

Beispiele von Dichteschätzung in R

Aus dem Boston Housing Datensatz:

- > Boston ← read.table("BostonHc.x",header=TRUE)
- > attach(Boston)
- > plot(density(AGE))
- > plot(density(AGE, kernel=c("epanechnikov")))
- > plot(density(AGE, bw=3, kernel=c("epanechnikov")))

Simulierte Daten:

- > x4 ← rchisq(200, df=4)
- > hist(x4)
- > plot(density(x4))
- > x2 ← rchisq(200, df=2)
- > hist(x2)
- > plot(density(x2))

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

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

Durch das Minimieren von *AMISE* kann man eine "optimale" Fensterbreite h bestimmen:

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

und

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

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

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

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

Probleme mit Kernfunktionen

Bias an den Grenzen

(z.B. eine Variable ohne negativen Werte)

o.B.d.A. sei $h = 1$. Dann ist $E[\hat{f}(x)]$ an einer Stelle $x = p$ ($0 < p < 1$)

$$\begin{aligned} E[\hat{f}(x)] &= \int_0^{1+p} K(u) f(x-u) du \\ &= f(x) \left(\int_{-1}^p K(u) du \right) \\ &\quad - f'(x) \left(\int_{-1}^p u K(u) du \right) \\ &\quad + f''(x) \frac{\left(\int_{-1}^p u^2 K(u) du \right)}{2} \\ &\quad + O(h^4) \end{aligned}$$

Wo, wie vorher, $f(x-u)$ mit einer Taylor Entwicklung um x approximiert worden ist.

Wenn wir die Grenze genau wissen, können wir "Designer" Kernfunktionen berechnen. Für eine Grenze $x = 0$ bestimmen wir gleitende (floating) Kernfunktionen im Intervall $[0, h]$ unter den Bedingungen

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

Dadurch kann man gute Resultate erzielen, wenn man nicht nur die Grenze genau kennt, aber auch die Form der Dichte in der Nähe der Grenze. Betrachten Sie z.B. χ^2 Verteilungen mit verschiedenen Freiheitsgraden.

Zwei spezifische Lösungsvorschläge sind

- (a) Normalisierung der Kernfunktion.
- (b) Spiegelung an der Grenze.

Aber beide weisen Bias auf.

Lokale Variabilität

h wird defaultmässig als Konstant gesetzt, so das lokale Variabilität schlecht modelliert werden kann.

h gross führt zu Überglätten (oversmoothing) wo viele Punkte mit viel Struktur sind.

h klein führt zu Unterglätten (undersmoothing) wo wenig Punkte sind (oft in den Schwänzen).

Gipfel und Täler

Für alle konstant h werden Gipfel und Täler abgeschwächt.

Wir möchten $h(x)$ lokal schätzen

$$\hat{f}(x) = \frac{\sum K\left(\frac{x-x_i}{h(x)}\right)}{nh(x)}$$

Lowess erreicht das, indem statt alle Punkte zwischen $(x - h)$ und $(x + h)$ zu nehmen, es die nächsten $f\%$ Punkte nimmt. Das führt zum Überglätten bei wenigen Punkten (meistens am Rand) und Unterglätten, wo viele sind.

Dichteschätzer und Standardplots

(Histogramme, Dotplots, Boxplots)

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.

Regressionsglättungsfunktionen

$$m(x) = \frac{\int y f(x, y) dy}{f_X(x)}$$

Wir werden $f(x, y)$ mit

$$\frac{\sum K_x\left(\frac{x-x_i}{h_x}\right) K_y\left(\frac{y-y_i}{h_y}\right)}{nh_x h_y}$$

obgleich dieser Schätzer kein multivariater Schätzer ist. Dann bekommen wir den sogenannten Nadaraya-Watson Schätzer für $m(x)$:

$$m_{NW}(x) = \frac{\sum y_i K_x\left(\frac{x-x_i}{h_x}\right)}{\sum K_x\left(\frac{x-x_i}{h_x}\right)}$$

eine gewichtete Summe der $\{y_i\}$, weil

$$\frac{1}{h_y} \int y_i K_y\left(\frac{y-y_i}{h_y}\right) dy = y_i$$

$m_{NW}(x)$ kann man nun als die Lösung, β_0 von einem gewichteten KQ-Problem erkennen:

$$\text{Min} \sum (y_i - \beta_0)^2 K\left(\frac{x - x_i}{h}\right)$$

was darauf hinweist, dass eine allgemeinere Methode wäre $m(x)$ mit einem lokalen Polynomial zu schätzen, statt mit einem Konstant.

Jetzt haben wir folgende Wahlmöglichkeiten:

1. Wahl der Kernfunktion
2. Wahl der lokalen Fensterbreite
3. Behandlung etwaiger Grenzen
4. Bestimmung der Ordnung der Polynomialfunktion

Und was machen wir mit Ausreißern?

Multivariate erklärende Variablen

Der Glättungsschätzer wird als Lösung von

$$\text{Min} \sum (y_i - \beta_0 - \beta_1(x_1 - x_{1i}) - \dots)^2 K_d(H^{-1}(x - x_i))$$

berechnet. K_d ist eine multivariate Kernfunktion und H ist eine multivariate Bandbreitenmatrix der Größe $(d + 1) \times (d + 1)$

z.B.

1. $H = hI$

2. $H = \text{diag}(h_i)$

3. $H = hS^{0.5}$ (S ist eine Schätzung der VarianzKovarianzmatrix von X ("Sphering"))

Das Hauptkonzept des Glättens ist, dass wir lokale Modelle schätzen (mit Einbeziehung der benachbarten Punkten). Aber in n Dimensionen gibt es nicht so viele Punkte:

1. Betrachten wir einen Hypercube $[-1,1]$ mit gleichverteilten Punkten.

d	% im Einheitskreis um 1
2	78,5
5	16
10	0,25

2. Für eine 10–dimensionale Normalverteilung liegt 99% der Dichte außerhalb des 10–dimensionalen Kugels mit Radius 1,16 um 0.

Die lokalen Berechnungen können mit einer festen Binbreite, mit einer lokalen Binbreite oder mit einer Nachbarschaft (wie Lowess) gemacht werden.

Statistische Aspekte

Eine Glättung ist das Resultat einer numerischen Optimierung.

Wie kann man die Güte messen? R^2 wäre eine erste Approximation, aber wieviele Parameter hat man geschätzt? Nichtparametrische Kriterien braucht man.

Wie kann man Variabilität schätzen?

- **Residuen - Punktweise**

Eine einfache Möglichkeit ist es, die Residuen ("Roughs") dazu zu benutzen und Anpassung und Variabilität entweder global oder lokal zu schätzen. Dann muß die Behandlung von Ausreißern natürlich konsequent sein.

- **Bootstrapping - Kurvenweise**

Interessanter (aber komplexer) ist es, eine Gruppe von Stichproben zu simulieren und für jede eine Glättung zu berechnen.