Estimación Robusta, Métodos Semiparamétricos, Datos Funcionales y de Alta Dimensión

Ana Bianco[1], Graciela Boente[1]

[1]Universidad de Buenos Aires y CONICET, Argentina

Fuentes de financiación

- Universidad de Buenos Aires. 20020170100022BA. Período: 01/01/2018 a 31/12/2020.

Fuentes de financiación

- Universidad de Buenos Aires. 20020170100022BA. Período: 01/01/2018 a 31/12/2020.

- Agencia Nacional de Promoción Científica y Tecnológica. PICT 2014-0351. Período: 15/01/2016 a 14/07/2019.

Fuentes de financiación

- Universidad de Buenos Aires. 20020170100022BA. Período: 01/01/2018 a 31/12/2020.

- Agencia Nacional de Promoción Científica y Tecnológica. PICT 2014-0351. Período: 15/01/2016 a 14/07/2019.

Algunas temáticas del proyecto están financiadas por

- Ministerio de Gobierno y Competitividad y FEDER (Fondos Europeos de Desarrollo Regional), España. INNPAR2D MTM2016-76969P. Período: 01/01/2017 a 31/12/2020.

  Dirigido por Wenceslao González Manteiga y Rosa María Crujeiras Casais.

Integrantes UBACYT 20020170100022BA

- **Lucas Bali**: Investigador CONICET & Instituto de Investigaciones Científicas y Técnicas para la Defensa, Ministerio de Defensa. JTP (DM)
- **Gonzalo Chebi***: Becario CONICET. JTP (DM)
- **Ricardo Maronna**: Investigador.
- **Alejandra Martínez**: Profesor Adjunto Universidad Nacional de Luján. JTP (DM)
- **Florencia Statti***: Becario CONICET.
- **Pablo Vena***: Becario CONICET. JTP (DM)

Integrantes UBACYT 20020170100022BA

- **Jemina García***: Estudiante de Maestría. Ayudante de Primera (IC).
- **Carlos Pita**: Estudiante de Maestría.

Integrantes UBACYT 20020170100022BA



Lucas



Gonzalo



Ricardo



Alejandra



Florencia



Pablo



Jemina



Carlos

Otros proyectos

- **Agustín Alvarez**: Universidad Nacional de General Sarmiento.
- **Stella Maris Donato**: Universidad Nacional de Cuyo.
- **Lucas Fernández Piana**: Universidad de San Andrés.
- **Nadia Kudraszow**: CONICET & Universidad Nacional La Plata.
- **Marcela Svarc**: CONICET & Universidad de San Andrés.

Colaboradores del exterior

- **Claudio Agostinelli**: Universitá di Trento.
- **Conceição Amado**: Universidad de Lisboa.
- **Wenceslao González Manteiga**: Universidad de Santiago de Compostela.
- **Juan Carlos Pardo–Fernández**: Universidad de Vigo.
- **Ana Pérez–González**: Universidad de Vigo.
- **Isabel Maria Alves Rodrigues**: Universidad de Lisboa.
- **Matias Salibian Barrera**: University of British Columbia.
- **Jane Ling Wang**: University of California at Davis.

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

- Desde el punto de vista estadístico, esto representa un nuevo desafío en varios sentidos.

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

- Desde el punto de vista estadístico, esto representa un nuevo desafío en varios sentidos. Es factible que
  - al tratar con tantas variables sea necesario considerar asociaciones ya sean conocidas salvo un número de parámetros o completamente desconocidas;

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

- Desde el punto de vista estadístico, esto representa un nuevo desafío en varios sentidos. Es factible que
  - al tratar con tantas variables sea necesario considerar asociaciones ya sean conocidas salvo un número de parámetros o completamente desconocidas;
  - que las observaciones correspondan a curvas suaves;

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

- Desde el punto de vista estadístico, esto representa un nuevo desafío en varios sentidos. Es factible que
  - al tratar con tantas variables sea necesario considerar asociaciones ya sean conocidas salvo un número de parámetros o completamente desconocidas;
  - que las observaciones correspondan a curvas suaves;
  - que haya más variables medidas que unidades experimentales o

Líneas de trabajo

- La evolución de los recursos tecnológicos ha aumentado la posibilidad de medir y almacenar variables en cada individuo, en muchos casos se registran muchas variables al mismo tiempo, permitiendo disponer de grandes masas de datos.

- Desde el punto de vista estadístico, esto representa un nuevo desafío en varios sentidos. Es factible que
  - al tratar con tantas variables sea necesario considerar asociaciones ya sean conocidas salvo un número de parámetros o completamente desconocidas;
  - que las observaciones correspondan a curvas suaves;
  - que haya más variables medidas que unidades experimentales o
  - que en algunos individuos haya mediciones faltantes para algunas de las variables de interés.

Líneas de trabajo

Para tratar estos problemas se han implementado distintas estrategias, tales como

- el uso de métodos de regresión lineal o no lineal

Líneas de trabajo

Para tratar estos problemas se han implementado distintas estrategias, tales como

- el uso de métodos de regresión lineal o no lineal
- modelos semi–paramétricos

Líneas de trabajo

Para tratar estos problemas se han implementado distintas estrategias, tales como

- el uso de métodos de regresión lineal o no lineal
- modelos semi–paramétricos
- el análisis de datos funcionales o

Líneas de trabajo

Para tratar estos problemas se han implementado distintas estrategias, tales como

- el uso de métodos de regresión lineal o no lineal
- modelos semi–paramétricos
- el análisis de datos funcionales o
- el desarrollo de técnicas para el tratamiento de datos de alta dimensión y/o datos faltantes, respectivamente.

Líneas de trabajo

Para tratar estos problemas se han implementado distintas estrategias, tales como

- el uso de métodos de regresión lineal o no lineal
- modelos semi–paramétricos
- el análisis de datos funcionales o
- el desarrollo de técnicas para el tratamiento de datos de alta dimensión y/o datos faltantes, respectivamente.

Es muy frecuente en cada una de estas situaciones, que un grupo de observaciones tenga un comportamiento desviado respecto del grueso de los datos. En este sentido, es imperativo el desarrollo de procedimientos robustos que sean a la vez resistentes a observaciones atípicas, eficientes bajo el modelo y computables.

Efecto de datos atípicos



Líneas de trabajo

En este sentido,

- Abordamos la búsqueda de procedimientos robustos para diversos modelos estadísticos paramétricos y semiparamétricos que incluyen el modelo lineal generalizado y parcialmente lineal.

- Se estudian también procedimientos robustos para datos funcionales, en particular, para componentes principales, correlación canónica y regresión semi–lineal.

- Por otra parte, como se ha mencionado, en muchos problemas se disponen de datos de alta dimensión, es decir, datos en los que la dimensión supera la cantidad de observaciones. Por esta razón, se estudian estimadores que sean adecuados cuando la dimensión de las covariables es baja respecto del tamaño muestral.

Líneas de trabajo

Más precisamente, entre nuestras líneas de trabajo podemos mencionar

1. En modelos paramétricos:
   (a) procedimientos robustos para comparar los parámetros de regresión cuando se tienen dos o más poblaciones, en el contexto de modelos de regresión generalizados. (A.B., G.B. & Isabel Rodrigues)

Líneas de trabajo

Más precisamente, entre nuestras líneas de trabajo podemos mencionar

1. En modelos paramétricos:
   (a) procedimientos robustos para comparar los parámetros de regresión cuando se tienen dos o más poblaciones, en el contexto de modelos de regresión generalizados. (A.B., G.B. & Isabel Rodrigues)

   (b) estimadores robustos en modelos de regresión no lineales heteroscedásticos. (Conceição Amado, A.B., G.B. & Isabel Rodrigues)

Líneas de trabajo

Más precisamente, entre nuestras líneas de trabajo podemos mencionar

1. En modelos paramétricos:
   (a) procedimientos robustos para comparar los parámetros de regresión cuando se tienen dos o más poblaciones, en el contexto de modelos de regresión generalizados. (A.B., G.B. & Isabel Rodrigues)

   (b) estimadores robustos en modelos de regresión no lineales heteroscedásticos. (Conceição Amado, A.B., G.B. & Isabel Rodrigues)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (a) test para estudiar la igualdad de funciones de varianza versus alternativas unilaterales en modelos de regresión noparamétrica heteroscedasticos. (G.B. & Juan Carlos Pardo–Fernández)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (a) test para estudiar la igualdad de funciones de varianza versus alternativas unilaterales en modelos de regresión noparamétrica heteroscedasticos. (G.B. & Juan Carlos Pardo–Fernández)

   (b) estimadores robustos en modelos generalizados parcialmente lineales isotónicos. (G.B., Daniela Rodriguez & Pablo Vena)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (a) test para estudiar la igualdad de funciones de varianza versus alternativas unilaterales en modelos de regresión noparamétrica heteroscedasticos. (G.B. & Juan Carlos Pardo–Fernández)

   (b) estimadores robustos en modelos generalizados parcialmente lineales isotónicos. (G.B., Daniela Rodriguez & Pablo Vena)

   (c) estimadores robustos para un parámetro marginal en modelos de regresión general donde las respuestas y algunas de las covariables pueden estar ausentes. (A.B., G.B., Ana Pérez–González & Wenceslao González Manteiga)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (a) test para estudiar la igualdad de funciones de varianza versus alternativas unilaterales en modelos de regresión noparamétrica heteroscedasticos. (G.B. & Juan Carlos Pardo–Fernández)

   (b) estimadores robustos en modelos generalizados parcialmente lineales isotónicos. (G.B., Daniela Rodriguez & Pablo Vena)

   (c) estimadores robustos para un parámetro marginal en modelos de regresión general donde las respuestas y algunas de las covariables pueden estar ausentes. (A.B., G.B., Ana Pérez–González & Wenceslao González Manteiga)

   (d) estimadores robustos en modelos parcialmente lineales aditivos. (G.B. & Alejandra Martínez)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (e) estimadores robustos de las curvas principales. (G.B. & Jemina García)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (e) estimadores robustos de las curvas principales. (G.B. & Jemina García)

   (f) estimadores robustos en modelos de índice simple cuando los errores son asimétricos. (Claudio Agostinelli, A.B. & G.B.)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (e) estimadores robustos de las curvas principales. (G.B. & Jemina García)

   (f) estimadores robustos en modelos de índice simple cuando los errores son asimétricos. (Claudio Agostinelli, A.B. & G.B.)

   (g) estimadores robustos en modelos parcialmente lineales de índice simple con parámetro de escala desconocido. (A.B. & Florencia Statti)

Líneas de trabajo

2. En el ámbito de modelos no–paramétricos y semi–paramétricos

   (e) estimadores robustos de las curvas principales. (G.B. & Jemina García)

   (f) estimadores robustos en modelos de índice simple cuando los errores son asimétricos. (Claudio Agostinelli, A.B. & G.B.)

   (g) estimadores robustos en modelos parcialmente lineales de índice simple con parámetro de escala desconocido. (A.B. & Florencia Statti)

   (h) estimadores robustos para el modelo aditivo esparso. (G.B. & Carlos Pita)

Líneas de trabajo

3. En el ámbito de datos de alta dimensión

   (a) estimadores robustos para los parámetros del modelo logístico que resulte consistente en el caso esparso y cuando p crece con n. (A.B., G.B. & Gonzalo Chebi)

Líneas de trabajo

4. En el ámbito de datos funcionales:

   (a) estimadores robustos en modelos parcialmente lineales funcionales y en modelos parcialmente lineales isotónicos funcionales. (G.B., Matías Salibián Barrera & Pablo Vena)

Líneas de trabajo

4. En el ámbito de datos funcionales:

   (a) estimadores robustos en modelos parcialmente lineales funcionales y en modelos parcialmente lineales isotónicos funcionales. (G.B., Matías Salibián Barrera & Pablo Vena)

   (b) estimadores del operador de dispersión basados en signos y test para proporcionalidad de operadores de dispersión entre varias poblaciones. (G.B., Daniela Rodriguez & Mariela Sued)

Líneas de trabajo

4. En el ámbito de datos funcionales:

   (a) estimadores robustos en modelos parcialmente lineales funcionales y en modelos parcialmente lineales isotónicos funcionales. (G.B., Matías Salibián Barrera & Pablo Vena)

   (b) estimadores del operador de dispersión basados en signos y test para proporcionalidad de operadores de dispersión entre varias poblaciones. (G.B., Daniela Rodriguez & Mariela Sued)

   (c) estimadores robustos para las componentes principales con datos ralos. (G.B. & Matías Salibián Barrera/ Ricardo Maronna)

Líneas de trabajo

4. En el ámbito de datos funcionales:

   (a) estimadores robustos en modelos parcialmente lineales funcionales y en modelos parcialmente lineales isotónicos funcionales. (G.B., Matías Salibián Barrera & Pablo Vena)

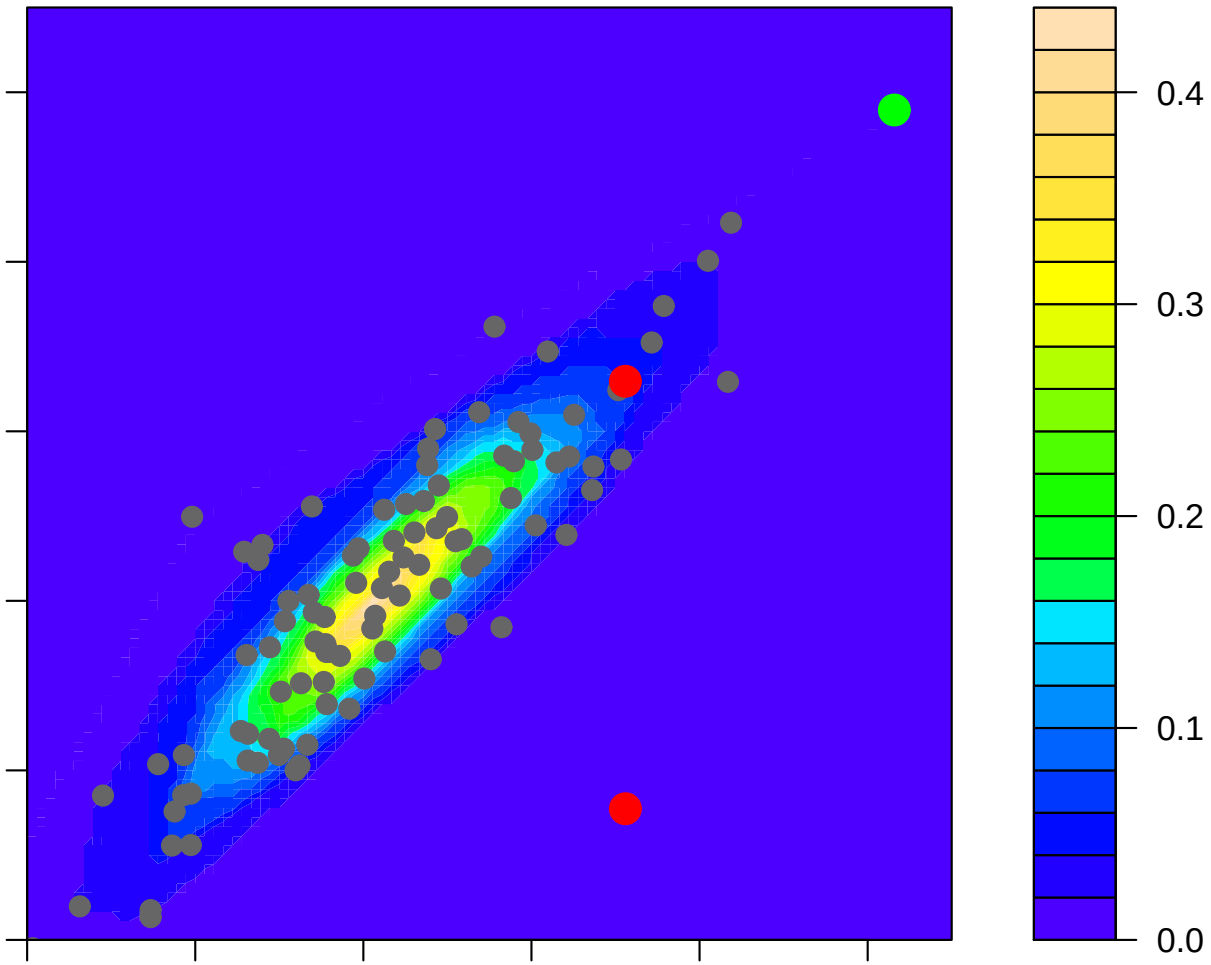
   (b) estimadores del operador de dispersión basados en signos y test para proporcionalidad de operadores de dispersión entre varias poblaciones. (G.B., Daniela Rodriguez & Mariela Sued)

   (c) estimadores robustos para las componentes principales con datos ralos. (G.B. & Matías Salibián Barrera/ Ricardo Maronna)

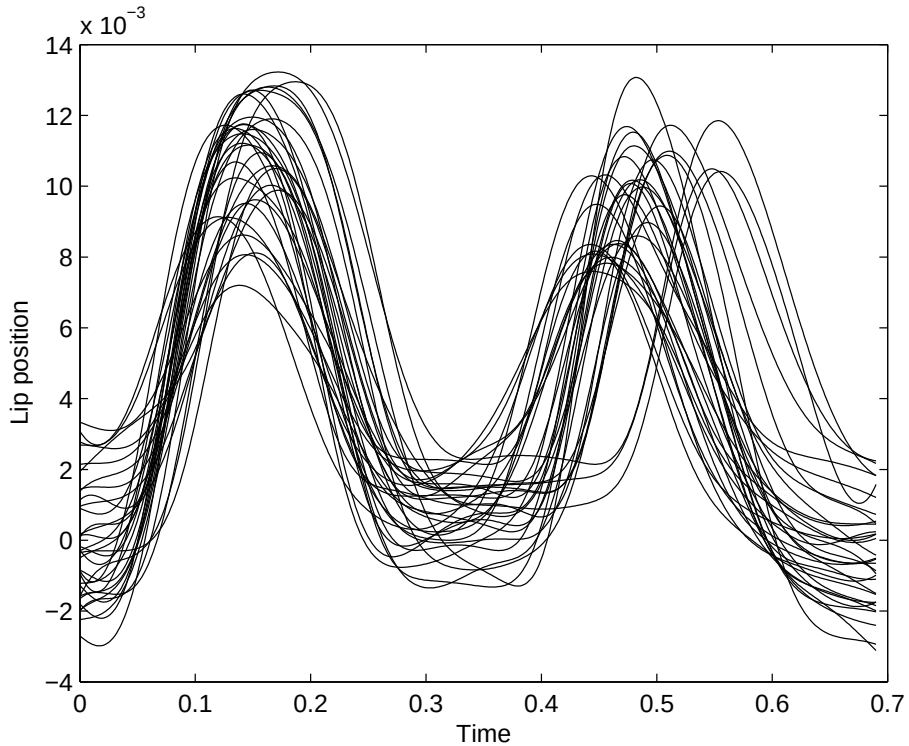
   (d) estimadores robustos de las correlaciones canónicas funcionales. (Agustín Alvarez, G.B. & Nadia Kudraszow)

Bienvenidos a los nuevos integrantes!!!

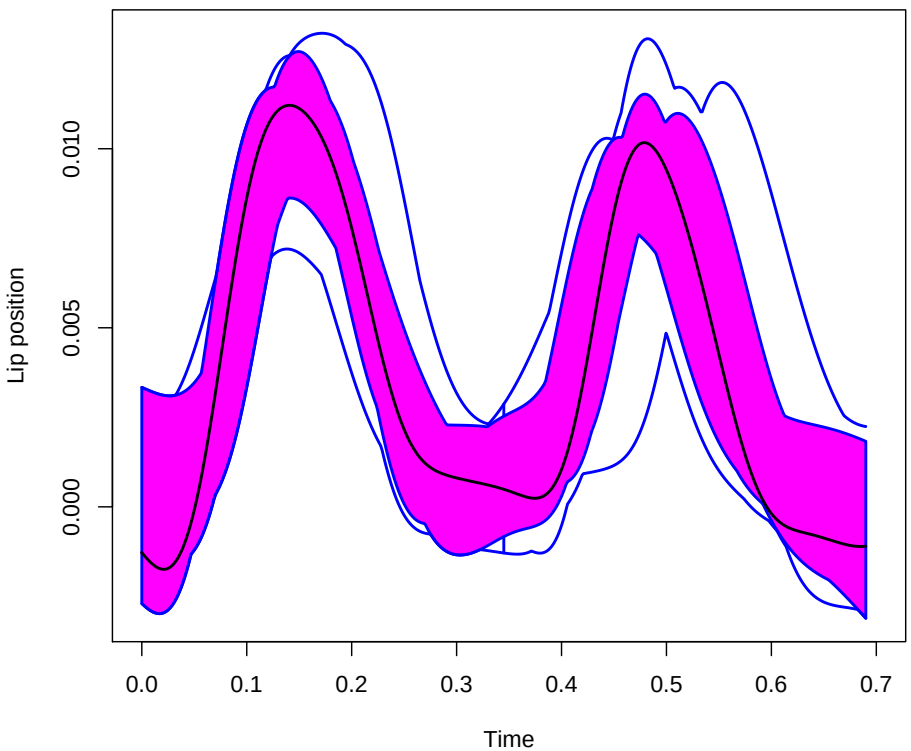
Un ejemplo de datos atípicos



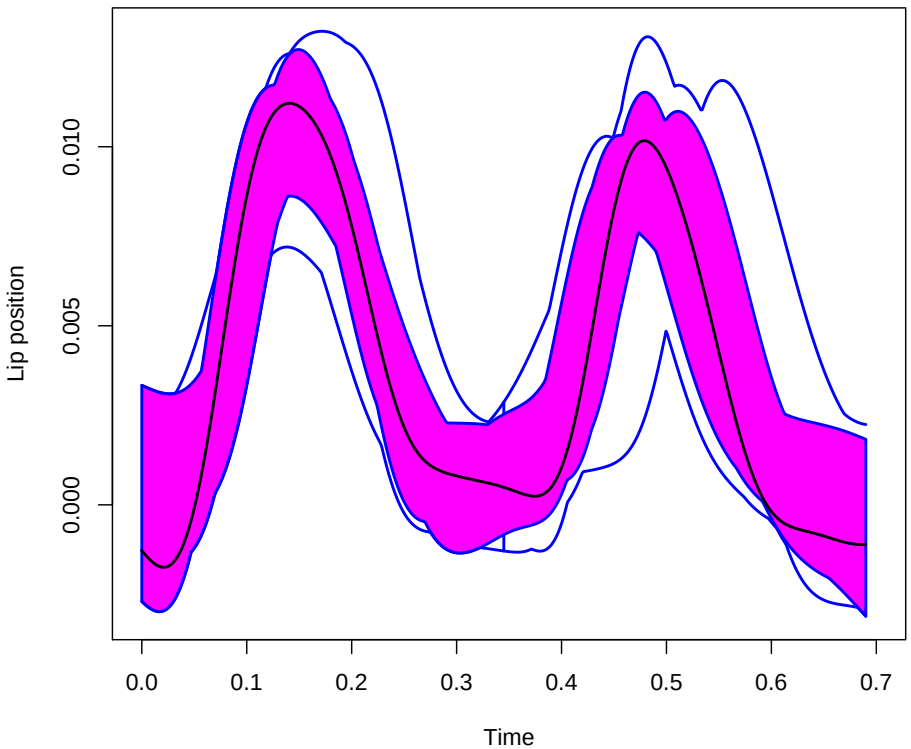
Un ejemplo en datos funcionales

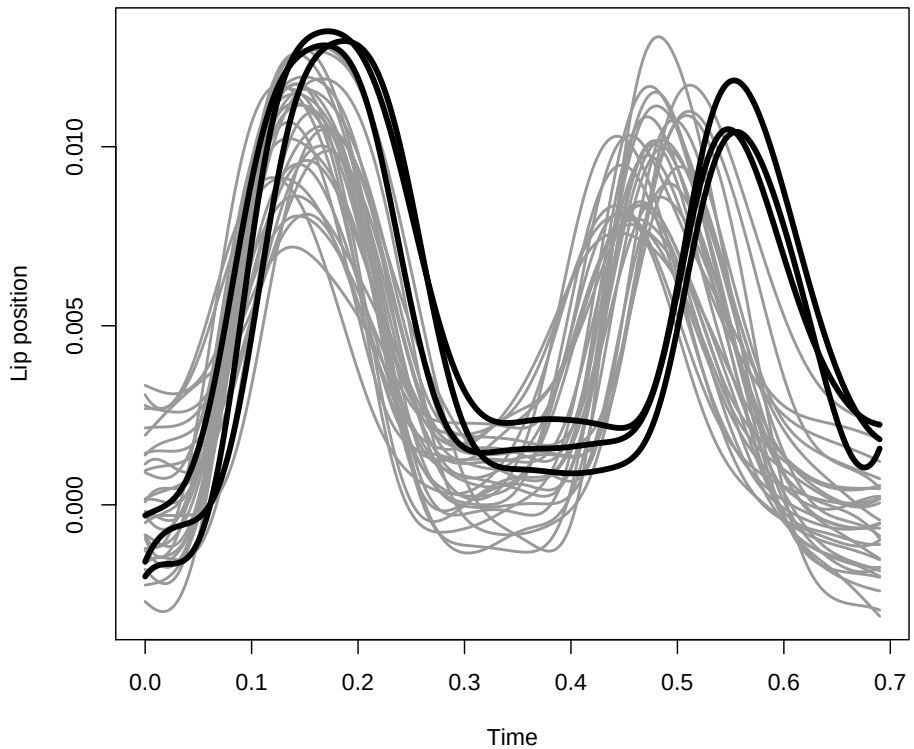


Un ejemplo en datos funcionales



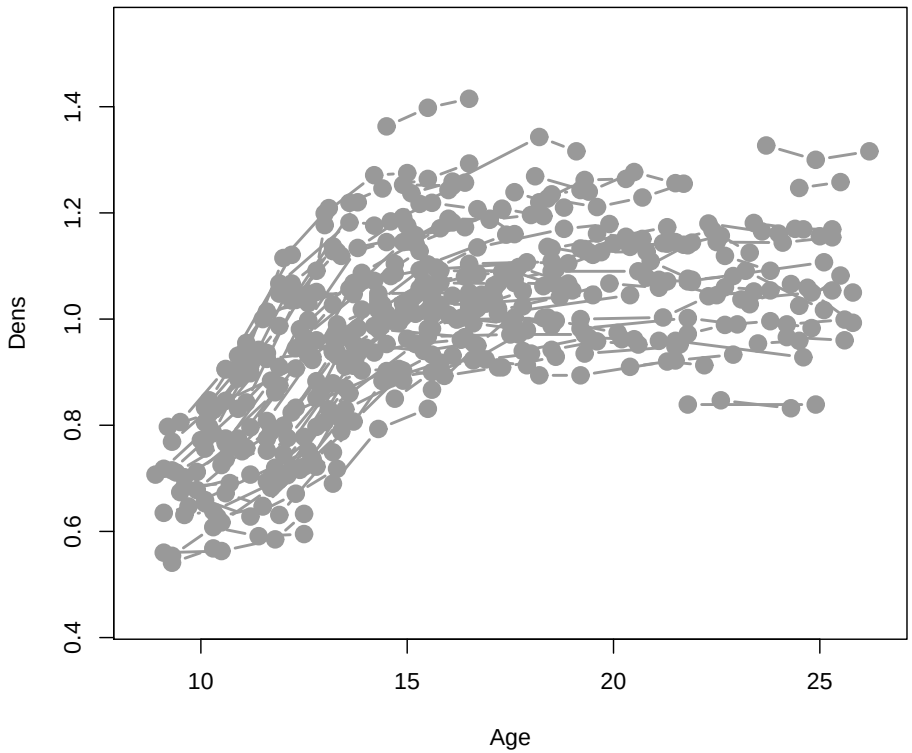
Un ejemplo en datos funcionales



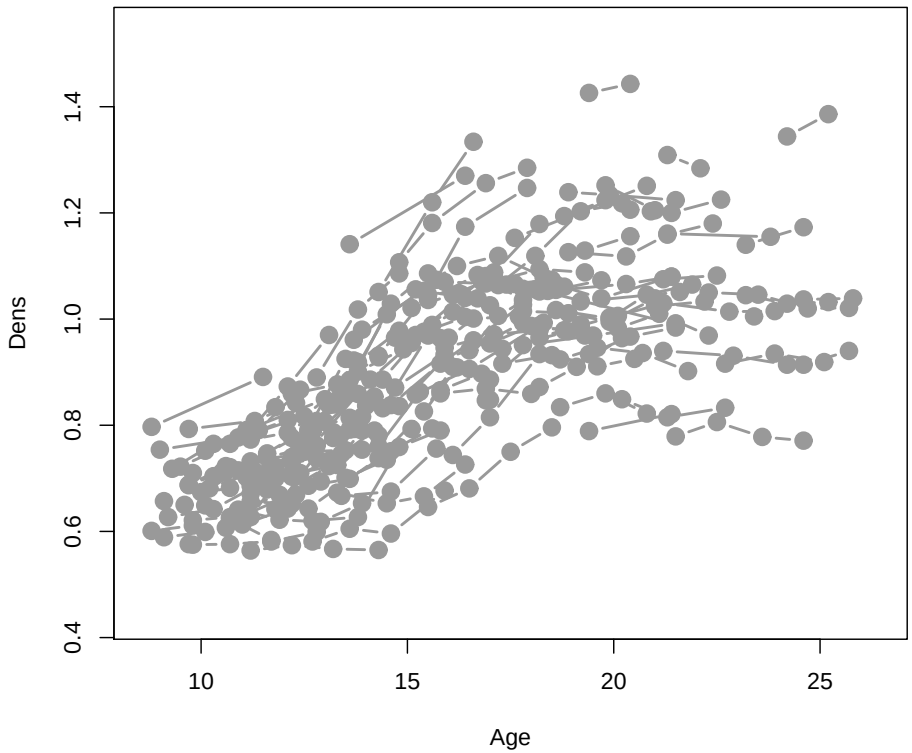


Datos ralos- Densidad de huesos

Mujeres ($n = 153$)



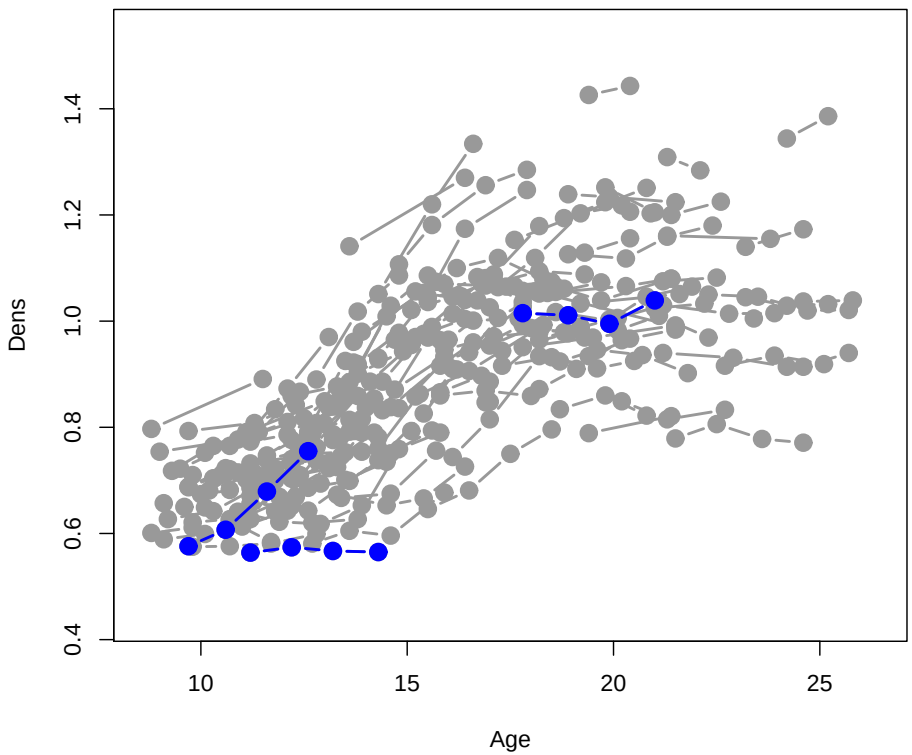
Hombres ($n = 127$)



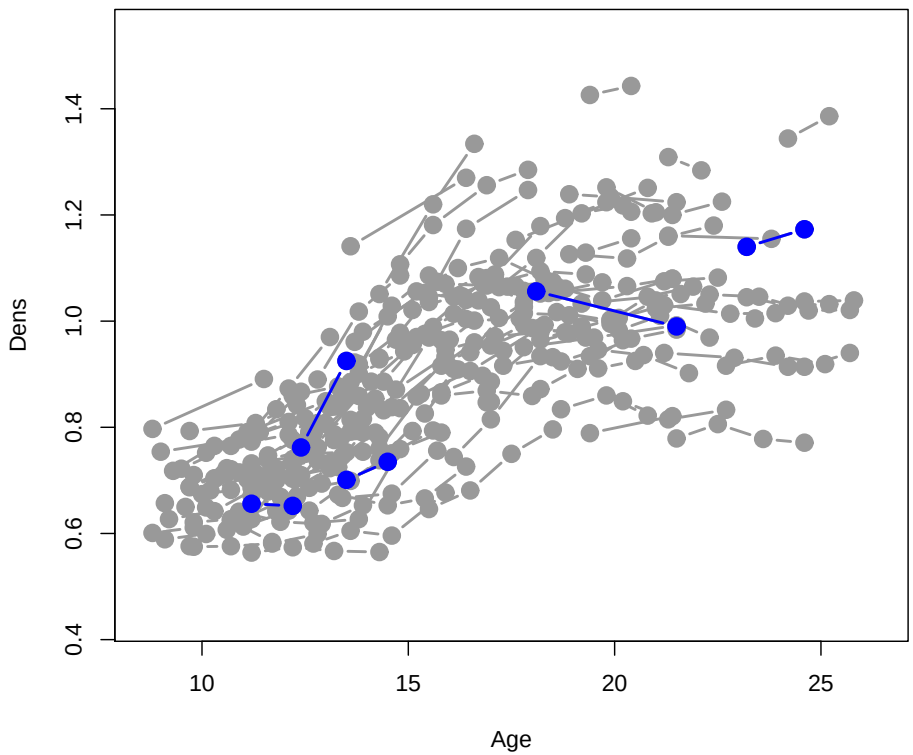
Densidad osea de la espina dorsal (g/cm^2) tomada en varias edades. En forma agregada, hay 860 observaciones medidas sobre un período de casi dos decadas

Datos ralos- Densidad de huesos- Hombres

Trayectorias largas



Trayectorias cortas



Un ejemplo en regresión logística

Observamos

$$(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n) \quad i.i.d.$$

- Y_i: $\{0, 1\}$
- $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$, $i = 1, ..., n$:

Un ejemplo en regresión logística

Observamos

$$(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n) \quad i.i.d.$$

- Y_i: $\{0, 1\}$
- $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$, $i = 1, ..., n$:

- Regresión logística:

$$\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}) = \pi(\mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0)$$

where

$$\pi(t) = \frac{1}{1 + e^{-t}}$$

- Objetivo: estimar $\boldsymbol{\beta}_0$.

Wisconsin Diagnostic Breast Cancer Dataset (WDBC)

Base de datos: 569 datos, 357 Tumores benignos (63%) y 212 Malignos (37%).

Covariables: 30 características medidas sobre una imágen tomada de una aspiración de masa mamaria. Describen características del núcleo de la célula relacionadas con

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ($perimeter^2/area - 1.0$)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension

Se calcularon estimadores con los datos originales y con datos contaminados artificialmente agregando

$n_{out} = 10, 20, 40, 60, 80$ observaciones mal clasificadas.

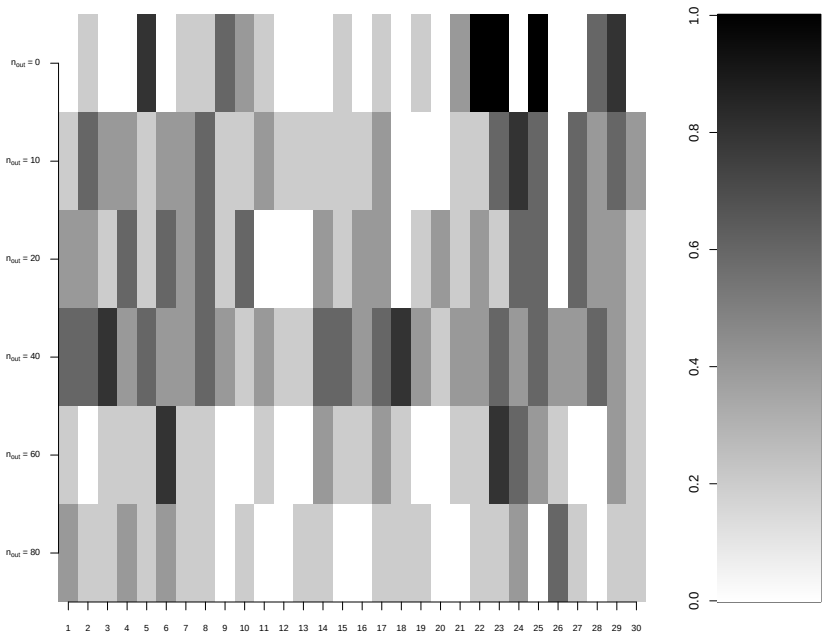
En cada caso, identificamos los coeficientes distintos de 0.

- Dividimos el conjunto en 10 grupos.
- Separamos uno de ellos y estimamos con los demás.
- Repetimos con cada grupo

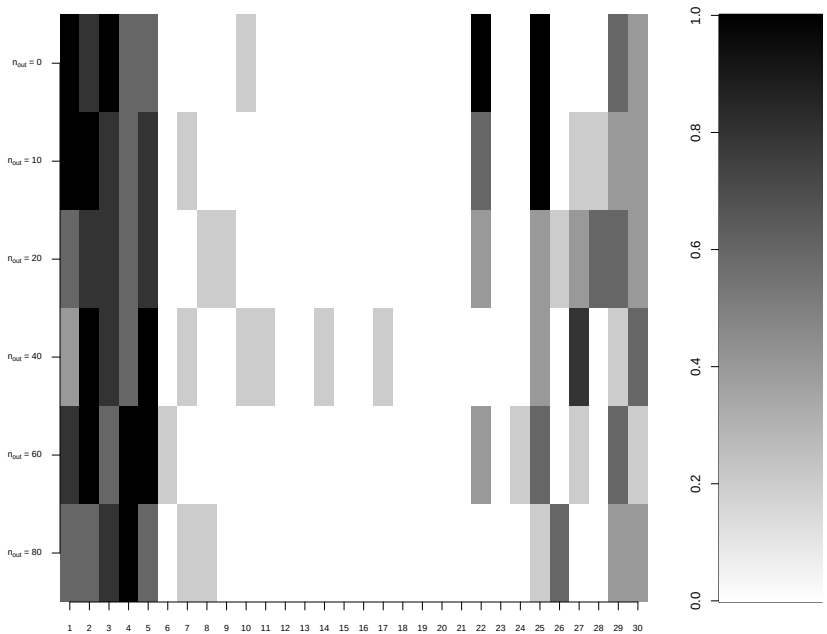
Contamos cuantas veces cada procedimiento elige cada variable, i.e., cuantas veces $\widehat{\beta}_j \neq 0$.

Porcentaje de veces que $\widehat{\beta}$ elige cada variable

Clásico



Robusto



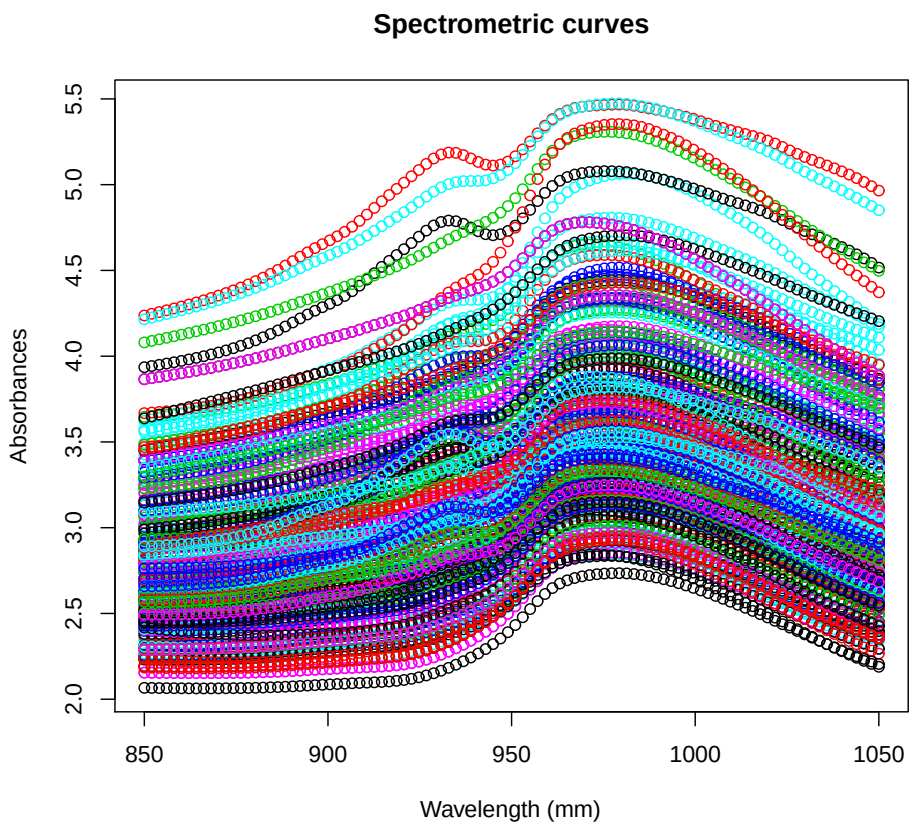
Datos espectrometricos

Los datos corresponden a un estudio de control de calidad en la industria de comestibles en la que se estudiaron 215 piezas de carne finamente triturada

Para cada unidad i, se observan:

- el contenido de grasa de la muestra medida por procedimientos analíticos (medida en porcentaje)
- el contenido de humedad y de proteínas de la muestra
- una curva espectrométrica $(\mathcal{X}_i)$ que corresponde a la absorbancia medida en 100 longitudes de onda.

Datos espectrometricos



Spectrometric data

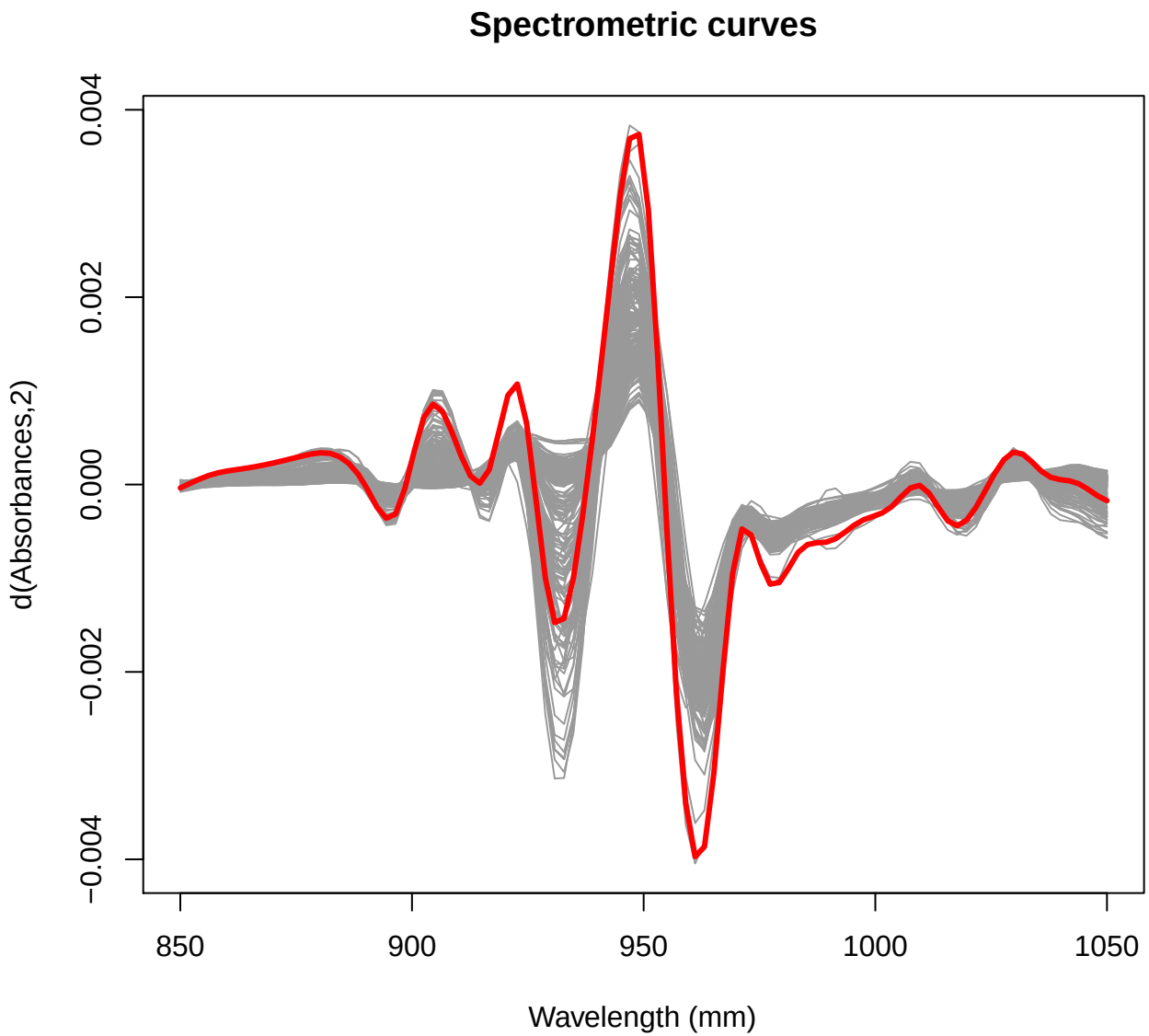
Como es más económico obtener la curva espectrométrica que el procedimiento analítico que determina el contenido de grasa, puede ser de interés considerar un modelo que premita predecir el contenido de grasa a partir del contenido de proteínas, humedad y del espectro de absorbancia.

Interesa modelar

$y =$ % de grasa y

- contenido de humedad z_1,
- porcentaje de proteínas, z_2
- alguna característica de la curva $\mathcal{X}$.

Datos espectrométricos: Segunda derivada del espectro

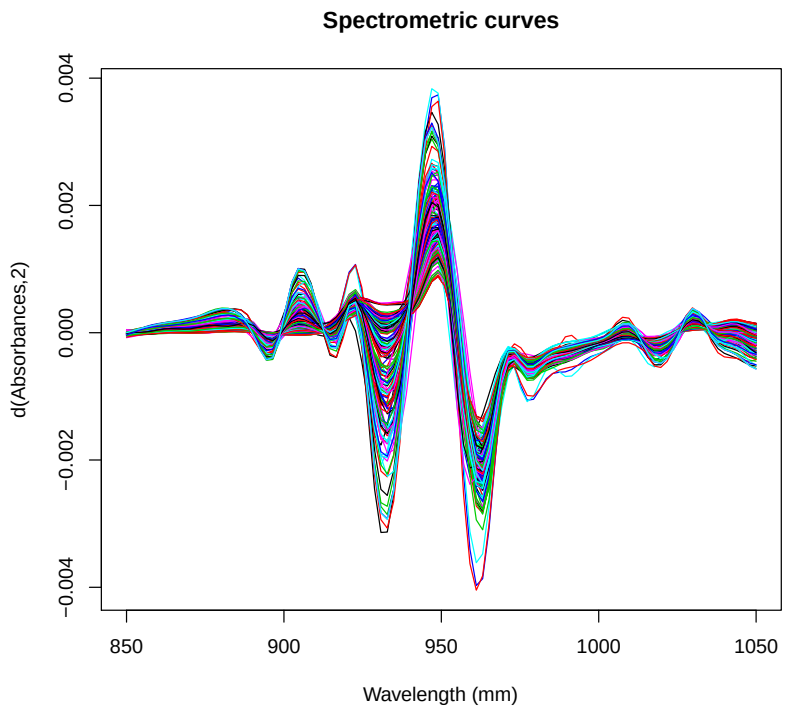


Datos espectrométricos, y = contenido de grasa

We considered the following model

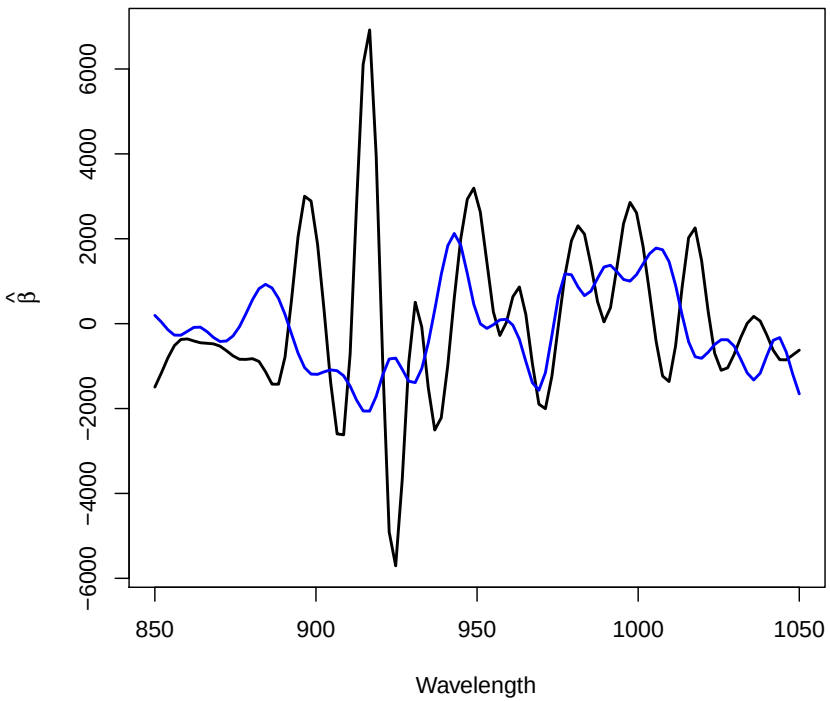
$$y = \langle X, \beta \rangle + z_1\, \eta(z_2) + \epsilon$$

- z_1 = humedad
- z_2 = proteínas
- X = derivada segunda de la curva del espectro $\mathcal{X}$

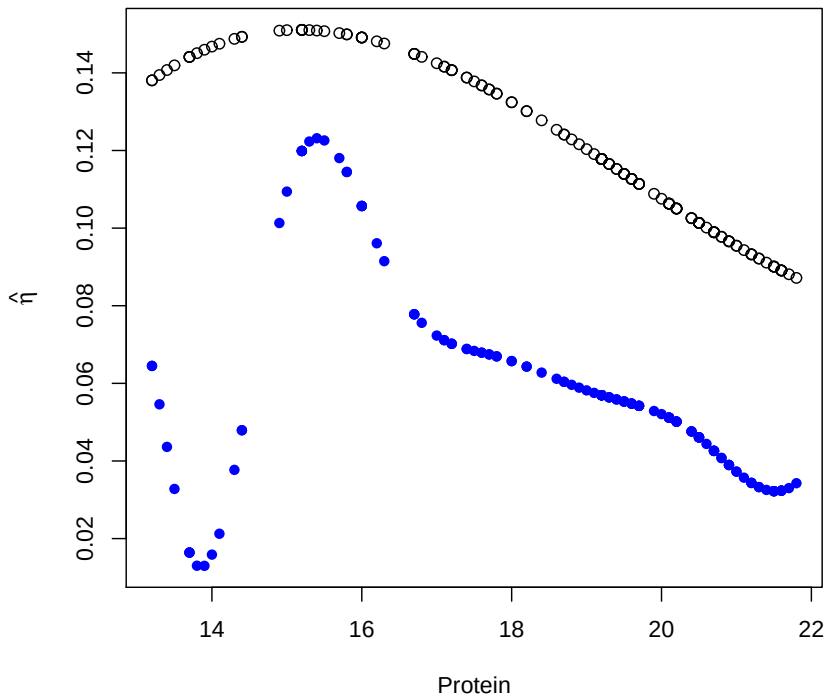


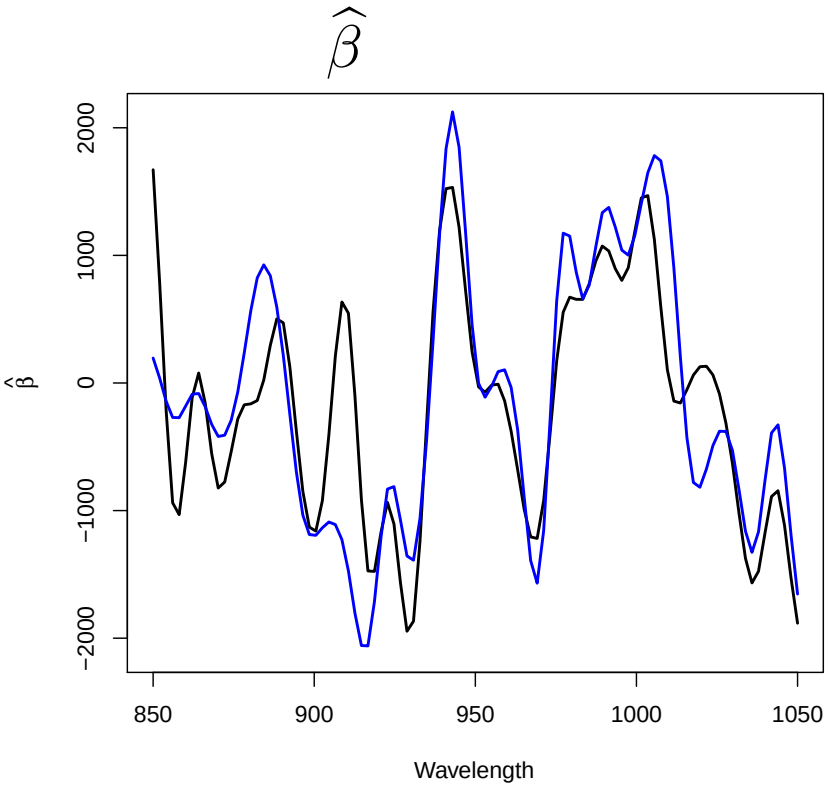
Datos espectrométricos: Clásico LS, Robusto

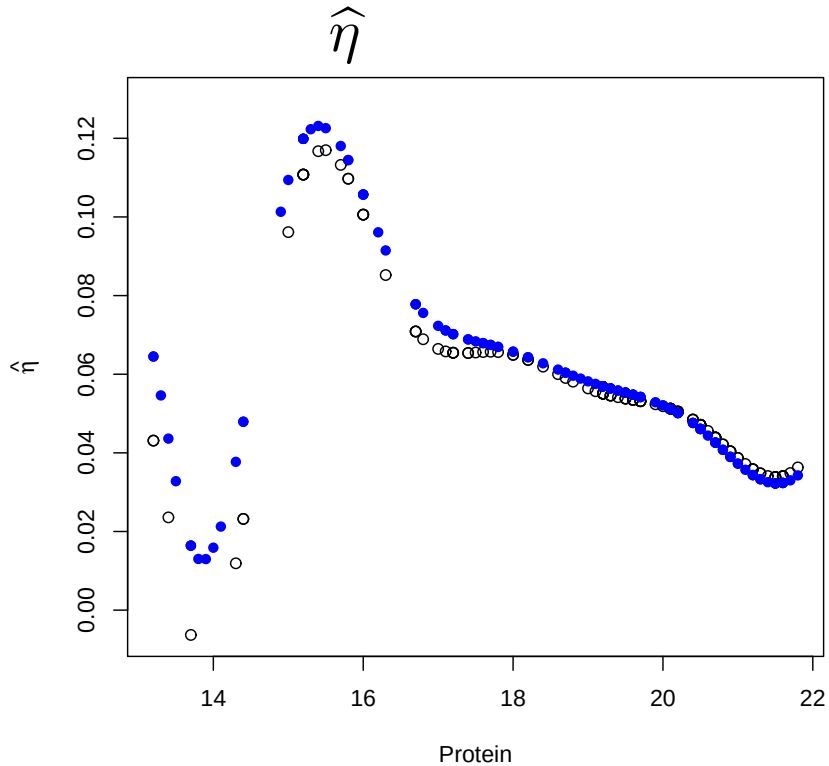
$\widehat{\beta}$



$\widehat{\eta}$




$\widehat{\beta}$
$\hat{\beta}$
2000
1000
0
−1000
−2000
850
900
950
1000
1050
Wavelength


$\widehat{\eta}$
$\hat{\eta}$
0.12
0.10
0.08
0.06
0.04
0.02
0.00
14
16
18
20
22
Protein

Datos de vino verde: 1599 Tintos, 4898 Blancos

Las variables medidas incluyen

- acidez fija/ acidez volatil/ acidez cítrica
- azúcar residual
- chlorides,
- Dióxido sulfúrico libre / Dióxido sulfúrico total
- density,
- pH,
- sulfatos
- alcohol

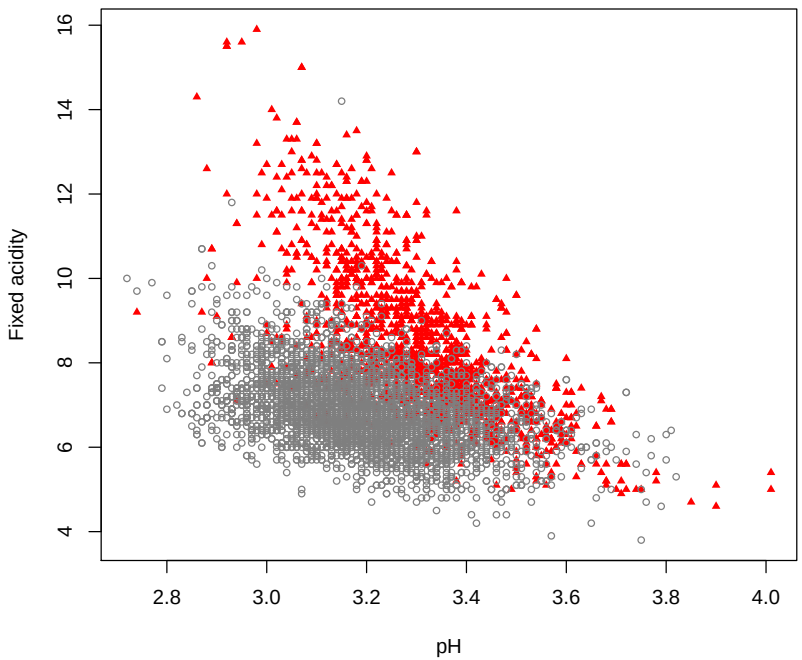
Nos interesa la relación entre pH y la acidez fija.

$$Y_j = m_j(X_j) + \sigma_j(X_j)\epsilon_j\,, \quad j = \text{Blanco}, \text{Tinto}$$

Datos de vino verde

Acidez fija versus pH, ▲= vinos tintos, ●= Vinos blancos

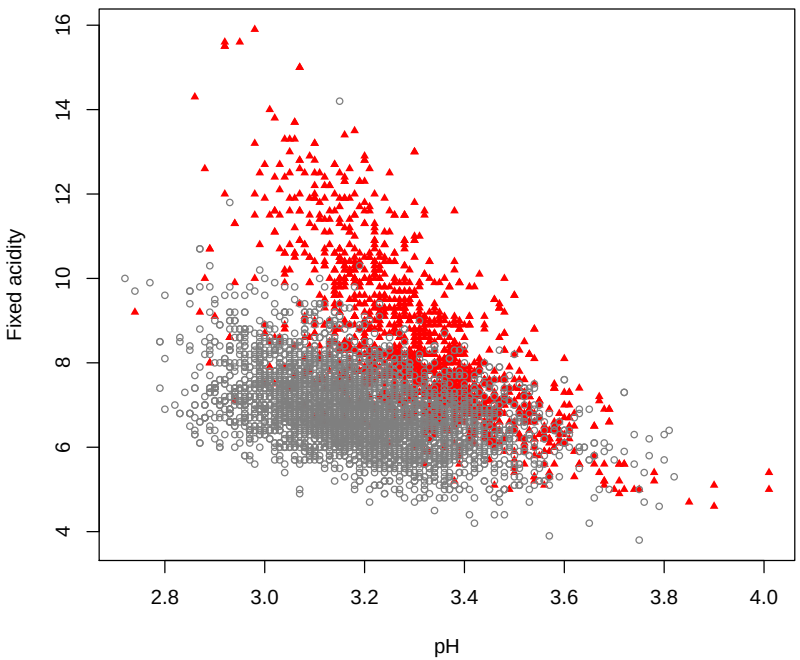
Datos completos Submuestras $n_1 = n_2 = 200$



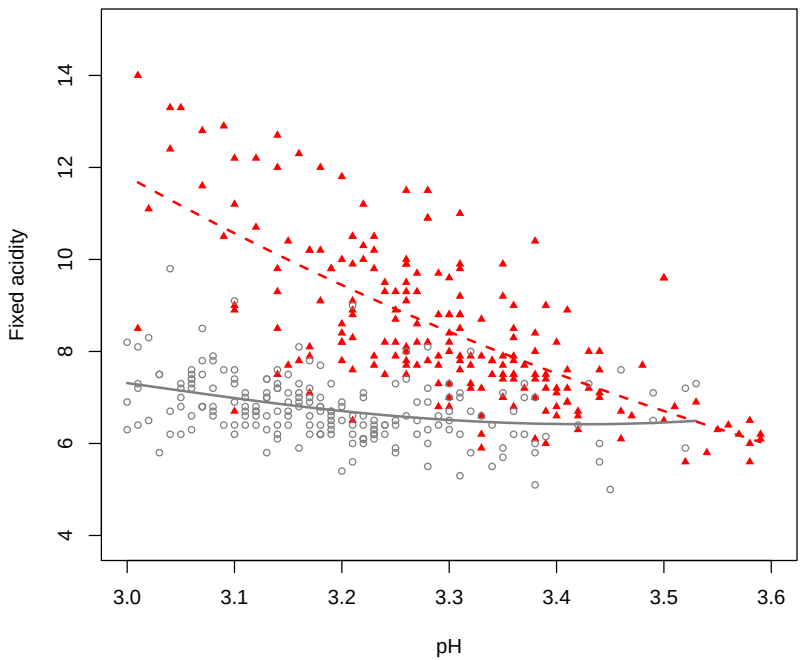
Datos de vino verde

Acidez fija versus pH, ▲= vinos tintos, ●= Vinos blancos

Datos completos



Submuestras $n_1 = n_2 = 200$



Datos de vino verde

Acidez fija versus pH, ▲= vinos tintos, ●= Vinos blancos

