

# Le spline di regressione

Data Mining

CLAMSES - University of Milano-Bicocca

Aldo Solari

# Riferimenti bibliografici

- Bowman, Evers. Lecture Notes on Nonparametric Smoothing. § 3.2.1, 3.2.2, 3.2.4
- AS § 4.4.1, 4.4.2
- LKA § 4.5, 4.7
- HTF § 5.1, 5.2, Appendix: B-splines

Si consideri il modello di regressione polinomiale

$$\mathbb{E}(y) = B\beta$$

dove  $y = (y_1, \dots, y_n)^t$  e

$$B = \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & \cdots & x_n^d \end{bmatrix}$$

La stima di  $\beta$  è data da

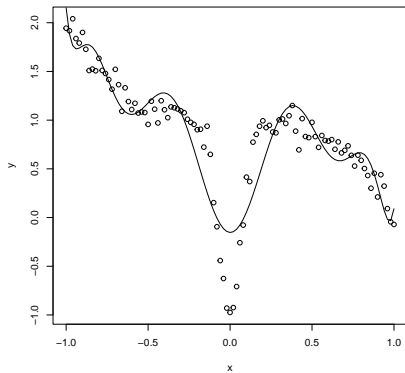
$$\hat{\beta} = (B^t B)^{-1} B^t y$$

Si stimi il modello per i dati simulati da

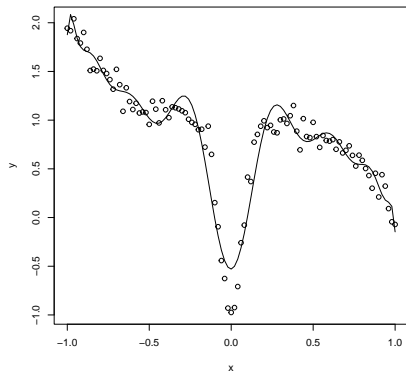
$$y_i = 1 - x_i^3 - 2 \exp(-100x_i^2) + \varepsilon_i, \quad i = 1, \dots, n$$

dove  $n = 101$ ,  $x = (-1, .98, \dots, 0.98, 1)^t$ , e  $\varepsilon_i \sim N(0, 0.1^2)$

Regressione polinomiale di grado 10



Regressione polinomiale di grado 17



La stima del modello è

$$\hat{y} = B\hat{\beta} = B(B^t B)^{-1} B^t y = S y$$

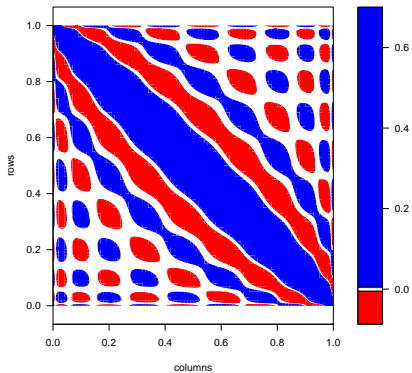
tuttavia la previsione  $\hat{y}_i$  in  $x_i$  non dipende solo da osservazioni vicine a  $x_i$ .

In altre parole, la matrice di liscio  $S$  non ha una banda diagonale, quindi la regressione polinomiale non è “locale”

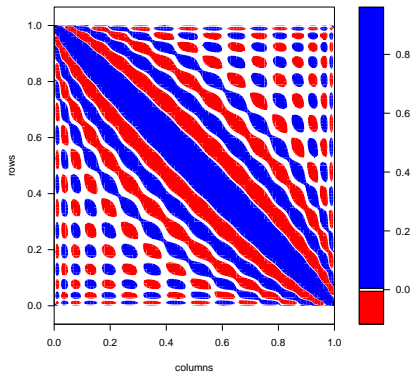
Un altro problema problema è la stima alle estremità, con curvatura molto elevata che in genere non è supportata dai dati

Infine, il *condition number* per  $B^t B$  con la regressione polinomiale di grado 17 è  $1.56 \cdot 10^{12}$

Regressione polinomiale di grado 10



Regressione polinomiale di grado 17



# Polinomio a tratti

- Si dividono i dati in sottoinsiemi determinati da valori detti nodi (*knots*) e si stima un modello polinomiale all'interno di quel sottoinsieme di dati
- Si specificano  $K$  nodi interni (*internal knots*) nel *range* di  $x$ :

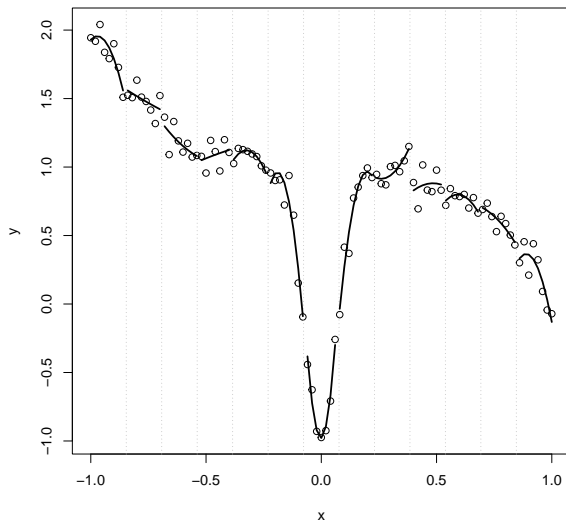
$$\min(x) < \xi_1 < \dots < \xi_K < \max(x)$$

che definiscono  $K + 1$  intervalli

- Stima un modello polinomiale di grado  $M$  su ciascuno degli intervalli  $K + 1$

$$[\min(x), \xi_1], (\xi_1, \xi_2], \dots, (\xi_{K-1}, \xi_K], (\xi_K, \max(x)]$$

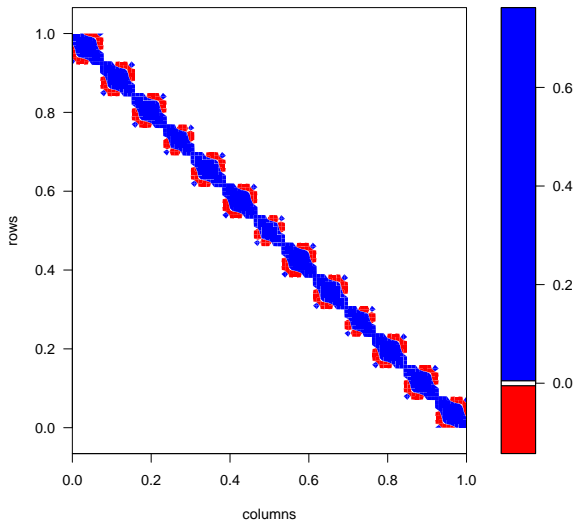
- Un difetto è che i valori previsti calcolati in corrispondenza di ogni nodo non saranno continui



Polinomio a tratti ( $M = 2$ )



### Polinomio a tratti di grado 2



Matrice di lisciamento del polinomio a tratti di grado 2

# Espansione di base

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

dove  $B_j(\cdot)$  sono funzioni note dette funzioni di base (*basis functions*).  
Per esempio

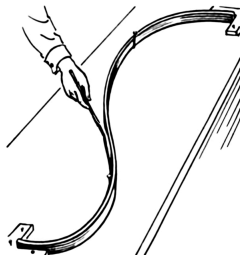
- Polinomio di grado 3:

$$B_1(x) = 1, B_2(x) = x, B_3(x) = x^2, B_4(x) = x^3$$

- Funzione a gradini con  $K$  nodi:

$$B_1(x) = \mathbb{1}\{x < \xi_1\}, B_2(x) = \mathbb{1}\{\xi_1 \leq x < \xi_2\}, \dots, \\ B_K(x) = \mathbb{1}\{\xi_{K-1} \leq x < \xi_K\}, B_{K+1}(x) = \mathbb{1}\{x \geq \xi_K\}$$

Storicamente, una spline era un righello elastico utilizzato per disegnare progetti tecnici, in particolare in costruzione navale e gli albori dell'ingegneria aeronautica



**Figure 3.6.** A spline.

Figura 3.6 da Bowman, Evers. Si veda anche <https://pages.cs.wisc.edu/deboor/draftspline.html>

# Spline di regressione

Una *spline* di grado  $M$  con nodi  $\xi_1, \dots, \xi_K$

- è un polinomio di grado  $M$  su ciascun intervallo

$$[\min(x), \xi_1], [\xi_1, \xi_2], [\xi_2, \xi_3], \dots, [\xi_{K-1}, \xi_K], [\xi_K, \max(x)]$$

- ha derivate continue di ordine  $0, \dots, M-1$  in ciascun nodo

$$f(\xi_k^-) = f(\xi_k^+), \quad \dots, \quad f^{(M-1)}(\xi_k^-) = f^{(M-1)}(\xi_k^+), \quad k = 1, \dots, K$$

dove  $\xi_k^+$  and  $\xi_k^-$  indica il limite sinistro e destro della funzione in  $\xi_k$

# Scelta del grado $M$

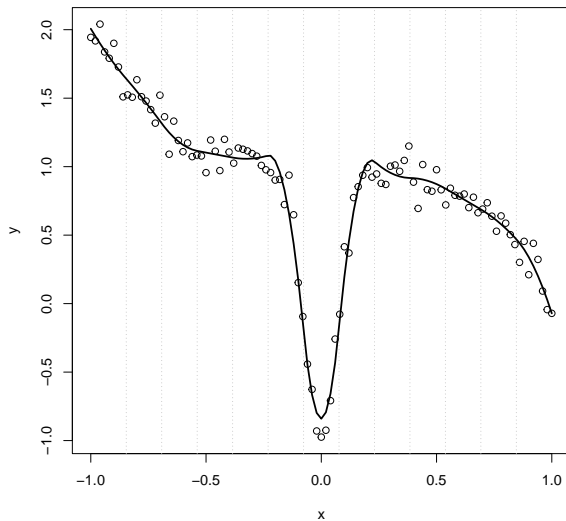
Il grado  $M$  della spline controlla il grado di lisciamiento.

Per  $M = 0$  la spline è una funzione a gradino discontinua.

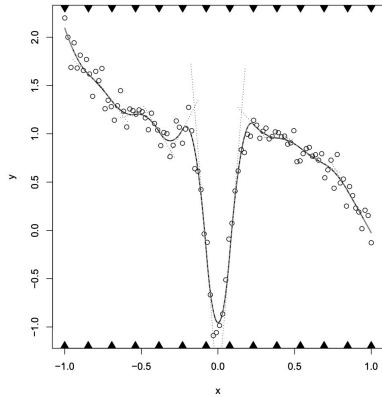
Per  $M = 1$  la spline è una linea poligonale.

Per valori maggiori di  $M$  la spline aumenta il grado di lisciamiento, ma si comporta sempre di più come un polinomio globale.

In pratica è raramente necessario andare oltre  $M = 3$ .



Spline di regressione di grado 2



(b) Piecewise polynomials which form a continuously differentiable function (derivatives at knots shown as dashed lines)

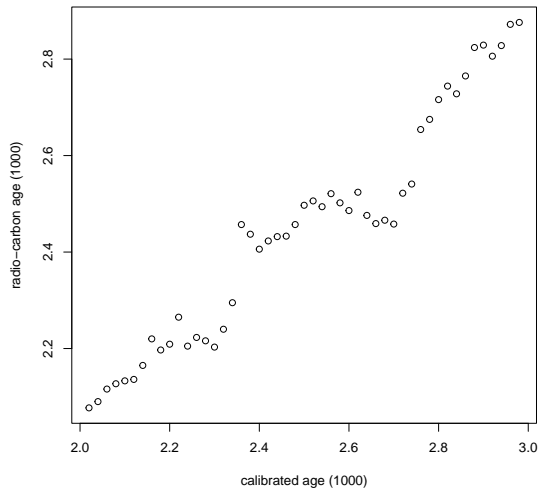
Figura 3.5(b) da Bowman, Evers

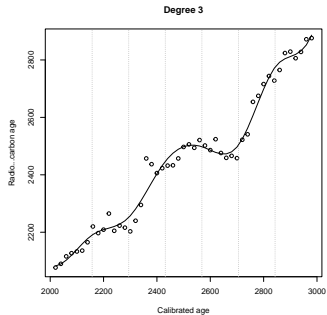
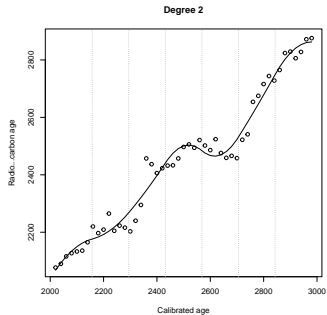
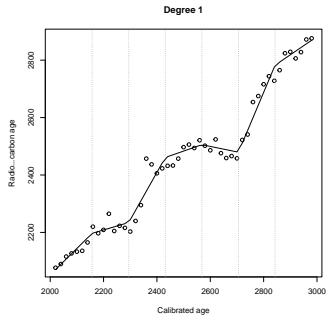
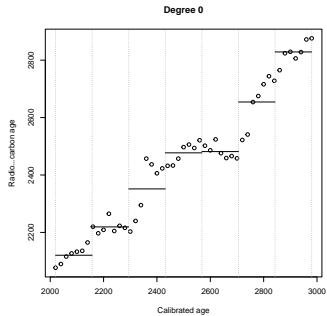
# Radiocarbon dating

In un esperimento scientifico sono state eseguite misurazioni ad alta precisione del radiocarbonio sulla quercia irlandese.

Per costruire una curva di calibrazione dobbiamo conoscere la relazione tra l'età (prevista) dal radiocarbonio e il calendario.







## Scelta del numero / posizione dei nodi

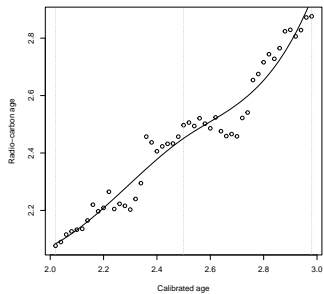
Maggiore è il numero di nodi, più flessibile è la funzione di regressione (meno bias ma più varianza).

Soprattutto quando il numero di nodi interni  $K$  non è elevato, il posizionamento dei nodi è molto importante.

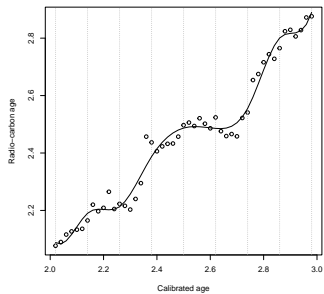
La strategia più semplice consiste nell'utilizzare un insieme di nodi equidistanti.

In alternativa, possiamo posizionare i nodi secondo il quantili della  $x$ . Questo rende la spline più flessibile nelle regioni con più dati (e quindi potenzialmente più informazioni) e meno flessibile nelle aree con meno dati (e potenzialmente meno informazioni).

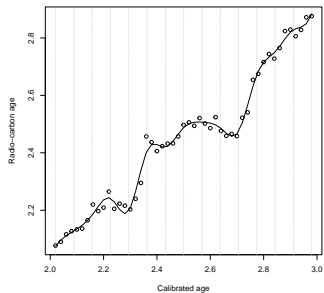
1 internal knots



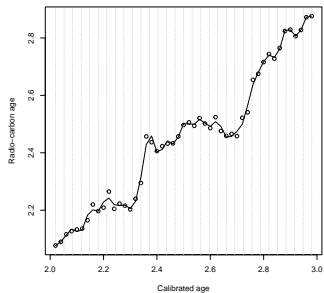
7 internal knots



13 internal knots



29 internal knots



# Spline come spazio vettoriale

Per un dato insieme di  $K$  nodi (interni) e dato il grado  $M$ , lo spazio delle spline (polinomiali) è uno spazio vettoriale.

Per trovare la dimensione di questo spazio vettoriale dobbiamo trovare il numero di parametri liberi.

Ogni polinomio ha  $M + 1$  parametri e ci sono  $K + 1$  polinomi.

Tuttavia non possiamo scegliere tutti questi parametri liberamente, poiché la funzione risultante deve soddisfare  $K$  vincoli di differenziabilità di ordine  $0, \dots, M - 1$

$$(M + 1)(K + 1) - KM = 1 + M + K$$

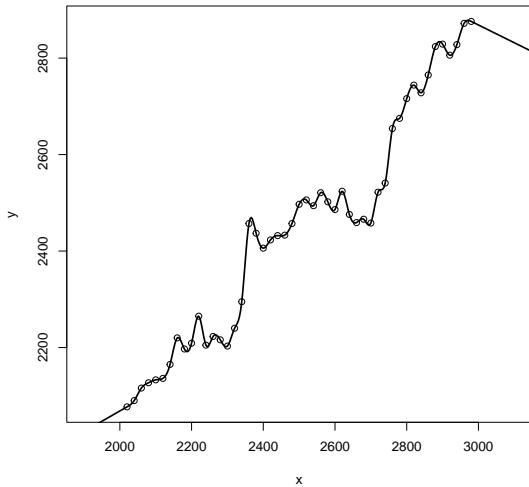
# Spline cubiche naturali

Una spline polinomiale  $f: [a, b] \mapsto \mathbb{R}$  di grado 3 è detta spline cubica naturale se  $f'(a) = f'(b) = 0$

Dato un insieme di  $K$  nodi interni lo spazio vettoriale di tutte le spline cubiche ha dimensione  $K + 4$ .

Le spline cubiche naturali introducono quattro vincoli aggiuntivi, e formano quindi uno spazio vettoriale di dimensione  $K$ . Ciò rende le spline cubiche naturali perfettamente adatte all'interpolazione

Un insieme di  $n$  punti  $(x_i, y_i)$  può essere interpolato esattamente usando una spline cubica naturale con  $x_{(2)} < \dots < x_{(n-1)}$  come nodi (interni). La spline cubica naturale di interpolazione è unica.



# Costruzione delle spline

Ora studieremo due modi di costruire una base per lo spazio vettoriale di spline polinomiali: la base delle potenze troncate e la base delle B-spline



# Base delle potenze troncate

Una spline di grado  $M$  con nodi  $\xi_1, \dots, \xi_K$  può essere definita dalla base delle potenze troncate (*truncated power basis*)

$$\begin{aligned} B_1(x) &= 1 \\ B_{j+1}(x) &= x^j, \quad j = 1, \dots, M \\ B_{M+k+1}(x) &= (x - \xi_k)_+^M, \quad k = 1, \dots, K \end{aligned}$$

dove  $(\cdot)_+$  è definito come

$$(x - \xi_k)_+^M = \begin{cases} (x - \xi_k)^M & x \geq \xi_k \\ 0 & \text{altrimenti} \end{cases}$$

Costruiamo quindi la matrice del disegno  $B$  di dimensione  $n \times (1 + M + K)$  con

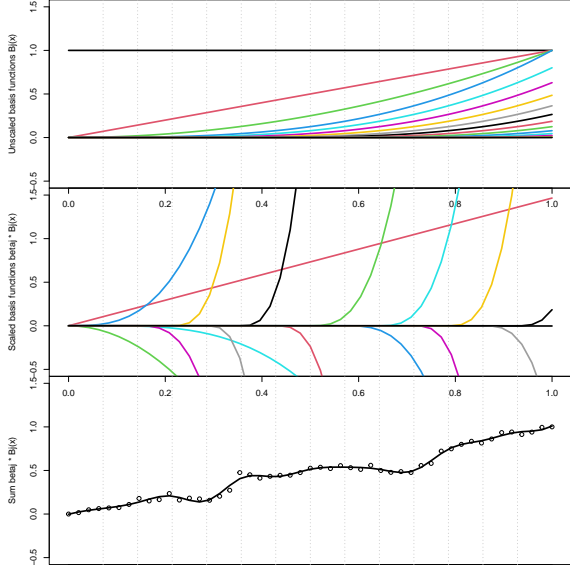
$$B_{i,j} = B_j(x_i) \quad i = 1, \dots, n, \quad j = 1, \dots, 1 + M + K$$

Possiamo scrivere la spline di regressione come

$$\hat{f}(x) = \sum_{j=1}^{1+M+K} \hat{\beta}_j B_j(x)$$

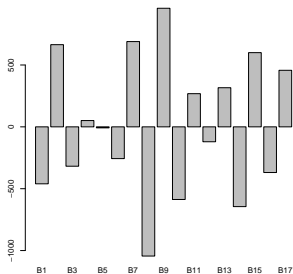
dove  $\hat{\beta}$  è dato da

$$\hat{\beta} = (B^t B)^{-1} B^t y$$



Alcune stime dei coefficienti sono molto elevate, e le funzioni di base sono scalate di un fattore superiore a 1000, con funzioni di base “vicine” di segno opposto.

La ragione di ciò è l'elevata correlazione tra le colonne della matrice del disegno  $B$  costruita con la base delle potenze troncate. La correlazione massima tra le colonne è 0.9990216.



# Base B-spline

Il *condition number* di  $B^t B$  è  $3.459309e+24$ , ovvero  $B^t B$  è quasi numericamente singolare. Questo comporta che la stima dei minimi quadrati dei coefficienti è prossima ad essere numericamente instabile

Le B-spline costituiscono una base numericamente più stabile. L'idea chiave delle B-spline è usare funzioni di base che sono locali, cioè diverse da zero per una “piccola” proporzione dell'intervallo della covariata e che sono delimitate superiormente.

Le B-spline si possono calcolare come differenza di funzioni di potenza troncate

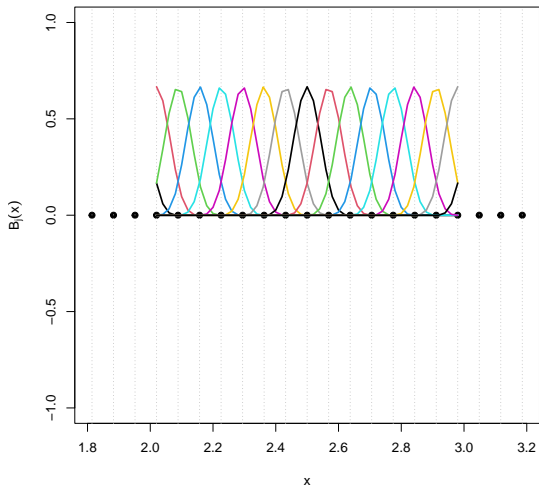
La formula generale per nodi equispaziati è

$$B_j(x) = \frac{(-1)^{M+1} \Delta^{M+1} f_j(x, M)}{h^M M!}$$

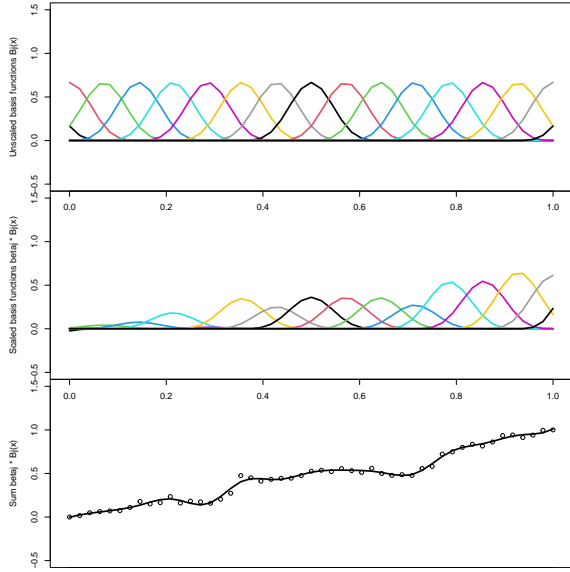
che soddisfano

$$\sum_j B_j(x) = 1$$

dove  $f_j(x, M) = (x - \xi_j)_+^M$ ,  $h$  è la distanza tra i nodi e  $\Delta^O$  è la differenza di ordine  $O$  con  $\Delta f_j(x, M) = f_j(x, M) - f_{j-1}(x, M)$ ,  
 $\Delta^2 f_j(x, M) = \Delta(\Delta f_j(x, M)) = f_j(x, M) - 2f_{j-1}(x, M) + f_{j-2}(x, M)$



Base *B*-spline con nodi equispaziati





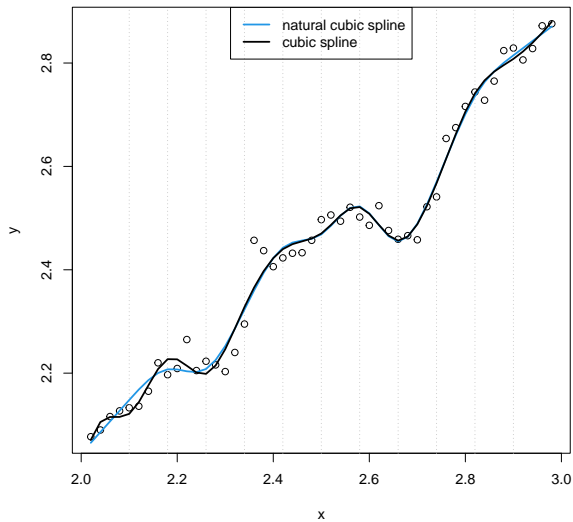
# Spline cubiche naturali

Una spline cubica naturale con  $K$  nodi è rappresentata da  $K$  funzioni di base

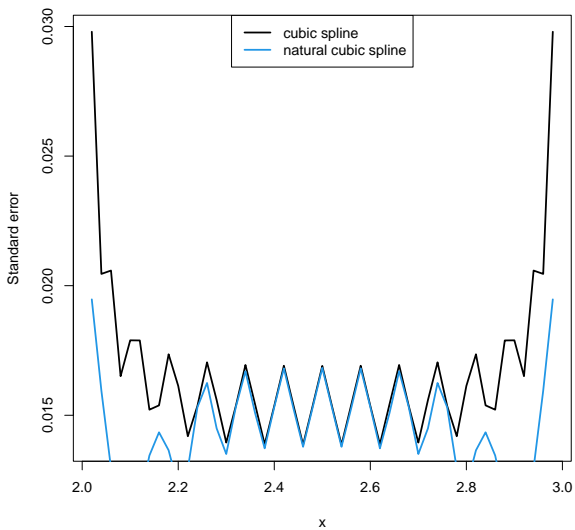
$$B_1(x) = 1$$

$$B_2(x) = x$$

$$B_{j+2}(x) = \frac{(x - \xi_j)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_j} - \frac{(x - \xi_{K-1})_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_{K-1}}$$
$$j = 1, \dots, K - 2$$



Standard error of the fit



Standard error of the fit