

6.Tutorium Generalisierte Regression

- Glättungsverfahren -

Cynthia Huber:

13.01.2015 und 20.01.2015

Michael Hanselmann:

15.01.2015 und 22.01.2015

Institut für Statistik, LMU München

Gliederung

- 1 Problemstellung
- 2 Regressionssplines
- 3 Glättungssplines
- 4 P-Splines

Gliederung

- 1 Problemstellung
- 2 Regressionssplines
- 3 Glättungssplines
- 4 P-Splines

Problemstellung

- bisher:
Betrachtung des **linearen** Prädiktors $\eta = x^T \beta$, als **Linearkombination** aus Kovariablen und Parametern
- jetzt:
Betrachtung einer allgemeineren, unbekannten **funktionalen** Form des Prädiktors:

$$\eta = f_1(x_1) + \dots + f_p(x_p)$$

→ semi- oder nonparametrische Regression!

- Schätzung von $f(x)$ mithilfe sogenannter **Glättungstechniken!**

Gliederung

- 1 Problemstellung
- 2 Regressionssplines**
- 3 Glättungssplines
- 4 P-Splines

Truncated Power Series Basis

Gegeben:

- (a) Hauptknoten $\tau_1, \dots, \tau_m$
- (b) Randknoten τ_0 bzw. τ_{m+1}
- (c) geordnete Beobachtungen $x_1 < x_2 < \dots < x_n$

$$f(x) = \delta_0 \underbrace{1}_{P_0(x)} + \delta_1 \underbrace{x}_{P_1(x)} + \delta_2 \underbrace{x^2}_{P_2(x)} + \dots + \delta_k \underbrace{x^k}_{P_k(x)} + \sum_{i=1}^m \delta_{k+i} \underbrace{(x - \tau_i)_+^k}_{P_{k+i}(x)}$$

$$= \sum_{i=0}^{k+m} \delta_i P_i(x) = \textcolor{red}{s}(x) \quad (\text{Spline-Funktion})$$

$$\text{mit } (x - \tau_i)_+ = \max\{0, x - \tau_i\}$$

→ **truncated power series**

Truncated Power Series Basis

Eigenschaften von $s(x)$:

- Polynom k -ten Grades auf $[\tau_i, \tau_{i+1}]$
- $s(x)$ $(k - 1)$ -mal stetig differenzierbar auf $[\tau_0, \tau_{m+1}]$
- Gewichtete Summe aus $m + k + 1$ Funktionen
- **Natürliche kubische Splines** auf $[\tau_0, \dots, \tau_{m+1}]$, falls $k = 3$:
→ $s(x)$ 2-mal stetig differenzierbar auf $[\tau_0, \tau_{m+1}]$
- Parameterschätzungen mittels linearer Regression

B-Splines

- Voraussetzungen wie zuvor, aber zusätzliche Randknoten

$\tau_{-k}, \dots, \tau_0$ bzw. $\tau_{m+1}, \dots, \tau_{m+1+k}$

- Basisfunktion 1. Ordnung (0-ten Grades):

$$B_{i,1} = \begin{cases} 1, & x \in [\tau_i, \tau_{i+1}] \\ 0, & \text{sonst.} \end{cases}$$

- Basisfunktion $(k+1)$ -ter Ordnung

$$B_{i,k+1}(x) = \frac{x - \tau_i}{\tau_{i+1} - \tau_i} B_{i,k}(x) + \frac{\tau_{i+k+1} - x}{\tau_{i+k+1} - \tau_{i+1}} B_{i+1,k}(x)$$

→ **rekursiv** definiert!

B-Splines

- Jede Spline-Funktion ist als Linearkombination von B-Splines darstellbar:

$$f(x) = \textcolor{red}{s}(x) = \sum_{i=1}^{m+k+1} \delta_i B_{i,k}(x)$$

- **Prinzip** von Regressionssplines:
Approximiere die gesuchte Funktion $f(x)$ durch gewichtete Basisfunktionen!

B-Splines

Eigenschaften von B-Splines k -ten Grades/ $(k + 1)$ -ter Ordnung:

- $k + 1$ Polynomabschnitte vom Grad k (Ordnung $k + 1$)
- positive Funktionswerte der Basisfunktionen in einem durch $k + 2$ Knoten aufgespannten Bereich, Null an allen anderen Stellen
- $k + 1$ B-Splines $\neq 0 \forall x$ (ausgenommen Knotenpunkte)
- Vorteile gegenüber *truncated power series*:
 - numerisch stabiler
 - höhere Genauigkeit durch lokale Wirksamkeit
- Genauigkeit vom Polynomgrad abhängig

Beachtenswertes

- Grundkonflikt zw. Bias und Varianz

Glattheit	↔	Datentreue
wenige Knoten		viele Knoten
niedrige Variabilität, d.h.		hohe Variabilität, d.h.
$\text{var}(\hat{f}(x))$ eher klein		$\text{var}(\hat{f}(x))$ eher groß
hoher Bias		geringer Bias

- **Wichtig:**
Anzahl und Lage der Basisfunktionen
- **Zweitrangig:**
Konkrete Form der Splines

Gliederung

- 1 Problemstellung
- 2 Regressionssplines
- 3 Glättungssplines**
- 4 P-Splines

Allgemeine Darstellungsweise

- Ziel:
Finde bei der Schätzung von $f(x)$ einen **Kompromiss** zwischen Glattheit und Datentreue der Funktion
- **Penalized-Least-Squares (PLS)**

$$\sum_{i=1}^n (y_i - f(x_i))^2 \underbrace{w_i}_{\text{bek. Gewichte}} + \underbrace{\lambda}_{\text{Glättungsparam.}} \underbrace{J(f)}_{\text{Funktional}} \xrightarrow{f(x_1), \dots, f(x_n)} \min$$

$$\lambda = 0 \Rightarrow \hat{f}(x_i) = y_i \text{ (Datentreue)}$$

$$\lambda \rightarrow \infty \Rightarrow \hat{f}(x_i) \text{ sehr starke Glättung}$$

Glättungssplines ohne Gewichtung ($w_i = 1$)

Vor.: $x_1 < x_2 < \dots < x_n$

f' und f'' existent und quadratisch integrierbar

PLS-Ansatz:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int \{f''(u)\}^2 du \xrightarrow{f(x_1), \dots, f(x_n)} \min$$

Lösung: natürliche kubische Splines, d.h. $\forall i$

→ $\hat{f}$ Polynom 3. Grades auf $[x_i, x_{i+1}]$

Glättungssplines ohne Gewichtung ($w_i = 1$)

PLS-Ansatz in Matrixform:

$$\text{PLS}(f) = (y - f)^T (y - f) + \lambda \underbrace{f^T K f}_{\text{quadratische Form}} \xrightarrow{f} \min$$

$$\text{mit } y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad f = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \quad \text{und} \quad K = D^T C^{-1} D$$

C und D abhängig von Distanzen $x_{i+1} - x_i$

(vgl. Green und Silverman, 1994)

Glättungssplines ohne Gewichtung ($w_i = 1$)

PLS-Schätzer:

$$\hat{f} = (I + \lambda K)^{-1} y = S_y y$$

Eigenschaften der **Smoothing-Matrix** S_y :

- $(n \times n)$ - Matrix entsprechend der Hat-Matrix $H = X(X^T X)^{-1} X^T$ im linearen Modell
- direkte Berechnung nur mit hohem numerischem Aufwand möglich
⇒ Lösungsansatz über Bandstruktur in C und D
(vgl. Reinsch, 1967)

Gliederung

- 1 Problemstellung
- 2 Regressionssplines
- 3 Glättungssplines
- 4 P-Splines**

Bildung von P-Splines

- Prinzip:

$$\text{P-Splines} = \text{B-Splines} + \text{Penalisierung}$$

- Vorgehen:

- 1 Entwicklung von $\underline{f}$ in B-Splines:

$$f(x_i) = \sum_{j=1}^m \overbrace{\delta_j}^{\text{Parameter}} \overbrace{B_j(x_i)}^{\text{bekannt}} \implies \underbrace{f}_{(n \times 1)} = \underbrace{B}_{(n \times m)} \cdot \underbrace{\delta}_{(m \times 1)}$$

hier: m = Anzahl der Basisfunktionen

Bildung von P-Splines

- ② Penalisierung aufeinanderfolgender Koeffizienten durch Minimierung von

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^m \delta_j B_j(x_i) \right)^2 w_i + \lambda \sum_{j=d+1}^m \left(\Delta^d \delta_j \right)^2$$

mit

$$\Delta^1 \delta_j = \delta_j - \delta_{j-1} \quad \text{Differenz 1. Ordnung}$$

$$\Delta^2 \delta_j = \Delta(\delta_j - \delta_{j-1}) \quad \text{Differenz 2. Ordnung}$$

$$= \delta_j - \delta_{j-1} - (\delta_{j-1} - \delta_{j-2})$$

$$= \delta_j - 2\delta_{j-1} + \delta_{j-2}$$

$$\vdots$$

Penalisierung in Matrixform

$$\begin{aligned}
 \text{(a) } d = 1 : \quad \sum_{j=2}^m (\delta_j - \delta_{j-1})^2 &= \left\| \begin{pmatrix} \delta_2 - \delta_1 \\ \delta_3 - \delta_2 \\ \vdots \\ \delta_m - \delta_{m-1} \end{pmatrix} \right\|^2 \\
 &= \left\| \begin{pmatrix} -1 & 1 & 0 & \cdots \\ 0 & -1 & 1 & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ \vdots & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_m \end{pmatrix} \right\|^2 = \delta^T D_1^T D_1 \delta = \delta^T K_1 \delta
 \end{aligned}$$

(b) Analoger Strafterm für alle $d = 1, 2, \dots$:

$$\delta^T D_d^T D_d \delta = \delta^T K_d \delta$$

$$(1 \times m) \underbrace{(m \times m - 1)(m - 1 \times m)}_{(m \times m)} (m \times 1) = (1 \times m)(m \times m)(m \times 1)$$

Penalized-Splines-Schätzung (PS)

$$PS(\underline{\delta}) = (y - \underbrace{B \delta}_f)^T (y - B \delta) + \lambda \delta^T \underbrace{K_d}_{D_d^T D_d} \delta \xrightarrow{\delta} \min$$

$$\hat{\delta} = (B^T B + \lambda K_d)^{-1} B^T y$$

Eigenschaften von P-Splines:

- Form des Strafterms ähnlich wie bei Glättungssplines
- Regressionssplines, da Knotenvorgabe nötig
- $\lambda = 0 \Rightarrow$ reiner Fit von Basisfunktionen (B-Splines);
keine Glättung
- $\lambda \rightarrow \infty$ & $d = k + 1 \Rightarrow$ Fit eines Polynoms vom Grad k