Stochastik–Praktikum
Nichtparametrische Dichteschätzung

Thorsten Dickhaus

Humboldt-Universität zu Berlin

12.10.2010

Einleitung

Zunächst:

$X_1, \ldots, X_n$ reellwertige i. i. d. Zufallsvariablen, deren Verteilung die Dichte f bezüglich des Lebesgue–Maßes besitzt.

Datenbeispiel:

272 beobachtete Ausbrüche des “Old Faithful”–Geysirs im Yellowstone National Park mit Eruptionsdauer sowie der Wartezeit bis zum nächsten Ausbruch

```
> data(faithful)
> er<-faithful$eruptions
```

Übersicht

Histogramm–Schätzer

Intervalle („bins“) I_s:

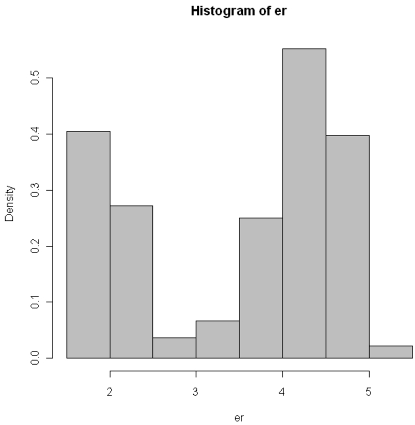
$$I_s = (a_{s-1}, a_s],\ s \in \{1, \dots, d\}$$

$$n_s := \#\{x_i \in I_s,\ i \in \{1, \dots, n\}\}$$

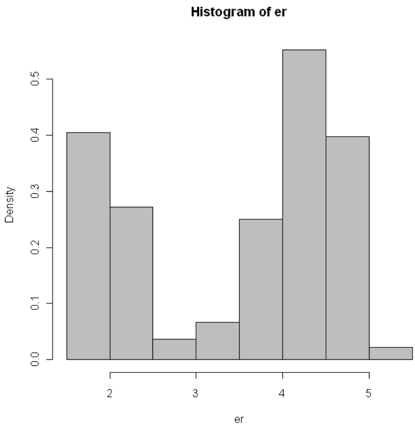
$$\hat{f}_{hist}(x) = \frac{n_s}{n} \frac{1}{a_s - a_{s-1}} \mathbb{1}_{\{I_s\}}(x)$$

Im Falle gleicher Intervalllängen mit
$a_s - a_{s-1} \equiv h \ \forall s \in \{1, \dots, d\}$:

$$\hat{f}_{hist}(x) = \frac{n_s}{nh} \mathbb{1}_{\{I_s\}}(x)$$




```
> hist(er,freq=FALSE,col="grey")
```



Nachteil des Histogramm–Schätzers:
Schätzer hängt von der Wahl der bin–Längen und des Startwertes ab!

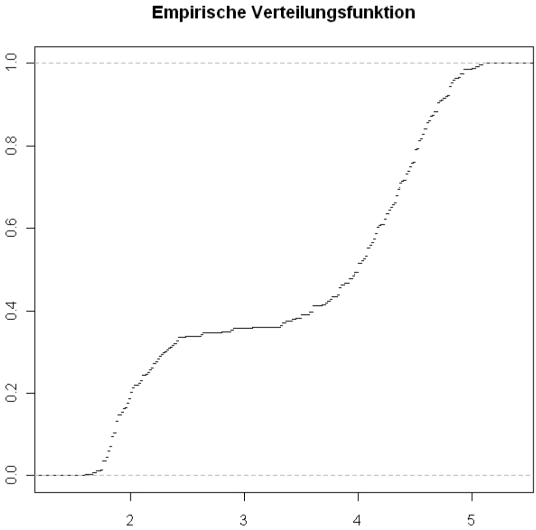
Empirische Verteilungsfunktion

$$F_n(x) = \frac{\#\{x_i | x_i \leq x,\ i \in \{1, \ldots, n\}}{n} = \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}_{(-\infty, x]}(x_i)\ .$$

Satz von Glivenko–Cantelli liefert fast sichere gleichmäßige Konvergenz:

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0 \quad \mathbb{P}_f - f.s.$$

```
> ecdf(er)
Empirical CDF
Call: ecdf(er)
 x[1:126] =    1.6,  1.667,    1.7,  ...,  5.067,    5.1
> summary(ecdf(er))
Empirical CDF:    126 unique values with summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.271   3.841   3.472   4.429   5.100
```

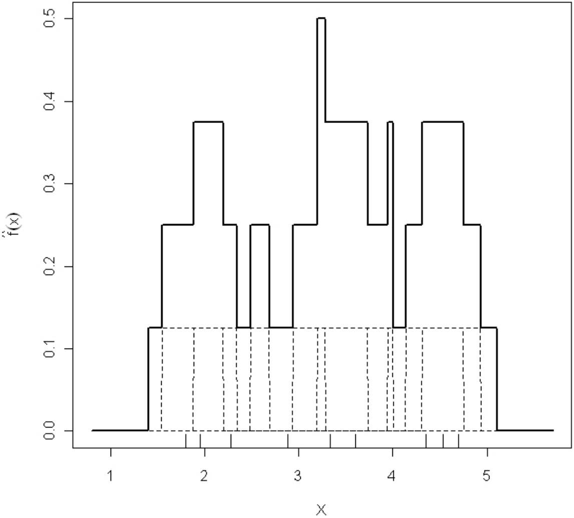

Empirische Verteilungsfunktion
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5

Gleitendes Histogramm

Durch den gleitenden Histogramm–Schätzer

$$\hat{f}_{GH}(x) := \frac{\#\{x_i | x_i \in (x-h, x+h]\}}{2hn} = \frac{F_n(x+h) - F_n(x-h)}{2h}$$
$$= \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}_R\left(\frac{x - x_i}{h}\right) \quad \text{mit} \quad \mathcal{K}_R(t) = (1/2)\mathbb{1}_{[-1,1]}(t),$$

bei dem jede Beobachtung Mittelpunkt eines bins ist, lässt sich das Startwertproblem lösen.


0.0
0.1
0.2
0.3
0.4
0.5
$\hat{f}(x)$
1
2
3
4
5
X

Übersicht

1. Histogramm–Schätzer und empirische Verteilungsfunktion
2. Univariate Kerndichteschätzung
3. Bandweitenwahl für Kernschätzer
4. Multivariate Kerndichteschätzung

Kernfunktionen

Definition

Eine Funktion $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ heißt Kern, falls gilt:

1. $\int \mathcal{K}(x)dx = 1$, $\mathcal{K}(x) \geq 0 \;\; \forall x \in \mathbb{R}$, $\mathcal{K}(x) = \mathcal{K}(-x)$

   Regularitätsbedingungen:

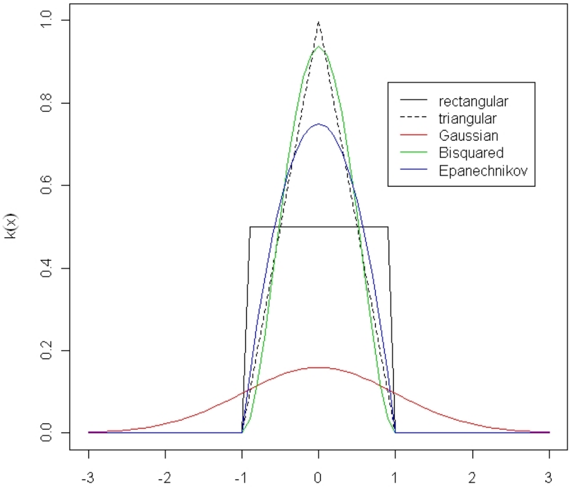
2. $\sup_{x \in \mathbb{R}} \mathcal{K}(x) = M < \infty$
3. $|x|\mathcal{K}(x) \to 0$ für $|x| \to 0$, $\int x^2 \mathcal{K}(x)dx =: k_2 < \infty$

.

Beispiele für Kernfunktionen:

Rechteckskern	$\mathcal{K}(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x),$
Dreieckskern	$\mathcal{K}(x) = (1 - \lvert x \rvert)\mathbb{1}_{[-1,1]}(x),$
Gaußkern	$\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$
Bisquarekern	$\mathcal{K}(x) = \frac{15}{16}(1 - x^2)^2\mathbb{1}_{[-1,1]}(x),$
Epanechnikovkern	$\mathcal{K}(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x).$

Grafische Darstellung der Kernfunktionen



Univariater Kerndichteschätzer

Definition

Sei $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern.

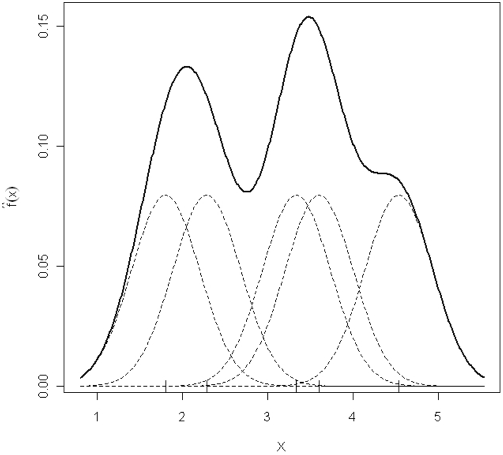
$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{t - x_i}{h}\right) = \int \frac{1}{h} \mathcal{K}\left(\frac{t - x}{h}\right) F_n(dx)$$

heißt (univariater) Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

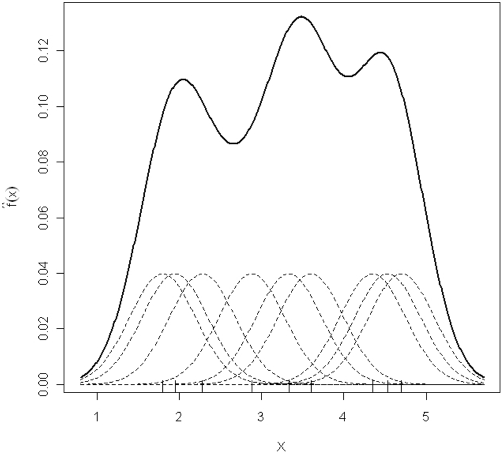
Mit $\tilde{\mathcal{K}}(t) := \int_{-\infty}^{t} \mathcal{K}(x) dx$ lässt sich auch F schätzen durch:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathcal{K}}\left(\frac{t - x_i}{h}\right) = \int \tilde{\mathcal{K}}\left(\frac{t - x}{h}\right) F_n(dx) .$$

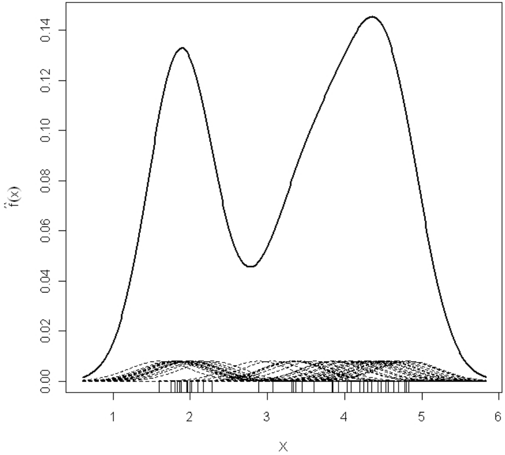
Gauß–Kernschätzer (fünf Moden)

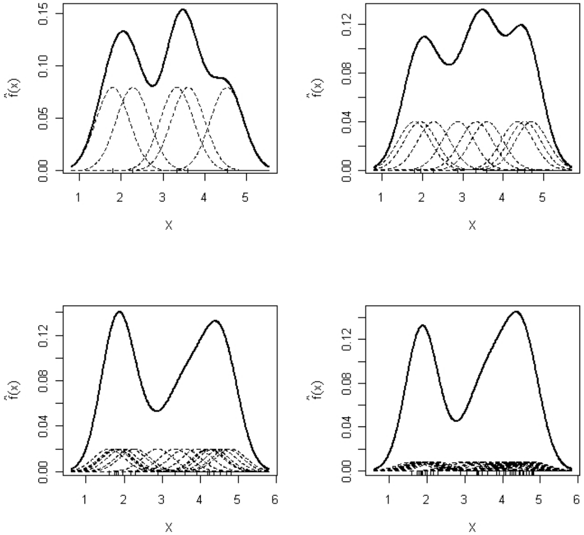


Gauß–Kernschätzer (neun Moden)



Gauß–Kernschätzer (50 Moden)





f̂(x)
X

Entscheidende Schwierigkeit: Wahl der Bandweite!

h zu groß $\longrightarrow$ *oversmoothing*
$\longrightarrow$ lokale Extrema werden nicht erkannt, zu glatt

h zu klein $\longrightarrow$ *undersmoothing*
$\longrightarrow$ lokale Moden, Schätzer ist „hairy“

Übersicht

1. Histogramm–Schätzer und empirische Verteilungsfunktion
2. Univariate Kerndichteschätzung
3. Bandweitenwahl für Kernschätzer
4. Multivariate Kerndichteschätzung

Bias

Satz

Wenn $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern ist, der die oben genannten Regularitätsbedingungen erfüllt, und $f \in \mathcal{C}^2(\mathbb{R})$, so gilt:

$$\mathbb{E}_f[\hat{f}_n(x)] - f(x) = \frac{h^2}{2} f''(x) \int x^2 \mathcal{K}(x) dx + \mathcal{o}(h^2) \; .$$

Bias

Beweis über Taylor–Entwicklung von f:

$$\begin{aligned}
f(x-ht) &= f(x) - htf'(x) + \frac{h^2t^2}{2}f''(x) + \mathcal{o}(h^2t^2) \\
\mathbb{E}_f[\hat{f}_n(x)] - f(x) &= \int \frac{1}{h}\mathcal{K}\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\
&= -hf'(x)\underbrace{\int y\mathcal{K}(y)dy}_{=0} + \\
&\quad \frac{h^2f''(x)}{2}\underbrace{\int y^2\mathcal{K}(y)dy}_{=k_2} + \mathcal{o}(h^2t^2) \\
&= \frac{h^2f''(x)}{2}k_2 + \mathcal{o}(h^2t^2).
\end{aligned}$$

Varianz

Satz

Wenn $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern ist, der die oben genannten Regularitätsbedingungen erfüllt, und $f \in \mathcal{C}(\mathbb{R})$, so gilt:

$$\mathbb{V}\mathrm{ar}_f\left(\hat{f}_n(x)\right) = \frac{1}{nh} f(x) \int \mathcal{K}^2(y)dy + \mathcal{o}\left(n^{-1}h^{-1}\right) .$$

Varianz

Beweis:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\hat{f}_n(x)\right) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left(\mathcal{K}\left(\frac{x-x_i}{h}\right)\right)\\
&= \frac{1}{nh^2}\int \mathcal{K}^2\left(\frac{x-y}{h}\right)f(y)dy - \frac{1}{n}\left(\mathbb{E}[\hat{f}_n(x)]\right)^2\\
&= \frac{1}{nh}\int \mathcal{K}^2(y)f(x-yh)dy - n^{-1}\left(\mathbb{E}[\hat{f}_n(x)]\right)^2\\
&= \frac{1}{nh}f(x)\int \mathcal{K}^2(y)dy + \mathcal{O}\left(n^{-1}h^{-1}\right).
\end{aligned}$$

Mean Squared Error: Bias–Varianz–Zerlegung

$$\mathbb{E}_f\left[\left(\hat{f}_n(x) - f(x)\right)^2\right] = \mathrm{Bias}(\hat{f}_n(x|h))^2 + \mathbb{V}\mathrm{ar}(\hat{f}_n(x|h))$$

$$= h^4\left(\frac{f''}{2}k_2\right)^2 + \frac{1}{nh}f(x)\int \mathcal{K}^2(y)dy + o\left(h^4 + n^{-1}h^{-1}\right) .$$

$\Rightarrow$ Trade–Off zwischen Bias und Varianz

Anmerkung:

Bias hängt nicht explizit vom Stichprobenumfang n ab.

Optimaler Kern

Minimierung des MISE bezüglich h ergibt optimale Bandweite:

$$h_{opt} = \frac{\left(\int \mathcal{K}^2(y)dy\right)^{1/5}}{n^{1/5} k_2^{1/5} \left(\int (f''(y))^2 \, dy\right)^{1/5}} \,. \tag{1}$$

Setzt man h_{opt} in den MISE ein, erhält man

$$\text{MISE} \approx \frac{5}{4} \left(k_2^{2/5} \left(\int \mathcal{K}^2(y)dy \right)^{4/5} \right) \left(\int (f''(y))^2 \, dy \right)^{1/5} \,.$$

Minimal für Epanechnikov–Kern:

$$\mathcal{K}_e(x) = \frac{3}{4\sqrt{5}} \left(1 - \frac{x^2}{5} \right) \mathbb{1}_{[-5,5]}(x) \,.$$

Optimaler Kern

Setzt man h_{opt} in den MISE ein, erhält man

$$\text{MISE} \approx \frac{5}{4}\left(k_2^{2/5}\left(\int \mathcal{K}^2(y)dy\right)^{4/5}\right)\left(\int (f''(y))^2\, dy\right)^{1/5}.$$

Minimal für Epanechnikov–Kern:

$$\mathcal{K}_e(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbb{1}_{[-5,5]}(x).$$

Effizienz $eff(\mathcal{K})$ für $\mathcal{K} \neq \mathcal{K}_e$ und n gegeben:
Zahl $eff(\mathcal{K})$ löst Gleichung $\text{MISE}(n, \mathcal{K}) = \text{MISE}(n \cdot eff(\mathcal{K}), \mathcal{K}_e)$.

Gauß–Kern: Effizienz von ca. 0.95,
Rechteck–Kern: Effizienz von ca. 0.93.

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb“
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb“
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb“
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb"
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb“
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Andere Methoden zur Bandweitenwahl

Methoden der ersten Generation:

- Silverman's „rule of thumb"
- biased und unbiased cross validation

Methoden der zweiten Generation:

- Plug-in Methode von Sheather and Jones
- smoothed Cross Validation
- Kontrastmethoden
- Bootstrap/Jackknife–Methoden

Silverman's „rule of thumb“

Unter Normalverteilungsannahme ($f = \varphi$) kann man die optimale Bandweite (1) in Abhängigkeit von σ berechnen:

$$\int (f''(x))^2\, dx = \sigma^{-5} \int (\varphi(x))^2\, dx = \frac{3}{8\sqrt{\pi}}\sigma^{-5} \approx 0.212\sigma^{-5},$$

$$\Rightarrow h_{opt} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \approx 1.06\sigma n^{-1/5}\ .$$

Schätzt man σ aus den Daten, lässt sich damit auch h_{opt} schätzen.

Kleinste–Quadrate–Kreuzvalidierung

Betrachte integrierten quadratischen Fehler
$\int \left(\hat{f} - f\right)^2 = \int \hat{f}^2 - 2\int \hat{f} f + \int f^2$.
Entscheidend: Mittleren Term schätzen.

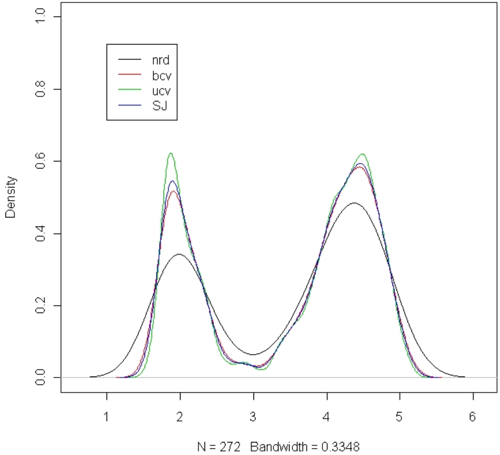
(Leave one out) Least Squares Cross Validation (LSCV):

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j \neq i} \mathcal{K}\left(\frac{x - X_j}{h}\right) .$$

Da der Bias nur von h und $\mathcal{K}$ abhängt, aber nicht von n, gilt:

$$\mathbb{E}_f\left[\frac{1}{n}\sum_i \hat{f}_{-i}(X_i)\right] = \mathbb{E}_f\left[\hat{f}_n(X_n)\right] = \mathbb{E}_f\left[\hat{f}(X)\right] .$$

Vergleich von Bandweiten



Übersicht

1. Histogramm–Schätzer und empirische Verteilungsfunktion
2. Univariate Kerndichteschätzung
3. Bandweitenwahl für Kernschätzer
4. Multivariate Kerndichteschätzung

mehrdimensionale Kernfunktionen

Modell: $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ i. i. d. $\sim f$.

Definition (p-dimensionaler Kern)

Eine Funktion $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ mit

$$\int_{\mathbb{R}^p} \mathcal{K}(\mathbf{y}) d\mathbf{y} = 1 \quad \text{und}$$

Regularitätsbedingungen:

- $\mathcal{K}$ ist radialsymmetrische Wahrscheinlichkeitsdichte
- Beschränkter Träger oder zumindest $|\mathbf{x}|\mathcal{K}(\mathbf{x}) \to 0$ für $|\mathbf{x}| \to 0$

Mehrdimensionale Kernfunktionen

Beispiele:

uniformer Kern $\mathcal{K}(\mathbf{x}) = \frac{1}{v_p}$ für $\mathbf{x}^T\mathbf{x} \leq 1$,

Gaußkern $\mathcal{K}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)$,

Epanechnikovkern $\mathcal{K}(\mathbf{x}) = \frac{1+p/2}{v_p}(1 - \mathbf{x}^T\mathbf{x})$, $\mathbf{x}^T\mathbf{x} \leq 1$.

Multivariater Kernschätzer

Definition

Sei $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ ein Kern.

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \quad x \in \mathbb{R}^p$$

heißt multivariater Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

Multivariater Kernschätzer

Definition

Sei $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ ein Kern.

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \quad x \in \mathbb{R}^p$$

heißt multivariater Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

Bias und Varianz:

$$Bias_h\left(\hat{f}_n(\mathbf{x})\right) = \frac{h^2}{2}\Delta f(\mathbf{x}) \int y_1^2 \mathcal{K}(\mathbf{y}) d\mathbf{y} + o(h^2)\,,$$

$$\mathbb{V}\mathrm{ar}_h\left(\hat{f}_n(\mathbf{x})\right) = \frac{1}{nh^p} \int \mathcal{K}^2(\mathbf{y}) d\mathbf{y} + o\left(n^{-1}h^{-p}\right)\,.$$

Multivariater Kernschätzer

Definition

Sei $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ ein Kern.

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \quad x \in \mathbb{R}^p$$

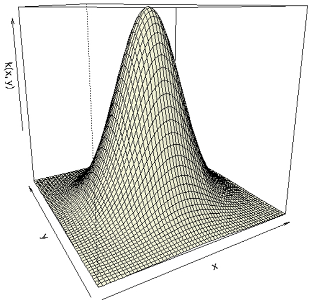
heißt multivariater Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

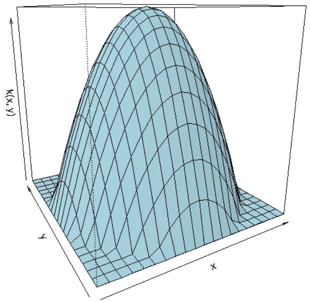
Minimierung des MISE:

$$h_{opt}^{p+4} = \frac{p}{n} \frac{\int \mathcal{K}^2(\mathbf{y})d\mathbf{y}}{\left(\int y_1^2 \mathcal{K}(\mathbf{y})d\mathbf{y}\right)^2} \left(\int (\Delta f(\mathbf{y}))^2 \, d\mathbf{y}\right) .$$

Darstellung zweidimensionaler Kernfunktionen

Gaußkern und Epanechnikovkern mit $p = 2$





Verschiedene Bandweiten in unterschiedliche Richtungen

Allgemeiner als in obiger Definition kann man den multivariaten Kerndichteschätzer mit einer Bandweitenmatrix H definieren:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n|H|^{1/2}} \mathcal{K}\left(H^{-1/2}(\mathbf{x} - \mathbf{X}_i)\right), \mathbf{x} \in \mathbb{R}^p.$$

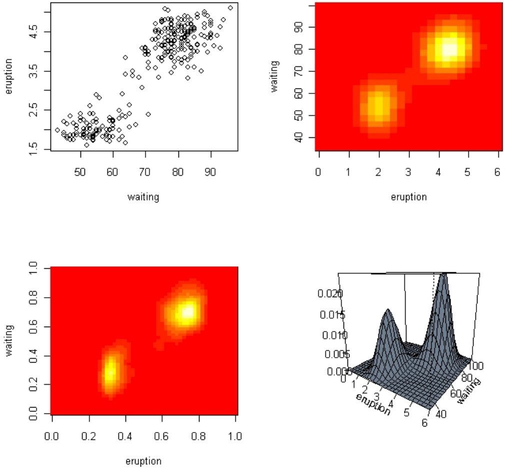
Zuvor: $H = h\mathbb{1}_p$, wobei $\mathbb{1}_p$ die p–dimensionale Einheitsmatrix bezeichnet

In R: Diagonalmatrix H angebbar.

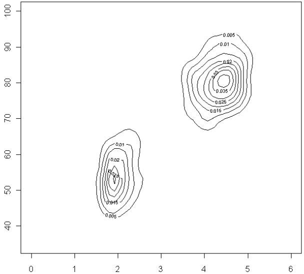
R Code: Zweidimensionale Kerndichteschätzung

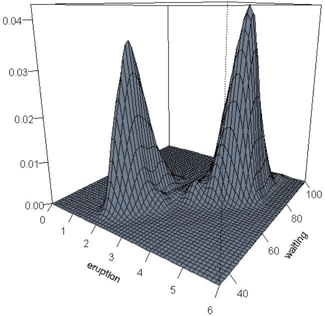
```
> library(MASS)
> library(KernSmooth)
> data(faithful)
> x<-faithful$eruptions
> y<-faithful$waiting
> par(mfrow=c(2,2),pty="m")
> plot(y,x,ylab="eruption",xlab="waiting")

> z<-kde2d(x,y,lims=c(0,6,35,100))
> zz<-bkde2D(faithful,range.x=list(c(0,6),c(35,100)),
+ bandwidth=c(bw.SJ(x),bw.SJ(y)))
> image(z,xlab="eruption",ylab="waiting")
> image(zz$fhat,xlab="eruption",ylab="waiting")
> persp(z,col="slategrey",theta=35,xlim=c(0,6),ylim=c(35,100),
+ ticktype="detailed",xlab="eruption",ylab="waiting",zlab="")
```

Zweidimensionale Kerndichteschätzung



```
> persp(zz$x1,zz$x2,zz$fhat,col="slategrey",
+ theta=35,xlim=c(0,6),ylim=c(35,100),
+ ticktype="detailed",xlab="eruption",
+ ylab="waiting",zlab="")
> contour(zz$x1,zz$x2,zz$fhat)
```





Literatur

- Bernard W. Silverman
  Density estimation for statistics and data analysis.
  Chapman and Hall/CRC, 1986.

- Simon J. Sheather and M.C. Jones
  A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation.
  Journal of the Royal Statistical Society, 53(3):683-690, 1991.

- David W. Scott
  Multivariate Density Estimation: Theory, Practice and Visualization.
  New York: Wiley, 1992.

Literatur

- J. Steven Marron and Peter Hall and Byeong U. Park
  Smoothed Cross–Validation.
  Probability Theory and Related Fields, 92:1-20, 2003.

- A. R. Mugdadi and Ibrahim A. Ahmad
  A Bandwidth Selection for Kernel Density Estimation of Functions of Random Variables.
  Computational Statistics and Data Analysis, 47(1):49-62, 1992.