Ejemplo Tortugas

En este ejemplo se miden la dimensiones del caparazón de las tortugas siendo

- $x_1 = 10 \log(\text{longitud del caparazón})$,
- $x_2 = 10 \log(\text{ancho del caparazón})$

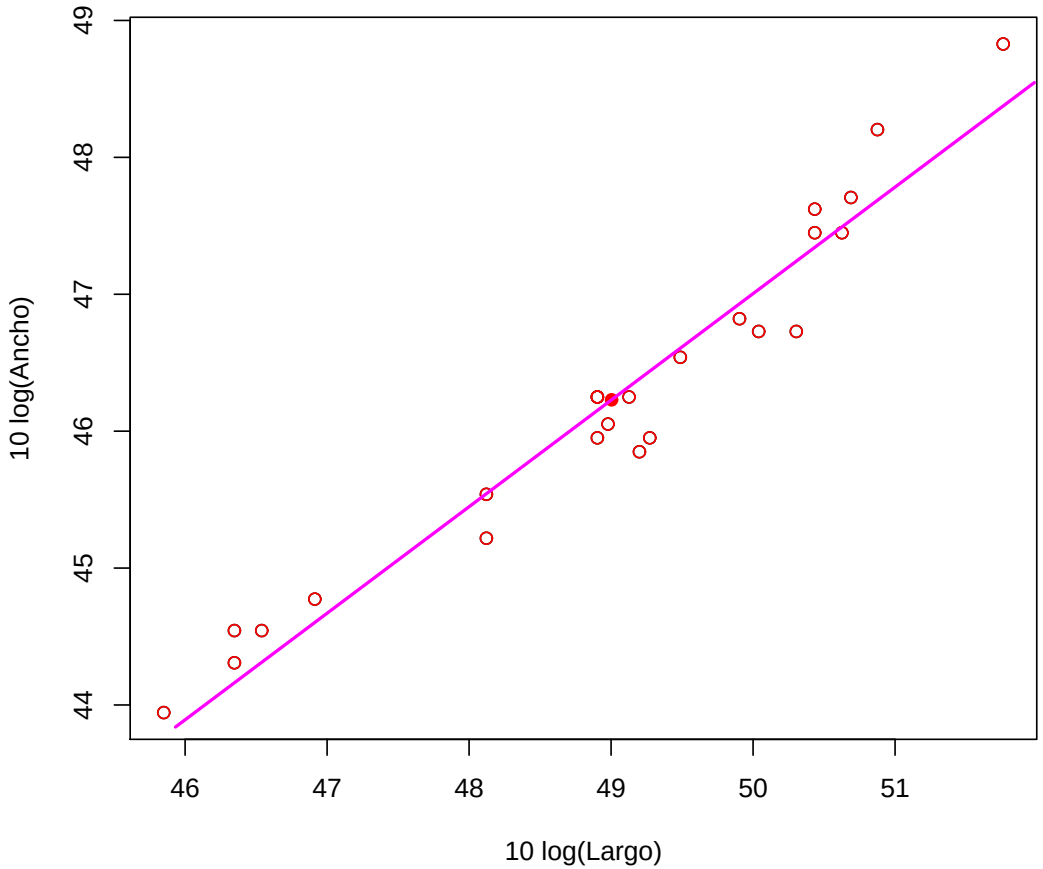
Se estudiaron 24 machos y 24 hembras.

$$\overline{\mathbf{x}}_{\mathrm{M}} = \begin{pmatrix} 47.254 \\ 44.776 \end{pmatrix} \qquad \overline{\mathbf{x}}_{\mathrm{H}} = \begin{pmatrix} 49.004 \\ 46.229 \end{pmatrix}$$

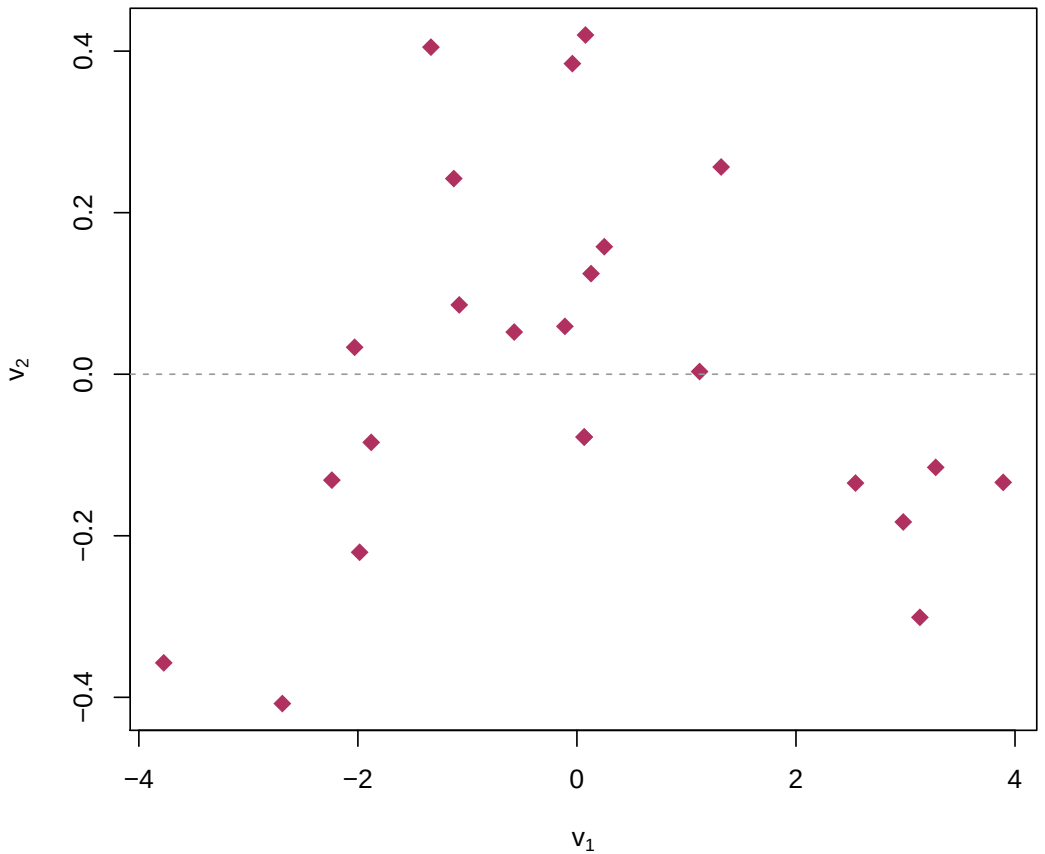
y

$$\widehat{\gamma}_{1,\mathrm{M}} = (0.7996, 0.6005)^{\mathrm{T}} \qquad \widehat{\gamma}_{1,\mathrm{H}} = (0.7892, 0.6141)^{\mathrm{T}}$$

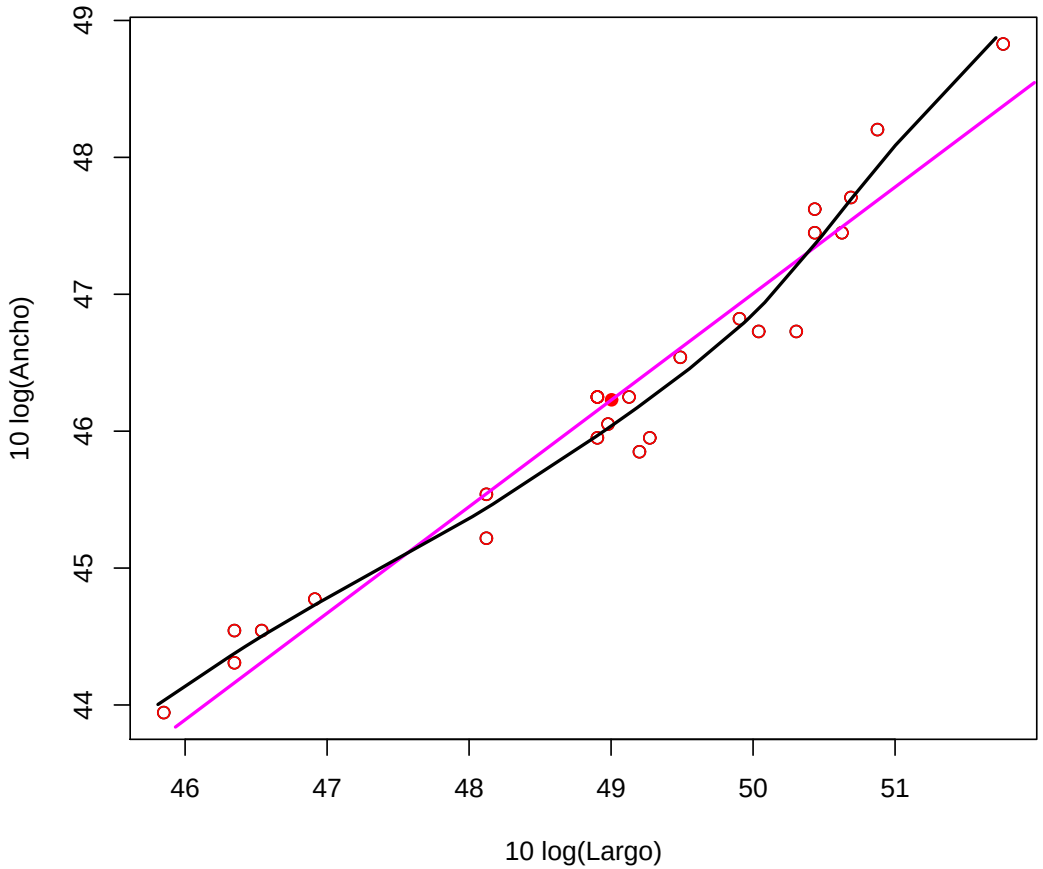
Hembras



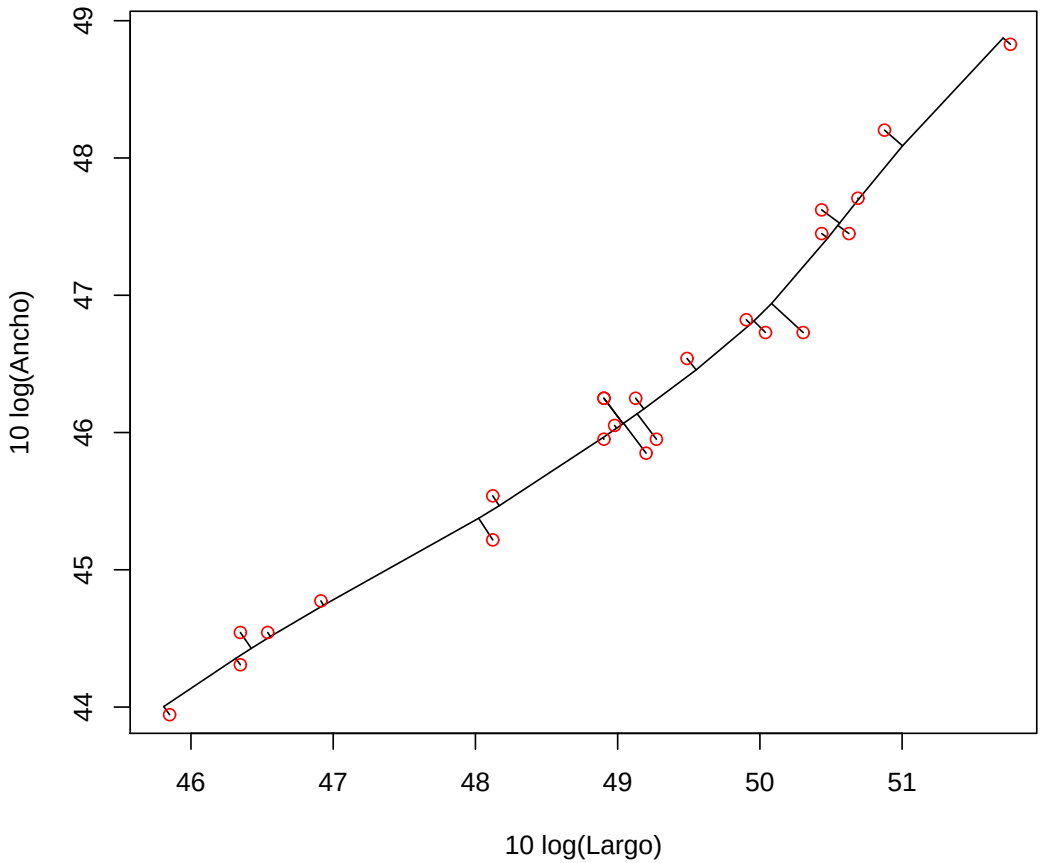
Ejemplo Tortugas Hembras



Ejemplo Tortugas Hembras



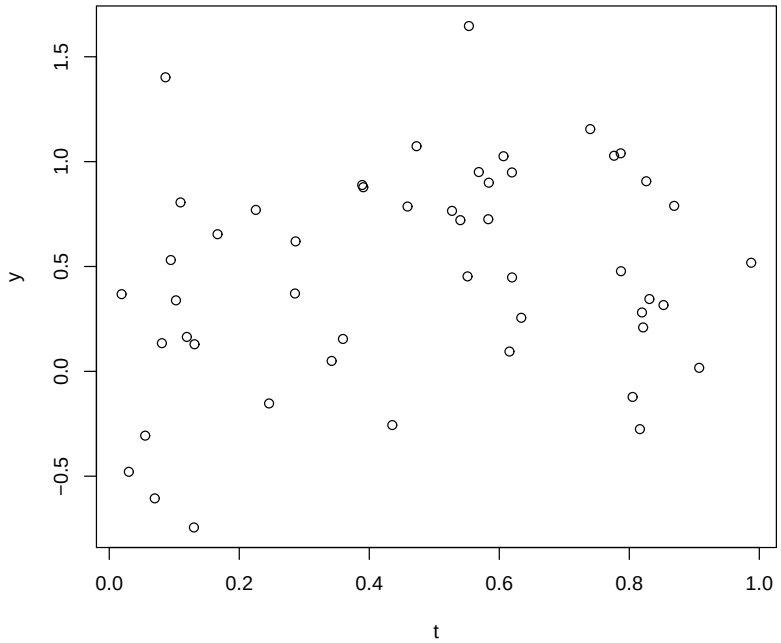
Ejemplo Tortugas Hembras



Regresión noparamétrica

$\{(t_i, y_i) : 1 \leq i \leq n\}$ vectores aleatorios independientes

$y_i = \eta(t_i) + \epsilon_i$



ϵ_i i.i.d. variables aleatorias, independientes de $\{t_i, i \geq 1\}$,

$$\mathbb{E}(\epsilon_i) = 0 \quad \text{and} \quad \mathbb{E}\epsilon_i^2 < \infty$$

Regresión noparamétrica

$\widehat{\eta}(t_0)$ se obtiene minimizando

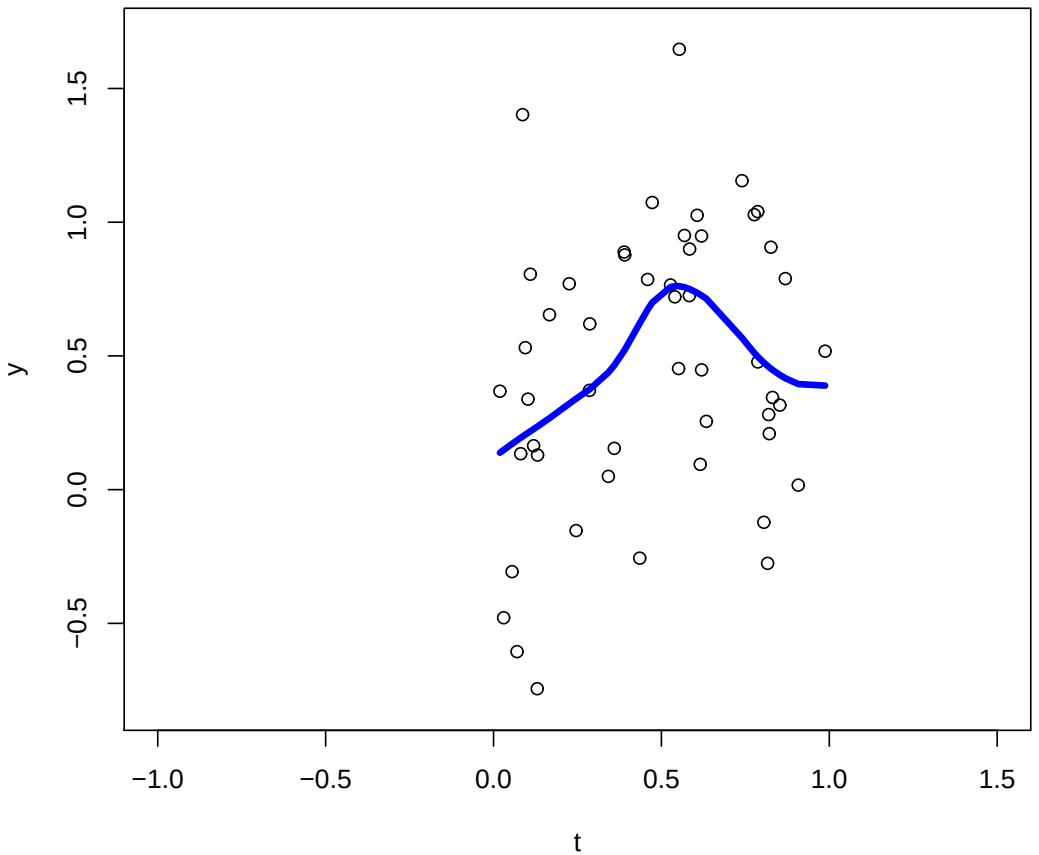
$$\sum_{i=1}^{n}(y_i - a)^2 w_i(t_0)$$

$$\mathbf{w_i(t) = \frac{K_{h_n}(t - t_i)}{\sum_{j=1}^{n} K_{h_n}(t - t_j)}}$$

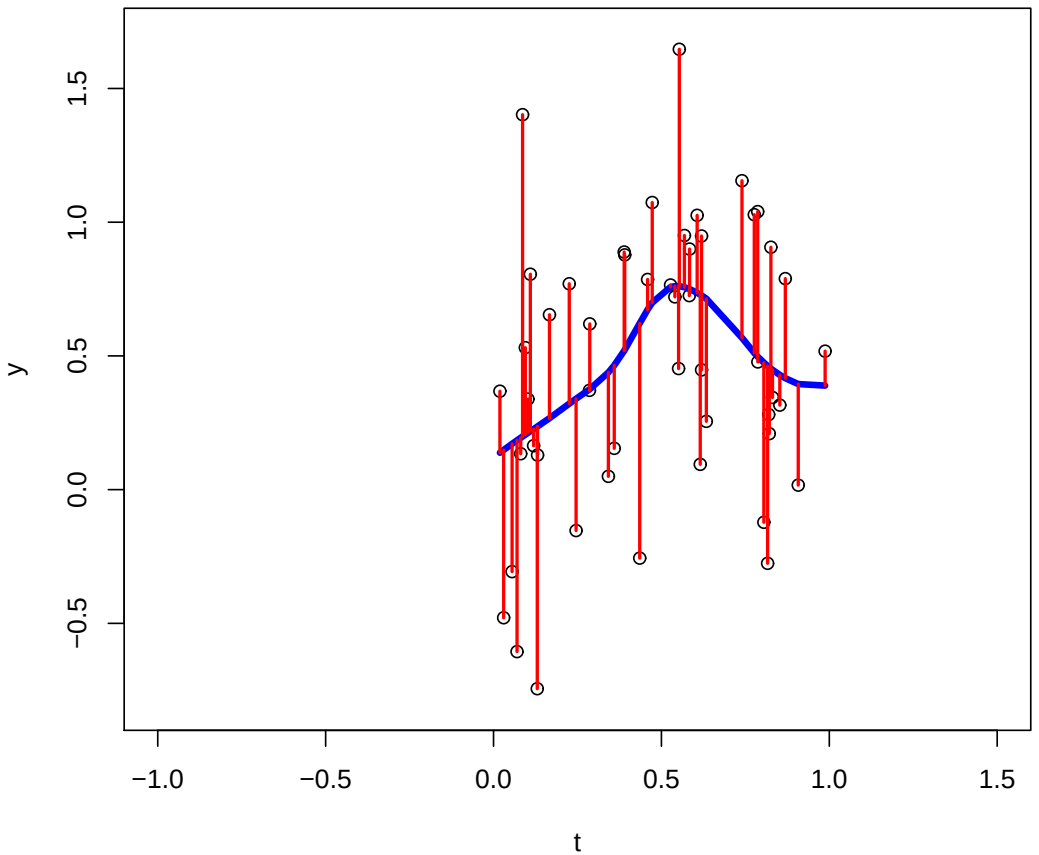
$$\mathbf{K_{h_n}(u) = \frac{1}{h_n} K\left(\frac{u}{h_n}\right)}$$

$$\mathbf{K \geq 0, \int K(u)\, du = 1}$$

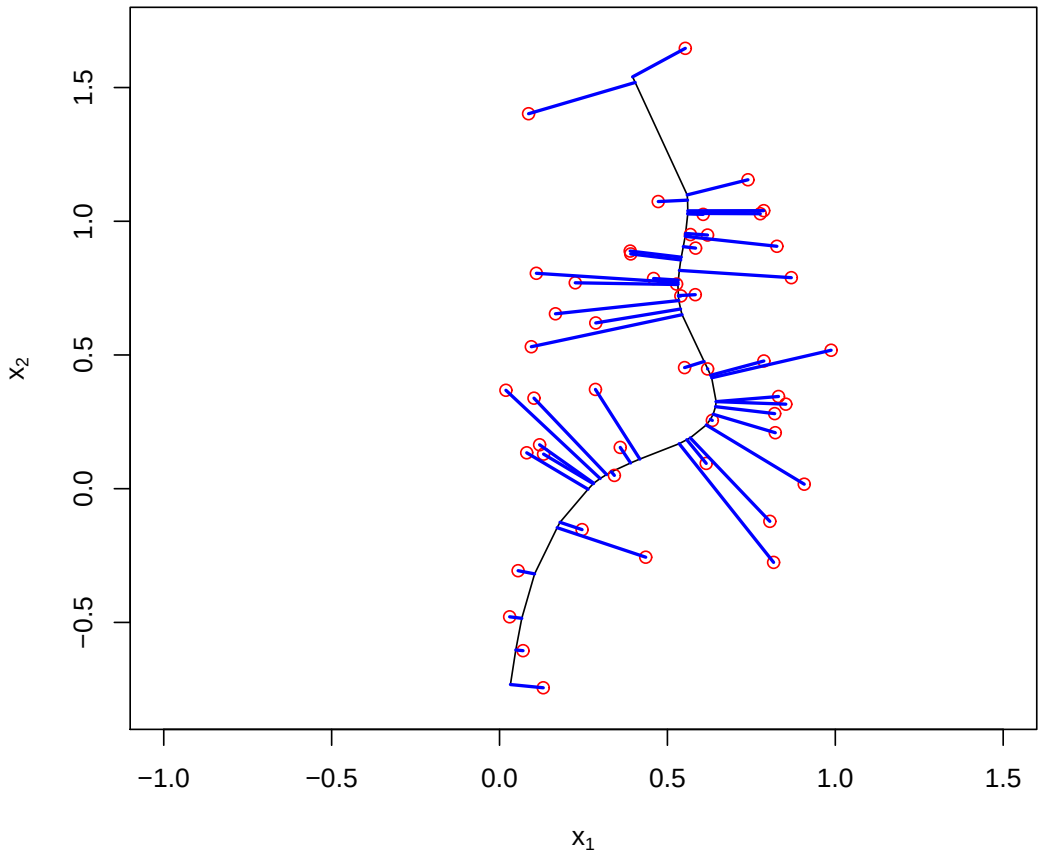
Regresión noparamétrica



Regresión noparamétrica



Curvas principales



Definimos las curvas principales como aquellas curvas que son *auto consistentes* para la distribución de $\mathbf{x}$ or para un conjunto de observaciones $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Es decir, que si agarramos un punto de la curva $\mathbf{z}$ y todos los datos $\mathbf{x}_i$ que se proyectan sobre $\mathbf{z}$, el promedio de esas observaciones es $\mathbf{z}$.

Observación

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independientes

$$\mathbf{x}_i = \boldsymbol{\mu} + \nu_i \boldsymbol{\gamma} + \mathbf{e}_i$$

- $\mathbb{E}(\mathbf{e}_i) = \mathbf{0}$
- $\mathrm{Cov}(\mathbf{e}_i) = \mathbf{I}_p$

entonces el estimador de mínimos cuadrados de $\boldsymbol{\gamma}$ es la primer componente principal, o sea,

$$\widehat{\boldsymbol{\gamma}}_1 = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu} - \nu_i \mathbf{a}\|^2$$

Observación

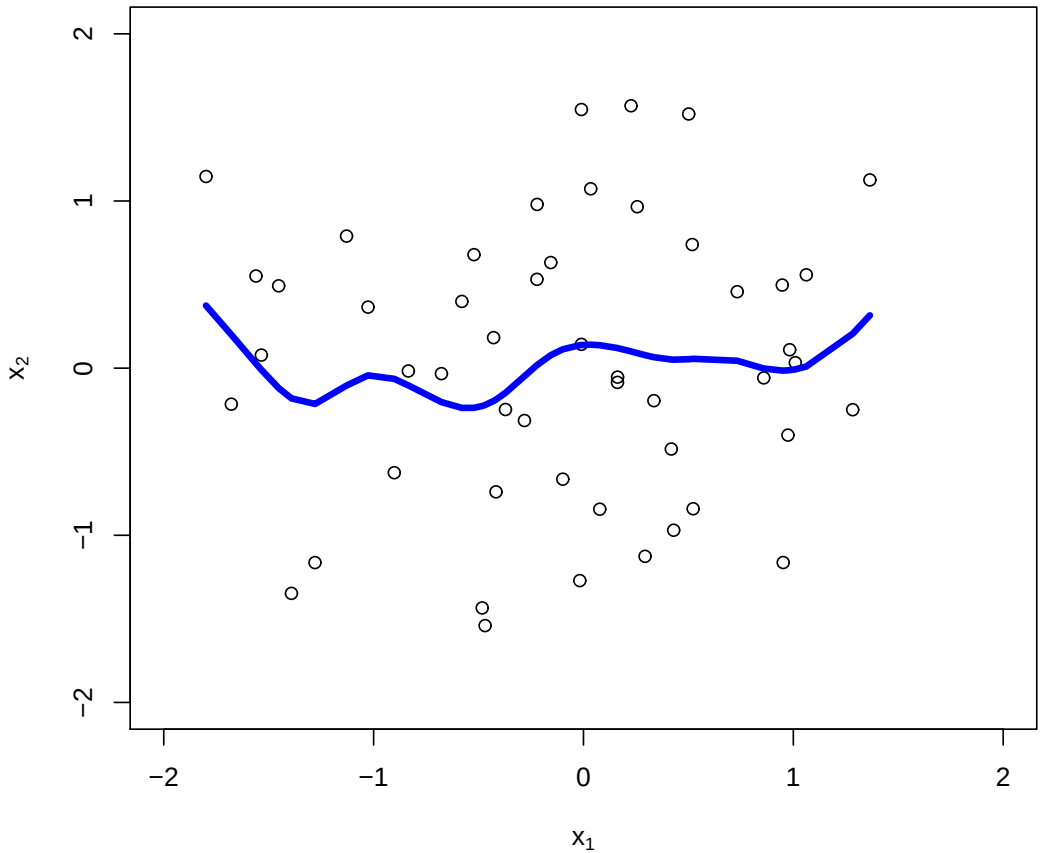
$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independientes

Una generalización del problema anterior es considerar el modelo

$$\mathbf{x}_i = \mathbf{f}(\nu_i) + \mathbf{e}_i$$

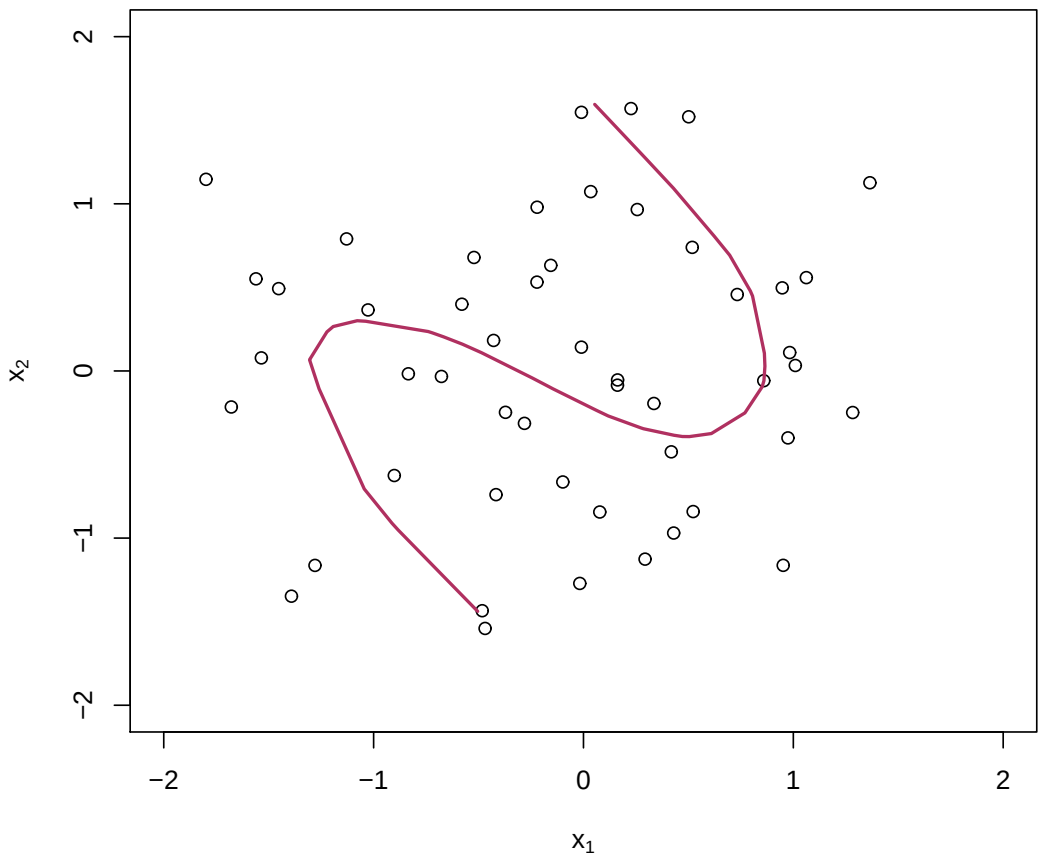
donde $\mathbf{f}(\nu) = (f_1(\nu), \ldots, f_p(\nu))^{\mathrm{T}}$ con f_j funciones suaves.

$$f_1(\nu) = \cos(\nu),\ f_2(\nu) = \sin(\nu),\ \nu \in [0, 2\pi]$$

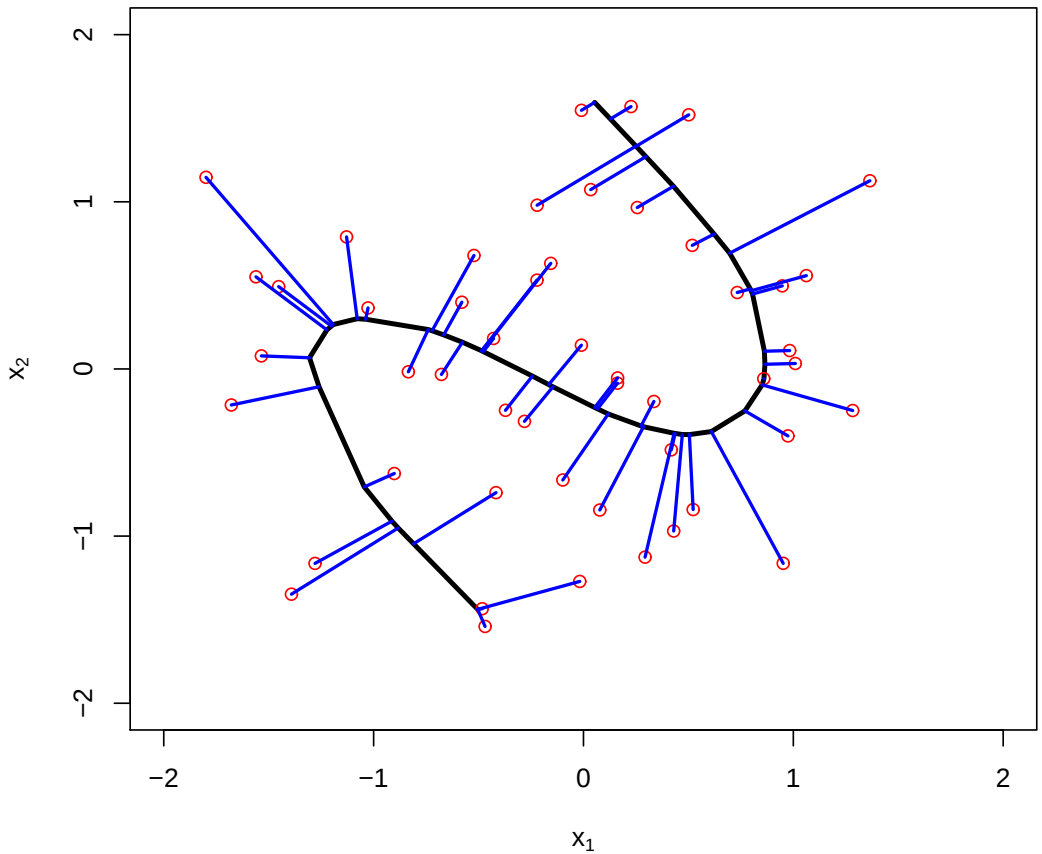


Ajuste no paramétrico

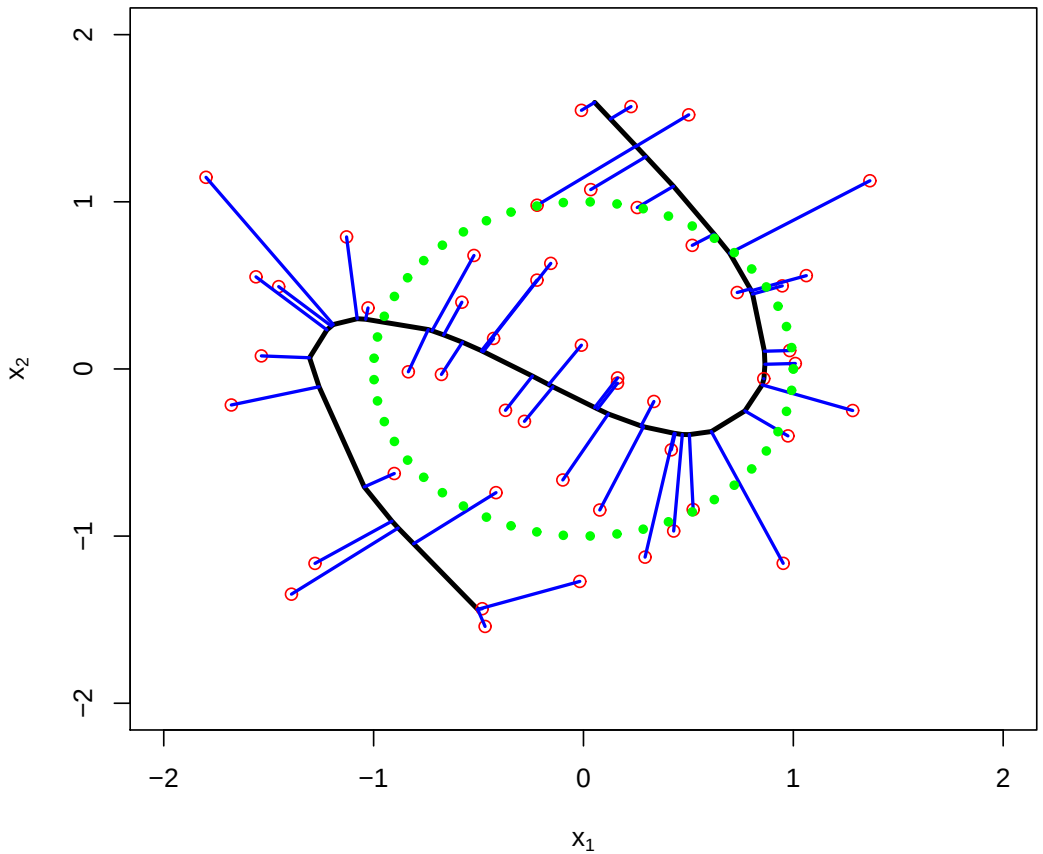
Curvas principales



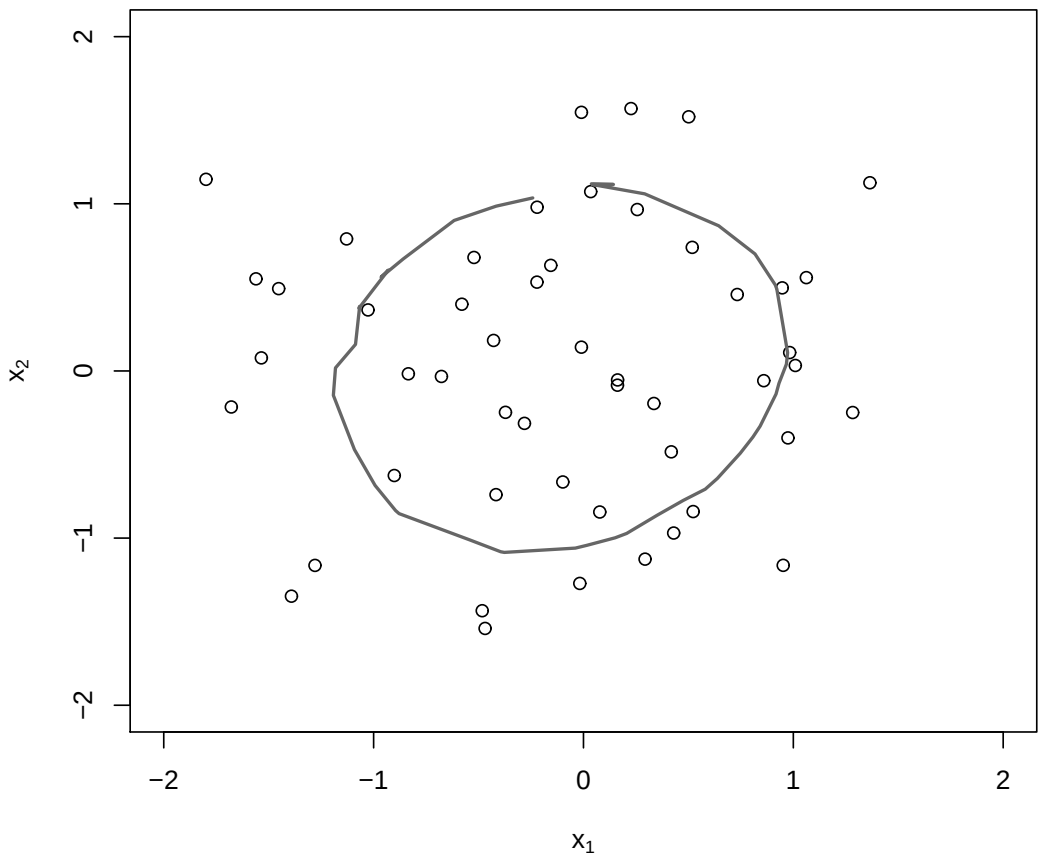
Curvas principales



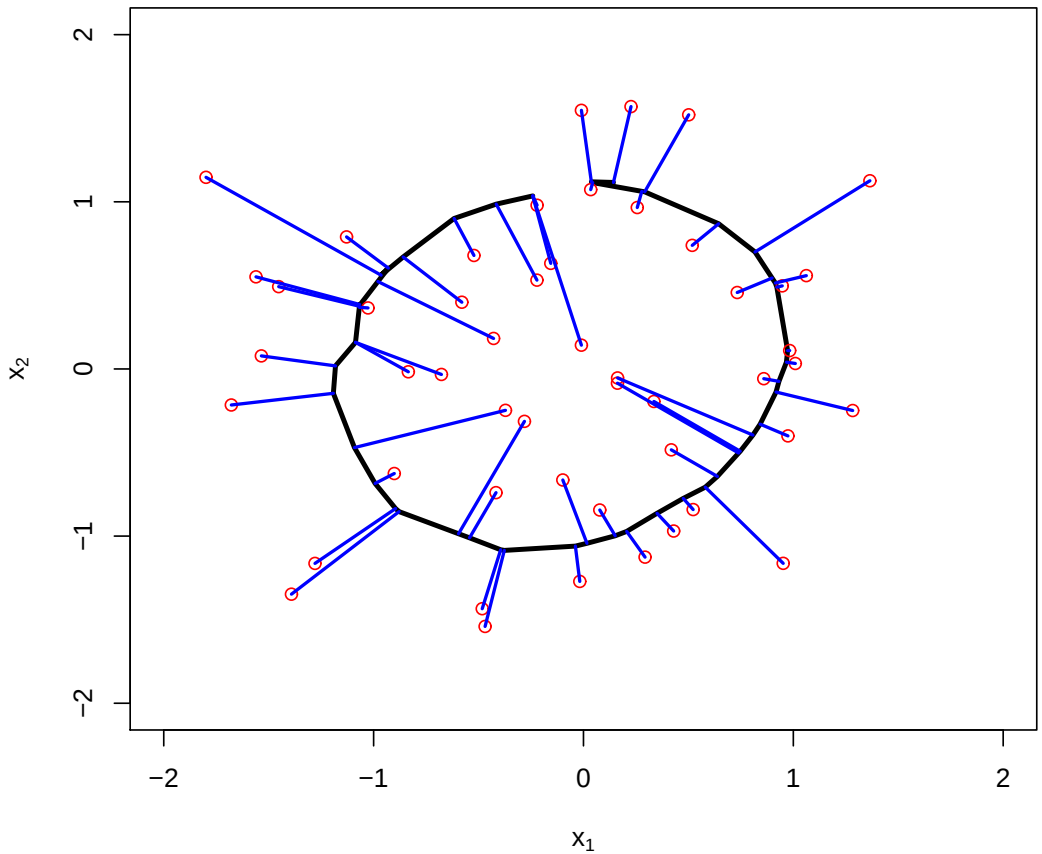
Curvas principales



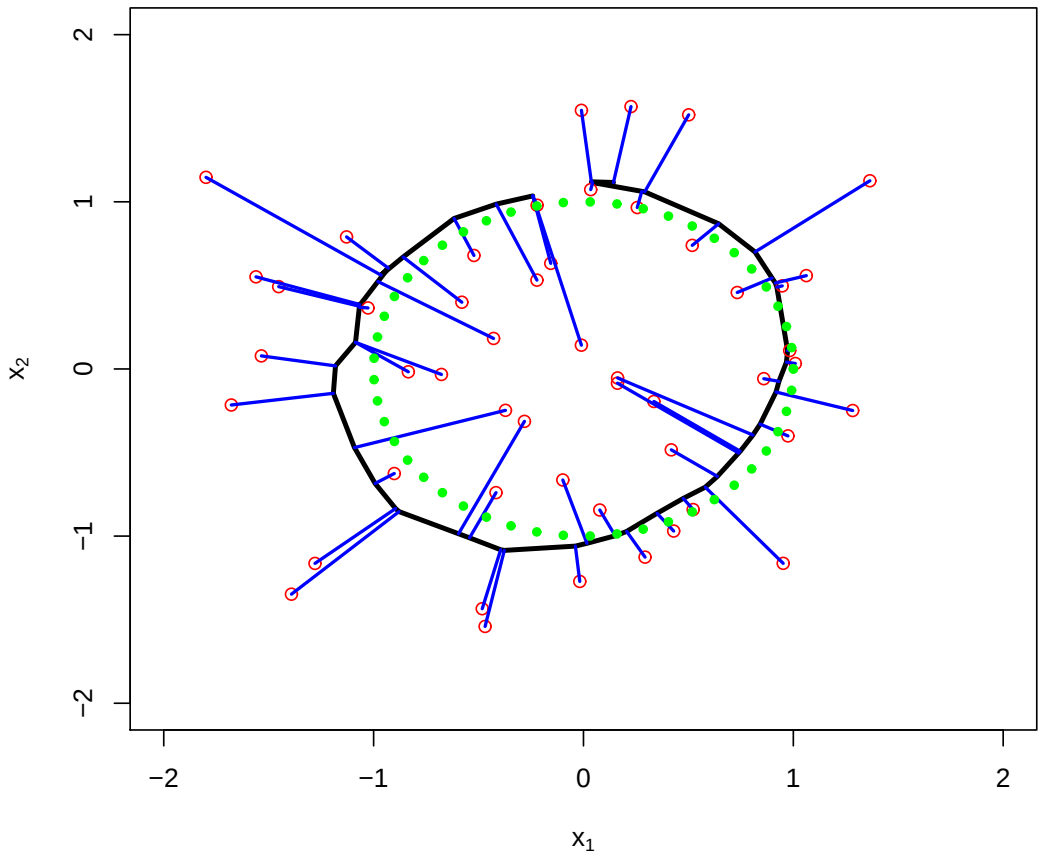
Curvas principales periódicas



Curvas principales periódicas



Curvas principales periódicas



Preliminares

- Una curva unidimensional en $\mathbb{R}^p$ es un vector $\mathbf{f}(\nu)$ de p funciones $f_1(\nu), \ldots, f_p(\nu)$ que dependen de una variable $\nu \in \mathcal{I} \subset \mathbb{R}$.
- $f_j(\nu)$ se llaman las funciones coordenadas
- ν provee un orden a lo largo de la curva.

- Si $h : \mathcal{I} \to \mathcal{J}$ es monótona podemos transformar ν por $\lambda = h(\nu)$
- Las funciones $f_j(\nu)$ se transforman como $f_j(h^{-1}(\lambda))$
- Existe una parametrización natural dada por el arco de curva

Preliminares

- El arco de una curva $\mathbf{f}$ de ν_0 a ν_1 se define como

$$\ell = \int_{\nu_0}^{\nu_1} \|\mathbf{f}'(z)\| dz$$

- $\mathbf{f}'(\nu)$ es la tangente en ν

- Una curva tal que $\|\mathbf{f}'(\nu)\| \equiv 1$ se dice una curva parametrizada con unidad de velocidad 1.

- Suavidad de la curva en tal caso se traduce en el aspecto visual de $\{\mathbf{f}(\nu) : \nu \in \mathcal{I}\}$ ya que no debe haber picos.

Preliminares

- El vector $\mathbf{f}''(\nu)$ es la aceleración de la curva.
- Si la curva tiene unidad de velocidad 1 entonces $\mathbf{f}''(\nu)$ es ortogonal al vector tangente $\mathbf{f}'(\nu)$.
- En ese caso,

$$\frac{\mathbf{f}''(\nu)}{\|\mathbf{f}''(\nu)\|}$$

se llama la normal principal de la curva en ν.

Preliminares

- Los vectores $\mathbf{f}''(\nu)$ y $\mathbf{f}'(\nu)$ generan un plano.
- Hay un único círculo de velocidad 1 en ese plano que pasa por $\mathbf{f}(\nu)$ y que tiene la misma velocidad y aceleración en $\mathbf{f}(\nu)$ que la curva.
- El radio $r_{\mathbf{f}}(\nu)$ de ese círculo se llama el radio de curvatura de la curva $\mathbf{f}$ en ν.

$$r_{\mathbf{f}}(\nu) = \frac{1}{\|\mathbf{f}''(\nu)\|}$$

- El centro $\mathbf{c}_{\mathbf{f}}(\nu)$ del círculo se llama el centro de curvatura de la curva $\mathbf{f}$ en ν.

Definición

- Sea $\mathbf{x} \in \mathbb{R}^p$ un vector aleatorio con densidad y tal que $\mathbb{E}(\mathbf{x}) = \mathbf{0}$, $\textsc{Var}(\mathbf{x}) = \mathbf{\Sigma}$ existe.

- Sea $\mathbf{f}$ una curva suave (C^∞) de unidad de velocidad 1 parametrizada sobre $\mathcal{I} \subset \mathbb{R}$

- Supondremos que
  - $\mathcal{I}$ es cerrado (puede ser infinito)
  - la curva no se interseca, o sea,

$$\nu_1 \neq \nu_2 \Longrightarrow \mathbf{f}(\nu_1) \neq \mathbf{f}(\nu_2)$$

Definición

- Definimos el índice de proyección $\nu_{\mathbf{f}} : \mathbb{R}^p \to \mathbb{R}$ como

$$\nu_{\mathbf{f}}(\mathbf{x}) = \sup\{\nu : \nu = \underset{\lambda}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{f}(\lambda)\|\}$$

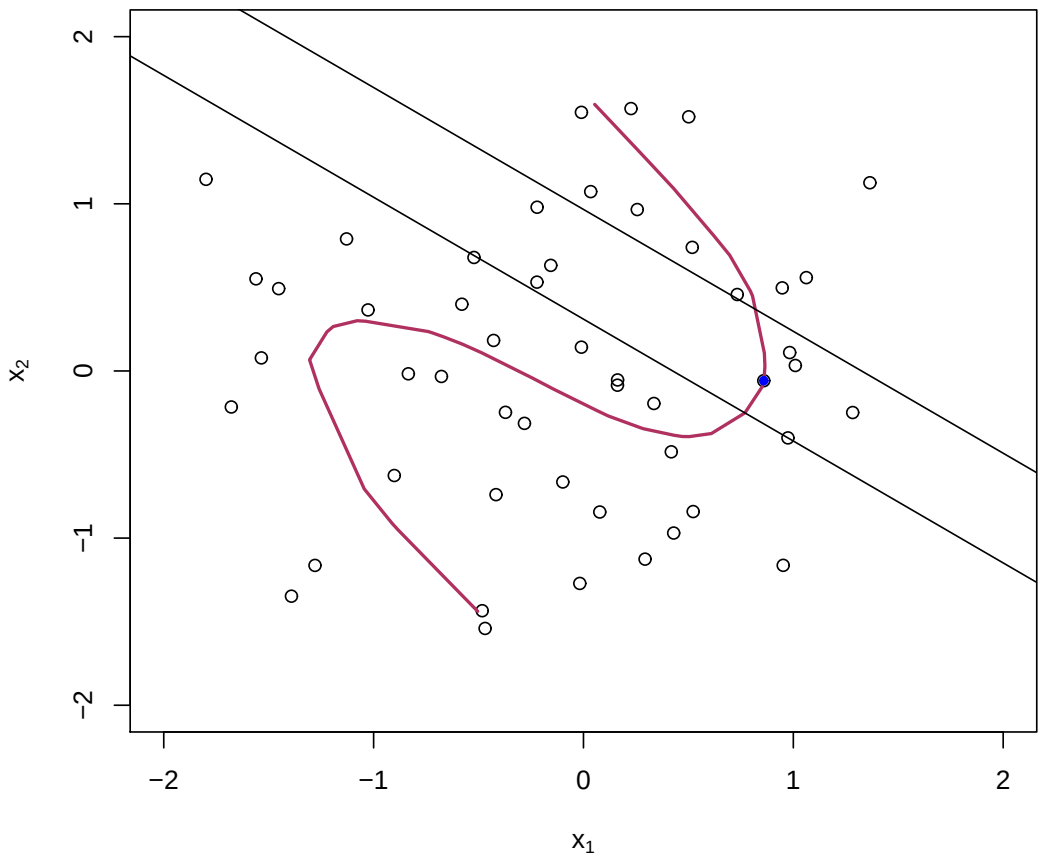
  El índice de proyección de $\mathbf{x}$ es el valor de ν para el cual $\mathbf{f}(\nu)$ está lo más cerca de $\mathbf{x}$. Si hay varios valores elegimos el mayor.

- $\nu_{\mathbf{f}}(\mathbf{x})$ está bien definido y es medible

Definición 1. La curva se dice auto–consistente o curva principal de $\mathbf{x}$ si

$$\mathbb{E}\left(\mathbf{x} \middle| \nu_{\mathbf{f}}(\mathbf{x}) = \nu\right) = \mathbf{f}(\nu) \quad \text{p.p. } \nu\,.$$

Definición



Preguntas

- Para qué tipo de distribuciones existen las curvas principales?
- Cuántas curvas principales hay para una distribución dad?
- Cuáles son sus propiedades?
- Delicado, P. (2001). Another Look at Principal Curves and Surfaces. *Journal of Multivariate Analysis*, **77**, 84-116.
- Hastie, T. and W. Stuetzle (1989) Principal curves. *Journal of the American Statistical Association*, **84**, 502-516.
- Sea $\mathbf{x} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$, $\mathrm{VAR}(\mathbf{x}) = \boldsymbol{\Sigma}$. Sea $\|\mathbf{a}\| = 1$, $\mathbf{x}_0 \in \mathbb{R}^p$ es un punto fijo y definamos $\mathbf{y} = \mathbf{x}_0 + \mathbf{a}\,\mathbf{a}^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0)$.
  Si $\mathbf{y}$ es auto–consistente para $\mathbf{x}$ entonces es una curva principal.

Preguntas

- Para distribuciones elípticas las componentes principales son curvas principales

- Para distribuciones esféricas $\mathcal{E}(\boldsymbol{\mu}, \mathbf{I}_p)$, o sea, su densidad es

$$a\, h(\|\mathbf{x} - \boldsymbol{\mu}\|)$$

  - Cualquier recta que pasa por $\boldsymbol{\mu}$ es una curva principal
  - Si $p = 2$ y $\boldsymbol{\mu} = \mathbf{0}$, el círculo de centro en el origen y radio $\mathbb{E}(\|\mathbf{x}\|)$ es una curva principal.
  - Si $\mathbf{x} \sim N_2(\mathbf{0}, \mathbf{I}_2)$, el círculo de centro en el origen y radio $\sqrt{\pi/2}$ es una curva principal.

- Si $\mathbf{x} = \mathbf{f}(\nu) + \mathbf{e}$ con $\mathbb{E}(\mathbf{e}) = \mathbf{0}$, puede haber casos en los que la curva principal no sea $\mathbf{f}$.

Una propiedad de distancias

Sea $d(\mathbf{x},)$ la distancia de $\mathbf{x}$ a su proyección en la curva $\mathbf{f}$

$$d(\mathbf{x},) = \|\mathbf{x} - \mathbf{f}(\nu_{\mathbf{f}}(\mathbf{x})\|$$

Si $\mathbf{x} \sim P$, definamos

$$D^2(P, \mathbf{f}) = \mathbb{E}d^2(\mathbf{x}, \mathbf{f})$$

- Consideremos la recta $\mathbf{f}(\nu) = \mathbf{x}_0 + \nu\,\mathbf{v}$ y recordemos que $\mathbb{E}(\mathbf{x}) = \mathbf{0}$
- La distancia $D^2(P,) = \mathbf{D^2(P, x_0, v)}$
- El gradiente de $D^2(P, \mathbf{x}_0, \mathbf{v}) = \mathbf{0}$ si y sólo si $\mathbf{u} = \mathbf{0}$ y $\mathbf{v}$ es un autovector de $\mathbf{\Sigma}$, o sea, la recta es la asociada a una componente principal.

Una propiedad de distancias

Sea $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\| = 1$ y supongamos que $\mathbf{x}_0^{\mathrm{T}}\mathbf{v} = 0$ sin pérdida de generalidad. Entonces,

$$D^2(P, \mathbf{x}_0, \mathbf{v}) = \mathrm{TR}(\mathbf{\Sigma}) + \|\mathbf{x}_0\|^2 - \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} - (\mathbf{x}_0^{\mathrm{T}}\mathbf{v})^2 = \mathrm{TR}(\mathbf{\Sigma}) + \|\mathbf{x}_0\|^2 - \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v}$$

Luego, para calcular el gradiente tenemos que agregar los vínculos $\|\mathbf{v}\| = 1$ y $\mathbf{x}_0^{\mathrm{T}}\mathbf{v} = 0$ es decir, debemos derivar

$$\Lambda(\mathbf{x}_0, \mathbf{v}) = D^2(P, \mathbf{x}_0, \mathbf{v}) + \lambda_1(\mathbf{v}^{\mathrm{T}}\mathbf{v} - 1) + \lambda_2\mathbf{x}_0^{\mathrm{T}}\mathbf{v}$$

Una propiedad de distancias

$$\frac{\partial}{\partial \mathbf{x}_0}\Lambda(\mathbf{x}_0, \mathbf{v}) = 2\mathbf{x}_0 + \lambda_2 \mathbf{v} \quad (1)$$

$$\frac{\partial}{\partial \mathbf{v}}\Lambda(\mathbf{x}_0, \mathbf{v}) = -2\mathbf{\Sigma}\mathbf{v} + \lambda_1 \mathbf{v} \quad (2)$$

Multiplicando (1) por $\mathbf{v}^{\mathrm{T}}$ e igualando a 0 obtenemos que $\lambda_2 = 0$ pues $\mathbf{x}_0^{\mathrm{T}}\mathbf{v} = 0$, de donde por (1) se deduce que $\mathbf{x}_0 = 0$

La ecuación (2) implica que si $\frac{\partial}{\partial \mathbf{v}}\Lambda(\mathbf{x}_0, \mathbf{v}) = 0$ entonces $\mathbf{\Sigma}\mathbf{v} = c\mathbf{v}$, o sea $\mathbf{v}$ es un autovector de $\mathbf{\Sigma}$.

Una propiedad de distancias

Para extenderlo a curvas principales, sea

$$\mathcal{G} = \text{el conjunto de las curvas parametrizadas sobre } \mathcal{I}$$

Para $g \in \mathcal{G}$ sea

$$\mathbf{f}_t = \mathbf{f} + t\mathbf{g}$$

Definición 2. La curva $\mathbf{f}$ se dice un punto crítico de la distancia de variación en la clase $\mathcal{G}$ si

$$\frac{\partial D^2(P, \mathbf{f}_t)}{\partial t}\Big|_{t=0} = 0 \qquad \forall \mathbf{g} \in \mathcal{G}$$

Una propiedad de distancias

Sea $\mathcal{G}_B$ la clase de las funciones suaves (C^∞) parametrizadas sobre $\mathcal{I}$ tales que

$$\|\mathbf{g}\|_\infty^2 = \sup_\nu \sum_{j=1}^{p} g_j^2(\nu) \leq 1 \qquad \|\mathbf{g}'\|_\infty^2 \leq 1\,.$$

Entonces, $\mathbf{f}$ es una curva principal para $\mathbf{x}$ si y sólo si $\mathbf{f}$ es un punto crítico de la distancia de variación en la clase $\mathcal{G}_B$.

La demostración se basa en probar que se puede intercambiar la derivada con la esperanza y usar que

$$\frac{\partial D^2(P, \mathbf{f}_t)}{\partial t}\Big|_{t=0} = -2\,\mathbb{E}\left\{\left[\mathbb{E}\left(\mathbf{x}|\boldsymbol{\nu}_\mathbf{f}(\mathbf{x})\right) - \mathbf{f}\left(\boldsymbol{\nu}_\mathbf{f}(\mathbf{x})\right)\right]\mathbf{g}\left(\boldsymbol{\nu}_\mathbf{f}(\mathbf{x})\right)\right\}$$

Algoritmo

Estamos interesados en encontrar curvas suaves asociadas a puntos críticos de $D^2(P, \mathbf{f}_t)$. La estrategia del algoritmo consiste en

- tomar una curva suave inicial, que será la asociada a la mayor componente principal lineal
- chequear si es una curva principal.

  Para ello, se proyectan los datos sobre la curva y evaluamos la esperanza condicional sobre los puntos proyectados.

  Si la esperanza condicional coincide con la curva, ya está.

  Sino, buscamos una nueva curva y seguimos.

Algoritmo

- Sea $\mathbf{f}^{(0)}(\nu) = \boldsymbol{\mu} + \nu\,\boldsymbol{\gamma}_1$ donde $\boldsymbol{\mu} = \mathbb{E}(\mathbf{x})$ y $\boldsymbol{\gamma}_1$ es el autovector de $\boldsymbol{\Sigma}$ asociado a su mayor autovalor.

  Sea $\nu^{(0)}(\mathbf{x}) = \nu_{\mathbf{f}^{(0)}}(\mathbf{x})$.

- Repita sobre el índice de iteración j los pasos a) a d)
  a) Sea $\mathbf{f}^{(j)}(\nu) = \mathbb{E}\left(\mathbf{x}\middle|\nu_{\mathbf{f}^{(j-1)}}(\mathbf{x}) = \nu\right)$
  b) Defina $\nu^{(j)}(\mathbf{x}) = \nu_{\mathbf{f}^{(j)}}(\mathbf{x})$
  c) Transforme $\nu^{(j)}(\mathbf{x})$ de modo que $\mathbf{f}^{(j)}$ sea de velocidad unitaria.
  d) Calcule

$$D^2(P, \mathbf{f}^{(j)}) = \mathbb{E}\|\mathbf{x} - \mathbf{f}^{(j)}(\nu^{(j)}(\mathbf{x})\|^2 = \mathbb{E}\left\{\mathbb{E}\left(\|\mathbf{x} - \mathbf{f}^{(j)}(\nu^{(j)}(\mathbf{x})\|^2\middle|\nu^{(j)}(\mathbf{x})\right)\right\}$$

- Hasta que la variación en $D^2(P, \mathbf{f}^{(j)})$ esté por debajo de un umbral.

Algoritmo

Si $\mathbf{x}_1, \ldots, \mathbf{x}_n$ son i.i.d., se reemplaza P por la empírica y la esperanza condicional por un suavizado, resultando el algoritmo.

- Sea $\mathbf{f}^{(0)}(\nu) = \overline{\mathbf{x}} + \nu\,\widehat{\gamma}_1$ donde $\widehat{\gamma}_1$ es el autovector de $\mathbf{S}$ asociado a su mayor autovalor.

  Sea $\nu^{(0)}(\mathbf{x}) = \nu_{\mathbf{f}^{(0)}}(\mathbf{x})$.

- Repita sobre el índice de iteración j los pasos a) a d)
  a) Sea $\mathbf{f}^{(j)}(\nu)$ un smoother (puede ser un polinomio local o basado en splines) de
     - *las respuestas* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ versus
     - los valores observados de la covariable $\nu_{\mathbf{f}^{(j-1)}}(\mathbf{x}_1), \ldots, \nu_{\mathbf{f}^{(j-1)}}(\mathbf{x}_n)$
  b) Defina $\nu^{(j)}(\mathbf{x}) = \nu_{\mathbf{f}^{(j)}}(\mathbf{x})$
  c) Transforme $\nu^{(j)}(\mathbf{x})$ de modo que $\mathbf{f}^{(j)}$ sea de velocidad unitaria.

Algoritmo ···

d) Calcule

$$D^2(P, \mathbf{f}^{(j)}) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{f}^{(j)}(\nu^{(j)}(\mathbf{x}_i)\|^2$$

- Hasta que la variación en $D^2(P, \mathbf{f}^{(j)})$, es decir,

$$\frac{|D^2(P, \mathbf{f}^{(j-1)}) - D^2(P, \mathbf{f}^{(j)})|}{D^2(P, \mathbf{f}^{(j-1)})}$$

esté por debajo de un umbral que en el algorimto por default es 0.001.

Algoritmo

Un punto clave en el algoritmo es el calculo de $\nu_{\mathbf{f}}(\mathbf{x}_i)$ para $\mathbf{f}(\cdot) = \mathbf{f}^{(j)}(\cdot)$ fija.

Queremos buscar para cada $1 \leq i \leq n$, el valor $\nu_i = \nu_{\mathbf{f}}(\mathbf{x}_i) = \nu_{\mathbf{f}^{(j)}}(\mathbf{x}_i)$.

En la iteración $j-1$ calculamos $\nu_i^{(j-1)} = \nu_{\mathbf{f}^{(j-1)}}(\mathbf{x}_i)$ y podemos calcular $\mathbf{f}(\nu_i^{(j-1)})$.

Definamos d_{ik} como la distancia entre $\mathbf{x}_i$ y su punto más cercano en el segmento que une $\mathbf{f}(\nu_k^{(j-1)})$ con $\mathbf{f}(\nu_{k+1}^{(j-1)})$.

Algoritmo

Asociado a d_{ik} existe un valor $\nu_{ik} \in [\nu_k^{(j-1)}, \nu_{k+1}^{(j-1)}]$

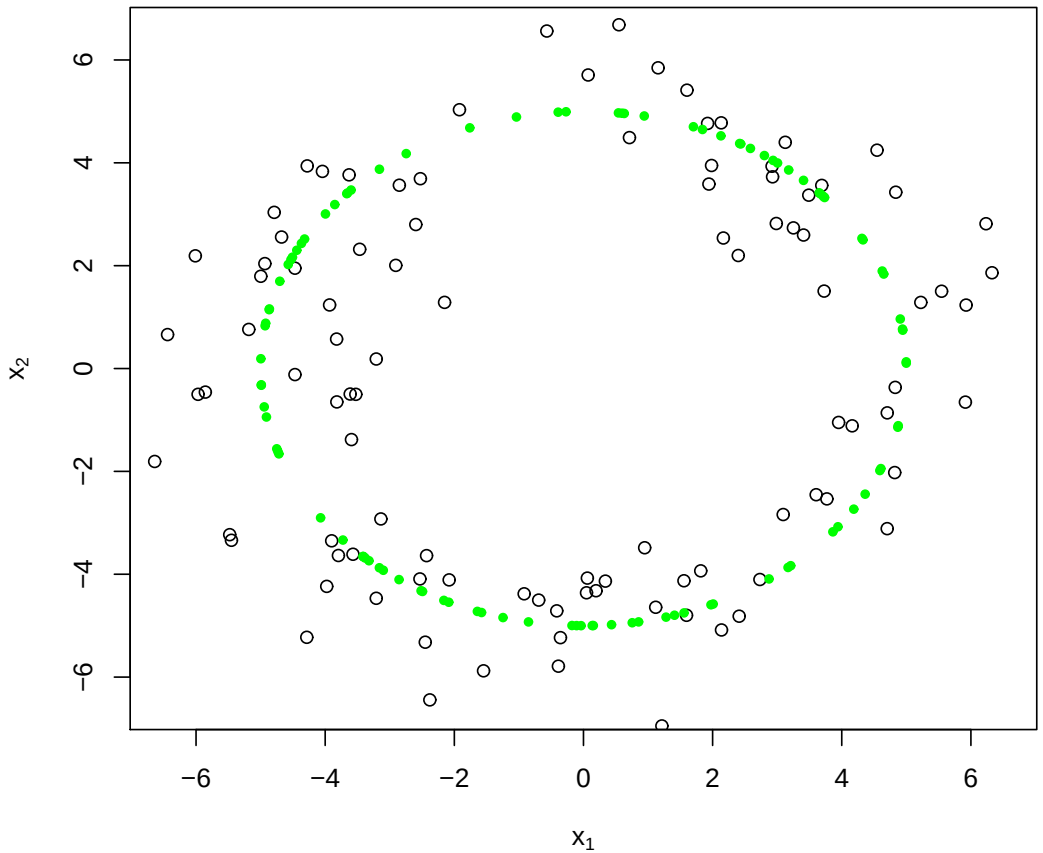
Definimos

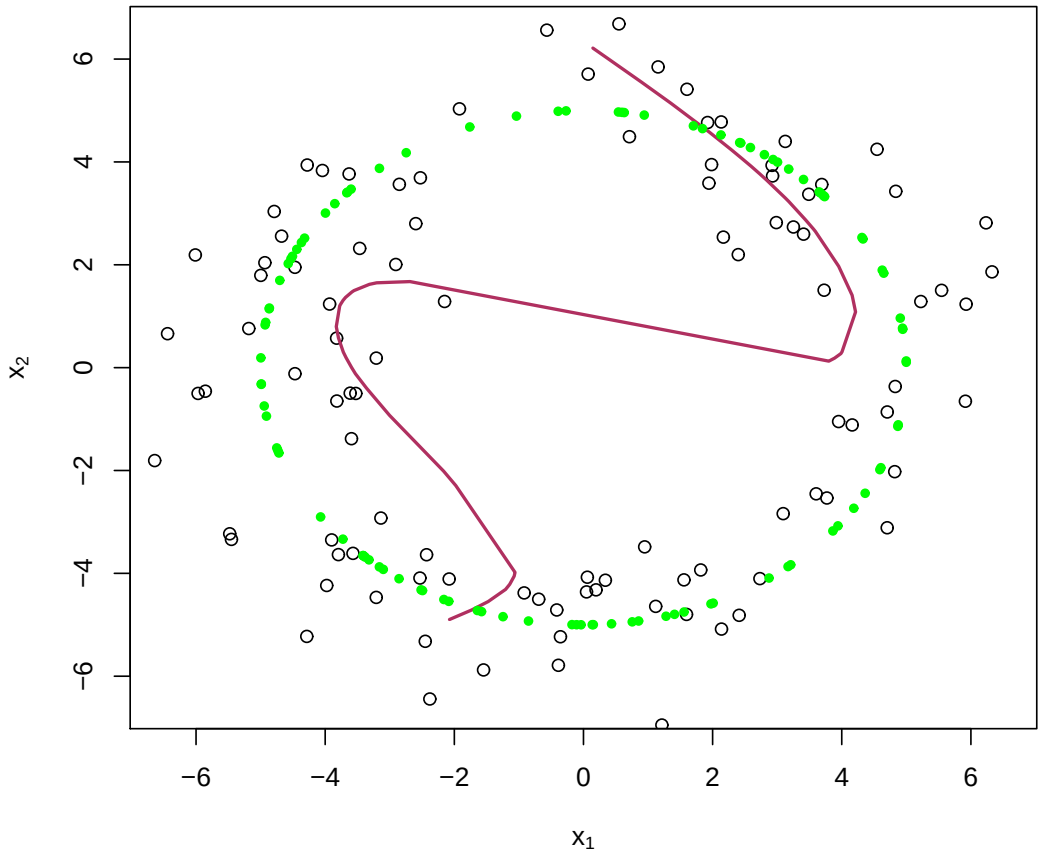
$$\nu_i = \nu_{ik^\star} \quad \text{donde} \quad d_{ik^\star} = \min_{1 \le k \le n-1} d_{ik}$$

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}_i \sim \mathbf{x}$ donde

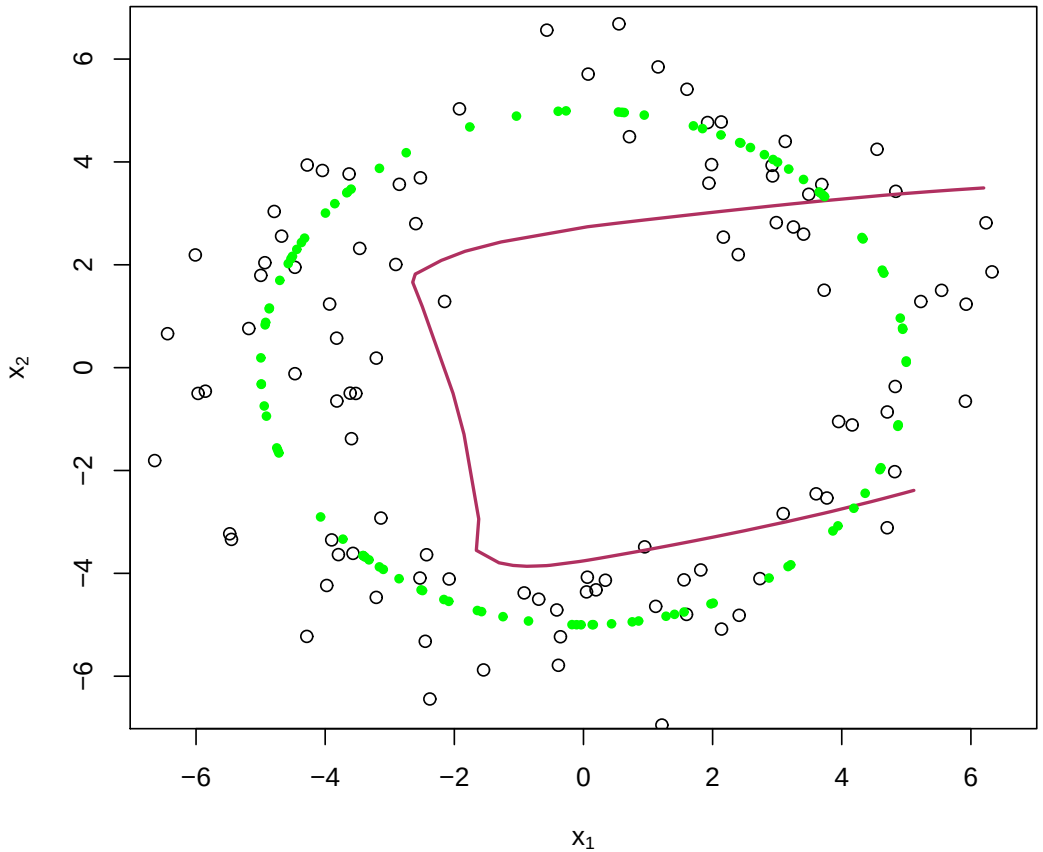
$$\mathbf{x} = 5 \begin{pmatrix} \sin(\nu) \\ \cos(\nu) \end{pmatrix} + \mathbf{e}$$

con $\nu \sim U[0, 2\pi)$ y $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_2)$.

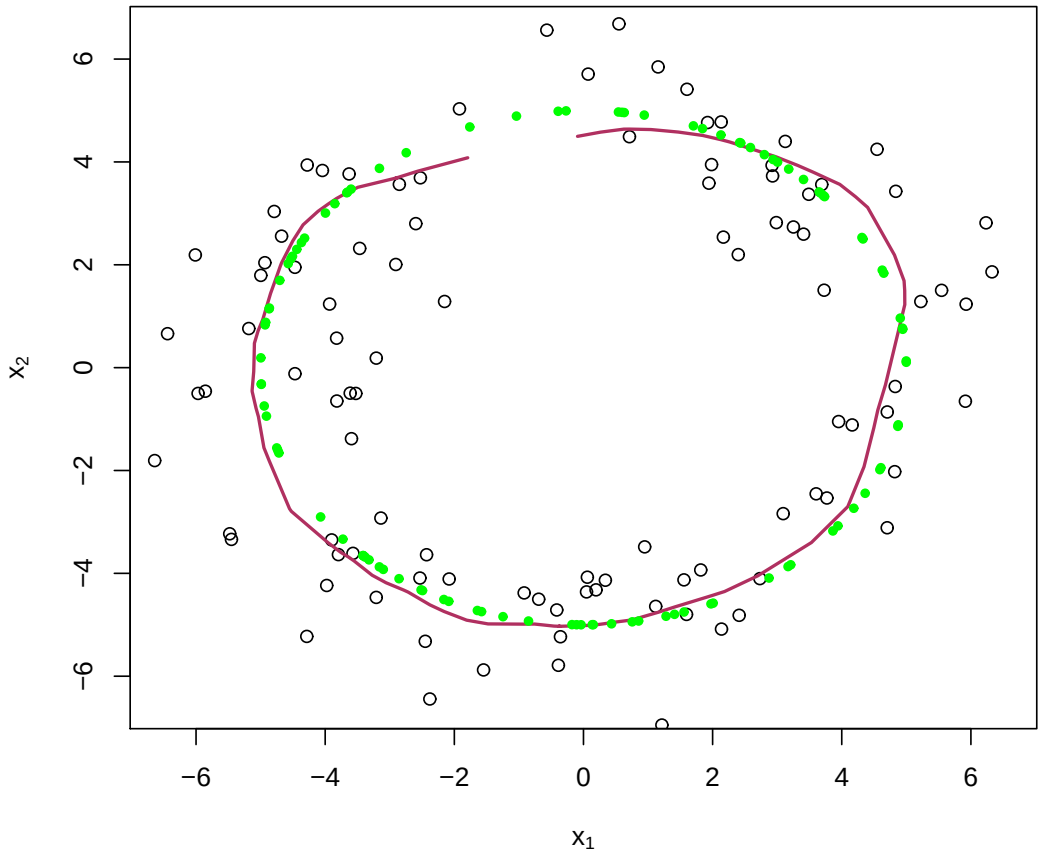

x2
x1
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6



Resultado de 10 iteraciones



Resultado de 10 iteraciones

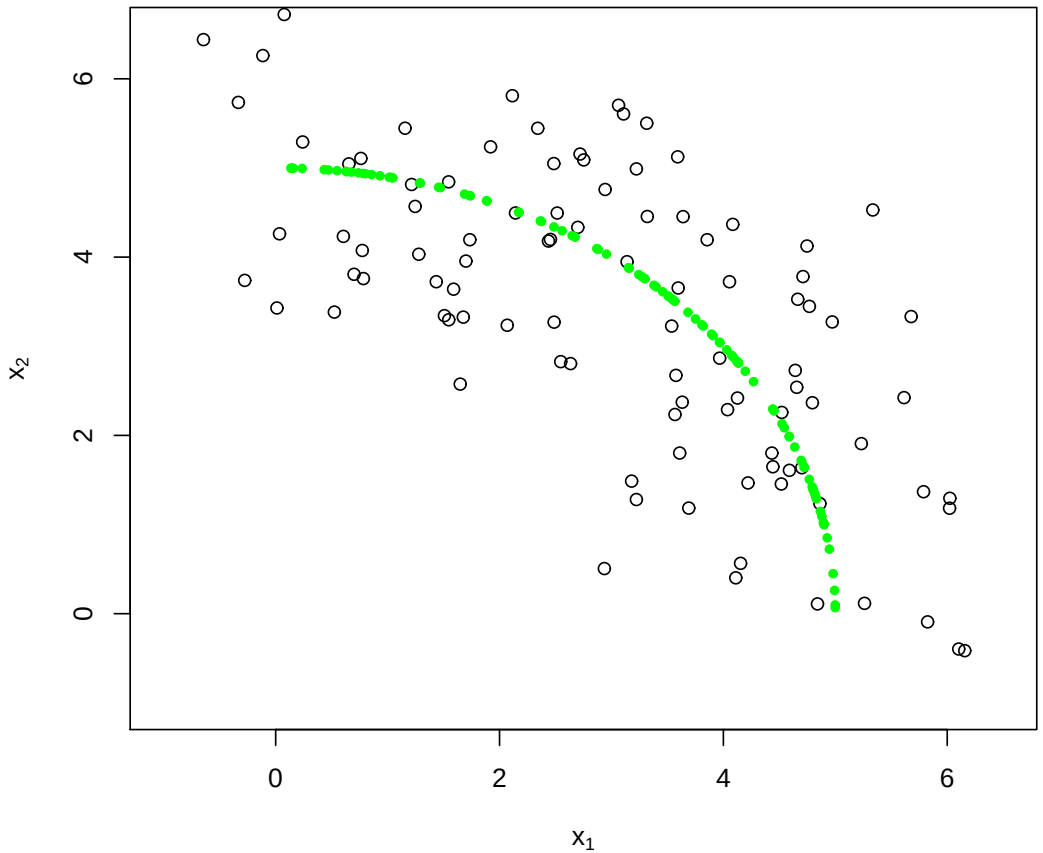


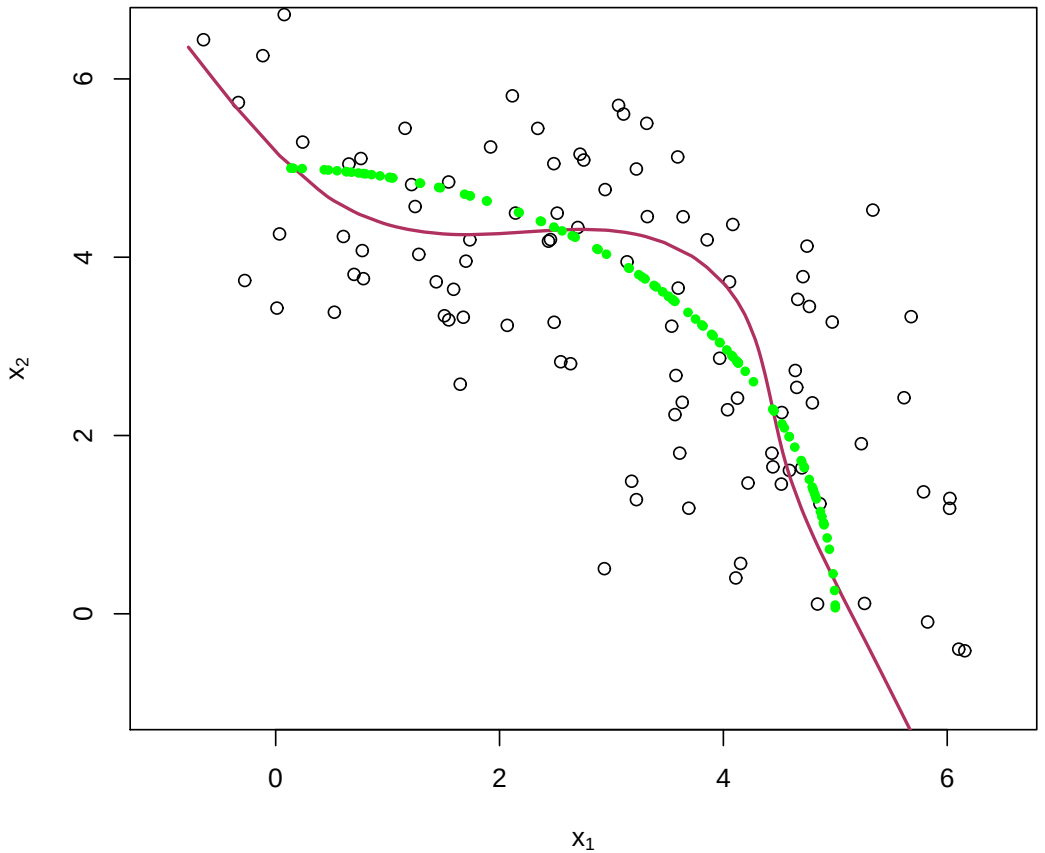
Resultado de 10 iteraciones

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}_i \sim \mathbf{x}$ donde

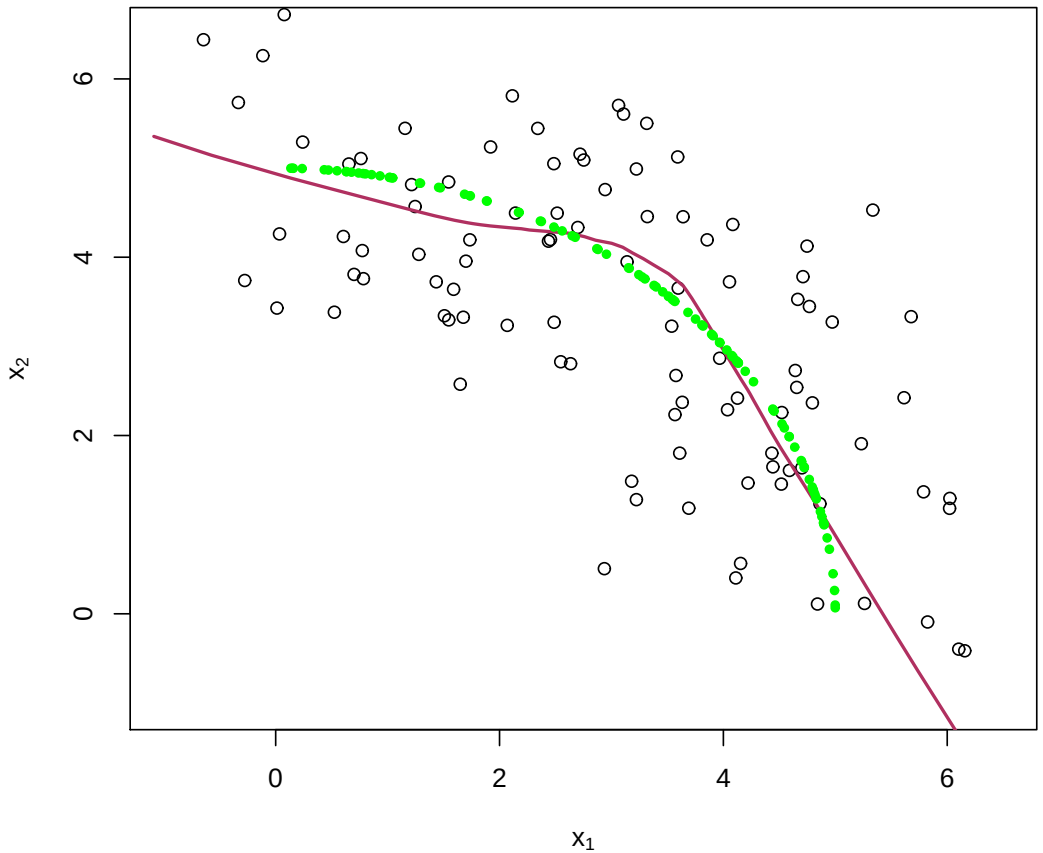
$$\mathbf{x} = 5 \begin{pmatrix} \sin(\nu) \\ \cos(\nu) \end{pmatrix} + \mathbf{e}$$

con $\nu \sim U[0, \pi/2)$ y $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_2)$.


x2
x1
0
2
4
6
0
2
4
6



Resultado de 10 iteraciones



Resultado de 10 iteraciones

Ejercicio: Repetir los ejemplos anteriores con 20, 100 y 200 iteraciones.

Qué observa?