

## K7 Glättung (Smoothing)

- Glättung für empirische Verteilungen  
(Dichteschätzung)  
d.h. Darstellung der Verteilung einer stetigen Variable
- Glättung für Zeitreihen  
—um den Trend herauszuholen
- Glättung für nichtlineare Modellierung  
—weil es keine allgemeine parametrische Familie gibt

Die Daten werden gefiltert.

## Warum Glättung?

1. Den Einfluß von Ausreißern entgegenzuwirken.
2. Modelle zu überprüfen  
(z.B. auf Linearität/Nichtlinearität).
3. Strukturänderungen hervorzuheben.
4. Lokale Variabilität zu schätzen.

Was bleibt nach einer Glättung wird der Smooth (Glättung) genannt. Der Rest wird Rough (die Residuen) genannt:

$$\text{Data} = \text{Smooth} + \text{Rough}$$

$$Y = f(x) + \epsilon$$

## 7.1 Dichteschätzung

Ein Kernschätzer für die eindimensionale Dichte einer Variable ist

$$\hat{f}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}{nh}$$

Die Kernfunktion  $K$  und die Bandbreite  $h$  müssen bestimmt werden.

$\hat{f}(x)$  ist eine Dichte  $\Rightarrow$

$$\begin{aligned}\int \hat{f}(x) dx &= 1 \\ \int \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) dx &= nh \\ \int K\left(\frac{x-x_i}{h}\right) dx &= h \quad \forall i \\ \int K(v) dv &= 1 \quad \text{für } v = \left(\frac{x-x_i}{h}\right)\end{aligned}$$

## 7.1.1 Beispiele von Kernfunktionen

### 1. Gleichverteilt

$$K(u) = \begin{cases} 0.5 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

### 2. Epanechnikov

$$K(u) = \begin{cases} 0.75(1 - u^2) & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

### 3. Biweight

$$K(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

### 4. Gauß

$$K(u) = (2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}$$

## 7.1.2 Eigenschaften von Kernfunktionen

$$\begin{aligned} K(u) &\geq 0 \\ \int K(u)du &= 1 \quad (\text{Gewicht}) \\ \int uK(u)du &= 0 \quad (\text{Bias}) \end{aligned}$$

## 7.1.3 Wie gut ist eine Kernfunktion?

Wir können die  $L_2$  Norm als Kriterium nehmen:

$$ISE = \int [\hat{f}(x) - f(x)]^2 dx$$

Es gibt auch die  $L_1$  Norm oder die  $L_\infty$  Norm

$$\sup |\hat{f}(x) - f(x)|$$

ISE könnte für einen bestimmten Datensatz und bekannten  $f(x)$  interessant sein. Für unbekannte  $f(x)$  betrachtet man statt dessen

$$E_f[ISE] = MISE$$

$$\begin{aligned}
MSE(x) &= E[(\hat{f}(x) - f(x))^2] \\
&= Var(\hat{f}(x)) + Bias^2(\hat{f}(x)) \\
Bias(\hat{f}(x)) &= E[\hat{f}(x)] - f(x) \\
E[\hat{f}(x)] &= \int_{x-h}^{x+h} K\left(\frac{x-y}{h}\right) f(y) dy \\
&= \int_{-h}^{+h} K\left(\frac{u}{h}\right) f(x-u) du
\end{aligned}$$

Für  $h$  klein ersetzen wir  $f(x-u)$  mit

$$f(x-u) = f(x) - uf'(x) + \frac{u^2}{2}f''(x) \dots$$

Für  $n \rightarrow \infty$  und  $h \rightarrow 0$

$$\begin{aligned}
E[\hat{f}(x)] &= f(x) + \frac{h^2}{2}f''(x)\sigma_K^2 + O(h^4) \\
\sigma_K^2 &= \int_{-h}^{+h} u^2 K\left(\frac{u}{h}\right) du
\end{aligned}$$

$$\Rightarrow Bias(\hat{f}(x)) \simeq \frac{h^2}{2}f''(x)\sigma_K^2$$

Auf ähnliche Weise kann man zeigen, dass

$$Var(\hat{f}(x)) = \frac{f(x) \int K^2(u) du}{nh} + O(n^{-1})$$

## 7.1.4 AMISE

Die theoretischen Resultate beschäftigen sich mit *AMISE*, die asymptotische *MISE*.

$$\begin{aligned} AMISE &= \int_{-\infty}^{\infty} MSE[\hat{f}(x)] dx \\ &= \frac{\int K^2(u) du}{nh} + \frac{h^4 \sigma_K^4 \int f''(x)^2 dx}{4} \end{aligned}$$

Durch das Minimieren von AMISE kann man eine “optimale” Fensterbreite  $h$  bestimmen:

$$h_0 = \left[ \frac{\int K^2(u) du}{\sigma_K^4 \int f''(x)^2 dx} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

und

$$AMISE_0 = \frac{5}{4} \left[ \sigma_K \int K^2(u) du \right]^{\frac{4}{5}} \left( \int f''(x)^2 dx \right)^{\frac{1}{5}} n^{-\frac{4}{5}}$$

Nur kennt man natürlich  $f''(x)$  nicht. Man könnte aber  $K(u)$  wählen, um AMISE kleiner zu machen. Dafür muss man  $\sigma_K \int K(u)^2 du$  minimieren.

Es kann gezeigt werden, dass die Epanechnikov Kernfunktion

$$K(u) = \begin{cases} 0.75(1 - u^2) & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

optimal ist. In diesem Fall

$$\sigma_K \int K^2(u) du = 3/(5\sqrt{5})$$

Das Verhältnis  $\sigma_K \int K^2(u) du / [3/(5\sqrt{5})]$  wird als Effizienzmaß genommen.

Kernel	Form	Effizienz
Epanechnikov	$0.75(1 - u^2)$	1
Biweight	$\frac{15}{16}(1 - u^2)^2$	1.0061
Triweight	$\frac{35}{32}(1 - u^2)^3$	1.0135
Gaussian	$(2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}$	1.0513
Gleichvert	$\frac{1}{2}$	1.0758



## 7.2 Dichteschätzer und Standardplots

Histogramme, Dotplots, Boxplots als Ergänzung zur Dichteschätzung.

Ein Dichteschätzer ist ein statistisches Werkzeug. Wir benutzen die empirischen Daten, um ein Modell für die vermutete dahinterliegende Verteilung zu schätzen. Wir bekommen eine geglättete Übersicht der Daten, die aber auf Sondereigenschaften oder Mischungen von Verteilungen hinweisen kann.

Die anderen Plots sind für datenanalytische Zwecke eingesetzt, um Eigenschaften wie Ausreißer, Lücken, Gipfel, Granularität, Grenzen... zu untersuchen.

## 7.3 Lowess (statt Regression)

(LOcally WEighted regression Scatterplot Smoothing)

### Erste Glättung

1. Für jeden Punkt  $j$  nimmt man die  $f\%$  nächstliegenden Punkte. Man berechnet die größte Entfernung unter diesen Punkten:

$$s_j = \max(x_j - x_{min}, x_{max} - x_j)$$

2. Tricube Gewichte werden für jeden Punkt  $i$  um  $j$  berechnet

$$w_i = \begin{cases} (1 - |u_i|^3)^3 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

wo

$$u_i = (x_j - x_i) / s_j$$

3. Eine gewichtete lineare Regression um  $j$  wird berechnet. Der erste geglättete Wert für  $j$  ist die Regressionsvoraussage für  $j$  aus der gewichteten Regression um  $j$ .

## Zweite (robuste) Glättung (nach Data Desk)

4. Nachdem man das für alle Punkte gemacht hat, wird eine zweite Glättung durchgeführt, um die Wirkung von Extremwerten zu beseitigen. Zuerst werden die Residuen  $r_j$  berechnet und dann der Median der absoluten Residuen in der Nachbarschaft des Punkts  $j$ :

$$MAD_j = \text{median}_j(|r_i|)$$

5. Neue Gewichtungsfaktoren werden berechnet

$$w_i = \begin{cases} (1 - u_i^2)^2 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

wo  $u_i = r_i/MAD_i/6$

6. Neue gewichtete Lokal-Regressionen werden an jedem Punkt berechnet, wo als Gewichtung das Produkt der zwei Gewichte genommen wird. Der endgültige geglättete Wert für Punkt  $j$  wird aus dieser entsprechenden Regression genommen.

## 7.4 Erwünschte Eigenschaften von Glättungen

1. Eine Glättung sollte durch die Mitte der Daten gehen.
2. Eine Glättung sollte nie die Richtung scharf ändern.
3. Eine Glättung sollte lange Wellen haben und nicht viel wackeln.
4. Ausreißer und kurze, scharfe Änderungen sollen nicht beachtet werden.

## 7.5 Wie soll man glätten?

$$y_i = m(x_i) + \epsilon_i$$

mit

$$m(x) = E[Y|X]$$

$$E[\epsilon|X] = 0$$

$$V[\epsilon|X] = \sigma^2(x)$$

$$\begin{aligned} \Rightarrow m(x) &= \int y f(y|x) dy \\ &= \frac{\int y f(x, y)}{f_X(x)} \end{aligned}$$

$f(y|x)$  ist die bedingte Dichte von  $Y$  gegeben  $X$

$f(x, y)$  ist die gemeinsame Dichte von  $X$  und  $Y$

$f_X(x)$  ist die marginelle Dichte von  $X$

Die Dichten sind unbekannt und müssen geschätzt werden.

## 7.6 lowess, loess, R und Mondrian

- **R lowess**

lowess ist nur für Streudiagramme:

```
alw ← lowess(x, y, f=0.3, iter=3, delta=0)
```

alw enthält x und die dazu geglätteten Werte.

- **R loess**

loess berücksichtigt mehrere erklärende Variablen:

```
al1 ← loess(y ~ x1 + x2 + x3, span=0.3, degree=1,  
family="symmetric", iterations=4)
```

al1 ist ein Objekt mit vielen Attributen.

N.B. die Formate und Defaults sind verschieden für lowess und loess (scatter.smooth verwendet loess für eine erklärende Variable, hat aber das (x,y) Format!)

### **Mondrian loess**

In Streudiagrammen verwendet Mondrian loess mit einer erklärenden Variable via Rserve.