Correlación lineal y correlación de distancias

José R. Berrendero
Departamento de Matemáticas
Universidad Autónoma de Madrid

Jornada Pearson, 8 de marzo de 2017

Esquema

- Ventajas e inconvenientes del coeficiente de correlación lineal.
- La covarianza y la correlación de distancias.
- Clasificación de datos funcionales.
- El uso de la correlación de distancias para seleccionar variables en clasificación de datos funcionales.

El concepto de correlación

Conceptualmente, la idea de correlación se debe a **Francis Galton**. En un artículo de 1888 trata de dar un significado formal al término.

I. "Co-relations and their Measurement, chiefly from Anthropometric Data." By FRANCIS GALTON, F.R.S. Received December 5, 1888.

"Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.

Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. Thus the length of the arm is

El coeficiente de correlación

Karl Pearson introdujo el coeficiente de correlación en dos artículos de 1896 y 1898.

(*b.*) *On the best Value of the Correlation Coefficient.*—This is the well-known Galtonian form of the frequency for two correlated variables, and r is the GALTON function or coefficient of correlation. The question now arises as to what is *practically* the best method of determining r. I do not feel satisfied that the

is then negative. Thus, it appears that the observed result is the most probable, when r is given the value $S(xy)/(n\sigma_1\sigma_2)$. This value presents no practical difficulty in calculation, and therefore we shall adopt it. It is the value given by BRAVAIS, but he does not show that it is the best.*

(*c.*) *Probable Error of the Correlation Coefficients.*—Assuming that r has this

Limitaciones

En 1920, Pearson escribe sobre las limitaciones del concepto de correlación:

As early as 1893 I dealt with quite a number of correlation tables for long series and was able to demonstrate

(i) by applying Galton's process of drawing contours of equal frequency that most smooth and definite systems of contours can arise from long series, obviously mathematical families of curves, which are (*a*) ovaloid, not ellipsoid, and (*b*) which do not possess—like the normal surface contours—more than one axis of symmetry,

(ii) that regression curves can be quite smooth mathematical curves differing widely from straight lines,

(iii) that in cases wherein (i) and (ii) hold, homoscedasticity is not the rule.

I obtained differential equations to such systems, but for more than 25 years while often returning to them, have failed to obtain their integration.

This seems to me the desideratum of the theory of correlation at the present time: the discovery of an appropriate system of surfaces, which will give bi-variate skew frequency. We want to free ourselves from the limitations of the normal surface, as we have from the normal curve of errors.

Coeficiente de correlación lineal

Si X e Y son dos v.a. con varianza finita:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

- $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\sigma(X)$, $\sigma(Y)$ son las desviaciones típicas de X e Y

Es fácil verificar

$$\rho(X, Y)^2 = \frac{\sigma^2(Y) - \min_{a,b} \mathbb{E}[(Y - (aX + b))^2]}{\sigma^2(Y)}.$$

- $-1 \leq \rho(X, Y) \leq 1$
- Cuando $\mathbb{P}(Y = aX + b) = 1$ se tiene $\rho(X, Y) = \mp 1$

¿Por qué es tan utilizado?

- Una interpretación clara:

$$\rho(X,Y)^2 = 1 - \frac{\min_{a,b} \mathbb{E}[(Y-(aX+b))^2]}{\min_a \mathbb{E}[(Y-a)^2]}.$$

- Fácil de calcular (solo requiere conocer segundos momentos).

- Fácil de estimar.

- Buenas propiedades ante transformaciones lineales:

$$\rho(aX+b, cY+d) = \text{signo}(a \cdot c)\rho(X,Y).$$

- Resulta natural como medida de dependencia **para vectores normales**.

Algunos defectos

- Solo está definido cuando las varianzas de X y de Y son finitas.

- $\rho(X, Y) = 0$ no implica en general que X e Y sean independientes.

  - $X \sim \mathrm{N}(0, 1)$
  - $\mathbb{P}(Z = 1) = \mathbb{P}(Z = -1) = 1/2$, independiente de X
  - $Y = ZX$
  - $Y \sim \mathrm{N}(0, 1)$, X e Y no son independientes, pero

$$\rho(X, Y) = \mathbb{E}(XY) = \mathbb{E}(X^2 Z) = 0.$$

Conclusión: X e Y normales y $\rho(X, Y) = 0$ no implica que X e Y sean independientes.

Hace falta que (X, Y) sea un vector normal bidimensional.

Correlación y distribuciones marginales

La correlación como medida de dependencia

Correlaciones pequeñas no deben interpretarse como una indicación de que la dependencia entre las variables es débil.

El rango de valores que puede tomar $\rho(X, Y)$ depende de las distribuciones de X e Y.

- Supongamos $X \sim \exp(1)$, $Y \sim \exp(1)$.
- Si $\rho(X, Y) = -1$, entonces $Y = aX + b$ c.s. con $a < 0$ y $b \in \mathbb{R}$,

$$\mathbb{P}(aX < b) > 0 \Rightarrow \mathbb{P}(Y < 0) > 0,$$

  en contradicción con $\mathbb{P}(Y \geq 0) = 1$.
- De hecho, si $U \sim \text{Unif}(0, 1)$, $F^{-1}(U) = -\log(1 - U) \sim \exp(1)$.

$$\rho(X, Y) \geq \rho(-\log(U), -\log(1 - U)) = 1 - \pi^2/6 \approx -0.6445$$

Correlación y distribuciones marginales

Teorema [Höffding (1940), Fréchet (1957]

Sea (X, Y) un vector aleatorio con marginales F_1 y F_2 con $0 < \sigma^2(X), \sigma^2(Y) < \infty$. Entonces

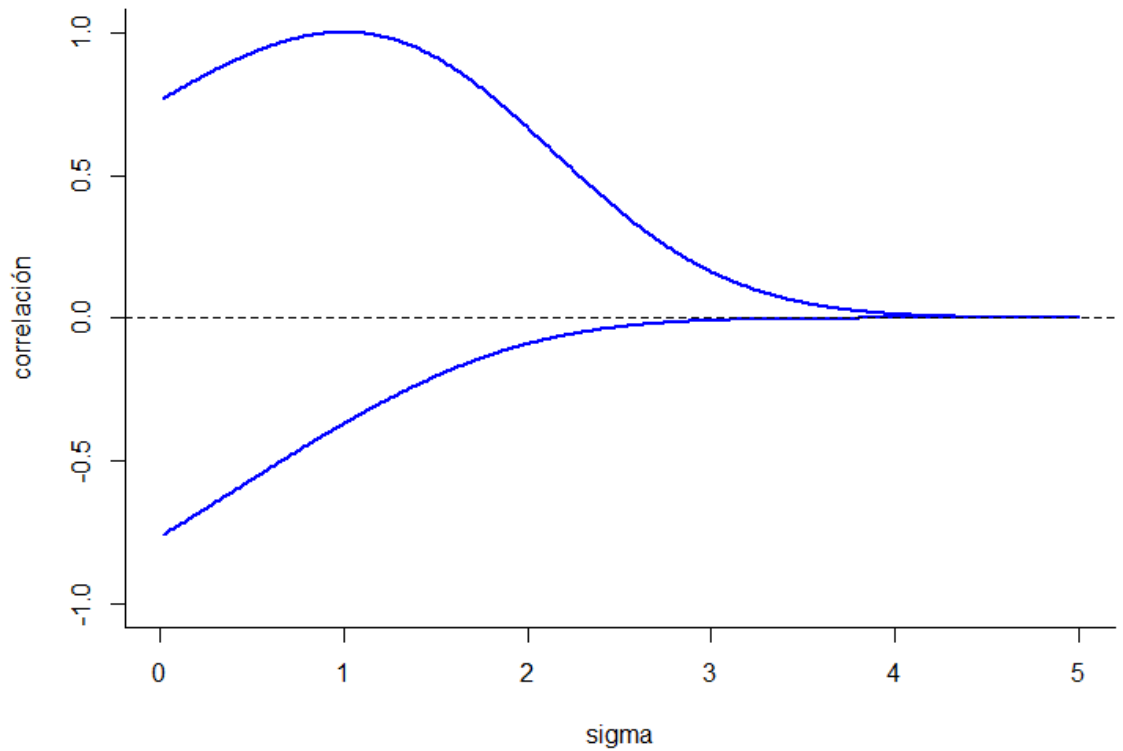
1. El conjunto de posibles correlaciones es un intervalo $[\rho_{\min}, \rho_{\max}]$ tal que $\rho_{\min} < 0 < \rho_{\max}$.

2. $\rho = \rho_{\max}$ si y solo si existe una v.a. Z y dos funciones monótonas crecientes u y v tales que $(X, Y) =_d (u(Z), v(Z))$.

3. $\rho = \rho_{\min}$ si y solo si existe una v.a. Z y dos funciones monótonas, una creciente y otra decreciente, tales que $(X, Y) =_d (u(Z), v(Z))$.

Correlación y distribuciones marginales

Si $X \sim \text{Lognormal}(0, 1)$, $Y \sim \text{Lognormal}(0, \sigma^2)$, entonces $\rho_{\text{min}} = \rho(e^Z, e^{-\sigma Z})$ y $\rho_{\text{max}} = \rho(e^Z, e^{\sigma Z})$, donde $Z \sim \text{N}(0, 1)$.



Si $\sigma = 2$, $\rho_{\text{min}} = -0.09$ y $\rho_{\text{max}} = 0.67$.

Nuevas medidas de dependencia/correlación

MATHEMATICS

A Correlation for the 21st Century

Terry Speed

Most scientists will be familiar with the use of Pearson's correlation coefficient r to measure the strength of association between a pair of variables: for example, between the height of a child and the average height of their parents ($r = 0.5$; see the figure, panel A), or between wheat yield and annual rainfall ($r = 0.75$, panel B). However, Pearson's r captures only linear association, and its usefulness is greatly reduced when associations are nonlinear. What has long been needed is a measure that quantifies associations between variables generally, one that reduces to Pearson's in the linear case, but that behaves as we'd like in the nonlinear Ysidro Edgeworth and later Karl Pearson gave us the modern formula for estimating r, and it very definitely required a manual or electromechanical calculator to convert 1000 pairs of values into a correlation coefficient. In marked contrast, the MIC requires a modern digital computer for its calculation; there is no simple formula, and no-one could compute it on any calculator. This is another instance of computer-intensive methods in statistics (*3*).

It is impossible to discuss measures of association without referring to the concept of independence. Events or measurements are termed probabilistically independent if information about some does not change the

Speed, T. (2011). A correlation for the 21st century. *Science*, 1502-1503.

Correlación de distancias

La **correlación de distancias** [Székely, Rizzo y Bakirov (2007)] es una medida que caracteriza la independencia:

$$\delta(X, Y) = 0 \Leftrightarrow X \text{ e } Y \text{ son independientes}$$

y además

- es fácil de estimar
- se define de forma natural para vectores aleatorios X e Y con valores en $\mathbb{R}^p$ y $\mathbb{R}^q$, donde en general $p \neq q$

Comenzamos describiendo la versión muestral.

Covarianza de distancias

Partimos de una muestra de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ de dos vectores X e Y.

Para calcular la **covarianza de distancias**:

- **Distancias:** calculamos dos matrices $n \times n$ de distancias, una para cada vector, cuyas entradas son $\|X_i - X_j\|$ e $\|Y_i - Y_j\|$.

- **Doble centrado:** a cada elemento de las matrices le restamos las medias de su fila y de su columna, y le sumamos la media de toda la matriz.

- **Covarianza:** se calculan las covarianzas entre las n^2 distancias centradas.

Covarianza y correlación de distancias

Más formalmente,

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..})(b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..})$$

donde $a_{ij} = \|X_i - X_j\|$ y $b_{ij} = \|Y_i - Y_j\|$.

La **correlación de distancias** es la correspondiente versión estandarizada:

$$\mathcal{R}_n^2(X,Y) = \frac{\mathcal{V}_n^2(X,Y)}{[\mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y)]^{1/2}}$$

¿Hacia dónde converge la covarianza de distancias?

Consistencia

Si $\mathbb{E}\|X\| < \infty$ y $\mathbb{E}\|Y\| < \infty$,

$$\lim_{n\to\infty} \mathcal{V}_n^2(X, Y) = \mathcal{V}^2(X, Y) \quad \text{c.s.}$$

- En términos de **distancias**: $\mathcal{V}^2(X, Y) = \mathbb{E}[d(X, X')d(Y, Y')]$,

$$d(X, X') = \|X - X'\| - m(X) - m(X') + \mathbb{E}(\|X - X'\|),$$
$$m(x) = \mathbb{E}(\|X - x\|)$$

- En términos de la **función característica**:

$$\mathcal{V}^2(X, Y) = \int |\varphi_{(X,Y)}(u, v) - \varphi_X(u)\varphi_Y(v)|^2 w(u, v)dudv,$$

para cierta función de pesos $w(u, v) \geq 0$.

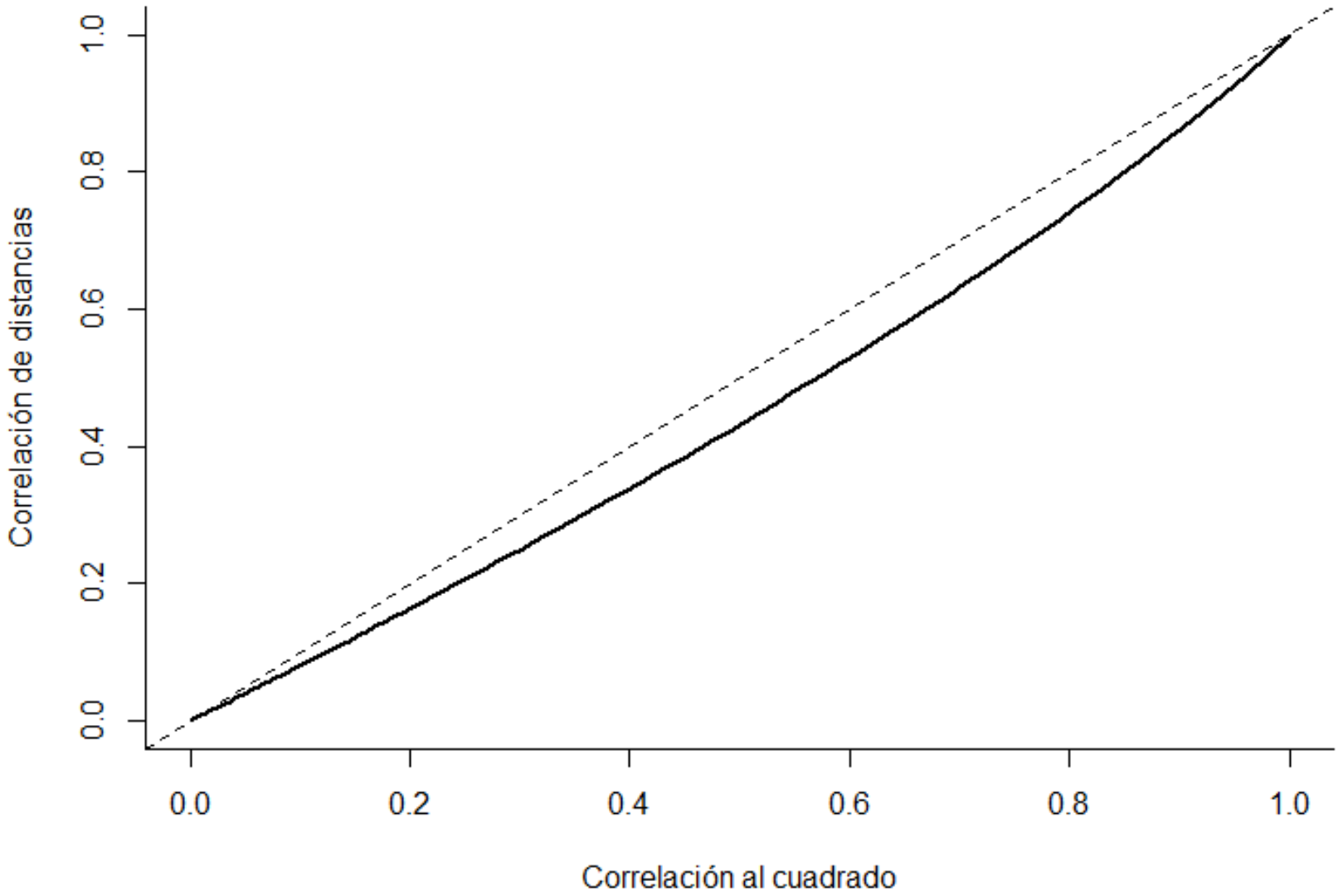
Propiedades básicas

Consecuencia importante

$$\mathcal{V}^2(X, Y) = 0 \Leftrightarrow X \text{ e } Y \text{ independientes}$$

- Se puede calcular la correlación de distancias para vectores de distinta dimensión.
- Es invariante por cambios de escala de los datos.
- Es invariante por rotaciones de los datos.
- Se puede generalizar a cualquier espacio métrico (aunque no en todos caracteriza la independencia).

Cuando (X, Y) es normal bidimensional



Distribución asintótica bajo independencia

$$H_0 : X \text{ e } Y \text{ independientes}$$

Distribución asintótica bajo H_0

Bajo H_0, si $\mathbb{E}\|X\| < \infty$ y $\mathbb{E}\|Y\| < \infty$, cuando $n \to \infty$

$$n\mathcal{V}_n^2(X, Y) \longrightarrow_d \sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

donde $Z_1, Z_2, \ldots$ son normales estándar independientes y $\lambda_1, \lambda_2, \ldots$ son constantes no negativas que dependen de la distribución de (X, Y).

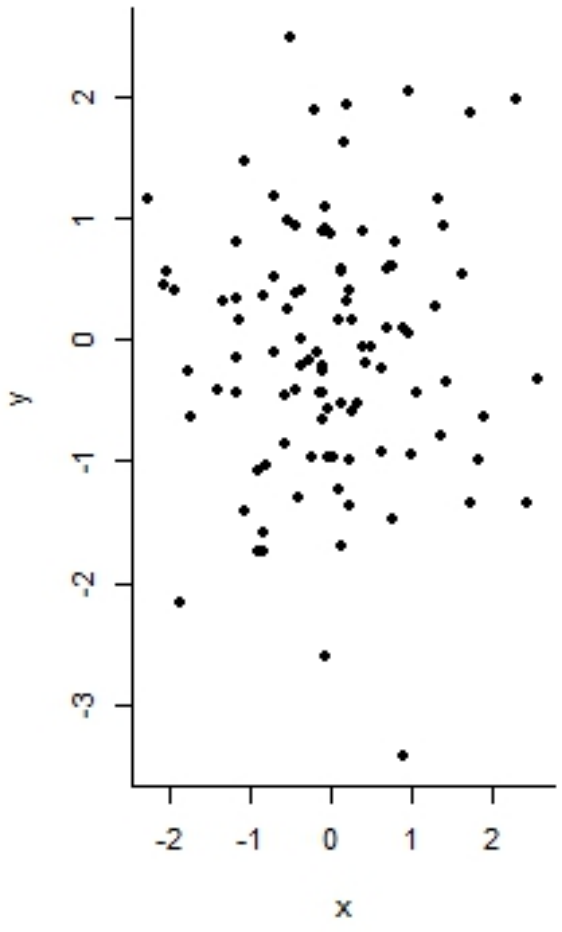
- Complicada de usar porque depende de la distribución de las variables.

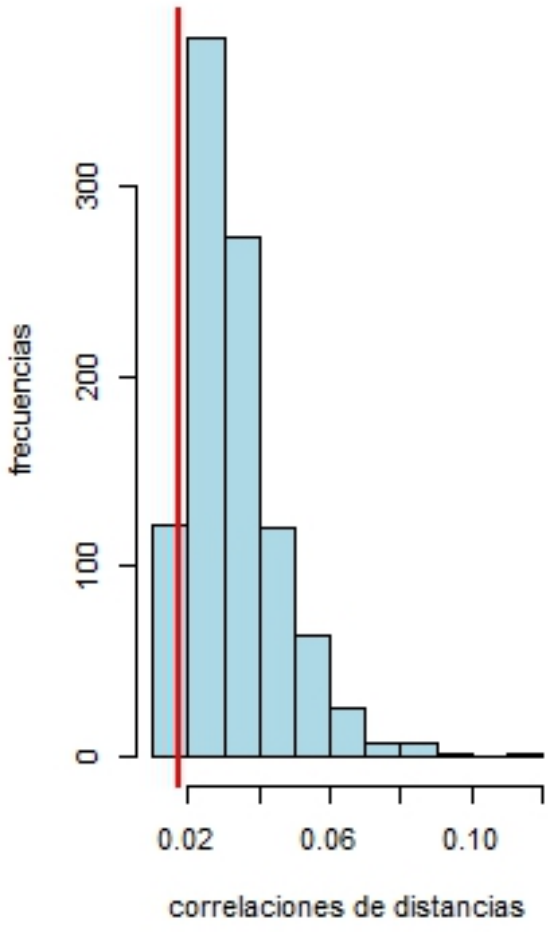
Contraste de independencia basado en permutaciones

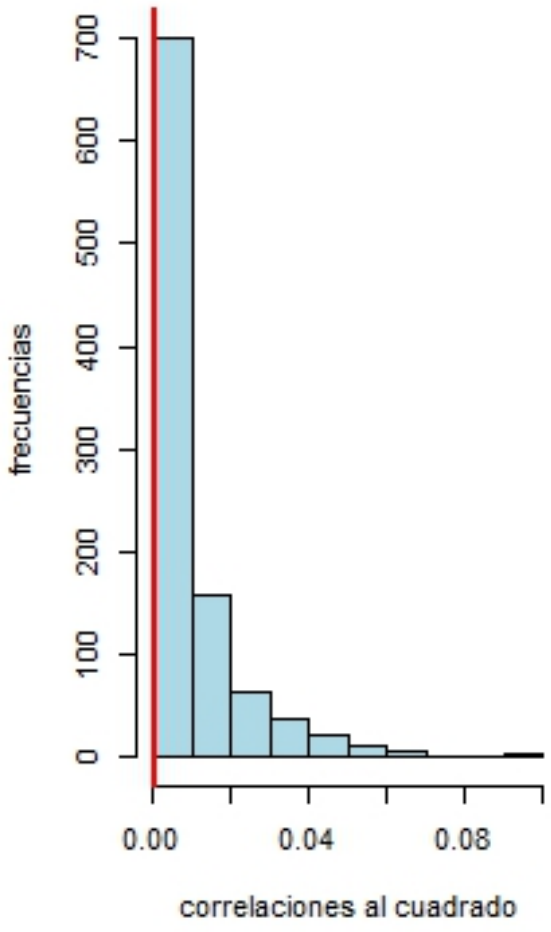
- Calculamos $\mathcal{R}_n^2 = \mathcal{R}_n^2[(X_1, Y_1), \ldots, (X_n, Y_n)]$
- Para $j = 1, \ldots, B$, $\pi_j : \{1, \ldots, n\} \to \{1, \ldots, n\}$ es una permutación.
- Calculamos $\mathcal{R}_n^2(j) = \mathcal{R}_n^2[(X_1, Y_{\pi_j(1)}), \ldots, (X_n, Y_{\pi_j(n)})]$
- Los valores $\mathcal{R}_n^2(j)$, para un número grande de permutaciones, sirven para aproximar la distribución de $\mathcal{R}_n^2$ bajo independencia.
- El p-valor del contraste es

$$p = \frac{\#\{j : \mathcal{R}_n^2(j) > \mathcal{R}_n^2\}}{B}.$$

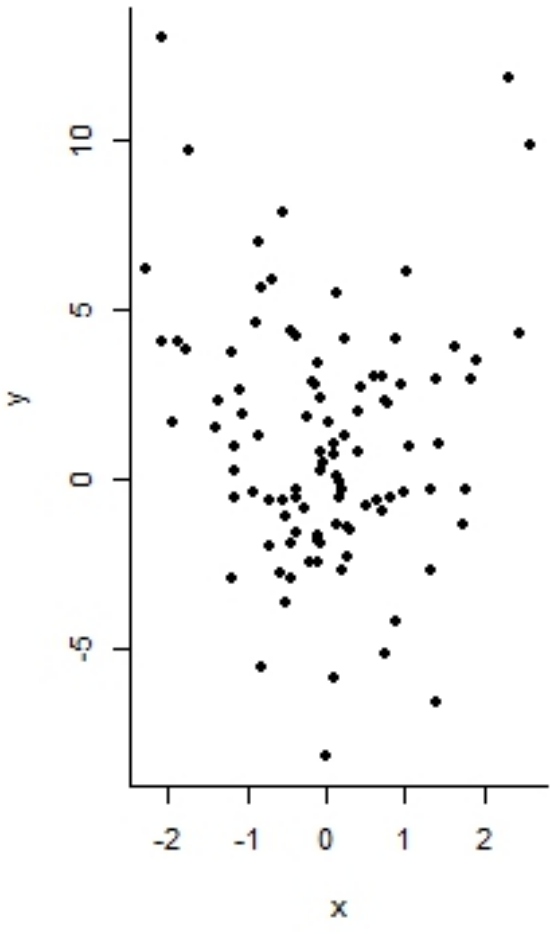
Ejemplo: datos independientes

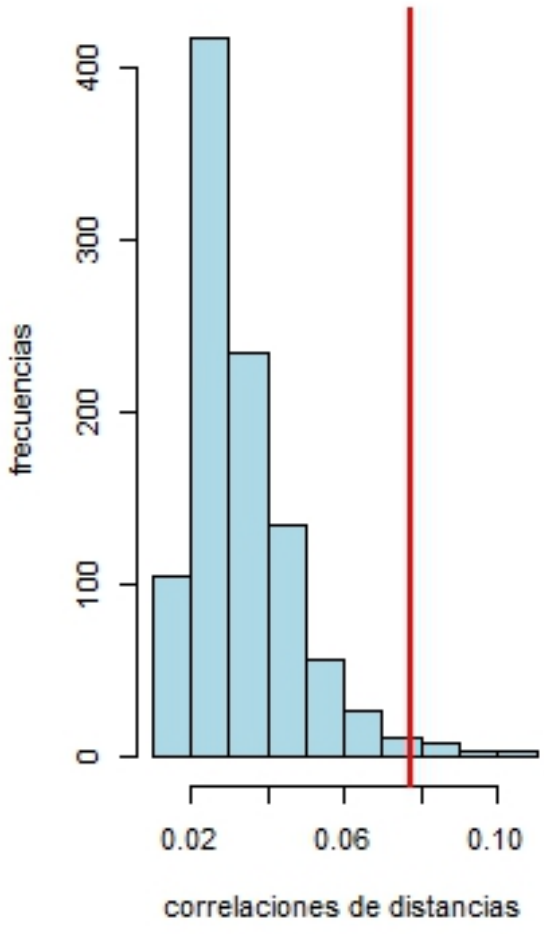


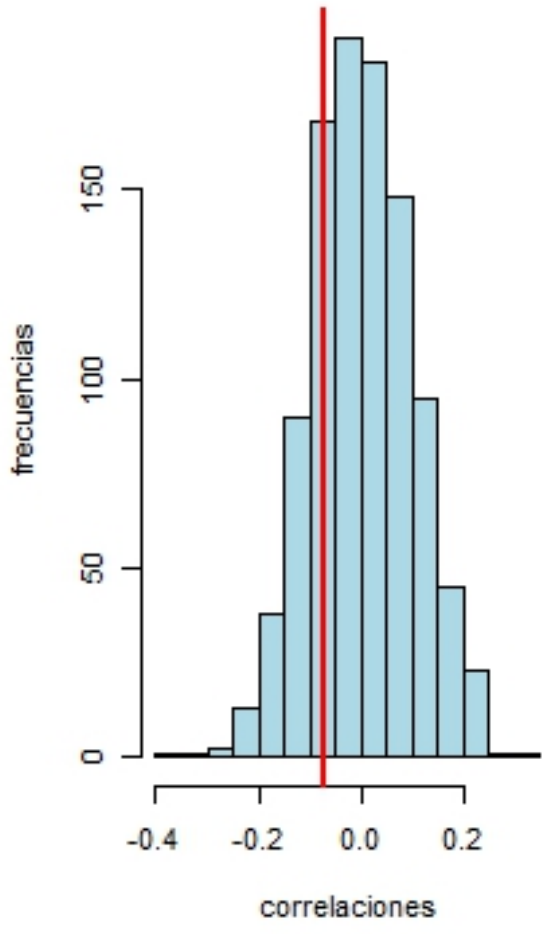




Ejemplo: relación cuadrática







Cuando Y tiene distribución de Bernoulli

El caso $Y \sim B(1, p)$ tiene interés en problemas de clasificación

Covarianza de distancias:

$$\mathcal{V}^2(X, Y) = 4p^2(1-p)^2 \left[D_{01} - \frac{D_{00} + D_{11}}{2} \right],$$

donde

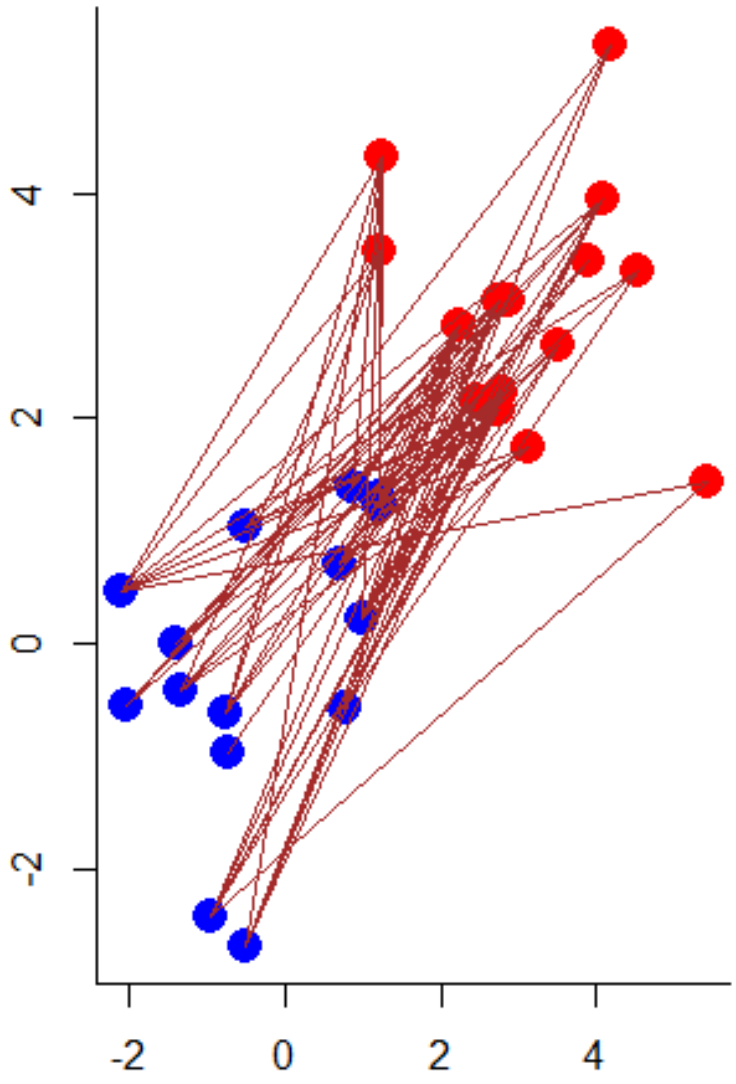
$$D_{ij} = \mathbb{E}(\|X - X'\| \mid Y = i, Y' = j).$$

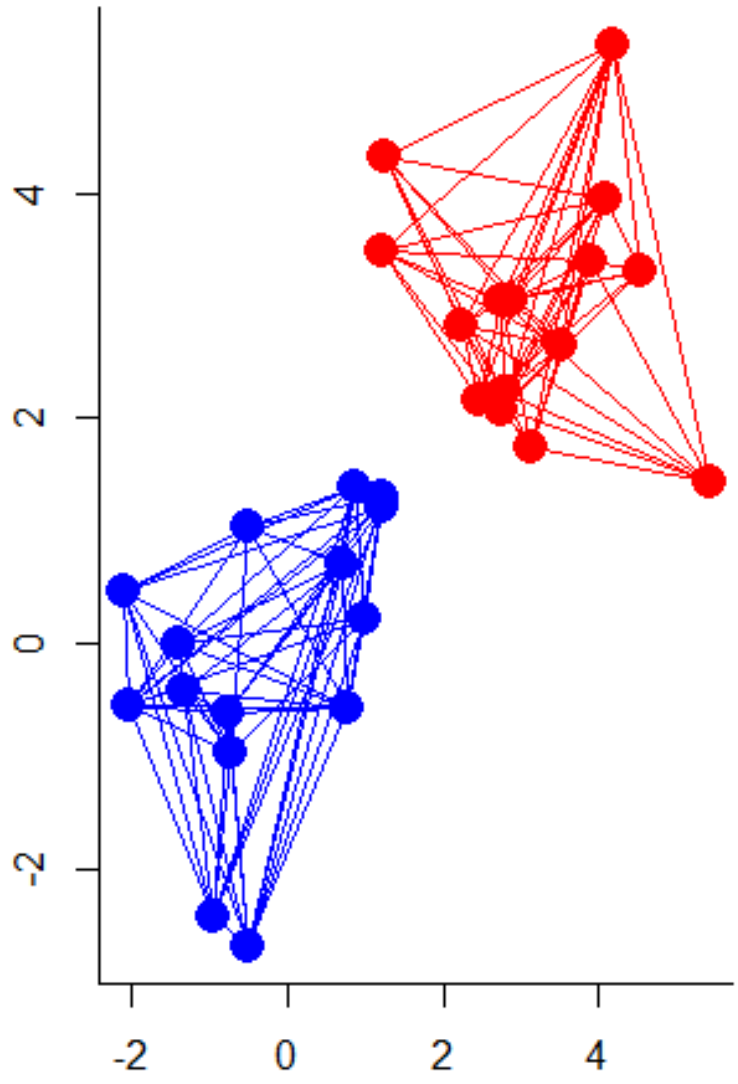
Covarianza:

$$\mathrm{Cov}(X, Y)^2 = p^2(1-p)^2(m_0 - m_1)^2,$$

donde $m_i = \mathbb{E}(X \mid Y = i)$.

Cuando Y tiene distribución de Bernoulli





Análisis de datos funcionales

Análisis de datos que consisten en funciones, imágenes, formas, ...

Se observan n funciones $X_1(t), \ldots, X_n(t)$, con $t \in [a, b]$.

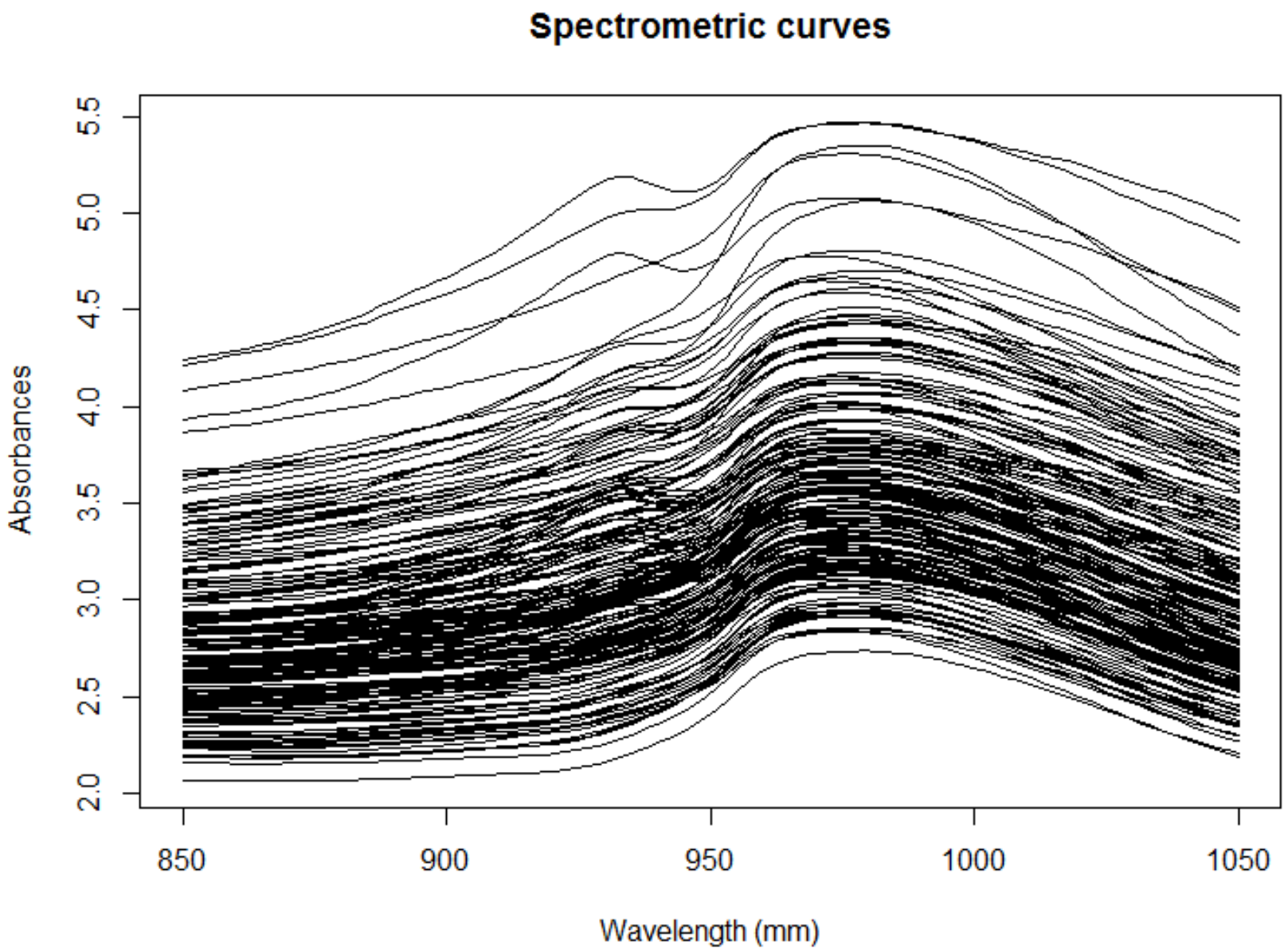
Se supone que estas funciones son trayectorias independientes de un cierto proceso estocástico.

En la práctica las funciones se observan en un grid $a \leq t_1 < \ldots < t_N \leq b$.

Datos funcionales $\neq$ datos multivariantes de alta dimensión:

- La continuidad de las funciones implica redundancia de información
- Puede definirse un orden y una métrica en el conjunto $t_1, \ldots, t_N$

Espectros de absorción de 215 piezas de carne



Clasificación con datos funcionales

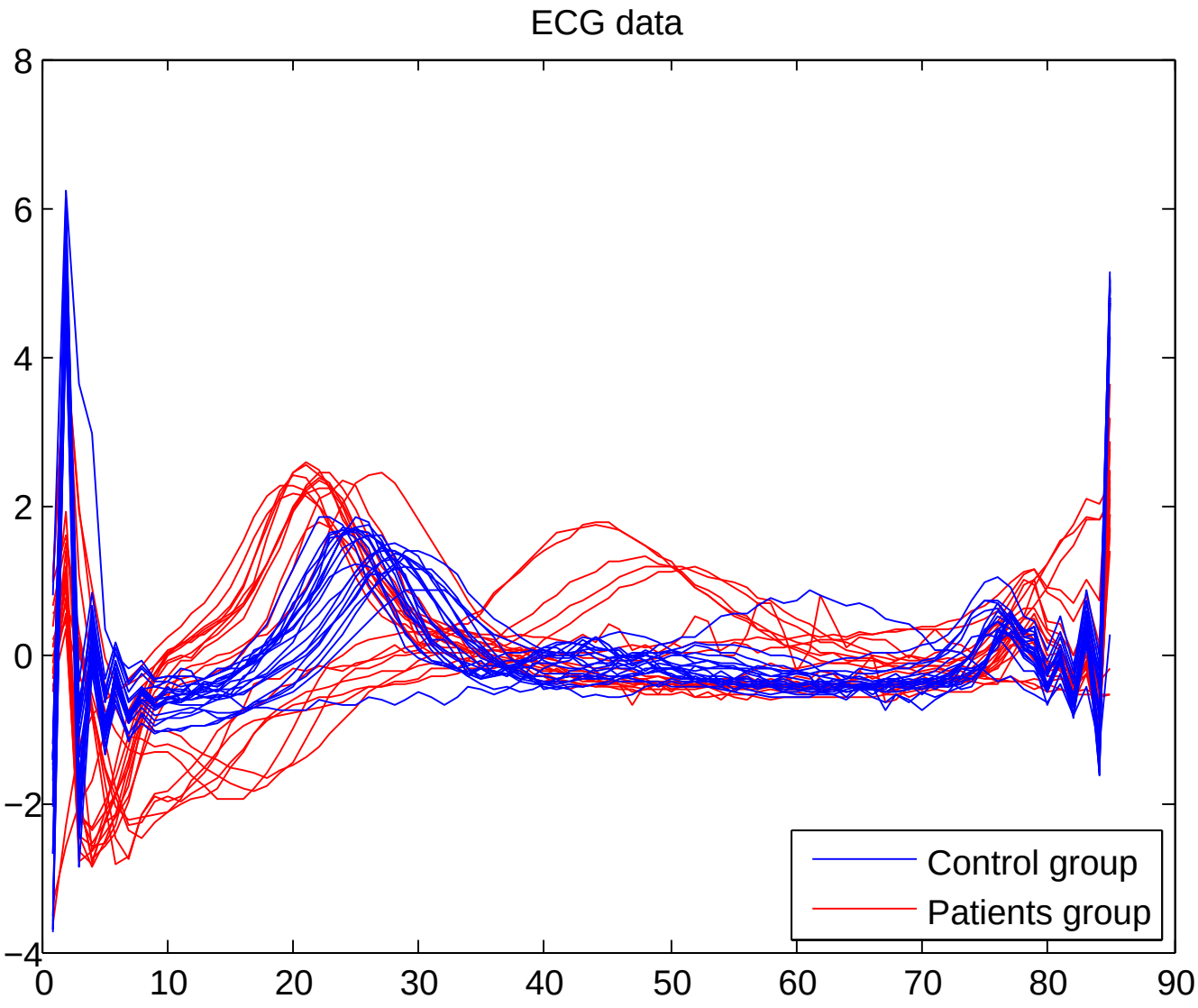
Los datos pueden proceder de una de dos posibles clases, 0 o 1.

Datos de entrenamiento: (X_i, Y_i), $i = 1, \ldots, n$, donde $Y_i \in \{0, 1\}$ es una variable que contine información sobre la clase a la que pertenece X_i.

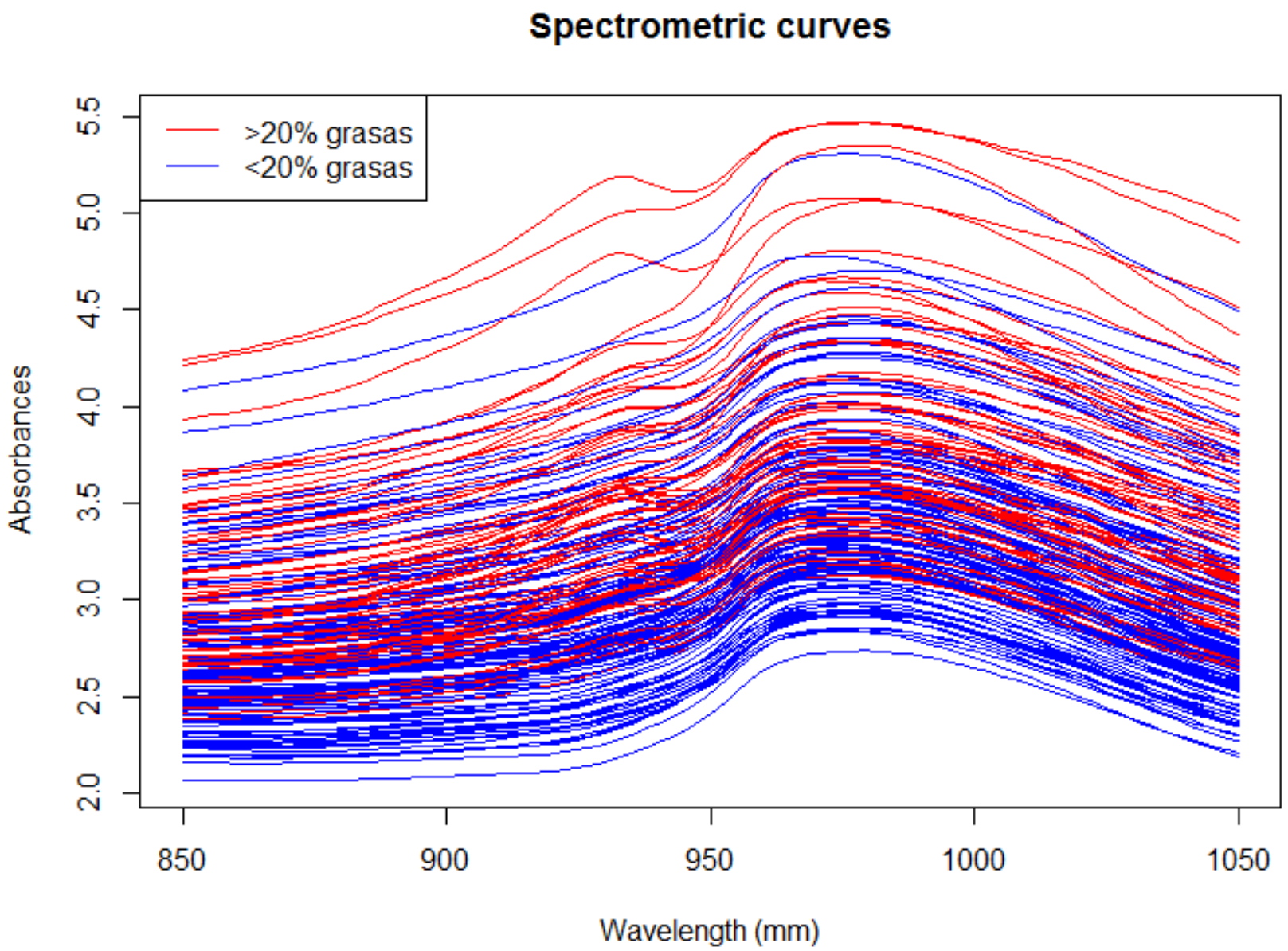
Objetivo: Clasificar una observación $(X, ?)$ independiente de las anteriores, pero que sigue el mismo modelo.

Consideramos el caso en que los valores X_i son funciones.

Clasificación con datos funcionales



Espectros según contenido en grasas



Selección de variables en clasificación de datos funcionales

Clasificar la función usando únicamente sus valores en un número pequeño de puntos cuidadosamente seleccionados

$$\{x(t),\ t \in [0,1]\} \rightsquigarrow (x(t_1^*), \ldots, x(t_d^*))$$

Una vez elegidos $t_1^*, \ldots, t_d^*$, se puede aplicar cualquier método de clasificación para datos multivariantes (Fisher, kNN, SVM,...)

Ventajas:

- Eliminar información redundante y ruido
- Suele mejorar el error de clasificación
- Interpretación fácil en comparación con otros métodos de reducción de dimensión

Relevancia y redundancia

Los puntos elegidos

- **Deben ser relevantes**, es decir, deben tener mucha información acerca de la clase a la que corresponden las funciones.
- **No deben ser redundantes**, es decir, la información que proporciona cada punto debe ser complementaria a la del resto.

Criterio para detectar puntos relevantes

Utilizamos como **criterio** la correlación (o la covarianza) de distancias.

$$\mathcal{R}^2(t) = \mathcal{R}^2(X(t), Y), \quad t \in [0,1]$$
$$\mathcal{R}_n^2(t) = \mathcal{R}_n^2(X(t), Y), \quad t \in [0,1]$$

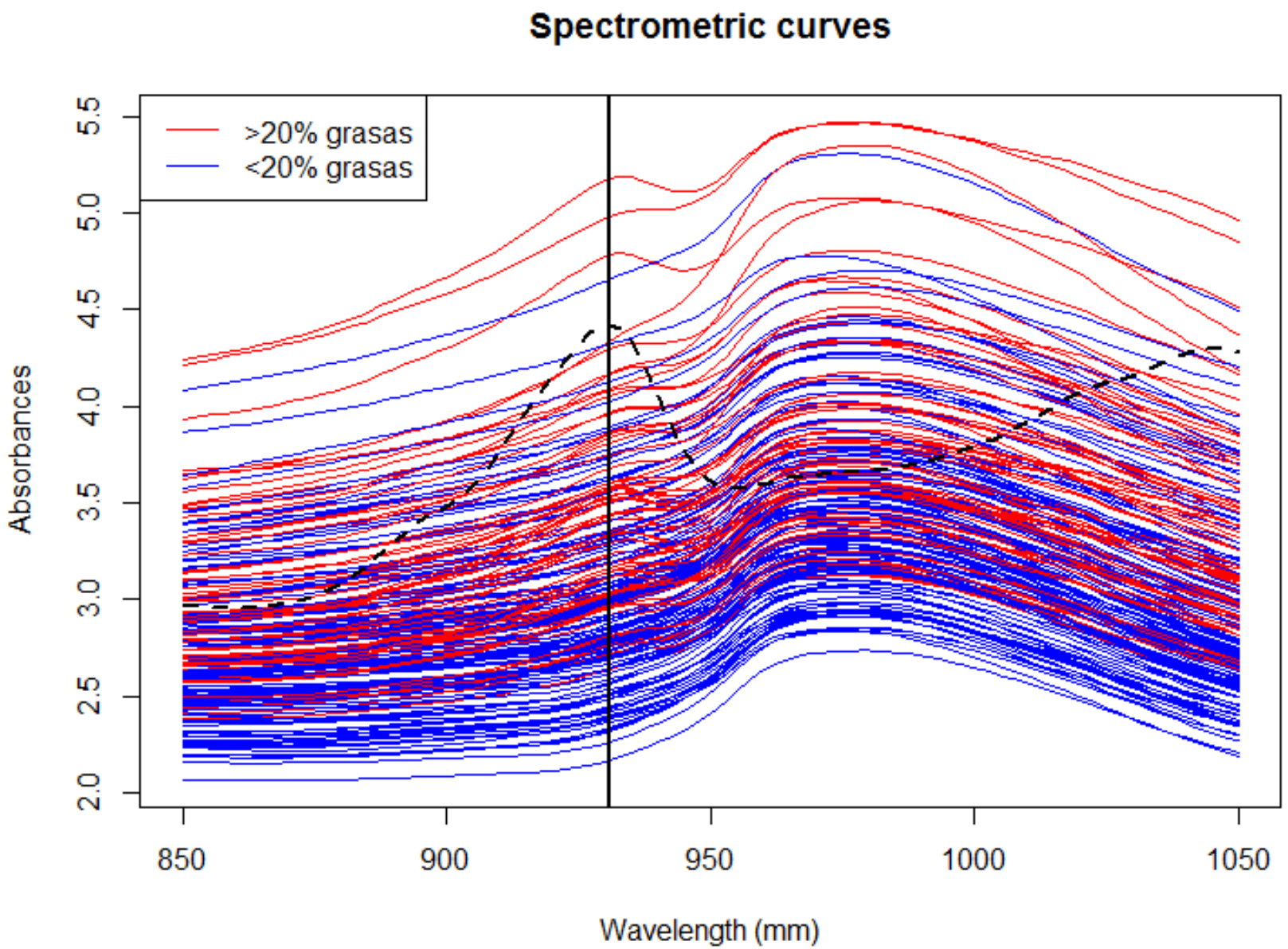
El punto más relevante es el punto t^* que verifica:

$$\mathcal{R}^2(t^*) \geq \mathcal{R}^2(t), \quad \text{para todo } t \in [0,1].$$

En la práctica t^* se estima mediante $\hat{t}^*$ tal que

$$\mathcal{R}_n^2(\hat{t}^*) \geq \mathcal{R}_n^2(t), \quad \text{para todo } t \in [0,1].$$

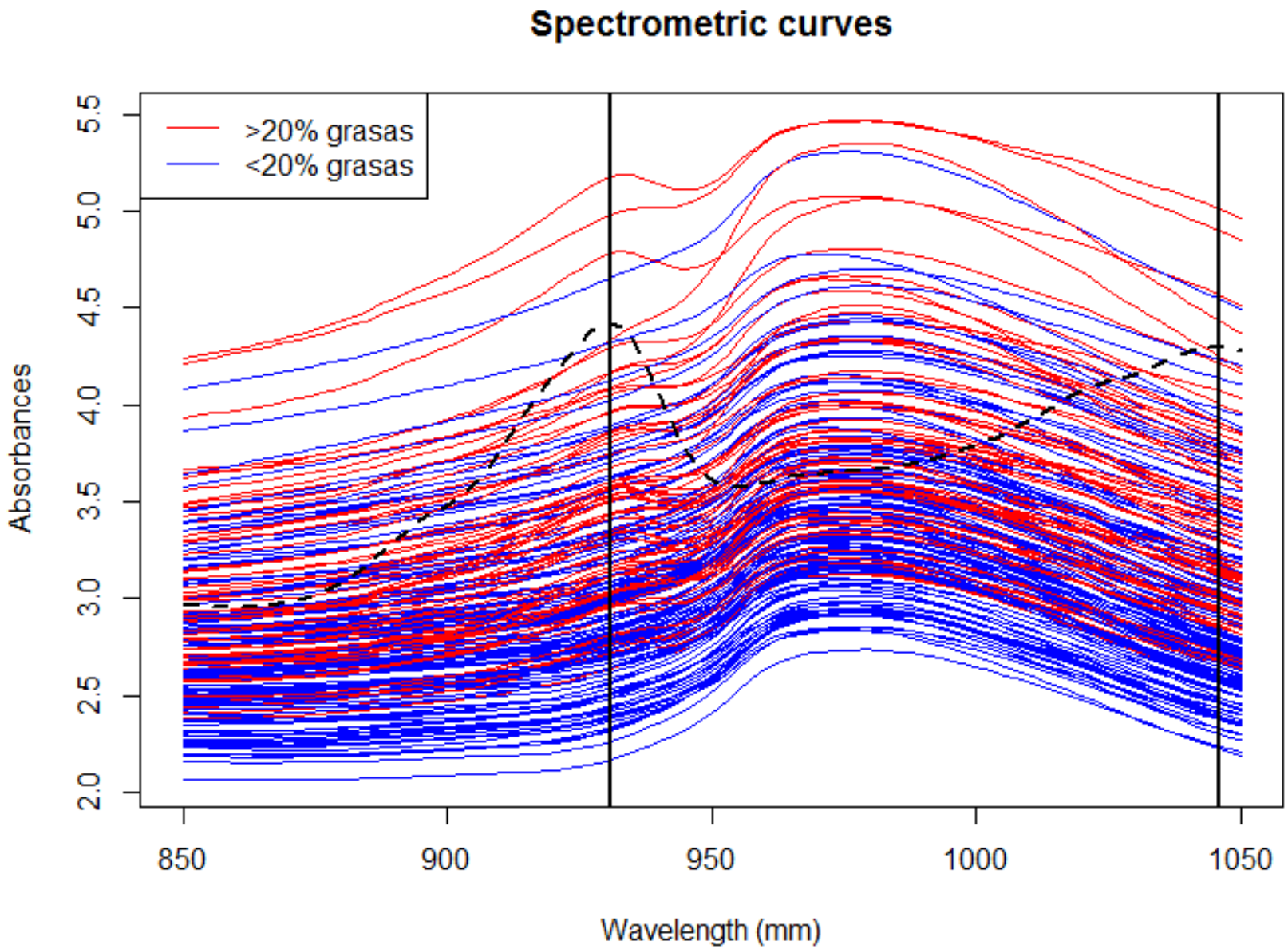
Ejemplo: la variable más relevante



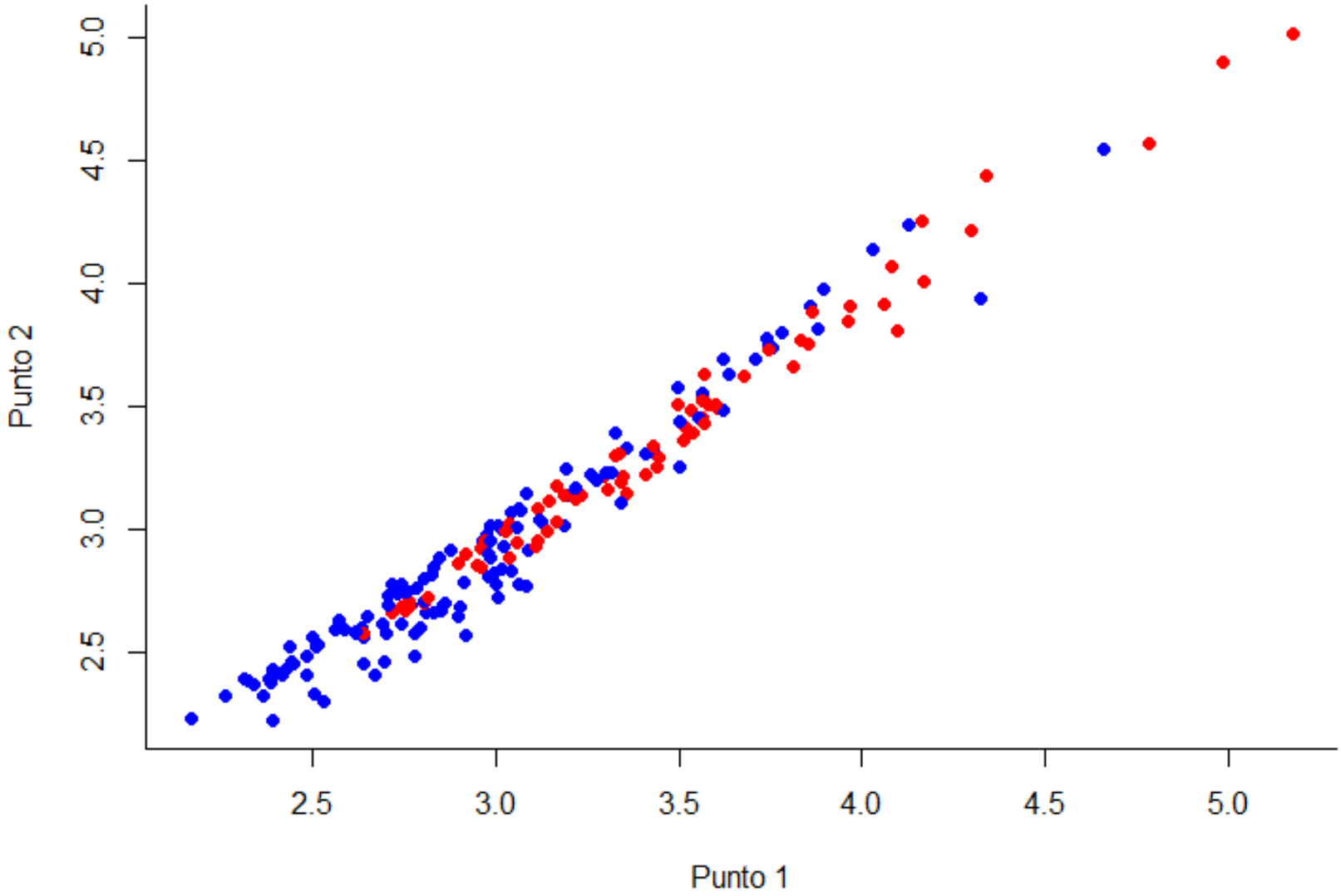
Caza de máximos

- ¿Qué otros puntos pueden ser relevantes?
- No es posible ordenar en función de la correlación de distancias porque solo seleccionaríamos puntos en un entorno del máximo global
- Estos puntos proporcionarían información redundante con la que ya tenemos.
- Nuestra propuesta es seleccionar los puntos que son máximos locales de $\mathcal{R}_n^2(t)$.
- En la práctica, un punto t_i se define como máximo local si da el mayor valor es un subgrid t_j, donde $j = i - h, i - h + 1, \ldots, i + h$,
- El valor de h se selecciona por validación cruzada.

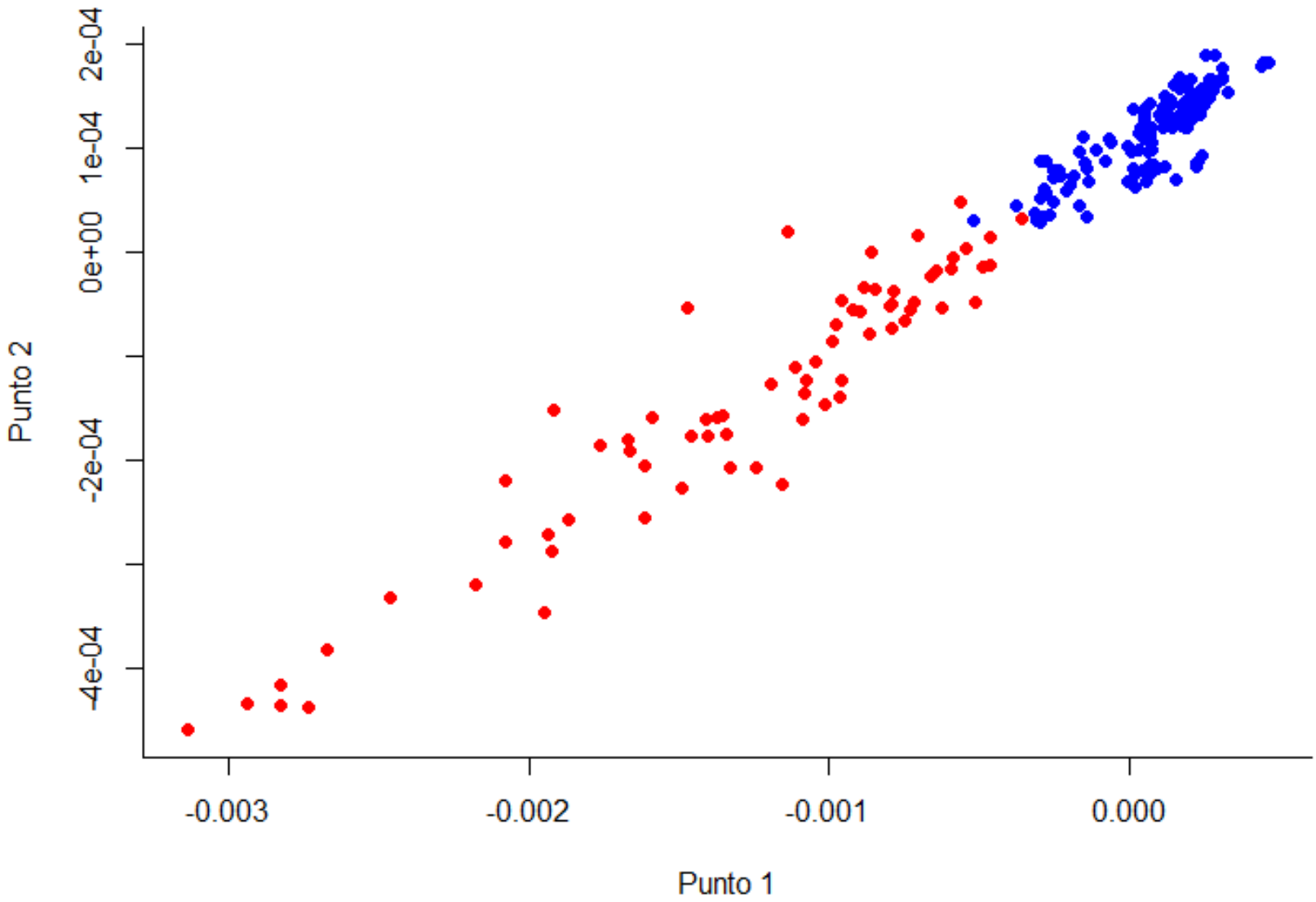
Ejemplo: las dos variables relevantes



Valores de las funciones en los dos puntos seleccionados



Valores de las segundas derivadas de las funciones



Consistencia

El estimador $\mathcal{V}_n^2(t)$ es uniformemente consistente en el sentido siguiente:

Consistencia uniforme

Bajo condiciones de regularidad no restrictivas,

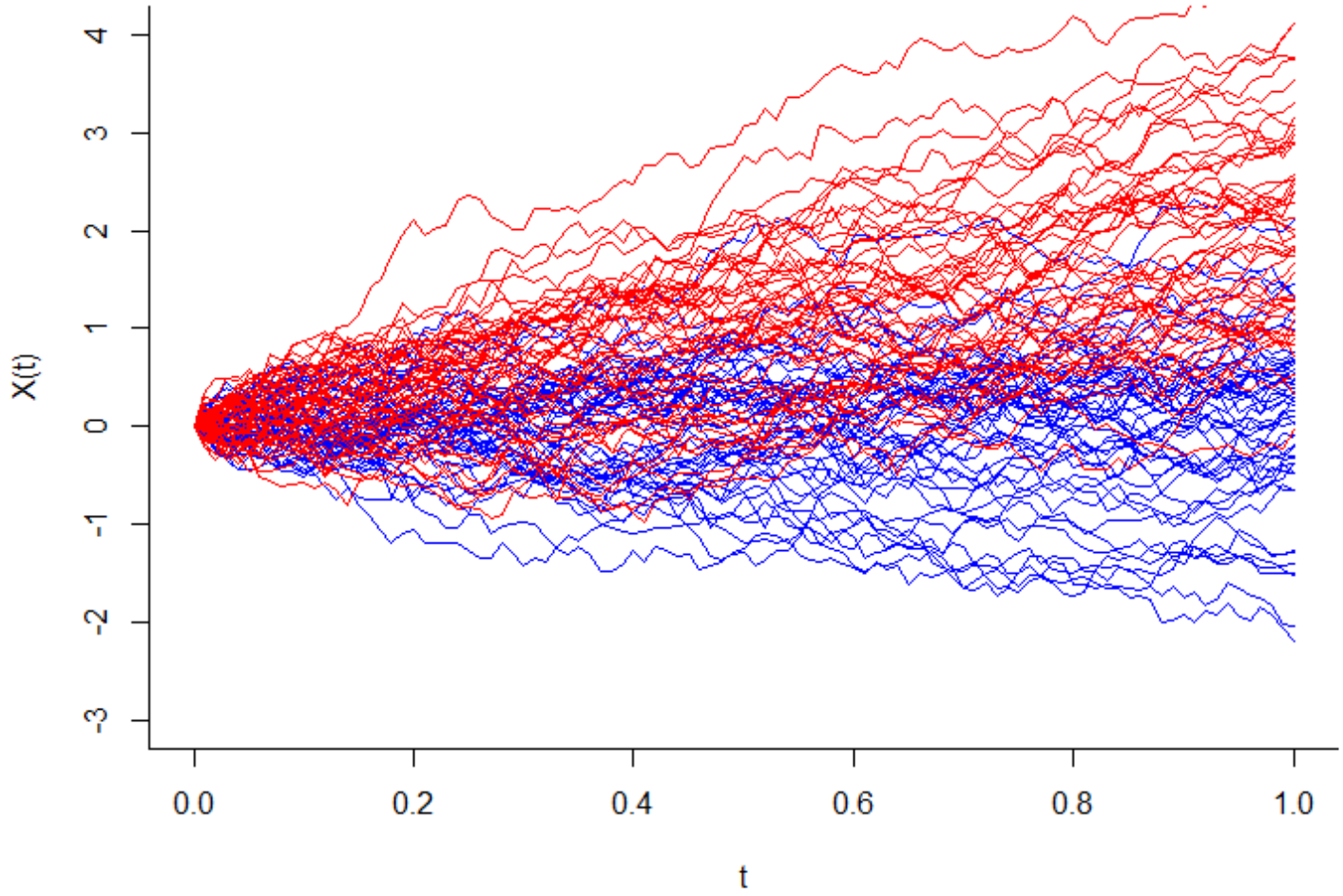
$$\sup_{t\in[a,b]} |\mathcal{V}_n^2(t) - \mathcal{V}^2(t)| \to 0 \quad \text{c.s., si } n \to \infty.$$

- Consecuencia: convergencia de $\hat{t}^*$ a t^*, para todos los máximos locales.
- Podemos aproximar arbirariamente bien los máximos locales con muestras suficientemente grandes.
- Esta propiedad tiene sentido para datos funcionales, pero no para datos multivariantes de alta dimensión.

¿Son relevantes los máximos locales?

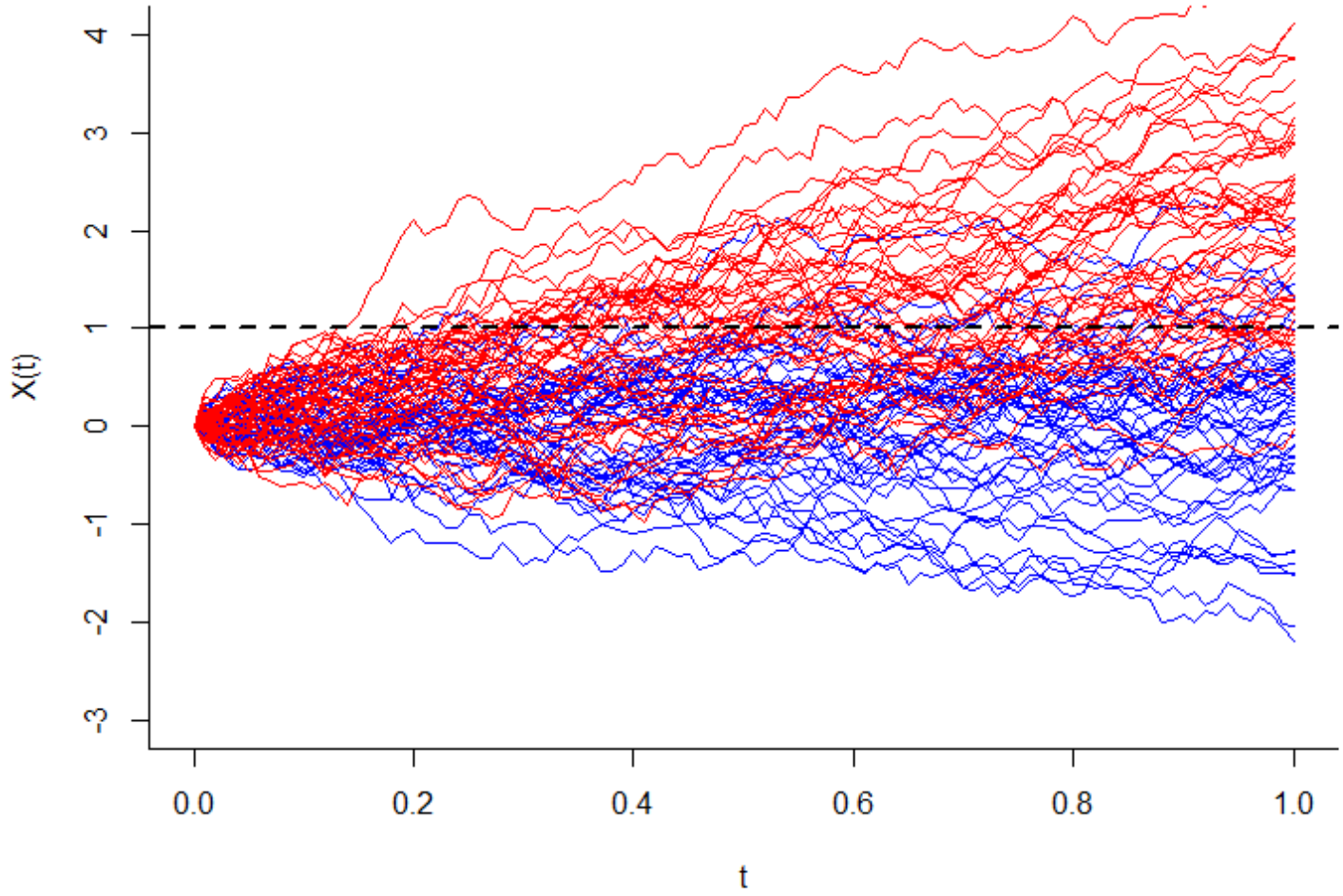
- Para algunos modelos (procesos gaussianos homocedásticos) es posible calcular la regla de clasificación óptima.
- Se pueden dar condiciones para que la regla óptima dependa de un número finito de puntos.
- En algunos casos también es posible calcular la correlación de distancias teórica y ver si sus máximos corresponden a los que determinan la regla óptima.
- En general, los máximos corresponden a puntos relevantes; pero a veces hay puntos relevantes que los máximos no detectan.
- La razón es que se consideran solo los efectos individuales de las variables seleccionadas.

Ejemplo: movimiento browniano



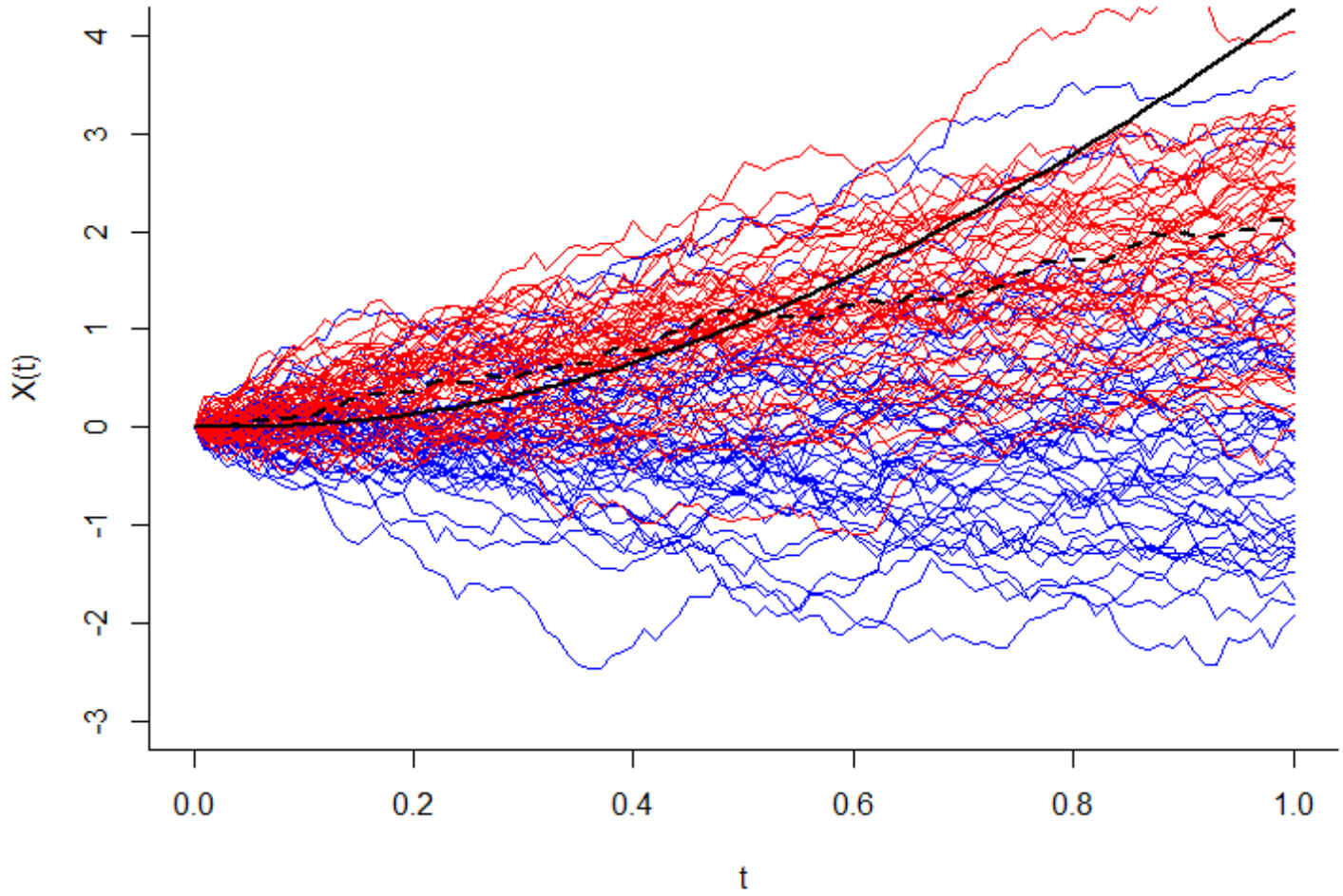
Trayectorias de $B(t)$ y de $B(t) + 2t$

Ejemplo: movimiento browniano



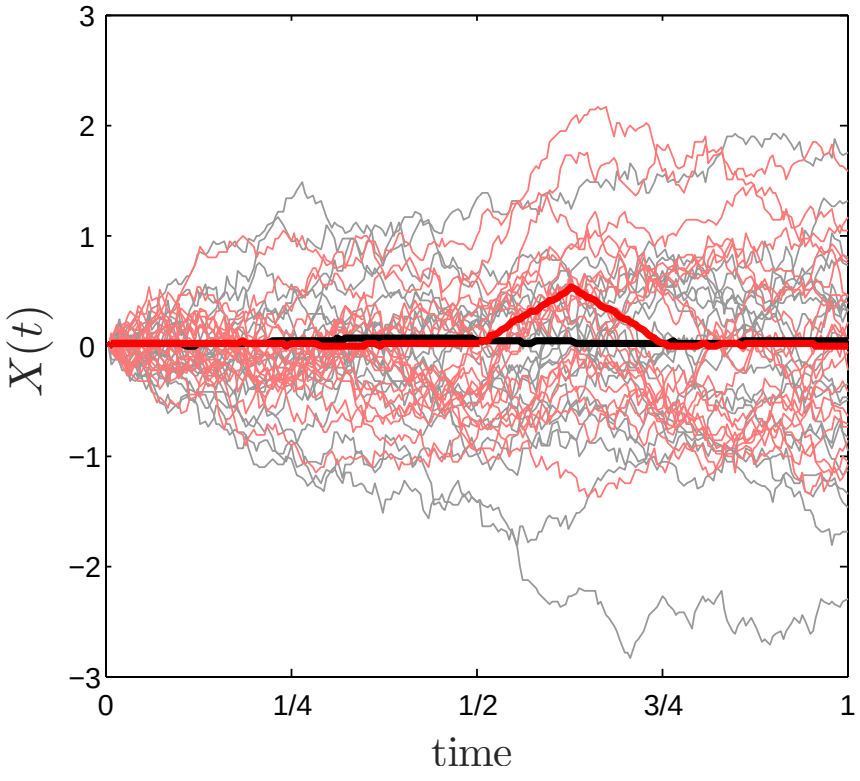
Regla óptima: $X(1) > 1$

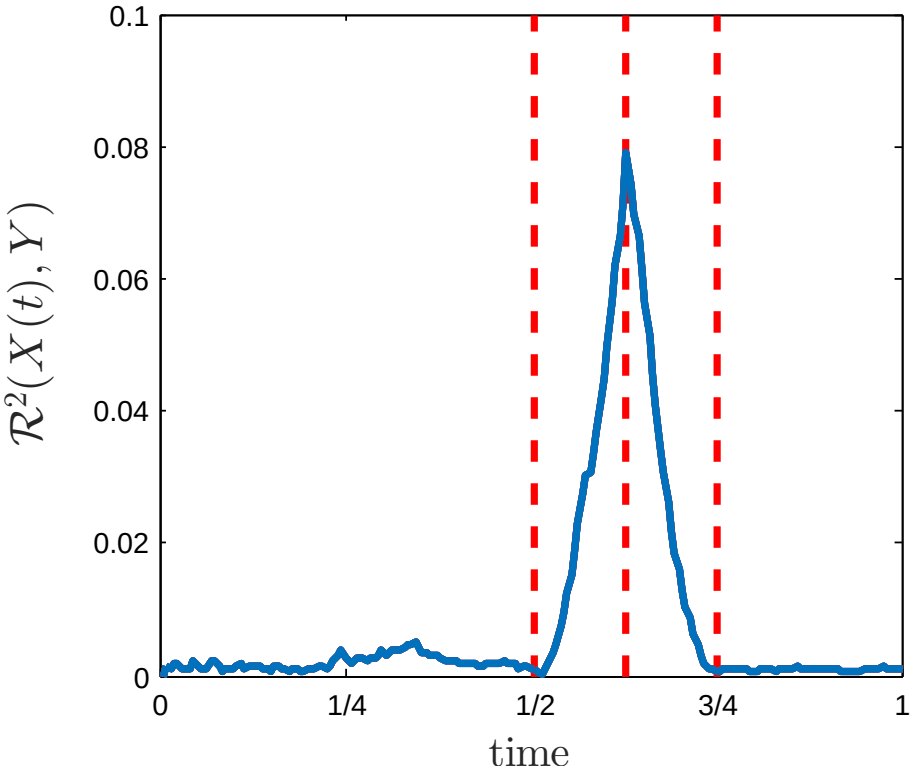
Ejemplo: movimiento browniano



Distancia de covarianzas teórica y estimada

Ejemplo: movimiento browniano





Regla óptima: $[X(5/8) - X(1/2)] + [X(5/8) - X(3/4)] > 1/4$

Referencias

- Berrendero, J. R., Cuevas, A., y Torrecilla, J. L. (2016). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, 619-638.

- Berrendero, J. R., Cuevas, A., y Torrecilla, J. L. (2016). The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation*, 891-907.

- Blyth, S. (1994). Karl Pearson and the correlation curve. *International Statistical Review*, 393-403.

- Embrechts, P., McNeil, A., y Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 176-223.

Referencias

- Lyons, R. (2013). Distance covariance in metric spaces. *Ann. Probab.*, 3284-3305.
- Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, 73-79.
- Székely, G. J., Rizzo, M. L., y Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 2769-2794.
- Székely, G. J., y Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.*, 1236-1265.

Gracias

José R. Berrendero (joser.berrendero@uam.es)