Skript zur Vorlesung :

Statistik und Numerik

Vorlesung SS 2024

Prof. Dr. Jens Timmer

16. Juli 2024

Inhaltsverzeichnis

1 Einleitung

Technicalities:

- Welche Semester ? Auch Master-Studenten ?
- Skript und Kommunikation über homepage
- Übungen
- Inhaltsverzeichnis
- Wenn etwas unklar: Fragen !

Literatur:

- Zur Statistik
  - A. Bevan. *Statistical Data Analysis for the Physical Science* [6]
  - J. Honerkamp. *Stochastic Dynamical Systems* [25] Kap. 1-3.
    Kondensierte Darstellung der für Physiker relevanten Grundlagen der Statitstik
  - J. Hartung. *Statistik* [24] Ein Klassiker, sehr detailliert
  - L. Sachs. *Applied Statistics* [59] Kompendium, angewandt
  - D.R. Cox, D.V. Hinkley. *Theoretical Statistics* [12] lesbares Theoriebuch
- Zur Numerik
  - W. Press et al. *Numerical Recipes* [50]:
    Die Bibel, optimal für Physiker
  - J. Stoer. *Einführung in die Numerische Mathematik I & II* [66, 67]
    Mathematisch orientierter Klassiker
  - Weitere Bücher aus dem Umfeld 'Computational Physics' und 'Monte-Carlo Methoden' : [8, 17, 18, 36] Franklin modern

Teil I
Statistik

Grundsätzliches zur Statistik :

- Ein paar Sachen muß man richtig verstehen.
- Vieles muß man kennen.
- Vieles muß man einfach nur nachschlagen können.
- Angewandte Statisitk ist kein Fall von Mathematik

2 Verteilungen

2.1 Zufallsvariablen

Zufallsvariable X:

- Etwas, das eine Wahrscheinlichkeitsdichte $p_X(x)$ hat
- Wahrscheinlichkeit, eine Realisierung x in $(x, x+dx)$ zu beobachten, ist $p_X(x)dx$
- $p_X(x) \geq 0$, $\int p_X(x)\, dx = 1$

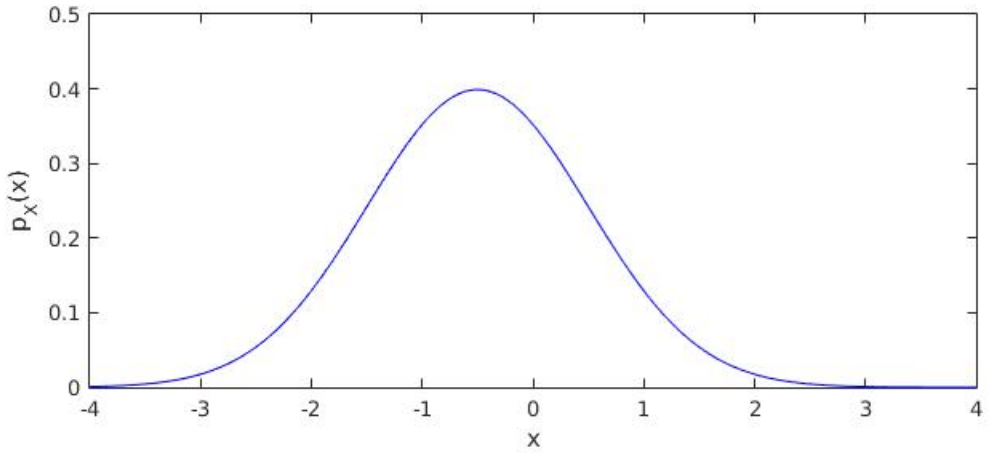


Abbildung 2.1: Beispiel für eine Wahrscheinlichkeitsdichte

- Physikalisch entsteht Zufall entweder durch
  - Quantenmechanik, eher selten in makroskopischen komplexen Systemen
  - Chaos, auch selten, bei Würfel realisiert
  - viele Einflüsse á la Brownian Motion, meistens.
- Es gibt auch diskrete Verteilungen $p(x_i)$, denke an Würfel

Im folgenden, wenn Bezug klar: $p_X(x) = p(x)$

2.2 Momente und Kumulanten

- Erwartungswert $\langle f(x) \rangle$

$$\langle f(x) \rangle = \int f(x)\, p(x)\, dx$$

  Erwartungswert ist eine Zahl

- Momente μ_k

$$\mu_k = \langle x^k \rangle = \int x^k\, p(x)\, dx$$

- 1. Moment: Mittelwert

$$\mu_1 = \bar{x} = \mu = \langle x \rangle = \int x\, p(x)\, dx$$

- 2. Moment

$$\mu_2 = \langle x^2 \rangle = \int x^2\, p(x)\, dx$$

  - Varianz: $\sigma^2 = \langle (x - \bar{x})^2 \rangle = \mu_2 - \mu_1^2$
  - Standardabweichung: σ
  - Addiert man unabhängige Zufallsvariablen, so sind Varianzen, nicht Standardabweichungen additiv.

- 3. Moment

$$\mu_3 =< x^3 >= \int x^3 p(x)\, dx$$

  Schiefe:

$$\kappa = \frac{\langle (x - \mu)^3 \rangle}{\sigma^3}$$

  Maß für Asymmetrie.

- 4. Moment

  Kurtosis (Bauchigkeit): $\gamma = \langle (x - \mu)^4 \rangle / \sigma^4 - 3$

  "3" erklärt sich weiter unten

- Charakteristische Funktion oder Erzeugende Funktion

$$G(k) =< e^{ikX} >= \int dx\, e^{ikx} p(x)$$

  Fouriertransformierte der Dichte $p(x)$

- Existieren die Momente, i.e. $< X^n >< \infty$, Taylor-Entwicklung

$$G(k) = \sum_{n=0}^{\infty} \frac{(i\,k)^n}{n!} < X^n >$$

Ist $G(k)$ bekannt, berechnen sich Momente durch:

$$\left.\frac{d^n G(k)}{dk^n}\right|_{k=0} = i^n < X^n >$$

- Entwickelt man $\log(G(k))$ nach k, folgt:

$$\log(G(k)) = \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} \kappa_n$$

mit den Kumulanten κ_i

$$\begin{aligned} \kappa_1 &= \mu_1 \\ \kappa_2 &= \mu_2 - \mu_1^2 = \sigma^2 \\ \kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \\ &\dots \end{aligned}$$

Wichtige Eigenschaften:

  - Kumulanten sind additiv, daher natürliche Größen
    Sei

$$Y = \sum_{i=1}^{N} X_i$$

    dann folgt

$$\kappa_n(Y) = \sum_{i=1}^{N} \kappa_n(X_i)$$

    Varianz ist additiv, nicht Standardabweichung
  - Man kann zeigen:
    * Entweder: Alle Kumulaten außer den ersten beiden verschwinden
    * Oder: Es existieren ∞ viele

2.3 Beispiele für Verteilungen

- Gaußverteilung oder Normalverteilung:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - Bezeichnung: $N(\mu, \sigma^2)$
    Standard-Gaußverteilung $\mu = 0$, $\sigma^2 = 1$: $N(0, 1)$
    In $\pm 1\sigma$ liegt 68 % der Masse
    In $\pm 1.96\sigma$ liegt 95 % der Masse
  - Momente von $N(0, 1)$:

$$\langle x^k \rangle = \begin{cases} 0 & \text{für k ungerade} \\ 1 \times 3 \times \cdots \times (k-1) & \text{für k gerade} \end{cases}$$

    Damit klar, warum bei Kurtosis "-3"
  - Charakteristische Funktion

$$G(k) = e^{i\mu k - \frac{1}{2}\sigma^2 k^2}$$

    Nur die ersten zwei Kumulanten sind ungleich null.
    Macht klar, warum Gaußverteilung so besonders

Zentraler Grenzwertsatz:

Existeren die ersten beiden Momente, so strebt die (normierte) Summe von unabhängigen, identisch verteilten (independent, identically distributed: iid) Zufallsvariablen gegen eine Gaußverteilung.

Betrachte N identische Zufallsvariablen X_i mit

- $\kappa_1(X_i) =< X_i >= 0$
- $\kappa_2(X_i) = \mu_2 - \mu_1^2 = \sigma^2$
- $\kappa_n(X_i) < \infty \quad \forall n$

Bilde:

$$Y = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$$

Für Kumulanten folgt:

$$\kappa_n(Y) = \frac{1}{N^{n/2}} \sum_{i=1}^{N} \kappa_n(X_i)$$

Speziell:

$$\kappa_2(Y) = \frac{1}{N} \sum_{i=1}^{N} \kappa_2(X_i) = \kappa_2 = \sigma^2$$

$$n > 2: \quad \kappa_n(Y) = \kappa_n(X_i) \frac{1}{N^{(n-2)/2}}$$

- Die höheren Kumulanten verschwinden mit N.
- Verteilung geht gegen Gauß-Verteilung
  Darum sind die beiden Größen μ und σ so wichtig.
- Gilt auch wenn X_i nicht identisch
- Konvergenzrate, i.e. wie schnell geht es zur Gaußverteilung, hängt i.w. von der Schiefe ab.
- Bedeutung von Zentralem Grenzwertsatz nicht zu unterschätzen.
- Bei Multiplikation führt es zur Log-Normal Verteilung

Mitteln:

$$Y = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\kappa_n(Y) = \kappa_n(X_i) \frac{1}{N^{n-1}}$$

$$\kappa_2(Y) = \frac{\sigma^2}{N}$$

- Gleichverteilung $U(a, b)$:

$$p(x) = \begin{cases} 1/(b-a) & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

- Exponential-Verteilung

$$p(x) = \frac{1}{\tau}\, e^{-x/\tau}$$

Es gilt:

$$\begin{aligned} \mu &= \tau \\ \sigma^2 &= \tau^2 \end{aligned}$$

Erwartungswert und Varianz sind keine unabhängigen Parameter.

Ergibt sich bei "konstanter Ausfallrate"

- χ^2_r Verteilung mit r Freiheitsgraden:
Summe von r quadrierten Gaußverteilungen

$$"\chi^2_r = \textstyle\sum_{i=1}^{r} (N(0,1))^2"$$

$$Y \sim \chi^2_r, \quad X_i \sim N(0,1) = p_G(x_i)$$

$$\begin{aligned} p(y) &= \int dx_1 \ldots dx_r\, \delta(y - (x_1^2 + \ldots + x_r^2)) \prod_{i=1}^{r} p_G(x_i) \\ &= \int dx_1 \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} dx_2 \frac{1}{\sqrt{2\pi}} e^{-x_2^2/2} \ldots dx_r \frac{1}{\sqrt{2\pi}} e^{-x_r^2/2}\, \delta(y - (x_1^2 + \ldots + x_r^2)) \\ &= \frac{y^{r/2-1} e^{-y/2}}{2^{r/2}\, \Gamma(r/2)}, \qquad \Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt \end{aligned}$$

Es gilt:

$$\begin{aligned} \langle \chi^2_r \rangle &= r \\ Var(\chi^2_r) &= 2r\,, \end{aligned}$$

d.h. Erwartungswert und Varianz sind keine unabhängigen Parameter.

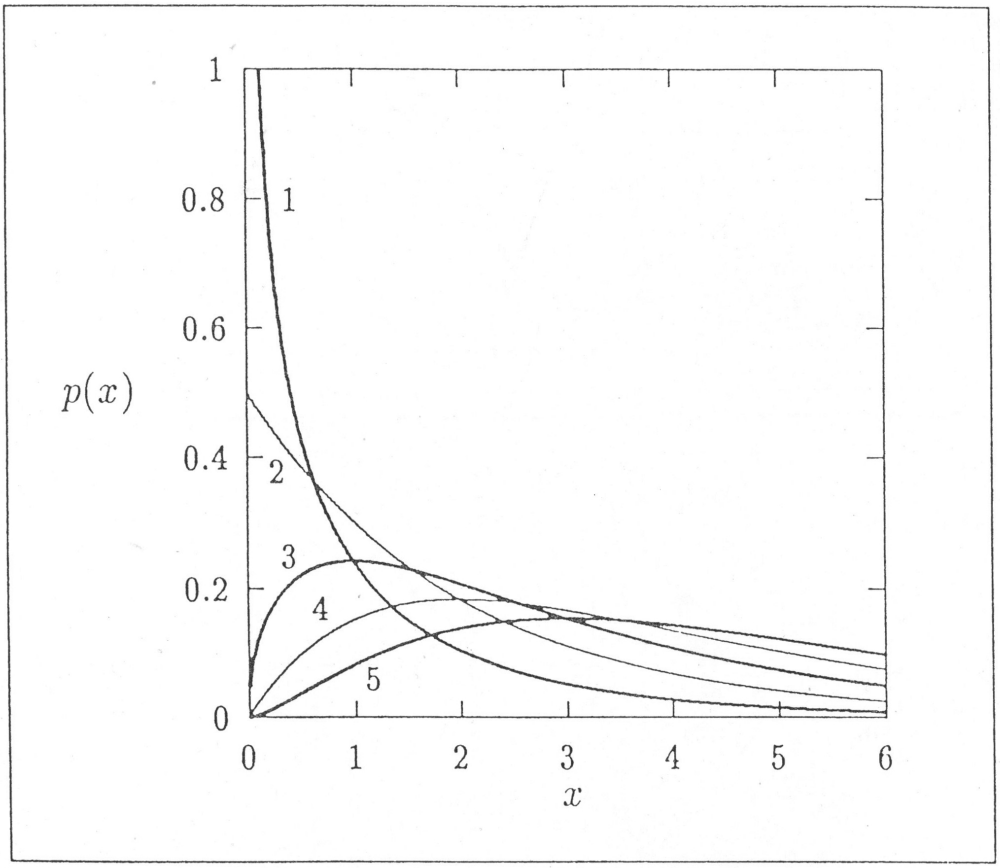


Abbildung 2.2: χ^2_r-Verteilungen mit $r = 1, 2, 3, 4, 5$ Freiheitsgraden

1.18

χ^2 Verteilungen sind additiv:

$$"\chi^2_{r_1} + \chi^2_{r_2} = \chi^2_{r_1+r_2}"$$

Aus dem zentralen Grenzwertsatz folgt:

$$\lim_{r\to\infty} \chi^2_r = N(r, 2r)$$

Bemerkungen :

  - $\chi^2_2 = \frac{1}{2}e^{-x/2}$ ist Exponential-Verteilung mit $\tau = 2$.
  - χ^2_2 wird wichtig in Kapitel 13 Spektralanalyse.

1.24

- t-Verteilung

$$t(r,x) = \frac{N(0,1)}{\sqrt{\chi^2_r/r}} = \frac{1}{\sqrt{r}}\frac{1}{B(1/2,r/2)}\left(1+\frac{x^2}{r}\right)^{-\frac{1}{2}(r+1)}, \quad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

  r: Anzahl der Freiheitsgrade

  Siehe Kapitel 3.1: t-Test: Test auf Gleichheit von Mittelwerten zweier Gaußverteilungen.

$$\lim_{r\to\infty} t(r,x) = N(0,1), \qquad \text{gute Näherung ab } r = 30$$

- F-Verteilung

$$F(r_1,r_2,x) = \frac{\chi^2_{r_1}/r_1}{\chi^2_{r_2}/r_2} = ...$$

  r_1, r_2: Jeweilige Anzahl der Freiheitsgrade

  F-Test: Test auf Gleichheit von Varianzen zweier Gaußverteilungen.

- Cauchy(Lorentz) Verteilung:

$$p_{Cauchy}(x,a,\gamma) = \frac{1}{\pi}\frac{\gamma^2}{(x-a)^2+\gamma^2}$$

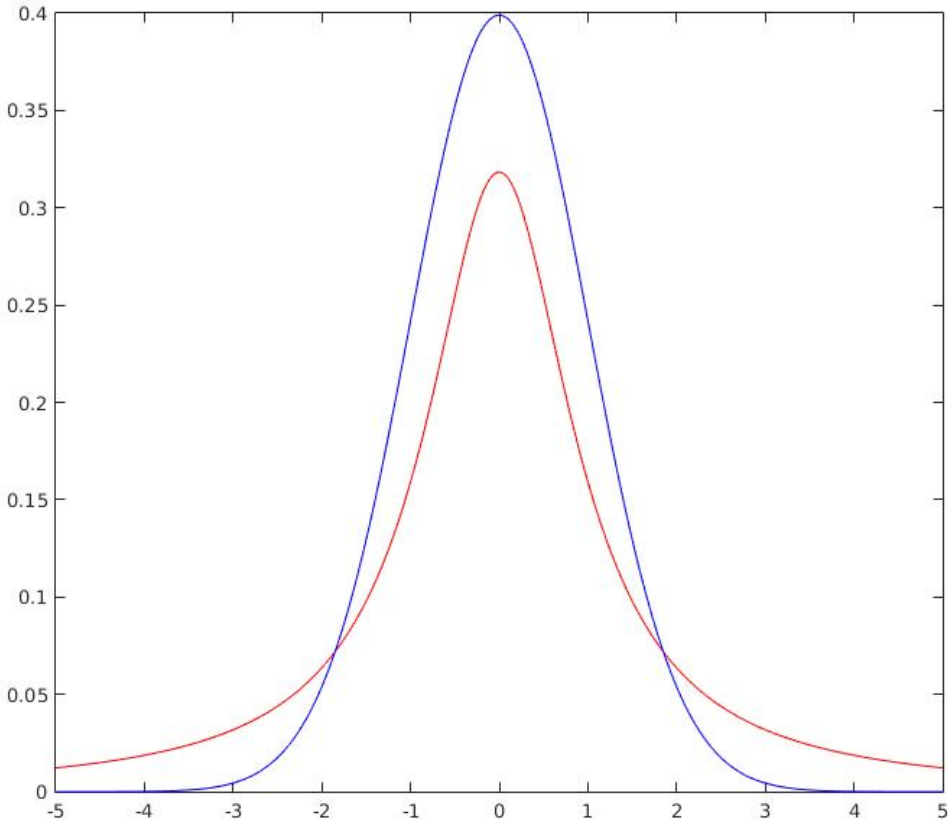


Abbildung 2.3: Cauchy-Verteilung (rot) im Vergleich zur Gaußveteilung (blau)

- Momente existieren nicht!
- Charakterisische Funktion:

$$G(k) = e^{ika-|k|\gamma}$$

  Es existiert keine Taylor-Entwicklung um $k = 0$
- a ist Lokalisationsparameter, aber kein Mittelwert.
- Zeichnung Normal- und Cauchy-Verteilung
- Cauchy-Verteilung spielt bei den Zuwächsen der Aktienkurse eine Rolle. Eventuell Ausflug: Black-Scholes

- Zentraler Grenzwertsatz gilt nicht für Cauchy-Verteilung.

  Aber auch für Verteilungen mit nicht-existenten Momenten gibt es Grenzwertsätze. Stichwort "Stabile Verteilungen"

- Einfachste Erzeugung:

$$''p_{Cauchy}(x) = \frac{N(0,1)}{N(0,1)}'' = \int dy \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \int dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, \delta(x - \frac{y}{z})$$

- Bezug zur t-Verteilung:

$$t(1,x) = p_{Cauchy}(x,0,1)$$

  Mit t-Verteilung kann man also zwischen Cauchy (keine Momente existieren) und Gauß (alle Momente existieren) hin- und herfahren.

- Cauchy-Verteilung in der Physik bekannt als Breit-Wigner-Verteilung.

- Multivariate Gaußverteilung

$$p(\vec{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{|C|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})\right), \qquad d = \dim(\vec{x})$$

  mit Kovarianzmatrix C

$$C = \langle (\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T \rangle$$

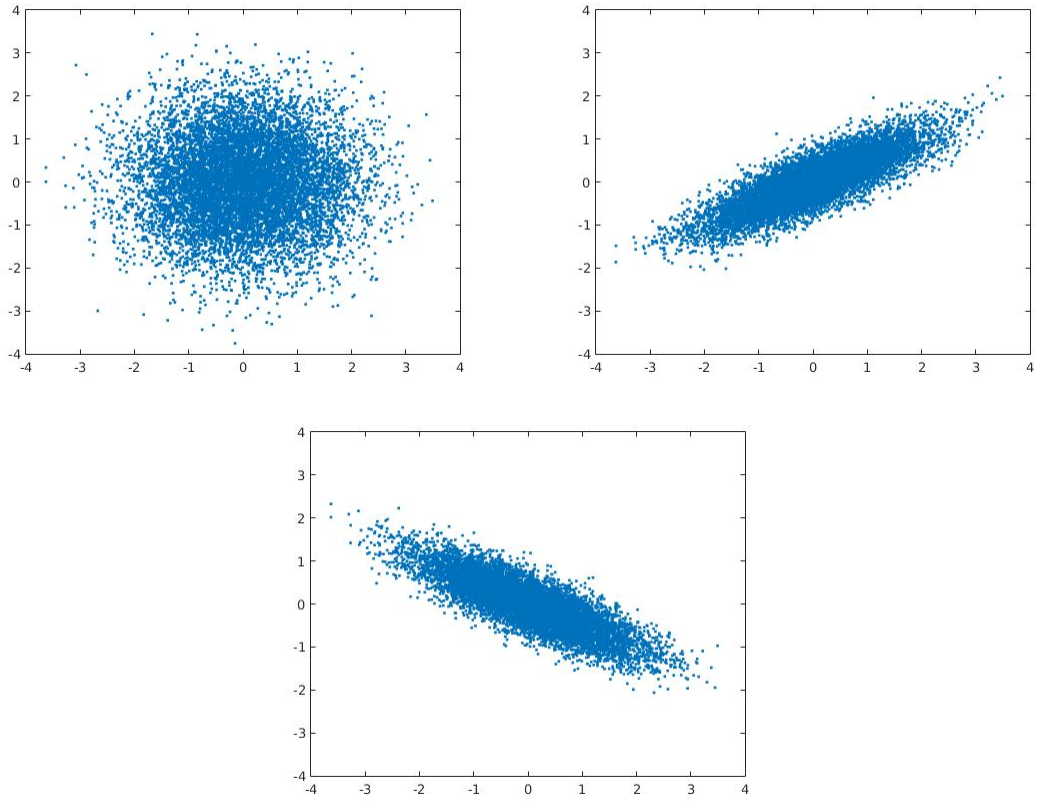


Abbildung 2.4: *2D*-Gaußverteilte-Zufallszahlen mit $C_1 = (0.71\ 0; 0\ 0.70)$, $C_2 = (0.78\ 0.39; 0.39\ 0.28)$, $C_3 = (0.79, -0.39;\ -0.39, 0.28)$

- 1 D Normalverteilung:
  68 % der Masse in $[-\sigma, \sigma]$
  99 % der Masse in $[-3\sigma, 3\sigma]$
- 10 D Normalverteilung, $C = \mathbb{1}$:
  99 % der Masse <u>außerhalb</u> der $[-3\sigma, 3\sigma]$-Kugel.
- Intuition:
  - Integration über die Winkel ausführen
  - Verbleibt, $d = \dim(\vec{x})$

$$\text{Masse innerhalb von Radius } r \sim \int_0^r r^{d-1}\, e^{-r^2/2}\, dr$$

– Es gibt praktisch nur die längsten Abstände, der Raum ist leer, "curse of dimensionality", kommt wieder in Kap. 12 Kernschätzer.

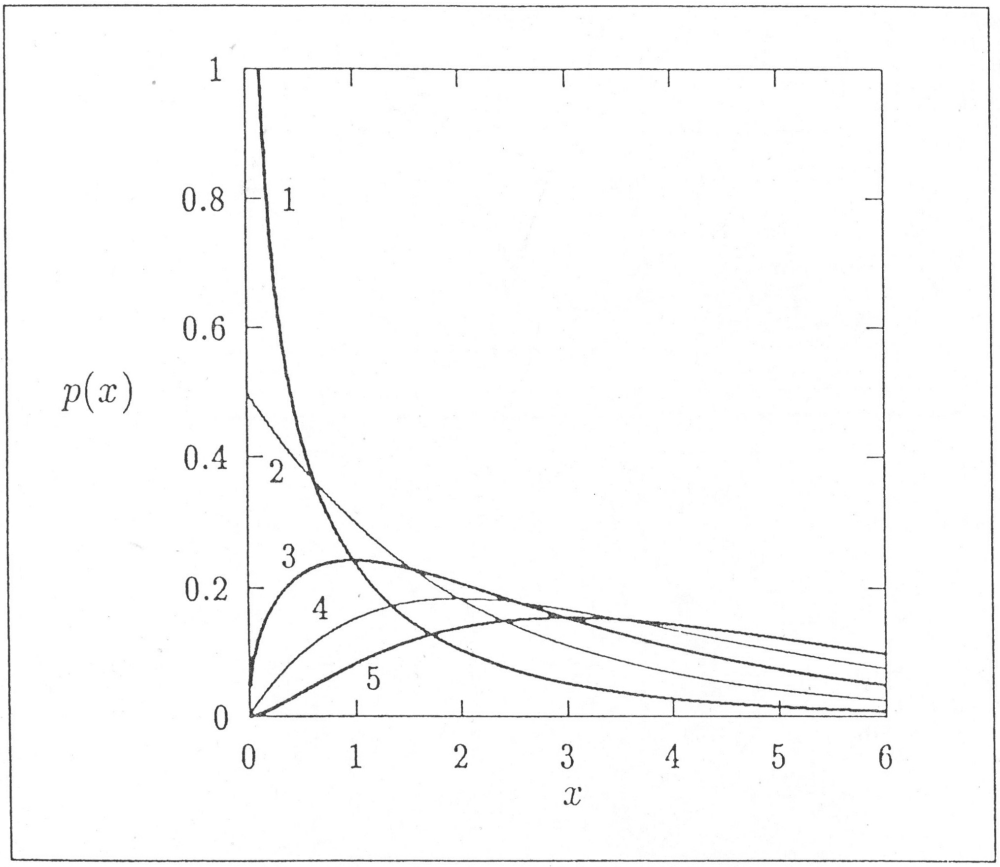


Abbildung 2.5: χ^2_r-Verteilungen mit $r = 1, 2, 3, 4, 5$ Freiheitsgraden

1. Woche

- Binomial Verteilung

$$B(n, p, k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots n$$

Zwei mögliche Ereignisse: x_1, x_2; $p = prob(x_1)$

Bei n-maliger Durchführung des Experimentes ist $B(n, p, k)$ die Wahrscheinlichkeit, x_1 k-mal zu realisieren.

- Poisson-Verteilung

$$P(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{N}_0,$$

  - Wahrscheinlichkeit für k Ereignisse in einem Zeitintervall
  - λ: Durchschnittliche Anzahl der Ereignisse in Zeitintervall
  - Wichtig für Punktprozesse mit konstanter Rate denke an Photonen-Zählprozesse, e-mails pro Stunde, feuernde Neuronen
  - Zusammenhang zu Dynamischen Systemen an Hand von integrate-and-fire Neuron erläutern

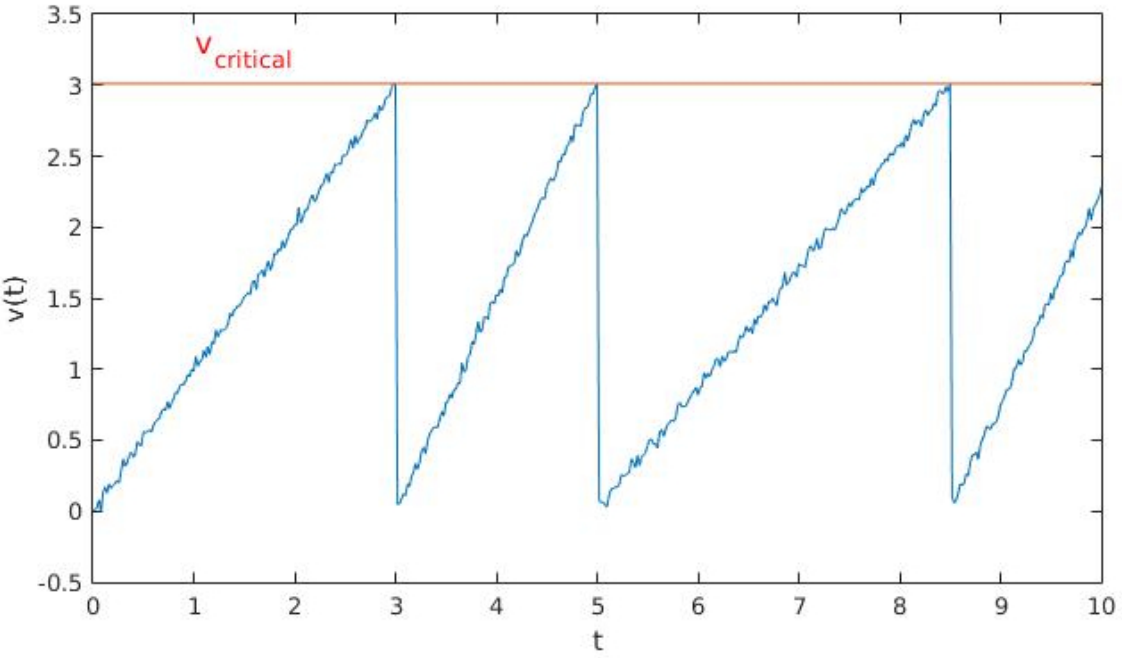


Abbildung 2.6: Integrate-and-Fire-Neuron

  - Für Poisson-Verteilung gilt

$$\mu = \sigma^2 = \lambda$$

- Ferner ist sie die Grenzverteilung der Binomialverteilung:

$$\lim_{n\to\infty} B(n,k,p) = P(k,\lambda) \text{ wobei } \lim_{n\to\infty} np = \lambda$$

  Beschreibt "Seltene Ereignisse"
- Poisson-Verteilung für kleine λ sehr unsymmetrisch. Für große, $\lambda > 30$, geht sie gegen Gauß-Verteilung

$$P(k,\lambda) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(k-\lambda)^2}{2\lambda}\right)$$

Kumulative Verteilungen

- Definition:

$$\text{cum}(x) = \int_{-\infty}^{x} dx' p(x')$$

- x_α mit

$$\text{cum}(x_\alpha) = \alpha$$

  nennt man das (100α) % Quantil.

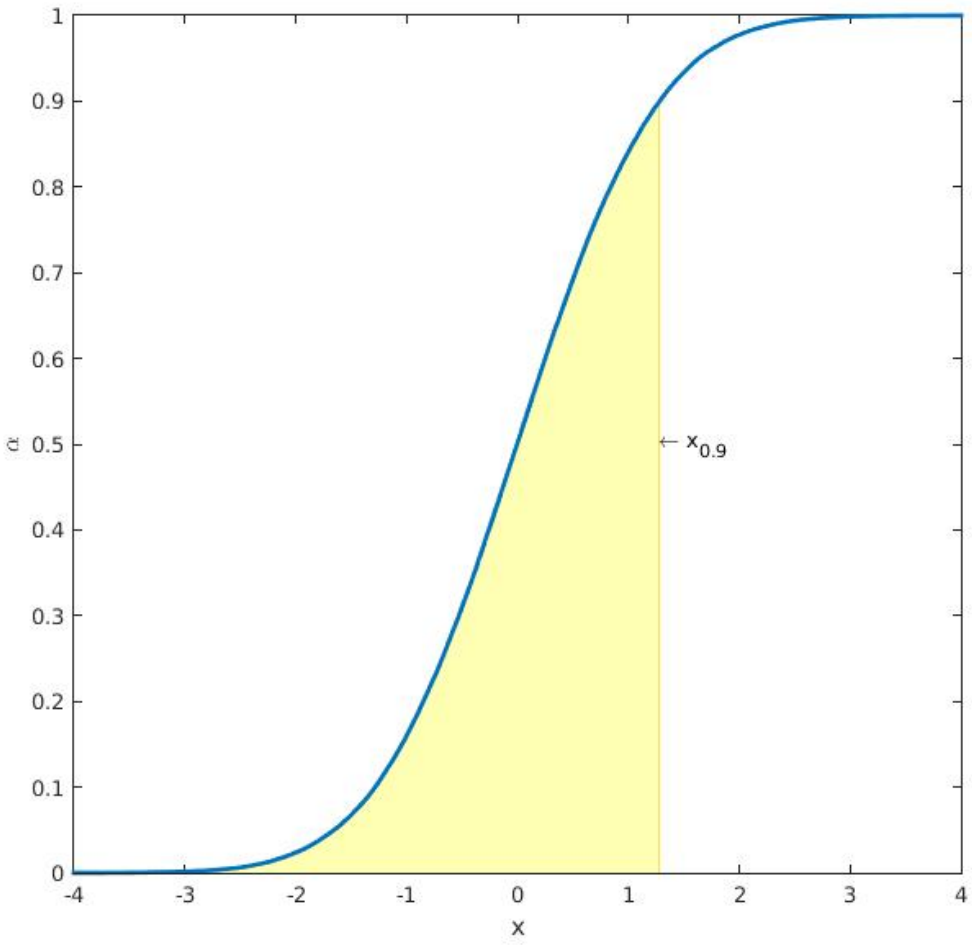


Abbildung 2.7: Kumulative Verteilung der Gauß-Verteilung mit 90%-Quantil

- Wichtig für Testtheorie, Kapitel. 3
- Definition des Medians:

$$\text{cum}(x_{Median}) = 0.5$$

Der mittlere Wert der Verteilung

2.24

2.4 Schätzen von Parametern von Verteilungen

Allgemeine Parameterschätztheorie in Kapitel 4

Definitionen:

- Wahrer Parameter : Θ_0
- Schätzer für Parameter : $\hat{\Theta}$, dies ist eine Zufallsvariable
- Bias (Verzerrung) : $\langle\hat{\Theta}\rangle - \Theta_0$
- Varianz des Schätzers : $\langle(\hat{\Theta} - \langle\hat{\Theta}\rangle)^2\rangle$
- Mittlerer Quadratischer Fehler : $\langle(\hat{\Theta} - \Theta_0)^2\rangle$ = bias^2 + Varianz des Schätzers
- Vertrauensintervall: Bereich um $\hat{\Theta}$, in dem wahrer Wert Θ_0 mit bestimmter Wahrscheinlichkeit liegt.

Gaußverteilung $N(\mu, \sigma^2)$:

- Sei jeweils $X \sim N(\mu_0, \sigma_0^2)$
- Schätzer für den Mittelwert μ

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Als Summe über Gauß-Verteilungen ist $\hat{\mu}$ Gauß-verteilt

$$< \hat{\mu} >= \frac{1}{N}\sum_{i=1}^{N} < X_i >=< X >= \mu_0$$

Schätzer ist unverzerrt oder unbiased. Liegt im Mittel richtig.

Varianz des Schätzers

$$Var(\hat{\mu}) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N}Var(X) = \frac{1}{N}\sigma_0^2$$

Zusammengefaßt: $\hat{\mu}$ ist Gauß'sche Zufallsvariable mit

$$\begin{aligned} <\hat{\mu}> &= \mu_0 \\ Var(\hat{\mu}) &= \frac{1}{N}\sigma_0^2 \\ \sigma(\hat{\mu}) &= \sqrt{\frac{1}{N}}\,\sigma_0 \quad \text{"Standard error of the mean"} \end{aligned}$$

Damit folgt: $\pm\sigma$ (=68%) Vertrauensintervall für wahres μ:

$$[\hat{\mu} - \sigma(\hat{\mu}), \hat{\mu} + \sigma(\hat{\mu})]$$

oder

$$\left[\hat{\mu} - \sqrt{\frac{1}{N}}\,\sigma, \hat{\mu} + \sqrt{\frac{1}{N}}\,\sigma\right]$$

Mit zunehmender Datenanzahl kann man den Mittelwert immer genauer wissen.

Schätzer unbiased und Vertrauensintervall wird mit $\sqrt{\frac{1}{N}}$ kleiner:
Schätzer ist konsistent.
Konsistent: Für $N \to \infty$ wird alles gut

- Drei Schätzer $\hat{S}_k^2$, $k = 1, 2, 3$, für die Varianz

  – Sei Mittelwert unbekannt
  First try:

  $$\hat{S}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2$$

  Betrachte einen Summanden, addiere geschickt 0

  $$\begin{aligned} <(X_i - \hat{\mu})^2> &= <((X_i - < X >) - (\hat{\mu} - < X >))^2> \\ &= Var(X) - 2 <(X_i - < X >)(\hat{\mu} - < X >)> \\ &\quad + Var(\hat{\mu}) \end{aligned}$$

Von oben: $Var(\hat{\mu}) = \frac{1}{N} Var(X)$
und

$$\begin{aligned} < (X_i - < X >)(\hat{\mu} - < X >) > &= \frac{1}{N} \sum_{j=1}^{N} < (X_i - < X >)(X_j - < X >) > \\ &= \frac{1}{N} < (X_i - < X >)^2 > \\ &= \frac{1}{N} Var(X) \end{aligned}$$

Alles zusammen

$$\begin{aligned} < (X_i - \hat{\mu})^2 > &= Var(X) - 2\frac{1}{N} Var(X) + \frac{1}{N} Var(X) \\ &= (1 - 1/N) Var(X) \\ &= \frac{N-1}{N} Var(X) = \frac{N-1}{N} \sigma_0^2 \end{aligned}$$

Somit

$$< \hat{S}_1^2 >= \frac{1}{N} \frac{N-1}{N} \sum_{i=1}^{N} \sigma_0^2 = \frac{N-1}{N} \sigma_0^2 = \sigma_0^2 - \frac{1}{N} \sigma_0^2$$

Ergo: Schätzer $\hat{S}_1^2$ hat einen

$$\text{Bias}(\hat{S}_1^2) = -\frac{1}{N} \sigma_0^2$$

Nur "asymptotisch unverzerrt", d.h. für $N \to \infty$
Diskussion Asymptotik

– Second try

$$\hat{S}_2^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Selbe Rechnung wie oben

$$< \hat{S}_2^2 >= \sigma_0^2$$

Begründung:

* Die Berechnung des Mittelwertes kostet einen Freiheitsgrad.
* $x_1, \ldots, x_N$ unterliegen Nebenbedingung:

$$\sum_{i=1}^{N} x_i = \hat{\mu}$$

* Faktor $\frac{1}{N-1}$ heißt Bessel-Korrektur

– Sei Mittelwert μ bekannt

$$\hat{S}_3^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Selbe Rechnung wie oben

$$< \hat{S}_3^2 >= \sigma_0^2$$

Vertrauensintervall für p der Binomial Verteilung

$$B(n, p, k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots n$$

- Mit $m = \#x_1$, ist der Schätzer

$$\hat{p} = \frac{m}{n}$$

asymptotisch gaußverteilt :

$$\hat{p} \sim N\left(p, \frac{1}{n} p(1-p)\right)$$

Gauß-verteilt wegen Zentralem Grenzwertsatz

- 95 % Konfidenzintervall:

$$\left[\frac{m}{n} - 1.96 \sqrt{\frac{1}{n} \frac{m}{n} (1 - \frac{m}{n})}, \frac{m}{n} + 1.96 \sqrt{\frac{1}{n} \frac{m}{n} (1 - \frac{m}{n})} \right]$$

- Asypmtotik gilt für $n\,p(1-p) > 10$

  Diskussion Asymptotik

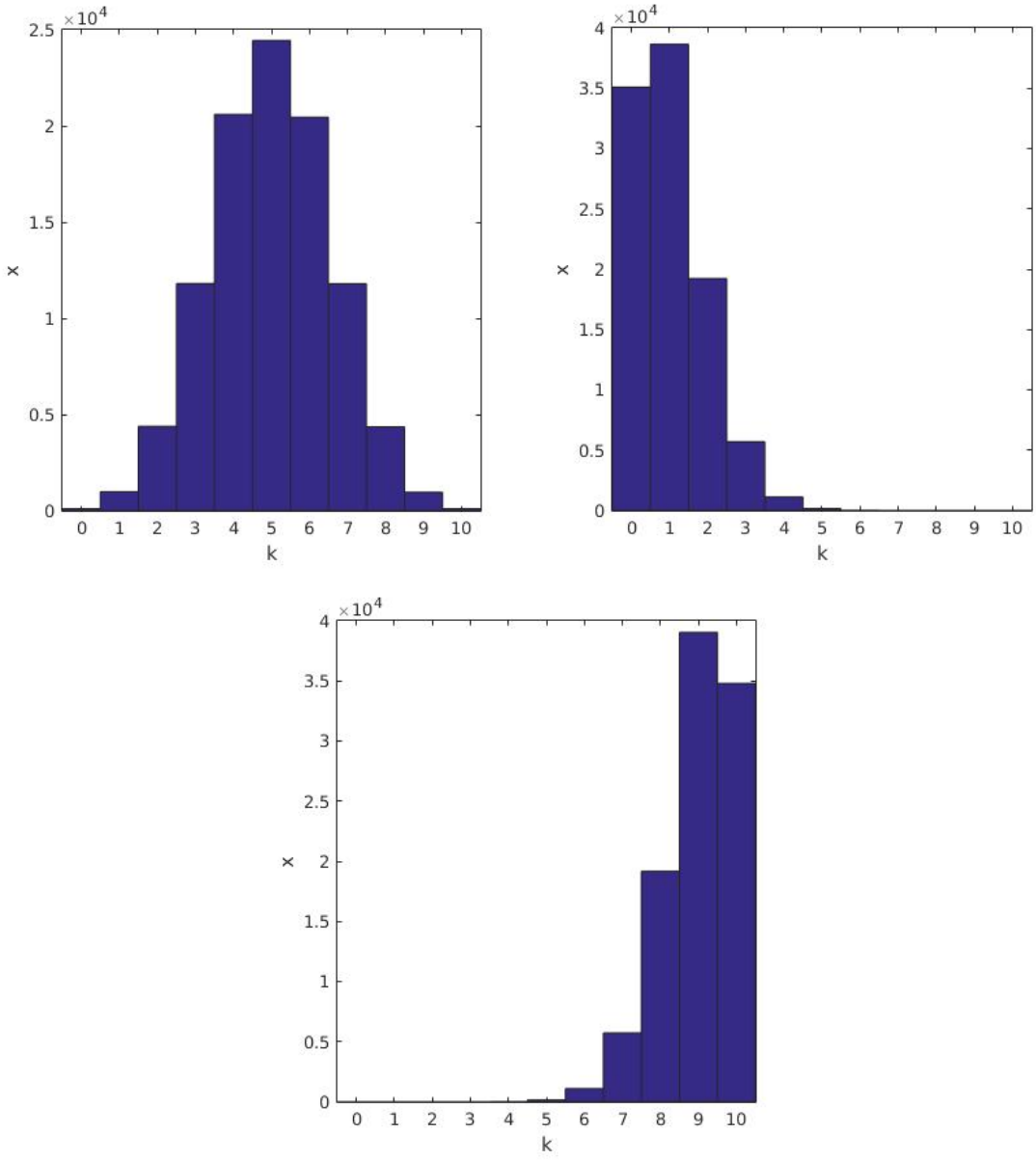


Abbildung 2.8: Binomial-Verteilung mit $p_1 = 0.5$, $p_2 = 0.1$, $p_3 = 0.9$

- Schiefe muß durch Mitteln klein werden, geht am Rand langsamer.

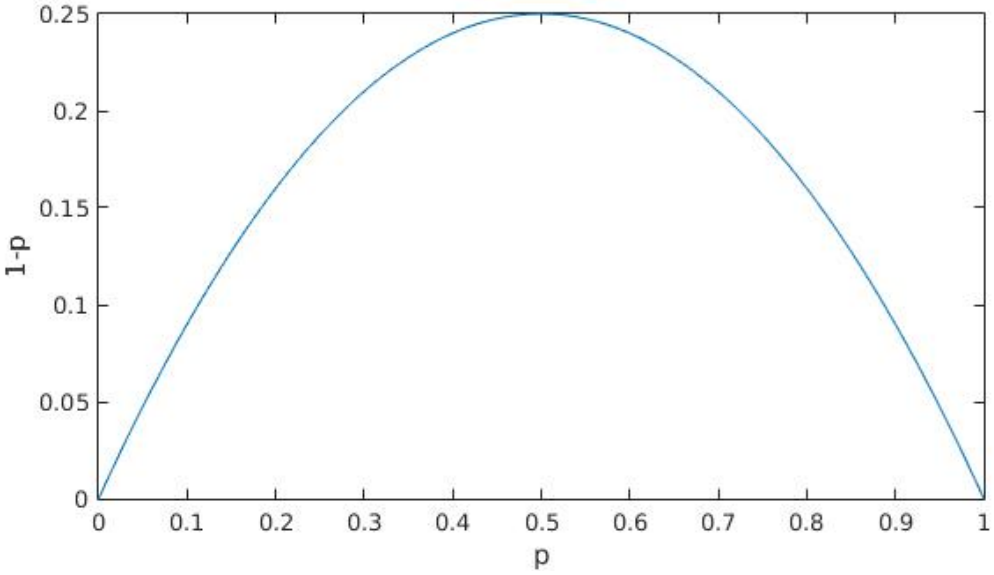


Abbildung 2.9: p(1-p): Verantwortlich für Varianz des Schätzers, "1-p" muss "Varianz" heißen

- Für $n\,p(1-p) < 10$: schlage Pearson-Clopper-Werte nach.

Lessons learned:

- Zufallsvariablen haben eine Verteilung, Realisierungen sind eine Zahl.
- Gauß-Verteilung ist wichtig wegen Zentralem Grenzwertsatz.
- Hochdimensionale Räume sind im wesentlichen leer.
- Schätzer sind Zufallsvariablen.
- Konsistente Schätzer sind toll.

3 Hypothesen Tests

... oder **Die fünf Dilemmata des Testens**

3.1 Parametrische Tests

Häufig laufen Fragestellungen auf statistische Tests hinaus.
Alles weitere am Beispiel des t-Tests.

Das Procedere

- Formuliere Nullhypothese H_0:

  Hier:

  Die Mittelwerte μ_1, μ_2 zweier Gauß-Verteilungen mit gleicher Varianz σ^2 sind gleich.

  Beachte: Hier stecken drei Annahmen drin

  Test ist parametrisch, weil eine parametrisierte Verteilungen, hier Gauß-Verteilungen, angenommen werden.

- Berechne (analytisch/simulativ) Verteilung einer Testgröße unter der Nullhypothese.

  Hier analytisch:

  - Schätze Mittelwerte $\hat{\mu}_1$, $\hat{\mu}_2$ für jeweils N Messwerte x_i^1 und x_i^2:

    $$\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \quad k = 1, 2$$

    Entsprechend Varianzen $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$:

    $$\hat{\sigma}_k^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i^k - \hat{\mu}_k)^2, \quad k = 1, 2$$

  - Berechne das Mittel:

    $$\hat{S}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}$$

    und den Standardfehler des Mittelwertes:

$$\hat{S}_M = \hat{S}\sqrt{\frac{2}{N}}$$

– Dann ist unter Gültigkeit von H_0

$$t := (\hat{\mu_1} - \hat{\mu_2})/\hat{S}_M$$

t-verteilt mit $r = 2N - 2$ Freiheitsgraden.
"-2" weil zwei Mittelwerte aus den Daten geschätzt

– Erinnere Definition t-Verteilung:

$$t(r, x) = \frac{N(0, 1)}{\sqrt{\chi_r^2/r}}$$

– Anschauliche Normierung in der Asymptotik:

$$\lim_{r\to\infty} t(r, x) = N(0, 1)$$

$$\tilde{t} \sim N\left(0, \frac{1}{2N-2}\,\sigma^2\right), \quad \mu = 0 \text{ wegen } \mu_1 = \mu_2$$

– Beachte:

* In der Regel will man H_0 ablehnen: Medikament ist besser als Placebo. Hier $\mu_1 \neq \mu_2$
* Unter der Alternative H_1 hat Test-Statistik (hoffentlich) eine andere Verteilung als unter H_0
* Hier: Anschaulich normierte asymptotische Verteilung von t unter Alternative $\mu_1 \neq \mu_2$:

$$\tilde{t} \sim N\left(\mu_1 - \mu_2, \frac{1}{2N-2}\,\sigma^2\right)$$

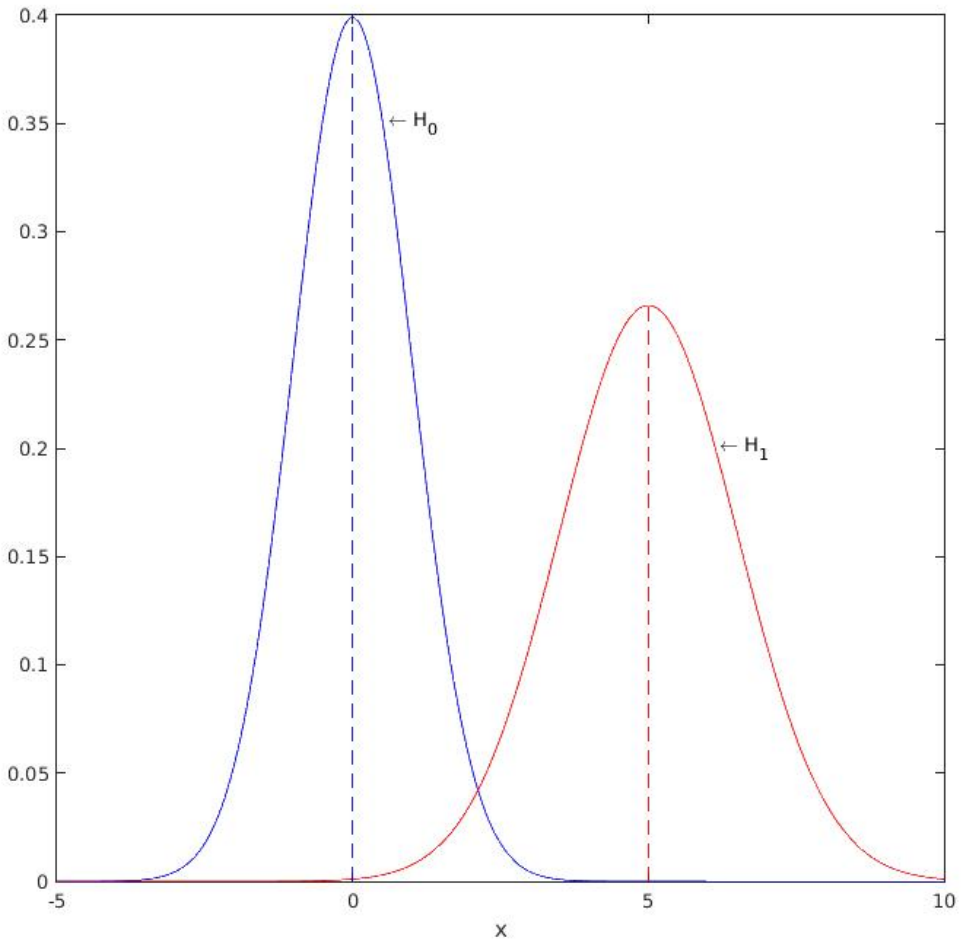


Abbildung 3.1: Null-Hypothese H_0 und Alternative H_1

- **Dilemma I** des Testens: Jeder kann dazugehören
  - Führt man den Test durch, ergeben sich konkrete Zahlen für $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{S}_M$ und damit für t
  - Test läuft auf Frage hinaus:
    Gehört ein Wert - der realisierte t-Wert - zu der - hier - t-Verteilung ?
  - Problem:

Das ist nicht zu verneinen !

Im Prinzip kann unter H_0 jeder Wert der Test-Statistik - hier t - vorkommen.

– p-Wert: Wahrscheinlichkeit für Wert größer t:

$$p = 1 - \int_{-\infty}^{t} p(x)\,dx$$

Per Konstruktion: Unter H_0: p-Wert der Teststatistik ist gleichverteilt auf [0,1].

– Ausweg: Verwerfe die Nullhypothese $H_0 : \mu_1 = \mu_2$ für extreme Ereignisse: p-Wert sehr klein, p-Wert sehr groß

– Dazu: Wähle Signifikanz-Niveaus α.
Verwerfe H_0, wenn p-Wert $< \alpha$

– Typische Werte für α: 0.05 oder 0.01
Angewandte Statistik ist kein Fall von Mathematik! 3.24

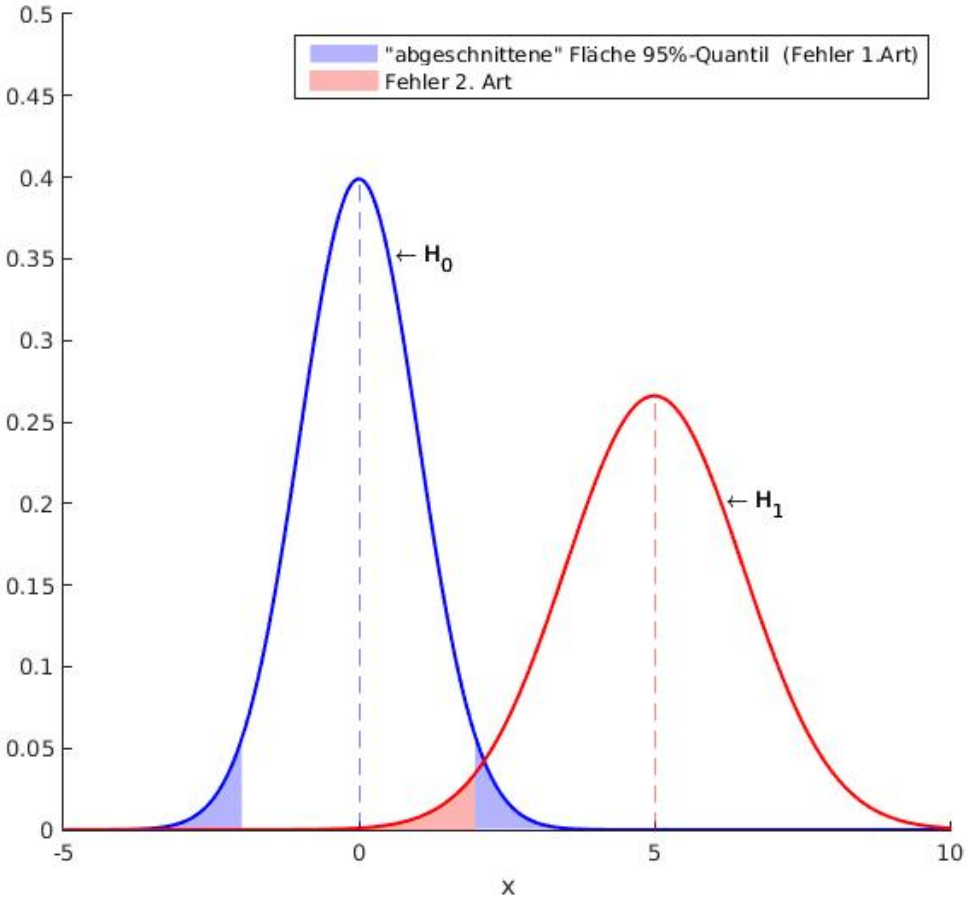


Abbildung 3.2: Nullhypothese H_0 und Alternativhypothese H_1 mit Signifikanzniveau α

  - Dabei können 2 Arten von Fehlern auftauchen:
    Fehler 1. Art: H_0 wird abgelehnt, obwohl sie wahr ist: falsch positiv
    Fehler 2. Art: H_0 wird nicht abgelehnt, obwohl sie falsch ist: falsch negativ
  - Fehler 2. Art kosten ein gutes Paper
  - Fehler 1. Art kosten die Karriere 2.18

- Power of the test: Häufigkeit der Ablehnung eines Tests, wenn H_0 falsch.

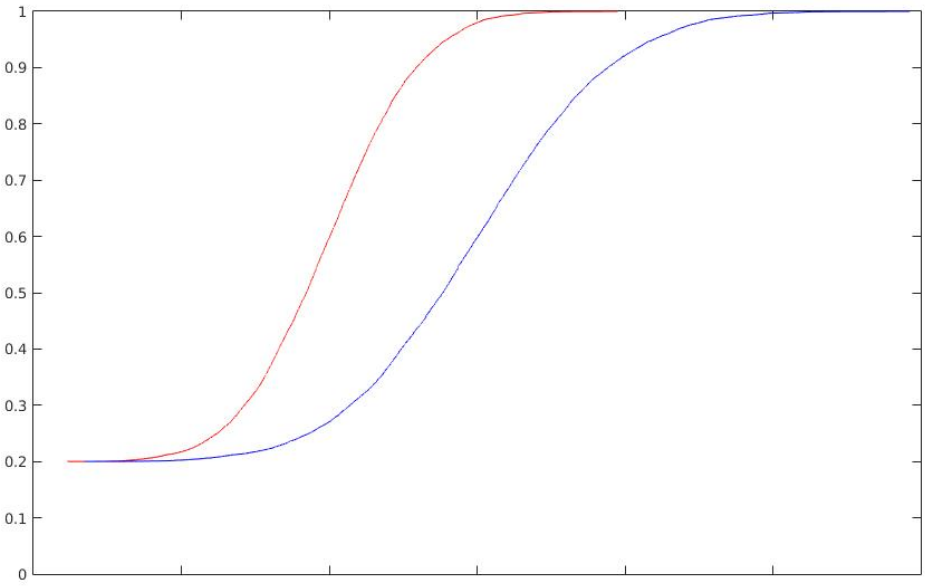


Abbildung 3.3: Power of the Test, α falsch, x-Achse: Verletzung von H_0, y-Achse: Power

- Tatsächliche Häufigkeit Fehler 1. Art $< \alpha$: Test ist konservativ.
- Tatsächliche Häufigkeit Fehler 1. Art $> \alpha$: Test ist Schrott.

Spielarten des t-Testes:

- Ist ein Mittelwert größer/kleiner als der andere ?

  Einseitiger und zweiseitiger Test, Power diskutieren
- Ein-Stichproben-t-Test: Ist eine Verteilung verträglich mit einem bestimmten Mittelwert ?
- Varianzen der beiden Gauß-Verteilugen unterschiedlich, Welch Test
- Stichprobenanzahl ist unterschiedlich
- Gepaarte Tests, später
- Keine Gaußannahme, später

2. Woche

Übung :
Die Power des t-Tests

Dilemma II: Dichotomie von Kakutani [28]

- Wenn H_0 nicht wahr ist, gilt eine Alternative H_1 mit Verteilung $p_{H_1}(x)$
- Nun gilt:

$$\lim_{N\to\infty} \int p_{H_0}(X)\, p_{H_1}(X)\, dX = \begin{cases} 0 \\ \text{oder} \\ 1 \end{cases}$$

  Für $N \to \infty$ werden Verteilungen $p_{H_0}(X)$ und $p_{H_1}(X)$ immer schmaler.

  - Falls $p_{H_0}(X) \neq p_{H_1}(X)$ gibt es irgendwann keinen gemeinsamen Träger mehr.
  - Falls $p_{H_0}(X) = p_{H_1}(X)$ passiert eh nix

- Wenn ein Test überhaupt Power hat, wird H_0 mit zunehmender Datenzahl immer abgelehnt

  "All null hypotheses are wrong" (Fischer, 1925) "... but some are useful!"

- Ionenkanal Beispiel

Dilemma III: Statistische Signifikanz vs. Inhaltliche Relevanz

- Patienten mit Puls 180 $\pm$ 10 Schläge/min
- Ein Medikament verringert Puls auf 170 $\pm$ 10 Schläge/min
- Führe t-test mit N Patienten durch:

$$\begin{array}{rcl} N{=}5 & : & \text{n.s.} \\ N{=}10 & : & p = 0.03 \\ N{=}100 & : & p < 10^{-7} \\ N{=}1000 & : & p < 10^{-20} \end{array}$$

  - Beliebig kleine Verletzungen der Nullhypothese führen zu signifikanten Unterschieden falls hinreichend viele Messungen N verfügbar sind.
  - Vor Durchführung des Testes überlegen, welches Ausmaß an Verletzung der Nullhypothese inhaltlich relevant ist.
  - Daraus lässt sich bestimmen, wieviele Messungen N notwendig sind, um eine sinnvolle Verletzung statistisch signifikant nachzuweisen.
  - Fallzahl Kalkulation

- Hier:
  - Was ist klinisch relevante Herabsetzung des Pulses ?
  - Wie viele Messungen/Patienten N braucht man, um Nullhypothese H_0: "Medikament hat keinen Effekt." abzulehnen ?
  - Wenn H_0 dann abgelehnt: Medikament ist sinnvoll.
  - Wenn H_0 nicht abgelehnt basierend auf N Messungen/Patienten: Gemessen in inhaltlich relevanter Skala hat das Medikament keinen Effekt.

4.24

Dilemma IV: Multiples testen

- Setting: An m Größen soll ge-t-testet werden, ob sich zwei Spezies unterscheiden.
- H_0 : Es gibt keine Unterschiede
- Procedere: Mache m t-Tests, jeweils zu Signifikanzniveau α.
- Wahrscheinlichkeit $\tilde{\alpha}$, H_0 abzulehnen:

$$\tilde{\alpha} = 1 - (1 - \alpha)^m \tag{1}$$

- Beipiel: $\alpha = 0.01$

$$\begin{aligned} m = 10 &\implies \tilde{\alpha} = 0.1 \\ m = 100 &\implies \tilde{\alpha} = 0.63 \\ m = 1000 &\implies \tilde{\alpha} = 0.99996 \end{aligned}$$

Lösung 1:

- Bonferroni - Korrektur:
- Löse Gl. (1) nach α auf:

$$\alpha = 1 - (1 - \tilde{\alpha})^{1/m} \approx \frac{\tilde{\alpha}}{m}$$

- Berechne für gewünschtes (globales) $\tilde{\alpha}$ das notwendige α für die Einzel-Tests.
- Problem :
  - α wird sehr klein,
  - Test sehr konservativ, keine Power
    $\Longrightarrow$ viele Fehler 2. Art
- Variation: Bonferroni-Holm: Korrigiere in jedem Schritt j mit j/m.

Lösung 2:

- Ein Experiment (m) zur Hypothesen Generierung,
- ergibt $m' \ll m$ Kandidaten.
  ein paar echt positive, ein paar falsch positive
- ein zweites Experiment zum Testen mit m'.

Variation über dieses Thema:

- AIDS Test
- Erst (billiger) sensitiver Test, der nicht hochspezifisch ist
- Falls positiv, dann mehrfach (teurer) Test, der hochspezifisch aber nicht so sensitiv ist.

Lösung 3:

- Benutze Binomial-Verteilung $B(m, \alpha, k)$ um Anzahl der erwarteten falsch-positiven abzuschätzen: False Discovery Rate

$$\langle \#(\text{Falsch positive})|_{H_0} \rangle = \sum_{k=1}^{m} kB(m, \alpha, k)$$

- Wenn's viel mehr positive sind, ist da ein Unterschied.

Spezialfall: ANOVA

- Betrachte: Experiment untersucht mehrere Bedinungen in der selben Hinsicht.
- Beispiel Placebo, Med_1, ..., Med_M in Bezug auf # Rote Blutkörperchen
- ANalysis Of VAriance (ANOVA) ist die Alternative zu $\frac{M(M-1)}{2}$ t-Tests.

Herleitung:

- H_0 : Kein Effekt.
- M Bedingungen, jeweils N Beobachtungen: x_{ij}
- Mittelwert pro Bedingung

$$\bar{x}_{i.} = \frac{1}{N} \sum_{j=1}^{N} x_{ij}$$

Mittelwert über alles:

$$\bar{x}_{..} = \frac{1}{M} \sum_{i=1}^{M} \bar{x}_{i.}$$

Varianz aller Daten, genannt SS_{total}, SS für sum of squares, ist:

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{M} \sum_{j=1}^{N} (x_{ij} - \bar{x}_{..})^2 \\ &= \sum_{i=1}^{M} N(\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{M} \sum_{j=1}^{N} (x_{ij} - \bar{x}_{i.})^2 \end{aligned}$$

- Erster Summand: Varianz der Gruppenmittel bezüglich des Gesamtmittels $SS_{between}$ mit $M-1$ Freiheitsgraden.

  Zweiter Summand: Varianz in den einzelnen Gruppen SS_{within} und hat $(N-1)M$ Freiheitsgrade.

- Unter Gültigkeit der Nullhypothese folgt ihr Quotient somit einer $F(M-1,(N-1)M)$ - Verteilung.

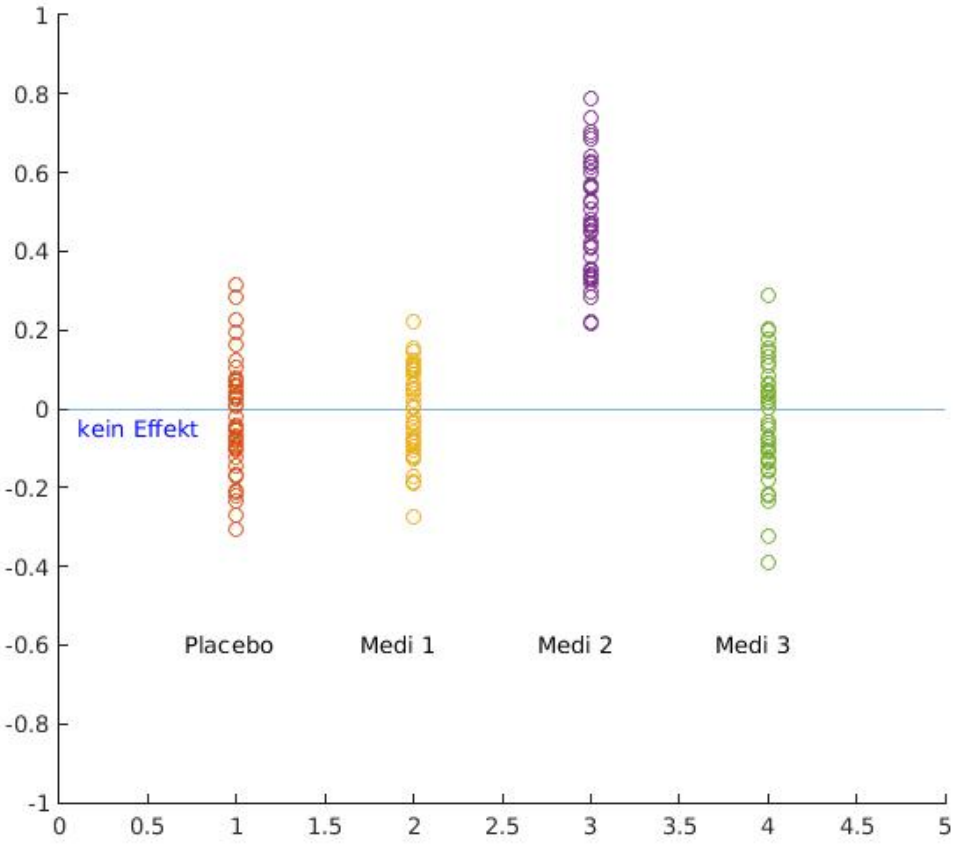


Abbildung 3.4: Verschiedene Medikamente im ANOVA-Test. Medikament 2 wirkt, Medikament 1 und 3 zeigen keine Wirkung

Im Falle der Ablehnung:

- Ist ANOVA signifikant, stellt sich Frage: Wer war's ?
- A posteriori Test (Tukey-Kramer oder Scheffé) gibt kritsche Differenz, die die Mittelwerte überschreiten müssen, um als signifikant verschieden angesehen werden zu können.
- Berücksichtig, daß die Daten für die ANOVA schon einmal statistisch "benutzt" worden sind.
- *A posteriori* Test hat immer kleinere Power als ein einzelner t - Test z.B. zwischen den grössten Mittelwertsdifferenzen. Es ist daher wichtig[1], sich *vor* einem Experiment zu überlegen, wie die minimal zu testende Hypothese lautet. Man läuft sonst Gefahr, vorliegende Effekte statistisch nicht nachweisen zu können.

Verallgemeinerungen auf

- unterschiedliche Varianzen
- unterschiedliche Stichprobenumfänge
- mehrerer Einflußgrößen, sogenannten Faktoren, Medikamente und Geschlecht

ANOVA in Kurzform: Vergleich der Varianz von Gruppenmittel bezüglich Gesamtmittel mit Varianz in einzelnen Gruppen (F-Test). So kann man sich $\frac{M(M-1)}{2}$ t-Tests sparen.

Gepaarte Tests

- Bisherige Annahme: Verteilungen sind unabhängig
- Werden Daten von den selben Individuen erhoben, so hat dieses eine Auswirkung auf die Varianz:

[1] aber leider nicht üblich

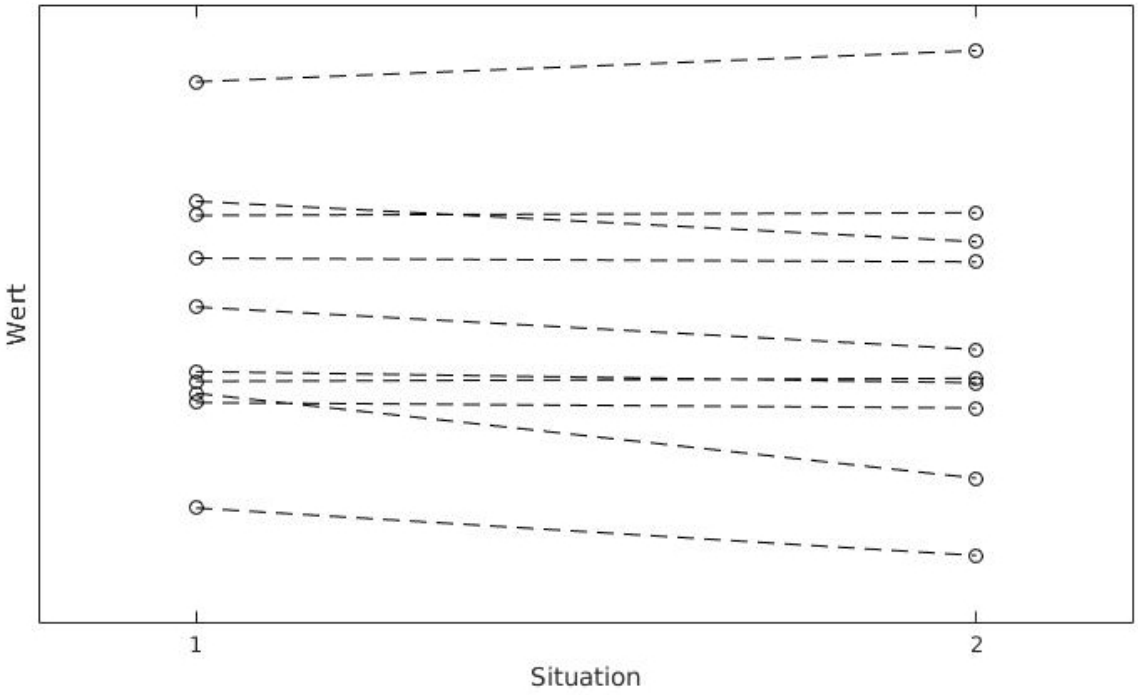


Abbildung 3.5: Gepaarte Werte

Dieses muß berücksichtigt werden.

- Man spricht von gepaarten Tests.
- Wie im Falle des gepaarten t - Tests läßt sich auch für die ANOVA eine Messwiederholung, i.e. die Abhängigkeit der Stichproben zu den verschiedenen Versuchsbedingungen, berücksichtigen. Dies geschieht z.B. durch die Greenhouse - Geisser - Korrektur.

3.18

3.2 Nichtparametrische Tests

Bisher:

- t-Test ging von Gaußverteilten Stichproben aus: Parametrischer Test
- Fällt Verteilung sehr viel langsamer als Gaußverteilung ab, z.B. Cauchy-Verteilung, verliert t-Test seine Power, siehe Übung

"Ausreißer" diskutieren, shit happpens

- Alternative: Nichtparametrische Tests
- Diese sind robust gegen Verletzungen der Verteilungsannahmen
- Statt Mittelwerts- dann Lokationsvergleich.

$$p_1(x) = p_2(x + \Delta) \quad H_0 : \Delta = 0$$

Hier:

- t-Test geht nach Wilcoxon Rangsummentest (oder U-Test, Mann-Whithney-Test)
- Nullhypothese: Die Verteilungen seien identisch.

  Anders formuliert: Die Ränge beider Stichproben sind gleichverteilt bezüglich der Gesamtheit.

  Sei $N_1 = N_2 = N$. Ränge R_i^k, $k = 1, 2$, $i = 1, \ldots N$: Sei

$$x_i^1 = (-6.7, -1, 5) \quad x_i^2 = (-5, -2.2, 7)$$

  dann

$$R_i^1 = (1, 4, 5) \quad R_i^2 = (2, 3, 6)$$

- Berechne die Ränge $R_1^1, \ldots, R_N^1$ der ersten Stichprobe bezüglich der Gesamtstichprobe.
- Unter H_0 gilt:

$$\left\langle \sum_{i=1}^{N} R_i^1 \right\rangle = N^2 + 0.5N, \quad Var\left(\sum_{i=1}^{N} R_i^1 \right) = \frac{1}{6}N^3 + \frac{1}{12}N^2$$

- Mit Zentralem Grenzwertsatz folgt:

$$W = \left(\sum_{i=1}^{N} R_i^1 - (N^2 + 0.5\,N) \right) / \sqrt{\frac{1}{6}N^3 + \frac{1}{12}N^2} \sim N(0, 1)$$

  Gute Näherung für $N \geq 20$

- Für $N < 20$, exakte Werte aus Kombinatorik, aber aufwendig zu berechnen, sind tabelliert
- Ursache der Robustheit: Niedrigster Wert hat Rang = 1, höchster Wert hat Rang = $2N$, egal ob Gauß- oder Cauchy-Verteilung zu Grunde liegt, oder "Ausreißer"
- Gepaarter Fall: Wilcoxon-Vorzeichen-Rang-Test

  Nichtparametrische ANOVA: Kruskal-Wallis Test oder auch H-Test.

Effizienz

- Sind Daten Gaußverteilt, erkennt parametrischer t - Test eine Mittelwertsdifferenz mit weniger Daten als Wilcoxon - Test, oder bei gleichem N bei kleinerer Differenz, t-Test hat höhere Power
- Geringere Power der nichtparametrischen im Vergleich zu den parametrischen Tests bei Gültigkeit der parametrischen Annahmen wird durch die sogenannte Effizienz angegeben:

  $$Eff = \frac{Power(NP - test)}{Power(P - test)}$$

  gegeben die Gültigkeit der parametrischen Verteilung.
- Wilcoxon-Test hat gegenüber t - Test eine Effizienz von 0.95, d.h. der t - Test hat für Gaußverteilte Daten mit 95% der Datenanzahl die selbe Power wie der Wilcoxon-Test.
- Da falsche Verteilungsannahmen zu einem Verlust der Power parametrischer Tests führen, nichtparametrische Tests aber eine Effizienz < 1 haben, folgt

  Dilemma V : Power vs. Effizienz
- Da Effizienz nahe 1 und nichtparametrische Tests robust gegen Verletzung von Verteilungsannahmen sind, werden sie heute in der Regel bevorzugt.

Lessons learned:

- Zentral: Herleiten der Verteilung einer Teststatisitk unter H_0
- Die fünf Dilemmata des Testens
  - Jeder kann dabei sein, Notwendigkeit eines Signifikanzniveaus
  - Mit zunehmender Datenanzahl i.d.R. wird jede Nullhypotese abgelehnt
  - Statistische Signifikanz vs. inhaltlicher Relevanz
  - Multiples Testen
  - Power vs. Effizienz bei parametrischen vs. nichtparametrischen Tests
- Signifikanzniveaus sind kein Fall von Mathematik, sondern Risikoabwägung

3. Woche 5.24

4 Parameterschätzung

Literatur:

- D.R. Cox and D.V. Hinkley: Theoretical Statistics [12]
- E.L. Lehmann: Theory of Point Estimation [39]

Motivation: Einfachstes Modell: Lineare Regression, siehe Kap. 10.1

$$y_i = ax_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Zwei Fragen, Antworten gehen auf Gauß zurück :

(i) Gegeben N Datenpaare (x_i, y_i), wie bestimmt man $\hat{a}$?
Stichwort: Punktschätzer

(ii) Wie genau ist $\hat{a}$ aus den Daten bestimmt ?
Stichwort: Konfidenzintervall

ad (i)

- Intuition: $\hat{a}$ so wählen, dass $y = \hat{a}x_i$ möglichst nah an den Daten liegt
- D.h. Abstand minimieren, z.B.

$$\hat{a} = argmin \sum_{i=1}^{N} (y_i - ax_i)^2$$

Kleinste Quadrate Schätzer oder Least squares estimator

ad (ii)

- Intuition: Wenn σ^2 groß, ist a schlechter bestimmt

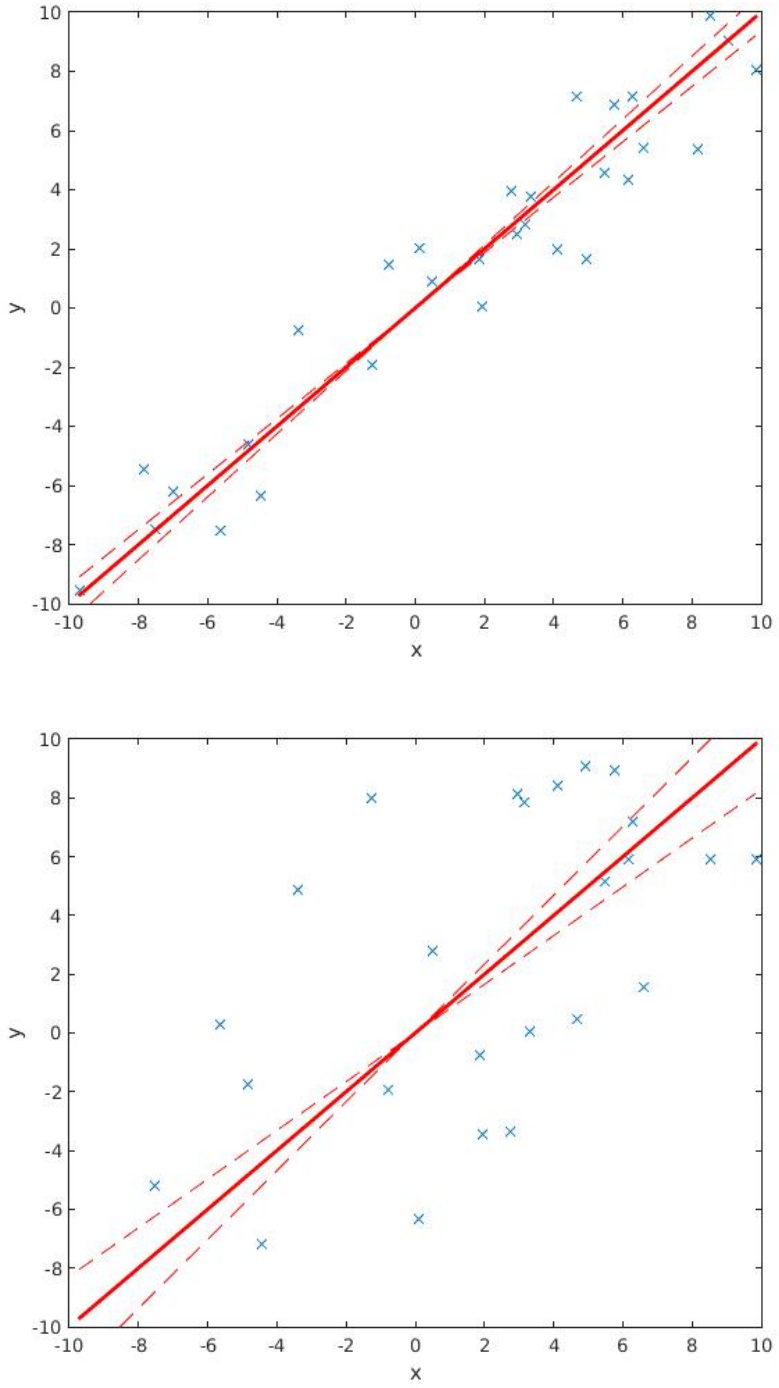


Abbildung 4.1: Lineare Regression an 2 Datensätzen mit (a) $\sigma^2 = 4$ und (b) $\sigma^2 = 36$

- Wenn σ_i^2 statt σ^2, gewichten

$$\hat{a} = argmin \sum_{i=1}^{N} \frac{(y_i - ax_i)^2}{\sigma_i^2}$$

Gewichteter Kleinste Quadrate Schätzer oder Weighted least squares estimator

$\sum_{i=1}^{N} \frac{(y_i - ax_i)^2}{\sigma_i^2}$ wird auch als χ^2 bezeichnet. Für wahres a ist es auch so verteilt.

Beispiele für Modelle

- Regressionsmodelle
  - Linear in den Parametern, nichtlinear in x, der unabhängigen Variablen

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^4 + \ldots + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$$p(y_i|a, x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \sum_{j=0}^{n} a_j x_i^j)^2}{2\sigma^2}}$$

    Siehe Kap. 10.2
  - Nichtlinear in den Parametern

$$y = \sin ax + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
$$p(y_i|a, x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \sin ax_i)^2}{2\sigma^2}}$$

    Siehe Kap. 10.3

- Dynamische Modelle
  - Partiell beobachtete gewöhnliche Differentialgleichungen

$$\begin{aligned} \dot{x} &= f(x, p), \quad x(0) = x_0 \quad dim(x) = n \\ y(t_i) &= g(x(t_i, p)) + \epsilon(t_i), \qquad dim(y) = m \\ m &< n \end{aligned}$$

$$p(y(t_i)|p, x_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y(t_i) - g(x(t_i, p, x_O))^2}{2\sigma^2}}$$

– Stochastische, partielle Differentialgleichungen

– Zustandsraummodell, zeit-diskret

$$\begin{aligned} x(t) &= Ax(t-1) + \epsilon(t), \quad \epsilon(t) \sim N(0, \sigma_\epsilon^2) \\ y(t) &= Cx(t) + \mu(t), \qquad \mu(t) \sim N(0, \sigma_\mu^2) \end{aligned}$$

Stichwort: Kalman-Filter

– Hidden Markov Modell, zeit-diskret
Diskrete Zustände $x_1, \ldots x_s$

$$p(x(t+1)|x(t), x(t-1), \ldots) = p(x(t+1)|x(t))$$

Übergangswahrscheinlichkeiten

$$a_{ij} = p(x(t+1) = j|x(t) = i)$$

Beobachtungen verrauscht

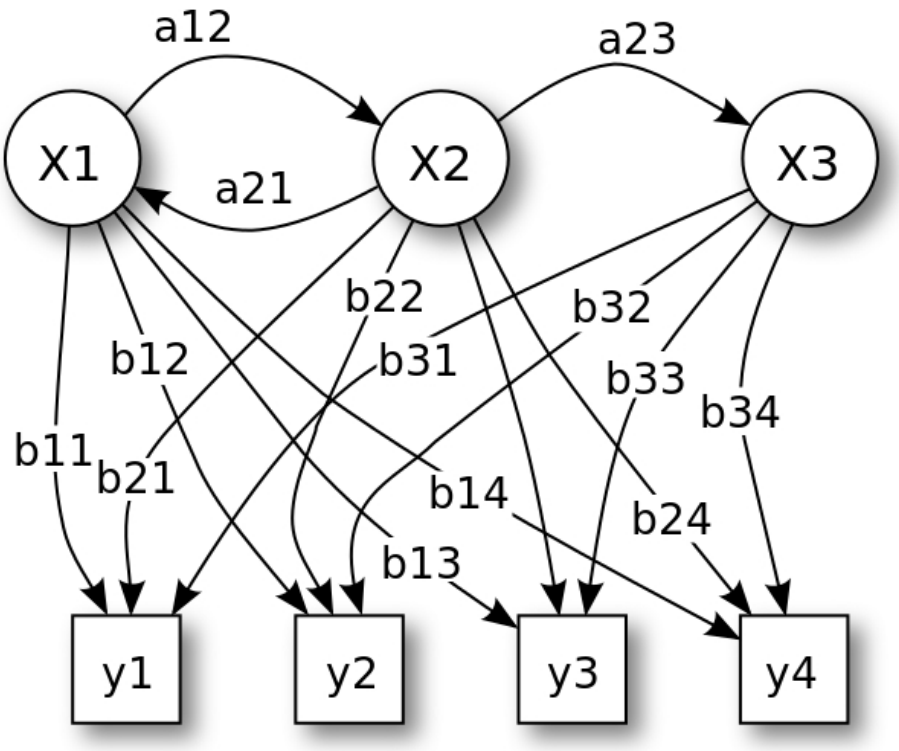


Abbildung 4.2: Hidden Markov Modell

Stichwort: Baum-Welsh Algorithmus, Viterbi Algorithmus

- Teilchenphysik

  Modell für Hintergrund bei Higgs-Boson Suche: Hochkomplexe Rechnungen und Simulationen

Gemeinsamkeit aller Modelle: Sie produzieren eine Wahrscheinlichkeit $p(z, a)$ für Beobachtungen z in Abhängigkeit der Parameter a

4.1 Maximum Likelihood Schätzer

Erinnere:

- Bias (Verzerrung) : $\langle\hat{\Theta}\rangle - \Theta_0$
- Varianz des Schätzers : $\langle(\hat{\Theta} - \langle\hat{\Theta}\rangle)^2\rangle$, bestimmt Vertrauensintervall
- Mittlerer Quadratischer Fehler : $\langle(\hat{\Theta} - \Theta_0)^2\rangle$ = bias^2 + Varianz des Schätzers

Sei X eine parametrisierte Zufallsvariable mit Dichte $p(x, a)$.

Normales Setting:

- a gegeben
- $p(x, a)$ gibt die Wahrscheinlichkkeit für x

Hier:
Gegeben N Realisierungen, dann heißt

$$L(x_1, \ldots, x_N|a) = \prod_{i=1}^{N} p(x_i, a)$$

die Likelihood

- $L(x_1, \ldots, x_N|a)$ ist in Abhängigkeit von a zu lesen
- Daten sind gegeben
- Likelihood: "Für ein angenommenes a, was ist die Wahrscheinlichkeit, dass die gegebenen Daten entstehen ?"
- Likelihood ist keine Wahrscheinlichkeit, da $\int p(x, a)\, da$ nicht normiert. Im Gegensatz zu $\int p(x, a)\, dx = 1$

Maximum Likelihood Schätzer (MLE):

- Wähle Parameter a so, dass Likelihood maximal wird
- Intiutiv sinnvoll
- Formal:

$$\frac{\partial L(x_1, \ldots, x_N|a)}{\partial a} = 0$$

- Logarithmiere:

$$\mathcal{L}(a) = \log L(a) = \sum_{i=1}^{N} \log p(x_i, a)$$

  Da Logarithmus monoton, ändert sich Wert für Maximum nicht

  Ersetzt schwierig handhabbares Produkt durch einfachere Summe

  Üblicher Weise noch ein Minuszeichen, ändert Wert für Maximum auch nicht.

  Minimierung der Log-Likelihood statt Maximierung der Likelihood

  Konvention: $-\log L$ wird wieder als Likelihood L bezeichnet

- M.k.z.: MLE-Schätzer unter milden Bedingungen asymptotisch unverzerrt. Widerspruchsbeweis (Cox/Hinkley p. 288f, hübsch)

- M.k.z.: MLE-Schätzer unter milden Bedingungen asymptotisch Gauß-verteilt

$$\sqrt{N}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, \Sigma)$$

  mit

$$\Sigma = -N \left(\frac{\partial^2 L(\hat{\theta})}{\partial \theta_i \partial \theta_j} \right)^{-1}$$

  Beweis folgt

  Intuition: Zeichung

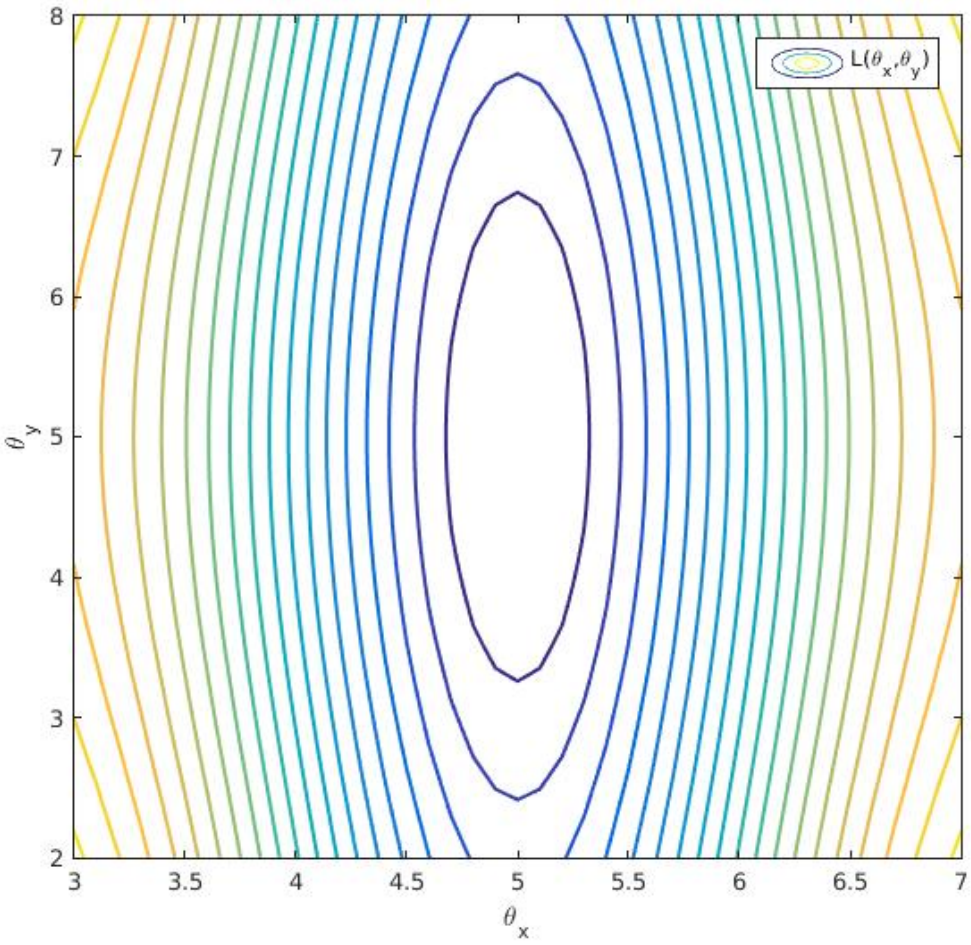


Abbildung 4.3: Maximum Likelihood Estimator im 2-Dimensionalen, korreliert malen

6.24

Cramér-Rao Schranke

- Im folgenden alle Indices unterdrückt
- Betrachte Score V:

$$V := \frac{\partial}{\partial a} L(x,a) = \frac{\partial}{\partial a} \log p(x,a) = \frac{1}{p(x,a)} \frac{\partial}{\partial a} p(x,a) \tag{2}$$

- Lemma 1: $< V >= 0$

$$\begin{aligned} < V > &= \int dx\, p(x,a) \frac{\partial}{\partial a} \log p(x,a) \\ &= \int dx\, p(x,a) \frac{1}{p(x,a)} \frac{\partial}{\partial a} p(x,a) \\ &= \int dx\, \frac{\partial}{\partial a} p(x,a) \\ &= \frac{\partial}{\partial a} \int dx\, p(x,a) \\ &= 0 \end{aligned}$$

- Lemma 2: $Var(V) = \left\langle -\frac{\partial^2}{\partial a^2} L(x,a) \right\rangle$

$$Var(V) = \left\langle \left(\frac{\partial}{\partial a} L(x,a) \right)^2 \right\rangle$$

Betrachte Ableitung nach a von

$$< V >= 0 = \int dx\, p(x,a) \frac{\partial}{\partial a} \log p(x,a)$$

$$0 = \int dx\, \frac{\partial}{\partial a} p(x,a) \frac{\partial}{\partial a} \log p(x,a) + \int dx\, p(x,a) \frac{\partial^2}{\partial a^2} \log p(x,a)$$

Mit Gl. (2) folgt für 1. Summanden:

$$\int dx\, p(x,a) \left(\frac{\partial}{\partial a} \log p(x,a) \right)^2 = Var(V)$$

und damit:

$$Var(V) = \left\langle -\frac{\partial^2}{\partial a^2} L(x,a) \right\rangle$$

$\left\langle -\frac{\partial^2}{\partial a^2} L(x,a) \right\rangle$ heißt Fischer Informations Matrix.

- Betrachte beliebigen, aber unverzerrten Schätzer $\hat{\theta}(x)$ für Parameter a, i.e. $< \hat{\theta}(x) >= a$.

  Lemma 3: $< V\hat{\theta}(x) >= 1$

$$\begin{aligned} < V\hat{\theta}(x) > &= \int dx\, p(x,a) \frac{1}{p(x,a)} \frac{\partial}{\partial a} p(x,a)\hat{\theta}(x) \\ &= \int dx\, \frac{\partial}{\partial a} p(x,a)\hat{\theta}(x) \\ &= \frac{\partial}{\partial a} \int dx\, p(x,a)\hat{\theta}(x) \\ &= \frac{\partial}{\partial a} < \hat{\theta}(x) > \\ &= \frac{\partial}{\partial a}\, a \\ &= 1 \end{aligned}$$

- Betrachte Cauchy-Schwarz Ungleichung:

$$\begin{aligned} < (V - < V >)(\hat{\theta} - < \hat{\theta} >) >^2 &\leq < (V - < V >)^2 >< (\hat{\theta} - < \hat{\theta} >)^2 > \\ < V\hat{\theta} - V < \hat{\theta} > - < V > \hat{\theta} + < V >< \hat{\theta} >>^2 &\leq Var(V)Var(\hat{\theta}) \\ < V\hat{\theta} >^2 &\leq Var(V)Var(\hat{\theta}) \end{aligned}$$

$$Var(\hat{\theta}) \geq \frac{1}{Var(V)} = \frac{1}{< \left(-\frac{\partial^2}{\partial a^2} L(x,a)\right) >}$$

Die Cramér-Rao Schranke

- Krümmung der Log-Likelihood bestimmt Varianz der Schätzer.

  Varianz des Schätzers ergibt Vertrauensintervalle.

  ZEICHNUNG

- M.k.z.: Maximum Likelihood Schätzer nimmt untere Grenze an

$$Var(\hat{\theta}_{MLE}) = \frac{1}{< \left\langle -\frac{\partial^2}{\partial a^2} L(x,a) \right\rangle >}$$

- Effizienz: Sei $\hat{\Theta}_.$ ein nicht-Maximum Likelihood Schätzer, dann gilt

$$Eff(\hat{\Theta}_.) = \frac{Var(\Theta_{MLE})}{Var(\Theta_.)} \leq 1$$

> MLE-Schätzer sind das beste Pferd im Stall, holen die meiste Information aus den Daten.

Konkrete Beispiele

- Gaußverteilung

$$p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

  Likelihood:

$$L(x_1,\ldots,x_N|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

  Log-Likelihood

$$L(\mu,\sigma) = -N\log\sigma - N\log\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

  – Schätzer für Mittelwert

$$\frac{\partial L(\mu,\sigma)}{\partial\mu} = \frac{1}{\sigma^2}\sum_i (x_i-\mu) \stackrel{!}{=} 0$$

  Damit:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i}^{N} x_i \quad \text{das beruhigt :-)}$$

  Varianz des Schätzers:

$$\frac{\partial^2 L(\mu,\sigma)}{\partial\mu^2} = \frac{1}{\sigma^2}\sum_i -1 = -\frac{N}{\sigma^2}$$

σ^2 bestimmt Krümmung der Likelihood: Je grösser σ^2, desto kleiner die Krümmung und damit um so die größer Varianz des Schätzers

$$\begin{aligned} Var(\hat{\mu}) &= -\frac{1}{\frac{\partial^2 L(\mu,\sigma)}{\partial \mu^2}} = \frac{\sigma^2}{N} \\ SEM(\hat{\mu}) &= \frac{1}{\sqrt{N}}\,\sigma \end{aligned}$$

Typische $\frac{1}{\sqrt{N}}$-Abhängigkeit

– Schätzer für Varianz

$$\begin{aligned} \frac{\partial L(\mu,\sigma)}{\partial \sigma} &= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \stackrel{!}{=} 0 \\ N\hat{\sigma}^2 &= \sum_i (x_i-\hat{\mu})^2 \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_i (x_i-\hat{\mu})^2 \end{aligned}$$

Erinnere Kap. 2.4: Unverzerrter Schätzer hatte $\frac{1}{N-1}$
MLE im allgemeinen nur asymptotisch unverzerrt.
Berechnung $Var(\sigma^2)$: Hausaufgabe

- Lineare Regression:

$$y_i = ax_i + \epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2)$$

$$p(y_i|a,x_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-ax_i)^2}{2\sigma^2}\right)$$

Log-Likelihood

$$L(a) \propto \sum_{i=1}^{N}(y_i-ax_i)^2$$

Rückwärts gelesen: Schätzt man auf Grund (gewichteter) kleinster Quadrate, hat man Gauß-Verteilung angenommen

$$\frac{\partial L(a)}{\partial a} = \sum (y_i - a x_i) x_i \stackrel{!}{=} 0$$

$$\sum (y_i x_i - a x_i^2) = 0$$

$$\hat{a}_{MLE} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Weitergehende Behandlung in Kap. 10.1 und 10.2

- Exponential-Verteilung

$$\begin{aligned}
p(x,\tau) &= \frac{1}{\tau} e^{-x/\tau}, \quad \lambda = \frac{1}{\tau}, \quad p(x,\lambda) = \lambda e^{-\lambda x} \\
L(\lambda) &= \prod_{i=1}^{N} \lambda e^{-\lambda x_i} \\
L(\lambda) &= \sum_{i=1}^{N} \log\left(\lambda e^{-\lambda x_i}\right) = \sum_{i=1}^{N} (\log\lambda - \lambda x_i) = N \log\lambda - \lambda \sum_{i=1}^{N} x_i \\
\frac{dL(\lambda)}{d\lambda} &= \frac{N}{\lambda} - \sum_{i=1}^{N} x_i \stackrel{!}{=} 0 \\
\hat{\lambda} &= \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}}, \quad \hat{\tau} = \bar{x} \\
\frac{d^2 L(\lambda)}{d\lambda^2} &= -\frac{N}{\lambda^2} \\
Var(\hat{\lambda}) &= \frac{\lambda^2}{N}
\end{aligned}$$

4.18

4.2 Methods of Moments

- Likelihood manchmal schwer, nur numerisch oder gar nicht zu berechnen
- Dann Methods of Moments eine Alternative

- Idee:

  Berechne Momente μ_k ...

  - ... aus den Daten: μ_k^{emp}
  - ... und aus Modell, parametrisierte theoretische Momente $\mu_k^{theo}(\theta)$.

- Definiere Schätzer über:

$$\mu_k^{emp} = \mu_k^{theo}(\hat{\theta}_{MM}), \quad k = 1, \ldots, m$$

  respektive

$$\hat{\theta}_{MM} = argmin \sum_{k=1}^{m} (\mu_k^{emp} - \mu_k^{theo}(\theta_{MM}))^2$$

- I.d.R. gilt

$$Var(\hat{\theta}_{MM}) \geq Var(\hat{\theta}_{MLE})$$

  Ist das Problem linear in den Parametern, die Fehler Gaußsch und betrachtet man 1. und 2. Momente, gilt:

$$\hat{\theta}_{MM} = \hat{\theta}_{MLE}, \quad Var(\hat{\theta}_{MM}) = Var(\hat{\theta}_{MLE})$$

7.24
4. Woche

4.3 Bayes'sche Ansätze

Bisher frequentistisch: Es gibt "wahre" Parameter

Bayes'sche Welt:

- Parameter sind auch Zufallsvariablen
- Alle Wahrscheinlichkeiten sind bedingte Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit

$$p(A|B) = \frac{p(A,B)}{p(B)} \quad \text{Wahrscheinlichkeit für } A \text{ gegeben } B$$

Betrachte Würfel: $A = \{1, 2\}$, $B = \{1, 2, 3\}$

- $p(A, B) = p(A \cap B)$
- $p(A) = 1/3$, $p(B) = 1/2$, $p(A, B) = 1/3$
- $p(A|B) = 2/3$

Satz von Bayes

$$\begin{aligned} p(A|B) &= \frac{p(A,B)}{p(B)} \\ p(B|A) &= \frac{p(A,B)}{p(A)} \\ p(A|B) &= \frac{p(B|A)\,p(A)}{p(B)} \end{aligned} \tag{9.24}$$

Mit $A = \theta$ und $B = Daten$ und $p(Daten) = const$ folgt

$$p(\theta|Daten) \propto p(Daten|\theta)\,p(\theta) \tag{3}$$

- Die Likelihood $p(Daten|\theta)$ wird durch den prior $p(\theta)$ verziert.
- Der prior $p(\theta)$ ist auch bedingte Wahrscheinlichkeit, bedingt auf das Vorwissen
- $p(\theta|Daten)$ heißt a posteriori Verteilung
- Liefert Maximum a posteriori (MAP)-Schätzer und seine Verteilung.

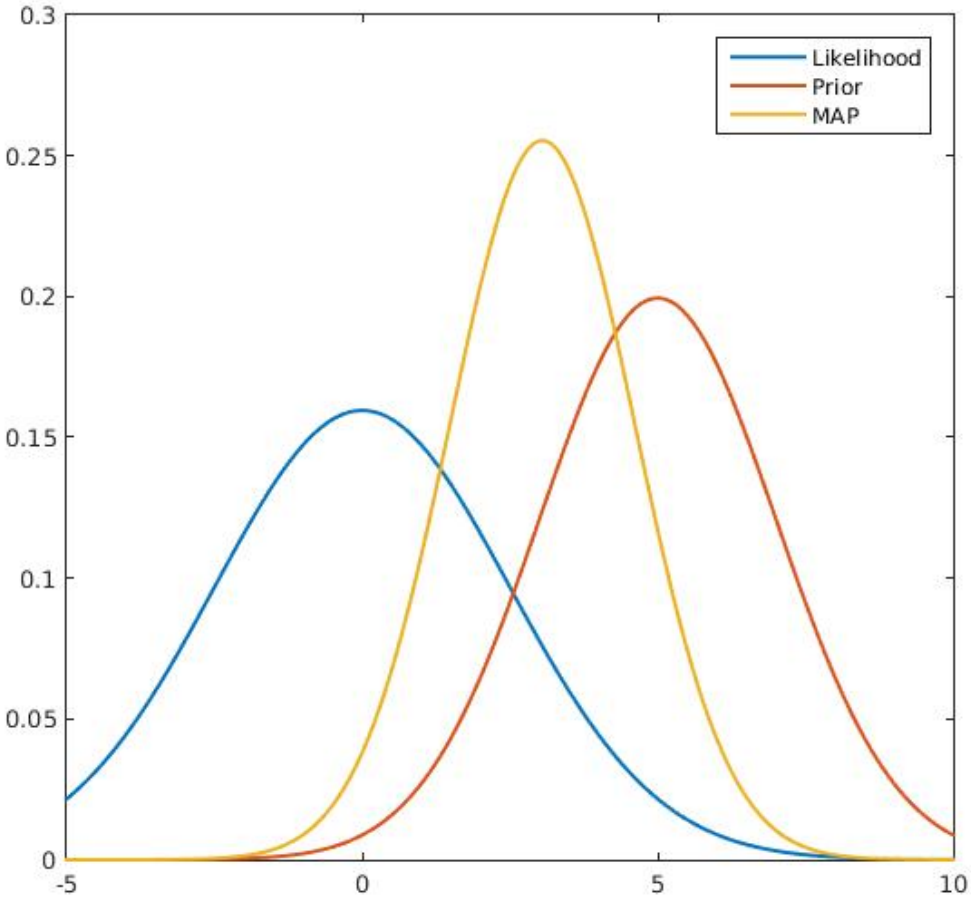


Abbildung 4.4: Einfluss des Priors bewirkt Bias, aber kleinere Varianz, aber nicht in dieser Abbildung :-)

- Logarithmiere Gl. (3)

$$\log p(\theta|Daten) \propto \log p(Daten|\theta) + \log p(\theta) = \sum_{i=1}^{N} \log p(x_i, \theta) + \log prior(\theta)$$

Einfluß log-Likelihood : $O(N)$, Einfluß Prior : $O(1)$,

Asymptotisch hat prior keinen Einfluss

- Problem: Prior im allgemeinen nicht bekannt
- Liefert im frequentistischen Sinne im endlichen verzerrte Schätzer
- Vorteil Bayes'scher Verfahren: Prior kann sinnvoll Vorwissen einbauen

  Akkumulation von Information bei Folge von Experimenten, experimentelle Versuchplanung [38], empirical Bayes

Speziell wichtig bei schlecht-gestellten inversen Problemen

- Einfachstes Beispiel:

$$\begin{aligned} y &= Ax + \epsilon \\ \hat{x} &= A^{-1}y \end{aligned}$$

  y wird gemessen, x soll ermittelt werden

- Ist A singulär oder schlecht konditioniert, i.e. fast singulär, grosse

$$\text{Konditionszahl} = \frac{\text{größter Eigenwert}}{\text{kleinster Eigenwert}}$$

  wird x zwar unverzerrt geschätzt, aber Schätzer hat riesengroße Varianz & damit großen mean square error

$$MSE =< (\hat{x} - x_0)^2 >= Bias^2 + Var(\hat{x})$$

- Prior kann $Var(\hat{x})$ (stark) reduzieren, bewirkt aber (nur kleinen) Bias.

  Stichwort: Regularisierung

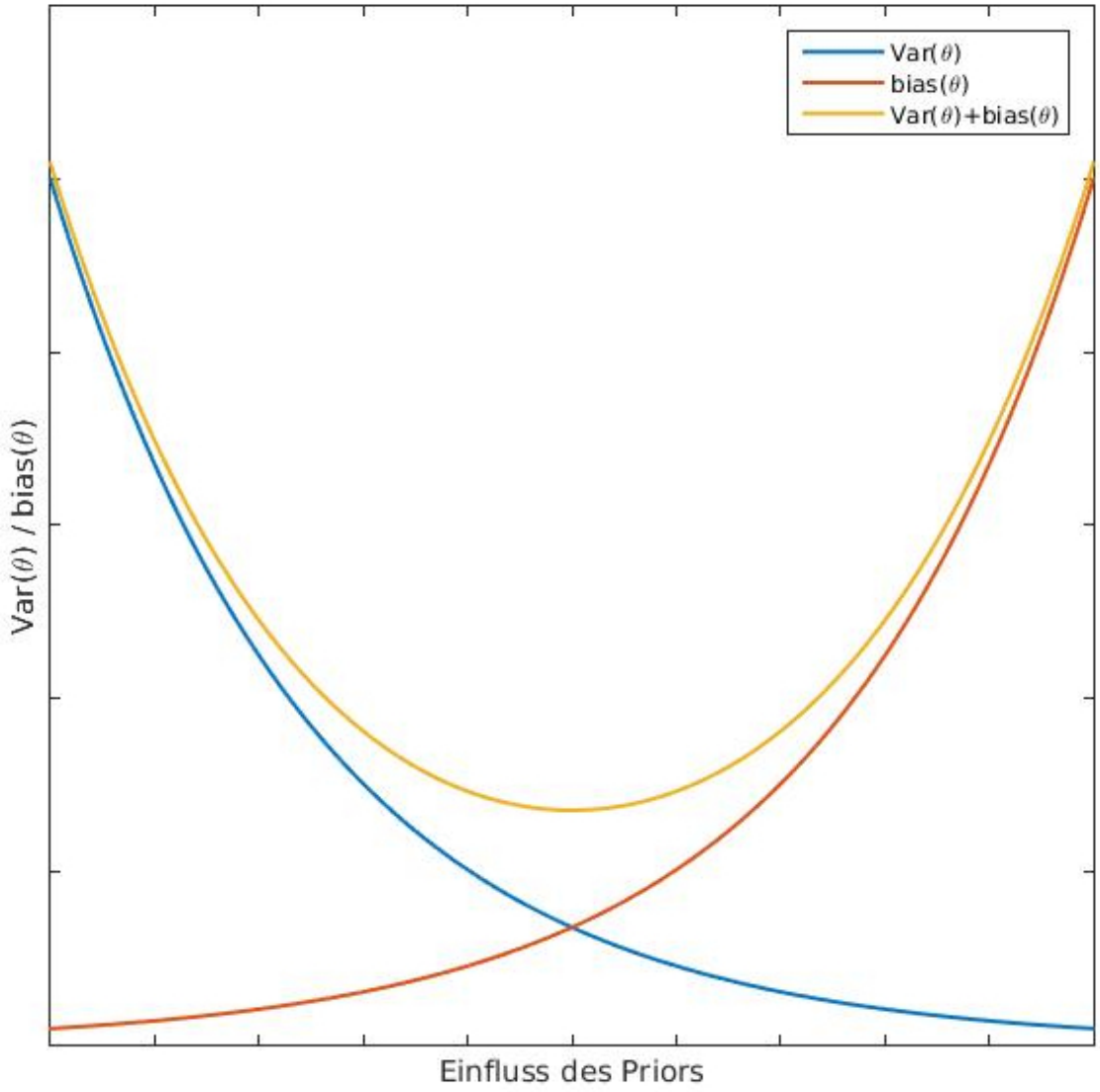


Abbildung 4.5: Verhalten von Bias^2 und Varianz je nach Einfluss des Priors, bias^2 muss bei Null anfangen

Sinnvolle priors:

- $|a|$ sei klein: $p(a) \propto e^{-|a|}$
- Schätzen einer Funktion $f(x, a)$.

$f(x,a)$ sei glatt: $p(f(x,a)) \propto exp(-\frac{\partial^2}{\partial x^2} f(x,a))$

Berechnung von $p(\theta|Daten)$

- Führt auf komplizierte hoch-dimensionale Integrale
- Monte Carlo Markov Chain Verfahren [20, 29, 49], siehe Kapitel 14
- Stochastischer Prozeß auf den Parametern
- Stationäre Dichte ist die gewünschte, sample davon

4.4 Profile Likelihood

Konfidenzintervalle basierend auf Fisher Information Matrix:

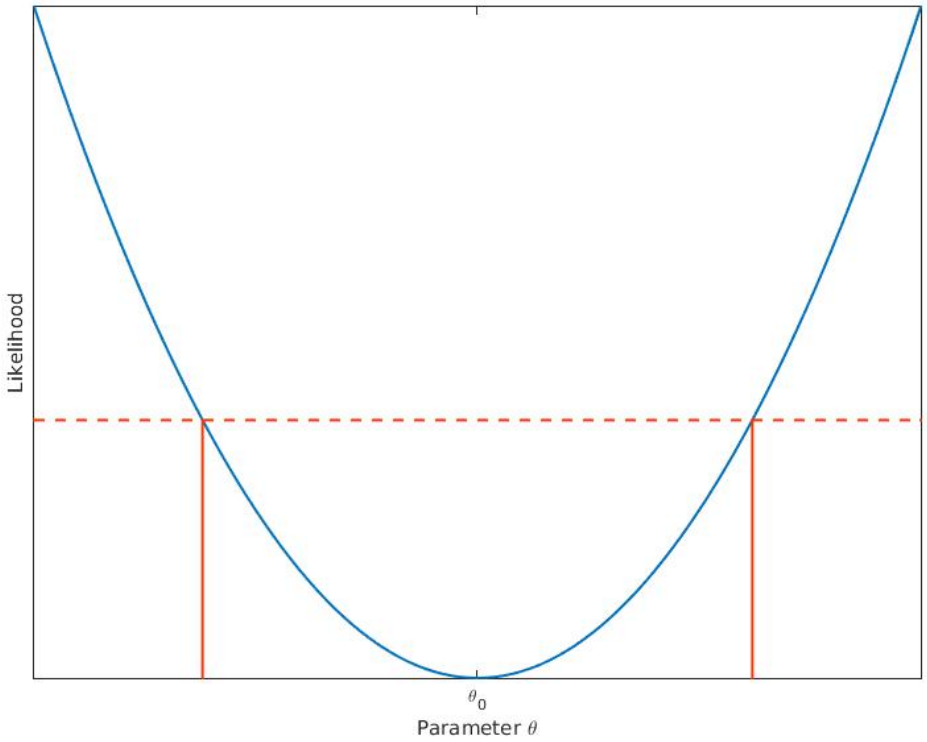


Abbildung 4.6: Fisher Information Matrix

- Starke Annahmen an die Asymptotik: Quadratische Näherung muss stimmen

  Gilt global nur für lineare Modelle

  $$y = \sum_i a_i x_+ \epsilon$$

  Sonst nur lokal um Optimum

- Gilt die Asymptotik, zwei mögliche Aussagen:
  - Quadratisch: Endliche Konfidenzintervale
  - Flach: Parameter nicht bestimmt: Strukturell nicht-idendifizierbar
    * Parameter kann auf Grund der Modellstruktur nicht bestimmt werden
    * (Triviales) Beispiel

      $$y = (ab)x$$

    * (Hochgradig) nichttriviale Beispiele z.B. bei partiell beobachteten Differentialgleichungen

- Nicht reparametrisierungsinvariant.

  Transformiert man Parameter, z.B. logarithmieren, ändern sich Konfidenzintervalle nicht entsprechend der Transformation

- Gilt die Asymptotik nicht, kann alles passieren

Alternative: Profile Likelihood

$$PL(\theta_i) = \max_{\theta_{j \neq i}} L(\theta)$$

Fahre jeden Parameter durch, reoptimiere die anderen

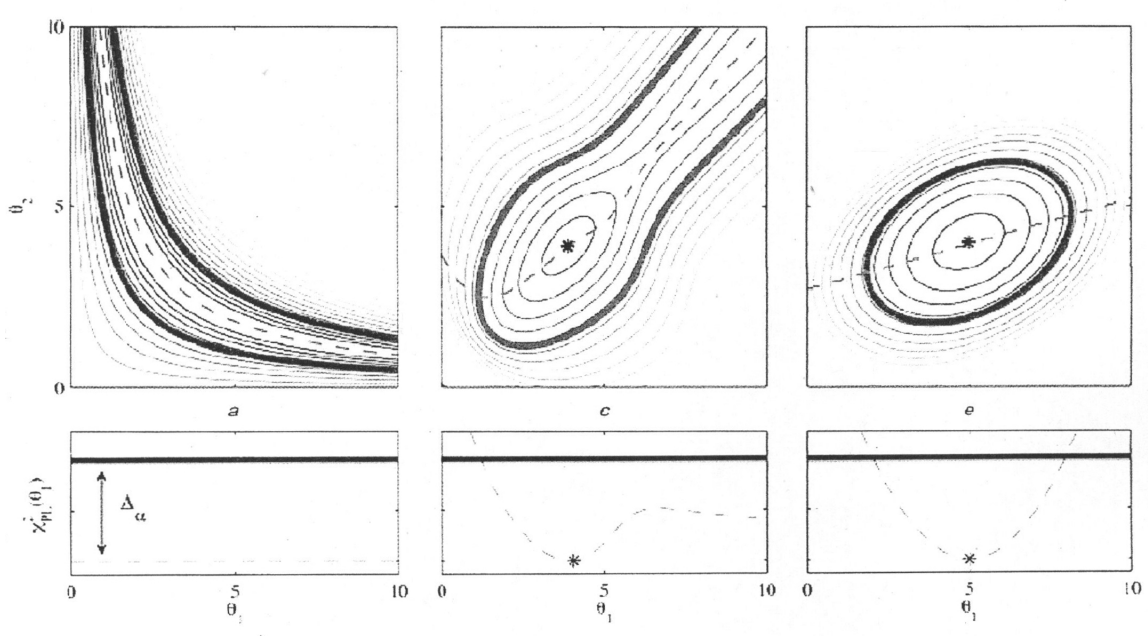


Abbildung 4.7: Profile Likelihood Schätzer, Wahl des Vertrauensintervalls

Konfidenzintervalle gegeben durch:

$$PL(\theta_i) - L(\hat{\vec{\theta}}) \leq \chi^2_{(1-\alpha,1)}$$

Begründung in Kap. 5.2

Eigenschaften:

- Sehr viel schwächere Asymptotik als Fisher Informations Matrix basierte Konfidenzintervalle. Konvexität der Likelihood ausreichend.
- Reparametrisierungsinvariant
- Erlaubt Aussagen, wenn quadratische Näherung nicht gültig
- Ermöglicht Modellreduktion
- Ermöglicht experimentelles Design

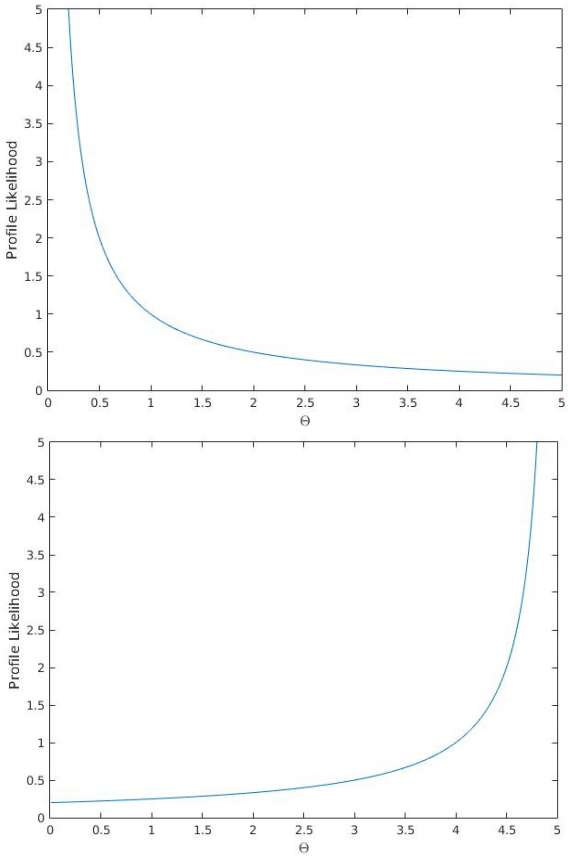


Abbildung 4.8: Mögliche Verläufe untere, obere Schranke s. Übung

- Ermöglicht Definition von praktischer Nicht-identifizierbarkeit [54]: Profile likelihood nicht flach, aber Konfidenz-Intervall unendlich.

  Ein Fall, der häufig mit mehr Daten zu lösen ist

Lessons learned:

- Maximum likelihood Schätzer das beste Pferd im Stall
- Gauß-verteilter Fehler => MLE = gewichtete kleinste Quadrate
- Cramér-Rao Schranke gibt maximal mögliche Genauigkeit
- Bayes'sche Verfahren können Vorinformation berücksichtigen
- Profile Likelihood ist hoch-informative Alternative zu asymptotischen Konfidenzintervallen

5 Modellselektion

In der Regel ist das wahre Modell nicht bekannt, eine endliche Anzahl von Kandidatenmodellen liegt vor

Zwei wichtige Fälle

- Geschachtelte Modelle (nested)
  - Sei M_1 ein Untermodell mit r_1 Freiheitsgraden (θ_1) von
  - wahrem Obermodell M_2 mit r_2 Freiheitsgraden (θ_2)
  - H_0: M_1 ist eine zulässige Vereinfachung von M_2

  Einfachster Fall:

  - M_1: 1. Komponente von $\theta_1 = 42$
  - M_2: 1. Komponente von $\theta_2 \subset R$
  - $r_1 = r_2 - 1$
- Nicht-geschachtelte Modelle (non-nested)
  - M_1 und M_2 streiten sich um Erklärung
  - M_1: $y = \sin ax$ vs. M_2: $y = \exp(bx) + cx^2$
  - M_1 und M_2 stehen für unterschiedliche Physik

Definition: Konsistentes Modellselektionsverfahren: Für $N \to \infty$ wird das wahre Modell mit Wahrscheinlichkeit 1 gefunden, sofern es unter den Kandidaten ist

Occam's Rasor: Mache es nur komplizierter, wenn es sein muss.

Alle Modellselektionsverfahren

- tragen dem Umstand Rechnung, dass ein größeres Modell immer mehr erklären kann
- bewerten statistisch, ob sich der größere Aufwand lohnt

5.1 F-Test

Mutter aller Modellselektions-Test:
Gegeben

- Regressionsmodelle, Gauß'sche Fehler, Least-Squares-Schätzproblem
- Modell M_1 mit k_1 Parametern, $\chi^2(M_1)$, Degress of Freedom: $N - k_1$
- Modell M_2 mit k_2 Parametern, $\chi^2(M_2)$, DoF: $N - k_2$
- Modelle geschachtelt mit $k_2 > k_1$
- M_2 wahres Obermodell
- H_0: M_1 ist zulässige Vereinfachung von M_2
- Unter H_0:
$$F = \frac{(\chi^2(M_1) - \chi^2(M_2))/(k_2 - k_1)}{\chi^2(M_2)/(N - k_2 - 1)}$$
ist F-verteilt mit $k_2 - k_1$ und $N - k_2 - 1$ Freiheitsgraden.
- Beispiel
  - M_1: $y = a + bx$
  - M_2: $y = a + bx + cx^2$

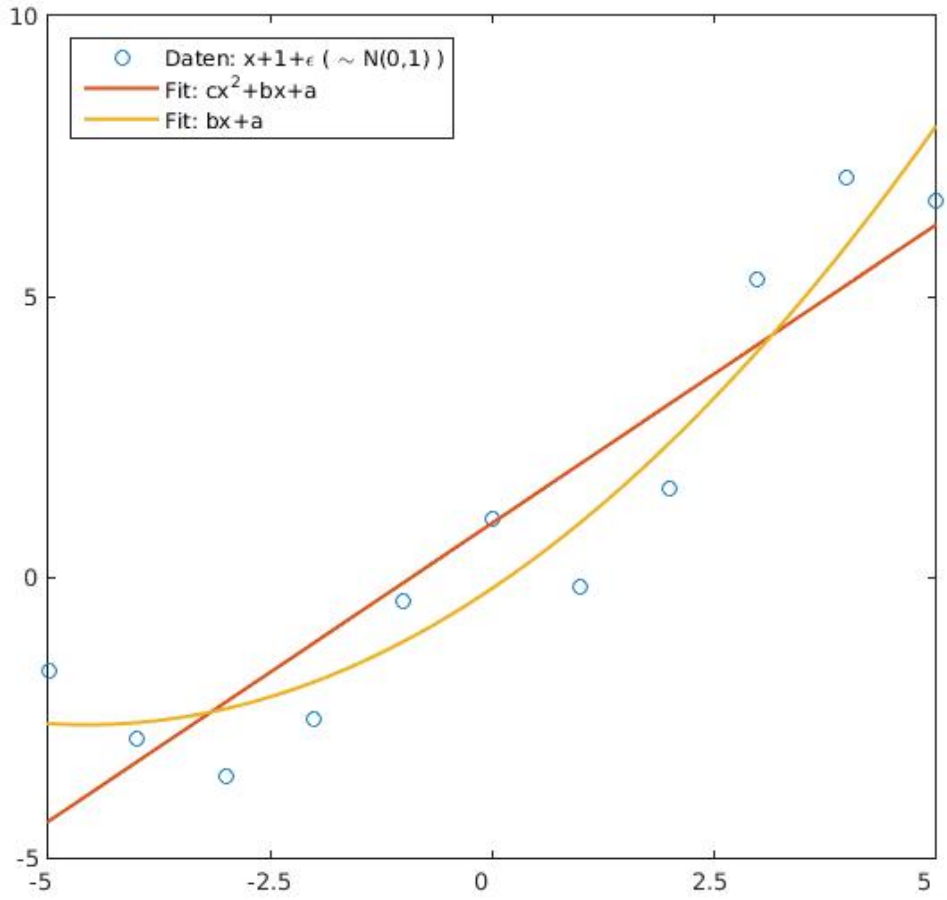


Abbildung 5.1: Verrauschte Daten eines linearen Verlaufs mit gefitteter Gerade und gefitteter Parabel. Zusätzliche Zeichnung mit Parabel als wahrem Modell.

  - F-Test misst Maß des overfitting

- Senkt man Signifikanzniveau α mit N, ergibt sich konsistentes Selektionsverfahren [4, 47]. Ist H_0 wahr, wird sie nicht abgelehnt

5.2 Likelihood Ratio Tests

Beste Theorie-Literatur: [12]

Nomenklatur:

- Gegeben Modell M mit Parametervektor $\theta \subset \mathbb{R}^r$.
- Wahrer Parameter: θ_0
- Geschätzter Parameter: $\hat{\theta}$

Erster LRT:

- H_0 : M ist wahr
- H_1 : M ist nicht wahr

Annahmen:

1. θ_0 liegt nicht auf dem Rand des Parameterraums. Dann MLEs asypmtotisch normal, i.e.:

   $$\sqrt{N}(\hat{\theta} - \theta_0) \sim N(0, \Sigma)$$

   mit

   $$\Sigma = -N \left(\frac{\partial^2 L(\hat{\theta})}{\partial \theta_i \partial \theta_j} \right)^{-1}$$

3. Das Modell sei identifizierbar, i.e. θ ist eindeutig aus den Daten zu bestimmen, siehe Identifizierbarkeit in Kap. 4.4.

dann gilt asymptotisch:

$$2\left(L(\hat{\theta}) - L(\theta_0)\right) \sim \chi^2_r \quad .$$

Differenz der log-likelihoods ist ratio der likelihoods

Beweis

$$\begin{aligned} L(\theta_0) &= L(\hat{\theta}) + \frac{\partial}{\partial\, \theta_i} L(\hat{\theta})(\theta_0 - \hat{\theta}) + \\ & \frac{1}{2}(\theta_0 - \hat{\theta})^T \frac{\partial^2}{\partial \theta_i \partial \theta_j} L(\hat{\theta})(\theta_0 - \hat{\theta}) + O(|\theta_0 - \hat{\theta}|^3) \quad . \end{aligned}$$

- 2. Term RHS = 0 wegen MLE.
- Vernachlässigung von Termen höherer Ordnung
- Σ^{-1} dreht die Kovarianzen der $\hat{\theta}$ raus.
- Quadratischer Term wird Summe über r quadrierte Standard Gauß-Verteilungen $\Longrightarrow \chi^2_r$ -Verteilung
- Löse auf nach $2(L(\hat{\theta}) - L(\theta_0))$
- Da $L(\theta_0)$ nicht bekannt, eher von theoretischem Interesse, macht aber Prinzip klar.

Aber [72]:

- Schätze θ_0 aus allen Daten
- Viele $\hat{\theta}$ aus Untermengen
- Teste, ob Verteilung $2(L(\hat{\theta}) - L(\theta_0))$ stimmt

2. LRT: Gegeben zwei Modelle
Annahmen:

1. Die Modelle sind geschachtelt, sei M_2 ist Obermodell von M_1.
2. Das Obermodell ist richtig spezifiziert
3. Die wahren Parameter liegen nicht auf dem Rand des Parameterraums
4. Alle Parameter sind identifizierbar unter der Nullhypothese.

H_0: M_1 ist eine zulässige Vereinfachung von M_2

Dann gilt asymptotisch, selbes Argument wie oben:

$$2\left[L(\hat{\theta}_2) - L(\hat{\theta}_1)\right] = 2\left[(L(\hat{\theta}_2) - L(\theta_0)) - (L(\hat{\theta}_1) - L(\theta_0))\right] \sim \chi^2_{r_2} - \chi^2_{r_1} = \chi^2_{r_2 - r_1}$$

Beweis

- Analog zu oben

- Rausdrehen der Korrelationen
- Summe über quadrierte Standard-Gaußverteilungen führt zu χ^2-Verteilung

5.18

Bemerkungen

- Verteilung der LRTs folgt aus asymptotischer Gaußianität der Schätzer
- LRT für Regressionsfall = F-Test, siehe z.B. [62]
- Konsistentes Modellselektionsverfahren:
  Für $N \to \infty$ und Signifikanzniveau α gegen Null wird das wahre Modell mit Wahrscheinlichkeit = 1 gefunden
- Verwandte Tests: Wald Test, Lagrange-Mulitplier Test
- Profile Likelihood ist LRT auf einen Parameter, daher χ^2_1 Verteilung

5. Woche

Bei vielen Modellen aufsteigender Komplexität: Selektionsstrategien für F-Test, LRTs:

- Forward Selection
  - Aufsteigend kompliziertere Modelle testen
  - Nachteile:
    - Falsch Negative $\Longrightarrow$ Early stopping
    - Es gibt mitunter keine natürliche Ordnung im Nichtlinearen
- Backward Selection
  - Vom generellsten Modell absteigend
  - Nachteil
    - Was ist das generellste Modell ?
    - Existenz des Obermodells ?
- Stepwise Selection

  Nach jedem Forward-Schritt, wieder Backward-Schritte

  Ist empfohlen

5.2.1 Nicht-Standard Testsituationen

"Nicht-Standard" meint: Annahmen für obige Resultate nicht erfüllt.

Häufigster Fall:

- Unter H_0 liegt Parameter auf dem Rand des Parameterraums [63, 73]
  Folge: Schätzer kann nicht Gauß-verteilt sein.
- Beispiel: a skalar, Parameterraum $a \geq 0$
  - H_0: $a = 0$
  - H_1: $a > 0$

  Unter H_0, statt (asymptotischer) Gaußverteilung :
  - Potentiell negative Werte werden 0
  - Potentiell positive Werte bleiben, wie gehabt

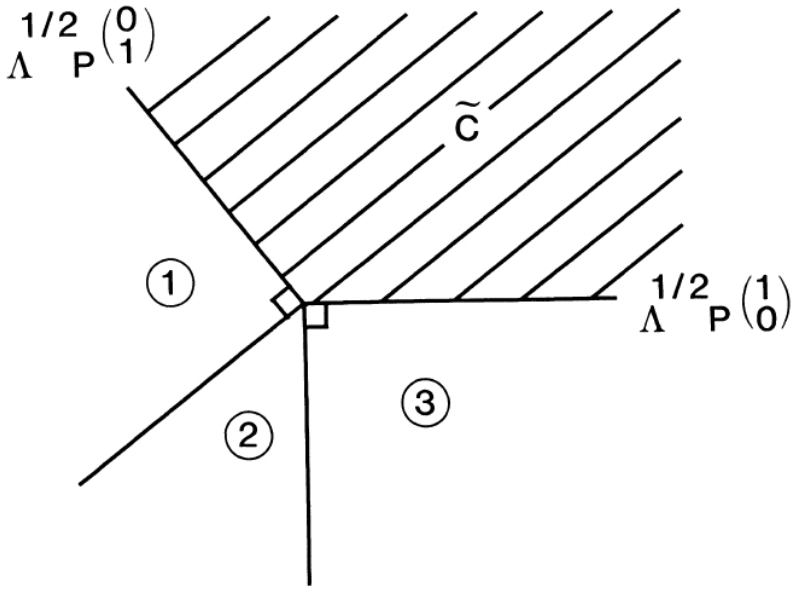


Abbildung 5.2: 2D-Parameterraum: $\widetilde{C}$ zeigt die zulässigen Parameterwerte unter H_1. Unter der Nullhypothese ist der Parameter am Ursprung lokalisiert. Die asymptotische Verteillung des Likelihood Ratio Test ist eine Kombination aus χ^2_0, χ^2_1 und χ^2_2 mit unterschiedlichen Wahrscheinlichkeiten, die vom Winkel in $\widetilde{C}$ abhängen. [63]

- Test Statistik:

$$2\left[L(\hat{\theta}_2) - L(\hat{\theta}_1)\right] \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2, \quad \chi_0^2 = \delta(0)$$

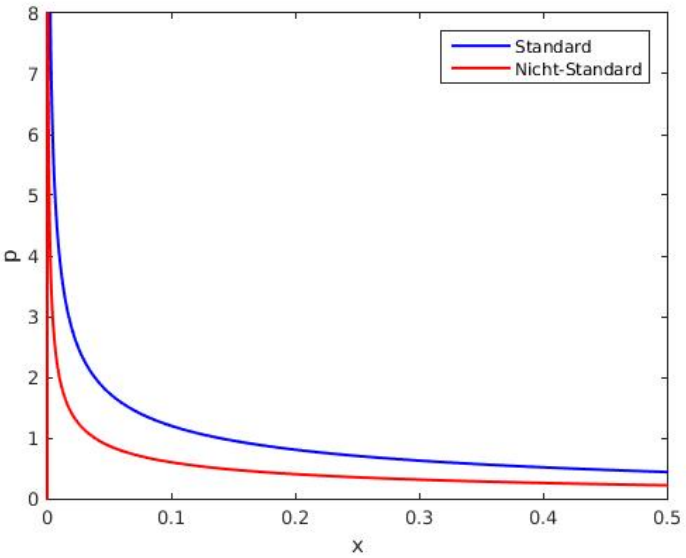


Abbildung 5.3: Normal vs. Nichtstandard

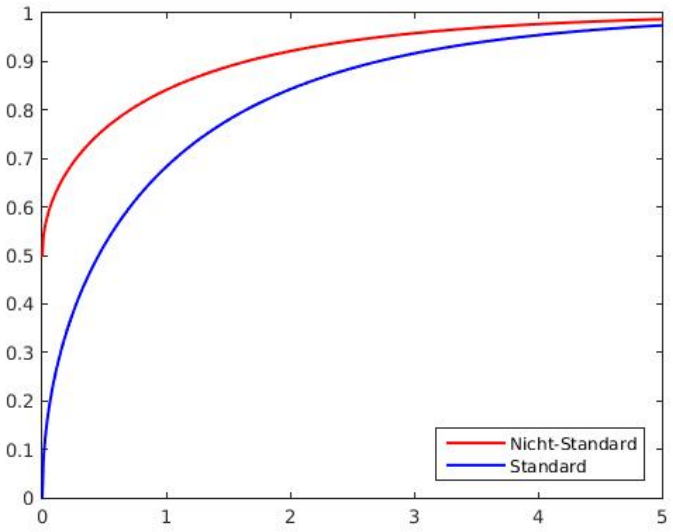


Abbildung 5.4: Kumulative Verteilungen

- Wichtig: Berücksichtigt man "Parameter auf dem Rand" nicht, wird der Standard-LRT konservativ.

5.2.2 Nicht geschachtelte Modelle [73]

Sind die Modelle nicht geschachelt (non-nested) könnte man allgemeines Obermodell als große Schachtel nehmen.
Verbietet sich im allgemeinen, da

- Anzahl der identifizierbaren Parameter begrenzt, siehe Kap. 4.4
- Dann ständig Nicht-Standard Test Situationen auftauchen würden

Simulativer Ausweg:

- Fitte Modell 1 und 2 an die Daten, berechne die Differenz ihrer Likelihoods
- Nimm an, Modell 1 ist korrekt
- Simuliere vielfach Daten aus Modell 1
- Fitte beide Modelle an die simulierten Daten

- Bestimme daraus die Verteilung der Differenz ihrer Likelihoods
- Checke, ob Originalfit-Likelihood-Differenz mit simulierter Differenz-Verteilung verträglich ist
- Mache das selbe ausgehend von Modell 2
- Vier Möglichkeiten
  - Beide Modelle werden abgelehnt
  - Modell 1 oder Modell 2 wird abgelehnt
  - Kein Modell kann abgelehnt werden

5.3 Akaike Information Criterion (AIC)

Akaike selber nannte es in [2] An Information Criterion, AIC :-)
Originalliteratur: [1, 2] schön dargestellt in [37]

Grundideen:

- Vereinigung von Parameterschätzung und Modellselektion
- Basierend auf einem Entropiemaß: Kullback-Leibler Distanz, einiges Ausintegrieren und Nähern
- Formal analog zu Cross-Validation [68]

Führt auf

$$AIC(M) = -2L(\hat{p}) + 2k, \quad k = dim(p)$$

Zur Modellwahl, wähle Modell mit kleinstem AIC, kein schrittweises Vorgehen.

Kommentare:

- Auf Grund der Einfachheit sehr beliebt.
- Aber: Betrachte geschachtelte Modelle M_1 und M_2 mit $\Delta k = 1$, in M_1 ein Parameter fixiert

$$\begin{aligned} AIC(M_1) &= -2(L(M_1)) + 2k_1 \\ AIC(M_2) &= -2(L(M_2)) + 2k_2 \\ AIC(M_1) - AIC(M_2) &= -2(L(M_1) - L(M_2)) + 2 \end{aligned}$$

Erinnere LRT:

Unter H_0

$$2(L(M_2) - L(M_1)) \sim \chi^2_1$$

Ergo: AIC ist LRT mit kritischem Wert α

$$\chi^2_1(2) = \alpha, \quad \text{ergibt } \alpha = 15.7\%$$

Im test-theoretischen Sinne: 15.7 % Fehler 1. Art

Führt systematisch zu zu komplexen Modellen

- Kein konsistentes Modellselektionsverfahren
- Aber gut, um Modell mit guter Vorhersagekraft zu finden
- Verhalten für Parameter auf dem Rand und bei nicht-identifizierbaren Parametern unklar

  Literatur: [64]

5.4 Bayes'sches Information Criterion (BIC)

- Geniales vier Seiten paper [61]
- Annahme: Schwächste Bayes'sche Priors und Vernachlässigung von Termen höherer Ordnung
- Ergibt

$$BIC = -2L(\hat{p}) + k\log(N), \quad k = dim(p)$$

  Berücksichtigt Datenanzahl

- Signifikanzniveau für einen Parameter Unterschied [70]

$$Prob(\chi^2_1 > log(N))$$

- Wahl des kleinsten BIC ergibt konsistentes Modellselektionsverfahren
- Vergleiche AIC vs. BIC, siehe [3, 7, 34, 45, 70]

Lessons learned:

- Modellselektionsverfahren bewerten höhere Erklärungsmöglichen komplexerer Modelle in Anbetracht der höheren Anzahl der Parameter (Occam's Rasor).
- F-Test und Likelihood Ratio Test legen Skala fest, Teststatistik
- AIC und BIC reihen einfach nur
- F-Test, LRT Test und BIC sind konsistente Modellselektionsverfahren
- AIC bevorzugt systematisch größere Modelle als notwendig

Teil II
Numerik

Es gibt zwei Sorten Numerik:

- Die, die man verstehen sollte, und ..
- ... die, die man nur zu kennen braucht

6 Erzeugung von Zufallszahlen

- Problem:

  Wie auf einem (deterministischem) Computer "zufällige" Zahlen erzeugen ?

- Diskussion: Erkennen von Zufall. Statistische Hypothese "5.6 ist zufällig" ist nicht abzulehnen. Ebenso "5.6 ist deterministisch.

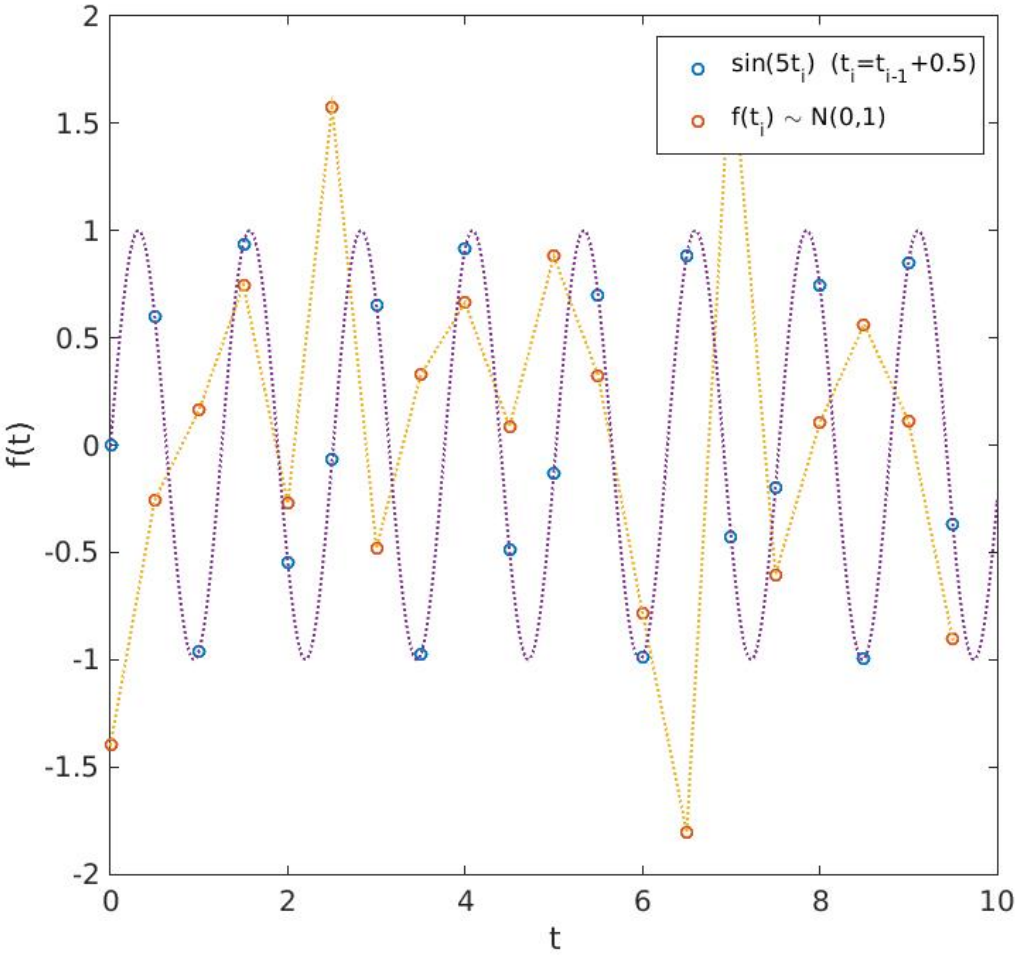


Abbildung 6.1: Sinus Reihe und weißes Rauschen

- Zufall = Nicht Vorhersagbar, de facto Definition
- Lösung :

  (Deterministische) chaotische dynamische Systeme zeigen Eigenschaften, die von echtem Zufall nicht zu unterscheiden sind.

  Mutter aller chaotischen Systeme: Das Lorenz-System

  – Langfristige Wettervorhersage nicht möglich

– Navier-Stokes Gleichungen in einem Kasten
– Randbedingungen: unten (Erde) warm, oben (Weltraum) kalt
– Fourierentwicklung für Temperatur, Druck, Dichte
– Nur erste Mode mitnehmen
– Ergibt:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z \end{aligned}$$

– Mittagessen Anekdote
– Chaos = Empfindliche Abhängigkeit von den Anfangsbedinungen. Positive Lyapunov-Exponenten

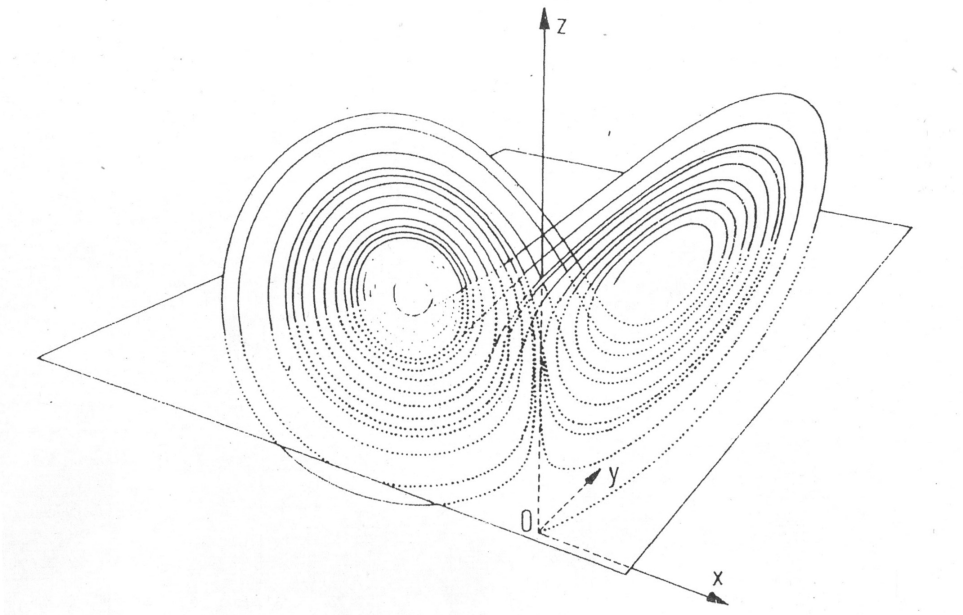


Abbildung 6.2: Lorenz 1963 'Deterministic Aperiodic Flow' [42]

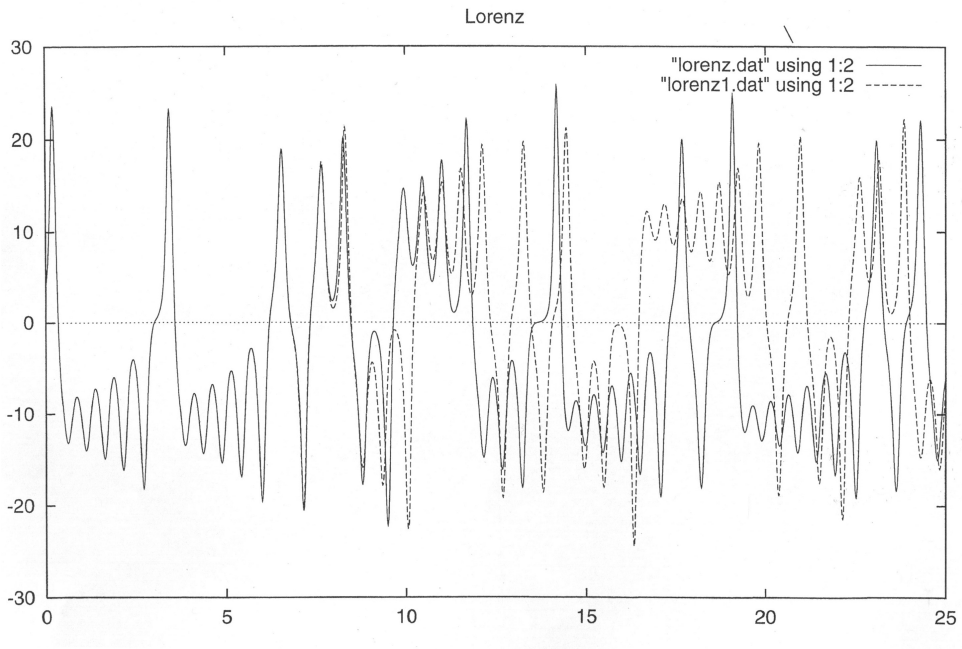


Abbildung 6.3: Lorenz 1963 'Deterministic Aperiodic Flow' [42]

- Zeichnung Max(i+1)/Max(i)
- (Pseudo-)Zufallszahlen-generator: Poincaré - Schnitt durch ein hochdimensionales deterministisches chaotisches System.
- Ähnliche Werte von $x(t)$ habe sehr verschiedene Werte $x(t+1)$.
- replace by Dreieck (0.,1;1)
- Alle ZV-Erzeugung basiert auf gleichverteilten Zufallsvariablen
- replace by

$$\begin{aligned} x(t+1) &= f(x(t)), \quad x(t) \subset [0,1] \\ x(t+1) &= a\,x(t) \bmod 1, \qquad a \text{ sehr große Zahl} \end{aligned}$$

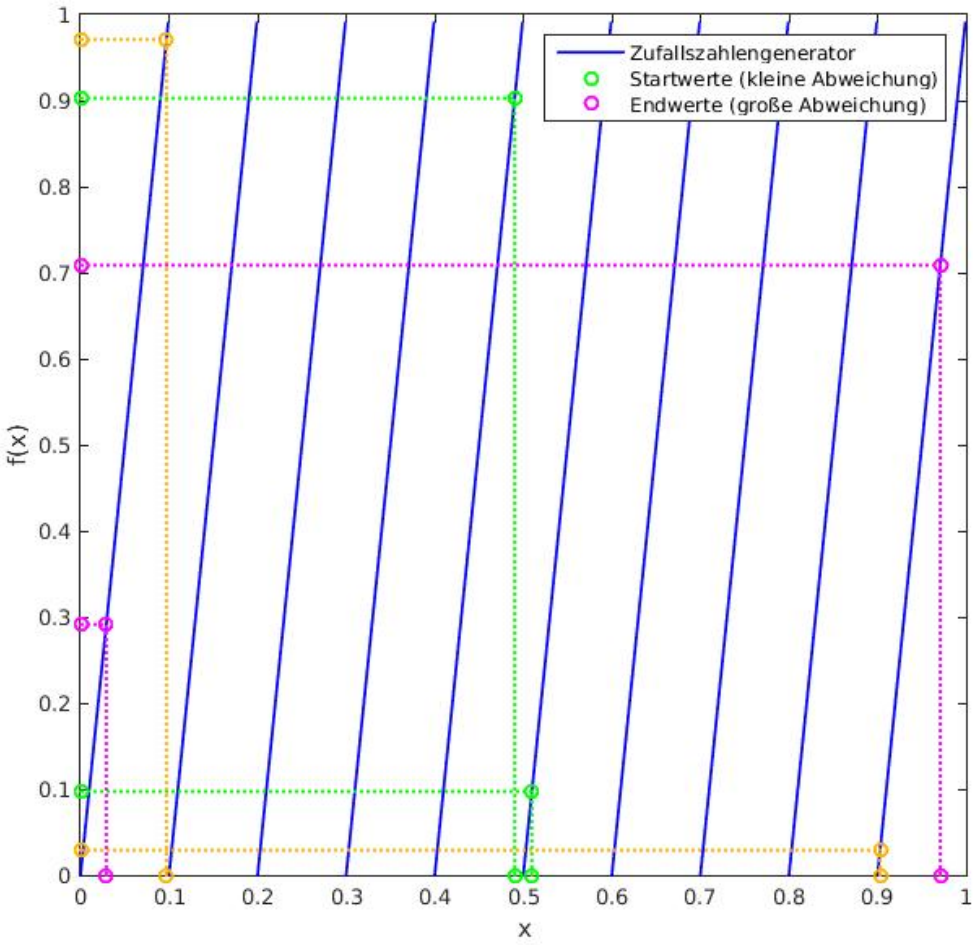


Abbildung 6.4: Prinzip eines Zufallszahlengenerators

Transformationsmethode:

- Zufallsvariable X
- Bilde $Y(X)$
- Es muß gelten

$$1 = \int dx\, p_X(x) = \int dy \left|\frac{dx}{dy}\right| p_X(x(y)) = \int dy\, p_Y(y)$$

ergo:

$$p_Y(y) = \left|\frac{dx}{dy}\right| p_X(x)$$

- In der Regel:
  - X gleichverteilt
  - $Y(X)$ klug gewählt
  - Für Verteilung von Y gilt somit

$$p_Y(y) = p_X(x) \left|\frac{dx}{dy}\right| = \left|\frac{dx}{dy}\right|$$

Beispiele:

- Exponentiell verteilte Zufallsvariablen

$$p(y) = \frac{1}{\tau} e^{-y/\tau}$$

  - Sei X gleichverteilt
  - Wähle $y(x) = -\log x$, $x = e^{-y}$
  - Gibt:

$$p(y) = \left|\frac{dx}{dy}\right| = e^{-y}$$

- Standard Gauß-verteilte ZV

$$p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

  - Hier zweidimensionale Transformationsmethode, Box-Müller Verfahren

$$p(y_1, y_2) = p(x_1, x_2) \left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right|$$

    $|.|$: Determinate der Jacobi Matrix
  - Sei X gleichverteilt

- Wähle klug:

$$\begin{aligned} y_1 &= \sqrt{-2\log x_1}\cos 2\pi x_2 \\ y_2 &= \sqrt{-2\log x_1}\sin 2\pi x_2 \end{aligned}$$

$$\begin{aligned} x_1 &= \exp\left[-\frac{1}{2}(y_1^2+y_2^2)\right] \\ x_2 &= \frac{1}{2\pi}\ \text{atan}\ \frac{y_2}{y_1} \end{aligned}$$

$$\left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| = \frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-y_2^2/2}$$

  - Liefert zwei Standard-Normal verteilte Zufallszahlen für zwei gleichverteilte.
  - Allgemeine Gauß-Verteilung durch Shift um μ und Skalierung mit σ

- Cauchy

$$p(y) = \frac{1}{\pi}\frac{\gamma}{(y-a)^2+\gamma^2}$$

  - Sei x gleichverteilt in $[-0.5, 0.5]$
  - Dann ist $y = \gamma\tan\pi x$ Cauchy-verteilt wegen

$$(\text{atan}\ x)' = \frac{1}{x^2+1}$$

$$p_{Cauchy}(x,0,1) = \tan(\pi(U[-1/2,1/2])$$

  - Es gilt auch:

$$''\text{Cauchy}(y,0,1) = \frac{N(0,1)''}{N(0,1)}$$

    Merke: Quotienten von Zufallsvariablen können garstig sein

Lessons learned:

- Zufallszahlenerzeugung in deterministischem Computer beruht auf Nichtlinearer Dynamik
- Gleichverteilung ist die Mutter aller Zufallszahlen
- Die anderen z.B. durch Transformationsmethode

7 Lösung linearer Gleichungssysteme

Gegeben Matrix A und Vektor b, finde Vektor x für:

$$Ax = b$$

Ubiquitäres Problem:

- Physik: Streuexperimente, Rheologie
- Numerik, siehe Kap. 9 Optimierung, Kap. 10 Nichtlineare Modellierung.
- b i.d.R. fehlerbelastet
$$Ax = b + \epsilon$$

Ziel :

$$x = A^{-1}b, \text{ oder manchmal: } \tilde{x} = \tilde{A}^{-1}b, \quad \tilde{A}^{-1} \text{ modifiziertes } A^{-1}$$

Probleme und Methoden unterscheiden sich nach Eigenschaften der Matrix A:

- A sei $N \times N$ Matrix (wichtigster Fall)

  Gute Chance für eindeutige Lösung.
  Mögliche Probleme:

  - Lineare Abhängigkeit von Zeilen/Spalten von A
    - Matrix singulär $\Longrightarrow$ keine eindeutige Lösung.
  - "Fast" lineare Abhängigkeit
    - Matrix schlecht-konditioniert.
    - Seien λ_i die absteigend sortieren Eigenwerte:
      Konditionszahl K:
      $$K = \frac{\lambda_1}{\lambda_N}$$
    - Großes K: Fehler auf b wird in Lösung x verstärkt, siehe unten.
  - N sehr groß:
    - Rundungsfehler können kumulieren.

- A sei $M \times N$ Matrix, $M < N$ (oder A sei singuläre $N \times N$ Matrix)

  - Unterbestimmtes Gleichungssystem.

- Lösung nicht eindeutig.
- Lösung durch Zusatzforderungen eindeutig machbar, siehe unten.

- A sei $M \times N$ Matrix, $M > N$
  - Überbestimmtes Gleichungssystem.
  - Suche einen Kompromiß, der Gleichungen simultan möglichst gut erfüllt.
  - Für "möglichst gut" im Sinne des Mittleren quadratischen Fehlers folgt eindeutige Lösung durch:

$$\begin{aligned} (A^T A)\, x &= A^T b \\ x &= (A^T A)^{-1}\, A^T\, b \end{aligned}$$

    $(A^T A)^{-1} A^T$ heißt Pseudo-Inverse oder auch Moore-Penrose-Inverse. Wird in Kap. 10 Nichtlineare Modellierung behandelt.
  - Schnack über "double descent"

6.18

7.1 Gauß-Jordan - Elimination

A sei $N \times N$ Matrix, gut konditioniert.

- Grundlage:

  Bilden von Linearkombinationen des Gleichungssystems ändert die Lösung nicht.

- Idee:

  Bringe System auf obere Dreiecksgestalt.

  Sei:

$$E_i : a_{i1}x_1 + \ldots + a_{in}x_n = b_i$$

  die i-te Zeile des Gleichungssystems.

- Eliminiere x_k in $E_{k+1}, \ldots, E_n$ durch:

  for $(k = 1, \ldots, N)$:

$$\begin{aligned} m_{ik} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k+1, \ldots, N \\ E_i^{(k+1)} &= E_i^{(k)} - m_{ik} E_k^{(k)} \end{aligned}$$

$a_{kk}^{(k)}$ heißt Pivot-Element.

- Ergebnis:

$$A^{(N)} = U, \quad b^{(N)} = g, \quad Ax = b \Longleftrightarrow Ux = g \quad U \text{ wie upper}$$

- Lösung x durch Backsubstitution
  for $(k = N, \ldots, 1)$:

$$x_i = \frac{1}{u_{ii}} \left[g_k - \sum_{j=k+1}^{N} u_{kj} x_j \right]$$

- Problem:
  Wenn $a_{kk}^{(k)}$ klein, führt das in

$$E_i^{(k+1)} = E_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} E_k^{(k)}$$

  zu Rundungsfehlern

- Lösung:
  - "Equilibriere" die Matrix A
  - Sortiere die Zeilen vorher so, daß immer große Zahlen auf der Diagonale stehen:
  - Heißt Pivoting, ausführliche Diskussion in Stoer Kap. 4.5
- Zeichnung
- Kompexität: $\mathcal{O}(N^3)$
- Nachteil: Muß für jedes neue b neu berechnet werden

7.2 Matrix-Zerlegungen

Matrix-Zerlegungen machen einem das Leben einfacher

7.2.1 LU-Zerlegung

$N \times N$ Matrix läßt sich schreiben als

$$A = LU$$

mit

- L: untere Dreiecksmatrix mit Einsen auf der Diagonale
- U: (beliebige) obere Dreiecksmatrix.
- Crout's Algorithmus macht Zerlegung elegant (mit pivoting).
- Stabil für gut konditionierte Matrizen
- Kompexität: $\mathcal{O}(N^3)$

Anwendungen:

- Lösung für $Ax = b$ durch Forward- und Backsubstitution

  $$\begin{aligned} Ax &= (LU)x = L(Ux) = Ly = b \\ Ly &= b \\ Ux &= y \end{aligned}$$

  Vorteil: Zerlegung muss für verschiedene bs nur einmal berechnet werden

- Berechnung von A^{-1} durch:
  for $(j = 1, \ldots, N)$

  $$b_i^{(j)} = \delta_{ij}$$

  Lösungen $x_i^{(j)}$ geben Spalten von A^{-1}.

- Sparsame Berechung der Determinante durch

  $$\det(A) = \det(LU) = \det(L)\det(U) = \prod_{i=1}^{N} U_{ii}$$

  mit $\mathcal{O}(N^3)$ statt $\mathcal{O}(N!)$ bei der definitionsbasierten Berechnung.

6. Woche

7.2.2 Cholesky-Zerlegung

- Sei A symmetrisch und positiv definit, i.e. $vAv > 0 \; \forall \, v \neq 0$
- "Wurzel" der Matriz:

$$A = LL^T \quad L \text{ wie lower triangular}$$

- Kovarianzmatrizen fallen in diese Klasse.

Erzeugung korrelierter Gaußscher Zufallsvariablen:

- Kovarianzmatrix, für $< x_i > = 0$

$$C_{ij} = \langle x_i x_j^T \rangle$$

  Bilde:

$$BB^T = C$$

- Erzeuge unkorreliere ZVs y_i und bilde korrelierte durch:

$$\vec{x} = B\vec{y}$$

- Beweis:

$$C =< xx^T >=< By(By)^T >= B < yy^T > B^T = B\delta_{ij}B^T = BB^T = C$$

7.2.3 Singulärwert-Zerlegung

A sei $N \times N$ Matrix, schlecht konditioniert

- Konditionszahl revisited:
  Konditionszahl K:

$$K = ||A||\,||A^{-1}|| = \frac{\lambda_1}{\lambda_N}$$

  mit $|\vec{x}|$ Euklidische Norm, ergibt sich Spektralnorm $||A||$

$$||A|| := \max_{x \neq 0} \frac{|Ax|}{|x|} = \max_{x \neq 0} \sqrt{\frac{x^T A^T A x}{x^T x}} = \sqrt{\lambda_{\max}(A^T A)} \quad \text{Klammer} = \text{"von"}$$

- Wie gehabt:
$$Ax = b$$
- Einfluß von Fehler Δb auf b auf das geschätzte $\hat{x} = x + \Delta x$:
  - Betrachte
$$A(x + \Delta x) = b + \Delta b$$
  - Aus
$$\Delta x = A^{-1}\Delta b$$
    folgt die Abschätzung:
$$|\Delta x| \leq ||A^{-1}||\ |\Delta b|$$
  - Für die relativen Fehler $|\Delta x|/|x|$ folgt mit
$$|b| = |Ax| \leq ||A||\ |x|, \quad \frac{1}{|x|} \leq \frac{||A||}{|b|}$$
    insgesamt:
$$\frac{|\Delta x|}{|x|} \leq ||A||\,||A^{-1}||\frac{|\Delta b|}{|b|} = K(A)\frac{|\Delta b|}{|b|}$$
  - Ergo, großes $K \Longrightarrow$ Verstärkung der Fehler auf b.
  - $K = 10^6$ ist desaströs in single precision.
- Singular-Value-Decomposition (SVD), Karhunen-Loéve-Transformation, Hauptkomponentenanalyse
  - Liefert
$$A = U\,[\mathrm{diag}(w_i)]\,V^T \quad ,$$
    mit
    * orthogonalem U, $N \times N$ Matrix
    * diagonaler $N \times N$ Matrix W mit Singulärwerten $w_i \geq 0$, Vorzeichen in U und V absorbiert [2]
    * orthogonalem V, $N \times N$ Matrix

[2] Ist A symmetrisch, sind die singulären Werte identisch mit den Eigenwerten.

- Für die Mathematik, siehe Stoer, Bulirsch [67] Kap. 6.7
- Inverse :

$$A^{-1} = V\,[\mathrm{diag}(1/w_i)]\,U^T$$

- Lösung von $Ax = b$

$$x = V\,[\mathrm{diag}(1/w_i)]\,U^T b \tag{4}$$

- Vorteil gegen Gauß-Jordan: b muss nicht vorher bekannt sein.
- Gehört zu den 5 wichtigsten Routinen, wo gibt.
- Geht auch für $M \times N$ Matrix, $M \neq N$.

Betrachte: A sei $N \times N$ Matrix, singulär oder schlecht konditioniert

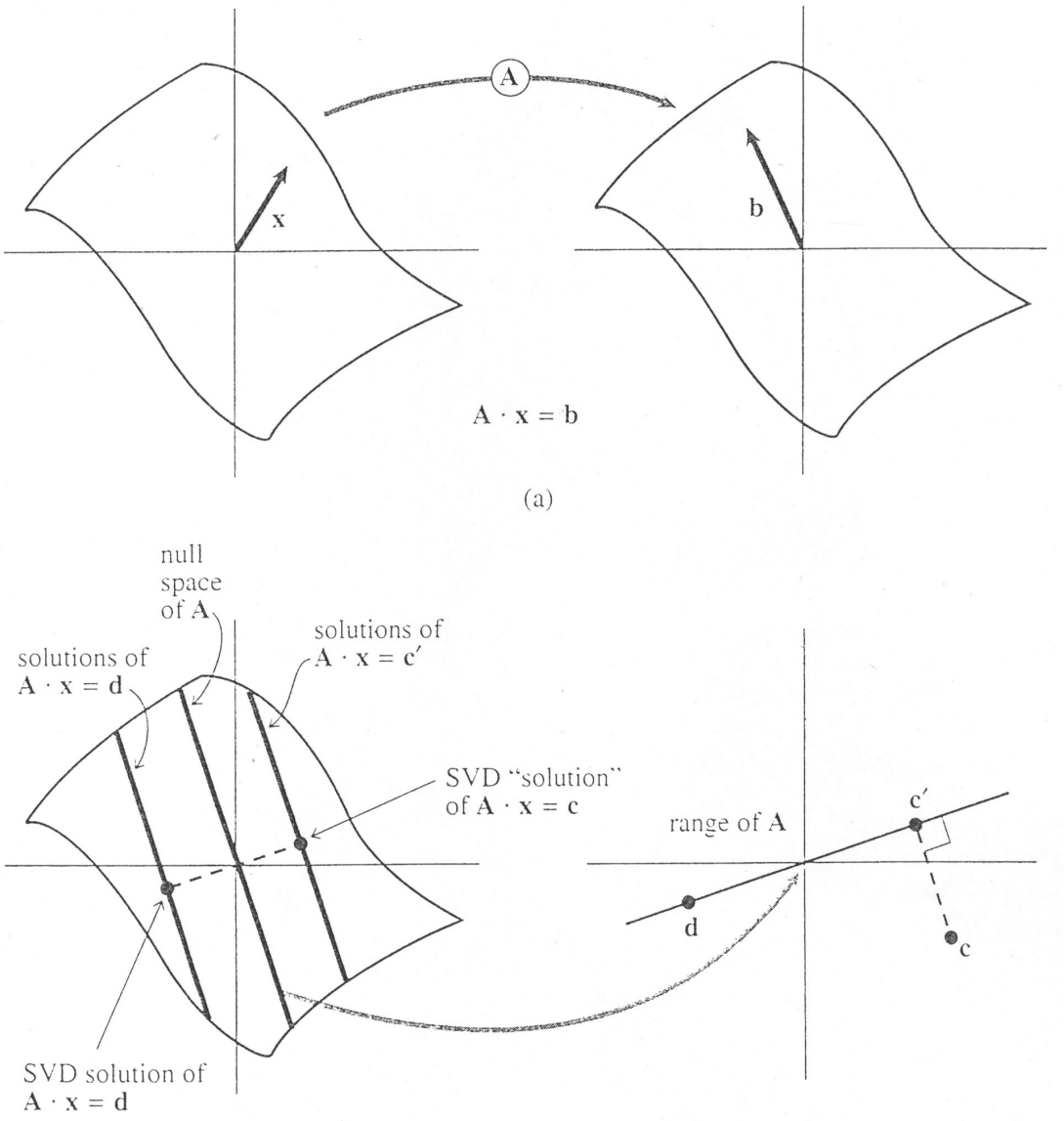


Abbildung 7.1: Singulärwertzerlegung

- Anschauung:
  - Die Eigenvektoren zu den 0- (oder kleinen) Eigenwerten machen beim Invertieren die Probleme.

  - Sorgen für große Fehler.
  - Lösung: Für kleine Eigenwerte w_i, setze in Gl. (4) $1/w_i = 0$ ($\infty = 0$:-)

- Mathematisch:
  - Es wird das x minimaler Norm bestimmt.
    * A singulär. Es gibt Kern x_k mit

      $$Ax_k = 0$$

      Bildraum (range) von A hat dim $< N$
    * Mit x_{nk} nicht aus dem Kern gilt

      $$x = x_{nk} + x_k$$

    * Gewählte Lösung : $x = x_{nk}$
  - Erinnere Bayesianismus, Stichwort: Regularisierung:
    * Zusatzinformation, um die Lösung eindeutig zu machen.
      Hier :

      $$Ax = b \quad " + " \ |x| \text{ minimal}$$

    * Regularisierung bewirkt:
      Reduktion der Varianz auf Kosten eines bias, siehe Übung

- Minimum norm äquivalent zu:

  Suche Lösung, für die gilt:

  $$\text{Suche } x, \text{das } r = |Ax - b|^2 \text{ minimiert}$$

Bemerkung : Alle besprochenen Algorithmen haben Aufwand $\mathcal{O}(N^3)$

Es gibt spezielle Methoden für

- Schwachbesetze große Matrizen. Stoer/Bulirsch Kap. 8
  Aufwand $\mathcal{O}(N)$ oder $\mathcal{O}(N^2)$
- Inverse für "leicht veränderte" Matrizen, Recipes Kap. 2.7:
  - Sherman-Morrison Gleichung
  - Woodbury Gleichung

- Speziell strukturierte Matrizen, Recipes Kap. 2.8 :
  - Matrizen mit Bandstruktur (z.B. bei Finite Element Methods)
  - Vandermonde Matrizen $a_{ij} = \alpha_i^{j-1}$
  - Toeplitz Matrizen $a_{ij} = \alpha_{i-j}$
- Bestimmung von Eigenwerten/Eigenvektoren
  RECIPES Kap. 11, Stoer/Bulirsch Kap. 6
  - Givens- und Householder Reduktionen
  - $A = QR$ Zerlegung, Q orthogonal, R obere Dreiecksmatrix
  - Hessenberg-Form, ab erster Unterdiagonale besetzt

Übung:
Bias und Varianz bei der Lösung schlecht-gestellter inverser Probleme.

Übung:
Erzeugung korrelierter Gaußscher Zufallsvektoren

Lessons learned:

- Gauß-Jordan Elimination
- Verschiedene Zerlegungen, die einem das Leben leicher machen:
  LU, Cholesky, Singulärwert, ...
- Singulärwert-Zerlegung liefert minimum norm Lösung bei schlecht-gestellten Problemen

8 Nullstellensuche

- Aufgabe: Gegeben $f(x)$, bestimme x_0, so daß:

$$f(x_0) = 0$$

- Geht i.a. nur iterativ.
- Wichtiger Begriff:

  Konvergenzordnung iterativer Algorithmen, auch wichtig für Kap. 9 Optimierung und Kap. 10 Nichtlineare Modellierung.

  Sei $\epsilon(i)$ die nach i Iterationen verbleibende Unsicherheit, so ist die Konvergenzordnung γ definiert durch:

$$\lim_{i \to \infty} \epsilon(i+1) = \text{const } \epsilon(i)^{\gamma}$$

Eindimensionaler Fall
Bisection

- Wähle zwei Punkte x_l und x_r, die die Nullstelle umschließen, i.e. $f(x_l)f(x_r) < 0$
- Bestimme $x_{Mitte} = \frac{x_r - x_l}{2}$
- Ersetze Ausgangspunkt mit dem selben Vorzeichen wie $f(x_{Mitte})$ durch x_{Mitte}.
- Iteriere dies, bis gewünschte Genauigkeit erreicht.
- Entwicklung der Unsicherheit:

$$\epsilon(i+1) = \frac{1}{2}\epsilon(i)$$

  somit lineare Konvergenzordnung γ.

- Anzahl n notwendiger Iterationen für gewünschte Genauigkeit ϵ bei Anfangsunsicherheit ϵ_0 :

$$n = \log_2 \frac{\epsilon_0}{\epsilon}$$

- Global konvergent, aber langsam.

Sekanten-Methode

- Setzt hinreichende Linearität voraus.

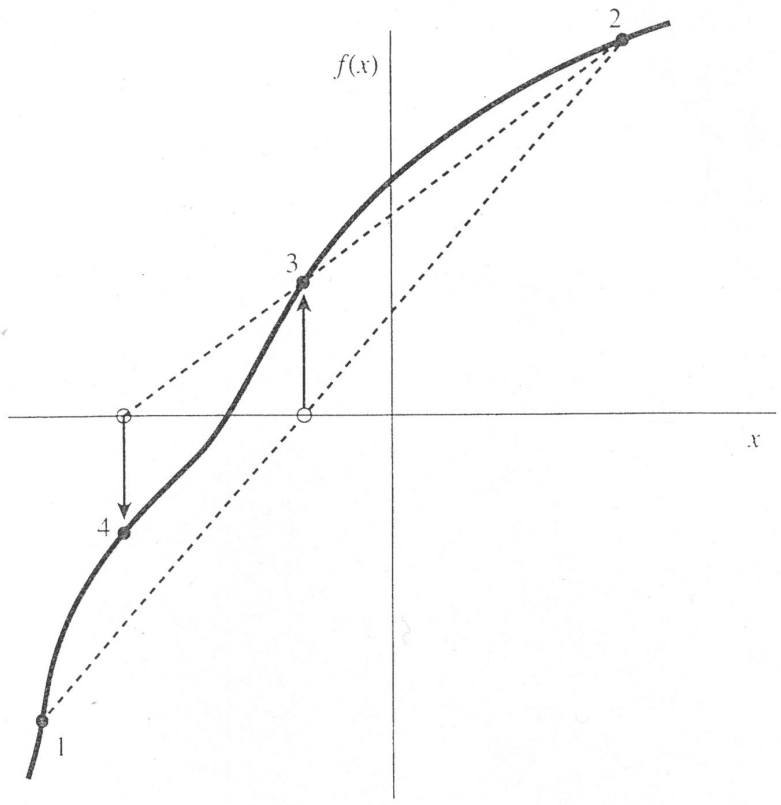


Abbildung 8.1: Sekanten-Methode

- Iteration:

$$x_{i+1} = \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})}$$

- Es gilt

$$\lim_{i\to\infty} \epsilon(i+1) = \text{const}\ \epsilon(i)^{\frac{\sqrt{5}+1}{2}}, \quad \frac{\sqrt{5}+1}{2} = 1.618... = \text{ Goldener Schnitt}$$

somit superlineare Konvergenz γ

- Nullstelle nicht notwendiger Weise umschlossen $\Longrightarrow$ Sekanten-Methode kann divergieren

Regula falsi

- Wie Sekanten-Methode, aber verwerfe x_l oder x_r je nachdem ob $f(x_l)f(x_{i+1}) > 0$ oder $f(x_r)f(x_{i+1}) > 0$

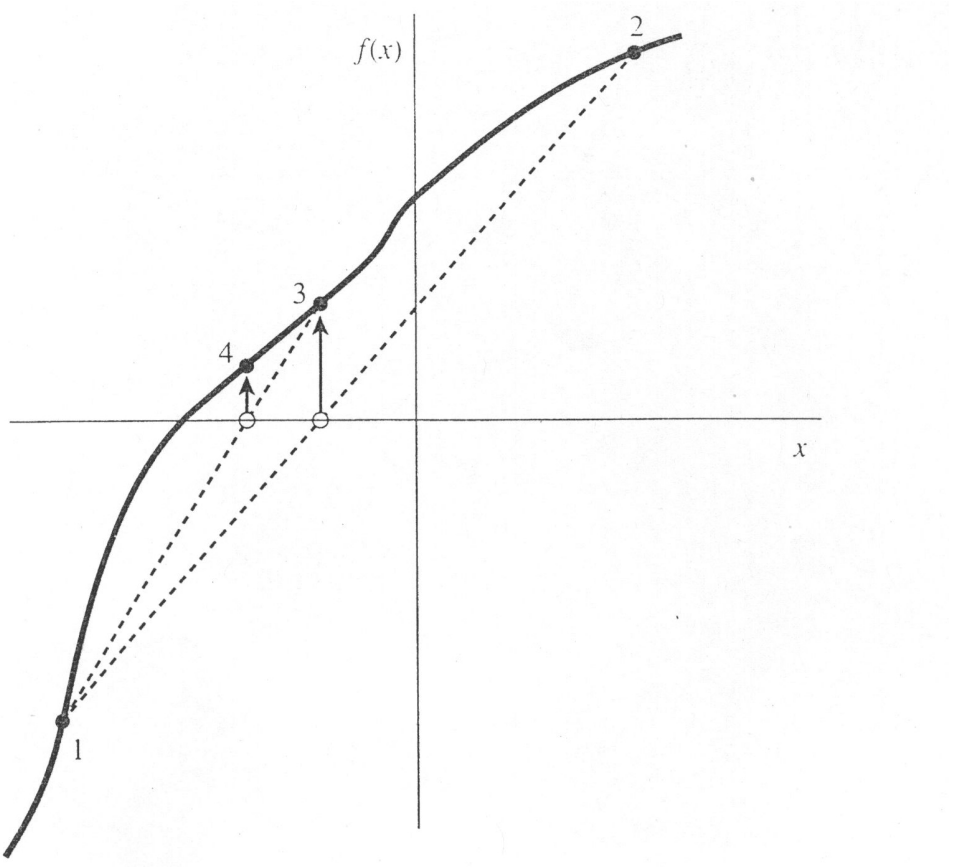


Abbildung 8.2: Regula-Falsi-Methode

- Konvergenzordnung $\gamma \geq 1$, i.a. langsamer als Sekanten-Methode, aber sicher

- Sekanten-Methode und Regula falsi können sehr langsam sein.

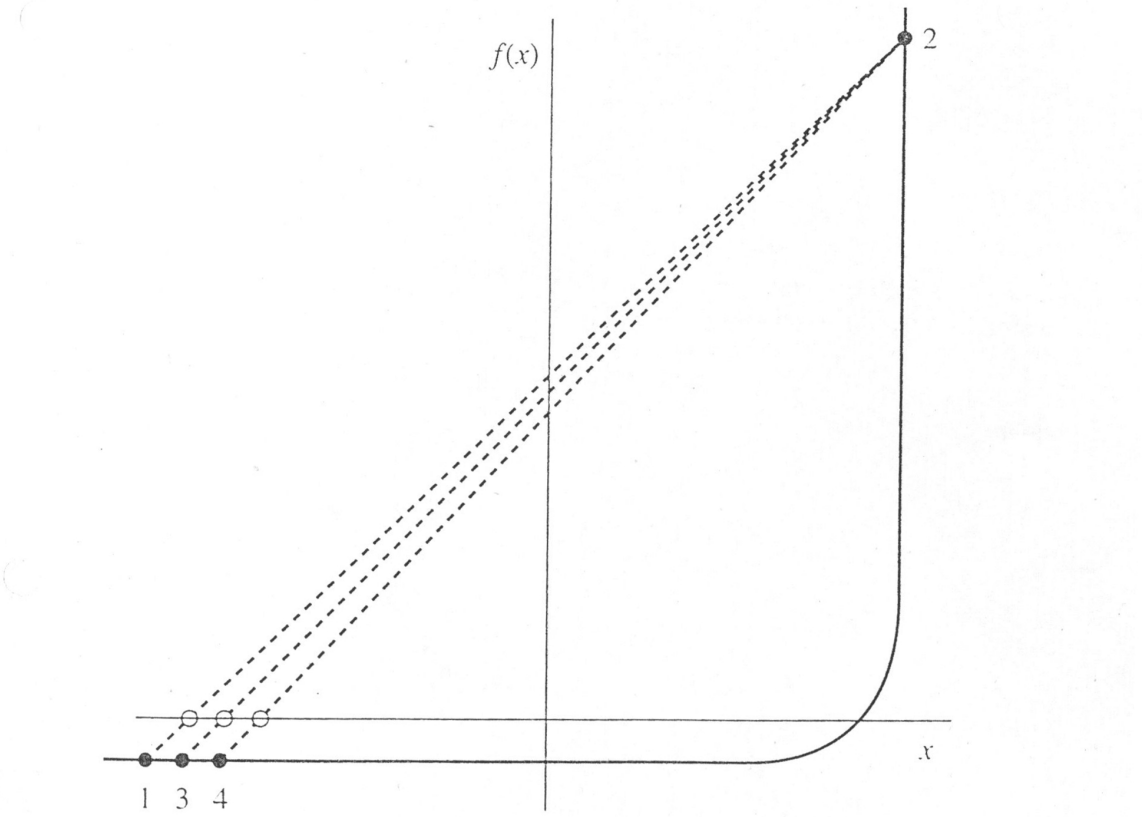


Abbildung 8.3: Beispiel bei dem Sekanten- und Regula-Falsi-Methode sehr viele Iterationen brauchen

Newton-Raphson

- Nutzt (und benötigt) 1. Ableitung
- Idee: Taylor-Entwicklung:

$$f(x_{i+1}) = f(x_i + \delta) \approx f(x_i) + f'(x_i)\delta + \frac{f''(x_i)}{2}\delta^2 + \dots$$

- Nah an der Nullstelle $f(x_i + \delta) = 0$, $\delta^2 \ll 1$, everything well behaved, folgt

$$\delta = -\frac{f(x_i)}{f'(x_i)} \implies x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

- Konvergenzordung bestimmen

$$\epsilon_{i+1} = \epsilon_i - \frac{f(x_i)}{f'(x_i)} \tag{5}$$

  - Taylorentwicklung für $f(x_i)$, $f'(x_i)$ um Nullstelle x_0 ergibt (alle Indices unterdrückt):

$$\begin{aligned} f(x+\epsilon) &= f(x) + \epsilon f'(x) + \epsilon^2 \frac{f''(x)}{2} + \ldots, \quad f(x) = 0 \\ f'(x+\epsilon) &= f'(x) + \epsilon f''(x) + \ldots \end{aligned}$$

  - Einsetzen in Gl. (5) ergibt:

$$\epsilon_{i+1} = \epsilon_i - \frac{\epsilon_i f'(x) + \epsilon_i^2 \frac{f''(x)}{2}}{f'(x) + \epsilon_i f''(x)}$$

    Erweitern:

$$\epsilon_{i+1} = \epsilon_i \frac{f'(x) + \epsilon_i f''(x)}{f'(x) + \epsilon_i f''(x)} - \frac{\epsilon_i f'(x) + \epsilon_i^2 \frac{f''(x)}{2}}{f'(x) + \epsilon_i f''(x)}$$

  - Mit $\epsilon_i f''(x) \ll f'(x)$ folgt:

$$\lim_{i \to \infty} \epsilon_{i+1} = \frac{f''(x)}{2f'(x)} \epsilon_i^2$$

    quadratische Konvergenzordnung

  - Aber nur lokal konvergent

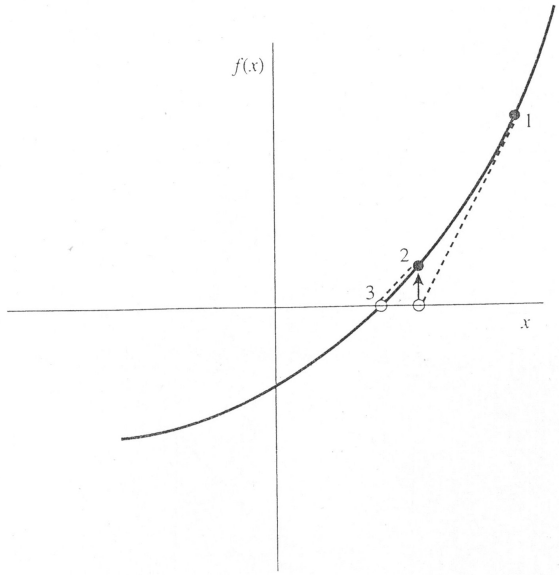


Abbildung 8.4: Newton-Raphson-Methode konvergiert

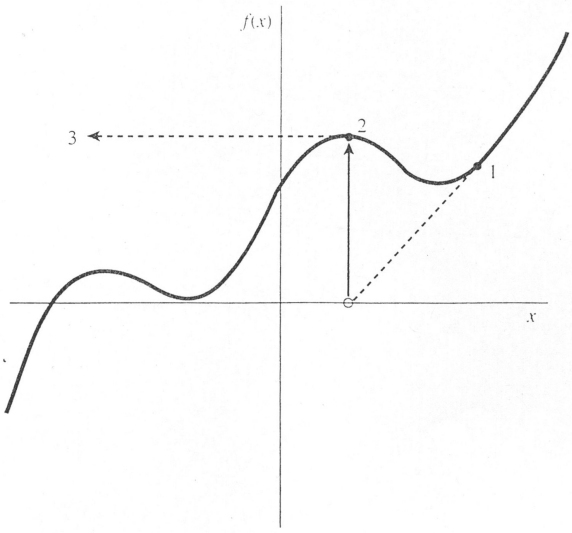


Abbildung 8.5: Newton-Raphson-Methode divergiert

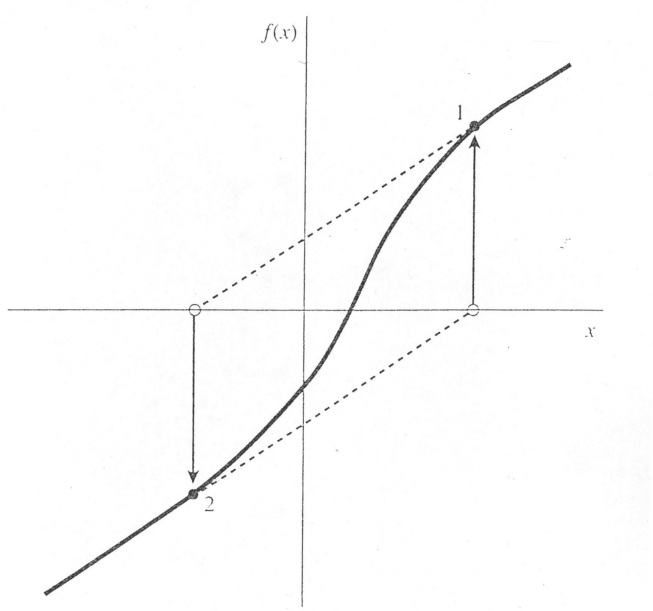


Abbildung 8.6: Newton-Raphson-Methode unglücklich gewählter Startpunkt

- Gut zum "Nachbrennen": Anfangs Bisection, später Newton-Raphson

Schmankerl, Chaos-Theorie revisited:
Finde Lösung von

$$z^3 - 1 = 0, \qquad z_0^1 = 1,\ z_0^{2,3} = \exp(\pm 2\pi i/3), \qquad z \in \mathbb{C}$$

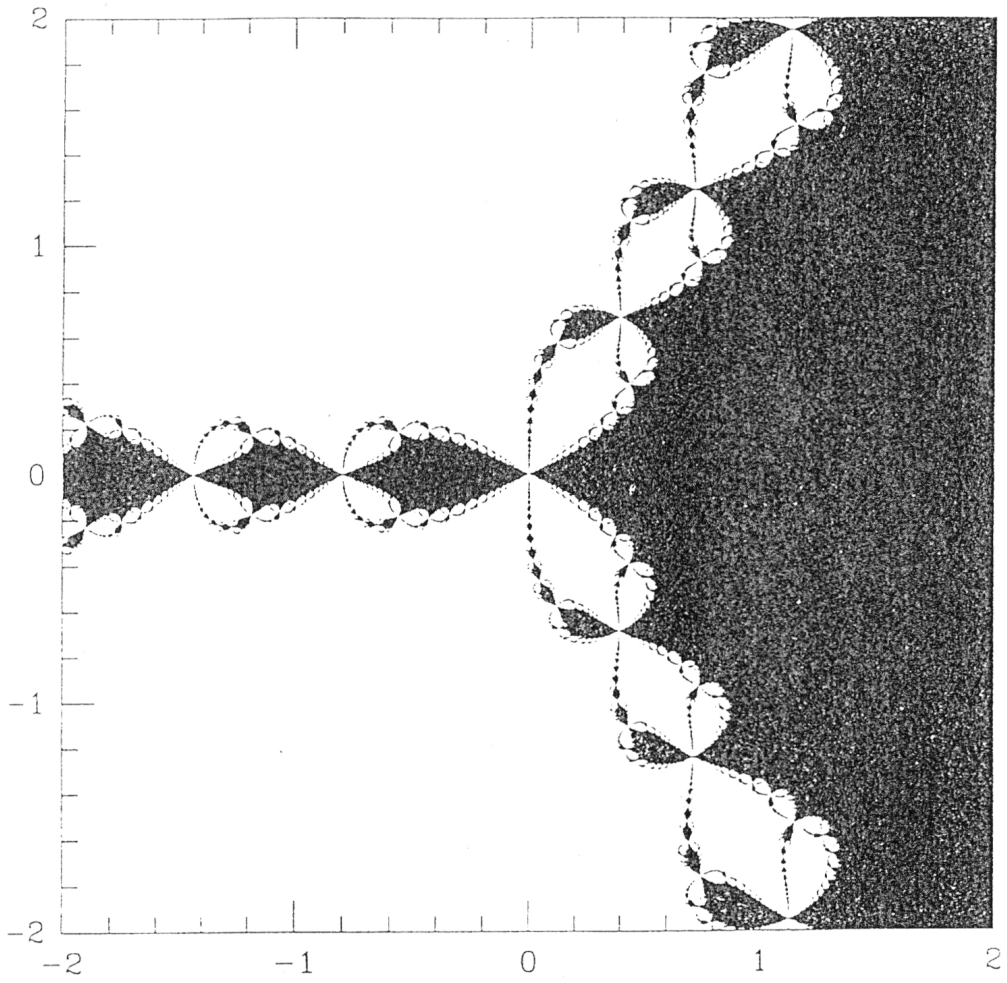


Abbildung 8.7: Fraktal: Im schwarzen Bereich konvergiert die Newton-Raphson-Methode auf $z = 1$.

Mehrdimensionaler Fall

$$\begin{aligned} f(x,y) &= 0 \\ g(x,y) &= 0 \end{aligned}$$

Haariges Problem, z.B. Anzahl der Lösungen a priori nicht klar, siehe Recipes Kap. 9.7

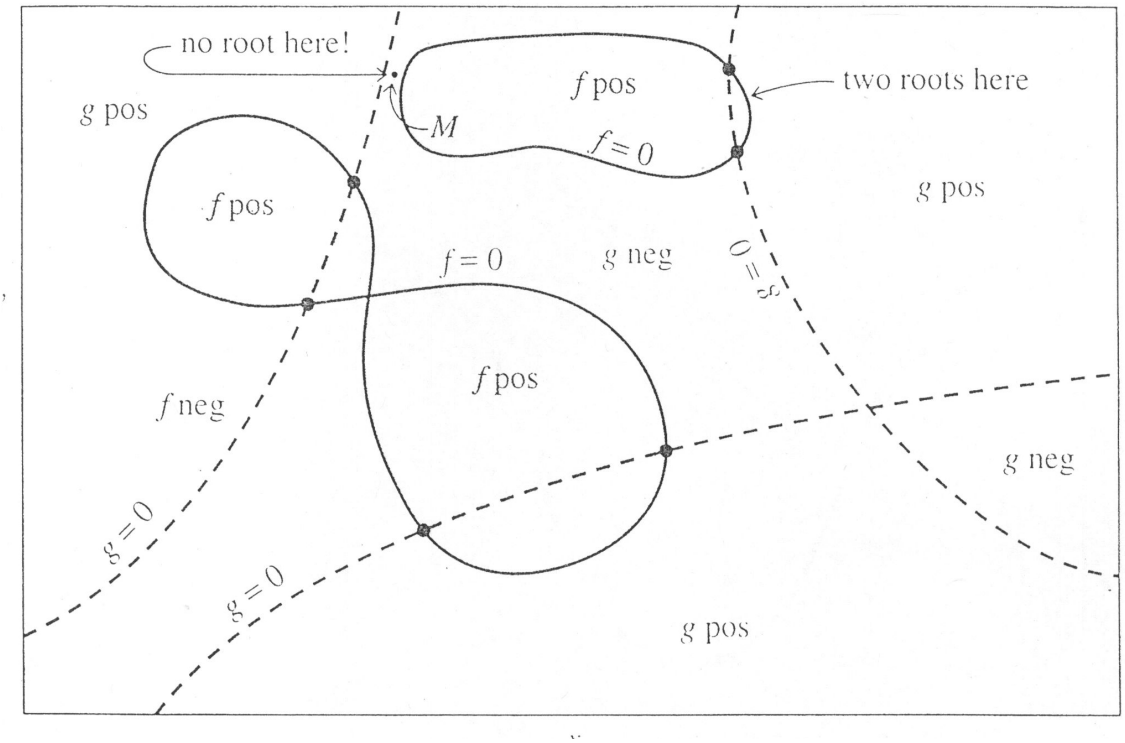


Abbildung 8.8: Lösung für zwei nichtlineare Gleichungen mit zwei Unbekanten

Übung:
Bestimmung der Quantile der Gaußverteilung

Lessons learned:

- Bei iterativen Algorithmen: Konvergenzordnung γ

$$\lim_{i \to \infty} \epsilon(i+1) = \text{const } \epsilon(i)^{\gamma}$$

- Bisection, Sekanten-Methode, Regula falsi, Newton-Raphson
- Trade-off: Konvergenzordnung vs. Konvergenzsicherheit

7.18

9 Optimierung

- Literatur: Recipes Kap. 10
- Aufgabe: Bestimme x, so daß:

$$f'(x) = 0, \quad f''(x) >< 0, \text{ je nach dem}$$

- Optimierung umfaßt Minimierung und Maximierung "one's f is the other's -f"
- Iterative Algorithmen

Unterschiedungen der Methoden:

- 1 D vs. N D
- Ableitungsinformation vorhanden oder nicht.
- Deterministische Verfahren: Konvergenz gegen lokales Optimum
- Stochastische Verfahren: Prinzipiell globale Konvergenz

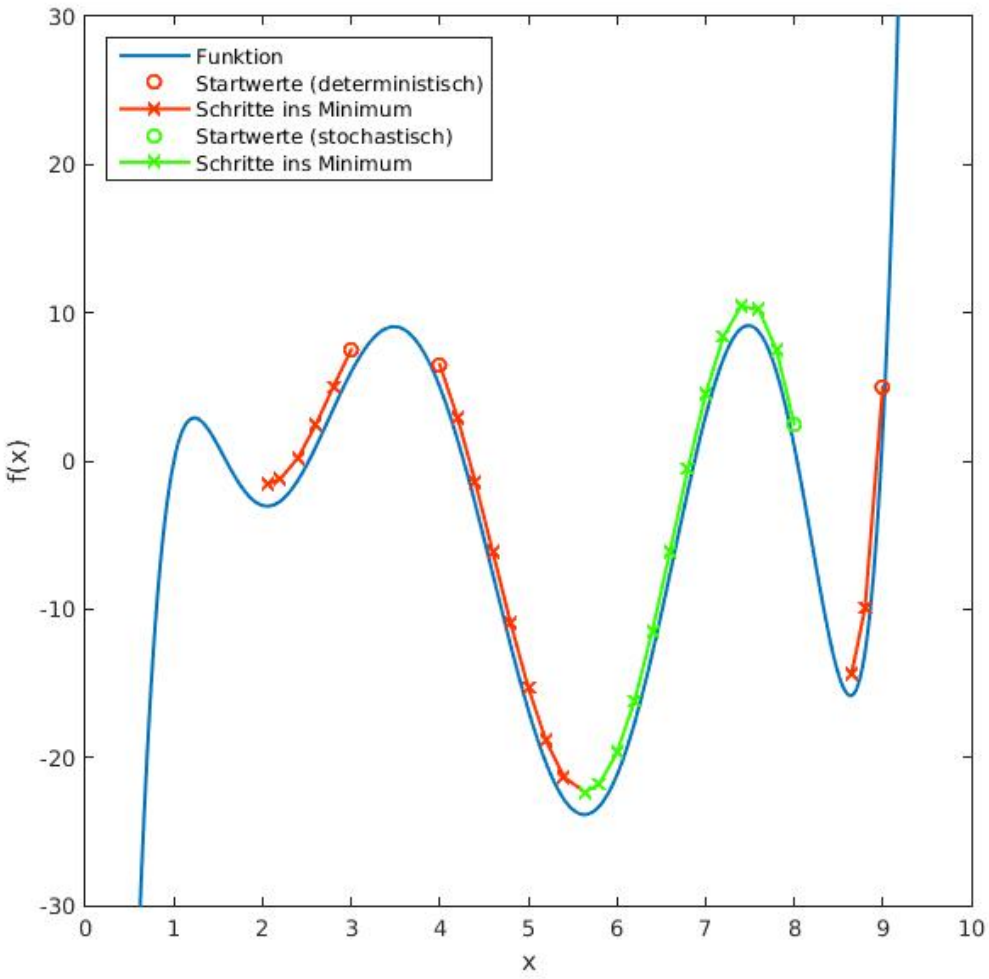


Abbildung 9.1: Unterschied zwischen deterministischen und stochastischen Verfahren

9.1 Eindimensionaler Fall

Betrachte jeweils Minimierung

9.1.1 Bracketing, Goldener Schnitt Suche

Analog zu Bisection in Kap. 8 Nullstellensuche
Beachte:

- Nullstellen-Bracketing braucht 2 Punkte
- Minima-Bracketing braucht 3 Punkte (a, b, c).

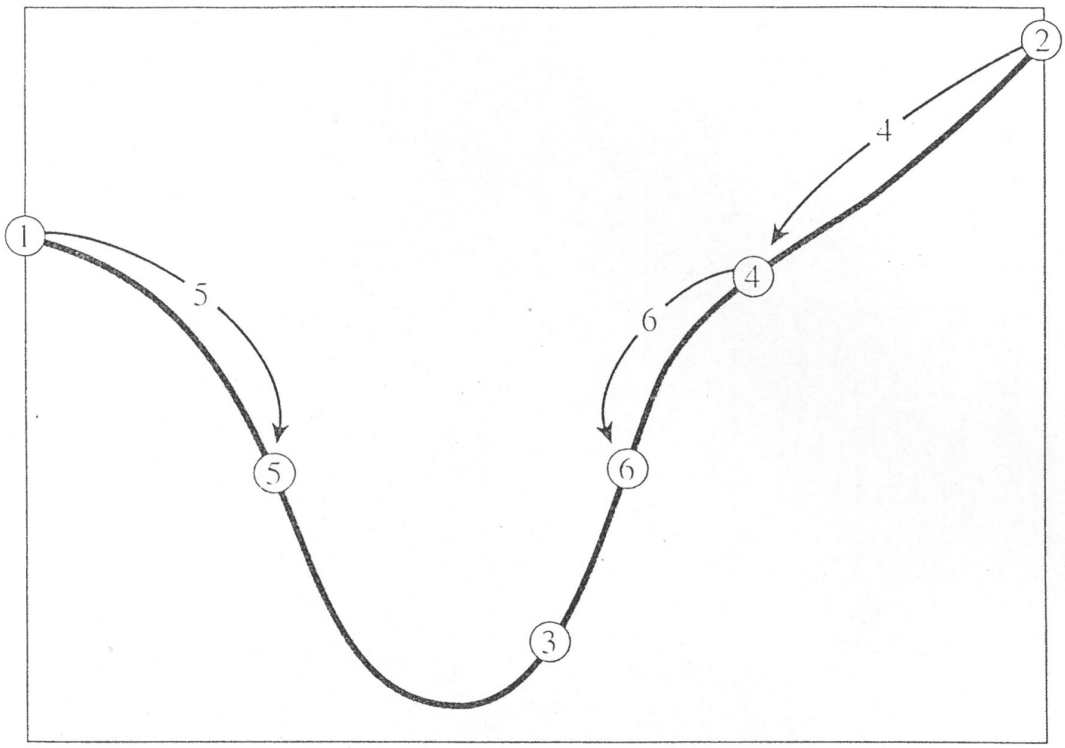


Abbildung 9.2: Minima-Bracketing

Wie neuen Zwischenpunkt x suchen, gegeben (a, b, c) ?

- Sei b ein Bruchteil w auf dem Weg von a nach c

$$w = \frac{b-a}{c-a}, \quad 1-w = \frac{c-b}{c-a}$$

- Neuer Punkt x sei hinter b bei zusätzlichem Bruchteil

$$z = \frac{x-b}{c-a}$$

Dann ist das nächste Bracketing Segment

  - entweder $w + z$
  - oder $1 - w$

relativ zum Bestehenden.

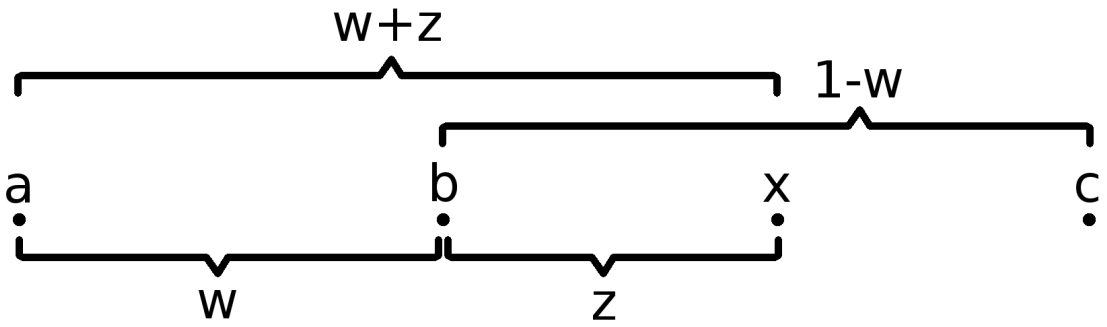


Abbildung 9.3: Skaleninvarianz des goldenen Schnitts

Zur Minimierung des schlimmsten Falles: Wähle z, so daß potentiell nächste Segmente gleichgroß:

$$w + z = 1 - w \quad \Longrightarrow \quad z = 1 - 2w \tag{6}$$

Per Konstruktion: x ist symmetrisch zu b im Ausgangsintervall $|b-a| = |x-c|$ $\Longrightarrow$ x liegt im längeren Segment

- Wie w wählen ?

  Nimm an, w war ebenso optimal, wie z sein soll

Skalenähnlichkeit: x gleicher Anteil an (b,c), wenn dies das längere Segment war, wie b war in (a,c)

$$\begin{aligned} \frac{x-b}{c-b} &= \frac{b-a}{c-a} \\ \frac{x-b}{c-b}\frac{c-a}{c-a} &= \frac{b-a}{c-a} \\ \frac{z}{1-w} &= w \end{aligned} \tag{7}$$

Gln. (6,7) zusammen:

$$\frac{1-2w}{1-w} = w$$

$$w^2 - 3w + 1 = 0, \quad \text{ergibt } w = \frac{3-\sqrt{5}}{2} \approx 0.38197$$

$$\frac{1-w}{w} = \text{Goldener Schnitt}$$

- Startet man mit beliebigen Punkten (a,b,c), konvergiert Verfahren zum Goldenen Schnitt
- Lineare Konvergenzordnung:

$$\epsilon(i+1) = 0.61803... \epsilon(i)$$

9.1.2 Parabolische Interpolation, Brent's Methode

Analog zu Regula falsi

- Regula falsi: Nah der Nullstelle ist lineare Näherung gut
- Parabolische Interpolation: Nah am Optimum ist quadratische Näherung gut.

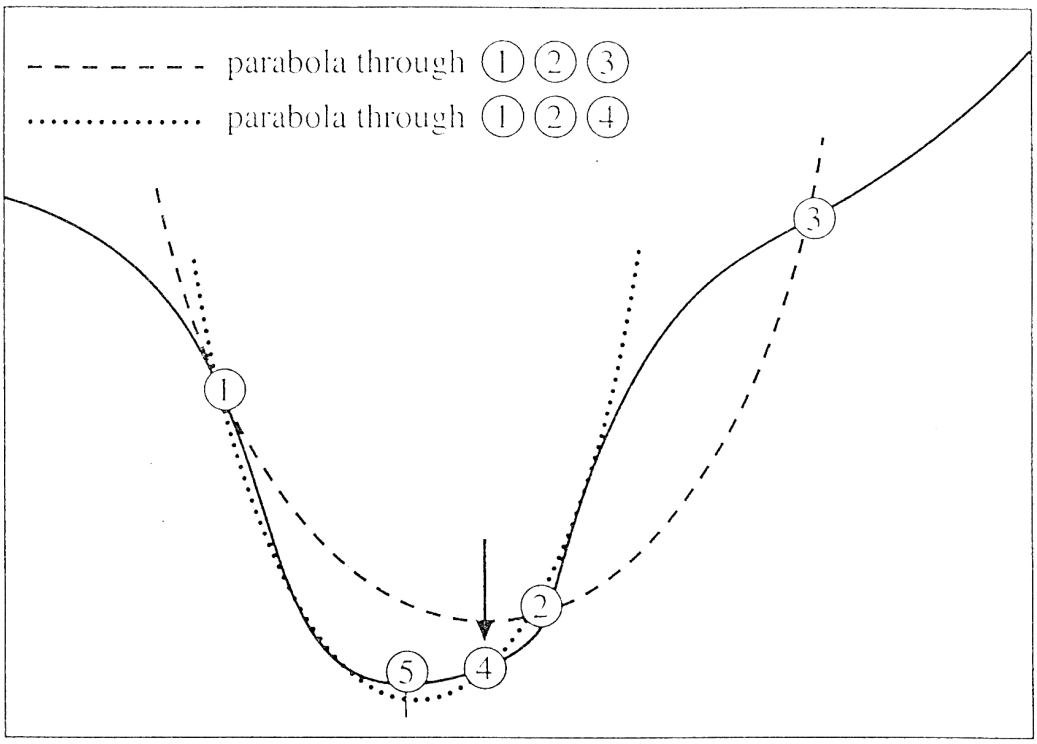


Abbildung 9.4: Konvergenz an ein Minimum durch parabolische Interpolation

Gegeben (a, b, c) und $f(a), f(b), f(c)$, neuer Punkt x durch:

$$x = b - \frac{(b-a)^2 [f(b) - f(c)] - (b-c)^2 [f(b) - f(a)]}{2(b-a) [f(b) - f(c)] - (b-c) [f(b) - f(a)]}$$

In 1D Ableitungsinformation i.d.R. nicht nötig.

9.2 N-dimensionaler Fall

9.2.1 Nur Funktionsauswertungen

Naivster Ansatz

1. Wähle Startpunkt

2. Laufe entlang einer Koordinatenachse, bis Minimum in dieser Richtung erreicht
3. Mache dies für alle Koordinatenachsen
4. go to 2.

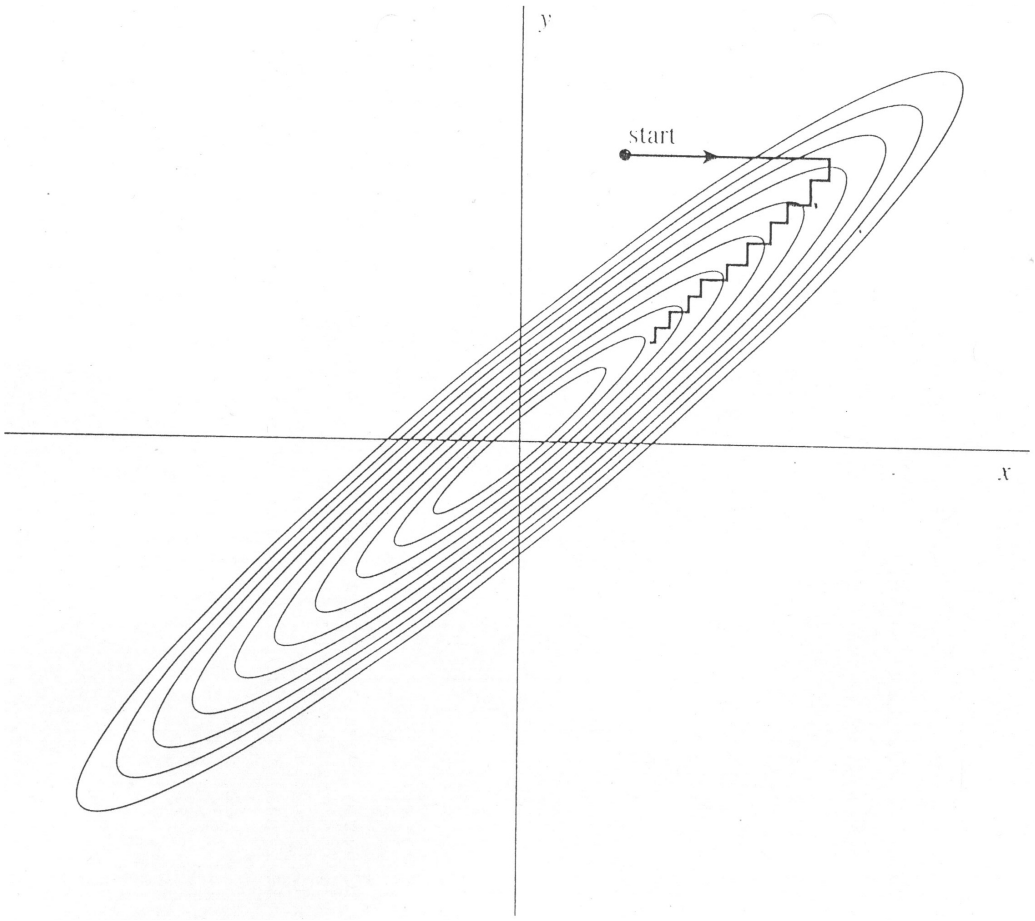


Abbildung 9.5: Sukzessive Minimierung entlang der Koordinatenachsen

Dies ist sehr ineffizient

Powell's Method
Basiert auf `linmin()` :

- Gegeben
  - zu minimierende Funktion $f(.)$
  - $\vec{P}$: Momentaner Punkt
  - $\vec{u}$: Suchrichtung
- Bracketiere Minimum in Richtung $\vec{P} + \mu\vec{u}$.
- Finde Skalar λ, so daß $f(\vec{P} + \lambda\vec{u})$ minimal.
  1D - Problem, siehe oben.
- Ersetze $\vec{P}$ durch $\vec{P} + \lambda\vec{u}$.

Idee:
Versuche suksessiv "gute" Abstiegsrichtungen $\vec{u}_i$, $i = 1, \ldots, N$ zu finden:

- Initialisiere: $\vec{u}_i = \vec{e}_i$, $i = 1, \ldots, N$
- Startposition: $\vec{P}_0$
- For $i = 1, \ldots, N$: $\vec{P}_i =$`linmin`$(\vec{P}_{i-1}, \vec{u}_i)$
- For $i = 1, \ldots, N-1$: Ersetzte $\vec{u}_i$ durch $\vec{u}_{i+1}$
- Setze $\vec{u}_N = \vec{P}_N - \vec{P}_0$, $\vec{P}_N - \vec{P}_0$: Mittlere Erfolgsrichtung
- $\vec{P}_0 =$`linmin`$(\vec{P}_N, \vec{u}_N)$
- Iteriere dies.

Konvergenzverhalten:

- Quadratische Näherung exakt: Verfahren nach N Iterationen im Optimum.
- Quadratische Näherung gut: Konvergenzordnung quadratisch.

7. Woche

9.2.2 Verwendung von Ableitungs-Information

Ableitung muß/sollte analytisch vorliegen. Approximationen durch z.B.

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + \Delta x_i) - f(x)}{\Delta x_i}$$

sind schwierig, wegen

- Auslöschung in $f(x + \Delta x_i) - f(x)$
- schwere Kontrollierbarkeit des "$\approx$"

Naivste Idee: Gradientenabstieg, Steepest Descent

- Startposition: $\vec{P}_0$
- Gehe von $\vec{P}_i$ nach $\vec{P}_{i+1}$ per Minimierung entlang der Richtung $-\nabla f(\vec{P}_i)$
- Iteriere zum Ziel

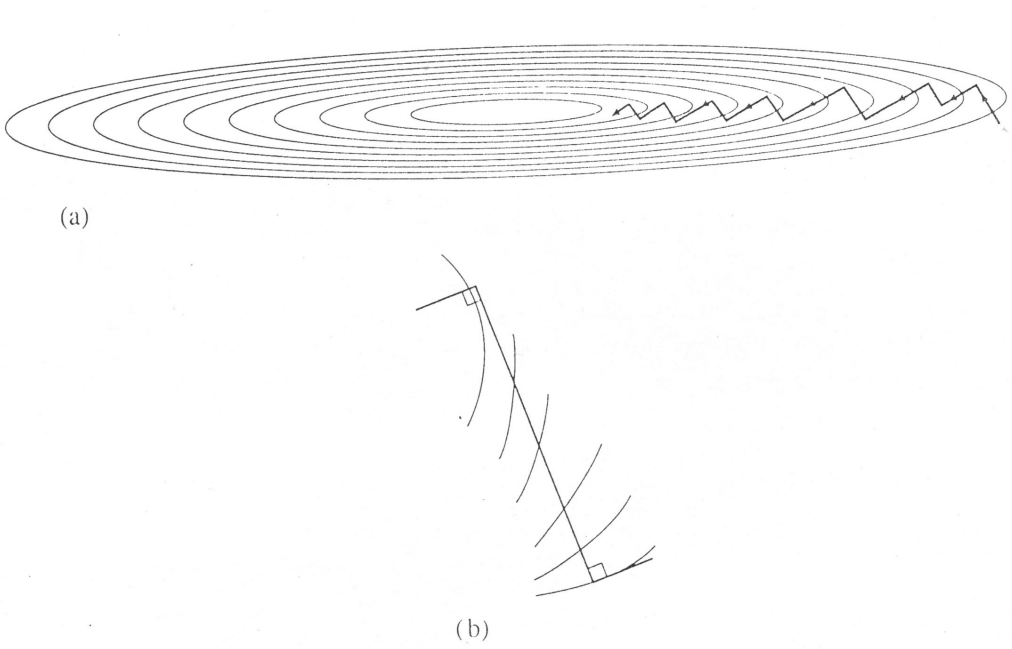


Abbildung 9.6: a) Steepest Descent-Methode in langem, engem Tal; b) Vergrößerung eines Schrittes

Do not do Steepest Descent

Grund:

- Keine Krümmungsinformation berücksichtigt
- Oder anders: Falsche Metrik
  Gradientenabstieg: Aufeinanderfolgende Suchrichtungen $\vec{u}_i$, $\vec{u}_{i+1}$ erfüllen:

$$\langle \vec{u}_{i+1}\, \vec{u}_i \rangle = 0 = \vec{u}_{i+1}^T\, 1\, \vec{u}_i$$

  Besser wäre :

$$0 = \langle \vec{u}_{i+1}\, A\vec{u}_i \rangle = \vec{u}_{i+1}^T\, A\, \vec{u}_i, \tag{8}$$

  mit

$$A = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \qquad \text{Hesse-Matrix.}$$

  Die Richtungen in Gl. (8) heißen dann konjugiert.

- Motivation
  - Wir seien bei P
    Taylorentwicklung:

$$\begin{aligned} f(x) &= f(P) + \nabla f(P)x + \frac{1}{2}x^T \frac{\partial^2 f(P)}{\partial x_i \partial x_j}\, x + \ldots \\ &\approx c + bx + \frac{1}{2}x^T A x \end{aligned}$$

    und damit

$$\nabla f(x) = b + Ax$$

  - Änderung von $\nabla f(x)$ bei Bewegung um δx:

$$\delta(\nabla f(x)) = A\, \delta x$$

– Hat man sich entlang Richtung u_i bis zum Minimum bewegt, sollte neue Richtung u_{i+1} maximal informativ sein:

$$0 =< u_{i+1}\delta(\nabla f(x)) >= u_{i+1}^T A u_i$$

- Powell's Method konstruiert konjugierte Richtungen
- Bemerkung:

  Wenn statt eindeutigem Minimum lange Wanne, dann A schlecht konditioniert, erinnere Kap. 7.2.3 Singulärwert-Zerlegung und Kap. 4.4 Nicht-Identifizierbarkeit.

Variable Metrik oder Quasi-Newton – Verfahren

- Taylorentwicklung um momentanen Punkt x_i:

$$f(x) = f(x_i) + (x - x_i)\nabla f(x_i) + \frac{1}{2}(x - x_i)A(x - x_i) + \ldots$$

  Ableiten:

$$\nabla f(x) = \nabla f(x_i) + A(x - x_i) \stackrel{!}{=} 0$$

  Gilt quadratische Näherung, direkt ins Ziel mit

$$x = x_i - A^{-1}\nabla f(x_i)$$

- Aber:
  - Berechnung von A^{-1} hat Aufwand $\mathcal{O}(N^3)$, erinnere Kap. 7
  - Weit weg vom Optimum kann A nicht positiv definit sein, Algorithmus divergiert
- Idee:

  Konstruiere während der Iteration positive definite Näherung der Hesse-Matrix nur aus ersten Ableitungen oder besser gleich das Inverse A^{-1}.

Das Procedere:

1. Wähle Startwert x_0
2. Wähle I_0, positiv definit, symmetrisch
3. Gehe $x_{i+1} = x_i - I_i \nabla f(x_i)$
4. Verwende die DFP oder BFGS updating formula ...

   RECIPES GL. 10.7.8 und 10.7.9

- Gehe zu 3.

Eigenschaften:

- Verwendet nur Gradienten-Information
- Es gilt:
$$\lim_{i \to \infty} I_i = A^{-1}$$
- Komplexität $\mathcal{O}(N^2)$
- Gehört zu den 5 wichtigsten Routinen, wo gibt.

Conjugierte Gradienten – Verfahren

- Bildet iterativ konjugierte Richtungen definiert in Gl. (8).
- Konstruiert nicht die (Inverse der) Hessematrix, was $\mathcal{O}(N^2)$ teuer ist.
- Methode der Wahl für hohe Dimensionen ($N > 100$)

Quasi-Newton und Conjugierte Gradienten Verfahren konvergieren quadratisch, wenn sie nah am Minimum sind.

Allgemein:

- Wann die Iteration beenden ?

  Abbruckkriterien:

i. Relative Änderung des Funktionswertes:

$$\left| \frac{f(x_{i+1}) - f(x_i)}{f(x_i)} \right| < \epsilon_1$$

ii. Relative Änderung von x :

$$\left| \frac{x_{i+1} - x_i}{x_i} \right| < \epsilon_2$$

Empfehlung: ii., wegen "langer Wannen".

- Bei allem bisherigen nur Konvergenz gegen lokales Optimum gesichert.
  Only cure: Verschiedene Startwerte testen.

9.2.3 Simulated annealing

Weitere Literatur:

- S.E. Koonin: *Computational Physics* Kap. 8.3 [36]
- Metropolis et al. 1953 [46]

Alle bisherigen Verfahren:

- deterministisch
- Zielort bestimmt durch Anfangswert, Weg zum Ziel durch Algorithmus
- Nur Konvergenz zu lokalem Optimum

Probabilistische/stochastische Optimierer, hier Minimierer
Namensgebung:

- Kühlt ("annealed") man Flüssigkeit schnell ab, erreicht der entstehende Kristall nicht das absolute Energieminimum, sondern nur lokales
- Es gibt viele lokale Minima, Konflikt: Nah-/Fernordnung.
- Kühlt man langsam, wird das Minmum mit hoher Wahrscheinlichkeit oder in guter Näherung angenommen.

- Grund: Bei langsamen Abkühlen können mit thermischer Energie (Boltzmann-Verteilung) Energiebarrieren übersprungen werden.

Idee für numerische Minimierer: Laufe auch mal bergauf.

Procedere:

- Wähle Startwert x_0
- Produziere zufällige Veränderung ϵ_i: $x_{i+1} = x_i + \epsilon_i$
- Falls $f(x_{i+1}) < f(x_i)$, akzeptiere x_{i+1}.
- Falls $f(x_{i+1}) > f(x_i)$, akzeptiere x_{i+1} mit Wahrscheinlichkeit

$$\text{prob } = \exp[-(f(x_{i+1}) - f(x_i))/T(i)]$$

  Erinnere: Boltzmann-Verteilung

- Wähle $T(i)$ anfangs groß, lass es mit den Iterationen gegen 0 gehen

Probleme:

- Wahl des Annealing-Schemas $T(i)$, z.B. $T(i) \propto 1/i$
- Wahl der Größe der Änderungen ϵ_i, z.B. $< \epsilon_i^2 > \propto 1/i$
- Beides braucht Vorwissen über das Problem: No free lunch - Theorem
- Das Vorwissen entspricht Gradienten- und Krümmungsinformation
  Zeichnungen dazu
- Spielt daher in "ernsten" Anwendungen keine große Rolle

Aber: Kann nicht-polynomial (NP) harte Probleme in sehr guter Näherung lösen

Beispiel: Travelling Salesman Problem $\mathcal{O}(N!)$

- N Städte mit Koordinaten (x_j, y_j)
- Suche Tour durch alle Städte, die minimale Länge hat
- Konfiguration $conf$ ist Permutation der Zahlen $j = 1, \ldots, N$

- Zu minimierendes Funktional: Weglänge

$$f(conf) = \sum_{j=1}^{N} \sqrt{(x_j - x_{j+1})^2 + (y_j - y_{j+1})^2}, \quad N+1 = 1$$

- "Veränderungen" ϵ: Lokale Änderungen der Permutation.

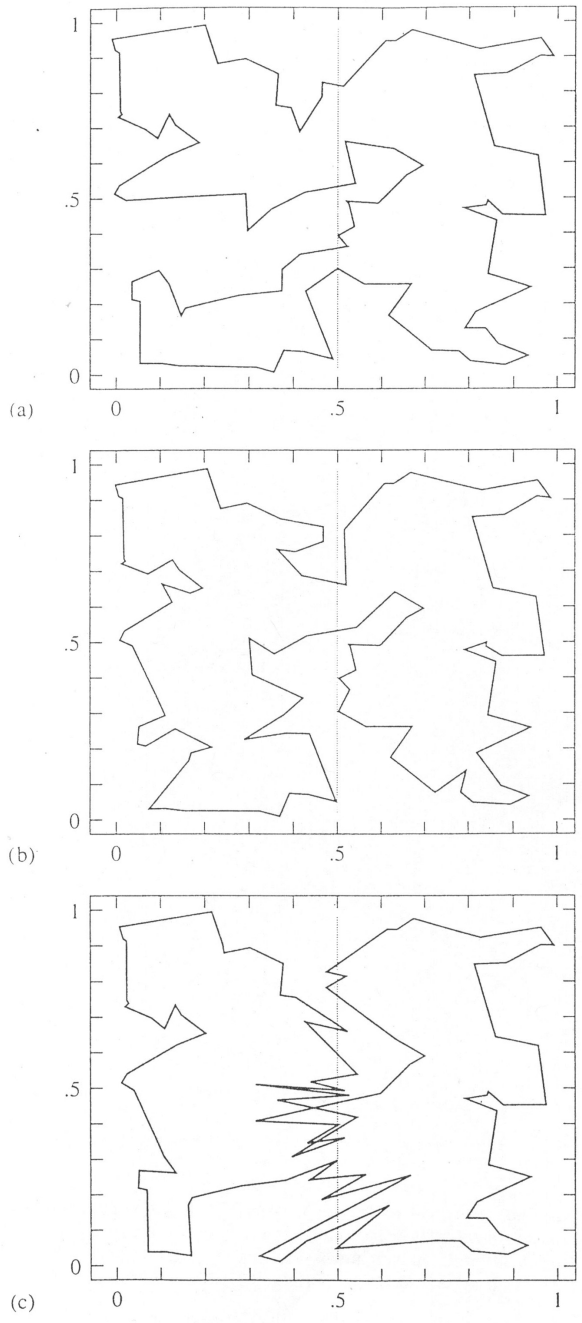


Abbildung 9.7: Travelling-Salesman a) ohne Nebenbedingung, b) mit möglichst wenigen Flusüberquerungen, c) mit möglichst vielen Flussüberquerungen

- 1. Travelling Salesman der Geschichte: Odysseus, 13 Stationen: 6.2×10^9 Möglichkeiten.

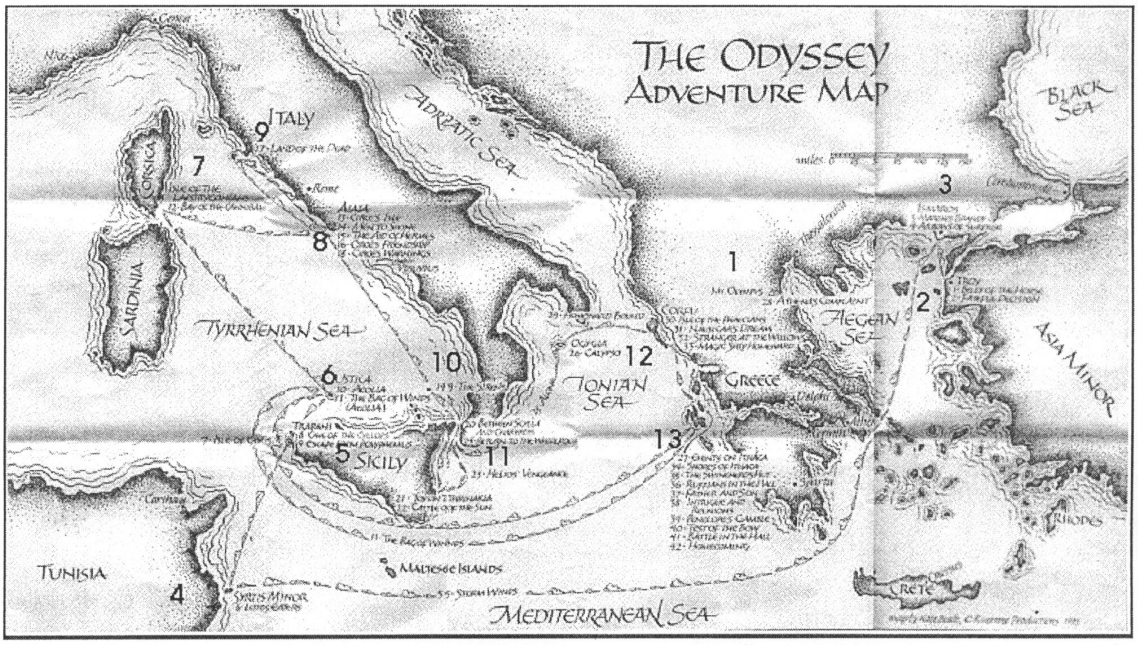


Abbildung 9.8: Odysseus' Reiseweg: His way home: 9000 km, kürzester 6000 km

- Flexibilität der Methode:
  Erweiterung des Funktionals um Strafterm, erinnere prior, Kap. 4.3
  Beispiel:
  - Nehme an: Fluss teilt das Gebiet.

  (i) Salesman hat Angst vor Flussüberquerung

  (ii) Salesman ist Schmuggler, will Fluss maximal oft überqueren

$$\mu_j = -1 \text{ für links vom Fluss}, \quad \mu_j = +1 \text{ für rechts vom Fluss}$$

$$f(conf) = \sum_{j=1}^{N} \sqrt{(x_j - x_{j+1})^2 + (y_j - y_{j+1})^2} + \lambda(\mu_j - \mu_{j+1})^2$$

  Für

(i) $\lambda > 0$

(ii) $\lambda < 0$

siehe Abbildung 9.7

Andere stochastische Optimierer

- Evolutionäre Algorithmen
- Genetische Algorithmen
- Particle Swarm Algorithmen

Übung:
Maximum-Entropie-Verteilung für diskrete Verteilungen

Lessons learned:

- Ein-dimensional: Goldener Schnitt
- Mehr-dimensional: Gradientenabstieg naheliegend, aber nicht gut
- Besser: Krümmungsinformation mitnehmen: Quasi-Newton
- Deterministische Verfahren: Lokal konvergent
- Stochastische Verfahren: Im Prinzip globale Konvergenz

8.18

10 Nichtlineare Modellierung

Literatur:

- *Numerical Recipes*, Kap. 15
- G.A.F. Seber and C.J. Wild. *Nonlinear Regression* [62] Der Klassiker
- G.J.S. Ross. *Nonlinear Estimation* [57] Klasse Buch

Motivation:

- Kap. 9 Optimierung: Allgemeine Suche von Optima
- Hier: Minimierung spezieller Funktionale

Ist

- $y(x) = y(x, a)$ eine mit a parametrisierte Funktion, z.B. first principle Gleichung mit freien Parametern
- y_i, $i = 1, \ldots, N$: N Messungen der Funktion $y(x, a)$ an Punkten x_i
- Messungen i. a. mit Fehler ϵ_i behaftet: $y_i = y(x_i, a) + \epsilon_i$, z.B. $\epsilon_i \sim N(0, \sigma_i^2)$
- Ziel: a basierend auf N Messungen (y_i, x_i) schätzen
- Anders formuliert: Zusammenhang (y_i, x_i) durch $y(x, a)$ modellieren
- Parameterschätzung durch Minimierung von:

$$\chi^2(a) = \sum_{i=1}^{N} \frac{(y_i - y(x_i, a))^2}{\sigma_i^2}$$

  Erinnere: Gewichteter Kleinste Quadrate Schätzer ist MLE für Gauß-verteilte Fehler

- Ist Modell richtig, k Anzahl der Parameter, so gilt:

$$\chi^2(\hat{a}) \sim \chi^2_{N-k}$$

Dies erlaubt goodness-of-fit Test:

- H_0: Das Modell stimmt.
- H_1: Das Modell stimmt nicht.

Erinnere :

$$\begin{aligned} \langle \chi^2_r \rangle &= r \\ Var(\chi^2_r) &= 2r\,, \end{aligned}$$

Unter H_0: $\chi^2(\hat{a})$ für 99% Vertrauensintervall im Bereich

$$[(N-k) - 3\sqrt{2(N-k)}, (N-k) + 3\sqrt{2(N-k)}]$$

- Ist $\chi^2(\hat{a})$ grösser, der natürliche Fall:
  - Model falsch ?

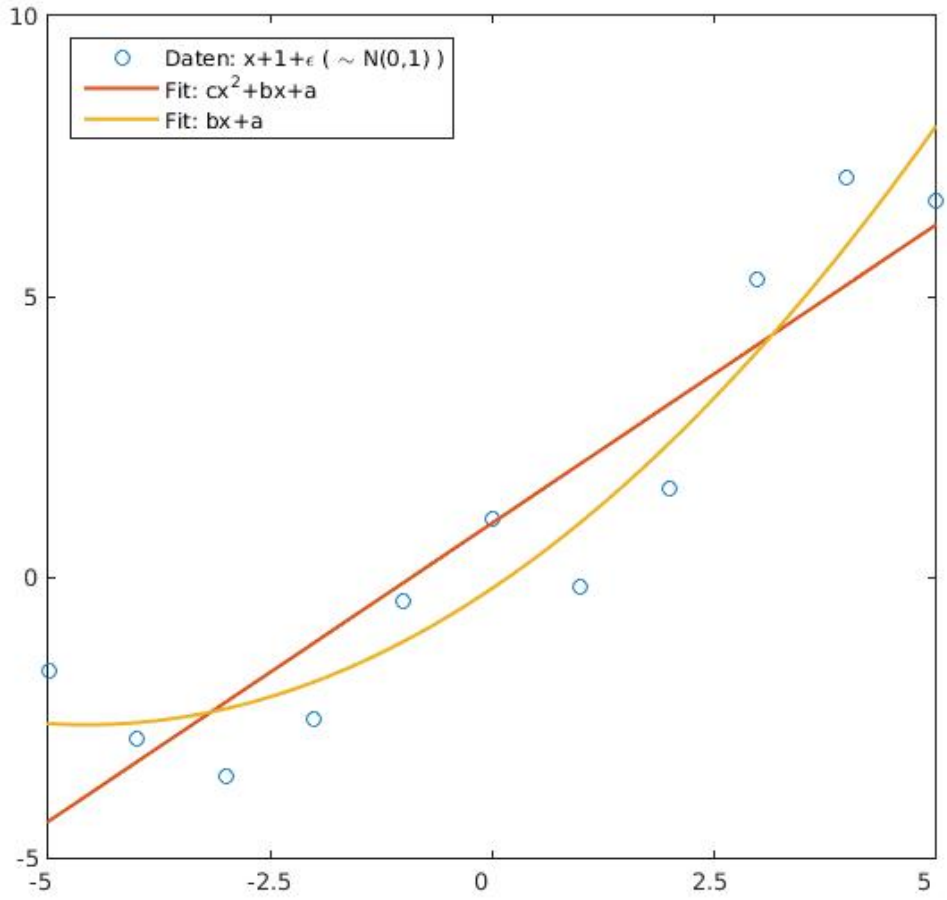


Abbildung 10.1

- σ_i fälschlich zu klein ?
- Fehler nicht Gaußsch ?

- Ist $\chi^2(\hat{a})$ kleiner, sollte eigentlich nicht vorkommen:
  - σ_i fälschlich zu groß ?
  - Fehler nicht Gaußsch ?

10.1 Lineare Regression

Annahme: Gaußscher Fehler:

$$y(x) = y(x, a, b) = a + bx + \epsilon, \quad \epsilon \sim N(0, \sigma_i^2)$$

Alles geht analytisch:

$$\chi^2(a, b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} \stackrel{!}{=} 0 \tag{9}$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i \left(y_i - a - bx_i \right)}{\sigma_i^2} \stackrel{!}{=} 0 \tag{10}$$

Mit

$$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \quad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \quad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \quad S_{xx} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}$$

folgt aus Gln. (9, 10)

$$\begin{aligned} aS + bS_x &= S_y \\ aS_x + bS_{xx} &= S_{xy} \end{aligned}$$

Mit Determinante Δ:

$$\Delta = SS_{xx} - S_x^2$$

folgt:

$$\begin{aligned} \hat{a} &= \frac{S_{xx} S_y - S_x S_{xy}}{\Delta} \\ \hat{b} &= \frac{S S_{xy} - S_x S_y}{\Delta} \end{aligned}$$

Gaußsche Fehlerfortpflanzung: Cramér-Rao Schranke

$$\sigma_{\hat{a}}^2 = \sum_{i=1}^{N} \left(\frac{\partial a}{\partial y_i} \right)^2 \sigma_i^2$$

Eingesetzt ergibt:

$$\begin{aligned} \sigma_a^2 &= S_{xx}/\Delta \\ \sigma_b^2 &= S/\Delta \end{aligned}$$

Aber: Es ist 2-D Schätzproblem:

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \sim N \left(\begin{pmatrix} a \\ b \end{pmatrix}, \Sigma \right)$$

mit

$$\Sigma = \begin{pmatrix} \sigma_a^2 & \sigma_{ab}^2 \\ \sigma_{ab}^2 & \sigma_b^2 \end{pmatrix}$$

Kovarianz σ_{ab}^2

$$\sigma_{ab}^2 = \frac{-S_x}{\Delta}$$

σ_{ab}^2, resp. Konditionszahl von Σ sagt, ob Schätzer abhängig

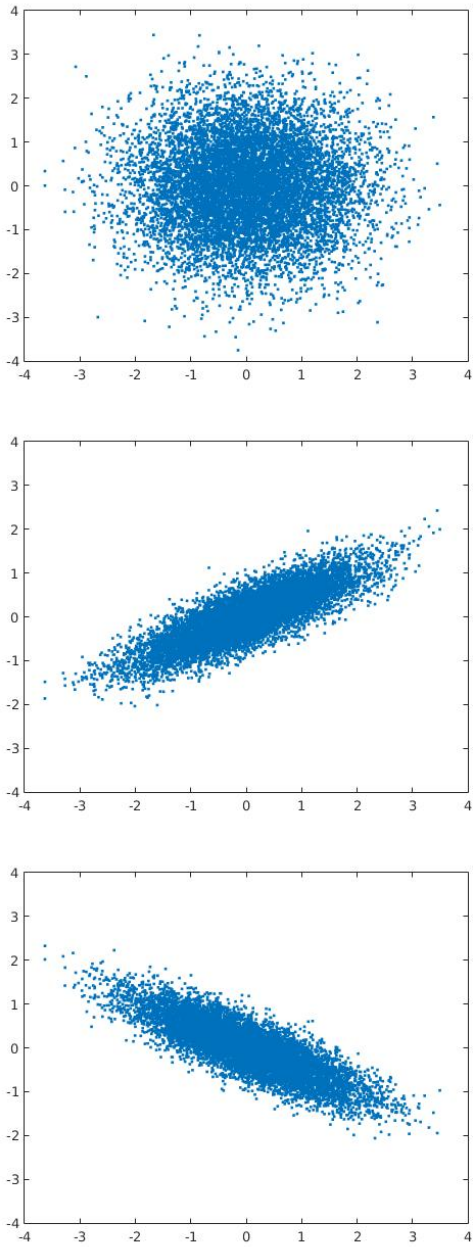


Abbildung 10.2: *2D*-Gaußverteilte-Zufallszahlen mit $C_1 = (0.71\ 0; 0\ 0.70)$, $C_2 = (0.78\ 0.39; 0.39\ 0.28)$, $C_3 = (0.79, -0.39;\ -0.39, 0.28)$

Kondensiert in: Korrelationen $r_{ab} \in [-1, 1]$ zwischen den Schätzfehlern

$$r_{ab} = \frac{-S_x}{\sqrt{SS_{xx}}}$$

Bemerkung:
Häufig: Summen über sehr viele Summanden, kann zu Rundungsfehlern führen
Lösung: Kahan-Summation [33]

Robuste lineare Regression

Zusätzliche Literatur:

- P. Huber: Robust Statistics [26]
- H. Rieder: Robust Statistics, Data Analysis, and Computer Intensive Methods [55]

Ist die Fehlerverteilung

- nicht gaußsch, ist χ^2 fitting nicht mehr MLE.
- symmetrisch, ist χ^2 fitting bias-frei, aber hat eine grössere Varianz als MLE.

  Erinnere Effizienz eines Schätzers:

  $$Eff(\hat{\Theta}_{\chi^2}) = \frac{Var(\hat{\Theta}_{MLE})}{Var(\hat{\Theta}_{\chi^2})} \leq 1$$

  siehe Übungsaufgabe.
- asymmetrisch, kann ein bias entstehen.
- langsamer abfallend als Gaußsch, fat-tailed, z.B. Cauchy, so "hängt" sich χ^2 fitting an diesen "Ausreißern" auf.
- Lösung: Robuste Verfahren. Hängen sich nicht an Ausreißern auf

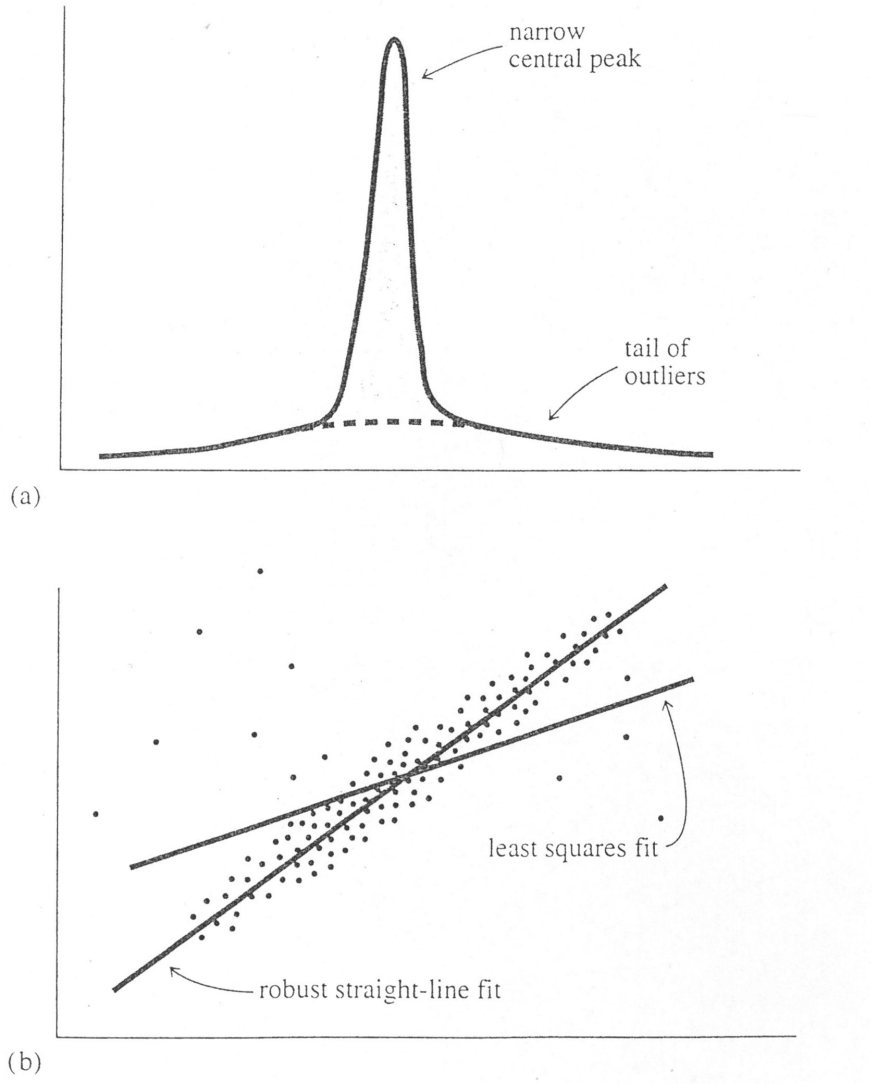


Abbildung 10.3: Beispiele für robuste statistische Methoden: (a) Eindimensionale Verteilung mit Ausreißern. (b) Zweidimensionale Verteilung an eine Gerade gefittet

8. Woche

Erinnere:

- Im Gauß-Fall war die Likelihood:

$$L(a) \propto \prod_{i=1}^{N} \exp\left(-\frac{(y_i - y(x_i, a))^2}{2\sigma_i^2}\right)$$

und log Likelihood

$$\mathcal{L}(a) \propto \sum_{i=1}^{N} \frac{(y_i - y(x_i, a))^2}{2\sigma_i^2}$$

Parameterschätzer durch Nullsetzen der Ableitung:

$$\sum_{i=1}^{N} \left(\frac{y_i - y(x_i, a)}{\sigma_i^2} \right) \left(\frac{\partial y(x_i, a)}{\partial a} \right) \stackrel{!}{=} 0 \tag{11}$$

Diskussion der Faktoren

- 1. Faktor: Einfluss der Daten
- 2. Faktor: Modell-spezifisch

- Im allgemeinen:

$$L(a) \propto \prod_{i=1}^{N} \exp(-\rho(y_i, y(x,a)), \quad \rho(.) = -\log p(.)$$

In der Regel

$$\rho(y_i, y(x_i, a)) = \rho\left(\frac{y_i - y(x_i, a)}{\sigma_i} \right) = \rho(z), \quad z = \left(\frac{y_i - y(x_i, a)}{\sigma_i} \right)$$

- Definiere:

$$\psi(z) = \frac{d\rho(z)}{dz}$$

$\psi(.)$ heißt Influence Function

- Ergibt in Verallgemeinerung von Gl. (11) MLE-Bedingung

$$\sum_{i=1}^{N} \frac{1}{\sigma_i} \psi \left(\frac{y_i - y(x_i, a)}{\sigma_i} \right) \left(\frac{\partial y(x_i, a)}{\partial a} \right) \stackrel{!}{=} 0 \tag{12}$$

- Spezialfall Gauß für:

$$\rho(z) = \frac{1}{2} z^2, \quad \psi(z) = z$$

- Ergo: Einfluß der Daten nimmt linear mit Abweichung zu.
- Darum nicht robust.

Andere Verteilungen:

- Doppel-Exponential Laplace Verteilung

$$p(y_i - y(x_i, a)) \sim \exp\left(-\left|\frac{y_i - y(x_i, a)}{\sigma_i}\right|\right)$$

$$\rho(z) = |z|, \quad \psi(z) = sign(z)$$

  Ergo: Einfluß der Daten auf MLE-Schätzer hängt nur vom Vorzeichen ab.
  Darum deutlich robuster !

- Beispiel: Cauchy Verteilung

$$p(y_i - y(x_i, a)) \sim \frac{1}{1 + \frac{1}{2}\left(\frac{y_i - y(x_i, a)}{\sigma_i}\right)^2}$$

$$\rho(z) = \log\left(1 + \frac{1}{2}z^2\right), \quad \psi(z) = \frac{z}{1 + \frac{1}{2}z^2}$$

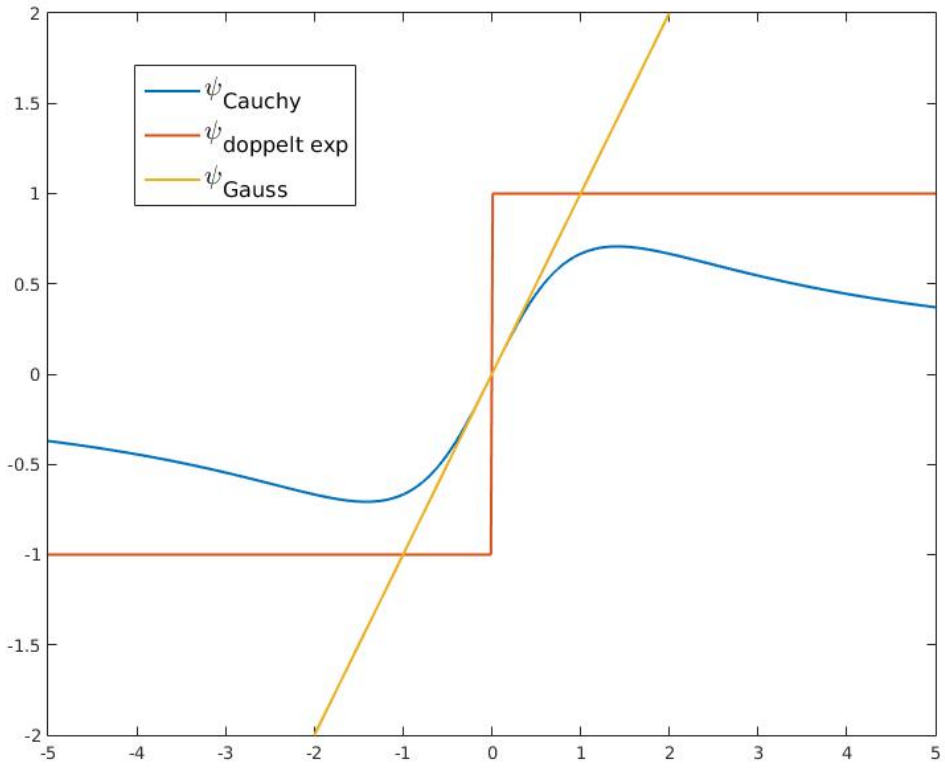


Abbildung 10.4: Influence Functions die sich aus unterschiedlichen Verteilungen ergeben

Ergo: Einfluß der Daten auf MLE-Schätzer nimmt mit größerem Abstand ab. Darum richtig robust !

- Drehe den Spieß um: Abnahme des Einflußes mit dem Abstand kann man nutzen zur Konstruktion von Influence functions für Fehlermodell = "well-behaved" "+" "Ausreißer"

  Andrews's sine

$$\psi(z) = \begin{cases} \sin(z/c) & |z| < c\pi \\ 0 & |z| > c\pi \end{cases}$$

$c = 2.1$

Tukey's biweight

$$\psi(z) = \begin{cases} z(1 - z^2/c^2)^2 & |z| < c \\ 0 & |z| > c \end{cases}$$

$c = 6.0$

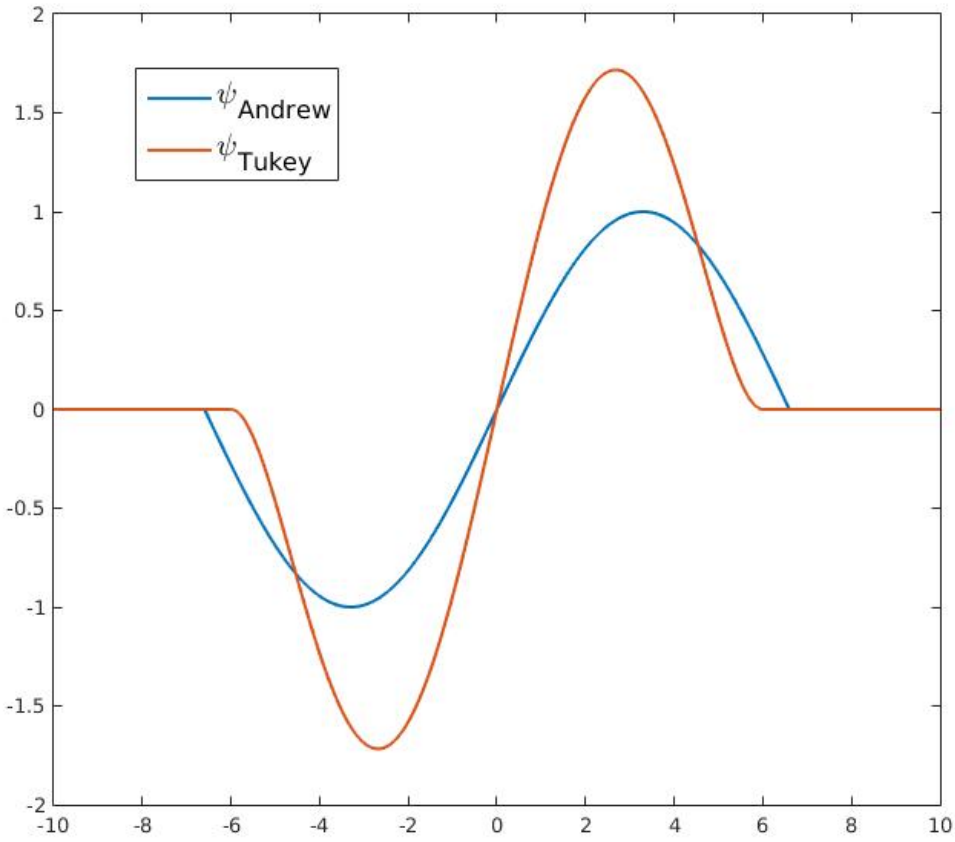


Abbildung 10.5: konstruierte Influence Functions

Beispiel für konkrete Berechnung:

- Lineare Regression mit double exponential - Fehlern

$$y(x, a, b) = a + bx + \epsilon, \quad p(\epsilon) = \frac{1}{2}\, e^{-|\epsilon|}$$

- Statt χ^2, ist die log-likelihood:

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - a - bx_i|$$

- Mentale Nebenrechnung:

  Definition Median:

  - Gegeben N Zahlen $\{z_i\}$.
  - Sortiere sie.
  - If N ungerade: $\text{med}\{z_i\} = z_M = z_{(N+1)/2}$
  - If N gerade: $\text{med}\{z_i\} = z_M = 0.5(z_{N/2+1} + z_{N/2})$

  Median z_M minimiert :

$$\sum_{i=1}^{N} |z_i - z_M|$$

  Beweis:

$$\frac{\partial}{\partial z_M} \sum_{i=1}^{N} |z_i - z_M| = -\sum_{i=1}^{N} \text{sign}(z_i - z_M) = 0$$

Damit:

- Iteratives Verfahren
- Wähle Startschätzungen $(\hat{a}_0, \hat{b}_0)$, z.B. aus Kleinste Quadrate Schätzer.
- Für gegebenes $\hat{b}_j$

$$\hat{a}_{j+1} = \text{med}\ \{y_i - \hat{b}_j x_i\}$$

dann analog Gl. (12), für gegebenes $\hat{a}_{j+1}$, folgt $\hat{b}_{j+1}$ aus:

$$0 = \sum_{i=1}^{N} x_i \, \text{sign}(y_i - \hat{a}_{j+1} - \hat{b}_{j+1} x_i)$$

Nullstellensuche.

Wird iterativ mit Bisection, siehe Kap. 8 Nullstellensuche, gelöst.

- Iteriere bis zur gewünschten Genauigkeit

Der Schnack "Seitdem am CERN robuste Statisitk verwendet wird, wurde kein neues Teilchen mehr gefunden" wurde jüngst widerlegt

Übung :
Suboptimalität des LS Schätzers bei nicht Gaußverteilten Daten

10.2 Nichtlineare Regression

Einfachste Fortsetzung von oben:

$$y(x, a) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_M x^M$$

oder genereller:

$$y(x, a) = \sum_{k=1}^{M} a_k X_k(x)$$

$X_k(x)$ Basisfunktionen, z.B. $\sin(\omega_k x)$, ω_k fix, denke an Fouriertransformation

Modell ist

- linear in den Parametern,
- hat aber nichtlineare Basisfunktionen.

Nun:

$$\chi^2(a) = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2$$

Definiere:

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i}, \quad b_i = \frac{y_i}{\sigma_i}$$

- A heißt Design-Matrix, ist $(N \times M)$,
- Sie legt den Versuch fest: Welche Basisfunktion wird wo gemessen.
- Erwähne Optimal Design, Experimental Design [52].

Maximum Likelihood Schätzer:

- Minimumsbedingung für χ^2 :

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(x_i) \stackrel{!}{=} 0 \tag{13}$$

- Mit

$$\alpha_{kj} = \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2}, \text{ oder } \alpha = A^T A$$

  α ist $(M \times M)$ Matrix

  und

$$\beta_k = \sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \text{ oder } \beta = A^T b$$

  und vertauschen der Summen in Gl. (13) folgt:

$$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k \tag{14}$$

- Die Gleichungen (13), resp. (14) heissen Normalen-Gleichungen

  und erinnern in der Form

$$(A^T A)\, a = A^T b$$

  an Kap. 7 Pseudo- oder Moore-Penrose - Inverse für dieses überbestimmte Gleichungssystem.

- Dies liefert den Punktschätzer.

Konfidenzintervalle für die Parameter:

- Definiere:

$$C = \alpha^{-1}$$

- Betrachte:

$$a_j = \sum_{k=1}^{M} \alpha_{jk}^{-1} \beta_k = \sum_{k=1}^{M} C_{jk} \left[\sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \right]$$

- Erinnere Fehlerfortpflanzung

$$\sigma^2(a_j) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2$$

  Mit

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=1}^{M} C_{jk} X_k(x_i) / \sigma_i^2$$

  folgt:

$$\sigma^2(a_j) = \sum_{k=1}^{M} \sum_{l=1}^{M} C_{jk} C_{jl} \left(\sum_{i=1}^{N} \frac{X_k(x_i) X_l(x_i)}{\sigma_i^2} \right)$$

- Der Term (.) ist aber das grade $\alpha = C^{-1}$, somit:

$$\sigma^2(a_j) = C_{jj}$$

  Entsprechend ergibt C_{jk} die Kovarianz zwischen den Schätzfehlern von a_j und a_k.

- Beachte:

  α, und damit C, ist unabhängig von y_i.

Damit: Optimales Design

- Da $\alpha = A^T A$, legt das Design die Fehler fest.
- Optimales Design: Betrachte lineare Regression auf Intervall [-1,1], man darf vier Mal messen.

  Wo soll man messen, damit Fehler der geschätzten Parameter möglichst klein ?
- Es gibt verschiedene Optimalitätskriterien: A bis D-optimal, ..., je nachdem ob Spur, Determinante oder ähnliches der Kovarianzmatrix klein werden soll.

Nichtlineare Regression und SVD
Normaler Weise:

- (Viel) mehr Daten als Parameter.
- Das System Gl. (14) sollte eigentlich gut gelöst werden können.

Aber:
Wenn Basisfunktionen nicht hinreichend unabhängig
$\Longrightarrow$
Problem schlecht konditioniert, egal wie viele Daten.

- Betrachte Monome 1, x, x^2, x^3, ... mit x gleichverteilt auf Intervall [0,1] als Basisfunktionen.
- Dann gilt für A:
$$A_{lm} = \sum_{i=1}^{N} x_i^l x_i^m \propto \frac{1}{l+m+1}$$
- Erinnere Hilbertmatrix, Übung Kap. 7 Lösung linearer Gleichungssysteme
- Bei bekannter Dichte $p(x)$ kann man bezüglich dieser Dichte orthogonale Polynome verwenden, und das Procedere wird stabil, weil A diagonal
- Beispiel $p(x) \sim$ Gleichverteilung [-1,1]: Legendre-Polynome
- Es gibt rekursive Konstruktionsvorschrift für Polynome, die auf empirisch gegebenen Daten orthogonal sind [16].
- Empfehlung: Verwende SVD, um Schlecht-Konditioniertheit zu checken, und gegebenenfalls zu behandeln. Die SVD generiert grade die empirisch orthogonalen Polynome.

10.3 Nichtlineare Modellierung

Reminder

- Lineare Regression: linear in Parametern und unabhängiger Variable x
- Nichtlineare Regression: linear in Parametern, nichtlinear in x

Nun: Nichtlinear auch in Parametern, z.B.:

$$y = e^{-\gamma x} \quad \text{oder} \quad y = x^b$$

- Iterative Verfahren, ähnlich zu Kap. 9 Optimierung.
- Erinnerung:
  Nah am Optimum gilt in guter Näherung die quadratische Näherung, und Newton-Schritt

$$a_{i+1} = a_i - A^{-1} \nabla f(a_i) \tag{15}$$

  führt zum Ziel.

- In Kap. 9 Optimierung: A^{-1} nicht bekannt/teuer zu berechnen
  - Quasi-Newton - Verfahren sammelt Information über A^{-1} während Iteration
  - Konjugierte Gradienten nähert $< \delta a_{i+1} \, A \, \delta a_i >$.

10.3.1 Levenberg-Marquardt Algorithmus

Hier weiß man mehr:

- Funktional:

$$f(a) = \chi^2(a) = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i, a)}{\sigma_i} \right]^2$$

- Gradient:

$$\frac{\partial \chi^2(a)}{\partial a_k} = -2 \sum_{i=1}^{N} \frac{(y_i - y(x_i, a))}{\sigma_i^2} \frac{\partial y(x_i, a)}{\partial a_k}, \quad k = 1, 2, \ldots, M$$

- Hesse-Matrix:

$$\frac{\partial^2\chi^2(a)}{\partial a_k \partial a_l} = 2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i,a)}{\partial a_k}\frac{\partial y(x_i,a)}{\partial a_l} - (y_i - y(x_i,a))\frac{\partial^2 y(x_i,a)}{\partial a_k \partial a_l}\right]$$

- Konvention:

$$\beta_k = -\frac{1}{2}\frac{\partial\chi^2(a)}{\partial a_k}, \quad \alpha_{kl} = \frac{1}{2}\frac{\partial^2\chi^2(a)}{\partial a_k \partial a_l}$$

- Wenn der Fit gut ist, gilt für 2. Term der Hesse-Matrix

$$\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(y_i - y(x_i,a))\frac{\partial^2 y(x_i,a)}{\partial a_k \partial a_l} \approx 0,$$

  da die Fehler $\epsilon_i = (y_i - y(x_i,a))$ unkorreliert sind.
  Also, definiere:

$$\alpha_{kl} := \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i,a)}{\partial a_k}\frac{\partial y(x_i,a)}{\partial a_l}\right]$$

- Damit und mit $\delta a_l = (a_{i+1} - a_i)_l$ wird Gl. (15) zu

$$\sum_{l=1}^{M}\alpha_{kl}\delta a_l = \beta_k \tag{16}$$

- Beachte : Steepest Descent lautet:

$$\delta a_l = \text{const } \beta_l \tag{17}$$

- Idee Levenberg-Marquardt Algorithmus:
  - Weit weg von Minimum mag Newton-Schritt Gl. (16) schlecht sein.
  - Mache Gradienten-Schritt Gl. (17). Wie "const" wählen ?
  - $\chi^2(a)$ dimensionslos, Dimension $[\beta_l]$ = Dimension $[1/\delta a_l]$, betrachte Gl. (16) $\Longrightarrow$
    $1/\alpha_{ll}$ ist Skalierungskanditat.

– Zur Sicherheit, damit Schritt nicht zu groß wird, wähle $\lambda \gg 1$ und setze:

$$\delta a_l = \frac{1}{\lambda \alpha_{ll}} \beta_l \quad \text{oder} \quad \lambda \alpha_{ll} \delta a_l = \beta_l \tag{18}$$

- Kombiniere Gradienten-Schritt Gl. (18) und Newton-Schritt Gl. (16) durch

$$\begin{aligned} \alpha'_{jj} &= \alpha_{jj}\,(1+\lambda) \\ \alpha'_{jk} &= \alpha_{jk}, \quad \text{für } j \neq k \end{aligned}$$

ergibt:

$$\sum_{l=1}^{M} \alpha'_{kl} \delta a_l = \beta_k \tag{19}$$

Bedeutung:

– Ist λ groß $\Longrightarrow$ α'_{kl} diagonal-dominant $\Longrightarrow$ kleiner Gradienten-Schritt
– Geht $\lambda \to 0$, Hesse-Schritt

Procedere:

1. Wähle Startschätzung für a, Berechne $\chi^2(a)$
2. Wähle ein kleines λ: $\lambda = 0.001$. Drückt Hoffnung aus
3. Löse Gl. (19) und berechne $\chi^2(a + \delta a)$
4. Wenn $\chi^2(a + \delta a) \geq \chi^2(a)$, verwerfe δa, wähle $\lambda = 10\lambda$, go to 3
5. Wenn $\chi^2(a + \delta a) < \chi^2(a)$, akzeptiere δa, wähle $\lambda = 0.1\lambda$, go to 3.

Interpretation:
Wenn Newton-Schritt

- gut, dann mehr davon,
- schlecht, dann mit Gradienten-Schritt auf Sicht fahren.

Bemerkungen:

- Gehört zu den 5 wichtigsten Routinen, wo gibt.

- Beachte:
  - Gleichung (19) kann schlecht-konditioniert sein
  - Wieder Mal die Wanne
  - Löse sie mit SVD.
- Abbruchkriterium:
  - Wenn nur noch kleine Änderungen in χ^2, Problem "Wanne"
  - Besser: Wenn $\lambda > 10^5$, entspricht: Keine Änderung in a mehr.

9.18

Nach Konvergenz:

- Asymptotische Kovarianzmatriz der Fehler in den geschätzten Parametern

$$C = \alpha^{-1} = \left\{ \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i, a)}{\partial a_k} \frac{\partial y(x_i, a)}{\partial a_l} \right] \right\}^{-1} \tag{20}$$

- Alternative zu Levenberg-Marquardt: Trust-Region Approach

9. Woche

Übung:
Nichtlineare Modellierung und Modelltest

10.3.2 Monte Carlo Vertrauensintervalle

- Erinnere Kap. 2.4. Die Standardabweichung der Parameterschätzer gibt - i.a. nur asymptotisch - Konfidenzintervall für die wahren Parameter, d.h. in

$$[\hat{a} - 1.96\,\sigma(\hat{a}), \hat{a} + 1.96\,\sigma(\hat{a})]$$

  liegt mit 95 % Wahrscheinlichkeit der wahre Wert.

- Die Kovarianzmatrix in Gl. (20) für die Fehler der geschätzten Parameter gilt nur asymptotisch
- Eine Alternative: Profile Likelihood, siehe Kap. 4.4
- Was man am liebsten hätte:

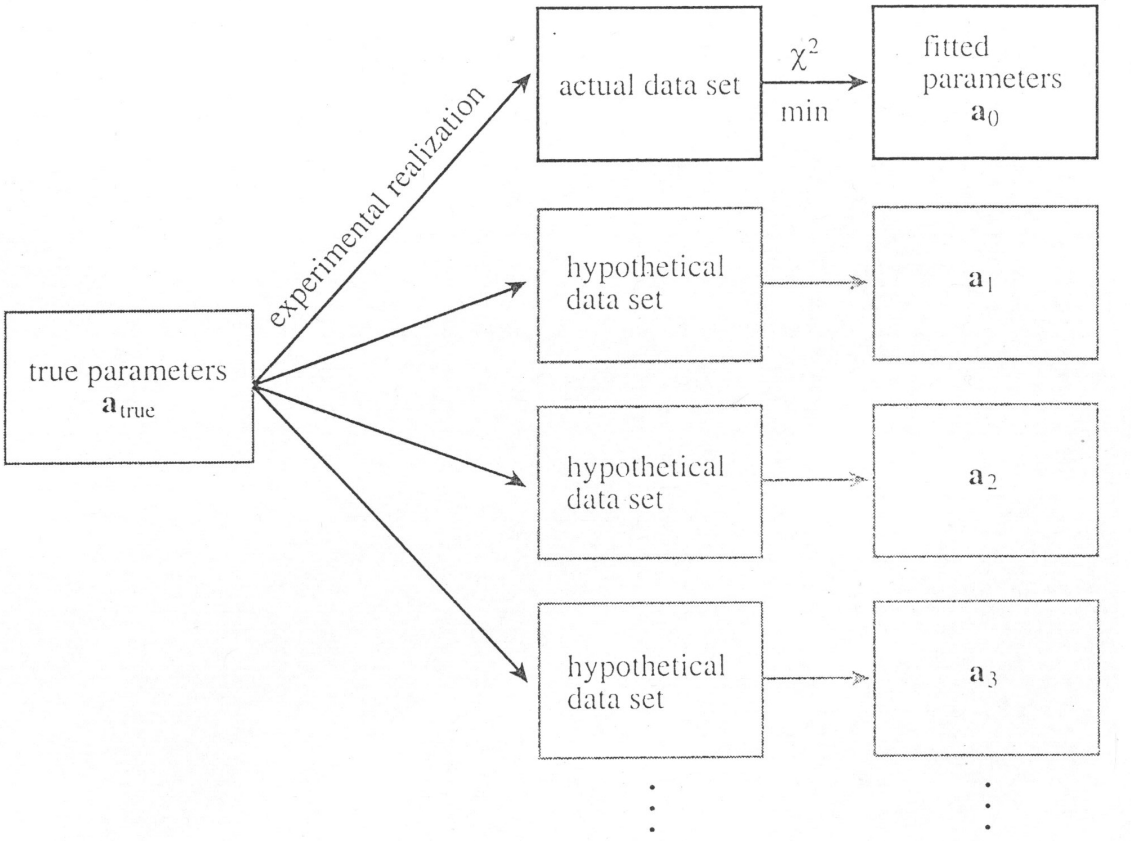


Abbildung 10.6: Ein statistisches Universum von Datensätzen für ein zugrundeliegendes Modell

- Gäbe die komplette Verteilung der geschätzten Parameter.
- Hat man aber nicht (und Datensplitten hilft nix).
- Man braucht Aussage über den wahren Wert, basierend auf einem (endlichen) Datensatz.

Asymptotische Konfidenzintervalle, Kap. 4.1

- Asymptotisch gilt unter milden Bedingungen

$$\sqrt{N}(\hat{a} - a) \sim N(0, \Sigma)$$

mit

$$\Sigma^{-1} = -\frac{1}{N}\frac{\partial^2 L(\hat{a})}{\partial a_i \partial a_j}$$

Ergibt Konfidenz-Ellipsen für die Parameter.

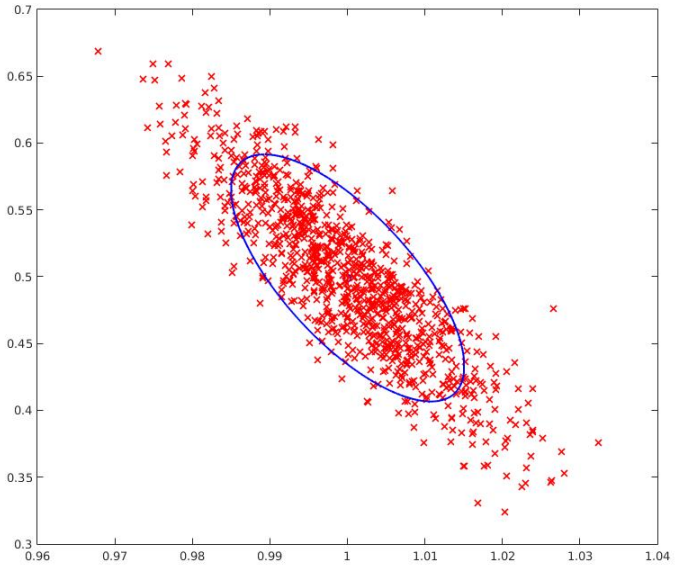


Abbildung 10.7

- Für Regression ist dies analog zur Fehlerfortpflanzung:

$$\sigma_a^2 = \sum_{i=1}^{N} \left(\frac{\partial a}{\partial y_i}\right)^2 \sigma_i^2$$

- In der nichtlinearen Modellierung gilt es nur asymptotisch.

Ein Ansatz im Endlichen: χ^2-Contour Konfidenzintervalle

- Konfidenzregionen über Iso-log-likelihood Kontouren

- Bestimmung durch Variation der Parameter um die geschätzten
- Ab, say, 10 Dimensionen nicht praktikabel

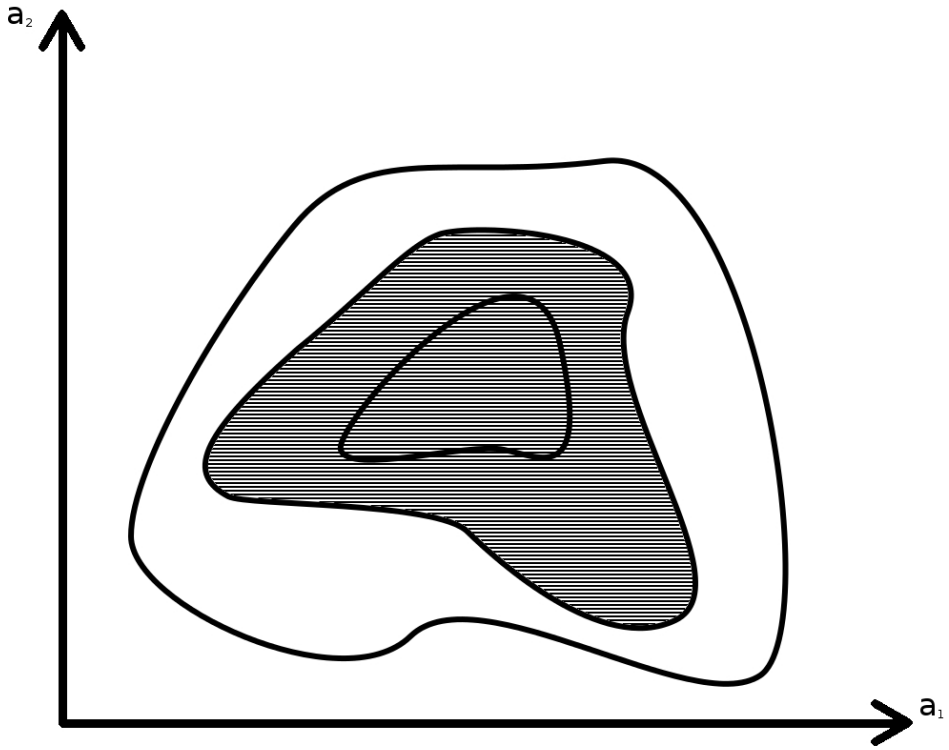


Abbildung 10.8: Konfidenzregionen über Iso-log-likelihood Kontouren

Grundsätzlich Alternative im Endlichen: Monte Carlo Konfidenzintervalle
(i) Parametrischer Bootstrap

- Bootstrap: Sich an den eigenen Stiefeln aus dem Sumpf ziehen © Münchhausen
- Schätze Parameter $\hat{a}$
- Produziere neue Datensätze mit
  - Parameter $\hat{a}$
  - neuen Fehlern unter parametrischer Annahme an ihre Verteilung
- Bestimme Konfidenzregion aus Verteilung der geschätzten $\hat{a}_i$.

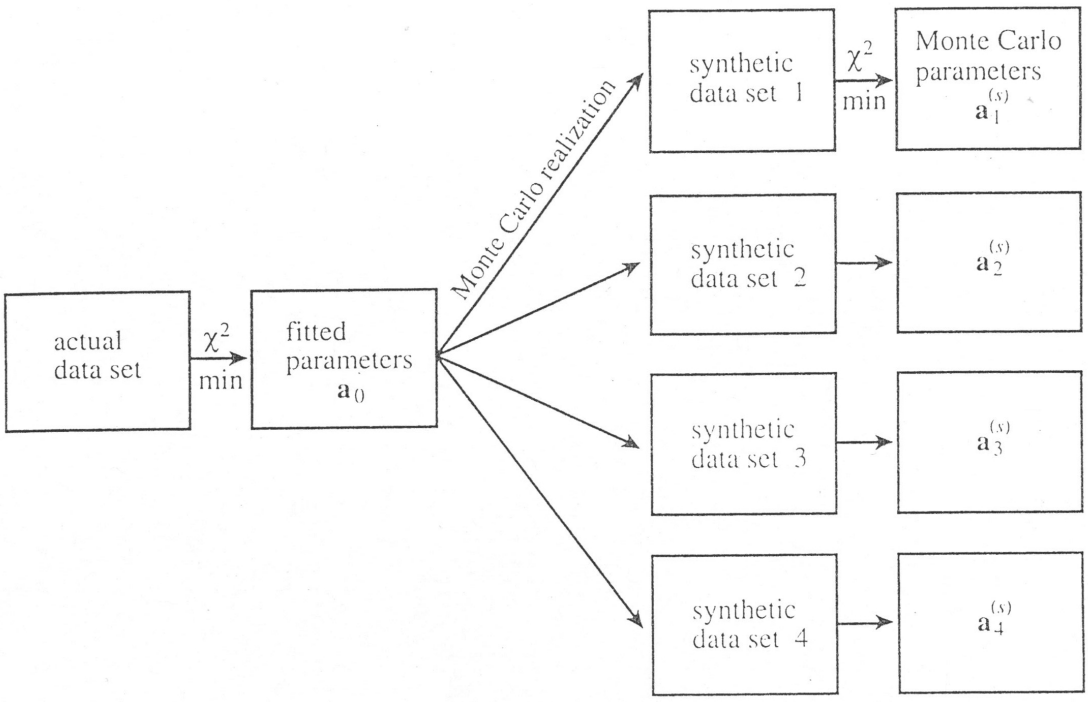


Abbildung 10.9: Monte-Carlo-Simulation eines Experiments

(ii) Nicht-parametrischer Bootstrap [14, 43]

- Bilde "neue" Datensätze durch Ziehen mit Zurücklegen aus den Originaldaten
- Motivation an Mittelwertschätzung bei diskreter Verteilung
- Dabei werden jeweils ≈ 32 % der Daten ausgetauscht.
- Manche kommen mehrfach vor
- Fitte jeweils Parameter.
- Konfidenzregion aus Verteilung der Fits
- Berücksichtigt empirische Verteilung der Fehler
- Korrektheit des Verfahrens: tiefsinnig.
  E. Mammen. When does the bootstrap work. Springer
- Here it does not work

- $x_i \sim U(0, a)$
- Bootstrap gibt kein Konfidenzintervall für a

Ein Beispiel zu (i), [62], p. 110 ff

- Betrachte das Modell, erinnere Aufgabenzettel 5:

$$y = \beta(1 - e^{-\gamma x}) + \sigma\epsilon, \qquad \text{mit } \beta = \gamma = 1$$

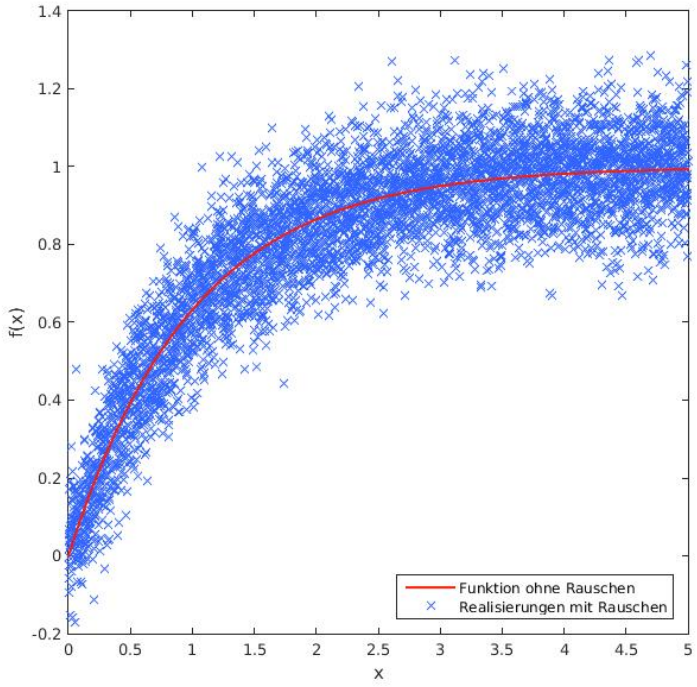


Abbildung 10.10: Zeichnung der Funktion mit $\sigma = 0.1$

- Diese Funktion hat 2 Charakteristika:
  - Steigung bei 0: $\frac{dy}{dx}|_{x=0} = \beta\gamma$
  - Sättigung für $x \to \infty$: β
- Wählt man $x \in [0, 1]$, so folgt:

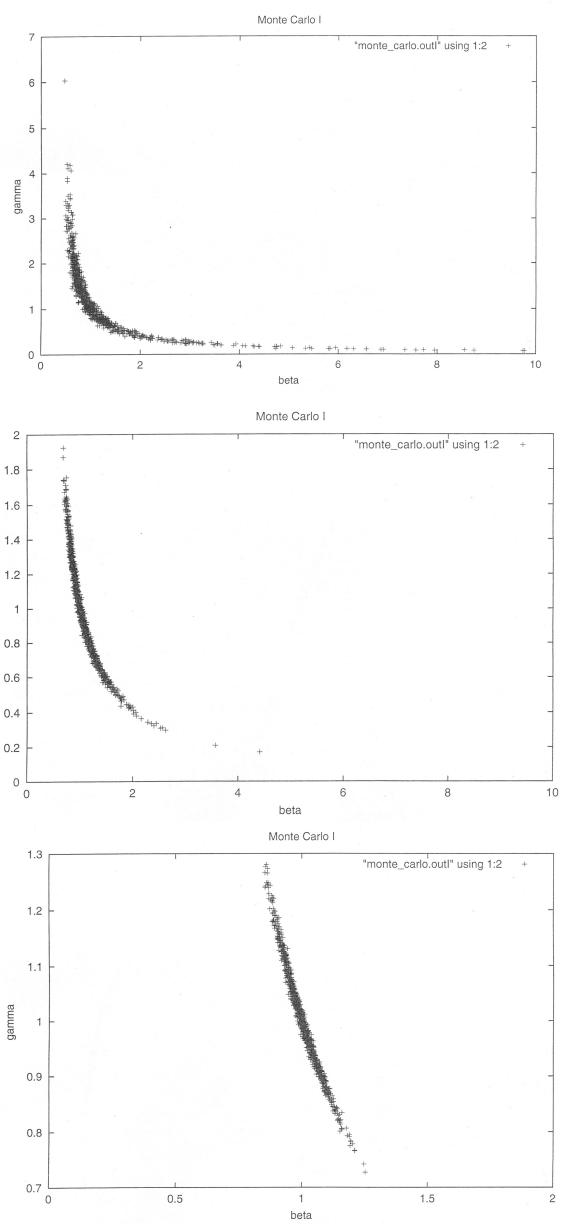


Abbildung 10.11: $x \in [0,1]$, $n_1 = 10, n_2 = 100, n_3 = 1000$

- Wählt man $x \in [0,5]$, so folgt:

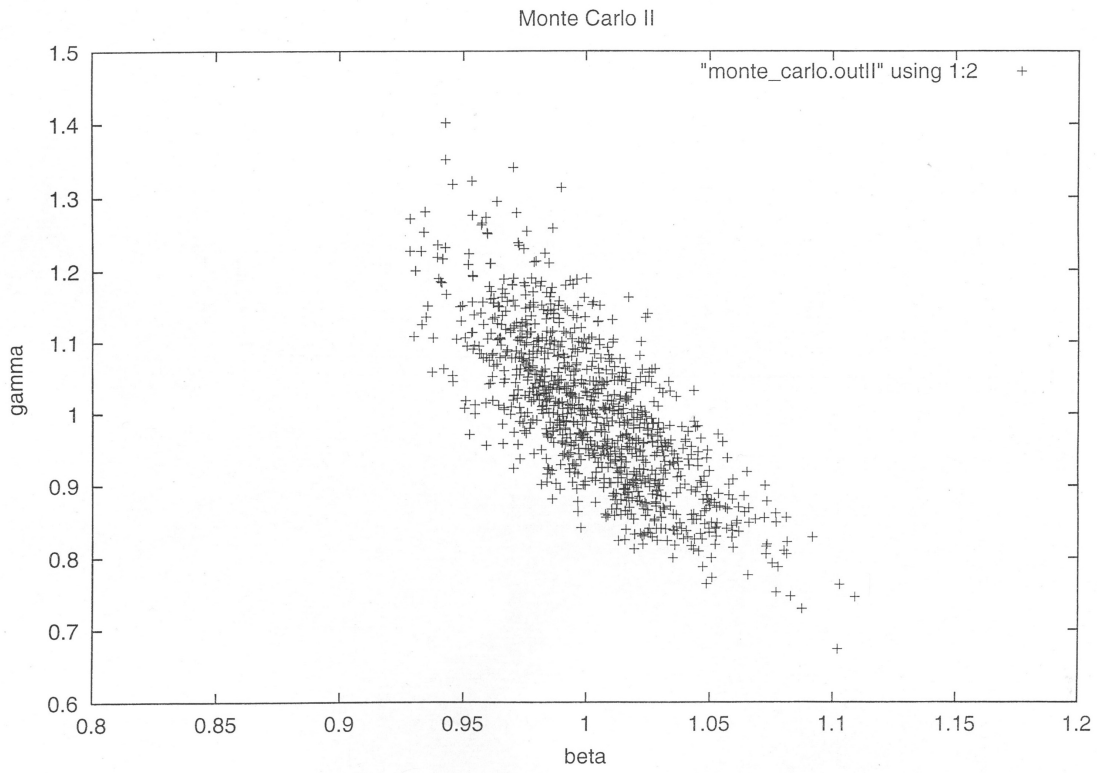


Abbildung 10.12: $x \in [0, 5]$, $n = 10$

Erklärung, $\Rightarrow$ Experimental Design [52].

Lessons learned:

- Lineare Regression und robuste Schätzer bei nicht-Gaußverteilung
- Nichtlineare Regression
- Nichtlineare Modellierung, Levenberg-Marquardt Algorithmus
- Vertrauensintervalle

11 Integration von Differentialgleichungen

11.1 Gewöhnliche Differentialgleichungen

Literatur:

- Recipes Kap. 16
- Stoer/Bulirsch Kap. 7

Aufgabe:

- Gegeben Dynamisches System:

$$\dot{\vec{x}} = \vec{f}(\vec{x}), \qquad \text{Anfangswert} : \vec{x}(t_0)$$

- Finde Trajektorie $\vec{x}(t)$, $t > t_0$, die bis auf einen kontrollierbaren Fehler mit wahrer Trajektorie übereinstimmt.

  Nomenklatur:

$$\frac{d}{dt} = \dot{}, \quad \frac{d}{dx} = ', \text{ Beachte: } \ddot{x} = \dot{f}(x) = f'(x)\dot{x} = f'(x)f(x) \tag{21}$$

11.1.1 Explizite Verfahren

Grundsätzliche Idee :

- Integrations-Schrittweite : h
- Taylor-Entwickung :

$$x_{t+h} = x_t + \dot{x}_t h + \frac{1}{2}\ddot{x}_t h^2 + \frac{1}{6}x_t^{(3)} h^3 + \mathcal{O}(h^4) \tag{22}$$

  $\dot{x}_t$ gegeben durch $f(x_t)$, aber $x_t^{(n)}$ möchte man nicht ausrechnen.

- Abbruch nach erster Ordnung: Euler-Verfahren:

$$x_{t+h} = x_t + f(x_t)h + \mathcal{O}(h^2)$$

  "Erster Ordnungs Verfahren"

- Idee: Höhere Ordnung durch geschickte Funktionsauswertungen.
  - Betrachte:

$$
\begin{aligned}
k_1 &= f(x_t)h \\
\text{Ansatz: } x_{t+h} &= x_t + f(x_t + \frac{1}{2}k_1)h \\
x_{t+h} &= x_t + f(x_t + \frac{1}{2}f(x_t)h)h \\
x_{t+h} &= x_t + f(x_t)h + f'(x_t)(\frac{1}{2}f(x_t)h)h \\
x_{t+h} &= x_t + f(x_t)h + \frac{1}{2}f'(x_t)f(x_t)h^2
\end{aligned}
$$

  - Mit Gl. (21) cancelled sich der 2. Ordnungs-Term in Gl. (22) und man erhält ein Verfahren 2. Ordnung (Midpoint Method).

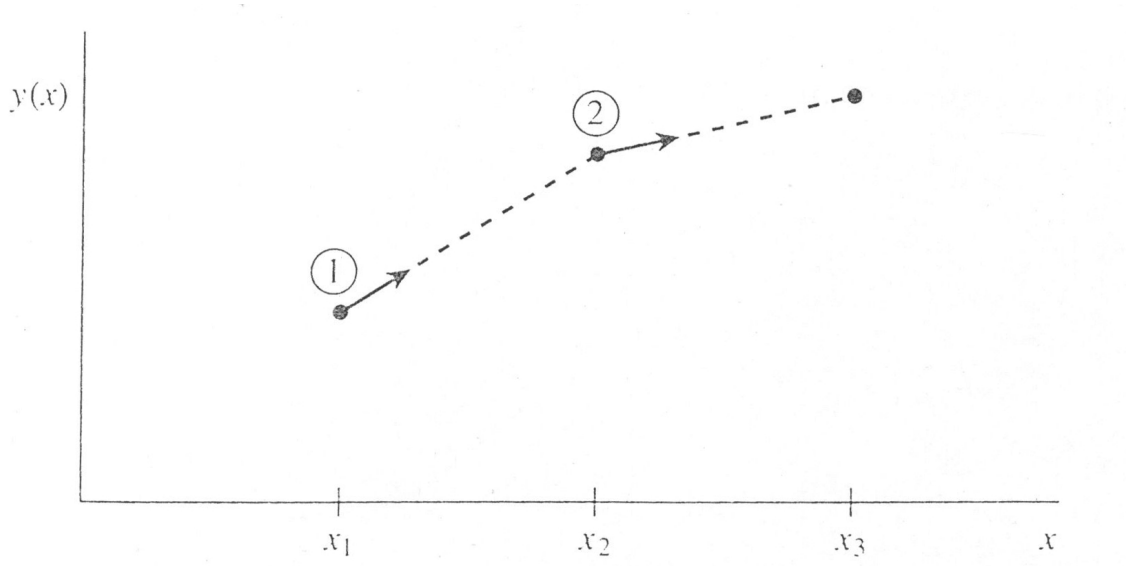


Abbildung 11.1: Euler-Verfahren. Einfachstes und ungenauestes Verfahren zum Integrieren von ODEs

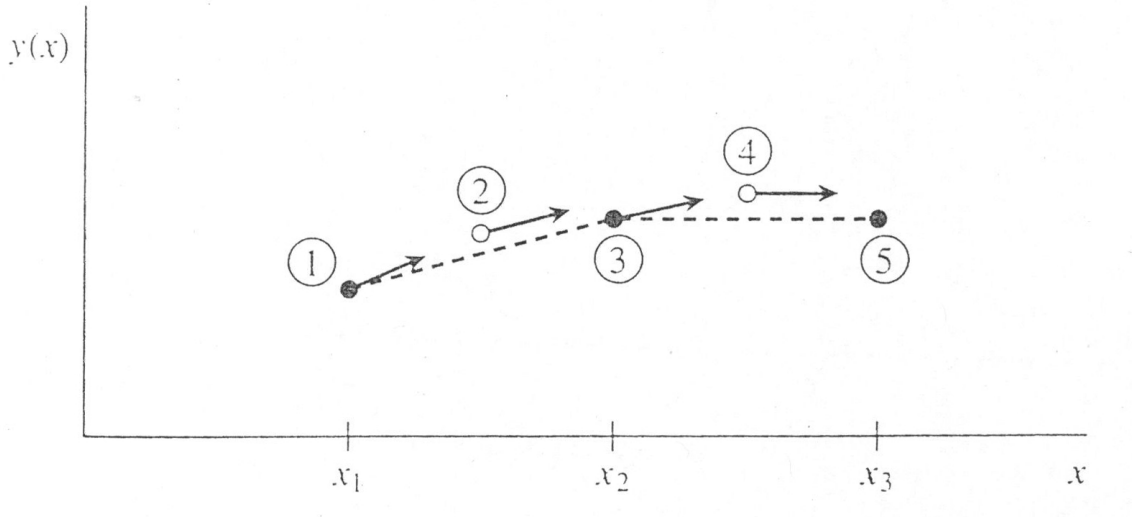


Abbildung 11.2: Midpoint-Verfahren. Verfahren 2. Ordnung

– Dieses läßt sich weiterspinnen
Allgemein:

$$\begin{aligned} x_{t+h} &= x_t + \sum_{j=1}^{p} \gamma_j k_j \\ k_1 &= f(x_t)\, h \\ k_j &= f(x_t + \sum_l \Gamma_{jl} k_l)\, h \end{aligned}$$

Speziell:

$$\begin{aligned} k_1 &= f(x_t)\, h \\ k_2 &= f(x_t + k_1/2)\, h \\ k_3 &= f(x_t + k_2/2)\, h \\ k_4 &= f(x_t + k_3)\, h \\ x_{t+h} &= x_t + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} + \mathcal{O}(h^5) \end{aligned}$$

heißt 4. Ordnung Runge-Kutta (1895)

- "Explizit", weil x_{t+h} explizit durch Werte zu früheren Zeitpunkten gegeben ist.

- Gehört zu den 5 wichtigsten Routinen, wo gibt.
- Im Allgemeinen:

  Ein 4. Ordnung Runge-Kutta Schritt mit h ist genauer als 2 Midpoint-Schritte mit $h/2$ ist genauer als 4 Euler-Schritte mit $h/4$.

Schrittweiten-Steuerung
"bis auf einen kontrollierbaren Fehler"
Näherungsfehler ist Funktion von $f(.)$, Schrittweite sollte adaptiert werden.

- Idee 1:

  Step Doubling: Integriere DGl mit Runge-Kutta 4. Ordnung mit

  (i) Schrittweite h: Ergebnis : $x_1(t+h)$

  (ii) zweischrittig mit $h/2$: Ergebnis : $x_2(t+h)$

  Die Differenz:

  $$\Delta = x_2 - x_1$$

  schätzt den Näherungsfehler, der von der Ordnung $\mathcal{O}(h^5)$ ist

- Idee 2:

  Embedded Runge-Kutta: Integriere DGl mit

  (i) 5. Ordnung Runge-Kutta Ergebnis : $x_5(t+h)$

  (ii) 4. Ordnung Runge-Kutta Ergebnis : $x_4(t+h)$
  ohne Mehraufwand (=embedded)

  Die Differenz:

  $$\Delta = x_5 - x_4$$

  schätzt den Näherungsfehler, der von der Ordnung $\mathcal{O}(h^5)$ ist

Praktisches Procedere:

- Wähle h und gewünschte Genauigkeit Δ_g

- Bestimme zu h gehörigen Fehler Δ
- Beachte, daß Δ mit h^5 skaliert.
- Wähle gewünschtes h_g nach:

$$h_g = h \left| \frac{\Delta_g}{\Delta} \right|^{0.2}$$

Wahl der gewünschten Genauigkeit Δ_g

- Relativer Fehler $\Delta_g = \epsilon |x_t|$
- $\Delta_g = \epsilon(|x_t| + |h\dot{x}_t|)$
- ...

Richardson Extrapolation, Stoer-Bulirsch Method
Idee:

- Der Fehler Δ ist eine Funktion von h mit $\Delta(0) = 0$.
- Berechne $\Delta(h_i)$, $h_i = h_0/i$ und extrapoliere nach $\Delta(0)$.

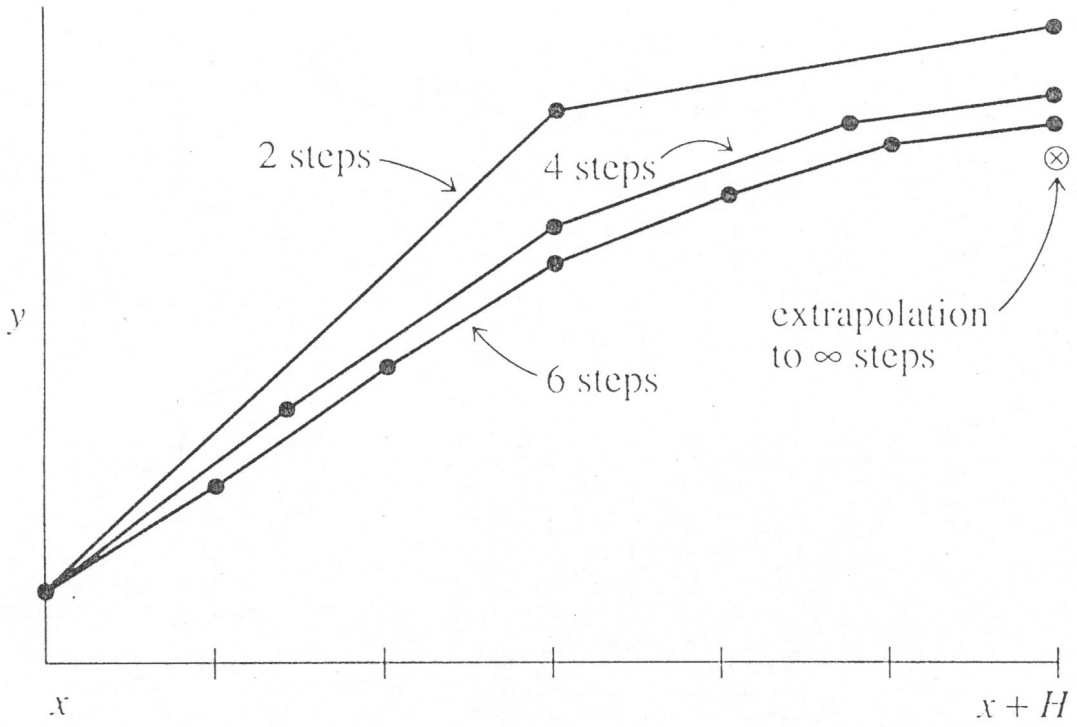


Abbildung 11.3: Richardson Extrapolation, wie in der Stoer-Burlisch-Methode benutzt

- Extrapolation liefert auch Schätzfehler der Extrapolation.

11.1.2 Implizite Verfahren

Das Problem:

- Betrachte das DGl-System:

$$\begin{aligned} \dot{x}_1 &= -\frac{\lambda_1 + \lambda_2}{2} x_1 - \frac{\lambda_1 - \lambda_2}{2} x_2 \\ \dot{x}_2 &= -\frac{\lambda_1 - \lambda_2}{2} x_1 - \frac{\lambda_1 + \lambda_2}{2} x_2 \end{aligned}$$

  mit $\lambda_i > 0$.

- Die allgemeine Lösung lautet:

$$x_1(t) = C_1 e^{-\lambda_1 t} + C_2 e^{-\lambda_2 t}$$
$$x_2(t) = C_1 e^{-\lambda_1 t} - C_2 e^{-\lambda_2 t}$$

- Integriert man Gleichungen mit Euler-Verfahren, so lauten die numerischen Trajektorien:

$$x_1(i) = C_1(1 - h\lambda_1)^i + C_2(1 - h\lambda_2)^i$$
$$x_2(i) = C_1(1 - h\lambda_1)^i - C_2(1 - h\lambda_2)^i$$

  Diese konvergieren nur, wenn gilt: $|1 - h\lambda_1| < 1$, $|1 - h\lambda_2| < 1$

- Sei $\lambda_2 \gg \lambda_1$, dann
  - ist Komponente $C_2 e^{-\lambda_2 t}$ für Lösung zu vernachlässen,
  - aber Schrittweite wird durch λ_2 diktiert.
- Systeme dieser Art nennt man steif. Schrittweitensteuerung läuft gegen $h = 0$.
- Obiges Argument gilt auch für Runge-Kutta und Stoer-Bulirsch.

Die Lösung:
Implizite Verfahren

- Betrachte 1D-Fall:

$$\dot{x} = -cx$$

  Das explizite (oder Vorwärts-) Euler-Verfahren lautet:

$$x_{t+h} = x_t + \dot{x}_t h = (1 - ch)x_t \tag{23}$$

  Erinnere: "Explizit", weil x_{t+h} hier explizit durch x_t gegeben.

- Methode ist unstabil, wenn $h > 2/c$, dann $|x_t| \to \infty$ für $t \to \infty$.

- Gl. (23) basiert auf:

$$\dot{x}_t \approx \frac{x_{t+h} - x_t}{h}$$

ebensogut gilt (implizites Differenzieren):

$$\dot{x}_{t+h} \approx \frac{x_{t+h} - x_t}{h}$$

Dies führt auf:

$$x_{t+h} = x_t + \dot{x}_{t+h}h = x_t - cx_{t+h}h \Longleftrightarrow x_{t+h} = \frac{x_t}{1+ch} \tag{24}$$

ein implizites Verfahren, weil x_{t+h} auf beiden Seiten der Gleichung steht.

- Dieses ist für alle h stabil, für lineare Systeme gibt es für $h \to \infty$ sogar die richtige asymptotische Lösung.
- Obiges Argument gilt auch für nichtlineare Systeme.
  - Für explizite Verfahren: Stabilität nur für

$$h < \frac{2}{\lambda_{\max}}, \quad \lambda_{\max} \text{ größter Eigenwert der Jakobi-Matrix von } f(.)$$

  - Implizite Verfahren: Immer stabil.
- Nicht all Systeme sind linear :-)

  Für

$$\dot{x} = f(x)$$

  lautet implizites Differenzieren:

$$x_{t+h} = x_t + f(x_{t+h})h \tag{25}$$

  Eine Selbstkonsistenz-Gleichung

  Versuche Linearisierung, erinnere Newton-Schritt aus Kap. 9 Optimierung :

$$x_{t+h} = x_t + \left(f(x_t) + \left.\frac{\partial f}{\partial x}\right|_{x_t} (x_{t+h} - x_t) \right) h$$

Sortieren ergibt:

$$x_{t+h} = x_t + h \left[\mathbf{1} - h \frac{\partial f}{\partial x} \right]^{-1} f(x_t)$$

- Hoffnung:
  h klein genug, dass dieses hinreichend gute Lösung für Gl. (25) ist.
- Merke:
  Jede Iteration benötigt eine Matrix-Inversion.

Es gibt Verallgemeinerungen für

- Runge-Kutta 4.Ordnung: Rosenbrock Methode
- Stoer-Bulirsch-Extrapolation: Bader-Deuflhard Methode

11.1.3 Integration Hamilton'scher Systeme

Kurzform Hamilton'sche Systeme

- Existieren für ein d-dimensionales Hamilton'sches System $d/2$ Erhaltungsgrößen, ist das System integrabel
- Dann Dynamik äquivalent zum Torus
- Ist das System integrabel, gehe man auf Winkel-Wirkungsvariablen und muß nur noch Sinusse auswerten.

Ansonsten:

- Für Hamilton'sche Systeme

  $$\dot{p} = -\frac{\partial H(x,p)}{\partial x}, \qquad \dot{x} = \frac{\partial H(x,p)}{\partial p}$$

  muß die Flußabbildung f_H^t :

$$\begin{pmatrix} p(t) \\ x(t) \end{pmatrix} = f_H^t \begin{pmatrix} p(0) \\ x(0) \end{pmatrix}$$

den Satz von Liouville und damit die Eigenschaft:

$$\det(Df_H^t) = 1, \qquad \text{mit } Df_H^t = \text{ Jacobimatrix}$$

erfüllen.

- Derartige Algorithmen heißen symplektische Integratoren, siehe z.B. [10, 15].
- Idee: Nach jedem Schritt wieder auf die erlaubte Energieschale projezieren.

Übung:
Integration des van der Pol Oszillators

10.18
10. Woche

11.2 Partielle Differentialgleichungen

Dieses Kapitel hat Daniel Lill beigesteuert.

PDEs sind Differentialgleichungen in mehreren Variablen, z.B. die Diffusionsgleichung:

$$\partial_t u(\vec{x}, t) = D \triangle u(\vec{x}, t)$$

mit Diffusionskonstante D.

Sie sind omnipräsent in der Physik:

- Wellengleichungen
- Maxwell-Gleichungen
- Schrödingergleichung

Allgemeines:

- Wie bei ODEs: Diskretisierung. Hier: mehrdimensionales Gitter.
- Eine eindeutige Lösung braucht passende Randbedingungen.
  Zwei wichtige Klassen:

– Anfangswertprobleme, z.B. Wellengleichung
  Jeder Zeitschritt kann nacheinander berechnet werden
– Randwertprobleme, z.B. Poissongleichung
  Simultane Lösung auf gesamtem Gitter

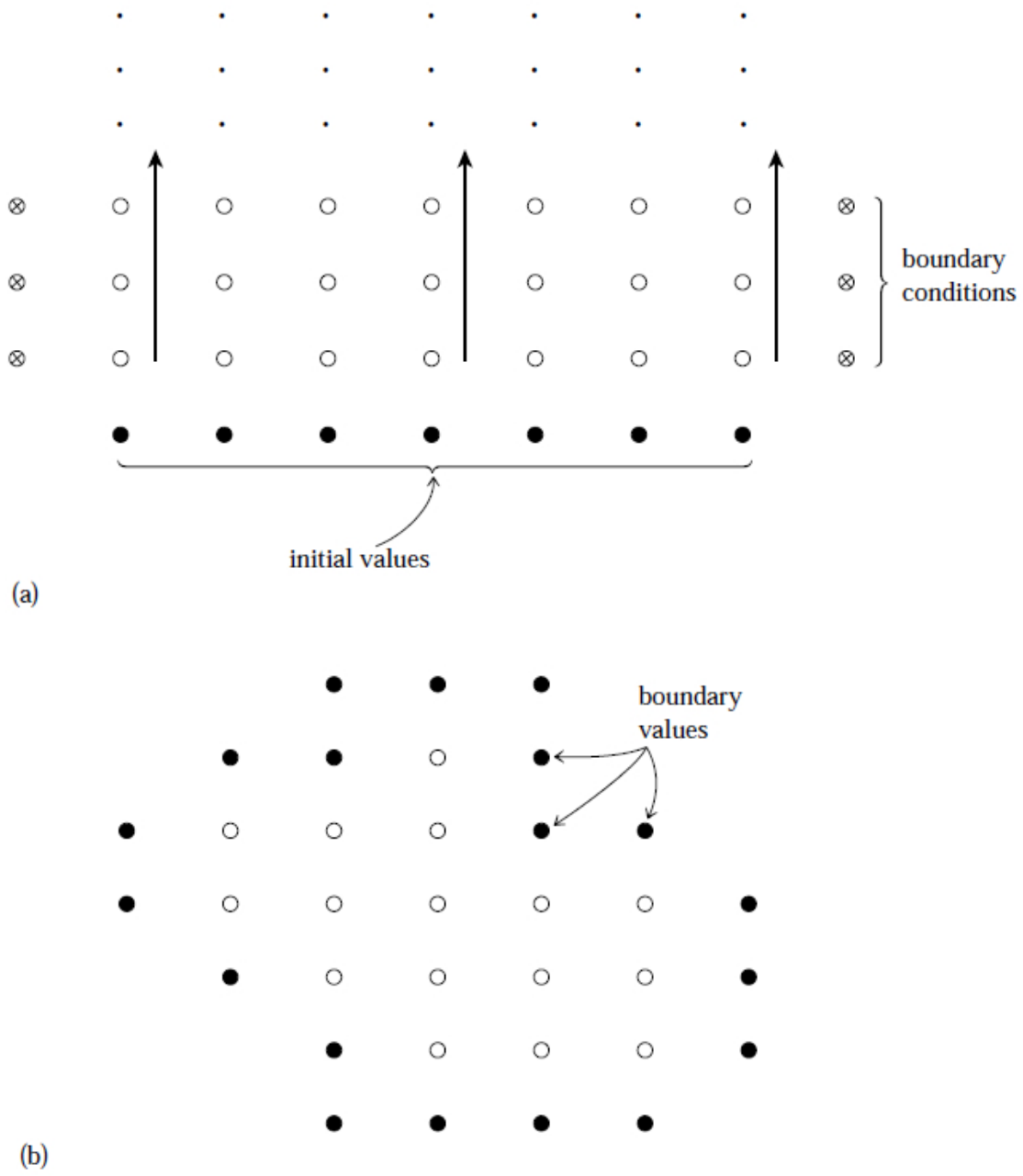


Abbildung 11.4: Zum Unterschied von Anfangs- und Randwertproblemen, aus Numerical Recipes, 3. ed

11.2.1 Anfangswertproblem am Beispiel der eindimensionalen Diffusionsgleichung

Die Diffusionsgleichung in einer Dimension:

$$\partial_t u(x,t) = D\,\partial_x^2 u(x,t)$$

mit Diffusionskonstante D und den Randbedingungen $u(x,0) = f(x)$

Finite Differenzen

- Erste Ableitung

$$\dot{u}(t) = \frac{u(t+\Delta t) - u(t)}{\Delta t} + \mathcal{O}(\Delta t)$$

- Zweite Ableitung

  Taylorentwicklung

$$u(x \pm \Delta x) = u(x) \pm u'(x)\Delta x + \frac{u''(x)}{2}\Delta x^2 \pm \frac{u'''(x)}{3!}\Delta x^3 + \mathcal{O}(\Delta x^4)$$

  Addition der Gleichung mit "+" und "−"

$$u(x+\Delta x) + u(x-\Delta x) = 2u(x) + u''(x)\Delta x^2 + \mathcal{O}(\Delta x^4)$$

  liefert eine Näherung der zweiten Ableitung:

$$u''(x) = \frac{u(x+\Delta x) - 2u(x) + u(x-\Delta x)}{\Delta x^2} + \mathcal{O}(\Delta x^2)$$

FTCS-Differenzenschema

- FTCS = Forward Time Centered Space Differenzenschema auf x-t-Gitter:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D\,\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) \tag{26}$$

  mit $u_j^n = u(j\Delta x, n\Delta t)$ mit $j = 1,\ldots,J$ und $n = 1,\ldots,T/\Delta t$

- Randbedingungen sind: $u_j^0 = f_j$ und z.B. $u_0^n = u_{J+1}^n = 0$

- Entscheidende Frage: Ist dieser Algorithmus stabil?

  Ein Algorithmus ist stabil, wenn er gegenüber Rundungsfehler unempfindlich ist.

von Neumannsche Stabilitätsanalyse:

- Sei $u_j^n = N_j^n + \epsilon_j^n$
  - u_j^n die exakte Lösung der Differenialgleichung
  - N_j^n die Lösung mit Rundungsfehler
  - $(-)\epsilon_j^n$ die Rundungsfehler
- Beachte: Auf Grund der Linearität löst Rundungsfehler ϵ_j^n auch die PDE.

  Rundungsfehler hat also dieselben Wachstumseigenschaften wie die Lösung selbst.
- Betrachte Separationsansatz
$$u_j^n = T_n X_j$$
  Damit Gl. (26)
$$T_{n+1}X_j - T_n X_j = s\, T_n(X_{j+1} - 2X_j + X_{j-1})$$
  Mit
$$s = \frac{D\Delta t}{\Delta x^2}$$
- Teilen durch T_n und X_j und sortieren ergibt:
$$\frac{T_{n+1}}{T_n} = 1 - s\left(2 - \frac{(X_{j+1} + X_{j-1})}{X_j}\right)$$
  Linke Seite hängt nur von n, rechte nur von j ab $\Longrightarrow$ beide Seiten müssen konstant sein.
$$\frac{T_{n+1}}{T_n} = g \quad \Longrightarrow \quad T_n = T_0 g^n$$
  g Wachstumsfaktor
$$1 - s\left(2 - \frac{X_{j+1} + X_{j-1}}{X_j}\right) = g \quad \Longrightarrow \quad g = 1 - 2s(1 - \cos(k\Delta x))$$
  Ergo: Stabil wenn $|g| < 1$, also $s < \frac{1}{2}$.

- Starke Einschränkung für die Schrittweite Δt, sie geht $\propto \Delta x^2$.

  Für ODEs reicht kleines Δt, hier zusätzliche Bedingung

Implizites Differenzenschema BTCS

- Das implizite Differenzenschema BTCS (Backward Time Centered Space)

$$u_j^{n+1} - u_j^n = s(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) + \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) \tag{27}$$

  hat Wachstumsfaktur

$$g = \frac{1}{1 + 4s\sin^2(k\Delta x/2)}$$

  und ist daher für alle s stabil.

- Für $\Delta t \to \infty$ stellt sich die Gleichgewichtsform der Gleichung ein:

$$\partial_{xx} u = 0$$

  und die Lösung ist wie bei impliziten ODE Lösern für lineare Systems asymptotisch richtig.

Crank-Nicolson-Schema

- Crank-Nicolson-Schema: Mittelwert aus FTCS- und BTCS-Schema.

$$u_j^{n+1} - u_j^n = \frac{s}{2}(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} + u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)$$

  Wachstumsfaktor

$$g = \frac{1 - s(1 - \cos k\Delta x)}{1 + s(1 - \cos k\Delta x)}$$

- Für alle s stabil
- Zeitliche Abschneidefehler ist $\mathcal{O}(\Delta t^2)$.

Für nichtlineare PDEs liefert von Neumannsche Stabilitätsanalyse nur notwendige, jedoch nicht immer hinreichende Bedingungen für die Stabilität.

11.2.2 Randwertprobleme

Beispiel: Laplace-Gleichung:

$$\triangle u(\vec{x}) = 0 \text{ für } \vec{x} \in V$$

mit Randbedingungen an $u(\partial V)$ oder $\frac{\partial u}{\partial n}(\partial V)$ gegeben.

- Finite Differenzen für 2D Laplace-Gleichung:

$$\frac{u_{j+1,i} - 2u_{j,i} + u_{j-1,i}}{\Delta x^2} + \frac{u_{j,i+1} - 2u_{j,i} + u_{j,i-1}}{\Delta y^2} + \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta y^2) = 0$$

  Sortieren ergibt:

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1})$$

  u_{ij} ist also der Mittelwert seiner nächsten Nachbarn.

- Naives Iterieren: Jacobi-Iteration

$$u_{i,j}^{(n+1)} = \frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)})$$

- Schneller ist das Gauß-Seidel-Verfahren:
  - Fange links unten an und berechne die Werte in der ersten Zeile von links nach rechts
  - Nutze die neuen Werte bereits für der nächsten Punkte

$$u_{i,j}^{(n+1)} = \frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)})$$

- Noch schneller ist das successive overrelaxation Verfahren:

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \omega(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)} - 4u_{i,j}^{(n)}) \tag{28}$$

  mit ω klug gewählt.

11.2.3 Methode der finiten Elemente

- Statt Näherung der PDE durch finite Differenzen ...
- Näherung der Lösung durch Linearkombination von Basisfunktionen

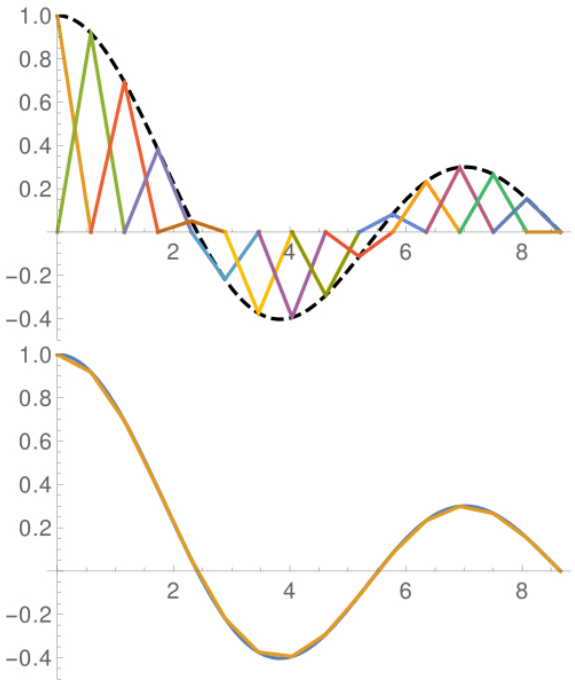


Abbildung 11.5: Die Besselfunktion wird durch eine Linearkombination der bunten Dreiecksfunktionen approximiert.

Beispiel:

- Poissongleichung in 1 D

$$u''(x) = -\rho, \qquad x \in [0,1]$$

Randbedingungen

$$u(0) = u(1) = 0$$

Formulierung des Problem in seiner schwachen Form in Bezug auf Basisfunktionen v_i:

$\forall v_i$ mit $v_i(0) = v_i(1) = 0$ gilt:

$$\int_0^1 -\rho v_i(x)dx = \int_0^1 u''(x)v_i(x)dx = u'(x)v_i(x)|_0^1 - \int_0^1 u'(x)v_i'(x)dx \qquad (29)$$

Der erste Term rechts verschwindet wegen der Randbedingungen von v_i.

- Teile den Bereich $[0, 1]$ in kleine Intervalle $[x_i, x_{i+1}]$ auf und ordne jedem Punkt x_i eine Dreiecksfunktion zu:

$$v_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & x \in [x_i, x_{i+1}] \\ 0 & sonst \end{cases}$$

Die Funktion $u(x)$ wird als Linearkombination dieser v_i ausgedrückt

$$u(x) = \sum_i a_i v_i(x) \qquad (30)$$

Statt unendlich jetzt endlich dimensional

- Gl. (29) wird mit Gl. (30) zu:

$$\forall j: \qquad \int_0^1 \rho v_j(x)dx = \sum_i a_i \int_0^1 v_i'(x)v_j'(x)dx$$

Ein lineares Gleichungssystem.

- Mit

$$M_{ij} = \int_0^1 v_i'(x)v_j'(x)dx$$

und

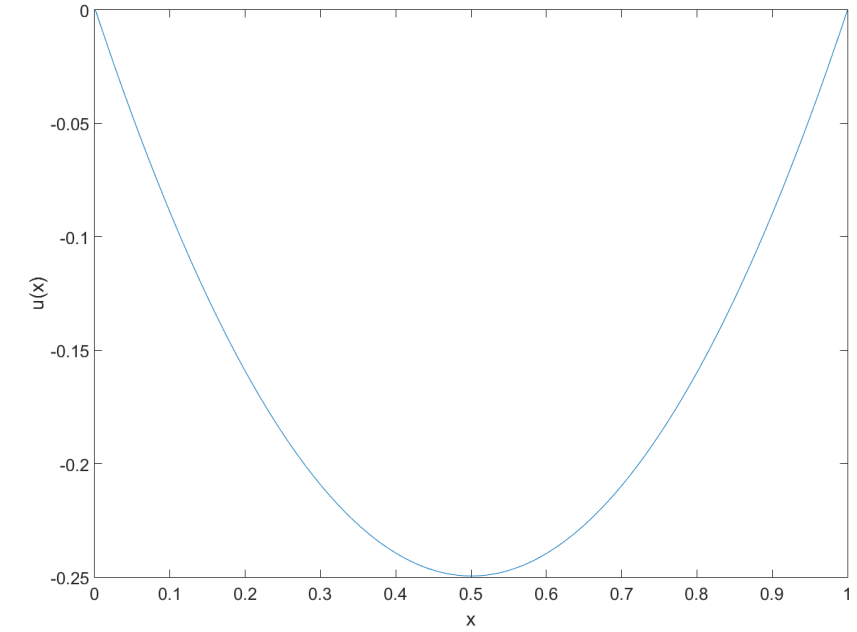
$$w_j = \int_0^1 \rho v_j(x)dx$$

folgt

$$\sum_i M_{ij} a_i = w_j \quad \Longrightarrow \vec{a} = M^{-1}\vec{w}$$

Beachte: Die Matrix M ist nur dünn besetzt, sodass Invertierung schnell geht.

In Abbildung 11.6 ist das 1D-Poisson-Problem mit $\rho = -2$ mit zwei unterschiedlichen Grids dargestellt. Die Parabel ist bei beiden gut zu erkennen, jedoch wurde in der rechten Abbildung das Grid zum Rand hin gröber gewählt.



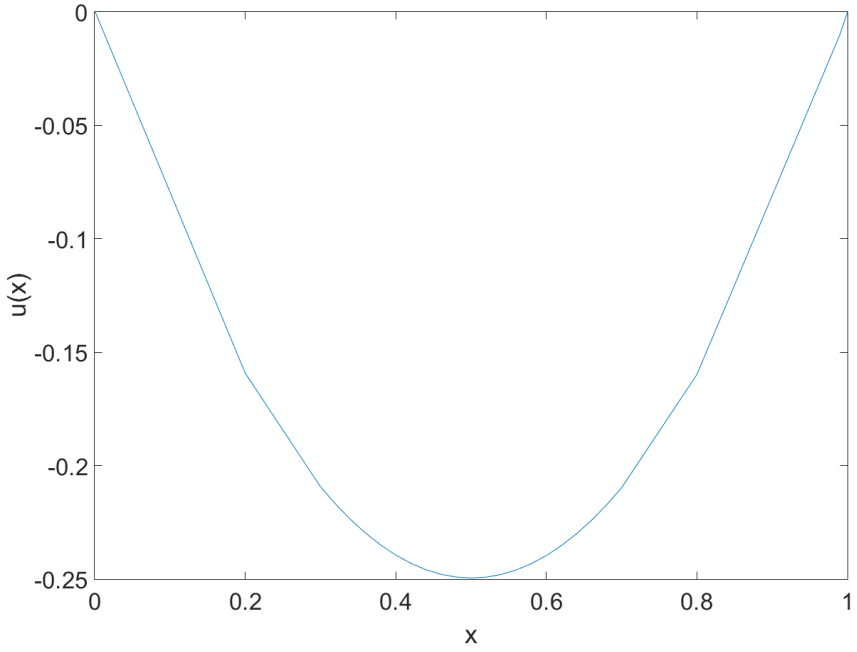


Abbildung 11.6: Das 1-d Poisson-Problem a) mit feinem Grid b) mit grobem Grid in der Nähe des Randes und feinem Grid beim Minimum der Parabel

Anmerkungen zur Methode der finiten Elemente

- Verallgemeinerung auf höhere Dimensionen: In 2D wird der Bereich trianguliert, in Dreiecke zerlegt. Zu jedem Punkt i gehören dann mehrere Dreiecke j, zu denen jeweils eine Basisfunktion $v_j(x, y)$ gehört. Analog für höhere Dimensionen.
- Grid kann flexibel auf die Geometrie des Problems angepasst werden. Beispiel: Crash-Simulationen, bei denen die Knautschzone feiner gerastert ist als das Heck des Fahrzeugs.
- Andere Basisfunktionen z.B. Polynome sind auch möglich
- Spektrale Methoden funktionieren ähnlich:
  - Die Funktion $u(.)$ wird wieder in endlich vielen Basisfunktionen entwickelt
  - Aber: Basisfunktionen haben vollen Träger, z.B. Fourierreihe
  - Abbruch bei der bestimmten Frequenz.

11.3 Stochastische Differentialgleichungen

Literatur:

- P.E. Kloeden, E. Platen. *Numerical Solution of Stochastic Differential Equations* [31], Mathematisch ausführlich
- P.E. Kloeden, E. Platen, H. Schurz. *The Numerical Solution of SDE through Computer Experiments* [32], mit Simulationssoftware
- B. Øksendal. *Stochastic Differential Equations* [48], gutes Buch
- J. Honerkamp. *Stochastic Dynamical Systems* [25] Kap. 10, Kondensierte Darstellung für Physiker

Stochastische Differentialgleichung (SDE), Physiker-Definition, Langevin-Gleichung

$$\dot{x} = f(x, \epsilon) = a(x) + b(x)\epsilon, \qquad \epsilon \sim N(0, 1)$$

- $a(x)$: Deterministischer Anteil: Drift-Term
- $b(x)\epsilon$: Stochastischer Anteil: Diffusions-Term
- ϵ: Dynamisches Rauschen
- Grundsätzliches Problem: $\dot{x}$ und x nicht glatt
- Mathematiker-Definition

  $$dx = a(x)dt + b(x)dW \tag{31}$$

  dazu unten mehr

Warum stochastische DGls ?

- Modellierung der äußeren Einflüsse in offenen (deterministischen) Systemen.

  Klassisches Beispiel: Brownian Motion:

  $$x(t) = x(t-1) + \sigma\epsilon(t), \quad \epsilon(t) \sim N(0, 1)$$

  Zeitskalenseparation zwischen träger Polle und schnellen Wasserteilchen

  Physikalische Interpretation

$$\begin{aligned} x(t) &= x(t-\Delta t) + \sigma\epsilon(t) \\ \frac{x(t)-x(t-\Delta t)}{\Delta t} &= \frac{\sigma\epsilon(t)}{\Delta t} \\ \lim_{\Delta t\to 0}: \quad \dot{x} &= \tilde{\epsilon} \end{aligned}$$

Geschwindigkeit ist weißes Rauschen mit Mittelwert 0, über $\tilde{\epsilon}$ denken wir unten nach

- Modellierung komplizierter Teilbereiche eines deterministischen Systems.
- Eigentlich in nicht-Hamilton'schen dissipativen Systemen immer nötig wegen Fluktations-Dissipations - Theorem : Wo Reibung, da auch Stochastik in der Dynamik [40].
- In Hamilton'schen Systemen führt dynamisches Rauschen zur Divergenz.

Bedeutung Term $b(x)\epsilon$:

- zustandsabhängige Varianz
- Parametrisches Rauschen:

$$\dot{x} = -(c+\epsilon)x = -cx - \epsilon x$$

Verrauschte Parameter

Integration von SDEs
Statt als Taylor-Entwicklung Gl. (22) lassen sich die verschiedenen Verfahren zur Integration von deterministischen DGls aus Kap. 11 auch als Näherungen von Integralen lesen:

$$\begin{aligned} \dot{x} &= f(x) \\ &\Longleftrightarrow \\ x_{t+h} &= x_t + \int_t^{t+h} f(x_{t'})dt' \end{aligned}$$

- Explizites Euler-Verfahren: $\int_t^{t+h} f(x'_t)dt' \approx f(x_t)h$

- Implizites Euler-Verfahren: $\int_t^{t+h} f(x'_t)dt' \approx f(x_{t+h})h$
- Runge-Kutta: Integralauswertung an mehreren Stellen

Für SDEs geht nur Zugang über Integralinterpretation.

$$x_{t+h} = x_t + \int_t^{t+h} f(x_{t'}, \epsilon_{t'})dt' = x_t + \int_t^{t+h} (a(x_{t'}) + b(x_{t'})\epsilon_{t'})dt'$$

Betrachte einfachstes Beispiel: Linear gedämpftes stochastisch getriebenes System

$$\begin{aligned} \dot{x} &= -\alpha x + \sigma\epsilon \\ x_{t+h} &= x_t + \int_t^{t+h} -\alpha x_{t'}dt' + \sigma \int_t^{t+h} \epsilon_{t'}dt' \end{aligned}$$

Aber was ist ein Integral über $\epsilon_{t'}$?

- Betrachte:

$$\int_t^{t+h} \epsilon_{t'}dt'$$

  Macht weder im Riemann- noch im Lebesgue-Sinne Sinn.

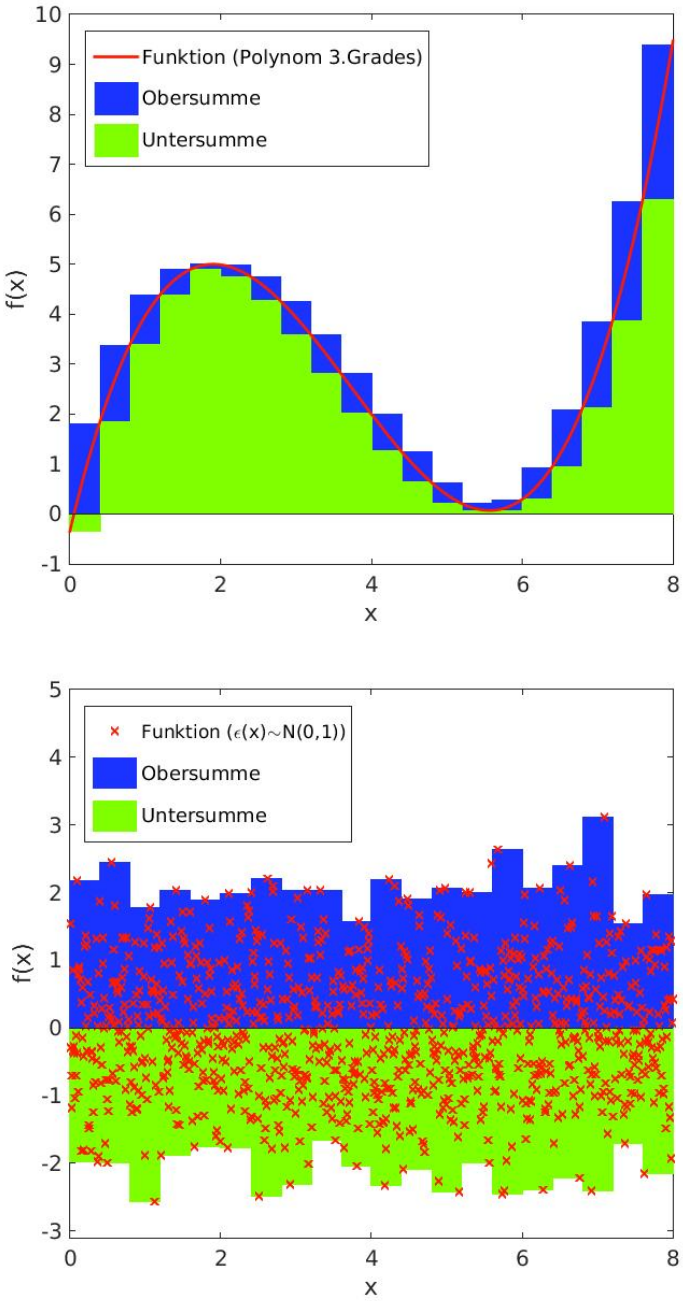


Abbildung 11.7: Ober- und Untersumme bei einem Polynom und einer stochastischen Funktion

- Beobachtung:
  Ergebnis des Integrals ist Brownian Motion
- Brownian Motion in diskreter Zeit ($\Delta t = 1$) ist:

$$x(t) = x(t-1) + \sigma\epsilon(t) \qquad x(0) = 0, \quad \epsilon(t) \sim N(0,1)$$

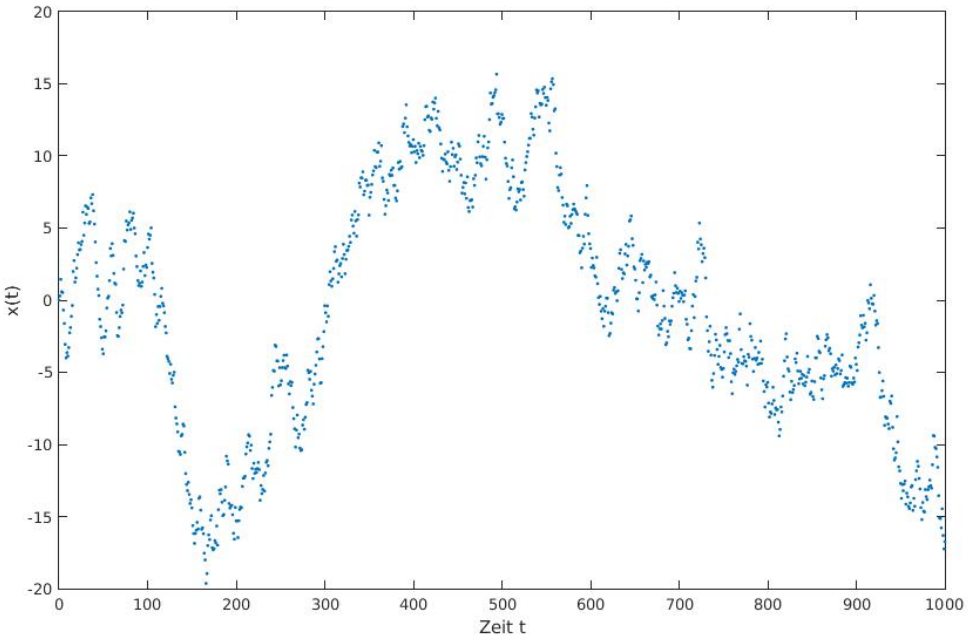


Abbildung 11.8: Brownian Motion

- Ineinander eingesetzt

$$\begin{aligned} x(t-1) &= x(t-2) + \sigma\epsilon(t-1) \\ x(t) &= \sigma \sum_{t'=1}^{t-1} \epsilon(t') \end{aligned}$$

- Da Varianzen additiv, gilt

$$\langle x^2(t) \rangle = \sigma^2 t, \quad \langle x(t) \rangle = 0 \tag{32}$$

$x(t)$ ist Gauß'sche Zufallsvariable mit Mittelwert 0 und Standardabweichung $\sigma\sqrt{t}$

- DEFINIERE:

$$\int_t^{t+h} dt'\, \epsilon_{t'} := \sqrt{h}\epsilon_t$$

- Bemerkung: Die Mathematiker drehen es herum:
  1. Definiere zeitkontinuierliche Brownian Motion durch Gl. (32), Wiener-Prozess
  2. Definiere "ϵ" als Inkremente, i.e. Zuwächse, des Wiener-Prozesses, dW in Gl. (31)

Für Genießer zum Selbststudium: Ito und Stratonovich Integral

- für additives Rauschen identisch
- für multiplikatives Rauschen unterschiedlich

Damit Euler-Verfahren für $\dot{x} = a(x) + b(x)\epsilon$

$$x_{t+h} = x_t + a(x_t)h + b(x_t)\epsilon_t\sqrt{h} + \mathcal{O}(h)$$

- Höhere Ordnung i.a. sehr schwierig, da komplizierte stochastische Integrale auftreten, siehe [25].
- Euler bewirkt: Integrationszeitschritt i.a. $\ll$ natürliche Samplingzeitschritt, siehe [71] speziell zur Integrationszeitschrittwahl.

Übung:
Integration des stochastischen van der Pol Oszillators

11.18
11. Woche

11.4 Gillespie-Algorithmus

Literatur:

- Original [21]
- See also: [19, 44, 53]
- Kritische Auseinandersetzung mit Grundlagen und Interpretation [75]

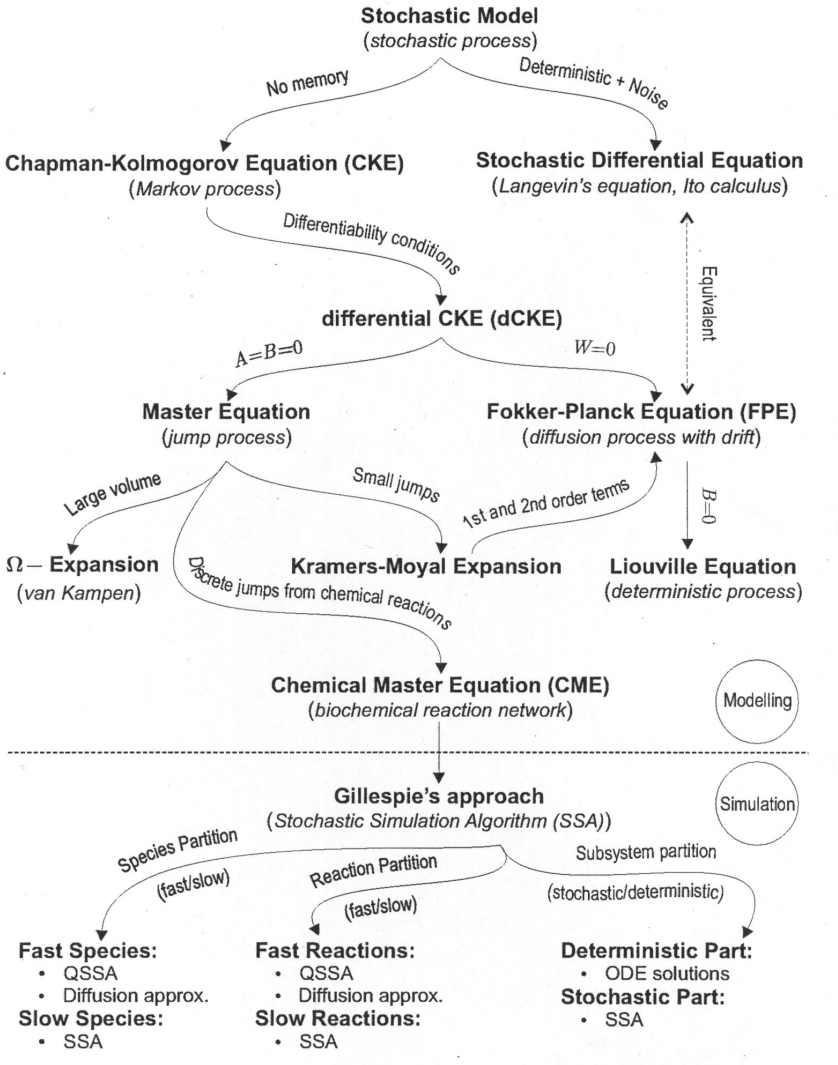


Abbildung 11.9: Überblick Stochastische Modelle

Alle Dynamik ist diskret-zahlig

- Populationsdynamik von Tiere
- Chemische Reaktionen zwischen Molekülen
- Bankverkehr
- Besetzungszahlen-Formalismus in der Quantenmechanik, a und $a^\dagger$
- Differentialgleichungen sind Grenzfall

Betrachte im Folgenden chemische Reaktionen

Sei S eine Spezies und

$$P_n(t) = Prob(\#S(t) = n \text{ zum Zeitpunkt } t)$$

Betrachte:

- Propensity $a_i(.)$: Wahrscheinlichkeit pro Zeiteinheit für Änderung des Zustandes
- Zufluß nach $P_n(t)$

  $$n - \nu_i \overset{a_i(n-\nu_i)}{\longrightarrow} n$$

  mit $a_i(n - \nu_i)$ der Rate der Änderung um ν_i, gegeben Zustand war in $n - \nu_i$
- Abfluß von $P_n(t)$

  $$n \overset{a_i(n)}{\longrightarrow} n + \nu_i$$

  mit $a_i(n)$ die Rate der Änderung um ν_i, gegeben Zustand war in n

Dann Chemische Master-Gleichung:

$$\dot{P}_n = \sum_{i=1}^{N} a_i(n - \nu_i)P_{n-\nu_i} - a_i(n)P_n$$

In der Regel

- mehr als eine Spezies: $P(S_1, S_2, \ldots, S_K)$

- mehrere mögliche Reaktionen $R_1, R_2, \ldots, R_M$
- nicht analytsich zu lösen.

Gillespie-Algorithmus: Statt analytischer Lösung

- Simuliere viele Trajektorien
- Ermittle Resultate durch Mitteln oder betrachte Verteilungen
- Man kann zeigen: Gillespie Algorithmus produziert die richtigen Verteilungen

Gillespie-Algorithmus beantwortet:

- Wann findet die nächste Reaktion statt ?
- Welche ist es ?

Zentrale Größe: Reaktions-Wahrscheinlichkeits-Funktion $P(i, \tau)$
$P(i, \tau)d\tau$: Wahrscheinlichkeit für Reaktion R_i im Intervall $(t+\tau, t+\tau+d\tau)$, gegeben System in Zustand $S(t)$

$$P(i, \tau)d\tau = P_0(\tau)\, P_i(d\tau) \tag{33}$$

mit

- $P_i(d\tau) = a_i d\tau$: Wahrscheinlichkeit, dass Reaktion R_i im Intervall $(t+\tau, t+\tau+d\tau)$ stattfindet.
- $P_0(\tau)$: Wahrscheinlichkeit, dass gegeben Zustand $S(t)$ keine Reaktion im Intervall $(t, t+\tau)$ stattfindet

Die Wahrscheinlichkeit, dass irgendeine Reaktion im Intervall $d\tau$ stattfindet, ist:

$$\sum_{i=1}^{M} a_i d\tau$$

- Definiere:

$$a^* = \sum_{i=1}^{M} a_i$$

Wahrscheinllichkeit für keine Reaktion im Intervall $d\tau$: $1 - a^* d\tau$.

- Daher

$$P_0(\tau + d\tau) = P_0(\tau)(1 - a^* d\tau) = P_0(\tau) - a^* P_o(\tau) d\tau$$

  Ergibt Differentialgleichung

$$\dot{P}_0 = -a^* P_0, \text{mit Lösung } P_0(\tau) = e^{-a^*\tau}$$

  $P_0(0) = 1$ o.k.

- Zusammengefaßt, mit Gl. (33):

$$P(i, \tau) = a_i e^{-a^*\tau}$$

Zentrale Fragen:

- Wann wird eine nächste Reaktion stattfinden ?
- Welche Reaktion ist die nächste ?

Wann ?
Summation über alle Reaktionen

$$\bar{P}(\tau) = \sum_{i=1}^{M} P(i, \tau) = a^* e^{-a^*\tau}$$

$\bar{P}(\tau)d\tau$: Wahrscheinlichkeit für irgendeine nächste Reaktion in Intervall $(t + \tau, t + \tau + d\tau)$

Welche Reaktion ?
Gegeben eine Reaktion geschieht im Intervall $(t + \tau, t + \tau + d\tau)$, gibt die bedingte Wahrscheinlichkeit

$$\tilde{P}(i|\tau) = \frac{P(i, \tau)}{\bar{P}(\tau)} = \frac{a_i e^{-a^*\tau}}{a^* e^{-a^*\tau}} = \frac{a_i}{a^*}$$

die Wahrscheinlichkeit, dass es Reaktion i ist.

Auf dem Weg zum Algorithmus:

- Wann ?

- Die kumulative Verteilung $F(\tau)$ für $\bar{P}(\tau)$ lautet:

$$F(\tau) = \int_0^\tau \bar{P}(\tau')d\tau' = a^* \int_0^\tau e^{-a^*\tau'}d\tau' = 1 - e^{-a^*\tau}$$

- Sei r_1 gleichverteilte Zufallszahl im Intervall $[0, 1]$
- Wählt man τ so, dass $F(\tau) = r_1$, ist die Wahrscheinlichkeitsdichte von τ die von $\bar{P}(\tau)$
- Man erhählt somit t durch

$$\tau = F^{-1}(r_1) = \frac{1}{a^*} \log\left(\frac{1}{1 - r_1}\right)$$

- Da r_1 genauso gleichverteilt ist wie $1 - r_1$, gilt für die Zufallsvariable der Zeit τ der nächsten Reaktion :

$$\tau = F^{-1}(r_1) = \frac{1}{a^*} \log\left(\frac{1}{r_1}\right) = \frac{1}{a^*} \ln r_1$$

- Welche ?
  - Sei r_2 eine gleichverteilte Zufallszahl in $[0, 1]$
  - Welche Reaktion stattfindet wird durch

$$\sum_{i=1}^{j-1} a_i \leq r_2 a^* < \sum_{i=1}^{j} a_i$$

    bestimmt

Bestimmung der Propensities a_i

- $c_i\, dt$: Wahrscheinlichkeit, dass eine bestimmte einzelne Reaktion R_i im nächsten Zeitschritt dt stattfindet.
- h_i: Anzahl Kombinationen der Reaktanten
- $a_i\, dt = h_i c_i\, dt$: Wahrscheinlichkeit der Reaktion R_i im nächsten Zeitschritt

- Beispiele

Reaktion R_i	c_i	h_i
$S_1 \xrightarrow{k} \dots$	k	$\#S_1$
$S_1 + S_2 \xrightarrow{k} \dots$	k/V	$\#S_1 \cdot \#S_2$
$2S_1 \xrightarrow{k} \dots$	$2k/V$	$\frac{1}{2}\#S_1 \cdot (\#S_1 - 1) = \begin{pmatrix} \#S_1 \\ 2 \end{pmatrix}$

Gillespie-Algorithmus:

1. Initialisierung
   - Setze $t = 0$
   - Wähle Anzahlen von Molekülen $\#S_i(0)$
2. Berechne Propensities
   - $a_i\,dt = h_i c_i\,dt$: Wahrscheinlichkeit der Reaktion R_i im nächsten Zeitschritt
   - Berechne $a^* = \sum_{i=1}^{M} a_i$
3. Ziehe zwei gleichverteilte Zufallszahlen r_1, r_2
   - Bestimme $\tau = -\frac{1}{a^*} \log r_1$
   - Bestimme j so, dass
   $$\sum_{i=1}^{j-1} a_i \leq r_2 a^* < \sum_{i=1}^{j} a_i$$
4. Update
   - Update der Anzahl der Moleküle nach dem Reaktionschema
   - Setze $t = t + \tau$
   - Gehe zu Punkt 2.

Übung:
Gillespie Algorithmus

Lessons learned:

- Runge-Kutta Integratoren für ODEs durch geschickte Funktionsauswertungen
- Steife Systeme brauchen implizite Integratoren
- Partielle DGLs, Kopplung von Δx und Δt
- Stochastische Differentialgleichungen, charakteristisches $\sqrt{h}$
- Gillespie-Algorithmus für die Chemische Master-Gleichung

12 Nichtparametrische Schätzer

12.1 Nichtparametrische Dichteschätzer

Literatur:

- B.W. Silverman. *Density Estimation* [65] Die Bibel

Aufgabe:

- Gegeben N Realisierungen x_i einer Zufallsvariablen X mit Dichte $\rho_X(x)$, schätze die Dichte.
- Parametrische Dichteschätzer
  - Bei Standardverteilungen wie Gauß, Exponential, χ^2_r schätze Parameter der Verteilung über Abgleich der Momente
  - Alternative: Fit an die kumulative Verteilung der Daten
- Nichtparametrische Dichteschätzer nehmen keine parametrisierte Verteilung an

Naivster Zugang: Histogramm

- Unterteile x-Achse in bins der Breite h ausgehend von Ankerpunkt x_0:

$$bin_m = [x_0 + mh, x_0 + (m+1)h], \qquad m \in \mathbb{Z}$$

- Schätze $\rho(x)$ durch

$$\hat{\rho}(x, x_0, h) = \frac{1}{Nh}(\text{Anzahl der } x_i \text{ in } bin_m) \tag{34}$$

- Problem 1: Wie den Ankerpunkt x_0 wählen ?
- Problem 2: Wie h wählen ?

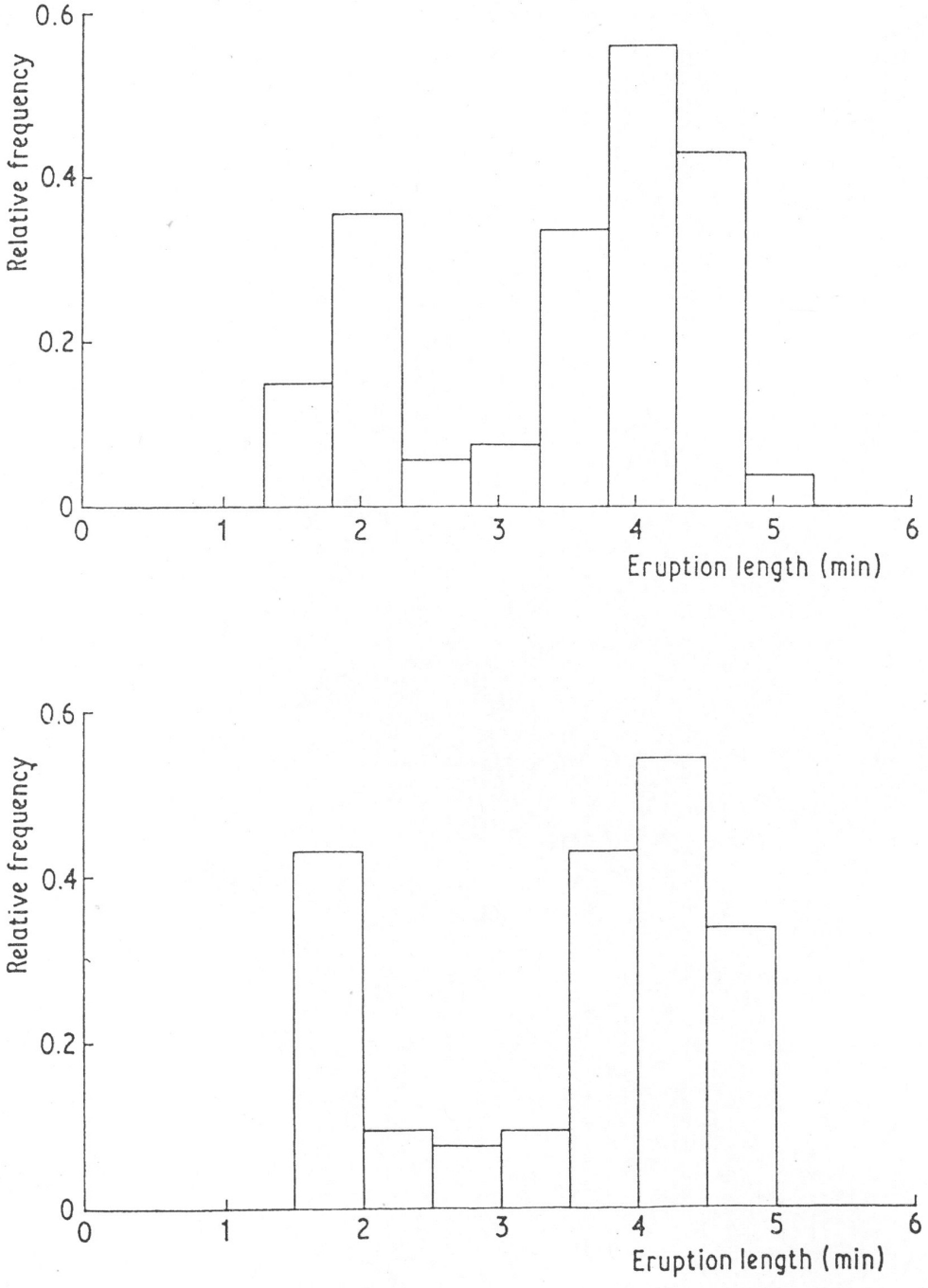


Abbildung 12.1: Histogramme der Eruptionslängen des Old Faithful Geysirs

Naiver Zugang:

- Ersetze Gl. (34) durch

$$\hat{\rho}(x,h) = \frac{1}{2Nh}(\text{Anzahl der } x_i \in [x-h, x+h])$$

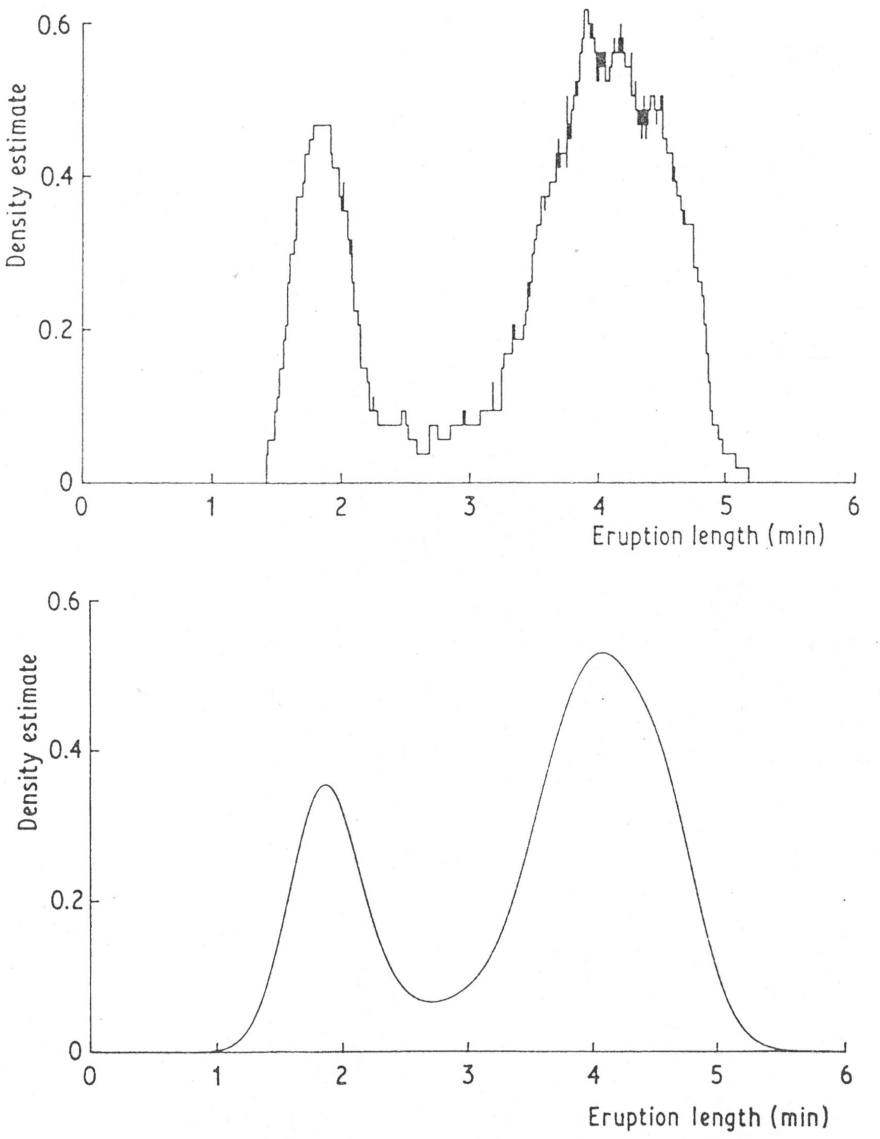


Abbildung 12.2: Kernschätzer für Daten des Old Faithful Geysirs

Kernschätzer (fixed size):

- Beachte, daß der naive Schätzer ausgedrückt werden kann durch:

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & sonst \end{cases}$$

$$\hat{\rho}(x,h) = \frac{1}{Nh} \sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right)$$

- Idee: Wähle statt Rechteckbox $w(x)$ eine glatte Funktion $K(x)$, K wie "Kern", die

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

  erfüllt und bis auf weiteres positiv ist.

$$\hat{\rho}_K(x,h) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

- Problem 2 : h ? bleibt
- "fixed size" erklären

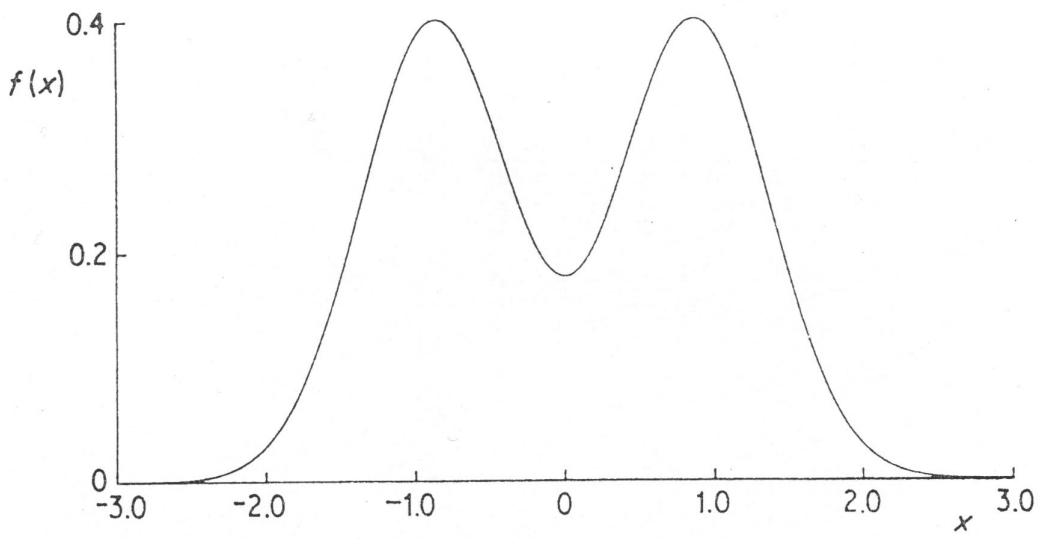


Abbildung 12.3: Wahre Dichte der Daten

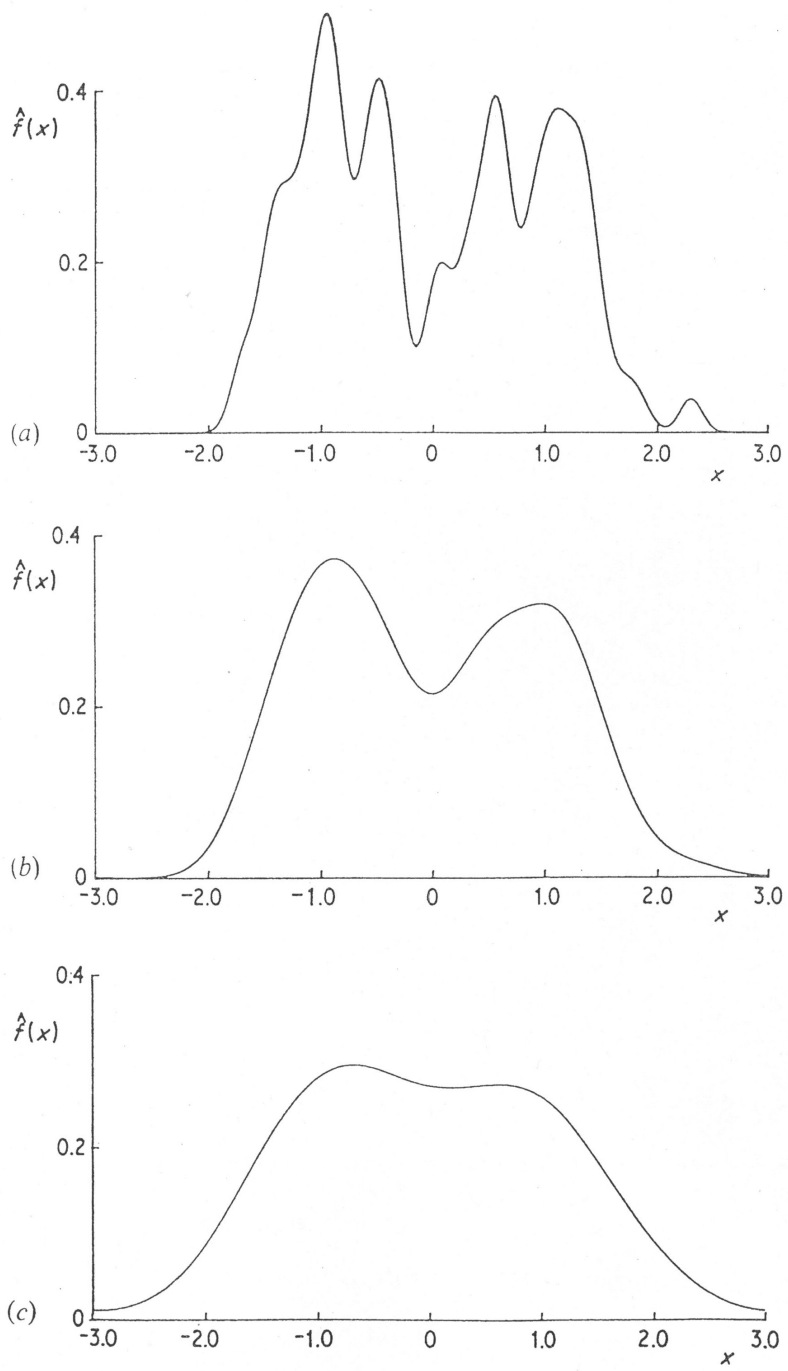


Abbildung 12.4: Geschätzte Dichte der Daten für 200 simulierte Datenpunkte mit (a) h=0.1; (b) h=0.3; (c) h=0.6

Nearest Neighbor method (fixed mass) :

- Idee:

  Wo viele Punkte liegen, wähle kleines h

- Wähle: Integer k

- Sei $d(x, x_i)$ der Abstand zwischen x und x_i

  Sortiere $d(x, x_i)$ aufsteigend: $d_1(x), d_2(x), \ldots, d_N(x)$

- und definiere den "k-th nearest neighbor" Schätzer:

$$\hat{\rho}_{NN}(x, k) = \frac{k}{2Nd_k(x)}$$

  Gleichung veranschaulichen. Aufgelöst ist $k = 2d_k(x)N\rho(x)$ die erwartete Anzahl.

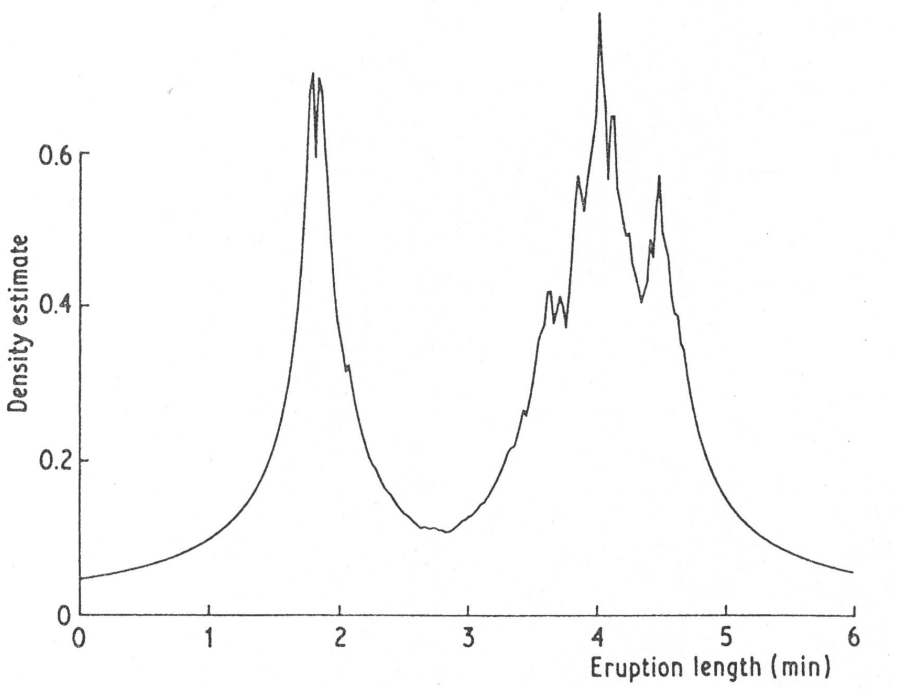


Abbildung 12.5: Nearest Neighbour-Schätzung für die Daten des Old Faithful Geysirs

- oder genereller:

$$\hat{\rho}_{NN}(x) = \frac{1}{N d_k(x)} \sum_{i=1}^{N} K\left(\frac{x - x_i}{d_k(x)}\right)$$

- Statt Problem 2, nun Problem 2a: Wahl von k ?
- "fixed mass" erklären

Mathematisierung (für Kernschätzer, für NN-Schätzer analog)

Annahmen:

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2 K(t)dt = k_2 \neq 0,$$

Berechnung des bias:

- Erwartungswert des Schätzers

$$\langle \hat{\rho}(x) \rangle = \frac{1}{Nh} \sum_{i=1}^{N} \left\langle K\left(\frac{x - x_i}{h}\right) \right\rangle = \frac{1}{h} \int K\left(\frac{x - y}{h}\right) \rho(y) dy$$

$$\begin{aligned} bias(x) &= \langle \hat{\rho}(x) \rangle - \rho(x) \\ &= \frac{1}{h} \int K\left(\frac{x - y}{h}\right) \rho(y) dy - \rho(x) \end{aligned}$$

- Variablentrafo: $y = x - ht$ und $\int K(t)dt = 1$:

$$\begin{aligned} bias(x) &= \int K(t)\rho(x - ht)dt - \rho(x) \\ &= \int K(t)(\rho(x - ht) - \rho(x))dt \end{aligned}$$

- Taylor-Entwicklung:

$$\rho(x - ht) = \rho(x) - ht\rho'(x) + \frac{1}{2}h^2 t^2 \rho''(x) + \ldots$$

$$\begin{aligned} bias(x) &= -h\rho'(x) \int tK(t)dt + \frac{1}{2}h^2 \rho''(x) \int t^2 K(t)dt + \ldots \\ &= \frac{1}{2}h^2 \rho''(x) k_2 + \mathcal{O}(h^3) \end{aligned}$$

Beobachtung:

- Bias hängt nicht von N ab.
- Nur von $\rho''(x)$ & h
- Anschaulich machen

Analoge Rechnung für die Varianz ergibt:

$$Var(\hat{\rho}(x)) = \frac{1}{Nh}\rho(x)\int K(t)^2 dt$$

- Varianz hängt von $\rho(x)$, N und h ab
  Zusammenhang mit Zählprozeß

Konsistenter Schätzer im Limes:

- $h \to 0$
- $Nh \to \infty$
- Ergo: h langsamer gegen 0 als N gegen ∞

Optimaler Kern

- Mean Square Error

$$MSE(\hat{\rho}(x)) = \langle(\hat{\rho}(x) - \rho(x))^2\rangle = bias(\hat{\rho}(x))^2 + Var(\hat{\rho}(x))$$

- Mean integrated square error

$$MISE(\hat{\rho}) = \int MSE(\hat{\rho}(x))dx$$

$$MISE(\hat{\rho}) = \frac{1}{4}h^4 k_2^2 \int \rho''(x)^2 dx + \frac{1}{Nh}\int K(t)^2 dt \tag{35}$$

- Minimierung des MISE bezüglich h: ergibt:

$$h_{opt} = k_2^{-2/5}\left(\int K(t)^2 dt\right)^2 \left(\int \rho''(x)dx\right)^{1/5} N^{-1/5} \tag{36}$$

  - Optimalerweise muß h skalieren mit $h \propto N^{-1/5}$,
  - Vorfaktor enthält leider Krümmung der wahren Dichte.

- Einsetzen von Gl. (36) in Gl. (35) ergibt:

$$MISE = \frac{5}{4} C(K) \left(\int \rho''(x) dx \right)^{1/5} N^{-4/5}$$

mit

$$C(K) = k_2^{2/5} \left(\int K(t)^2 dt \right)^{4/5}$$

Unter obigen Annahmen an $K(t)$ wird das minimiert durch Epanechnikow-Kern

$$K_{Ep}(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} t^2\right) & \text{if } -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{sonst} \end{cases}$$

- Effizienz eines Kernes K :

$$Eff(K) = C(K_{Ep}) / C(K)$$

Kern	K(t)	Effizienz
Dreieck	$1 - \lvert t \rvert$ für $\lvert t \rvert < 1$	0.986
Gauß	klar	0.951
Rechteck	$1/2$ für $\lvert t \rvert < 1$	0.930

Conclusio:

- Rechteck ist schlecht
- Gauß hat keinen endlichen Träger, auch schlecht
- Dreieck ist o.k.

Wahl von h: Crossvalidation

- Idee:

- Angenommen man hätte eine weitere Beobachtung x_{N+1}
- Dann wäre die log likelihood : $\mathcal{L}(h) = \log \hat{\rho}_h(x_{N+1})$, und man könnte sie in Abhängigkeit von h maximieren.
- Leider hat man keine, darum:

- Definiere den "leave-one-out" Schätzer:

$$\hat{\rho}_h^{-i}(x_i) := \frac{1}{(N-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right)$$

und die Cross-Validation Funktion $CV(h)$:

$$CV(h) := \frac{1}{N} \sum_{i=1}^{N} \log \hat{\rho}_h^{-i}(x_i)$$

- Bestimme "optimales" h durch maximales $CV(h)$.

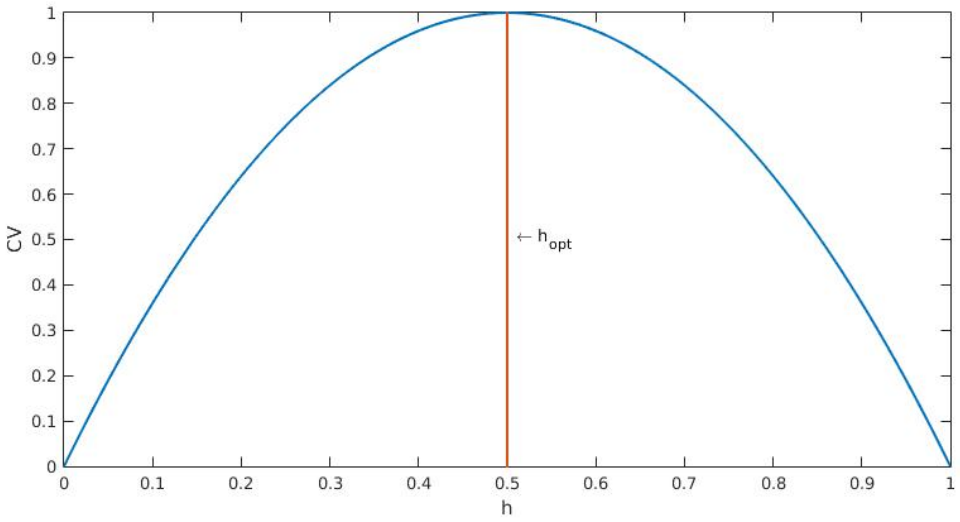


Abbildung 12.6: Gewünschtes Verhalten von $CV(h)$

- Es gibt viele andere heuristische Ideen, jede hat ihre Probleme.

12.2 Nichtparametrische Regression

Literatur:

- W. Härdle. *Applied Nonparametric Regression* [23]

Das Setting:

- Aufgabe:

  Gegeben N Realisierungen des Modells

  $$y = m(x) + \epsilon, \qquad ''m()'', \text{ weil das der } mean \text{ von } y \text{ ist}, \quad \epsilon \sim N(0, \sigma^2)$$

  Schätze $m(x)$ nichtparametrisch, i.e. ohne Annahme eines parametrisierten Modells wie in Kap. 6, auf Grund von Messungen (y_i, x_i).

- Ansatz, wiederum Kernschätzer:

  $$\hat{m}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i$$

  oder auch Nadaraya-Watson Kernschätzer:

  $$\hat{m}(x) = \frac{\sum_{i=1}^{N} K((x - x_i)/h)}{\sum_{i=1}^{N} K((x - x_i)/h)} y_i$$

  wegen Normierung.

- Mit

  $$\begin{aligned} c_K &= \int K^2(u)du \\ d_K &= \int u^2 K^2(u)du \end{aligned}$$

  gilt für den Mean Square Error analog zu oben:

  $$MSE(x) = \frac{\sigma^2 c_K}{Nh} + h^4 d_K^2 \frac{m''(x)^2}{4}$$

- Wie bei Dichte-Kernschätzern:

  Breite des Kerns kontrolliert den trade-off zwischen bias und Varianz

- Wiederum konsistenter Schätzer für $h \to 0$, $Nh \to \infty$

- Bei Dichte-Kernschätzern waren positive Kerne natürlich, das ist hier nicht mehr notwendig, siehe unten.

- Zur Wahl von h, siehe Übung.

Äquivalenz von Kernschätzern und lokaler nichtlinearer Regression

- Betrachte Rechteckskern[3]:

$$K(u) = \begin{cases} \frac{1}{2h} & \text{if } |u| < h \\ 0 & sonst \end{cases}$$

  Betrachte für festes x:

$$\frac{1}{N} \min_{a,b} \sum_{i=1}^{N} K(x - x_i)(y_i - a - b(x - x_i)^2)^2$$

  den lokalen Parabelfit an das durch uniformen Kern festgelegte Interval.

  Ergebnis wird sein:

$$\hat{m}(x) = \hat{a}$$

- Die Normalengleichungen (remember Kap. 10.2) lauten

$$\frac{\partial}{\partial a}: \qquad \frac{1}{N} \sum_i K(x - x_i)(y_i - \hat{a} - \hat{b}(x - x_i)^2) = 0$$

$$\frac{\partial}{\partial b}: \qquad \frac{1}{N} \sum_i K(x - x_i)(y_i - \hat{a} - \hat{b}(x - x_i)^2)(x_i - x)^2 = 0$$

- Definiere

$$\tilde{y}(x) := \sum_i K(x - x_i) y_i$$

[3]Etwas anders definiert als sonst, um ständiges Teilen durch h zu vermeiden

Nehme an x_i sei gleich-verteilt und beachte:

$$\frac{1}{N}\sum_i K(x-x_i) \approx 1$$

- Nähere

$$\frac{1}{N}\sum_i K(x-x_i)(x-x_i)^2 \approx \int_{-\infty}^{\infty} K(x-u)(x-u)^2 du = \int_{-1/2h}^{1/2h} (x-u)^2 du = h^3/3$$

Analog

$$\frac{1}{N}\sum_i K(x-x_i)(x-x_i)^4 \approx \int_{-\infty}^{\infty} K(x-u)(x-u)^4 du = \int_{-1/2h}^{1/2h} (x-u)^4 du = h^5/5$$

- Mit

$$A = \frac{1}{N}\sum_i K(x-x_i)(x-x_i)^2 y_i$$

lauten die Normalengleichungen dann:

$$\begin{aligned} 0 &= \tilde{y} - \hat{a} - \frac{h^3}{3}\hat{b} \\ 0 &= A - \frac{h^3}{3}\hat{a} - \frac{h^5}{5}\hat{b} \end{aligned}$$

Führt für $\hat{a}$ auf:

$$0 = 3h^2\tilde{y} - 5A + \frac{4}{3}h^2\hat{a}$$

- Alles eingesetzt:

$$\hat{a} = \frac{3}{4N}\sum_i K(x-x_i)\left(3 - 5\left(\frac{(x-x_i}{h}\right)^2\right) y_i$$

- Scharfes Hinsehen zeigt:

$$\hat{m}(x) = \hat{a} = \frac{1}{N} \sum_i K^*(x - x_i) y_i$$

mit

$$K^*(u) = \begin{cases} 3/8(3 - 5(u/h)^2 & \text{if } |u| < h \\ 0 & sonst \end{cases}$$

einem parabolischem Kern.

- Umgedreht: Parabolischer Kern entspricht lokalem Parabel-Fit

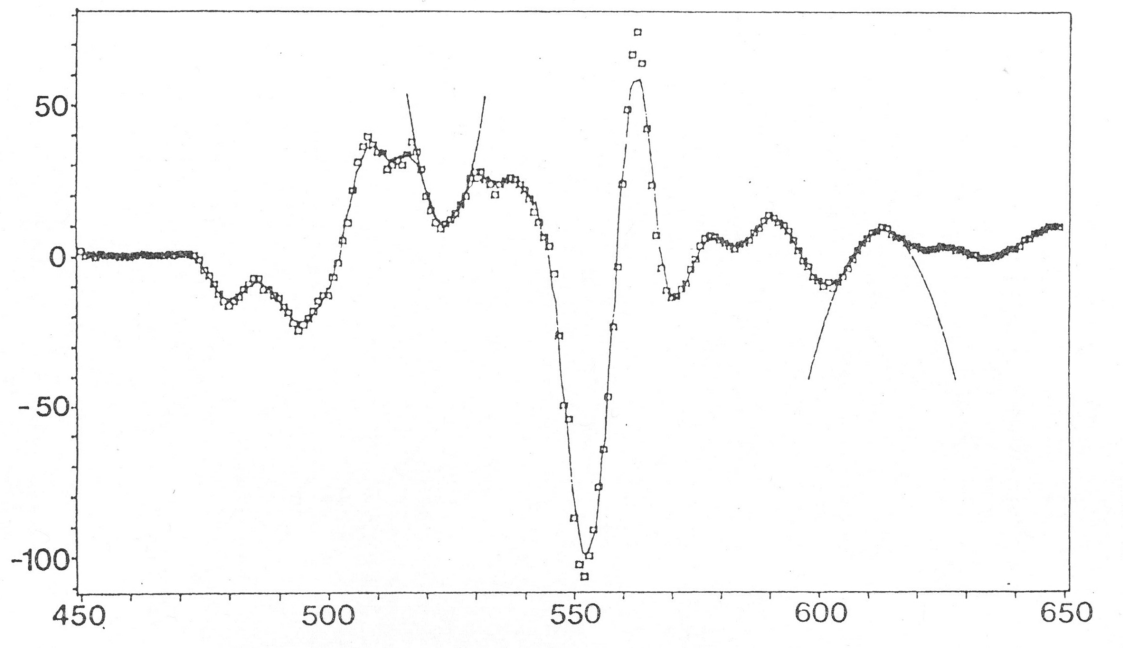


Abbildung 12.7: Lokale Parabel-Fits im Vergleich zum Kernschätzer

- Für andere Kerne höhere Ordnungen
- Merke: Nichtparametrische Regression = Parametrisch mit vielen Parametern

Savitzky-Golay - Filter

- Idee: Drehe es herum

  Bestimmung der Kernkoeffizienten aus Polynomfit

  Daten seien äquidistant, $\Delta x = 1$.

$$\hat{m}(x_i) = \sum_{j=-h}^{h} c_j y_{i+j}$$

- Wähle c_j so, daß es einem Polynomfit mit

$$y_i = a_0 + a_1 i + a_2 i^2 + \ldots a_M i^M$$

  an die Daten $y = (y_{i-h}, \ldots, y_{i+h})$ entspricht, dann analog zu oben:

$$\hat{m}(x_i) = \hat{a}_0$$

- Erinnere Kap. 6 Nichtlineare Regression

  Die Designmatrix A lautet:

$$A_{il} = i^l$$

  und die Normalengleichungen führen auf:

$$A^T A a = A^T y \text{ oder } a = (A^T A)^{-1} A^T y$$

  Praktisch: Koeffizienten a sind linear in den Daten.

- Somit ergibt sich c_j als a_0, wenn man y durch die Einheitsvektoren e_j ersetzt:

$$c_j = \{(A^T A)^{-1} A^T e_j\}_0 = \sum_{m=0}^{M} \left\{(A^T A)^{-1}\right\}_{0m} j^m$$

  Für M=2, h=2, lauten Koeffizienten:
  $-.0086, 0.343, 0.486, 0.343, -.0086$ und sind nicht positiv.

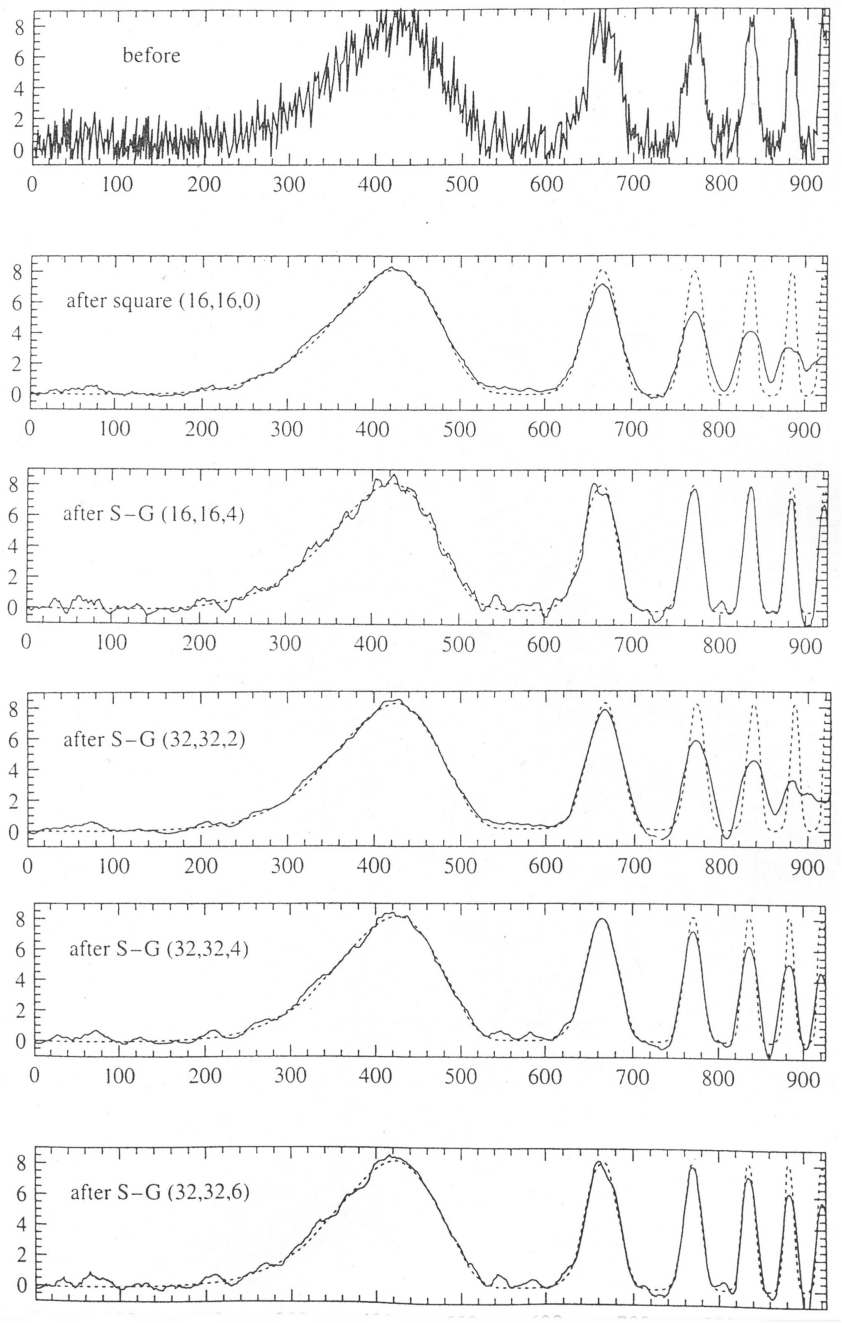


Abbildung 12.8: (a) Verrauschte Daten; (b) Gefittet ohne S-G-Filter, 16 Punkte links und rechts; (c) mit S-G-Filter vom Grad 4, 16 Punkte links und rechts; (d) mit S-G-Filter vom Grad 2, 32 Punkte links und rechts; (e) mit S-G-Filter vom Grad 4, 32 Punkte links und rechts; (f) mit S-G-Filter vom Grad 6, 32 Punkte links und rechts

Schätzung von Ableitungen
Macht man dasselbe wie eben für a_1, erhählt man eine Schätzung der 1. Ableitung, usw.

Spline smoothing

- Paßt man eine Funktion $g(x)$ mit vielen Freiheitsgraden, z.B. Polynom sehr hoher Ordnung, mittels least squares

$$a = argmin \sum_i (y_i - g(x_i, a))^2$$

  an Daten y_i an, wird $g(x, a)$ die Daten interpolieren und lokal sehr variabel sein.

- Idee:

  Fordere eine gewisse Glattheit von $g(x, a)$.

  Glattheit kann man messen durch:

$$\int (g''(x, a))^2 dx$$

- Definiere, erinnere Regularisierung, Kap. 4.3

$$S_\lambda(g) = \sum_{i=1}^{N} (y_i - g(x_i, a))^2 + \lambda \int (g''(x, a))^2 dx$$

- Beachte
  - $\lambda = 0$: Interpolation
  - $\lambda = \infty$: Lineare Regression

- Minimierung von $S_\lambda(g, a)$ über alle zweimal differenzierbaren Funktionen hat eine eindeutige Lösung:

  $\hat{m}_\lambda(x)$ ist:
  - kubisches Polynom zwischen aufeinanderfolgenden x_i-Werten.
  - stetig an den x_i-Werten

- 1. und 2. Ableitung stetig, 3. Ableitung unstetig.
- 2. Ableitung = 0 an x_1 und x_N.

- Heißt Spline: "längliches dünnes Holz oder Metall" (Langenscheidt)
  Biegt man das, ist es schön glatt.
- Kennt man Fehler auf den Daten, kann man λ fixieren.
- Läßt sich (unschön) auch als Kernschätzer formulieren. λ wird zu h. Bestimmung durch Cross-Validation

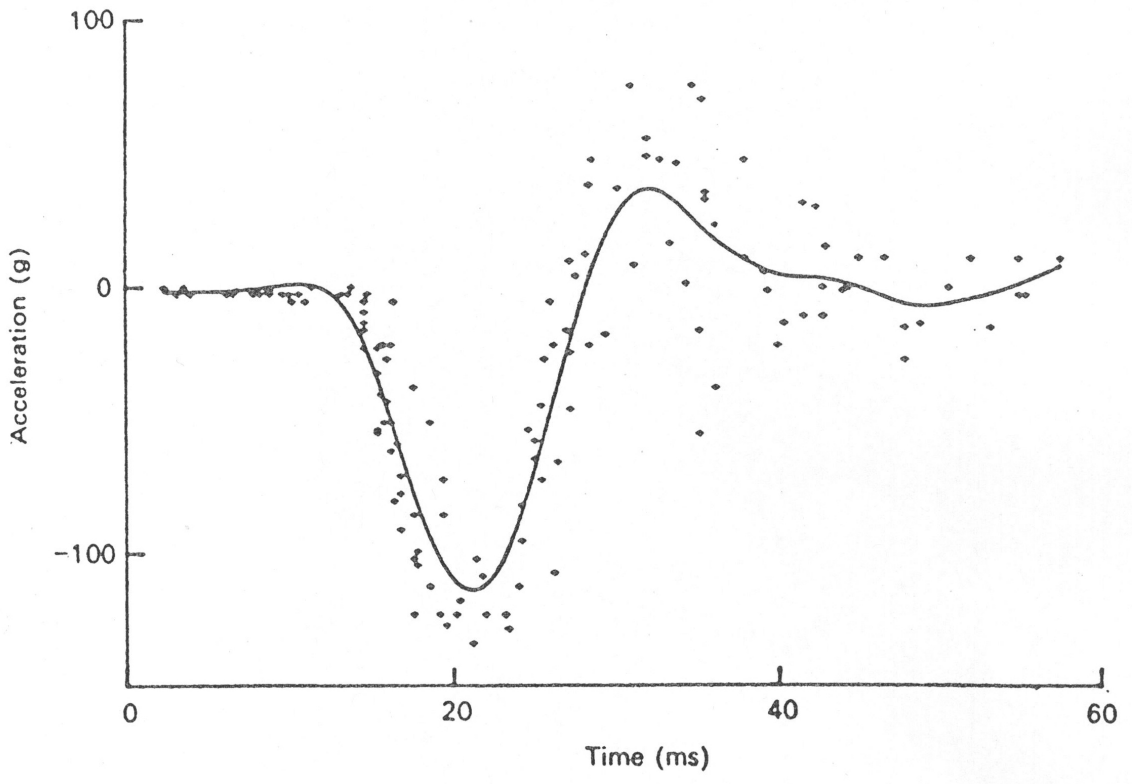


Abbildung 12.9: Spline smoothing eines Datensatzes

Robust smoothing
Sind die Fehler nicht Gauß'sch, kann der Median-Filter:

$$\hat{m}_M(x) = \text{med}\{y_i\}, \quad \{y_i | x_i \in [x-h, x+h]\}$$

von Nutzen sein, z.B. in Rauschunterdrückung in S/W-Bildern.

Der Fluch der hohen Dimension

Verteilt man N Punkte äquidistant im d-dimensionalen Einheitskubus $[0,1]^D$, so beträgt der Abstand $dist_{NN}$ zwischen 2 benachbarten Punkten:

$$dist_{NN} = N^{-1/D}$$

Beispiel $N = 10000$:

D	$dist_{NN}$
1	1/10000
2	1/100
3	1/21.54
4	1/10
5	1/6.31=0.16
10	1/2.51=0.4

I.e. in 10 Dimensionen hat jeder Punkt in jeder Richtung 2.5 Nachbarn, also i.w. keine.

Übung: Crossvalidation

12. Woche

Lessons learned:

- Nichtparametrische Dichteschätzung: Kernschätzer und Nearest-Neighbor-Schätzer
- Bias und Varianz der Schätzer
- Optimaler Kern und optimales h
- Nichtparametrische Regression = Parametrisch mit vielen Parametern
- Savitzky-Golay - Filter & Splines

13 Spektralanalyse

Literatur:

- M.B. Priestley *Spectral Analysis and Time Series* [51].
  Der mathematische Klassiker No. 1
- P.J. Brockwell, R.A. Davis *Time Series: Theory and Methods* [9].
  Der mathematische Klassiker No. 2
- J. Honerkamp *Stochastic Dynamical Systems* [25] Kap. 13.3
  Kondensierte Darstellung für Physiker

Definition Autokovarianzfunktion:
Sei $x(t)$ ein stationärer Prozeß mit $\langle x(t)\rangle = 0$, dann ist die Autokovarianzfunktion:

$$ACF(\tau) = \langle x(t)x(t+\tau)\rangle$$

Definition Spektrum:

$$S(\omega) = \int d\tau\, e^{-i\omega\tau} ACF(\tau) = \langle |f(\omega)|^2\rangle$$

mit

$$f(\omega) = \int dt\, e^{-i\omega t} x(t)$$

In Analogie zu weißes Rauschen vs. Wiener Prozessist die erste Formulierung bei Mathematikern bevorzugt, die zweite bei Physikern.

Die Fouriertransformation ist orthogonal (alle Eigenwerte Betrag 1), d.h.:

$$\int S(\omega)d\omega = \mathrm{Var}(x(t))$$

Spektrum ist "Varianz pro Frequenz" Darstellung des Prozesses.

Zeit-diskrete Prozesse:
Betrachte:

$$x(i) = ax(i-1) + \sigma\epsilon(i), \qquad 0 < a < 1, \quad \epsilon(i) \sim N(0,1)$$

- Wenn $\sigma = 0$

$$x(i) = x(0)e^{-i/\tau}$$

ein Relaxator mit $\tau = -1/\log a$

- $\sigma \neq 0$: Prozeß wird vom Rauschen immer wieder aus der Gleichgewichtslage um 0 weggetrieben.

Physikalisch: Stochastisch getriebener Relaxator

- Prozeß heißt Autoregressiver Prozeß erster Ordnung, AR[1].

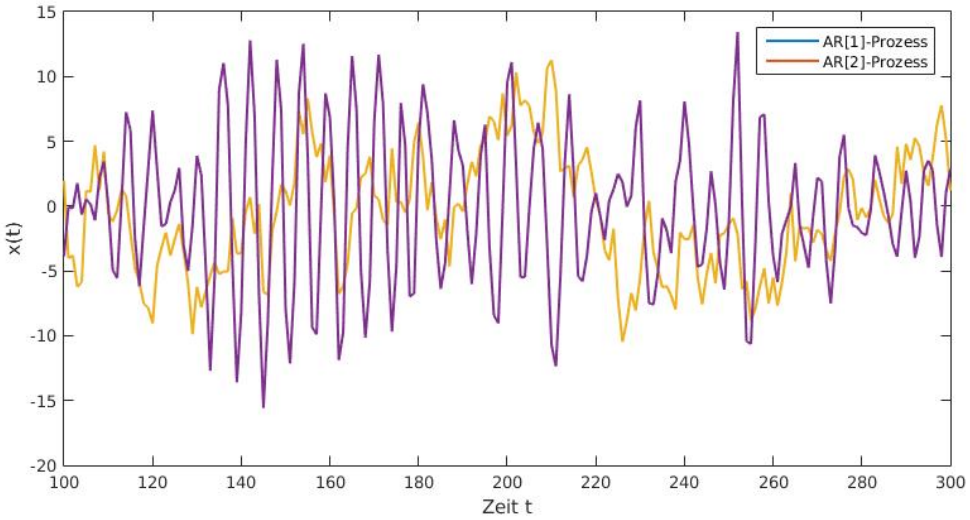


Abbildung 13.1: Realisierung linearer stochastischer Prozesse 1. und 2. Ordnung

- AR[2] Prozeß:

$$x(i) = a_1 x(i-1) + a_2 x(i-2) + \epsilon(i)$$

ergibt mit:

$$\begin{aligned} a_1 &= 2\cos(2\pi/T)e^{-1/\tau} \\ a_2 &= -e^{-2/\tau} \end{aligned}$$

einen stochastisch getriebenen, gedämpften Oszillator der Periode T und Relaxationszeit τ.

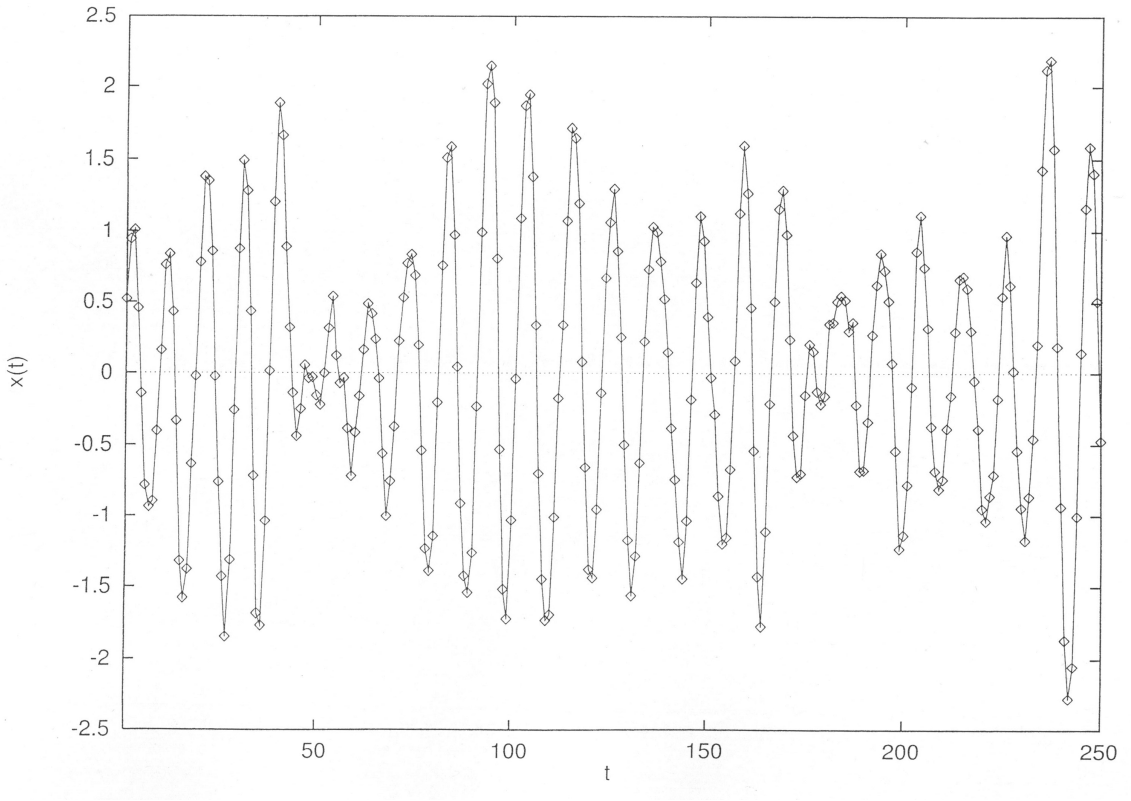


Abbildung 13.2: Realisierung eines linearen stochastischen Prozesses 2. Ordnung

13.1 Spektren von AR[p] Prozessen

- Definiere Backshift-Operator:

$$B(x(t)) = x(t-1)$$

- Sei

$$f(\omega) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} e^{-i\omega t} x(t)$$

(Normierung fehlt ab jetzt) dann ist

$$\sum_{t=1}^{N} e^{-i\omega t} B(x(t)) = \sum_{t=1}^{N} e^{-i\omega t} x(t-1) = e^{-i\omega} \sum_{t=1}^{N} e^{-i\omega t} x(t) = e^{-i\omega} f(\omega)$$

Allgemein:

$$\sum_{t=1}^{N} e^{-i\omega t} B^d(x(t)) = e^{-id\omega} f(\omega)$$

- AR[p] Prozeß:

$$x(t) - \sum_{j=1}^{p} a_j B^j(x(t)) = \epsilon(t)$$

- Fouriertrafo:

$$f(\omega)(1 - \sum_{j=1}^{p} a_j e^{-ij\omega}) = \tilde{\epsilon}$$

- Spektrum:

$$S(\omega) = \langle |f(\omega)|^2 \rangle = \frac{1}{2\pi} \frac{\sigma^2}{|1 - \sum_{j=1}^{p} a_j e^{-ij\omega}|^2}$$

Wichtig:
Spektrum von AR[p]-Prozess ist glatt.

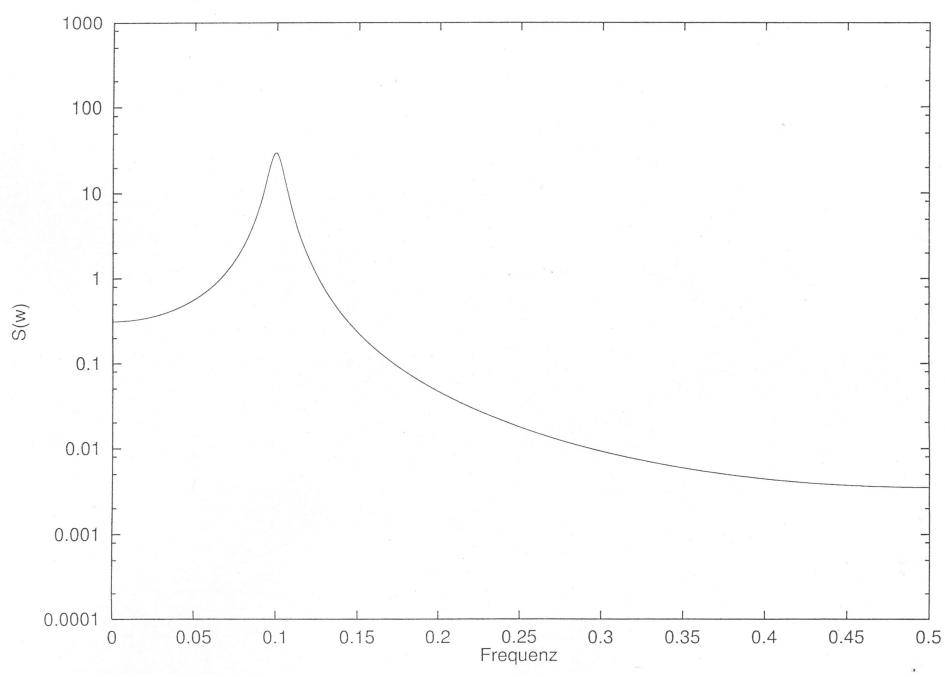


Abbildung 13.3: Spektrum eines linearen stochastischen Prozesses 2. Ordnung

Erinnere AP I: Getriebener gedämpfter Oszillator
Dort Anregung mit einer Frequenz nach der anderen. Hier alle gleichzeitig. Superpositionsprinzip.

Glattheit gilt auch

- für alle nichtlinearen stochastischen Prozesse.
- i.w. für alle chaotischen Prozesse.

Allgemein: Immer, wenn die ACF zerfällt, i.e. der Prozess vergeßlich, mischend ist.
An Faltung erklären.
Glattheit gilt nicht für

- integrable Systeme
- Grenzzyklen

13.2 Fast Fourier Transform (FFT)

Cooley & Tukey, 1965 [11].

- Die Berechnung der Fourier-Transformierten

$$f(\omega_k) = \sum_{t=0}^{N-1} e^{-i\omega_k t} x(t)$$

für alle Fourierfrequenzen

$$\omega_k = \frac{2\pi k}{N}, \quad k = -N/2 \ldots, 0, \ldots, N/2$$

ist von der Komplexität $\mathcal{O}(N^2)$.

- Für $x(t)$ reell:

$$f(\omega_k) = f^*(-\omega_k)$$

Freiheitsgrade abzählen.

- Divide and Conquer - Strategie

Sei $N = 2^n$

$$\begin{aligned} f(\omega_k) = f_k &= \sum_{t=0}^{N-1} e^{-i\omega_k t} x(t) \\ &= \sum_{t=0}^{N/2-1} e^{-i\omega_k (2t)} x(2t) + \sum_{t=0}^{N/2-1} e^{-i\omega_k (2t+1)} x(2t+1) \\ &= \sum_{t=0}^{N/2-1} e^{-i2\omega_k t} x(2t) + e^{-i\omega_k} \sum_{t=0}^{N/2+1} e^{-i2\omega_k t} x(2t+1) \\ &= f_k^e + e^{-i\omega_k} f_k^o \quad e \text{ wie even} \quad o \text{ wie odd} \end{aligned} \tag{37}$$

- f_k^e und f_k^o periodisch in k mit Periode $N/2$

- Für f_k^e und f_k^o kann man Zerlegung wiederholen.
  Ergibt : f_k^{ee}, f_k^{eo}, f_k^{oe} und f_k^{oo}.
  Effektive FT-Länge jeweils $N/4$, der Rest ist periodisch.
- Iteriere dies, bis die Länge der Fourier-Trafo = 1 ist.
- Aber

$$\sum_{t=0}^{0} e^{i\omega t} x(0) = x(0)$$

  D.h., es gibt Darstellung:

$$f_k^{eooeeeo..oe} = f^{eooeeeo..oe} = x(t) \quad \forall\, t \tag{38}$$

  hängt nicht von k ab, da periodisch in k mit Periode 1.
- Länge der Kette *eooeeeo..oe*: $\log_2 N$
- Nun zentraler Punkt: Bitreversal:
  - Welche Folge von e's und o's gehört zu welchem t
  - Drehe die Ordnung der e's und o's um
  - Ersetze die Folge *eo..oeeeooe* mit $e = 0$ und $o = 1$
  - Ergibt die binary-Darstellung für jedes t.
  - FFT Algorithmus Gl. (37)
  - even/odd-Zerlegung bitreversed baut von unten binary Darstellung auf
  - Beispiel 4

	00 0	01 1	10 2	11 3	BR	binary
ee	x				ee	00
eo			x		oe	10
oe		x			eo	01
oo				x	oo	11

  - Beispiel 8:

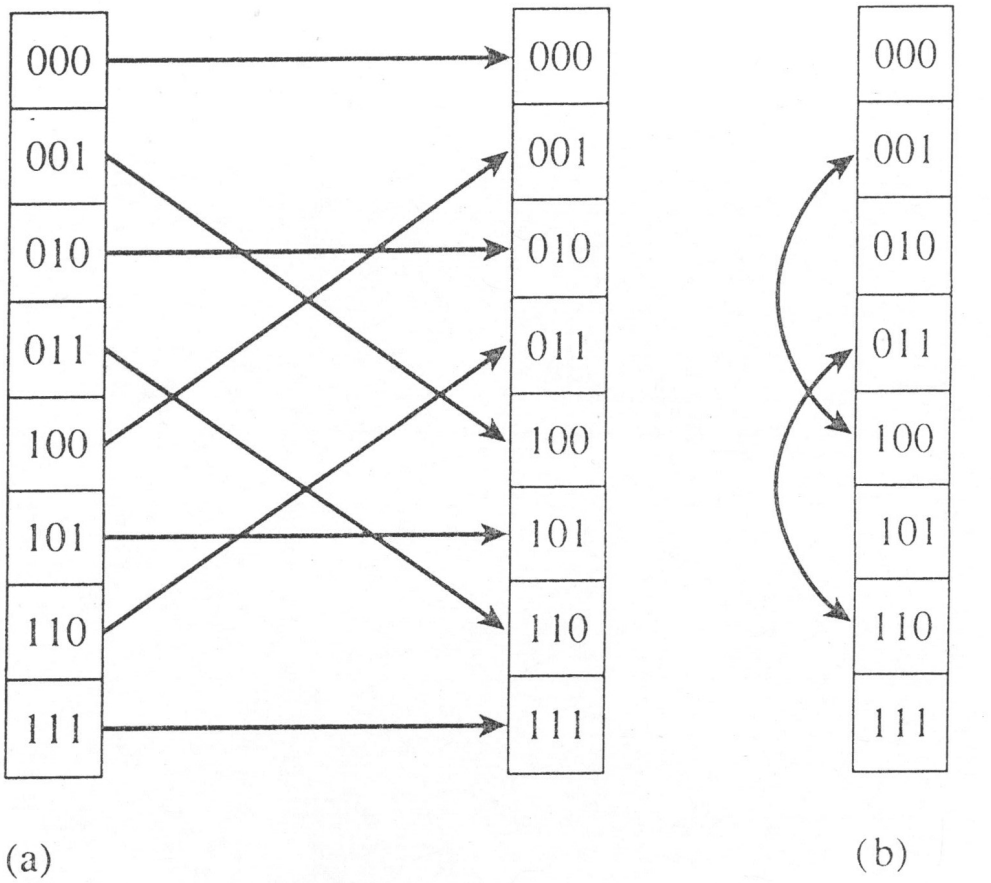


Abbildung 13.4: Umordnung eines Arrays durch bit reversal, (a) zwischen zwei Arrays und (b) in einem Array

  - Beachte: Braucht nur Paar-Vertauschung

- Startpunkt für die Invertierung von Gl. (38) mittels $\log N$-facher Anwendung von Gl. (37).

- Buchhaltung sehr einfach:

  - Sortiere Daten in bit-reversed Ordnung: Ein-Punkt-Transformation
  - Kombiniere benachbarte Punkte
    * Zwei-Punkt-Trafo

* Beispiel $N = 64$, $n = 6$

$$f_k^{eooee} = f^{eooeee} + e^{-i\omega_k} f^{eooeeo}$$

* Ergebnis braucht 2 Punkte Platz
* Komplexität: $\mathcal{O}(N)$

– Kombiniere benachbarte Punkte-Paare

* 4-Punkt-trafo
* Braucht 4 Punkte Platz
* Komplexität: $\mathcal{O}(N)$

– Kombiniere benachbarte Quadrupel

– usw.

- Von N Datenpunkten ausgehend $\log N$ mal Gl. (37), gibt Aufwand $\mathcal{O}(N \log N)$
- Es gibt FFTs auch für $N = 2^n\, 3^k\, 5^l$
- Die selbe Teile-und-Herrsche Idee ist auch in anderen Situationen anwendbar.

Einer der fünf wichtigsten Algorithmen, wo gibt

13.3 Spektralanalyse zeitdiskreter Prozesse

Ein Schätzer $\hat{\Theta}_N$ basierend auf N Daten heißt konsistent, wenn gilt

$$\lim_{N\to\infty} (\hat{\Theta}_N - \Theta) \to_{prob} 0$$

i.e. bias und Variance gehen mit N gegen 0.

Periodogramm von weißem Rauschen

- Sei $x(t) = \epsilon(t) \sim N(0, \sigma^2)$
- Dann lautet $f(\omega_k)$

$$f(\omega_k) = \frac{1}{N} \sum_{t=1}^{N} e^{-i\omega_k t} \epsilon(t) = \frac{1}{N} \sum_{t=1}^{N} (\cos(\omega_k t) + \sin(\omega_k t))\, \epsilon(t)$$

- Mit

$$|e^{-i\omega_k t}| = 1 \text{ und } \langle \epsilon(t_i)\epsilon(t_j)\rangle = \delta_{ij}$$

folgt:

$$f(\omega_k) \sim N_C(0, \sigma^2)$$

  - unabhängig von ω_k (darum "weißes" Rauschen)
  - mit unabhängigem Real- und Imaginärteil (weil $\sin(\omega_k t)$ und $\cos(\omega_k t)$ orthogonal)
  -

$$\langle f(\omega_k), f(\omega_l)\rangle = \sigma^2 \delta_{kl}$$

$\underline{f(\omega_k) \text{ ist unabhängig komplex normal verteilt.}}$

- Spektrum war

$$S(\omega) = \langle |f(\omega)|^2 \rangle$$

- $|f(\omega)|^2$ hat extra Namen: $\underline{\text{Periodogramm}}$

$$Per(\omega_k) = |f(\omega)|^2$$

- Da

$$Per(\omega_k) = |f(\omega_k)|^2 = (\mathrm{Re}(f(\omega_k)))^2 + (\mathrm{Im}(f(\omega_k)))^2$$

gilt für $x(t) = \epsilon(t)$:

$$Per(\omega_k) \sim \chi_2^2$$

Für nichtweiße (i.a. nichtlineare) Prozesse hilft der Zentrale Grenzwertsatz, und es gilt allgemein (mit richtigem Vorfaktor) :

$$Per(\omega_k) \sim \frac{1}{2} S(\omega_k)\chi_2^2, \qquad \omega_k \neq 0, \pi$$

unabhängig von N. (Für $\omega_k = 0, \pi$: $Per(\omega_k) \sim S(\omega_k)\chi_1^2$, da nur $\cos(\omega_k t)$ beiträgt).

- Da
$$\langle \chi^2_2 \rangle = 2, \qquad Var(\chi^2_2) = 4 \qquad SD(\chi^2_2) = 2$$
ist die Periodogramm unverzerrter Schätzer,

- aber : Standardabweichung des Periodogramms ist unabhängig von N (und gleich Erwartungswert)

Damit ist das Periodogramm kein konsistenter Schätzer für das Spektrum !

- Zunehmende Datenmenge:
Statt kleiner werdender Varianz des Schätzers erhält man bessere Auflösung im Frequenzraum.

Zentral:
Weil das (wahre) Spektrum glatt, kann man Spektren schätzen, in dem man das Periodogramm glättet:

$$\hat{S}(\omega_k) = \sum_{l=-h}^{h} W_l Per(\omega_{k+l})$$

Dieses ergibt mit $N \to \infty$ $h \to \infty$, und $h/N \to 0$ einen konsistenten Schätzer.

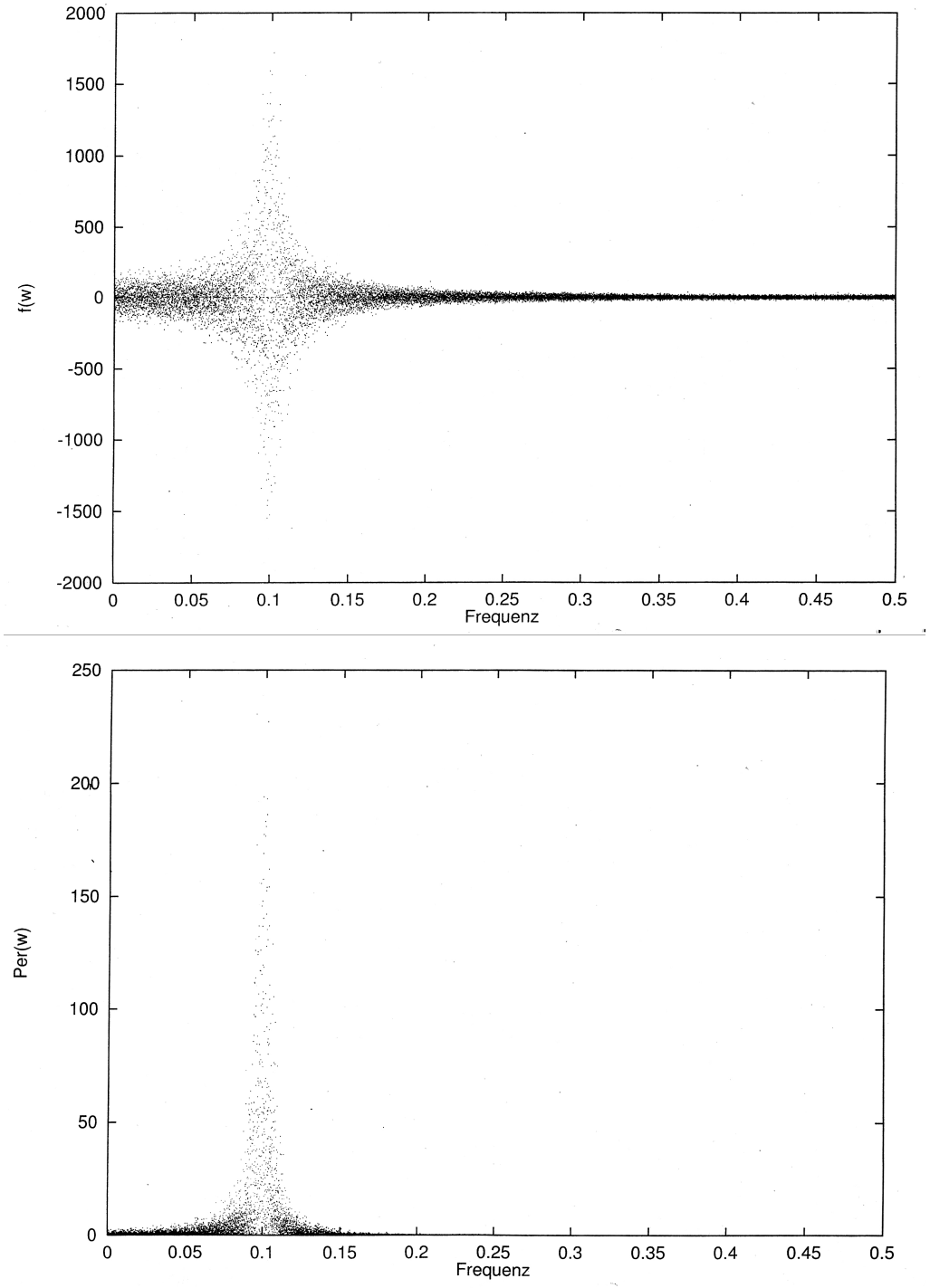


Abbildung 13.5: Linearer stochastischer Prozess 2. Ordnung

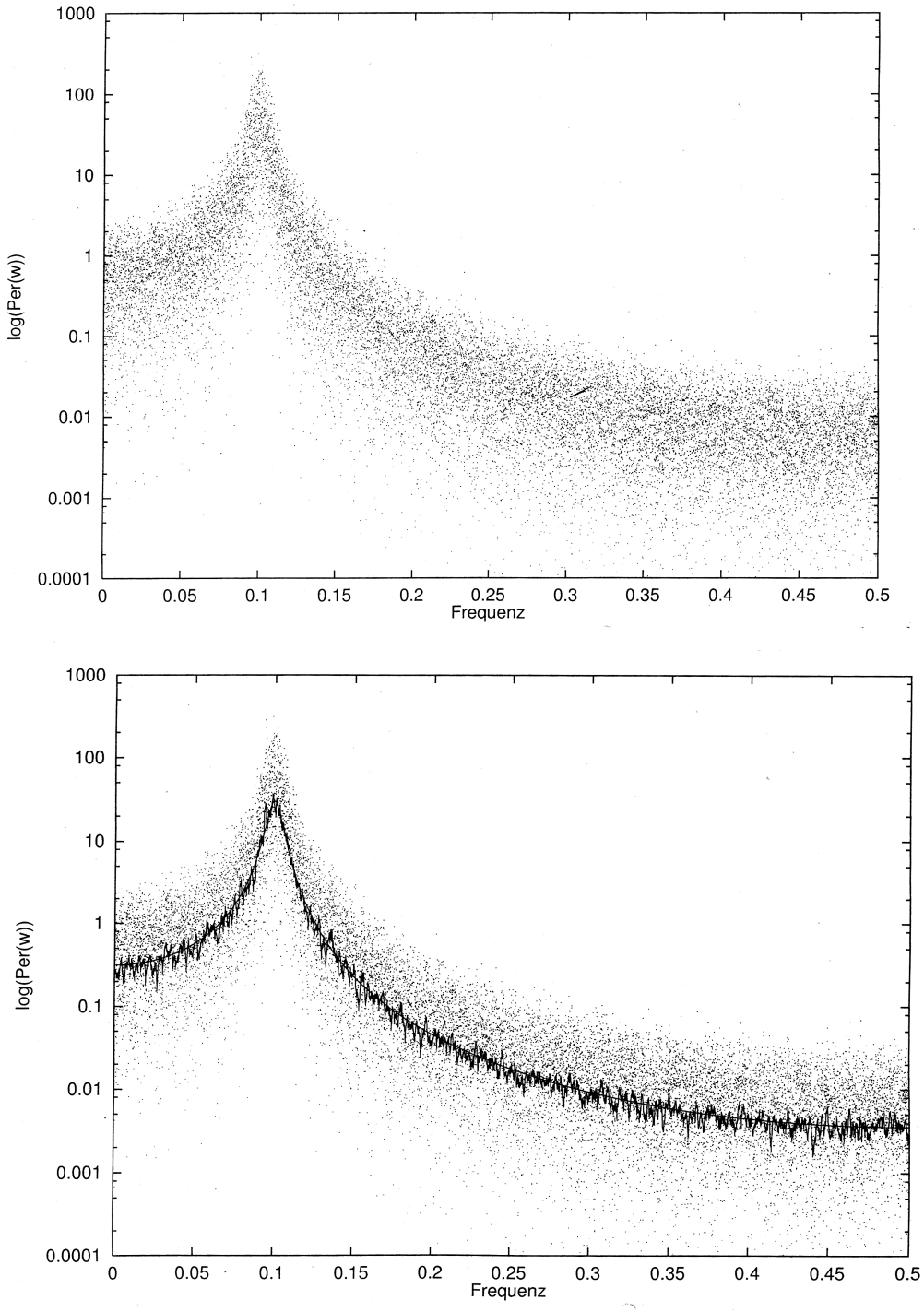


Abbildung 13.6: Linearer stochastischer Prozess 2. Ordnung

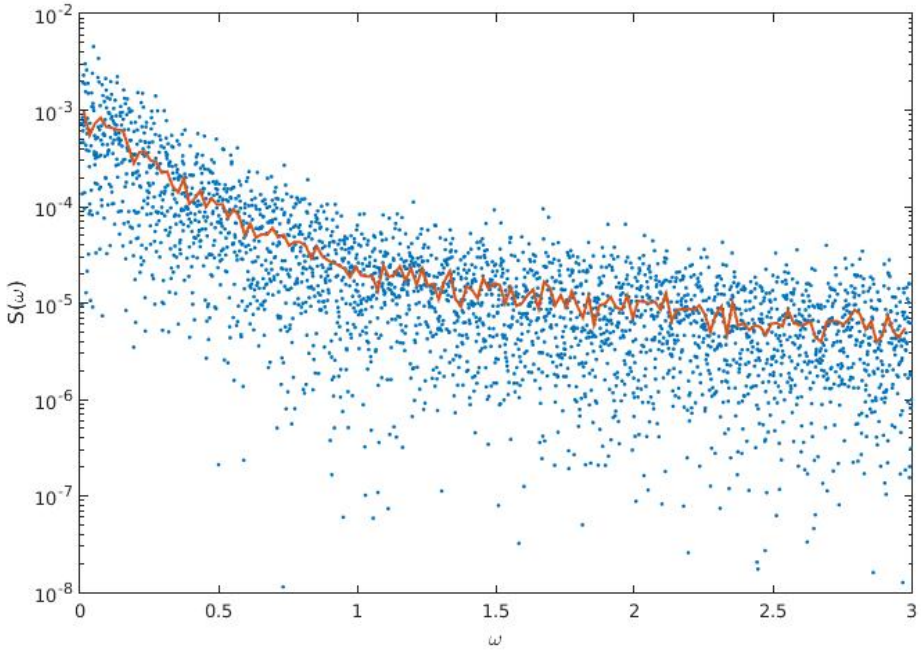


Abbildung 13.7: Periodogamm und geschätztes Spektrum eines linearen stochastischen Prozesses 1. Ordnung

Spektrum war definiert als:

$$S(\omega) = \sum e^{-i\omega\tau} ACF(\tau) =< |f(\omega)|^2 >$$

Wählt man geschätzte $\widehat{ACF}$

$$\widehat{ACF}(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} x(t)x(t+\tau)$$

ergibt sich als Schätzer für $S(\omega)$

$$\hat{S}(\omega) = \sum e^{-i\omega\tau} \widehat{ACF}(\tau) = |f(\omega)|^2$$

Periodogramm $|f(\omega)|^2$ zappelt wie blöde, $\widehat{ACF}(\tau)$ ist schön glatt. Wie paßt das zusammen ?

ZEICHNUNG wahre ACF

ZEICHNUNG geschätzte ACF

$$< (\widehat{ACF}(\tau_1) - ACF(\tau_1))(\widehat{ACF}(\tau_2) - ACF(\tau_2)) >=$$
$$\frac{1}{N}\sum_k \{ACF(k+\tau_1) + ACF(k-\tau_1) - 2ACF(\tau_1)ACF(k)\} +$$
$$\{ACF(k+\tau_2) + ACF(k-\tau_2) - 2ACF(\tau_2)ACF(k)\}$$

Die Zappelei des Periodogramms steckt in den korrelierten Fehlern der geschätzen ACF
Fouriertransformation macht sie im Frequenzraum unabhängig.

Andere Methoden der Spektralschätzung:

- Zeitreihe in L Stücke schneiden und deren Periodogramme mitteln
  Sei $M = N/L$
  $$Per_l(\omega_k) = |\sum_{t=1}^{M} e^{-i\omega_k t} x((l-1)M + t)|^2$$

  $$\hat{S}(\omega_k) = \frac{1}{L}\sum_{l=1}^{L} Per_l(\omega_k)$$

- ACF Fenstern, Erinnere QM: Faltung im Frequenzraum ist Multiplikation im Zeitraum und vice versa.

  $$\hat{S}(\omega_k) = \sum_{\tau=1}^{N} w(\tau) e^{-i\omega_k \tau} \widehat{ACF}(\tau)$$

  $w(\tau) = 0$ für $\tau > \tau_{max}$. $\tau_{max} \propto 1/h$. Methode der Wahl vor FFT Erfindung.

Im Falle

- linearer Prozesse bleiben die Fourierkomponenten unabhängig
- nichtlinearer Prozesse ergeben sich Korrelationen.
- Siehe next semester for details

Vergleich Fourierreihen vs. FT(mischender Prozeß)
Für z.B. Sägezahn:

$$y = x \text{ für } -\pi < x < \pi, \quad \text{und periodisch fortgesetzt}$$

gilt:

$$y = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \dots\right)$$

Das "Periodogramm (=Spektrum)" lautet also:

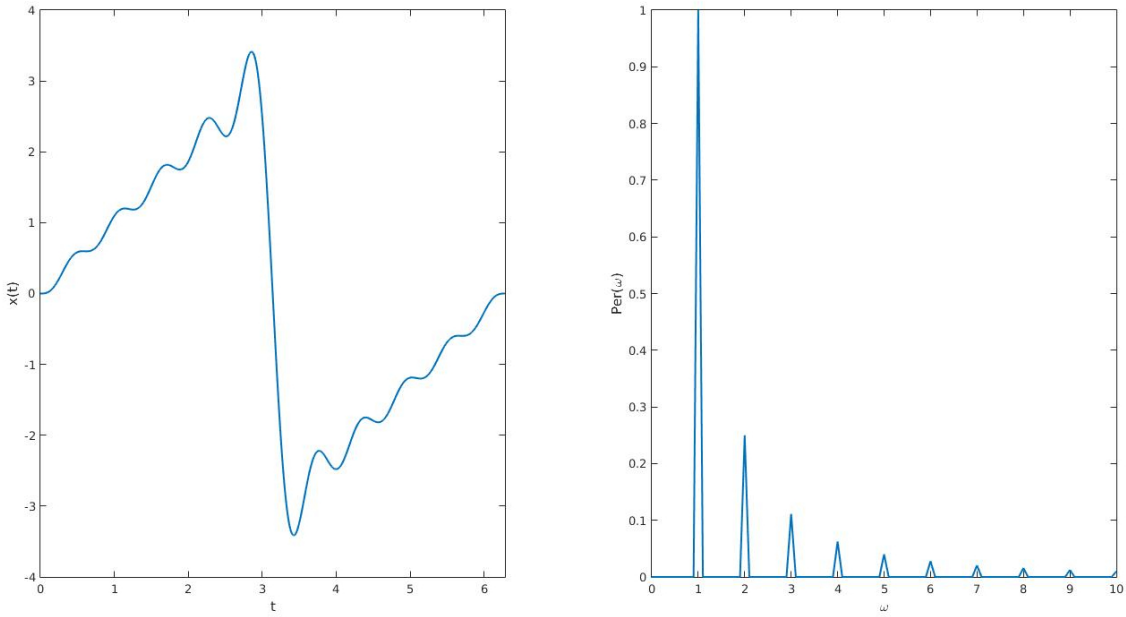


Abbildung 13.8: Periodogramm einer Sägezahnkurve

Betrachte van der Pol Oszillator:

$$\ddot{x} = \mu(1 - x^2)\dot{x} - x$$

Kubische Nichtlinearität, Störungstheorie, Höhere Harmonische bei (2i+1) fachem der Grundfrequenz.

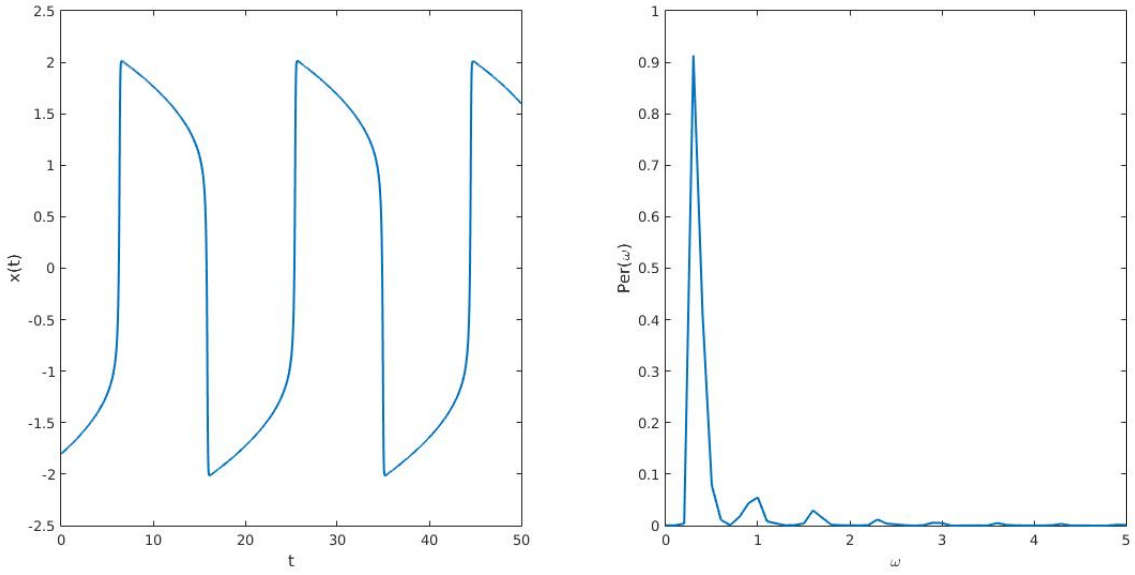


Abbildung 13.9: Periodogramm des van-der-Pol-Oszillators

Leakage und Tapern:
Die FT sieht die Zeitreihe $x(t), t = 1, \ldots T$ als Ausschnitt einer unendlich langen Reihe $y(t), t \in Z$:

$$x(t) = w_u(t)y(t), \quad w_u(t) = \begin{cases} 1 & \text{if } 1 \leq t \leq T \\ 0 & \text{sonst} \end{cases}$$

Effekt Leakage:

- Multiplikation im Zeitraum ist Faltung im Frequenzraum
- Periodogramm hat einen bias
- Spektralschätzung "verschmiert"
- Masse wird von Peaks in Täler transportiert.
- "Leakage": Durchsickern
- Ist für $w_u(t)$ maximal schlimm.

Behandlung:

- Wähle $w(t)$, das sanfter verläuft, z.B. Bartlett window

$$w_B(t) = \begin{cases} 1 - \left| \frac{t - \frac{1}{2}T}{\frac{1}{2}T} \right| & \text{if } 1 \leq t \leq T \\ 0 & \text{sonst} \end{cases}$$

- Dieses nennt man Tapern.

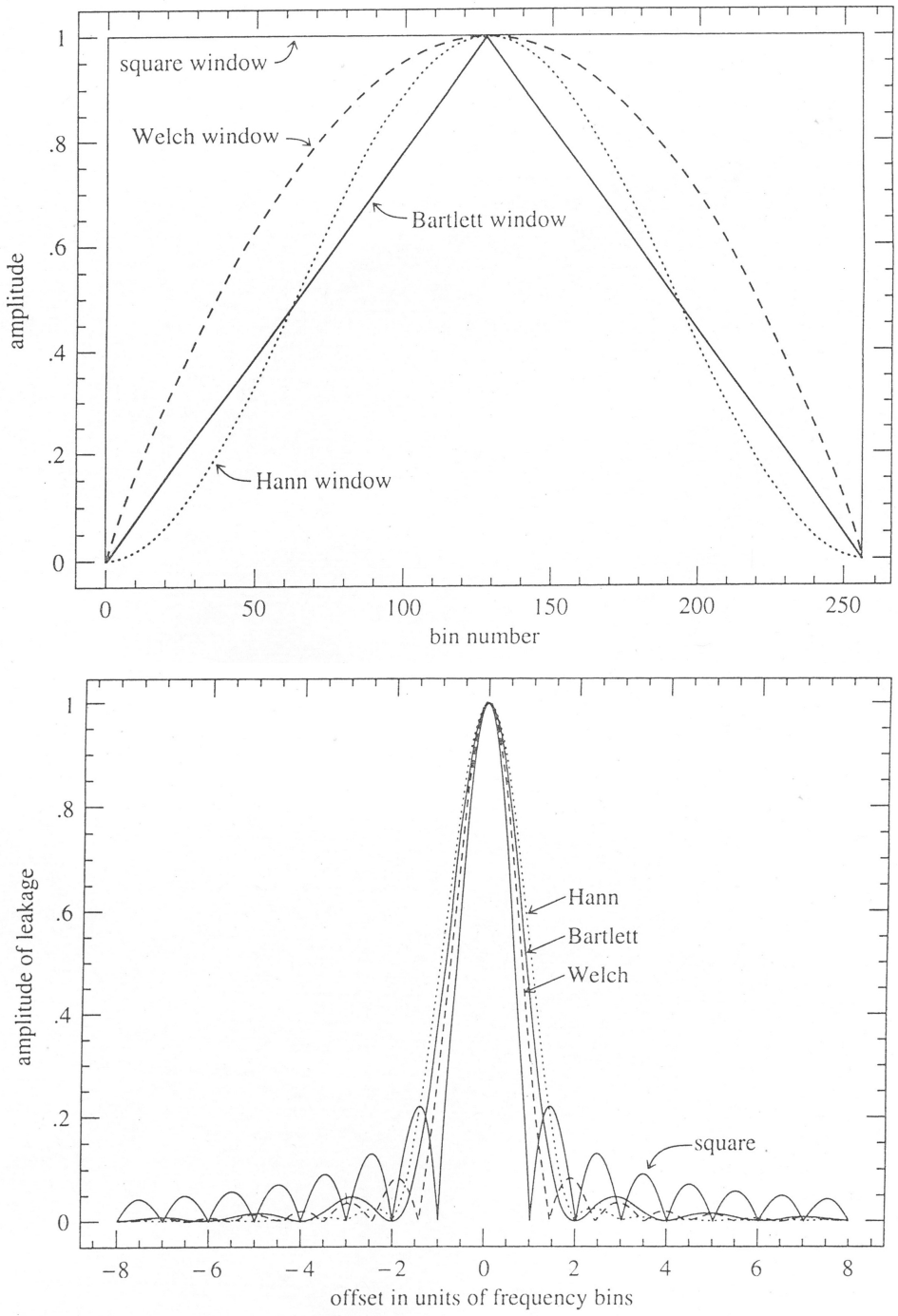


Abbildung 13.10

Das resultierende Spektrum ist mit

$$g = \frac{T}{\sum_{t=1}^{T} w^2(t)}$$

zu normieren.

Übung:
Simulation und Spekralschätzung für AR[2]-Prozeß

Lessons learned:

- Fast-Fourier Transformation
- χ^2_2-Verteilung des Periodogramms mischender Prozesse
- Konsistente Schätzer für das Spektrum

14 Markov Chain Monte Carlo Verfahren

Literatur:

- W.R. Gilks et al. *Markov chain Monte Carlo in practice* [20]
- J.J.K. ÓRuanaidh, W. Fitzgerald *Numerical Bayesian methods applied to signal processing* [49]
- R.E. Kass et al.: Markov Chain Monte Carlo in Practice: A Roundtable Diskussion [29]

Bayes'scher Ansatz (biased version):

- Es gibt keine "wahren" Parameter.
- Parameter sind Zufallsvariablen.
- Jede Wahrscheinlichkeit ist eine bedingte Wahrscheinlichkeit.
- Vorwissen ist die Bedingung.

Bayes'sches Theorem:
Aus

$$p(a,b) = p(a|b)p(b) = p(b|a)p(a)$$

folgt

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)}$$

erlaubt "Umschaufeln" von $p(b|a)$ auf $p(a|b)$.
Seien b die Parameter, a die Daten, dann war die MLE Idee: $p(a|b)$ als Funktion von b zu lesen.

Für Bayesianer macht aber $p(b|a)$ Sinn. $p(b)$ stellt sein Vorwissen dar. $p(a)$ ist konstant und wird weggelassen.

$$p(b|a) \propto p(a|b)p(b) = \text{ Likelihood } \times \text{ Vorwissen}$$

Ist das Fehlermodell Gauß'sch und besagt das Vorwissen, der prior, daß der Betrag von b eher klein ist, z.B. durch:

$$p(b) \propto e^{-\lambda b^2}$$

so ergibt logarithmieren:

$$p(b|a) \propto \sum_{i=1}^{N} \frac{(a_i - a(x_i, b))^2}{\sigma_i^2} + \lambda b^2$$

die Minimum Norm Regularisierung der SVD aus Kap. 7 Lösung linearer Gleichungssysteme.

Gibbs-Sampler
Die Gleichung

$$p(b|a) \propto p(a|b)p(b)$$

bietet die Möglichkeit, Parameter eines Modells im Bayes'schen Kontext zu schätzen.
Problem: Die hochdimensionalen Integrale.

Gleichung

Ausweg: Der Gibbs Sampler
Man kann zeigen: Ziehe einzelne Parameter funktioniert.

ZEICHNUNG Schema

Konvergenz: Lasse 2 Prozesse parallel laufen, wenn Intravarianz = Intervarianz, dann konvergiert.

Wahl des priors

- Wenn der prior den Type der Verteilungsklasse der Likelihood nicht ändert, nennt man ihn conjugate prior. Das macht vieles einfacher.
- Wählt man als prior eine Verteilung, die sehr breit ist, nennt man ihn nicht-informativ
- Im Falle nicht-informativer prior, ist das Ganze MLE, und nur eine Integrationstechnik.

15 Klassifikation

Literatur:

- D.J. Hand, *Discrimination and Classification* [22]
- O. Duda and P.E. Hart *Pattern classification and scene analysis* [13]
- T. Kohonen *Self-organizing maps* [35]

Fischer'sche Diskriminanz Analyse
Mahalanobis Distanz

Clustering

Kohonen map
Optimierung der Trans Information [41]
MDS und projection pursuit
Literatur:

- J.W. Sammon *A nonlinear mapping for data structure analysis* [60]
- P.J. Huber *Projection Pursuit* [27]

Aufgabe:
Gegeben ein hochdimensionaler Datensatz, suche nach Strukturen.
Siehe auch:
ISOMAP [69]
LLE [58]
Was fehlt
An wesentlichem :

- Integralberechnung, Recipes Kap. 4 und 7.6, Stoer Kap. 3
- Stochastische Approximation [30, 56], Die grosse Flut, Thresholding

Literatur

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974.

[3] A.C. Atkinson. Likelihood ratios, posterior odds and information criteria. *J. Econometrics*, 16:15–20, 1981.

[4] P. Bauer, B.B. Pötscher, and P. Hackl. Model selection by multiple test procedures. *Statistics*, 1:39–44, 1988.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, 57:289–300, 1995.

[6] A. Bevan. *Statistical Data Analysis for the Physical Science*. Cambridge University Press, Cambridge, 2013.

[7] R.J. Bhansali and D.Y. Downham. Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, 64:547–551, 1977.

[8] K. Binder and D.W. Hermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Springer, New York, 1997.

[9] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, New York, 1998.

[10] J. Candy and W. Rozmus. A symplectic integration algorithm for separable hamiltonian functions. *J. Computational Physics*, 92:230–256, 1991.

[11] J.W. Cooley and J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19:297, 1965.

[12] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman & Hall, London, 1994.

[13] O. Duda and P.E. Hart. *Pattern classification and Scene Analysis*. Wiley, New York, 1973.

[14] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1998.

[15] E. Forest and R.D. Ruth. Fourth-order symplectic integration. *Physica D*, 43:105–117, 1990.

[16] G.E. Forsythe. Generation and use of orthogonal polynomials for data – fitting with a digital computer. *J. Soc. Indust. Appl. Math.*, 5:74–88, 1957.

[17] J. Franklin. *Computational Methods for Physics*. Cambridge University Press, Cambridge, 2013.

[18] R. Frühwirth and M. Regler. *Monte-Carlo-Methoden*. B.I. Wissenschaftsverlag, Mannheim, 1983.

[19] M.A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem.*, 104:1876–1889, 2000.

[20] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, London, 1997.

[21] D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Physics*, 22:403–434, 1976.

[22] D.J. Hand. *Discrimination and Classification*. Wiley, New York, 1992.

[23] W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1989.

[24] J. Hartung. *Statistik*. Oldenbourg, München, 1989.

[25] J. Honerkamp. *Stochastic Dynamical Systems*. VCH, New York, 1993.

[26] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[27] P.J. Huber. Projection pursuit. *Ann. Stat.*, 13:435–475, 1985.

[28] S. Kakutani. On equivalence of inifinite product measures. *Ann. Math.*, 49:214–224, 1948.

[29] R.E. Kass, B.P. Carlin, A. Gelman, and R.M. Neal. Markov Chain Monte Carlo in practice: A roundtable diskussion. *The American Statistician*, 52:93–100, 1998.

[30] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.*, 23:462–466, 1952?

[31] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics*. Springer, New York, 1992.

[32] P.E. Kloeden, E. Platen, and H. Schurz. *The numerical solution of SDE through computer experiments*. Springer, New York, 1994.

[33] D.E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, Reading, 1981.

[34] A.B. Koehler and E.S. Murphee. A comparison of the Akaike and Schwarz criteria for selcting model order. *Appl. Statist.*, 37:187–195, 1988.

[35] T. Kohonen. *Self-organizing maps*. Springer, Berlin, 1995.

[36] S.E. Koonin. *Computational Physics*. The Benjamin/Cummings Publishing Company, Inc., Menlo Park, 1986.

[37] S. Kotz and N.L. Johnson. *Breakthroughs in Statistics 1 & 2*. Springer, New York, 1992.

[38] S. Lauer, J. Timmer, D. von Calker, D. Maier, and J. Honerkamp. Optimal weighted Bayesian design applied to dose-response-curve analysis. *Communications in Statistics - Theory and Methods*, 26:2879–2903, 1997.

[39] E.L. Lehmann. *Theory of Point Estimation*. Wadsworth Inc., New York, 1991.

[40] Q. Li and H. Wang. Has chaos implied by macrovariable equations been justified. *Phys. Rev. E*, 58:R1191–1194, 1998.

[41] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1:402–411, 1989.

[42] E.N. Lorenz. Deterministic aperiodic flow. *J. Atmos. Sci.*, 20:130, 1963.

[43] E. Mammen. *When does bootstrap work? : asymptotic results and simulations*. Number 77 in Lecture notes in statistics. Springer, New York, 1992.

[44] H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad Sci.*, 100:15522–15527, 1997.

[45] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, London, 1995.

[46] N. Metropolis, A. Rosenbluth, M.Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing maschines. *J. Chem. Physics*, 21:1087–1092, 1953.

[47] L. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A*, 231:289–337, 1933.

[48] B. Øksendal. *Stochastic Differential Equations.* Springer, New York, 1998.

[49] J.J.K. ÓRuanaidh and W. Fitzgerald. *Numerical Bayesian methods applied to signal processing.* Springer, New York, 1996.

[50] W.H. Press, B.P. Flannery, S.A. Saul, and W.T. Vetterling. *Numerical Recipes.* Cambridge University Press, Cambridge, 1992.

[51] M.B. Priestley. *Spectral Analysis and Time Series.* Academic Press, London, 1989.

[52] F. Pukelsheim. *Optimal Design of Experiments.* Wiley, New York, 1993.

[53] C.V. Rao and A.P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.*, 118:4999–5010, 2003.

[54] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25:1923–1929, 2009.

[55] H. Rieder. *Robust statistics, data analysis, and computer intensive methods: in honor of Peter HuberÄs 60th birthday.* Number 109 in Lecture Notes in Statistics. Springer, New York, 1996.

[56] H. Robbins and S. Monro. A stochastic approximation method. *Annals Math. Stat.*, 22:400–407, 1952.

[57] G.J.S. Ross. *Nonlinear Estimation.* Springer, New York, 1990.

[58] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[59] L. Sachs. *Applied Statistics.* Springer, Heidelberg, 1984.

[60] J.W: Sammon. A nonlinear mapping for data structure analysis. *IEEE TRANSACTIONS ON COMPUTERS*, 18:401–409, 1969.

[61] G. Schwarz. Estimating the dimension of a model. *Annals Statistics*, 6:461–464, 1978.

[62] G.A.F. Seber and C.J. Wild. *Nonlinear regression.* Wiley, New York, 1989.

[63] S. G. Self and K. Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Ass.*, 82:605–610, 1987.

[64] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[65] B.W. Silverman. *Density Estimation.* Chapman and Hall, London, 1986.

[66] J. Stoer. *Einführung in die Numerische Mathematik I.* Springer, Heidelberg, 1983.

[67] J. Stoer and R. Bulirsch. *Einführung in die Numerische Mathematik II.* Springer, Heidelberg, 1983.

[68] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Stat. Soc. B*, 39:44–47, 1977.

[69] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[70] T. Teräsvirta and I. Mellin. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, 13:159–171, 1986.

[71] J. Timmer. Estimating parameters in nonlinear stochastic differential equations. *Chaos, Solitons & Fractals*, 11:2571–2578, 2000.

[72] J. Timmer and S. Klein. Testing the Markov condition in ion channel recordings. *Phys. Rev. E*, 55:3306–3310, 1997.

[73] Q. H. Vuong. Likelihood ratio tests for modelselection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

[74] Westfall and Young. *Resampling based Multiple Testing: Examples and Methods for p-value Adjustment.* Wiley, New York, 1993.

[75] O. Wolkenhauer, M. Ullah, W. Kolch, and K.-H. Cho. Modelling and simulation of intracellular dynamics: Choosing an appropriate framework. *IEEE Transactions on NanoBioScience*, 3:200–207, 2004.