AEGIS. Primeiro comezaremos coa clasificación non supervisada. Este tipo de análise, ó igual que a maioría do traballo, realízase cos comandos do paquete **fda.usc**; en concreto co comando *kmeans*. Ademais, utilizamos a variable *dm* que nos indica se un individuo do estudo foi diagnosticado como diabético. Polo tanto, despois de dividir a mostra en dous grupos con dito método comprobaremos se a separación que fai é con respecto a individuos diabéticos e non diabéticos. En R, todo isto conséguese do seguinte xeito:

```
> clasf1 <- kmeans.fd(funcional,ncl=2,dfunc=func.mean,draw=TRUE)
> a <- lista$x$dm
> for(i in 1:length(a)){if(a[i]==0){a[i]=1} else {a[i]=2}}
> tabla <- table(clasf1$cluster,factor(a))
> tabla
     1    2
1   73    1
2  253 2413
> prop.table(tabla,1)
            1           2
1 0.98648649 0.01351351
2 0.09489872 0.90510128
```

No caso de que os grupos obtidos seguisen este patrón, o algoritmo de $k-medias$ para $k=2$ clasificaría ben o 91 % dos datos para o grupo 2 e o 99 % para o grupo 1. Podemos ver dita clasificación na Figura 3.1.

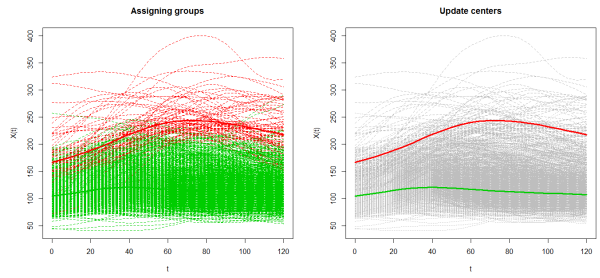

Figura 3.1: Clasificación obtida considerando o algoritmo de k-medias e co tipo de medida a media funcional. Á dereita están en cores os centroides de cada un dos grupos e á esquerda os individuos dividido en grupos.

O seguinte paso será intentar a clasificación supervisada. Para isto, empregaranse todos os individuos considerados anteriormente menos 6, que se extraerán da mostra para poder así realizar unha calibración da regra, é dicir, comprobarase se os que son diabéticos son considerados como tales e viceversa. Polo tanto, escribiremos o seguinte código de R:

```
>n<-dim(espectro)[1]
>clasif<-classif.np(dmregra,funcionalregra,h=seq(70,100,length=7))
>predict(clasif,funcionaltest,type="probs")
```

co que se obteñen os resultados presentados no Cadro 3.1.

Obs.	Prob. de grupo 0	Prob. de grupo 1	Clasificado como	Clasificación real
2665	0.0000019	0.9999809	1	1
2666	0.0018269	0.9981730	1	1
2667	0.0000025	0.9999751	1	1
2680	0.9130805	0.0869195	0	0
2681	0.9854161	0.0145839	0	0
2682	0.9555297	0.0444703	0	0

Cadro 3.1: Estimación dunha regra de clasificación e aplicación nos 6 datos extraídos para a calibración.

Como podemos observar, parece que as observacións do grupo 0, é dicir, dos non diabéticos son clasificadas correctamente aínda que con unha probabilidade lixeiramente menor que para os clasificados como diabéticos ou grupo 1.

Outra opción que se barallou foi a do DD-plot. Na Figura 3.2 pódese ver 4 DD-plot, usando unha regresión loxística dos modelos glm (dereita) ou usando k veciños máis próximos (esquerda). Os gráficos de arriba correspóndense coa profundidade de Fraiman-Muniz, mentres que os de abaixo coa profundidade de proxeccións aleatorias.

```
>ctrl=list(fine=51,draw=TRUE,col=c("red","blue"))
>res.DD=classif.DD(dmregra,funcionalregra,depth="FM",classif="glm")
>res.DD2=classif.DD(dmregra,funcionalregra,depth="FM",classif="knn",
                    par.classif=list(knn=5))
>res=classif.DD(dmregra,funcionalregra,depth="RP",control=ctrl)
>res.DD3=classif.DD(dmregra,funcionalregra,depth="RP",classif="knn",
                    par.classif=list(knn=5))
```

As proporcións da mala clasificación, resultado destes modelos, veñen explicadas no Cadro 3.2.

	glm	k veciños máis próximos
FMD	0.06	0.08
RPD	0.09	0.06

Cadro 3.2: Proporción de mala clasificación do DD-plot segundo a profundidade usada (Fraiman-Muniz ou proxeccións aleatorias) e segundo o método usado (regresión loxística dos modelos glm ou k veciños máis próximos).

Tanto do Cadro 3.2 coma a Figura 3.2 se extrae que o DD-plot ten unha razoable utilidade á hora de clasificar pois a proporción de mala clasificación é boa. Aínda así, utilizar a clasificación con k veciños máis próximos coa profundidade RPD ou o modelo loxístico coa profundidade de FM parece máis razoable.

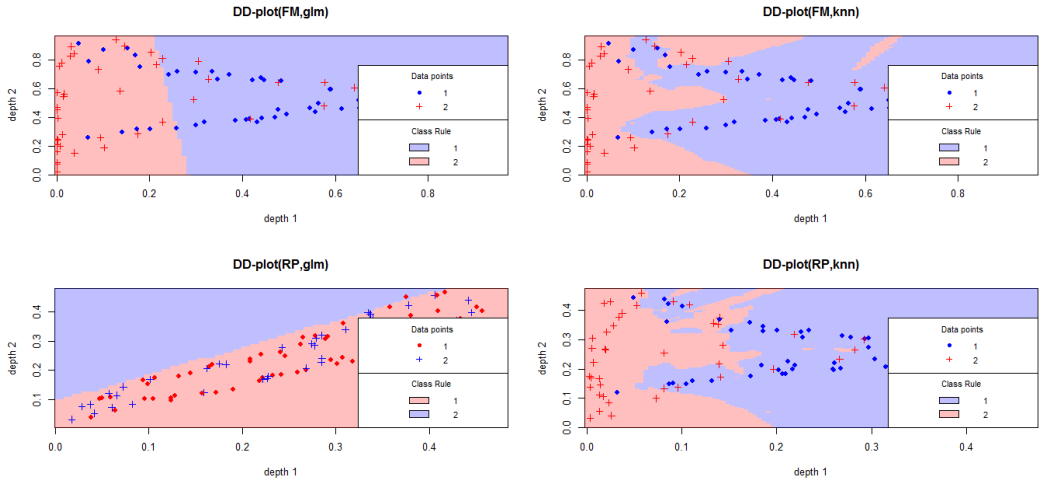

Figura 3.2: DD-plot usando a regresión loxística dos modelos glm (á dereita) ou usando k veciños máis próximos (á esquerda). Os gráficos de arriba correspóndense coa profundidade de Fraiman-Muniz, mentres que os de abaixo coa profundidade de proxeccións aleatorias. Aplicación ós datos de *espectro* clasificando nos grupos 1 e 2.

Por último, atendendo ós resultados anteriores, usarase o DD-plot non paramétrico con k veciños máis próximos e coa profundidade RPD para establecer unha regra de clasificación e así poder obter unha clasificación dos 6 individuos extraidos. O gráfico resultante pódese ver na Figura 3.3 mentres que no Cadro 3.3 se pode ver a clasificación real e a feita polo algoritmo.

	Reclasificados como grupo 1	Reclasificados como grupo 2
Grupo 1	3	0
Grupo 2	1	2

Cadro 3.3: Clasificación real fronte a clasificación realizada mediante o DD-plot coa profundidade RPD e kernel non paramétrico.

No Cadro 3.3 pódese observar que a clasificación que realiza o algoritmo é a correcta. Isto é debido á proporción de mala clasificación. Na Figura 3.3 vese que o DD-plot é moi efectivo á hora de clasificar.

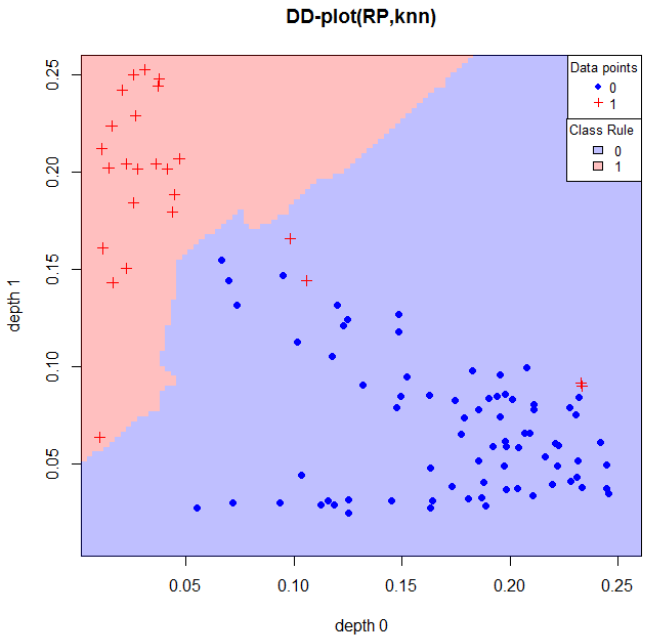

Figura 3.3: DD-plot usando kernel non paramétrico e a profundidade de proxeccións aleatorias.

Capítulo 4

ANOVA funcional (FANOVA)

No capítulo anterior tratouse de clasificar datos segundo as súas características. Pero agora intentarase propoñer un test para decidir se existen ou non diferencias no proceso (que será o dato funcional) con respecto a unha ou varias variables factores que poden afectalo.

Neste capítulo propoñerase un procedemento flexible e xenérico para resolver deseños complexos ANOVA con datos funcionais baseados en proxeccións aleatorias. O procedemento é sinxelo e fácil de aplicar co paquete de R **fda.usc**.

4.1. ANOVA dun factor

Para realizar un test deste tipo seguimos o artigo Cuevas A. et al (2007). Comezarase considerando a posibilidade de usar unha versión análoga ó F-test usando o estatístico:

$$F_n = \frac{\sum_{i=1}^{k} n_i ||X_{i\cdot} - \bar{X}_{\cdot\cdot}||^2/(k-1)}{\sum_{ij} ||X_{ij} - \bar{X}_{i\cdot}||^2/(n-k)}$$

onde

$\bar{X}_{i\cdot} = \bar{X}_{i\cdot}(t) = \sum_{j=1}^{n_i} \frac{X_{ij}(t)}{n_i}$ $\qquad$ $\bar{X}_{\cdot\cdot} = \bar{X}_{\cdot\cdot}(t) = \sum_{j=1}^{n_i} \frac{n_i \bar{X}_{ij}(t)}{n}$ $\qquad$ $n = \sum_{i=1}^{k} n_i$

e $\|.\|$ é a norma usual de $\mathcal{L}_2$.

De igual xeito que se podía facer no clásico ANOVA, podemos interpretar o numerador como a medición da variabilidade externa entre as diferentes mostras e o denominador como a variabilidade interna entre as mostras. Por suposto, rexeitaríamos H_0, a un nivel α, cando $F_n > F_{n,\alpha}$, onde $F_{n,\alpha}$ é tal que $\mathbb{P}_{H_0}\{F_n > F_{n,\alpha}\} = \alpha$. Pero non podemos seguir utilizando o caso clásico xa que non podemos saber a distribución de F_n baixo a hipótese nula. Non obstante, se estamos dispostos a empregar un test asintótico, renunciando así a esixir niveis de significación, a estrutura do clásico F estatístico suxire unha alternativa que pode adaptarse facilmente á análise funcional. En Cuevas A. et al (2004), chégase á conclusión de que se pode reformular dito estatístico para que só sexa necesario calcular a distribución do numerador multiplicado por σ^{-2} e logo reemplazar σ^2 por un estimador desta distribución. Polo tanto, podemos pensar en usar o test baseado en:

$$T_n = \sum_{i=1}^{k} n_i ||\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}||^2$$

ou equivalentemente

$$V_n = \sum_{i<j} n_i ||\bar{X}_{i.} - \bar{X}_{..}||^2$$

Para desfacernos da suposición de homocedasticidade, normalmente presente nos test ANOVA clásicos, recurrimos o seguinte teorema:

Teorema 4.1.1. *(Cuevas A. et al (2004)) Asúmese que $n_i, n \to \infty$ de maneira que $n_i/n \to p_i > 0$ para $i = 1, \dots, k$. Tamén supoñemos que temos observacións $X_{ij}(t)$, con $j = 1, \dots, n_i$, correspondendo a k mostras independentes de tamaños n_i de k procesos en $\mathcal{L}_2$ con media 0 e covarianzas $K_i(s,t) = Cov(X_i(s), X_i(t))$. Entón, a distribución asintótica de V_n baixo H_0 coincide coa do estatístico*

$$V := \sum_{i<j}^{k} ||Z_i(t) - C_{ij} Z_j(t)||^2$$

onde $C_{ij} = (p_i/p_j)^{1/2}$ e $Z_1(t), \dots, Z_k(t)$ son procesos gausianos independentes con media 0 e funcións de covarianza $K_i(s,t)$.

Nótese que o test baseado no estatístico V_n é consistente se V_n tende a infinito cando H_0 non se cumpre.

Chegados a este punto, para un tamaño mostral suficiente, rexeitarase H_0, a un nivel α, cando $V_n > V_\alpha$ onde $\mathbb{P}_{H_0}\{V > V_\alpha\} = \alpha$. Nótese que a distribución de V baixo H_0 é coñecida cando o son as funcións de covarianza $K_i(s,t)$. Non obstante, isto non sempre ocorre, polo que as estimaremos, baixo a hipótese nula, do seguinte xeito:

$$\hat{K}_i(s,t) = \sum_{j=1}^{n_i} \frac{(X_{ij}(s) - \bar{X}_{i.}(s))(X_{ij}(t) - \bar{X}_{i.}(t))}{n_i - 1}$$

Pero como a distribución de V segue a ser difícil de manexar, na práctica utilízase un procedemento asintótico Monte Carlo para calculala, de xeito que a distribución de V e de V_α baixo H_0 é aproximada pola distribución empírica da mostra $\tilde{V}_1, \dots, \tilde{V}_N$ sendo estas as replicacións artificiais de V.

4.2. ANOVA de varios factores

Neste caso imos tratar con modelos con covariables escalares. Asumimos que para cada $r = 1, \dots, R$, $s = 1, \dots, S$ existe $\mathcal{X}_i^{r,s}$, $i = 1, \dots, n_{r,s} \in \mathbb{N}$ funcións aleatorias no espazo citado $\mathbb{H}$, tal que

$$\mathcal{X}_i^{r,s}(t) = m(t) + f^r(t) + g^s(t) + h^{r,s}(t) + \gamma(t) Y_i^{r,s} + \epsilon_i^{r,s}(t),\ t \in [0,1]$$

onde:

- A función m é non aleatoria e describe a forma global do proceso.
- As funcións f^r, g^s, $h^{r,s}$ pertencen a $\mathbb{H}$ e refírense ó primeiro factor, ó segundo e a interacción entre eles.
- A $Y_i^{r,s} \in \mathbb{R}$ son cantidades aleatorias e coñecidas que inflúen no proceso de acordo cos pesos dados pola función $\gamma \in \mathbb{H}$ (non aleatoria e descoñecida).
- As traxectorias aleatorias $\epsilon_i^{r,s}$ asúmense independentes e centradas en media. Ademais, para cada r, s fixos, $\epsilon_i^{r,s}$, $i = 1, \dots, n_{r,s}$ son identicamente distribuídas.

Entón, nun modelo no que hai varios factores, pódense contrastar as seguintes hipóteses nulas:

1. O primeiro factor non ten influencia: $H_0^A:\ f^1 = \cdots = f^R = 0$
2. O segundo factor non ten influencia: $H_0^B:\ g^1 = \cdots = g^S = 0$
3. Non hai interacción: $H_0^I:\ h^{1,1} = \cdots = h^{R,S} = 0$
4. A covariable non ten influencia: $H_0^C:\ \gamma = 0$

Para cada un dos anteriores contrastes, podemos plantexar o seguinte teorema:

Teorema 4.2.1. *Sexa μ a distribución Gaussiana en $\mathbb{H}$ tal que cada das súas proxeccións unidimensionais son non dexeneradas. Entón:*

1. *Se existe r_1, r_2 tal que $f^{r_1} \neq f^{r_2}$ entón:*
$$\mu\{v \in \mathbb{H}\text{: tal que } < v, f^1 >= \cdots =< v, f^R >\} = 0$$
2. *Se existe s_1, s_2 tal que $g^{s_1} \neq g^{s_2}$ entón:*
$$\mu\{v \in \mathbb{H}\text{: tal que } < v, g^1 >= \cdots =< v, g^S >\} = 0$$
3. *Se existe $(r_1, s_1),\ (r_2, s_2)$ tal que $h^{r_1,s_1} \neq h^{r_2,s_2}$ entón:*
$$\mu\{v \in \mathbb{H}\text{: tal que } < v, h^{1,1} >= \cdots =< v, h^{R,S} >\} = 0$$
4. *Se $\gamma \neq 0$ entón:*
$$\mu\{v \in \mathbb{H}\text{: tal que} < v, \gamma >= 0\} = 0$$

Porén, calquera das hipóteses nulas (H_0^A,H_0^B,H_0^I,H_0^C) no modelo son falsas con:

$$\mathcal{X}_i^{r,s}(t) = m(t) + f^r(t) + g^s(t) + h^{r,s}(t) + \gamma(t)Y_i^{r,s} + \epsilon_i * r, s(t)$$

se e só se calquera das hipóteses nulas ($H_0^{A,v}$,$H_0^{B,v}$,$H_0^{I,v}$,$H_0^{C,v}$) no modelo:

$$< \mathcal{X}_i^{r,s}, v >=< m, v > + < f^r, v > + < g^s, v > + < h^{r,s}, v > + < \gamma, v > Y_i^{r,s} + < \epsilon_i * r, s, v >$$

son falsas.

Polo tanto, o problema en $\mathbb{H}$ está resolto se podemos solucionar o problema proxectado en $\mathbb{R}$. O número de proxección que se necesitarán é 1, segundo Cuesta-Albertos J. et al (2007).

En conclusión, úsase un test sinxelo, fácil de aplicar, flexible pero cun inconveniente: cunha única proxección pérdese poder en cando a hipótese alternativa. Por esta razón,na práctica habitual, lánzanse varias proxeccións corrixindo o resultado conxunto mediante Bonferroni, Bootstrap ou False Discovery Rate (FDR)[1].

[1]Resultado de Benjamini e Yekutieli (2001) que nos leva ó seguinte procedemento: dado cada p-valor ordenado, $p_{(1)}, \ldots, p_{(k)}$, rexeitar a hipótese nula a un nivel $\alpha \geqslant \text{ínf}\{\frac{k}{i}p_{(i)},\ i = 1, \ldots, k\}$. Entón, escollerase o p-valor corrixido como a cantidade ínf$\{\frac{k}{i}p_{(i)},\ i = 1, \ldots, k\}$

4.3. Exemplo práctico

Unha vez presentado o test para un análise ANOVA con datos funcionais, volvamos ó xa coñecido exemplo das observacións de glicosa. Comezaremos co ANOVA dun factor. Para isto, volverase ó paquete **fda.usc** e utilizarase o comando *anova.onefactor*:

```
>res.anova<-anova.onefactor(funcional,factor(lista$x$dm),nboot=200,plot=TRUE)
>res.anova$pvalue # p-valor 0
```

Ó levar a cabo este test coa variable indicadora de diabetes *dm* (cun número de mostras bootstrap igual a 200), tense que hai diferenzas significativas entre os grupos (p-valor moi próximo a 0). Na Figura 4.1 pódense ver, nos dous gráficos da esquerda, as medias de cada grupo e a media global. No gráfico do medio tamén se aprecian as remostraxes supoñendo que non hai diferenzas de grupos, habendo dúas das medias grupais que non se parecen a estas curvas. Na terceira gráfica tense a estimación da distribución baixo a hipótese de non diferenza nos grupos.

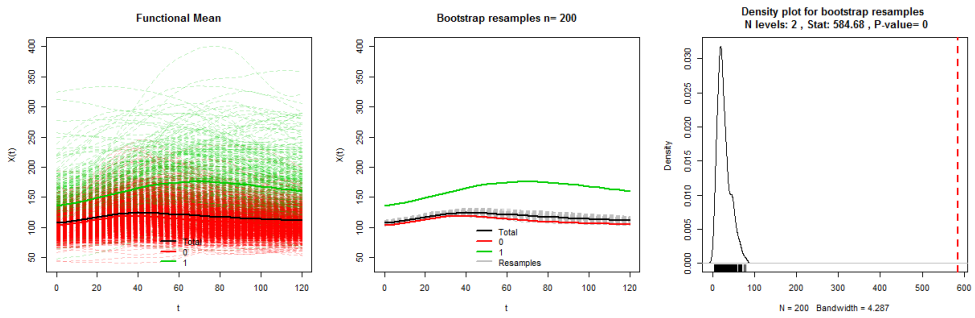

Figura 4.1: Gráficos resultantes do axuste do ANOVA dun factor mediante a metodoloxía bootstrap ós datos de glicosa, con grupos 0 e 1 coa variable *dm*.

Logo, volveuse a aplicar este test, pero neste caso coa variable *sex*, que nos indica o sexo do paciente para cada curva. Este test, segundo os investigadores, non debería dar significativo xa que para os niveis de glicosa este factor debería ser irrelevante. Non obstante, isto non ocorre e dano un p-valor moi próximo a 0 novamente. As gráficas resultantes da aplicación deste test móstranse na Figura 4.2.

```
> res.anova_consex<-anova.onefactor(funcional,lista$x$sex,nboot=200,plot=TRUE)
> res.anova_consex$pvalue # p-valor 0.0
```

O seguinte paso é aplicar o ANOVA de varios factores. Recórdese que neste test xa non só se fai mostras bootstrap, senón que se usan proxeccións aleatorias de forma que para cada proxección se ten un modelo ANOVA univariante. Entón, realizando o test, conseguimos o Cadro 4.1 onde se nos mostran os p-valores de cada factor segundo os distintos métodos de corrección.

De dito Cadro 4.1 podemos ver como a variable *dm* sempre é significativa para todos os métodos e todas as proxeccións, mentres que para a segunda variable con 1 proxección aleatoria non.

Por último, podemos considerar a posibilidade de que haxa interacción entre as variables factor. Para isto, implementamos en R a seguinte liña de comandos para podelo ter en conta:

```
res.anova3<-anova.RPm(funcional,~sex+dm+sex:dm,data.fac=dataf,RP=c(1,6,12),nboot=200)
```

Os distintos resultados para cada factor e interacción recóllense no Cadro 4.2

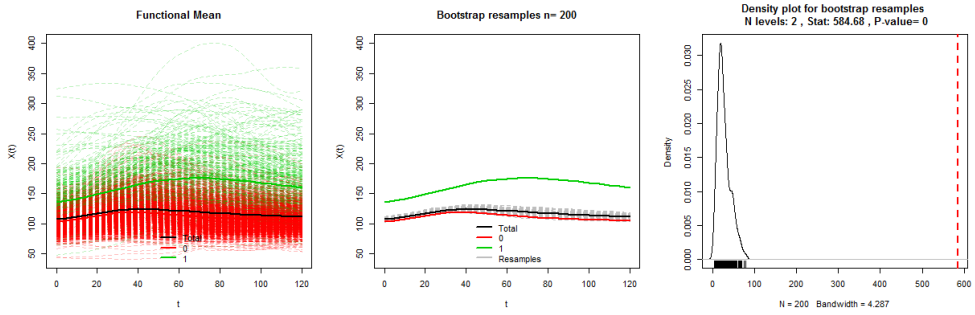

Figura 4.2: Gráficos resultantes do axuste do ANOVA dun factor mediante a metodoloxía bootstrap ós datos de glicosa, con grupos 1 e 2 coa variable *sex*.

	p-valores por Bonferroni		p-valores por FDR		p-valores por bootstrap	
	sex	*dm*	*sex*	*dm*	*sex*	*dm*
RP1	0.14958	0	0.14958	0	0.14958	0
RP6	0.00431	0	0.00432	0	0.00431	0
RP12	0.00432	0	0.00713	0	0.00432	0

Cadro 4.1: p-valores para cada factor e cada método resultantes do axuste dun ANOVA de dous factores (*sex* e *dm*) á variable funcional $funcional$.

	p-valores por Bonferroni			p-valores por FDR			p-valores por bootstrap		
	sex	*dm*	*sex* : *dm*	*sex*	*dm*	*sex* : *dm*	*sex*	*dm*	*sex* : *dm*
RP1	0.5349	0	0.1673	0.0444	0	0.0171	0.530	0	0.170
RP6	0.1426	0	0.0766	0.0005	0	0.0005	0.065	0	0.025
RP12	0.0035	0	0.0262	0.0016	0	0.0016	0	0	0.025

Cadro 4.2: p-valores para cada factor, e interacción entre eles; e cada método resultantes do axuste dun ANOVA de dous factores (*iy* e *iy*2) á variable funcional *espectro*.

Capítulo 5

Aplicación a datos reais: proxecto AEGIS

Durante á estancia no departamento de epidemioloxía do Hospital Clínico de Santiago de Compostela apareceu, entre outros, un estudo sobre un proxecto que se realizou no municipio de A Estrada[1].

O estudo foi dirixido polos doutores Arturo González-Quintela (Medicina Interna) e Francisco Gude (Epidemioloxía Clínica) e conta coa participación activa dun amplo grupo de profesionais pertencentes a diferentes disciplinas como a atención primaria, alergoloxía, bioloxía, bioquímica, enfermaría, endocrinoloxía ou bioestatística.

5.1. Características xerais do proxecto

O Proxecto da Estrada baséase nun estudo de base poboacional, nunha mostra representativa da poboación xeral de adultos, cun amplo tamaño mostral, extensa fenotipación e documentación individual e con almacenamento reglado de mostras biolóxicas (suero, ouriños e sangue). Os participantes foron escollidos do xeito que se mostra no diagrama de fluxo da Figura 5.1.

[1] O municipio de A Estrada pertence á comunidade autónoma de Pontevedra. Conta cun total de poboación de 22.362 persoas, onde 10.538 son homes e 11.824 mulleres, segundo a páxina oficial do concello (Concello da Estrada 2016). A súa superficie é duns 281.8 quilómetros onde se estima que cerca dun cuarto da poboación vive na cidade e o resto nun entorno rural.

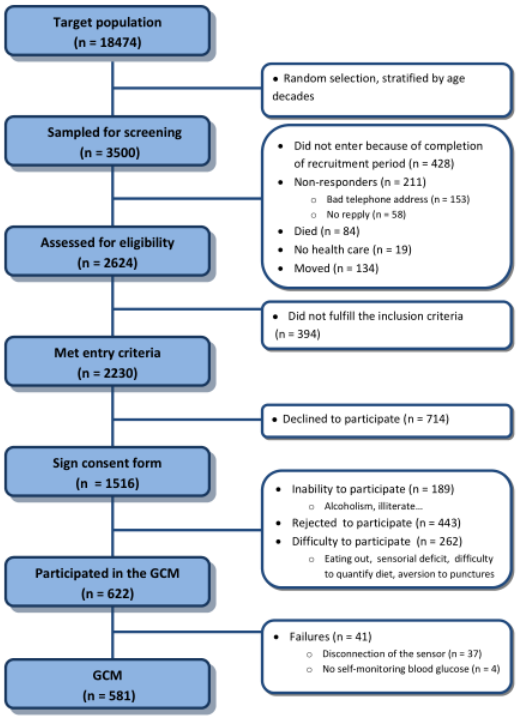

Figura 5.1: Diagrama de fluxo representando o procedemento que se seguiu para a selección dos participantes.

Como se mostra, para a selección estratificouse a poboación total en ideades desde os 18 ata os 80 anos e xerouse unha mostra de 500 persoas en cada grupo de idade, obtendo 3500 individuos en total. De estes, excluíronse persoas por razóns como que faleceran, por non ter asistencia sanitaria, por presentar demencia, atraso mental, enfermidades cerebrovasculares graves, cáncer ou outros motivos (véxase a Figura 5.1). Dos resultantes, un total de 1516 persoas (55 % mulleres e 45 % homes) accederon a participar. Ó final, de entre eles, fíxoselle a monitorización continua de glicosa (CGM) a un total de 581 pacientes. A partir de agora imos centrarnos nestes últimos participantes.

Todos os pacientes acudiron a unha consulta ó centro de saúde de A Estrada para a realización dunha entrevista clínica e determinacións que incluía: cuestionarios estruturados con datos demográficos e antropométricos, estilos de vida con rexistro da actividade física, inxesta dietética, consumo de tabaco e alcohol; unha batería de test psicolóxicos, exame periodontal, probas alérxicas, mostras sanguíneas e a insertación dun dispositivo de monitorización continua de glicosa[2]. Este último, disponse na rexión abdominal de cada paciente durante a súa visita inicial e para realizar a calibración do dispositivo, solicítase ós participantes que realicen polo menos tres determinacións diarias de glicemia capilar[3]. Así, o dispositivo de CGM (monitorización continua de glicosa) almacena os niveis de glicosa durante as 24 horas do día, durante 6 días cada 5 minutos. Por suposto, os pacientes deben levar unha vida normal, anotando á hora de inxesta de alimentos e a actividade física que se realizou cada un dos 6 días nos cadernos que se lles proporcionou na primeira visita.

Nótese que o estudo foi levado a cabo de acordo cos principios da Declaración de Helsinki (proposta de principios éticos para a investigación médica en seres humanos, incluída a investigación do material

[2]En concreto o iPro®, deseñado por Medtronic

[3]Para realizar esta medición empregouse glucómetros, que utilizan tiras reactivas que deben ser introducidas no aparato cunha pequena mostra de sangue.

humano e información identificables) e coa lexislación vixente. Ademais, foi aprobado polo Comité Ético de Investigación Clínica de Galicia, Santiago de Compostela.

5.2. Introdución e preparación das bases de datos

Unha vez realizado a medición da glicosa durante estes días, os datos obtidos dispóñense nun documentos *.csv* onde aparece por orde as columnas *id*, que reflexa o número do paciente; *Día*, que toma valores de entre 1 e 6; $Fecha$, indicando a fecha da monitorización; $Hora$, indicando a hora de cada medición; $Marca.de.fecha.y.hora$ que indica conxuntamente as dúas últimas columnas citadas; *Glucosa* que indica o nivel de glicosa no momento da medición e unha variable chamada *come*. Na primeira parte das prácticas, esta última columna foi a protagonista, xa que ós investigadores do hospital interesábanlles recoller a diferencia que había entre cada unha das fases que pasaba o paciente: horas de sono, diferencia entre a hora entre que se ergueron e almorzaron, diferenza entre o almorzo e a seguinte comida, etc.

Para isto, tívose que revisar os 581 cadernos que cada paciente tiña e anotar na base de datos a hora correspondente, e para cada un dos 6 días, os datos que faltaban como a hora que se deitaban, a hora que se levantaban e a hora de almorzo (xunto con unha corrección dos datos xa implementados). Se se desexa ver o deseño dos cadernos, unha folla en branco pode verse en Apéndice A.

Unha vez realizado estes pasos, tívose que modificar e aumentar a función que se tiña no departamento para ler e sacar as variables desexadas do documento *.csv*. A versión á que se chegou foi a que se pode ver no Apéndice B. En concreto, esta función, ademais de poder ler e extraer unha serie de datos funcionais para cada individuo e cada día, extrae os seguintes tempos:

- $dif07$: tempo en minutos que durmiu o paciente cada día.
- $dif01$: tempo en minutos que o paciente tardou en almorzar desde que despertou.
- $dif123$: tempo que pasou entre o almorzo e a seguinte comida (media mañá, que está como representado como 2; ou comida, denotado por 3) medido en minutos.
- $dif345$: tempo que transcorreu entre a comida e a merenda (4) ou cea (5), medido en minutos.
- $dif56$: tempo que pasou entre a cea e a seguinte comida (se a houbo), medida en minutos.

Estes datos que se calculan aportan gran información ó estudo da glicosa, xa que se o suxeito come durante un tramo de glicosa que estamos a analizar, este verase afectado. Se non tivésemos este dato, poderíamos pensar que están afectando outras variables en vez da inxesta calórica.

Unha das complicación que tivo calcular estes datos foi que, como é habitual, algunhas persoas non se levantaban e deitaban nun mesmo día (por exemplo, as persoas que traballan ata altas horas da madrugada), pero isto solventouse construíndo un vector en cada día e comprobando a que momento pertencía cada sinalización. Por exemplo, se había nun mesmo día dous 7 (hora de deitarse) no día 2 e ningún no día 1, significa que no día 1, esa persoa deitouse despois das 12 da noite.

Outro dato interesante que se pode sacar destas novas variables é a causa pola que un paciente se saltou o almorzo. Ata o de agora, se un paciente se saltaba o almorzo, podíase ver no aporte calórico dese día a esa hora. Non obstante, non tiñamos maneira de saber se foi porque o paciente estaba esperto pero non almorzou ou se quedou durmindo ata horas o suficientemente tardes como para que a comida que fixo ó levantarse encaixara máis na hora de comida. Coa variable $dif01$ isto xa se pode saber porque esta deseñado de tal maneira que se o seu valor é negativo para algún día dun individuo, isto significa que non estaba desperto á horas nas que debería almorzar (tanto pola súa rutina como polas horas en sí).

Logo disto, púidose unificar as variables tanto para o almorzo, comida ou cea coas demais variables das que se dispón (aparte das xa citadas). As variables que se dispoñen para os tres momentos son:

- *age*: anos do paciente no momento do estudo.
- *sex*: sexo do paciente.
- *dm*: variable factor que indica se o paciente está diagnosticado como diabético.
- *ado*: antidiabéticos orais.
- *insul*: variable indicadora se o paciente estaba cun tratamento de insulina.
- *weight*: peso do paciente (kg).
- *heigth*: altura do paciente (cm).
- *bmi*: índice de masa corporal calculado como:

$$bmi = \frac{Peso\ (kg)}{Altura^2\ (m)}$$

- *waist*: medida de cintura do paciente (cm).
- *ipq*: variable estratificada extraida do test *ipaq* indicando a actividade física que realiza o paciente. O grupo 1 correspóndese a individuos inactivos, o grupo 2 a minimamente activos e o grupo 3 a individuos altamente activos.
- *mett*: variable que estima o gasto metabólico.
- *tab*012: variable que divide a mostra en fumadores (grupo 2), exfumadores[4] (grupo 1) e non fumadores (grupo 0).
- *oh*4: variable factor indicando a cantidade de alcohol que o paciente toma na súa vida cotiá: grupo 0 para os abstemios ou os que beben en ocasións especiais, grupo 1 para os que beben menos de 140 gramos á semana, grupo 2 para os que toman entre 140 e 280 gramos e grupo 3 para máis de 280 gramos.
- *glu*: medida de glicosa no momento da primeira visita.
- *mdrd*: medida que estima a modificación da dieta en enfermidade renal. Esta é calculada a través das variables: creatinina en soro, idade, grupo étnico e xénero.
- *a*1*c*1: hemoglobina glicada. Esta variable vai ser de grande interese, xa que un valor alto desta está moi relacionada coa presenza de glicosa alta nos dous meses anteriores.
- $fru1$: fructosamina (proteína glicada como a hemoglobina, pero que ten unha vida media menor).
- *sm*: síndrome metabólico. Esta variable indica se no paciente se mostran polo menos as seguintes alteracións metabólicas: obesidade abdominal, trastorno de lípicos en sangue, alteración da glicosa (hiperglicemia) e aumento da presión arterial.
- Variables relacionadas coa dieta de todo o día do paciente: *energia*, *proteinas*, *lipidos*, *grasas_sat*, *grasas_mono*, *grasas_poli*, *colesterol*, *carbohidratos*, $fibra$, *calcio*, *hierro*, *magnesio*, $fosforo$, *potasio* e *sodio*.

[4]Unha persoa considérase exfumadora se leva máis dun ano sen fumar.

Non obstante, para o almorzo temos variables específicas relacionadas coa inxesta realizada: *prot*, cantidade de proteídas no almorzo; HCg, cantidade de hidratos de carbono; $fibrag$, cantidade de fibra, *grasag*, graxa do almorzo; e *valener*, valor enerxético do almorzo.

5.3. Análise exploratoria dos datos

Para comezar, fagamos unha revisión a cada unha das variables coas que imos traballar. Unha pequena táboa resumo das variables continuas atópase no Cadro 5.1, mentres que, para as variables factor se observa que:

	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
age	18.00	37.00	48.00	48.03	60.00	87.00
weight	41.00	63.05	73.95	75.57	84.65	145.00
bmi	17.36	24.25	27.54	28.15	31.30	52.54
waist	56.00	80.00	91.00	91.17	100.00	141.00
mett	0	495	1386	2365	2946	18190
glu	63.00	81.00	88.00	93.36	98.00	254.00
mdrd	40.61	89.69	102.60	103.90	116.50	267.70
$a1c1$	3.100	5.200	5.400	5.565	5.600	10.100
$fru1$	105.0	186.0	215.0	222.5	247.0	526.0

Cadro 5.1: Mínimo, primeiro cuartil, mediana, media, terceiro cuartil e máximo de cada un dos datos escalares cuantitativos cos que se vai traballar.

Tamén podemos ver como son as variables factor.

- Na mostra contamos cun 62.2 % de mulleres fronte a un 37.8 % de homes.
- Do total da poboación estudada hai un 88 % de persoas non diabéticas.
- En canto a actividade física, un 35.7 % realiza actividade física baixa, un 38.7 % realiza actividade moderada e o resto alta.
- Un 53.5 % da poboación é non fumadora, un 26.5 % é fumadora e o resto exfumadora.
- En canto ó consumo de alcohol, as porcentaxes de menor a maior consumo son: 39.5 %, 40.6 %, 14.5 % e 5.4 %.
- Por último, un 81.5 % da poboación non ten síndrome metabólico.

Por outra banda, o que primeiro se debería facer é representar graficamente as curvas. Na Figura 5.2 pode verse as curvas de glicosa para os non diabéticos na primeira fila tanto para almorzos, comidas e ceas. Na segunda fila represéntanse as mesmas curvas pero considerando todos os individuos (non só os non diabéticos). Nótese que as curvas para os non diabéticos parecen estar máis compactas mentres que cando aparecen os diabéticos o rango dispárase (non só se dispara, senón que parece que hai maior variabilidade).

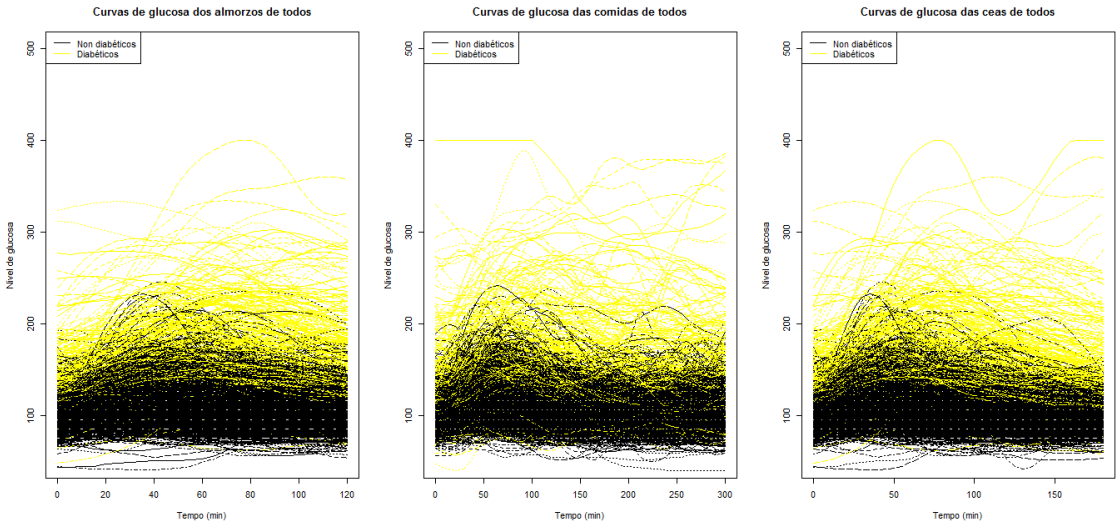

Figura 5.2: Representación gráfica das curvas de glicosa para almorzos, comidas e ceas para os diabéticos e non diabéticos. En negro represéntase os non diabéticos e en amarelo os diabéticos.

Analogamente, como xa se explicou, o paquete **fda.usc** permite a opción de obter a derivada dos datos funcionais coa función $fdata.deriv$. Por criterios médicos, o que máis interesa nas curvas de glicosa son as flutuacións ou crecementos e decrecementos extremos para identificar os posibles casos de diabetes ou pre-diabetes. Entón derívase os datos funcionais e obtemos as curvas que se representan na Figura 5.3. De novo, a primeira fila correspóndese con individuos non diabéticos e na segunda todos as persoas da mostra. Nótese que como pasaba no caso das curvas sen derivar, as flutuacións son máis atenuadas, sobre todo durante a cea, con toda a poboación.

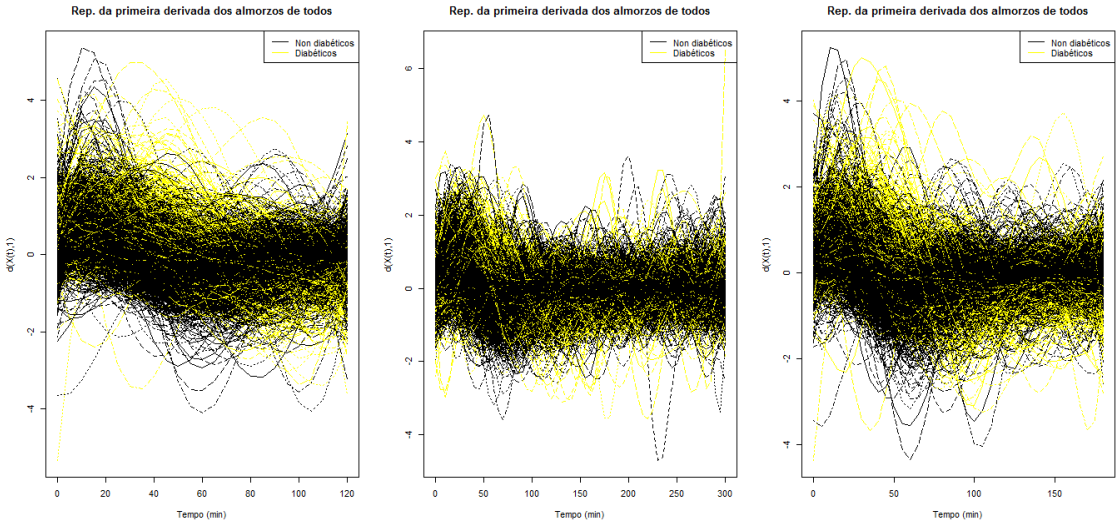

Figura 5.3: Representación gráfica da derivada das curvas de glicosa para almorzos, comidas e ceas para os non diabéticos e toda a mostra. En negro represéntase os non diabéticos e en amarelo os diabéticos.

O seguinte paso que se realizou neste traballo, foi a busca dunha boa representación dos datos funcionais. Ó igual que se fixo no capítulo de representación, escolleuse unha curva e representouse por cada un dos métodos explicados con anterioridade. O resultado foi a Figura 5.4, onde se escolleu un individuo diabético (individuo 10) e un non diabético (individuo 1002) e, cos métodos máis importantes, intentouse representar dito dato para almorzo, comida e cea. Un dato a destacar de dita figura é que, cunha simple ollada, podemos saber cal dos dous é o diabético debido ó rango de valores que toma. Con respecto ás aproximacións, todas axustan bastante ben a curva. Se tiveramos que descartar algunha sería a base de B-splines xa que parece que non capta de todo as pequenas flutuacións de glicosa, como no almorzo do individuo 10.

Ademais, isto pódese realizar para todas as curvas mediante o comando de R *min.basis* e *min.np* tanto para bases como para representación non paramétrica, respectivamente. Logo, extráese os parámetros óptimos en cada segundo cada tipo de método.

Os parámetros empregados para cada unha destas aproximacións son:

- Para o individuo 10:
  - Para o almorzo:
    - $h = 1$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
    - $\lambda = 0.1$ e o un número de bases igual a 6 para o método de bases B-spline
  - Para a comida:
    - $h = 1$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
    - $\lambda = 10$ e o un número de bases igual a 16 para o método de bases B-spline

- Para a cea:
  - $h = 1$ para a aproximación kernel co método de Nadaraya-Watson.
  - $h = 1.857143$ para a aproximación kernel con regresión linear local.
  - $h = 1.857143$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
  - $\lambda = 10$ e o un número de bases igual a 13 para o método de bases B-spline

- Para o individuo 1002:
  - Para o almorzo:
    - $h = 1$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
    - $\lambda = 10$ e o un número de bases igual a 13 para o método de bases B-spline
  - Para a comida:
    - $h = 1.857143$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.857143$ para a aproximación kernel con regresión linear local.
    - $h = 1.857143$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
    - $\lambda = 1$ e o un número de bases igual a 21 para o método de bases B-spline
  - Para a cea:
    - $h = 1$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local.
    - $h = 1.428571$ para a aproximación kernel con regresión linear local utilizando o criterio Rice.
    - $\lambda = 10$ e o un número de bases igual a 16 para o método de bases B-spline

- Para toda a mostra:
  - Para o almorzo:
    - $h = 2.247475$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 1.686869$ para a aproximación kernel con regresión linear local, coincidindo coa aproximación co criterio Rice.
    - $\lambda = 128$ e o un número de bases igual a 11 para o método de bases de Fourier.
    - $\lambda = 32$ e o un número de bases igual. a 29 para o método de bases B-spline
  - Para a comida:
    - $h = 3.75$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 3.75$ para a aproximación kernel con regresión linear local, coincidindo coa aproximación co criterio Rice.
    - $\lambda = 128$ e o un número de bases igual a 23 para o método de bases de Fourier.
    - $\lambda = 32$ e o un número de bases igual. a 29 para o método de bases B-spline
  - Para a cea:
    - $h = 2.25$ para a aproximación kernel co método de Nadaraya-Watson.
    - $h = 2.25$ para a aproximación kernel con regresión linear local, coincidindo coa aproximación co criterio Rice.

- $\lambda = 128$ e o un número de bases igual a 17 para o método de bases de Fourier.
- $\lambda = 32$ e o un número de bases igual. a 23 para o método de bases B-spline

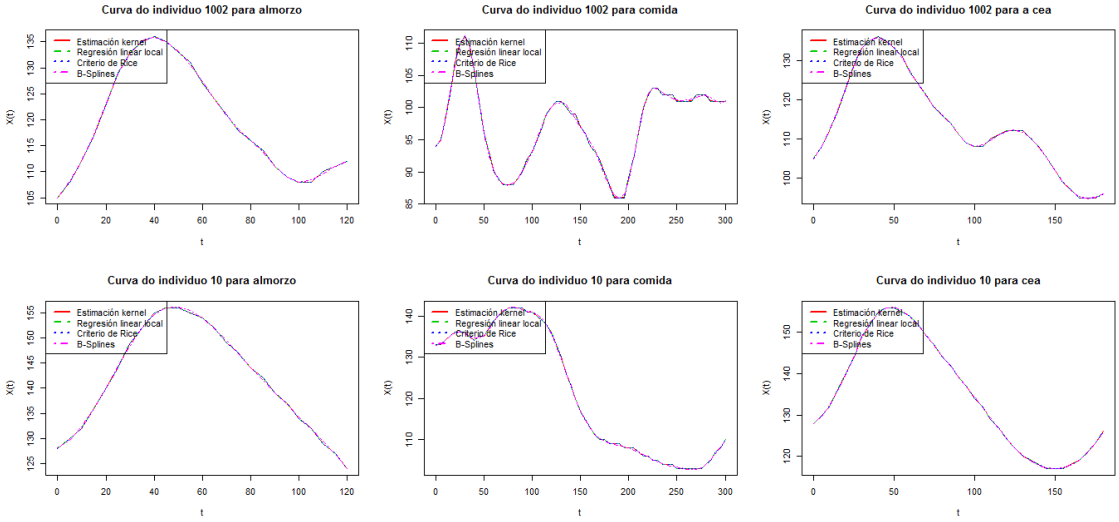

Figura 5.4: Curvas de glicosa para o almorzo, comida e cea ó longo do tempo para o individuo 10 (diabético) e 1002 (non diabético), xunto coas súas aproximacións.

Por outra banda, tamén cabe a posibilidade de representar as curvas mediante compoñentes principais. Entón calculamos cantas compoñentes precisaríamos para cada momento do día (intentando chegar ó redor do 90 % da variabilidade explicada) xunto coa variabilidade que explica cada unha delas. Os resultados para almorzos, comidas e ceas para individuos non diabéticos da mostras represéntanse no Cadro 5.2. Nótese que existe unha gran diferencia entre a variabilidade explicada polas compoñentes do almorzo e cea e a variabilidade explicada das comidas. Isto pode deberse, probablemente, polos hábitos de consumo dos individuos; é dicir, os almorzos e as ceas deben ser máis homoxéneas mentres que nas comidas hai máis variabilidade de alimentos.

	Comp. necesarias	Comp.1	Comp.2	Comp.3	Comp.4	Variabilidade total
Almorzo	3	72.89	15.57	7.58	-	96.03
Comida	4	60.46	14.97	8.81	4.78	89.01
Cea	3	67.11	16.30	8.34	-	91.75

Cadro 5.2: Cadro resumo da aplicación das compoñentes principais ás curvas de almorzo, comida e cea. Nel están o número de compoñentes a considerar, a variabilidade explicada por cada unha delas e variabilidade total explicada.

Logo de ver dito cadro, representamos as compoñentes principais ó longo do tempo na Figura 5.5, de novo para os tres momentos do día e para os non diabéticos. Nesta pódense realizar as seguintes observacións para cada momento:

- No almorzo, a primeira compoñente principal está sempre por debaixo da media, polo que as curvas que máis puntúan nela son os niveis de glicosa baixos. En canto á segunda, ten unha subida aproximadamente á hora de comezar o rexistro con respecto á media debido, posiblemente, á inxesta do almorzo; e na terceira contan as curvas iguais que para a segunda pero desprazados 45 minutos, posiblemente almorzaron tarde ou fixeron unha comida á media mañá.

- Na comida, os resultados son máis difíciles de discutir por mor do número de compoñentes a considerar. De todos xeitos, a primeira está claramente por encima da media mentres que as demais comezan por debaixo da media e teñen unha subida de glicosa, ou sexa, mostran distintos comportamentos á hora da inxesta (de feito, parece que experimentan unha subida de glicosa pero atrasada no tempo).

- Por último, na cea podemos distinguir (ó igual que fixemos na comida) entre unha que está lonxe con respecto á media, a primeira, e que parece que non se ve afectado polo aporte de azucre da cea mentres que as demais sí que lle acontece unha subida (aproximadamente ós 30-45 minutos). Nótese tamén que a terceira, ó final do tempo, volve a estar por encima da media, mentres que as demais non. Isto pode ser un indicador preliminar de que individuos posúen maior variabilidade das curvas, pois serán os que máis puntuación teñan nesta compoñente.

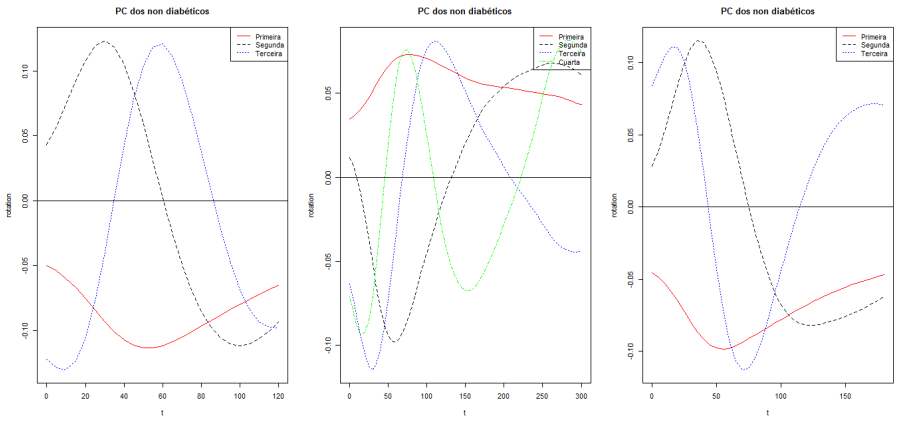

Figura 5.5: Representación das curvas principais para o almorzo, comida e cea cuxa variabilidade explicada está no Cadro 5.2.

Por outra banda, pódese comprobar a correlación que poida existir entre as puntuacións das compoñentes principais de todos os momentos (que se expresan por $scorexy$ sendo x o momento do día e y o número da compoñente principal). Isto móstrase na Figura 5.6, xunto coas correlacións entre as variables que consideramos. Como podemos ver, non existe unha gran correlación entre as compoñentes e as demais variables, aínda que sí que se pode ver algo de relación entre os scores das compoñentes e as variables age, $weight$, $a1c1$ e glu.

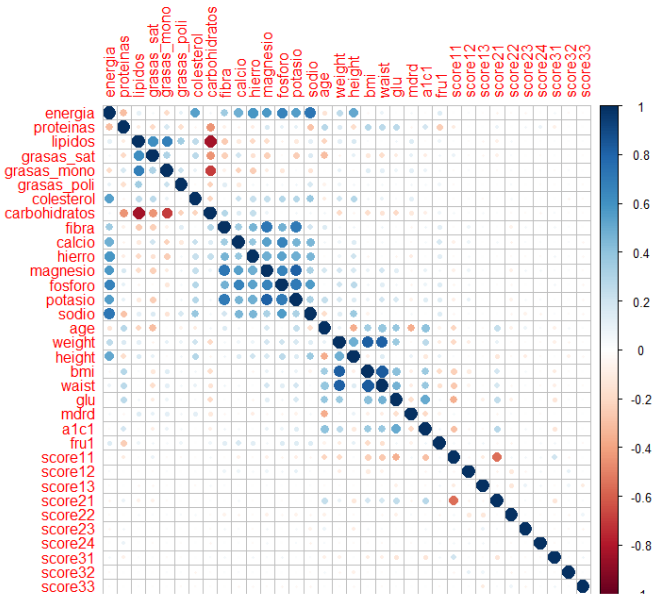

Figura 5.6: Cadro resumo das correlacións existentes entre as variables que temos e as compoñentes principais para cada momento do día (están representados como $scorexy$, onde x é o momento e y a compoñente).

Ó igual que se podía facer co exemplo dos anteriores capítulos, podemos extraer as medidas de centralización. O primeiro será calcular e representar a media teórica e as medias mostrais calculadas coa distancia do supremo e a do espazo $\mathcal{L}_2$. Isto pode verse na Figura 5.7.

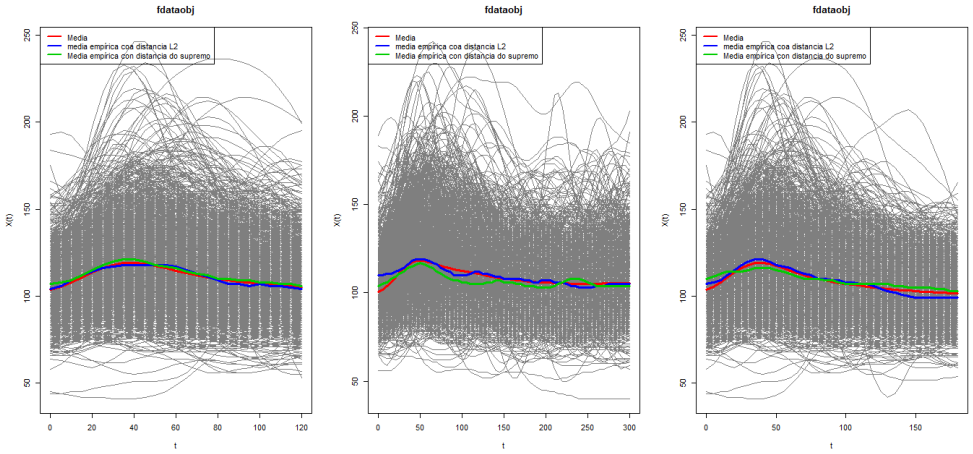

Figura 5.7: Representación gráfica dos datos de glicosa para o almorzo, comida e cea coa media teórica (en vermello), a media mostral coa distancia de $\mathcal{L}_2$ e a media mostral coa distancia do supremo (verde).

De igual xeito, pódese realizar unha representación da varianza dos datos funcionais para a mostra cos individuos non diagnosticados como diabéticos, xunto coas varianzas das curvas máis profundas segundo o método que queiramos elixir. Isto vese na Figura 5.8. Nótese que a pesar de non ter o mesmo

rango que a varianza total, a varianza das curvas máis profundas ten a mesma forma e que a que máis variabilidade mostra é a escollida pola profundidade de Fraiman-Muniz para todos os momentos do día (almorzo, comida e cea). Por último, destacar que nestes tres momentos, todas parecen seguir o mesmo patrón debido a como son extraidas (despois de cadansúa comida).

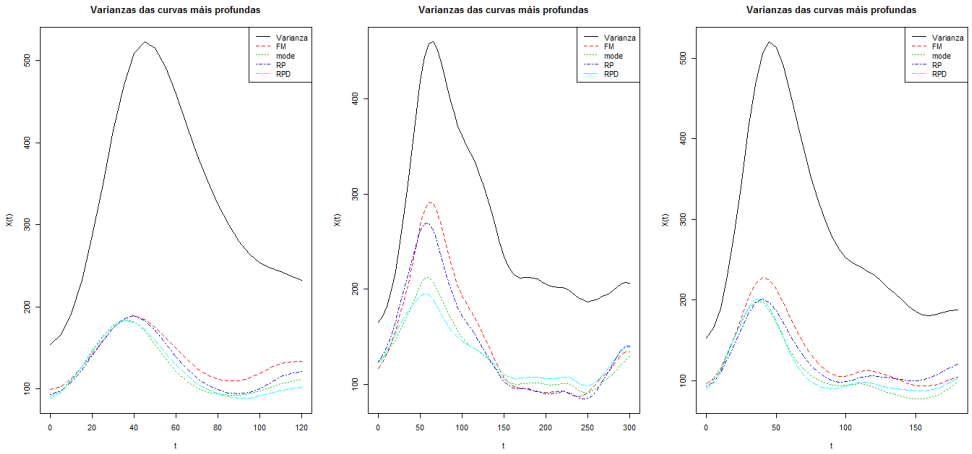

Figura 5.8: Representación da varianza dos datos funcionais (en negro), xunto coa varianza das curvas máis profundas segundo o método (FM, modal, RP ou RPD).

Por último, calcularase a matriz de covarianzas da poboación non diabética para almorzos, comidas e ceas. A representación desta atópase na Figura 5.9. Por regra xeral, existe unha maior correlación entre os datos recollidos en minutos próximos. En concreto, nas horas do almorzo a correlación esténdese a máis minutos que nas horas das comidas, que a correlación parece estar máis restrinxida ós minutos máis pretos. Nótese que isto é unha das razóns polas que se precisan máis compoñentes para a comida que para o resto.

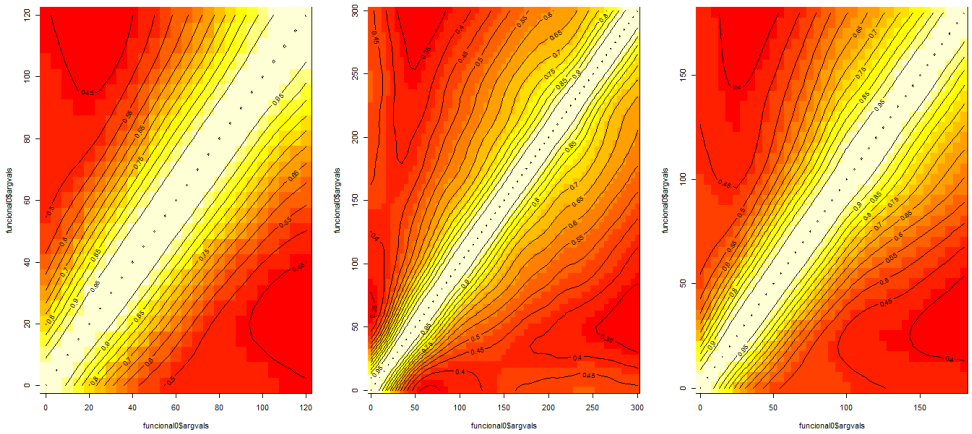

Figura 5.9: Representación gráfica da matriz de covarianzas dos datos de glicosa nos almorzos, comidas e ceas.

Unha vez vistas estas medidas descritivas, comezaremos calculando as profundidades dos datos

funcionais. O primeiro que se fará é calcular a Figura 5.10. Aquí móstrase a media do 25 % das curvas máis profundas da poboación non diabética, calculadas cos métodos FM, modal, RP e RPD. Non parece que haxa diferenzas significativas entre as curvas calculadas cos distintos métodos.

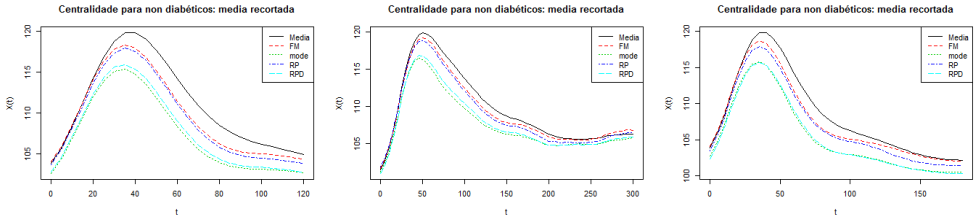

Figura 5.10: Representación da media do 25 % das curvas máis profundas seguindo os criterios que aparecen na lenda; para todos os tres momentos do día.

Outra interesante medida a calcular son as distintas profundidades. Na Figura 5.20 calcúlase as medidas de profundidade para unha lista de obxectos de datos funcionais. Neste caso, utilizouse o dato funcional e a súa derivada, xa que medicamente interesa tanto as curvas como as súas flutuacións, é dicir, a súa primeira derivada. Para ver con detalle cales son os pasos que se realizan véxase Febrero Bande M. (2016). Como pasaba no caso sinxelo do capítulo 2, as curvas máis profundas debúxanse cunha cor máis escura, mentres que as menos profundas cunha cor clara. Ademais, represéntase a mediana e a curva máis profunda segundo o método elixido. Por último, cabe destacar que as curvas con maior profundidade son as obtidas pola profundidade modal, tanto para os almorzos, comidas e ceas. Este método é moito máis fiable que o que se usou nos anteriores capítulos xa que ten en conta á vez ambos conxuntos de datos funcionais, non só os datos orixinais. Pódese facer unha comparación coa profundidade por RP, por exemplo, que se mostra na Figura 5.11. Nesta última, non parece haber ningún dato funcional que teña pouca profundidade e que atravese a mediana. Non obstante, na profundidade por RP da Figura 5.20 (terceira fila) podemos comprobar a simple vista que isto si que acontece.

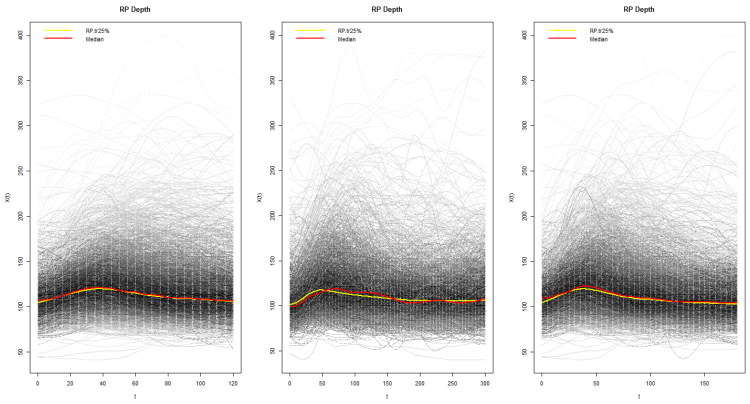

Figura 5.11: Profundidade con todos os individuos da mostra (diabéticos e non diabéticos) coa profundidade RP.

Traballemos un pouco entón coa profundidade calculada por proxeccións aleatorias. Consideremos entón todos os individuos, diabéticos e non. A representación destas está na Figura 5.11.

Ademais podemos comprobar as bandas bootstrap de confianza para ditas medidas. Na Figura 5.12

móstranse ditas bandas para a mediana e para a profundidade calculada por proxeccións aleatorias coa mostra dos individuos non diabéticos. Nótese que as bandas son menos anchas para a mediana que para a profundidade RP.

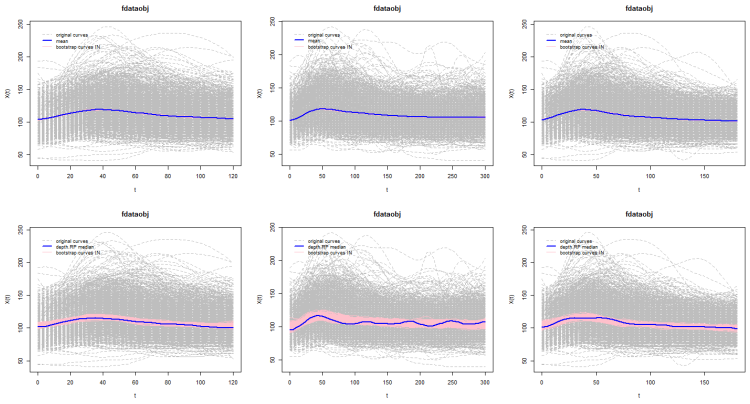

Figura 5.12: Bandas de confianza bootstrap para a poboación non diabética.

5.4. Cálculo de datos atípicos

Despois de calcular e interpretar as profundidades, buscamos as curvas atípicas. Por que sospeitamos que existen outliers na nosa mostra? Observemos a Figura 5.13. Nel están representados a profundidade fronte as variables $fibra$ e $dif123$, categorizadas en 4 grupos. Aquí vemos que polo menos existe un outlier en cada gráfica, polo que unha análise de datos atípicos é razoable.

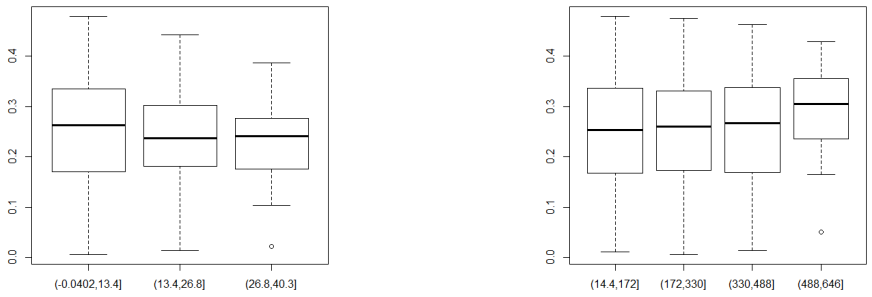

Figura 5.13: Diagrama de caixas das variables $fibra$ e $dif123$ no almorzo segundo a profundidade RP.

Polo tanto, comprobaremos cales son estas curvas. Comezaremos estudando os outliers segundo as profundidades FM, RP e RT. O resultado de aplicar este estudo é a Figura 5.14. En verde están representados os outliers calculados polo método de recortes mentres que en cor vermello co método de

ponderacións. Nas filas están os momentos do día (almorzos, comidas e ceas) e por columnas os distintos métodos. Nótese que para a profundidade FM atópanse moitos menos outliers que para os outros dous métodos. En canto a diferenzas entre os momentos, non parece que haxa grandes diferenzas.

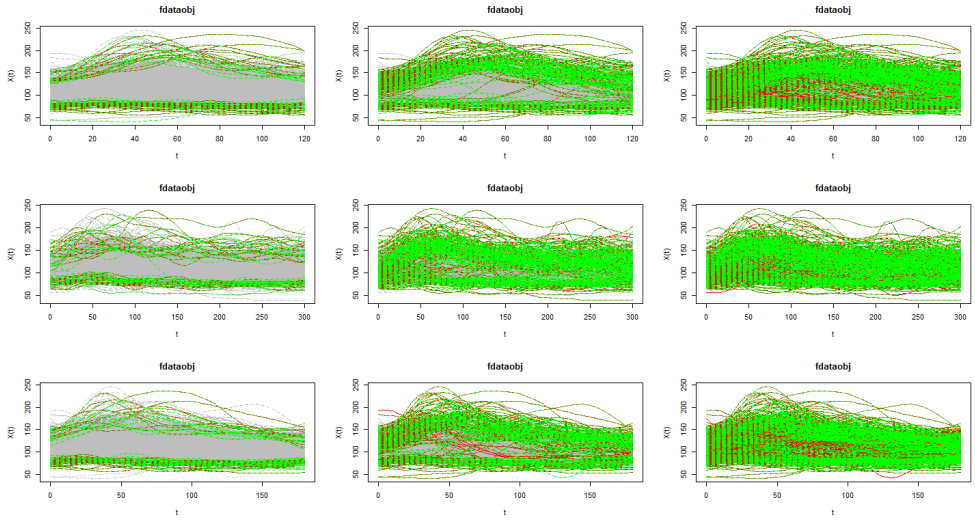

Figura 5.14: Datos atípicos extraidos segundo os métodos FM, RP e RT (por columnas) para os almorzos, comidas e ceas (dividido por filas). De cor verde están calculados polo método recortado e en cor vermello co método ponderado.

Outra maneira de calcular outliers ou datos atípicos é mediante os métodos HDR e HS (coa representación mediante o bagplot). Realizando dito estudo[5] obtemos a Figura 5.15. Os resultados están representados tanto de xeito bivariante como en forma de dato funcional. Por filas atópanse os momentos do día, como na anterior figura, e por columnas os distintos métodos: HDR e HS, respectivamente. Ademais, tanto nas lendas (na representación funcional) como no gráfico bivariante atópanse o número da curva pola que se identifican os datos funcionais.

[5]Realizado co paquete **rainbow** en ver do **fda.usc**, co que estivemos traballando ata o de agora.

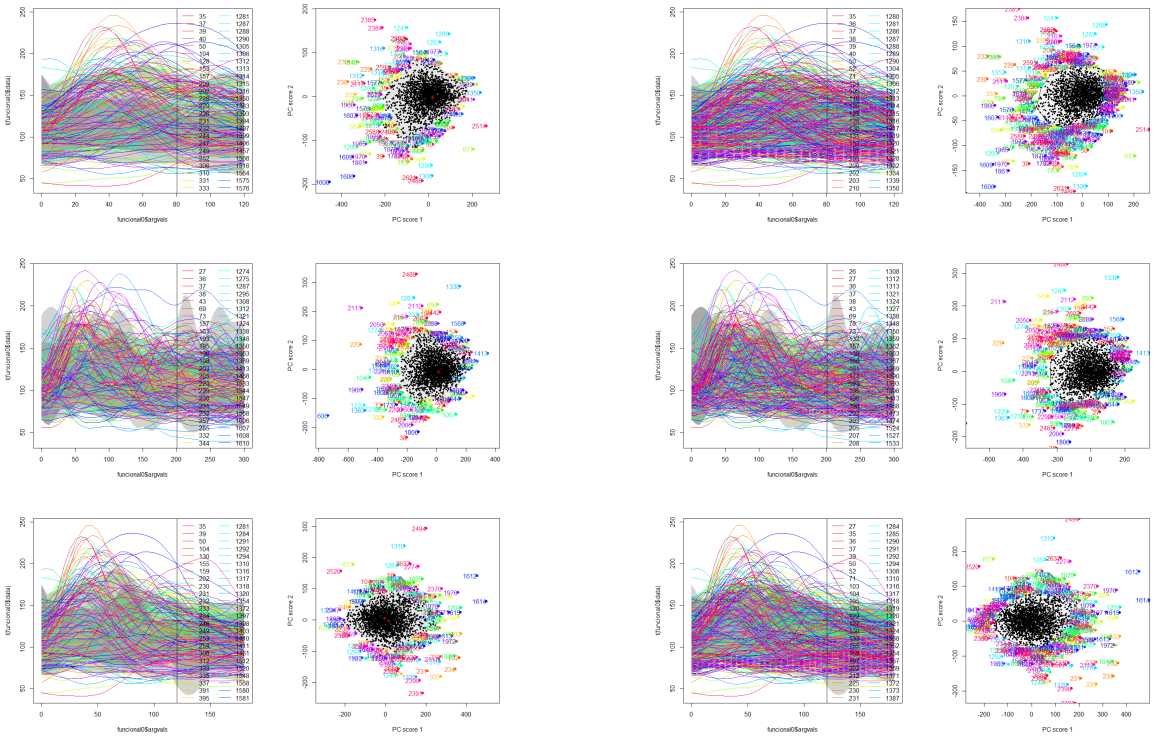

Figura 5.15: Datos atípicos calculados a través dos métodos HDR e HS.

Non obstante, estas análises compórtanse como se todas as curvas procedesen de individuos ou casos homoxéneos, que non están influídos por nada máis que a forma das curvas. Isto non ocorre neste caso, onde cada curva de glicosa ven modificada tanto por variables propias de cada individuo como pola inxesta de glicosa. Por isto, non pode ser comparable unha curva dun individuo que non almorza con outra cuxo individuo realiza unha inxesta calórica rica en glicosa. Polo tanto, o lóxico sería pensar en realizar a análise de datos atípicos en subconxuntos de curvas coas mesmas (ou razoablemente parecidas) características e cun tamaño suficiente como para poder buscar outliers. Así, tendo en conta variables como se o individuo foi diagnosticado con diabetes, se ten un índice de masa corporal elevado ou a cantidade de hidratos de carbono inxeridos son altos; realizamos un estudo de datos atípicos coa profundidade calculada polo método de proxeccións aleatorias (RP). O resultado disto é a Figura 5.16. Aquí móstranse os datos atípicos que se considerarían máis fiables que os anteriormente calculados para os almorzos, comidas e ceas. Esta análise tamén se pode realizar coa poboación unicamente non diabética, cuxo resultado se pode ver na Figura 5.17. Estes datos gardaranse para logo estudalos a conciencia na seguinte sección.

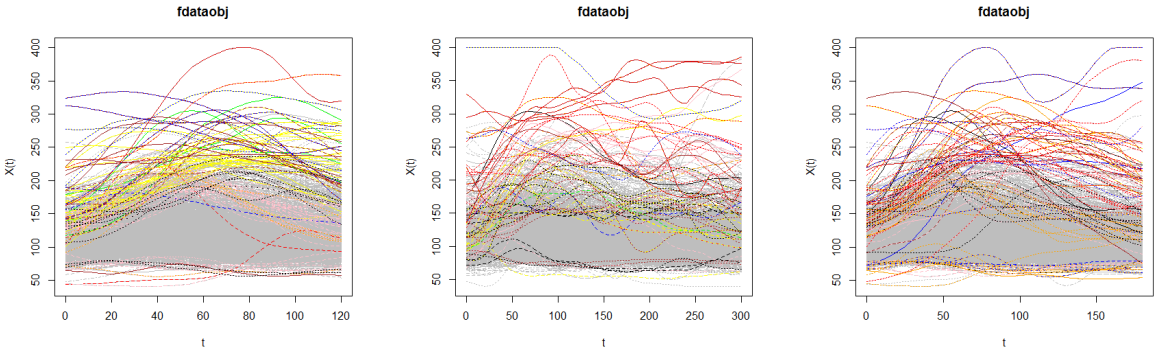

Figura 5.16: Datos atípicos de toda a poboación calculados separando a mostra por características comúns.

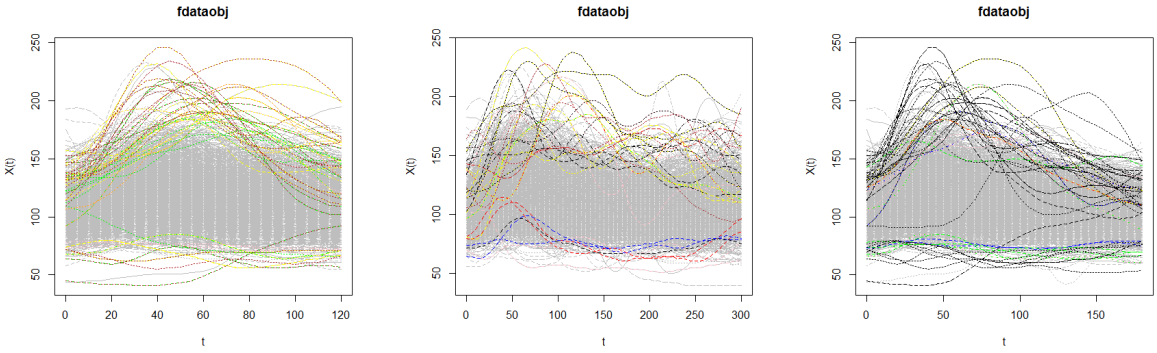

Figura 5.17: Datos atípicos da poboación non diabética calculados separando a mostra por características comúns.

5.5. Clasificación

Nesta sección aplicaremos os dous tipos que se estudaron anteriormente: a clasificación non supervisada e a non supervisada. Comezaremos por esta última.

5.5.1. Clasificación non supervisada

Logo de conversar cos investigadores, chegouse a conclusión de que o máis razoable sería dividir ou clasificar os datos funcionais en 4 grupos ó realizar unha clasificación non supervisada, xa que existe a posibilidade de que os grupos se mostren do seguinte xeito: un grupo no que aparecerían os diagnosticados como pacientes diabéticos pero que non seguen un control adecuado da enfermidade, outro no que aparecerían os pacientes diabéticos e a poboación prediabética (ou sexa, individuos que non están diagnosticados como diabéticos pero sí que teñen unha alta probabilidade de selo ó longo

da súa vida), outro no que aparecería a poboación normoglucémica que, tras a inxesta viron afectado o seu nivel de glicosa; e os diabéticos ben controlados e, por último, a poboación normoglucémica que non ven afectados (ou sí, pero a un xeito moito máis leve) as súas curvas de glicosa tras a inxesta, tanto porque non se realizou dita acción ou por causas debidas ó metabolismo de cada individuo. Despois de realizar esta reflexión, volvemos a empregar o algoritmo de k-medias e conseguimos a Figura 5.18.

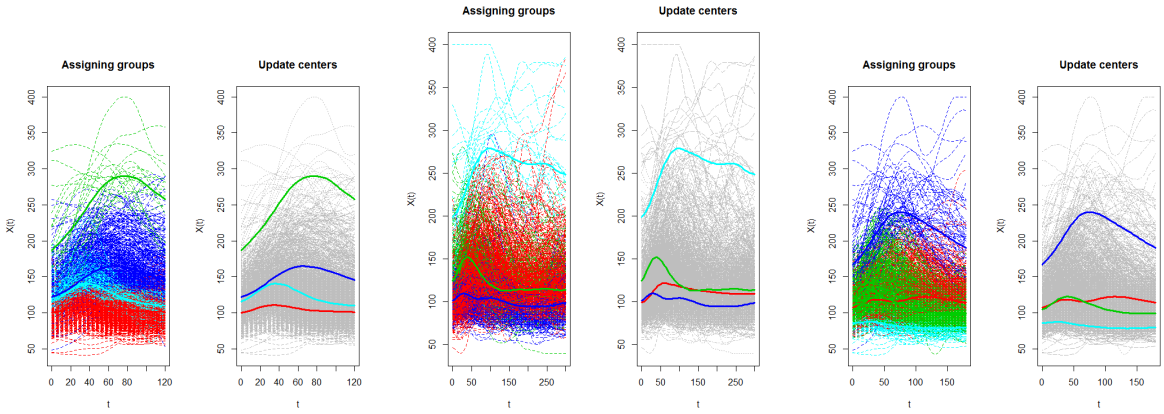

Figura 5.18: Resultado da clasificación non supervisada con 4 grupos para a mostra completa para almorzos, comidas e ceas.

En dita figura parece que se conseguiu a clasificación desexada, sobre todo nos almorzos. Vistosamente, nos outros dous momentos do día, esta clasificación non parece tan clara, posiblemente debido á gran variabilidade de inxestas que os individuos poidan realizar. No caso dos almorzos, os distintos tipos de alimentos que soen tomar esta poboación é máis reducida, xeralmente. Polo tanto, para ser máis concisos nesta clasificación, podemos confrontar os grupos obtidos coa variable indicadora *dm*, para ver se estamos realizando un razoamento correcto. Os datos obtidos móstranse no Cadro 5.3. Aquí é posible comprobar como é un razoamento probable, sobre todo nos grupos extremos xa que, por exemplo, tanto no grupo 2 para os nos almorzos, no grupo 4 nas comidas e no grupo 3 nas ceas; non se encontra ningún individuo non diabético.

		Non diabéticos	Diabéticos
Almorzos	Grupo 2	0	21
	Grupo 3	245	227
	Grupo 4	476	34
	Grupo 1	1693	44
Comidas	Grupo 4	0	23
	Grupo 2	196	55
	Grupo 1	1589	220
	Grupo 3	629	28
Ceas	Grupo 3	0	78
	Grupo 1	673	119
	Grupo 2	1627	126
	Grupo 4	120	3

Cadro 5.3: Clasificación non supervisada con 4 grupos para toda a mostra.

De igual xeito que se realizou esta análise, podemos considerar realizar a mesma clasificación non supervisada co algoritmo de k-medias, pero usando as curvas que máis puntúan nas compoñentes principais (calculadas anteriormente) como centroides. Isto ven a causa de que nelas atopábase os distintos patróns de comportamentos que motivou o anterior razoamento. Non obstante, isto non dou os resultados esperados, polo que se omiten tanto os pasos como os resultados.

5.5.2. Clasificación supervisada

Nesta subsección, rescataremos os datos outliers obtidos mediante o método de proxeccións aleatorias e, despois de calcular unha regra de decisión decidiremos se eran outliers por ter comportamento de diabético ou por ser atípico dentro dos non diabéticos.

Recordemos entón que obtivemos os outliers a través de separar as curvas segundo as características individuais e calcular os datos atípicos de cada un dos subconxuntos. Ditas curvas están representadas na Figura 5.16.

Entón, estableceremos regras de clasificación. Primeiro, calculamos un DD-plot da poboación total. A regra de clasificación e a representación dos datos xunto co seu grupo pódese ver na Figura 5.19. Isto faise a través da profundidade de proxeccións aleatorias mediante un modelo gam (Generalized additive model). Para ver a probabilidade de clasificación correcta véxase o Cadro 5.4. Ademais podemos realizar predición cos datos atípicos dos tres momentos do día. Igualmente, podemos ver os resultados de dita predición no cadro citado. Nótese que este método é o que menos probabilidade de clasificación correcta ten. Esta diminución é debida á mala clasificación do grupo de diabéticos. Loxicamente esta análise é realizada con distintos métodos como a estimación tipo kernel ou considerando outra profundidade;

	Modelos	Prob. clas. correcta			nº clasificados como...	
		0	1	Total	0	1
Almorzos	GLM funcional	0.99	0.60	0.9431	35	18
	GAM funcional	0.98	0.60	0.9445	34	19
	DD-plot	0.99	0.54	0.9315	30	23
Comidas	GLM funcional	0.99	0.48	0.9318	48	22
	GAM funcional	0.99	0.47	0.9326	47	23
	DD-plot	0.99	0.49	0.9273	49	21
Ceas	GLM funcional	0.99	0.61	0.9461	30	26
	GAM funcional	0.99	0.61	0.9465	30	26
	DD-plot	0.99	0.59	0.9428	30	26

Cadro 5.4: Resultados obtidos tras aplicar distintos métodos de clasificación supervisada ós datos atípicos que se chegou anteriormente.

non obstante isto non mellorou a clasificación significativamente.

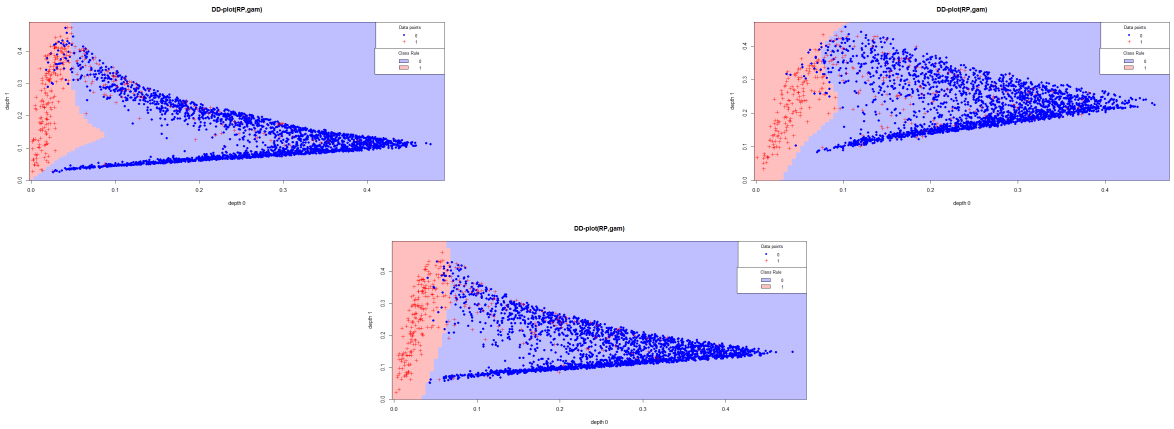

Figura 5.19: DD-plot conseguido coa clasificación supervisada coa variable *dm* para todos os momentos do día cun modelo *gam* e utilizando a profundidade RP.

Por último, cabe destacar que despois da realización de distintos métodos (explicados no anterior capítulos), cos que mellores resultados obtivemos fixemos un cadro resumo (Cadro 5.4) onde se mostra a probabilidade de clasificación correcta e o número de observacións que foron preditas en cada grupo.

	dm	*oh*4	*oh*3	*ipq*	*sex*	*tab*012	*sm*
Almorzos	0	0.075	0.005	0.005	0.005	0	0
Comidas	0	0	0.03	0	0	0	0
Ceas	0	0	0	0.02	0	0	0
Almorzos	0	0.141	0.06	0.001	0	0	0
Comidas	0	0.018	0	0	0.054	0.025	0
Ceas	0	0.034	0.027	0.003	0	0	0

Cadro 5.5: p-valores obtidos despois da aplicación dun test anova dun factor segundo as variables escritas na primeira columna tanto para a a mostra con todos os individuos (tres primeiras filas) como con só os non diabéticos (tres últimas filas).

5.6. Anova

Neste apartado terase en conta as variables como *dm* ou *oh*4, que son factores, pretendendo así ver se inflúe ou non nas curvas de glicosa.

Realizando un test anova dun factor, que se explicou no anterior capítulo do traballo, das curvas de glicosa fronte as demais variables factores que hai na nosa base de datos obtéñense os p-valores representados no Cadro 5.5. Nótese que aparece unha variable nova: *oh*3. Esta é unha modificación de *oh*4 onde se unen os dous últimos grupos. Ademais, nas tres primeiras filas móstranse os resultados coa poboación total mentres que nas tres últimas considéranse a poboación non diabética.

Para un nivel de confianza ó 95 %, a maioría dos test mostran diferenzas significativas (a non ser variables como *oh*4 para os almorzos ou para as comidas). Isto medicamente pode non ter sentido, sobre todo coa variable *sex*; pero débese ter en conta que son tests cun único factor. Seguramente que estas diferenzas están relacionadas con outras levándonos a un caso de problemas coa colinealidade.

Debido a isto, realizamos test de varios factores así como test considerando interacción entre as variables. Unha vez realizado estas análises, chegamos á conclusión de que:

- Para os almorzos, incluíndo a variable *dm* conseguimos que as variables *age* e *sex* deixen de ser significativas. Ademais, a iteración entre as variables *ipq* con *oh*3 e *ipq* con *tab*012 son significativas, ó igual que estas variables por separado. É dicir, chegamos a conclusión de que existen diferenzas considerando as variables *dm*, *tab*012, *oh*3, *ipq* e a iteración entre esta última e as dúas anteriores.

- Para as comidas, obtivemos resultados análogas pero desta vez tamén se debe considerar como significativa a diferenza usando a variable *age* pero ningunha das iteracións deben ser tratadas.

- En canto as ceas, o *ipq*, *oh*3 e *tab*012 saen significativas mentres que as iteracións de *oh*3 con *tab*012 non. As demais sí que se deben ter en conta.

Nestas análises descartouse a variable *sm* por producir un fenómeno de confusión, xa que para o seu cálculo precísase os valores de *dm*.

Véxase que agora variables que non tiñan sentido clinicamente falando xa se deixaron de considerar.

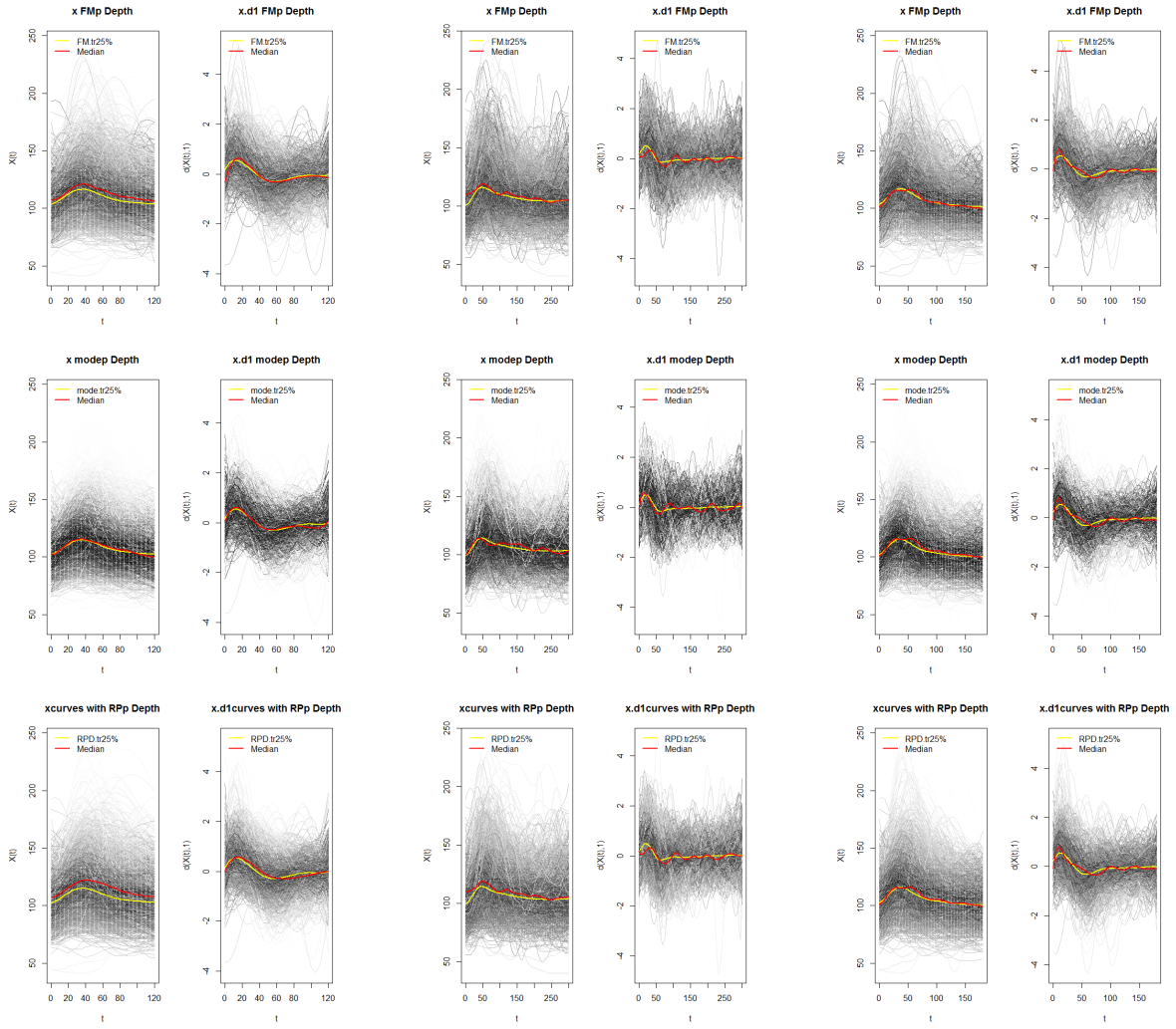

Figura 5.20: Profundidades das curvas e das derivadas da poboación non diabética en cada momento do día segundo as profundidades FM (primeira fila), modal (segunda fila) e RP (terceira fila).

Apéndice A

Folla a cubrir polo paciente

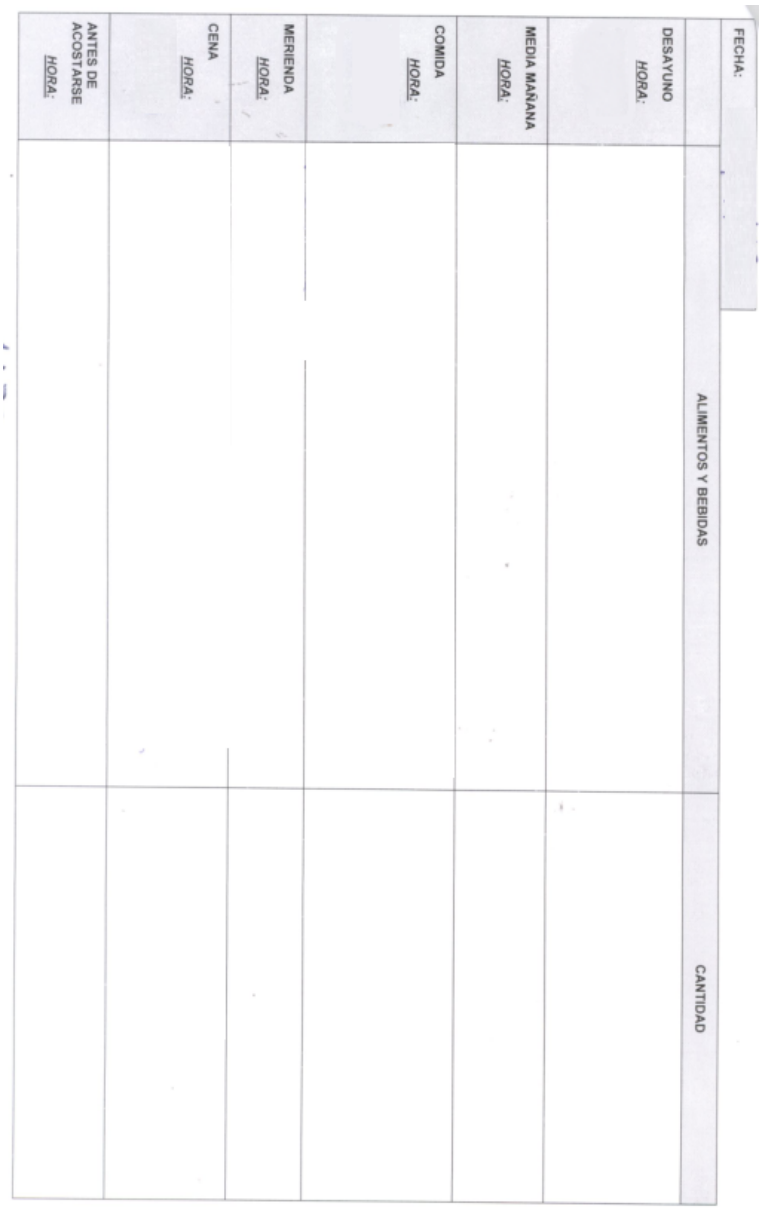

Figura A.1: Folla que debían de cubrir o paciente cada vez que realizaba unha comida ou exercicio físico.

Apéndice B

Función para ler a monitorización

```
# d  = 5 minutes, diferencia entre las tomas

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))
  dif123 = numeric()
  dif01=numeric()
  dif07=numeric()
  dif345=numeric()
  dif56=numeric()
  dif567=numeric()
  da = NULL
  library(chron)
  for(i in 1:(N-1)) {
cat(i, "\n")
y<-which(data$come==7)[i]
z<-which(data$come[data$Día==i]==0)
a<-which(data$come[data$Día==i]==1)
b<-which(data$come[data$Día==i]==2)[1]
c<-which(data$come[data$Día==i]==3)
d<-which(data$come[data$Día==i]==4)[1]
e<-which(data$come==5)[i]
e2<-which(data$come==7)[i+1]
f<-which(data$come==6)[which(which(data$come==6)>e & which(data$come==6)<e2)][1]
if(length(a)==1 & length(z)==1){
dif01[i]<-hours(times(as.character(data$Hora[a])))*60+
minutes(times(as.character(data$Hora[a])))-
hours(times(as.character(data$Hora[z])))*60-minutes(times(as.character(data$Hora[z])))
} else {dif01[i]<-NA}
if(length(y)==1 & length(z)==1){
if((hours(times(as.character(data$Hora[y])))*60+
minutes(times(as.character(data$Hora[y]))))>1140)
{dif07[i]<-1440-((hours(times(as.character(data$Hora[y])))*60+minutes(times(as.character(
data$Hora[y])))))+(hours(times(as.character(data$Hora[z])))*60+
```

```
minutes(times(as.character(data$Hora[z])))))}
else {dif07[i]<-hours(times(as.character(data$Hora[z])))*60+
minutes(times(as.character(data$Hora[z])))-hours(times(as.character(data$Hora[y])))*60-
minutes(times(as.character(data$Hora[y])))
}} else {dif07[i]<-NA}
if(is.na(b)!=TRUE){
dif123[i]<-hours(times(as.character(data$Hora[b])))*60+
minutes(times(as.character(data$Hora[b])))-hours(times(as.character(data$Hora[a])))*60-
minutes(times(as.character(data$Hora[a])))
} else {dif123[i]<-hours(times(as.character(data$Hora[c])))*60+
minutes(times(as.character(data$Hora[c])))-hours(times(as.character(data$Hora[a])))*60-
minutes(times(as.character(data$Hora[a])))}

if(is.na(d)==TRUE){
if((hours(times(as.character(data$Hora[e])))*60+minutes(times(as.character(
data$Hora[e])))))>900)
{dif345[i]<-((hours(times(as.character(data$Hora[e])))*60+
minutes(times(as.character(data$Hora[e])))))-(hours(times(as.character(data$Hora[c])))*60+
minutes(times(as.character(data$Hora[c])))))}
else {dif345[i]<-(hours(times(as.character(data$Hora[e])))*60+
minutes(times(as.character(data$Hora[e]))))+
(1440-hours(times(as.character(data$Hora[c])))*60-
minutes(times(as.character(data$Hora[c]))))
}} else {dif345[i]<-hours(times(as.character(data$Hora[d])))*60+
minutes(times(as.character(data$Hora[d])))-hours(times(as.character(data$Hora[c])))*60-
minutes(times(as.character(data$Hora[c])))}
if(length(e)==1 & is.na(f)!=TRUE &
 (hours(times(as.character(data$Hora[e])))*60+minutes(times(as.character(data$Hora[e])))))>
900 &(hours(times(as.character(data$Hora[f])))*60+
minutes(times(as.character(data$Hora[f])))))>900 ){
dif56[i]<-hours(times(as.character(data$Hora[f])))*60+
minutes(times(as.character(data$Hora[f])))-hours(times(as.character(data$Hora[e])))*60-
minutes(times(as.character(data$Hora[e])))
} else if(length(e)==1 & is.na(f)!=TRUE &
 (hours(times(as.character(data$Hora[e])))*60+minutes(times(as.character(data$Hora[e])))))>
900 &(hours(times(as.character(data$Hora[f])))*60+
minutes(times(as.character(data$Hora[f])))))<900 ){
dif56[i]<-1440-(hours(times(as.character(data$Hora[e])))*60+
minutes(times(as.character(data$Hora[e]))))+hours(times(as.character(data$Hora[f])))*60+
minutes(times(as.character(data$Hora[f])))
} else if(length(e)==1 & is.na(f)!=TRUE &
 (hours(times(as.character(data$Hora[e])))*60+minutes(times(as.character(data$Hora[e])))))<
900){dif56[i]<-hours(times(as.character(data$Hora[f])))*60+
minutes(times(as.character(data$Hora[f])))-hours(times(as.character(data$Hora[e])))*60-
minutes(times(as.character(data$Hora[e])))
} else {dif56[i]<-NA}

cc = n[i]:(n[i] + h)
data[cc, ]$Día[data[cc, ]$Día == i+1] = i #Sólo necesario para cenas
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")
datafinal = cbind(datafinal,dif07,dif01,dif123,dif345,dif56)
return(datafinal)
}

### HORAS: 3, 4 y 8 para desayuno, comida y cena, respectivamente.
# 3 horas despois do desayuno, 4 despois da comida e 8 despois da cena
### CÓDIGO: 1, 3 y 5 para desayuno, comida y cena, respectivamente.

lista2 = list.files(pattern = ".csv")
cena = NULL
for(k in 1:(length(lista2)-1)){
cena= rbind(cena,particion(read.csv2(lista2[k]), 5, 5))
cat(lista2[k],"\n")
}
```

Bibliografía

[1] Brownlee M. (2005) The pathobiology of diabetic complications: a unifying mechanism. Diabetes 54:1615-1625.

[2] Cadarso C.M. (2015). Apuntes de Estadística non Paramétrica.

[3] Concello da Estrada. A Estrada en datos. http://www.aestrada.com/index.php/gl/. Accedido o 3 de Xuño do 2016.

[4] Cuesta-Albertos J., Fraiman R. and Ransford T.(2007). A sharp form of the Cram´ er-Wold theorem. J. Theoret. Probab. 20,201?209.

[5] Cuesta-Albertos J.A. e Nieto-Reyes A. (2008). The Random Tukey Depth. Computational Statistics & Data Analysis.

[6] Cuevas A., Febrero M. e Fraiman R. (2007). Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics, 22(3):481-496.

[7] Cuevas A., Febrero M., e Fraiman R. (2004). An anova test for functional data. Computational statistics & data analysis, 47(1), 111-122

[8] de Boor C. (1978). A Practical Guide to Splines. Springer.

[9] Deza M. (2014). Encyclopedia of Distances. Springer.

[10] Eslami S., Taherzadeh Z, Schultz MJ, Abu-Hanna A. (2011) Glucose variability measures and their effect on mortality: a systematic review. Intensive Care Med 37:583-593.

[11] Febrero-Bande M. e González-Manteiga W. (2013). Generalized additive models for functional data. TEST, pages 1?15.

[12] Febrero Bande M., Oviedo de la Fuente M., Galeano P., Nieto A., Garcia-Portugues E. (2016) Functional Data Analysis and Utilities for Statistical Computing. R package version 1.2.3. https://cran.r-project.org/web/packages/fda.usc/fda.usc. Accedido o 24 de xuño do 2016

[13] Febrero M., Galeano P. e González-Manteiga W. (2007a). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. Environmetrics, 19(4):331?345.

[14] Fernandez F.J. (2012). Apuntes de Series de Fourier.

[15] Ferraty, F. and Vieu, P. (2006) NonParametric Functional Data Analysis Springer Series in Statistics http://www.math.univ-toulouse.fr/staph/npfda. Accedido o 20 de maio

[16] Gillies C.L., Abrams KR, Lambert PC (2007) Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. BMJ 334:299.

[17] Hyndman R.J., Sang H.L. (2015) Functional data sets. R package version 1.7. https://cran.r-project.org/web/packages/fds/fds. Accedido o 1 de xuño do 2016.

[18] Kilpatrick E.S., Rigby A.S., Atkin S.L. (2009) Effect of glucose variability on the long-term risk of microvascular complications in type 1 diabetes. Diabetes Care; 32: 1901?1903

[19] Preda C. e Saporta G. (2005). Pls regression on a stochastic process. Computational Statistics & Data Analysis, 48(1):149-158.

[20] Ramsay J.O e Silverman B.W. (2005). Functional Data Analysis. Springer.

[21] Rousseeuw P.J., Ruts I. e Tukey J.W. (2012). The Bagplot: A Bivariate Boxplot. The American Statistician.

[22] Service F.J., Molnar G.D., Rosevear J.W., Ackerman E., Gatewood L.C., Taylor W.F. (1970) Mean amplitude of glycemic excursions, a measure of diabetic instability. Diabetes 19:644-655.

[23] Shang H.L., Hyndman R.J. (2016) Rainbow Plots, Bagplots and Boxplots for Functional Data. R package version 3.4. https://cran.r-project.org/web/packages/rainbow/rainbow. Accedido o 25 de xuño do 2016.

[24] Wickham H., Ramsay J.O., Graves S. e Hooker G. (2015) Functional Data Analysis. R package version 2.4.4. https://cran.r-project.org/web/packages/fda/fda.pdf. Accedido o 29 de xuño do 2016.

[25] Whiting DR, Guariguata L,Weil C (2011) IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes Res Clin Pract 94:311-321.