Script Following the Lecture :

Statistics and Numerics

Lecture SS 21

Prof. Dr. Jens Timmer

December 22, 2021

Contents

1 Introduction

Technicalities:

- Which semester ? Any master students ?
- Script and communication on homepage
- Exercises ,Wednesday flexible, not a Hacker course
- Slides available after a few weeks
- Script is meant as a table of contents
- If something is not clear $\Rightarrow$**Ask Questions!**

Literature:

- On statistics
  - A. Bevan. *Statistical Data Analysis for the Physical Science* [6]
  - J. Honerkamp. *Stochastic Dynamical Systems* [25] Kap. 1-3.
    Condensed showcase of the basics of statistics relevant for physicists
  - J. Hartung. *Statistik* [24] A classic, very detailed
  - L. Sachs. *Applied Statistics* [59] Compendium, applied
  - D.R. Cox, D.V. Hinkley. *Theoretical Statistics* [12] easy to read theoretical literature
- On numerics
  - W. Press et al. *Numerical Recipes* [50]:
    The Bibel, optimal for physicists
  - J. Stoer. *Einführung in die Numerische Mathematik I & II* [66,67]
    Mathematically orientated classic
  - Additional books from the field of 'Computational Physics' and 'Monte-Carlo Methods' : [8, 17, 18, 36] Franklin modern

Part I
Statistics

Fundamentals on the topic of statistics:

- Some things need to be understood.
- Much should be known.
- Many things you just have to be able to look up.
- Applied statistics is not a case of mathematics.

2 Distributions

2.1 Random variables

Random variable X:

- Something with a probability distribution $p_X(x)$
- Probability to observe a realization of x in $(x, x + dx)$ is $p_X(x)dx$
- $p_X(x) \geq 0$, $\int p_X(x)\, dx = 1$

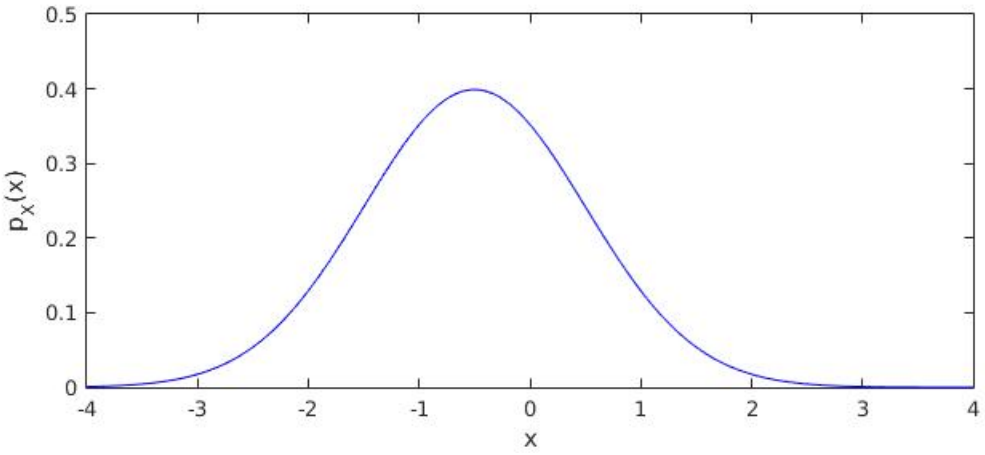

Figure 2.1: Example of a probability distribution

- In physics, coincidence comes from:
  - Qunatum mechanics, rather rare in macroscopic complex systems
  - Chaos, also rare, realized e.g. by rolling a dice
  - A lot of influences like Brownian Motion, most common.
- There are also discrete distributions $p(x_i)$, (think back to the dice)

In the following, if the relation is clear: $p_X(x) = p(x)$, $X = x$

2.2 Moments and Cumulants

- Expectation value $\langle f(x) \rangle$

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

  Expectation value is a number

- Moment μ_k

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

- 1. Moment: Mean
$$\mu_1 = \bar{x} = \mu = \langle x \rangle = \int x\, p(x)\, dx$$
- 2. Moment
$$\mu_2 = \langle x^2 \rangle = \int x^2\, p(x)\, dx$$
  - Variance: $\sigma^2 = \langle (x - \bar{x})^2 \rangle = \mu_2 - \mu_1^2$
  - Standard deviation: σ
  - While adding independent random variables, variances, as opposed to standard deviations, are additive.
- 3. Moment
$$\mu_3 =< x^3 >= \int x^3 p(x)\, dx$$
  Skewness:
$$\kappa = \langle (x - \mu)^3 \rangle$$
  Measure of asymmetry.
- 4. Moment
  Curtosis (bellyness):
$$\gamma = \langle (x - \mu)^4 \rangle / \sigma^4 - 3$$
  "-3" will become clear further down the road.
- Characteristic function or generating function
$$G(k) =< e^{ikX} >= \int dx\, e^{ikx} p(x)$$
  Fourier transform of the density $p(x)$
- If the moments exist, i.e. $< X^n >< \infty$, Taylor evolution
$$G(k) = \sum_{n=0}^{\infty} \frac{(i\,k)^n}{n!} < X^n >$$

is $G(k)$ known, the moments are calculated via:

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

- Evolution of $\log(G(k))$ by k, follows:

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

with the Accumulants κ_i

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

Important characteristics:

- Accumulants are additive , therefor natural values
  Let

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

  then follows

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

  variance is additive, not standard deviation
- It can be shown:
  - Either: All accumulants except the first two disappear
  - Or there exist ∞ many

2.3 Examples of distributions

- Gaussian or normal distribution:

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

  - Notation: $N(\mu, \sigma^2)$
    Standard Gaussian distribution (normal distribution) $\mu = 0$, $\sigma^2 = 1$: $N(0, 1)$
    In $\pm 1\sigma$ lies 68 % of the mass
    In $\pm 1.96\sigma$ lies 95 % of the mass
  - Moments of $N(0, 1)$:

$$\langle x^k \rangle = \begin{cases} 0 & \text{for k uneven} \\ 1 \times 3 \times \cdots \times (k-1) & \text{for k even} \end{cases}$$

    Therefor it is clear where the "-3" in the kurtosis comes from.
  - Characteristic function:

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

    Only the first two accumulants are $\neq 0$
    Shows why the SDG is so special!

Central limit theorem:

If the first two moments exist, the (normalized) sum of independent, identically distributed (iid) random variables strives toward a normal distribution.

Consider N identical random variables X_i with

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

Form:

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

For accumulants follows:

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

Especially:

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

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

- The higher accumulants disappear with N.
- Distribution tends towards normal distribution, which is why μ and σ are so important.
- Holds also for non-identical X_i
- Convergence rate, i.e. how quickly the convergence to the normal distribution happens, depends on the skewness.

Averaging:

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

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

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

The importance of the central limit theorem is not to be underestimated.

- Even distribution $U(a, b)$: χ^2_r-distribution with $r = 1, 2, 3, 4, 5$ degrees of freedom.

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

- Exponential distribution

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

It holds:

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

Estimation value and variance are not independent parameters.

Obtained for "constant decay rate"

- χ^2_r distribution with r degrees of freedom:
Sum of r squared normal distributions

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

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

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

It holds:

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

Meaning expectation value and variance are not independent parameters

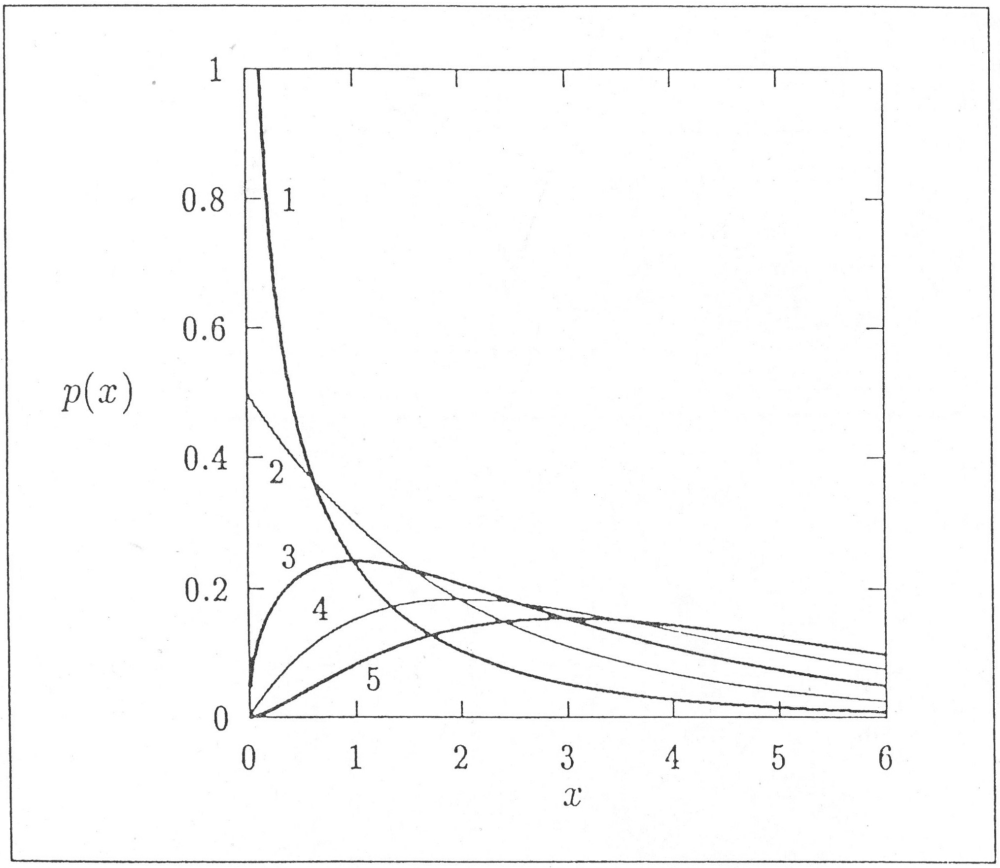

Figure 2.2: χ^2_r distribution with $r = 1, 2, 3, 4, 5$ degrees of freedom

χ^2 distributions are additive:

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

From the central limit theorem follows:

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

Remarks :

- $\chi^2_2 = \frac{1}{2}e^{-x/2}$ is an exponential distribution with $\tau = 2$.
- χ^2_2 will be important in 13 Spectral analysis.

- t-distribution

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

r: Number of degrees of freedom

See chapter 3.1: t-Test: Tests for equality of averages of two normal distributions.

$$\lim_{r\to\infty} t(r,x) = N(0,1), \qquad \text{good approximation for } r = 30$$

- F distribution

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

r_1, r_2: Respective number of degrees of freedom

F test: Tests for equality of variances of two normal distributions.

- Cauchy(Lorenz) distribution:

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

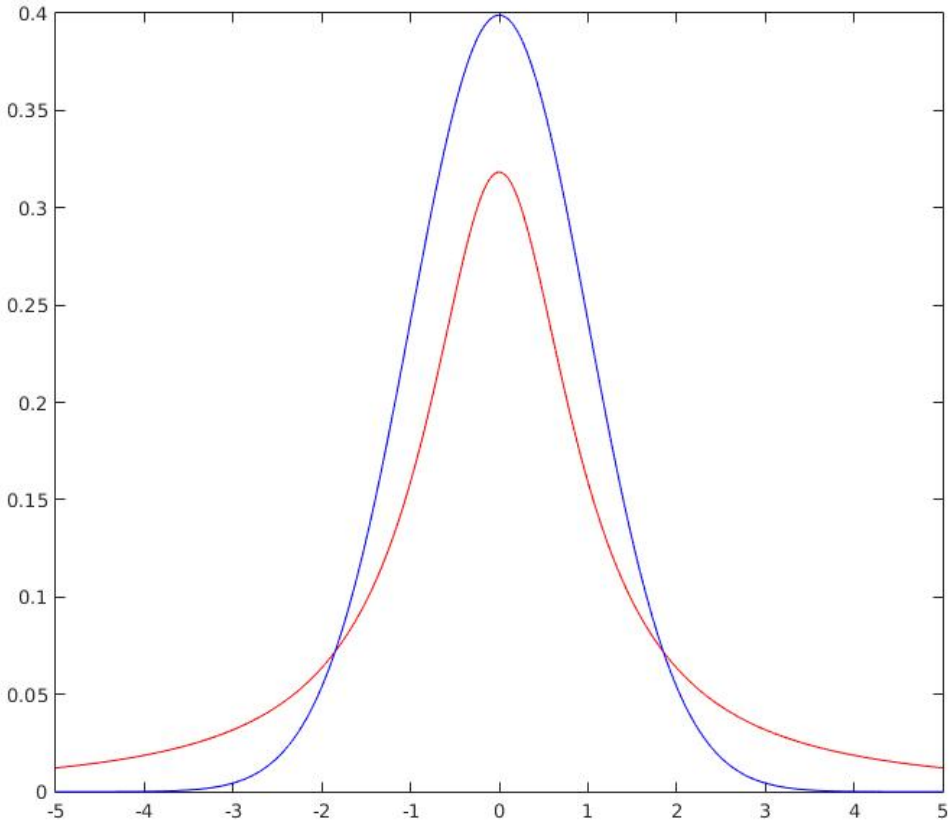

Figure 2.3: Cauchy distribution (red) in comparison to the normal distribution (blue)

- Moments don't exist!
- Characteristic function:

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

  There exists no Taylor evolution around $k = 0$
- a is a Localization parameter, but no mean.
- Cauchy-distribution plays a role in the increase of share prices.
  Optional excursion: Black-Scholes

- Central limit theorem is not valid for the Cauchy- distribution.

  However there are limit theorems for distributions with non-existent moments. Keyword „stable distributions “

- Reference to the t-distribution:

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

  With t distribution one can transition between Cauchy (no moments exist) and Gaussian (all moments exist).

- Cauchy distribution known in physics as Breit-Wigner distribution.

- Multivariate normal distribution

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

  with covariance matrix C

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

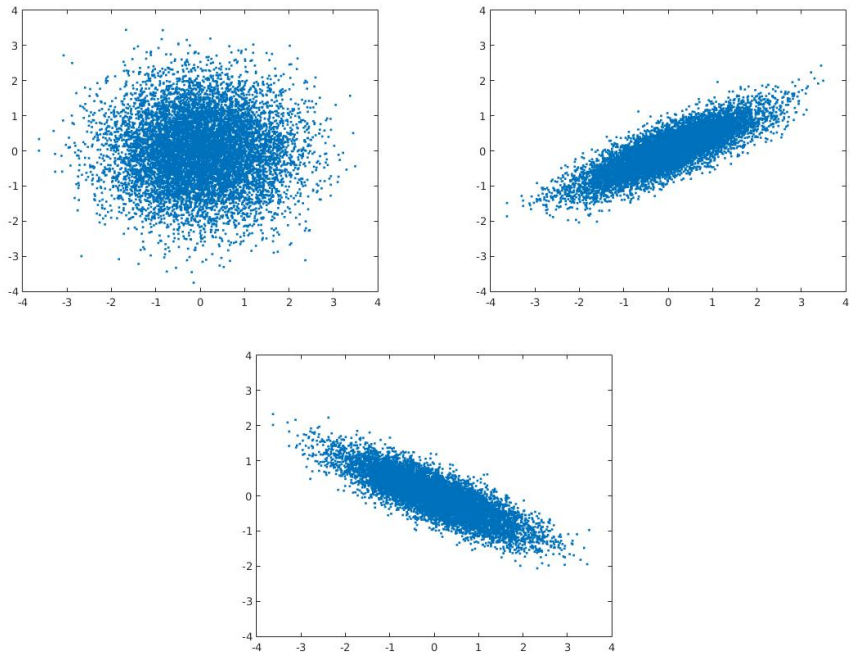

Figure 2.4: *2D*-normal distributed random numbers with $C_1 = (0.71\ 0; 0\ 0.70)$, $C_2 = (0.78\ 0.39; 0.39\ 0.28)$, $C_3 = (0.79, -0.39;\ -0.39, 0.28)$

- 1 D Normal distribution:
  68 % of the mass in $[-\sigma, \sigma]$
  99 % of the mass in $[-3\sigma, 3\sigma]$
- 10 D Normal distribution, $C = \mathbb{1}$:
  99 % of the mass outside of the $[-3\sigma, 3\sigma]$-sphere.
- Intuition:
  - Integration over the angles
  - Leaves, $d = \dim(\vec{x})$

$$\text{Mass inside of radius } r \sim \int_0^r r^{d-1}\, e^{-r^2/2}\, dr$$

- There are practically only the longest distances, the space is empty, "curse of dimensionality", comes back in Chap. 12 Core estimator.

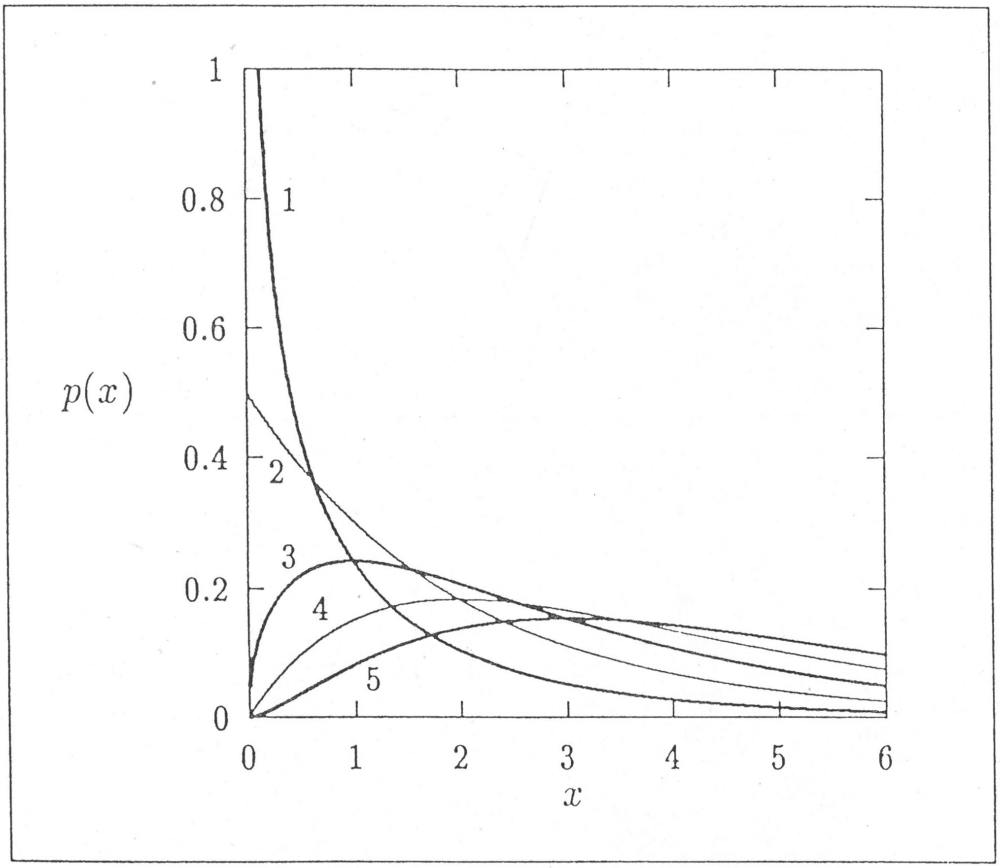

Figure 2.5: χ^2_r-distribution with $r = 1, 2, 3, 4, 5$ d.o.f.

1. week

- Binomial distribution

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

Two possible events: x_1, x_2; $p = prob(x_1)$

For n executions of the experiment, $B(n, p, k)$ is the probability of realizing x_1 k times.

- Poisson distribution

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

  - Probability for k events in a time interval
  - λ: Average number of events in time interval
  - Important for point-processes with constant rate, think of firing neurons or photon counting processes
  - Explain connection to dynamical systems by means of integrate-and-fire neuron

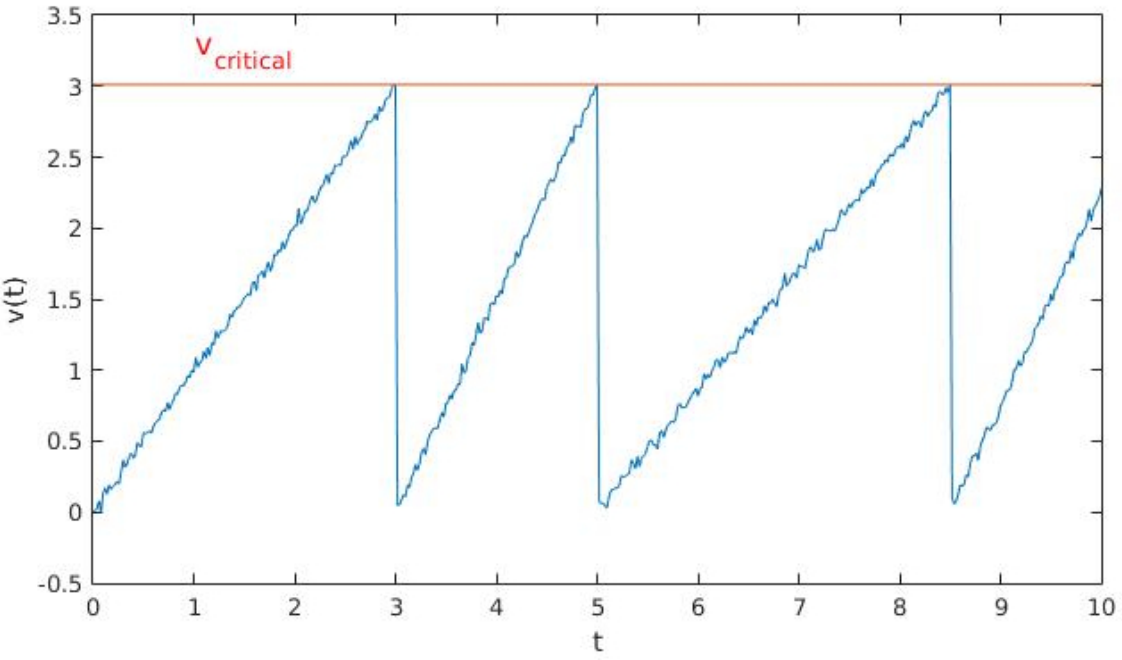

Figure 2.6: Integrate-and-Fire-neuron

  - For Poisson distribution holds

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

  - Furthermore it is the limit distribution of the binomial distribution:

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

    Describes "rare events"

– Poisson distribution for small λ very asymmetric. For large λ>30, it tends towards a normal distribution

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

Cumulative distributions

- Definition:

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

- x_α with

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

  is called $\underline{(100\alpha)\ \%\ \text{Quanta}}$.

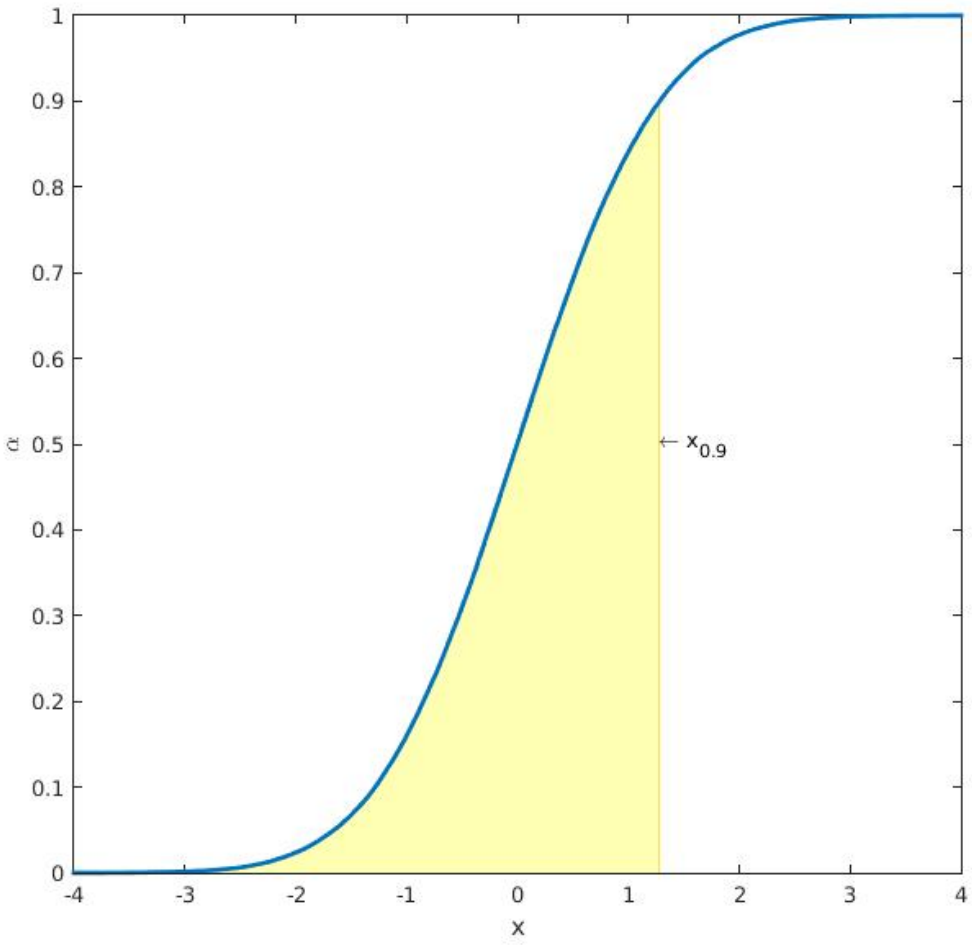

Figure 2.7: Cumulative distribution of the normal distribution with 90%-quanta

- Important for test theory, Chap. 3
- Definition of the median:

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

The mean value of the distribution

2.4 Estimation of parameters from distributions

General parameter estimation theory in chapter 4

Definitions:

- True parameter : Θ_0
- Estimator for parameter : $\hat{\Theta}$, this is a random variable
- Bias : $\langle\hat{\Theta}\rangle - \Theta_0$
- Variance of the estimator : $\langle(\hat{\Theta} - \langle\hat{\Theta}\rangle)^2\rangle$
- Mean quadratic error : $\langle(\hat{\Theta} - \Theta_0)^2\rangle = \text{bias}^2 + \text{variance of the estimator}$
- Confidence interval: Area around $\hat{\Theta}$, where the true value lies Θ_0 with a certain probability.

Gaussian distribution $N(\mu, \sigma^2)$:

- Let every $X \sim N(\mu, \sigma^2)$
- Estimator for the mean μ

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

As sum over normal distributions $\hat{\mu}$ is normal distributed

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

Estimator is unbiased. Is correct on average.

Variance of the estimator

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

In summary: $\hat{\mu}$ is a normal random variable with

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

With this follows: $\pm\sigma$ (=68%) confidence interval for true μ:

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

or

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

With increasing amounts of data points the mean can be determined ever more accurately.

Estimator unbiased and confidence interval decreases with $\sqrt{\frac{1}{N}}$: Estimator is consistent. Consistent: For $N \to \infty$ everything is going to be fine

- Three estimators S_k^2, $k = 1, 2, 3$, for the variance
  - Let the mean be unknown
    First try:
    $$S_1^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2$$
    Looking at one of the summants and skillfully adding 0
    $$\begin{aligned} <(X_i - \hat{\mu})^2> &= <((X_i - < X >) - (\hat{\mu} - < X >))^2> \\ &= Var(X) - 2 < (X_i - < X >)(\hat{\mu} - < X >) > \\ &\quad + Var(\hat{\mu}) \end{aligned}$$

From before: $Var(\hat{\mu}) = \frac{1}{N} Var(X)$
and

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

All together

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

Therefore:

$$< S_1^2 >= \frac{1}{N} \frac{N-1}{N} \sum_{i=1}^{N} Var(X) = \frac{N-1}{N} Var(X) = Var(X) - \frac{1}{N} Var(X)$$

Ergo: Estimator S_1^2 has a bias of

$$\text{Bias}(s_1^2) = \frac{1}{N} Var(X)$$

Only "asymptotically undistorted ", meaning for $N \to \infty$
Discussion asymptotic

– Second try

$$S_2^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Same calculation as above

$$< S_2^2 >= Var(X)$$

Justification:

* The calculation of the mean costs one d.o.f.
* $x_1, \ldots, x_N$ underlie the constraints :

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

* Factor $\frac{1}{N-1}$ is called Bessel correction

– Let the mean μ be known

$$S_3^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Same calculations as before

$$< S_3^2 >= Var(X)$$

Confidence interval for p of the binomial distribution

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

- With $m = \#x_1$, the estimator is

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

asymptotically normal distributed :

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

Normal distributed because of the central limit theorem

- 95 % confidence interval:

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

- Asymptotic holds for $n\,p(1-p) > 10$

  Discussion asymptotic

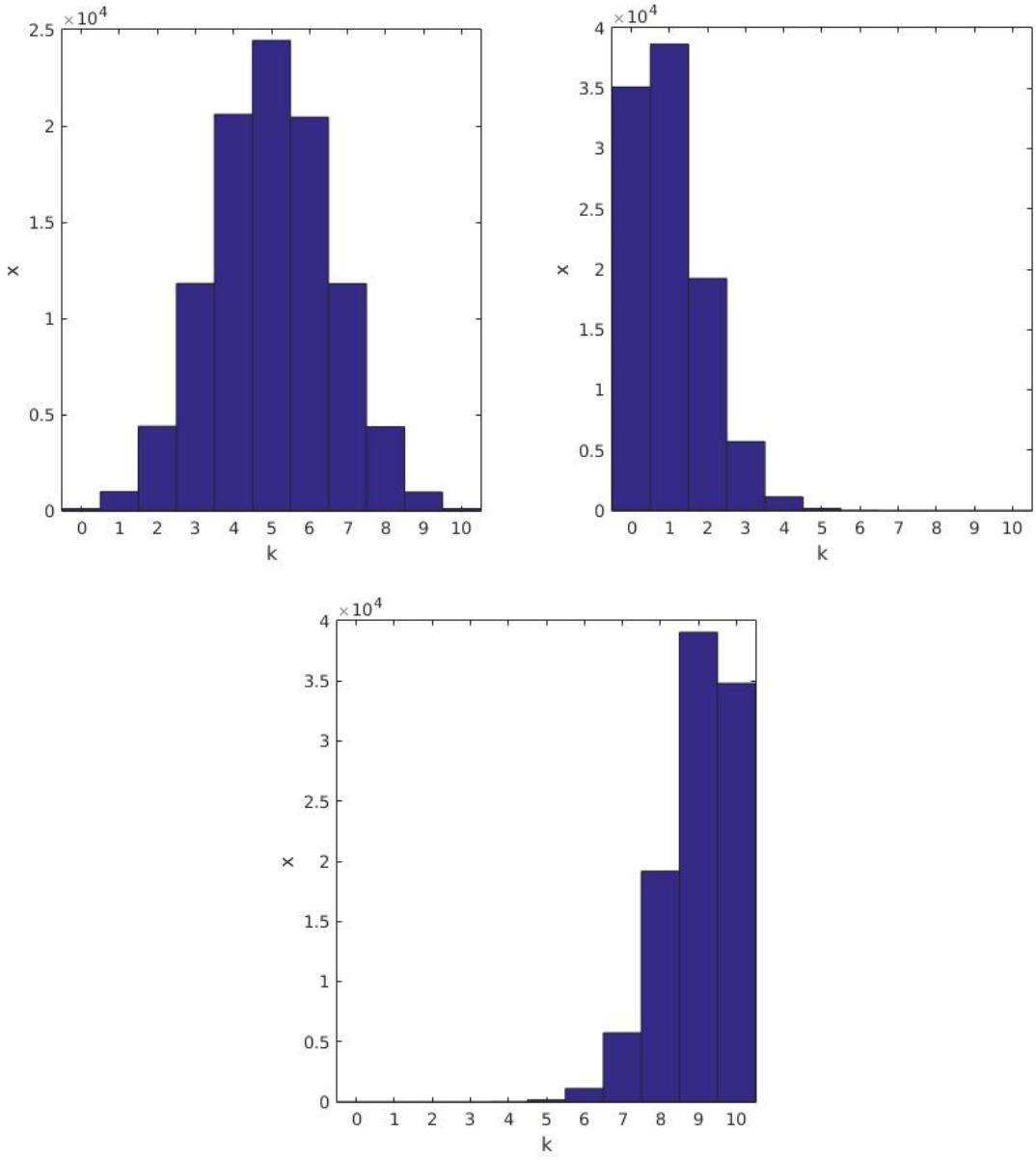

Figure 2.8: Binomial distribution with $p_1 = 0.5$, $p_2 = 0.1$, $p_3 = 0.9$

- Scewness must become smaller by averaging, is slower on the edge.

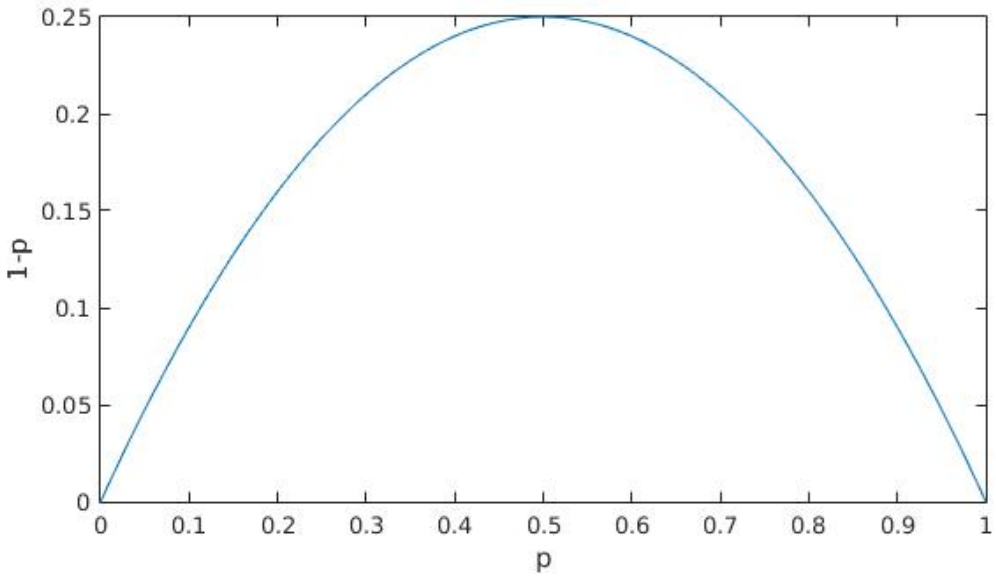

Figure 2.9: p(1-p): Responsible for variance of the estimator, "1-p" must be called "variance"

- For $n\,p(1-p) < 10$: Look up Pearson-Clopper values.

Lessons learned:

- Random variables have a distribution, realizations are a number.
- Normal distribution is important because of the central limit theorem.
- High dimensional spaces are basically empty.
- Estimators are random variables.
- Consistent estimators are great.

3 Hypothesis tests

... or **The five dilemmas of testing**

3.1 Parametric tests

More often than not questions will amount to statistical tests.
Everything else will be shown with the example of the t-test.

The procedure

- Formulate a null-hypothesis H_0:

  Here:

  The means μ_1, μ_2 of 2 normal distributions with equal variance σ^2 are equal.

  Note: This contains three assumptions

  Test is parametric, because parametric distributions, here normal distributions, are assumed.

- Calculate (analytic/simulate) distribution of a test size under the null hypothesis.

  Here analytical:

  - Estimate means $\hat{\mu}_1$, $\hat{\mu}_2$ for N measurements x_i^1 and x_i^2:

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

    Corresponding variances $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$:

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

  - Calculate the mean :

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

    And the standard error of the mean:

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

– Under the validity of H_0

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

t-distributed with $r = 2N - 2$ d.o.f.

"-2" because 2 means are estimated from the data

– Reminder: Definition of t-distribution:

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

– Normalization in the asymptotic:

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

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

– Consider:

* Usually one wants to reject H_0: Drug is better than placebo. Here $\mu_1 \neq \mu_2$
* Under the alternative H_1 test statistic (hopefully) has a different distribution than under H_0
* Here: Normalized asymptotic distribution of t under alternative $\mu_1 \neq \mu_2$:

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

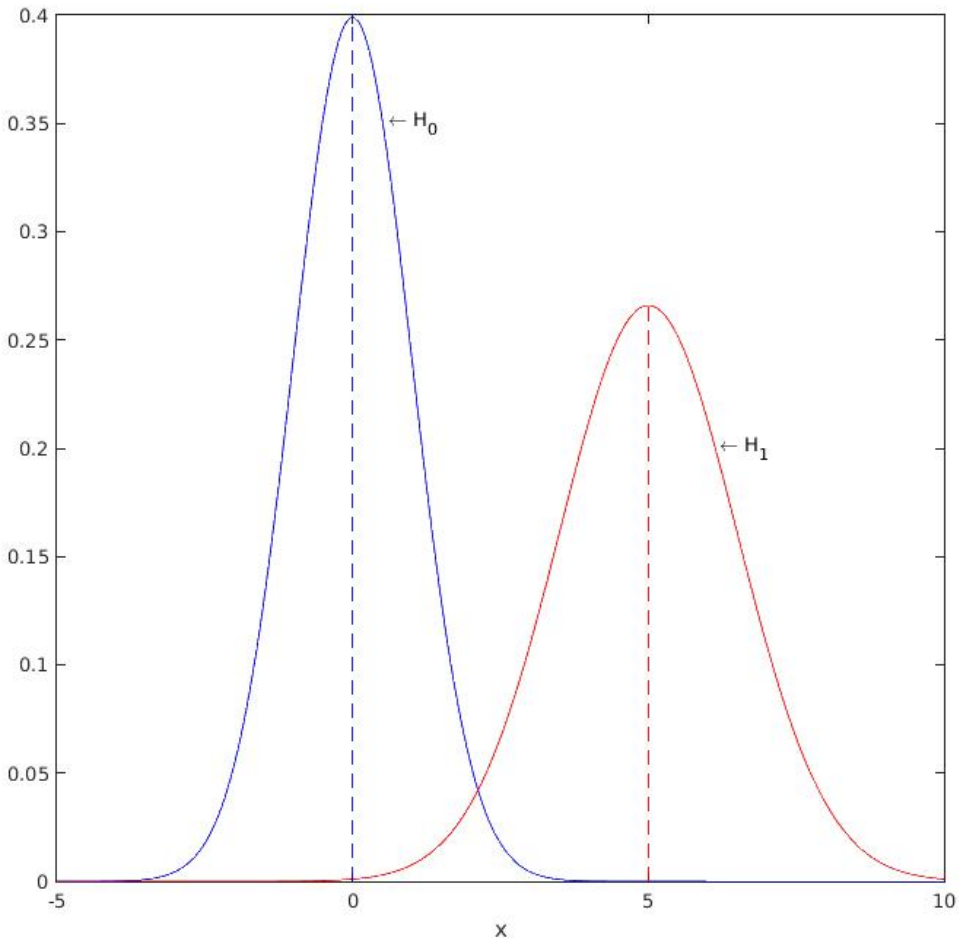

Figure 3.1: Null-Hypothesis H_0 and alternative H_1

- **Dilemma I** of testing: Everybody can belong
  - Execution of the test, yield concrete numbers for $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{S}_M$ and thus for t
  - Test surmounts to question:
    Does a value belong to - the realized t-value - to the - here - t-distribution ?
  - Problem:

This cannot be denied !

In principal every value of the test-statistic - here t- can occur under H_0.

- p-value: Probability for value bigger t:

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

  Per construction: Under H_0: p value of the test-statistic is equally distributed on [0,1].

- Way out: discard null hypothesis $H_0 : \mu_1 = \mu_2$ for extreme events:
  p-value very small, p-value very large
- Therefor: choose significance levels α.
  Reject H_0, when p-value $< \alpha$
- Typical values for α: 0.05 or 0.01
  Applied statistic is not a case of mathematics!

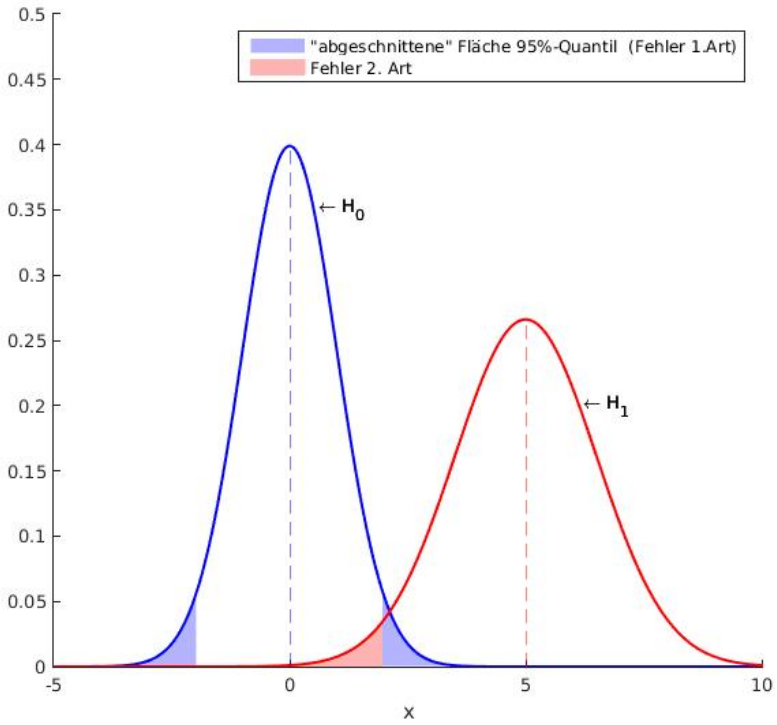

Figure 3.2: Null hypothesis H_0 and alternative hypothesis H_1 with significance levels α

  - Two kinds of errors can happen:
    Error of the 1. kind: H_0 is rejected even though true: False positive
    Error of the 2. kind: H_0 is not rejected even though false: False negative
  - Error of the 2. kind costs a good paper
  - Error of the 1. kind costs the career

- Power of the test: Frequency of rejections of a test, when H_0 is false.

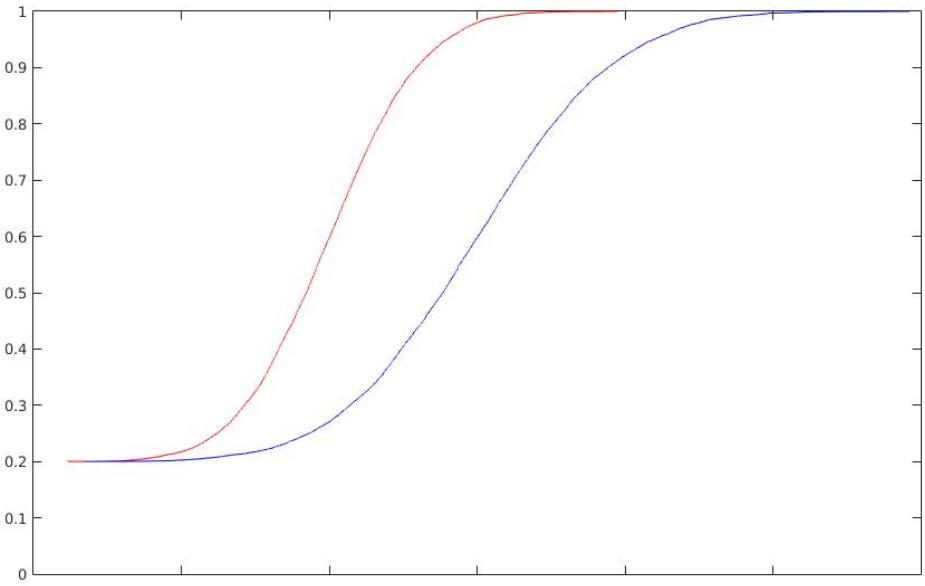

Figure 3.3: Power of the Test

- Actual frequency of errors of 1. kind $< \alpha$: Test is conservative.
- Actual frequency of errors of 1. kind $> \alpha$: Test is garbage.

Varieties of the t-test:

- Is one mean larger than another ?
  One sided and two sided test, discussing power
- Random sample-t-test: Is a distribution in agreement with a certain mean?
- Variances of two normal distributions are different
- Number of samples is different

2. week

Exercise :
The power of the t-test

Dilemma II: Dichotomy of Kakutani [28]

- When H_0 is not true, an alternative H_1 with distribution $p_{H_1}(x)$ applies
- Now holds:

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

  For $N \to \infty$ distributions $p_{H_0}(X)$ and $p_{H_1}(X)$ become ever narrower.

  - If $p_{H_0}(X) \neq p_{H_1}(X)$ there will eventually not be any common carriers left.
  - If $p_{H_0}(X) = p_{H_1}(X)$ nothing happens anyway.

- When a test has any power at all, H_0 will always be rejected with increasing number of data points.

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

Dilemma III: Statistical significance vs. content relevance

- Patients with pulses of 180 ± 10 beats/min
- A drug reduces pulse to 170 ± 10 beats/min
- Perform t-test with N patients through:

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

  - Any small violation of the null hypothesis lead to significant differences if a sufficient amount of measurements N are available.
  - Before performing the test one should consider, to what extend a violation of the null hypothesis is relevant for content.
  - From this it can be determined how many measurements N are necessary to proof a sensible violation.
  - Case number calculation

- Here:
  - What is a clinically relevant decrease in pulse?
  - How many measurements/patients N are needed to reject the null hypothesis H_0: "Drug has no effect." ?
  - If H_0 is afterwards rejected: Drug is useful.
  - If H_0 is not rejected based on N measurements/patients:
    On a content relevant scale the drug has no effect.

Dilemma IV: Multiples tests

- Setting: Based on m parameters it shall be tested , whether two species differ.
- H_0 : There is a difference
- Procedure: Perform m t-Tests, each at significance level α.
- Probability $\tilde{\alpha}$, to reject H_0 :

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

- Example: $\alpha = 0.01$

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

Solution 1:

- Bonferroni - correction:
- Solve eq. (1) for α:

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

- Calculate for desired (global) $\tilde{\alpha}$ the needed α for the single tests.

- Problem :
  - α becomes very small,
  - Test become very conservative, no power
    $\Longrightarrow$ many errors of the 2. kind.
- Variation: Bonferroni-Holm: Correct in every step j with j/m.

Solution 2:

- An experiment (m) to generate hypothesis,
- Yields $m' \ll m$ candidates.
  Some correctly positive, some false positive.
- A second experiment to test with m'.

Variation for this topic:

- AIDS Test
- First (cheap) sensitive test, which is not highly specific
- If positive, then multiples (expensive) tests, which are highly specific but not as sensitive.

Solution 3:

- Use binomial distribution $B(m, \alpha, k)$ to estimate the number of false positives: False discovery rate

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

- If there are many more positives, there is a difference.

Or Bootstrap-method : [5, 74]

Special case: ANOVA

- Consider: Experiment examines several conditions in the same respect.
- For example placebo, drug_1, ..., drug_M with respect to # red blood cells
- ANalysis Of VAriance (ANOVA) is the alternative to $\frac{M(M-1)}{2}$ t-tests.

Derivation:

- H_0 : No effect.
- M conditions, N observations each: x_{ij}
- Average per condition

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

Average over all:

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

Variance of all data, called SS_{total}, SS for sum of squares, is:

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

- First summant: Variance of the group means with respect to total average $SS_{between}$ with $M-1$ d.o.f. .

  Second summant: Variance in the different group means SS_{within} with $(N-1)M$ d.o.f.
- Under validity of the null hypothesis their quotients follow a $F(M-1, (N-1)M)$ distribution.

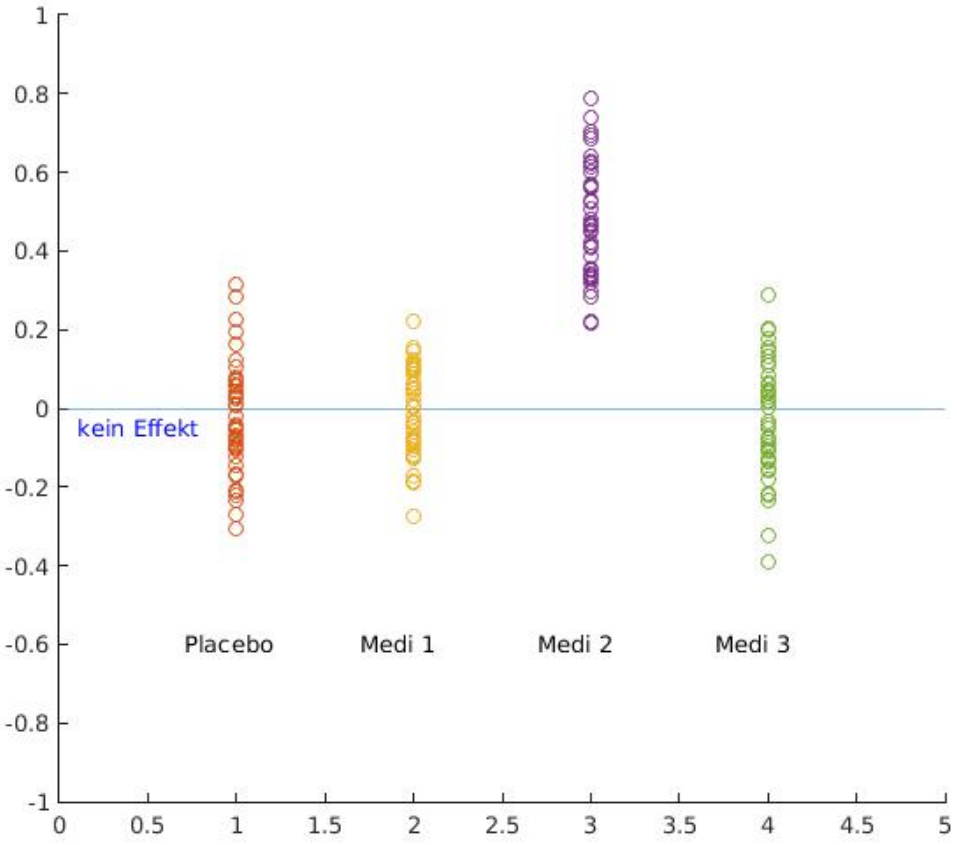

Figure 3.4: Different drugs in ANOVA-test: Drug 2 works, drugs 1 and 3 and the placebo show no effect

In case of rejection:

- If ANOVA is significant, it poses the question: Who did it?
- A posteriori test (Tukey-Kramer or Scheffé) yields critical difference which the means must surpass to be considered significant.
- Considers that the data used in ANOVA have already been statistically used once.

- *A posteriori* test always has smaller power than a single t-test between for example the largest mean differences. It is therefor important [1] to determine *prior* to an experiment what the minimum hypothesis to be tested should be. Otherwise one runs the risk of not being able to statistically prove present effects.

Generalization for

- Different variances
- Different sample sizes
- Multiples parameters, so called factors, drugs and genders

Short version of ANOVA: Compare ,the variance of group means to the total mean, to the variance of the different group means (F-test). That way one can save $\frac{M(M-1)}{2}$ t-tests.

Paired tests

- Previous assumptions: Distributions are independent.
- If data is recorded from the same individuals, this has an effect on the variance:

[1] but sadly not common

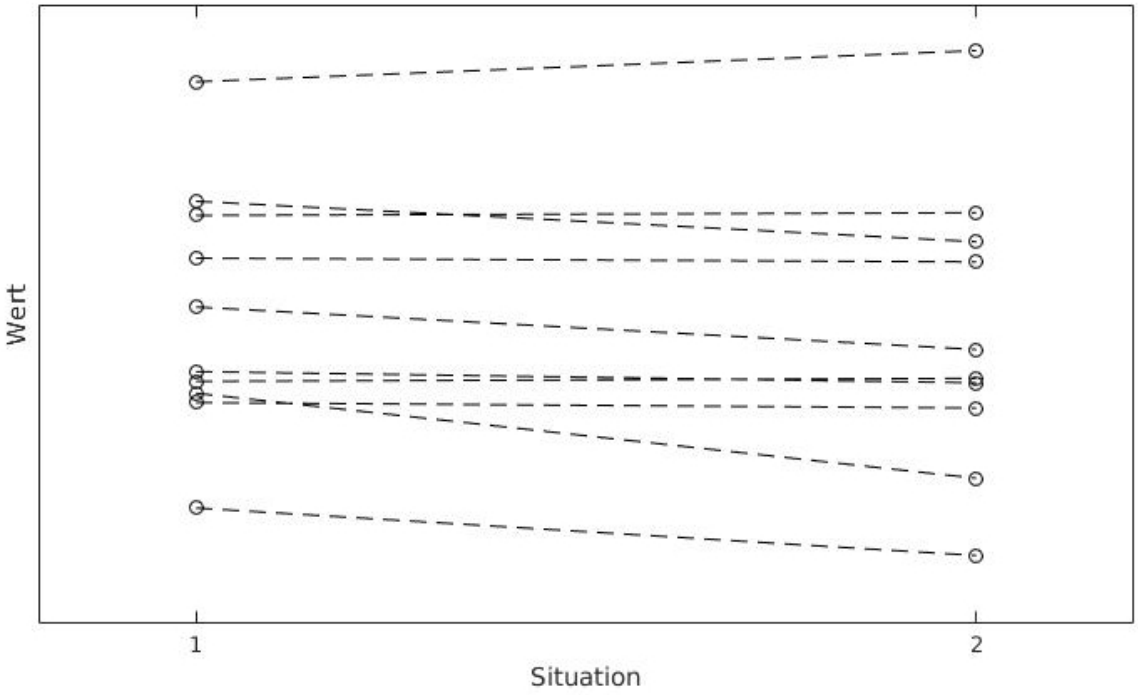

Figure 3.5: Paired values

This must be considered.

- One can speaks of paired tests.
- As in the case of paired t-tests, a repetition of tests can also be considered for ANOVA e.g. for the dependency of sample size to the different experimental conditions. This happens for example through the Greenhouse-Geisser-correction.

3.18

3.2 Non-parametric tests

Previously:

- t-test assumed normally distributed samples: Parametric Test
- If distribution drops a lot slower than normal distribution e.g. Cauchy-distribution, t-test loses it's power ,see exercise

Discuss "Outliers", shit happens

- Alternative: Non-parametric tests
- Those are robust against violations of the assumption distribution
- Instead of mean comparison $\Rightarrow$ location comparison.

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

Here :

- t-test works following Wilcoxon rank sum test (or U-test, Mann-Whithney-test)
- Null hypothesis: The distributions are identical.

  Or: The ranks of both samples are equally distributed with respect to the ensemble

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

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

  then

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

- Calculate the ranks $R_1^1, \ldots, R_N^1$ of the first sample with respect to the total.
- Under H_0 applies:

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

- With central limit theorem:

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

  Good approximation for $N \geq 20$

- For $N < 20$, exact values can be obtained through combinatorics but are time-consuming to calculate, are tabulated

- Reason for robustness: Lowest value has rank= 1, highest value has rank= $2N$, no matter whether normal or Cauchy distribution are underlying or if there are "outliers".

- Paired case: Wilcoxon-sign-rank-test

  Non-parametric ANOVA: Kruskal-Wallis test or also H-test.

Efficiency

- If data is normal distributed, parametric t-test recognizes a difference in averages with fewer data as Wilcoxon-test or with equal N with smaller differences, t-test has higher power.

- The smaller power of non-parametric test as compared to parametric test if parametric assumption is valid, is given by the efficiency:

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

  given the validity of the parametric distribution.

- Wilcoxon-test has an efficiency of 0.95 compared to the t-test, meaning the t-test has with 95% of the data the same power as the Wilcoxon-test is the data is normally distributed.

- Since wrong distribution assumptions lead to a loss of power in parametric tests but non-parametric tests have an efficiency< 1, it follows:

  Dilemma V : Power vs. Efficiency

- Since they have an efficiency near 1 and are robust against violations of distribution assumptions, non-parametric tests are preferred nowadays.

Lessons learned:

- Crucial: Derivation of the distribution of a test statistic under H_0
- The five Dilemmata of testing:
  - Everyone can be part of it, necessity of a significance level
  - With increasing number of data points every null hypothesis will be rejected
  - Statistical significance vs. content relevance
  - Multiples tests
  - Power vs. efficiency in parametric vs. non-parametric tests
- Significance levels are not a case of mathematics but of risk assessment

3. week

4 Parameter estimation

Literature:

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

Motivation: Easiest model: Linear regression, see Chap. 10.1

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

Two questions, answers are going back to Gauß:

(i) Given N data pairs (x_i, y_i), how does one determine a ?
Keyword: Point estimation

(ii) How precisely is a determined by the data?
Keyword: Confidence interval

ad (i)

- Intuition: Choose $\hat{a}$ in a way that $y = ax_i$ lies as close as possible to the data
- This means minimizing the distances, i.e.

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

Least squares estimator

ad (ii)

- Intuition: When σ^2 large, a is badly determined

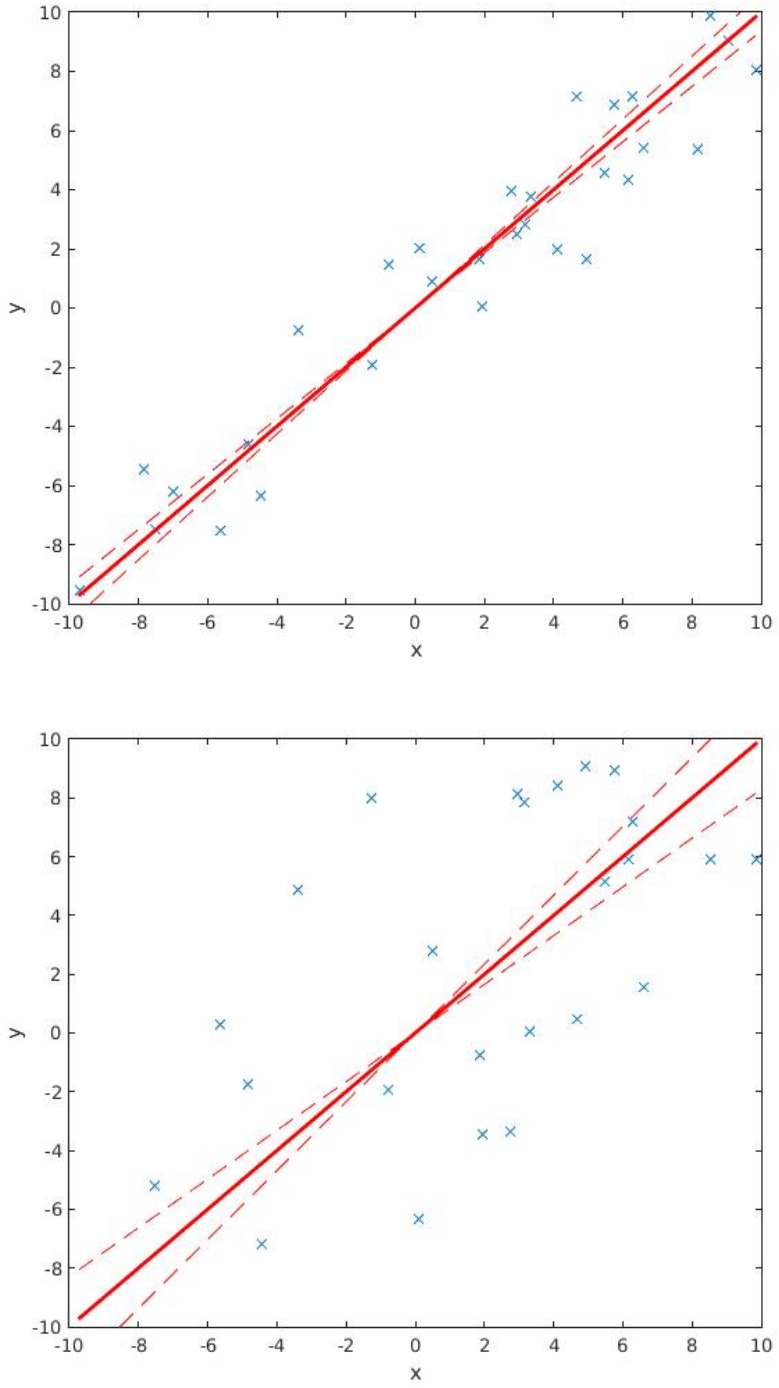

Figure 4.1: Linear regression on two data samples with (a) $\sigma^2 = 4$ and (b) $\sigma^2 = 36$

- If σ_i^2 is weighted instead of σ^2

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

Weighted least squares estimator

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

aka χ^2. For true a it is also distributed as such.

Examples for models:

- Regression models
  - Linear in parameters, nonlinear in x, the independent variable

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

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

    see Chap. 10.2
  - Nonlinear in parameters

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

    see Chap. 10.3
- Dynamical models
  - Partially observed ordinary differential equations

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

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

- Stochastic partial differential equation
- Time discrete state space model

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

  Keyword: Kalman-Filter

- Hidden Markov model, time discrete
  Discrete states $x_1, \ldots x_s$

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

  Transition probabilities

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

  Noisy observations

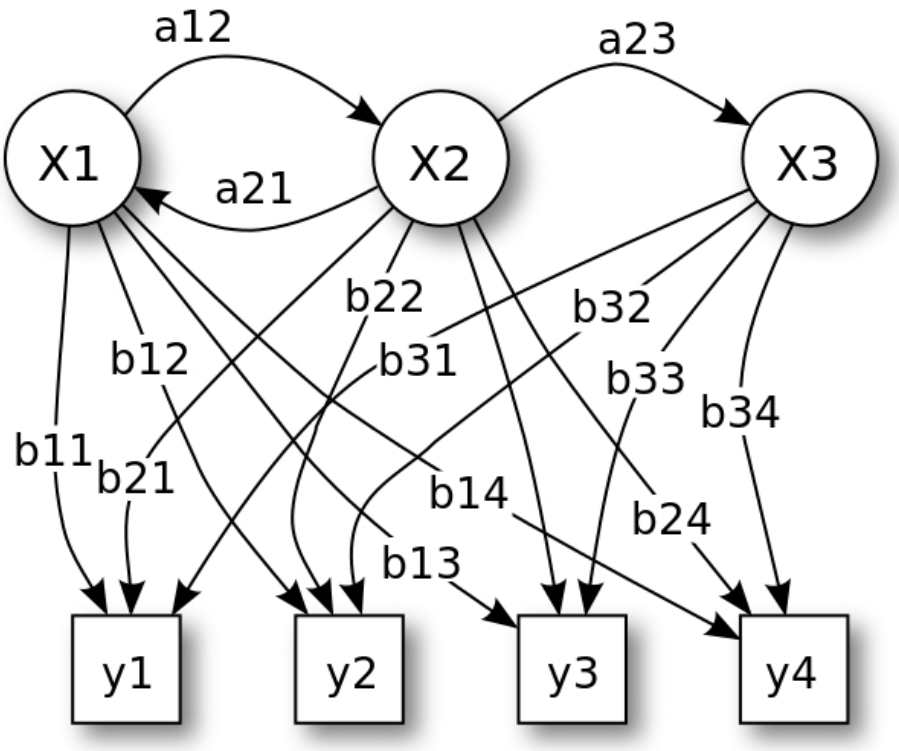

Figure 4.2: Hidden Markov model

Keyword: Baum-Welsh algorithm, Viterbi algorithm

- Particle physics

  Model for background in Higgs boson search: Highly complex calculations and simulations

Common feature of all models: They produce a probability $p(z, a)$ for observations z dependent on the parameter a

4.1 Maximum Likelihood Estimator

Remember:

- Bias (Distortion) : $\langle\hat{\Theta}\rangle - \Theta$
- Variance of the estimator : $\langle(\hat{\Theta} - \langle\hat{\Theta}\rangle)^2\rangle$, determines confidence interval
- Mean square error : $\langle(\hat{\Theta} - \Theta)^2\rangle$ = bias^2 + variance of the estimator

Let X be a parametric random variable with density $p(x, a)$.
Given N realizations

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

is called the Likelihood

- $L(x_1, \ldots, x_N|a)$ is to be read in dependence of a
- Data is given
- Likelihood: "For an assumed a, what is the probability given the data?"
- Likelihood is not a probability, since $\int p(x, a)\,\mathrm{d}a$ not normalized. As opposed to $\int p(x, a)\,\mathrm{d}x = 1$

Maximum Likelihood Estimator (MLE):

- Choose parameter a so that Likelihood is maxed
- Intuitively sensible
- Formally:
$$\frac{\partial L(x_1, \ldots, x_N|a)}{\partial a} = 0$$
- Logarithmic:
$$\mathcal{L}(a) = \log L(a) = \sum_{i=1}^{N} \log p(x_i, a)$$

  Since logarithm monotonous, the value of the maximum does not change.

Replaces difficult multiplication with manageable sum.

Usually addition of a minus sign, doesn't change value for maximum anyway.

Minimization of log-likelihood instead of maximization of likelihood

- M.k.z.: MLE under mild conditions, asymptotically unbiased. Proof of contradiction (Cox/Hinkley p. 288f, pretty)
- M.k.z.: MLE under mild conditions, asymptotically normal distributed.

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

with

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

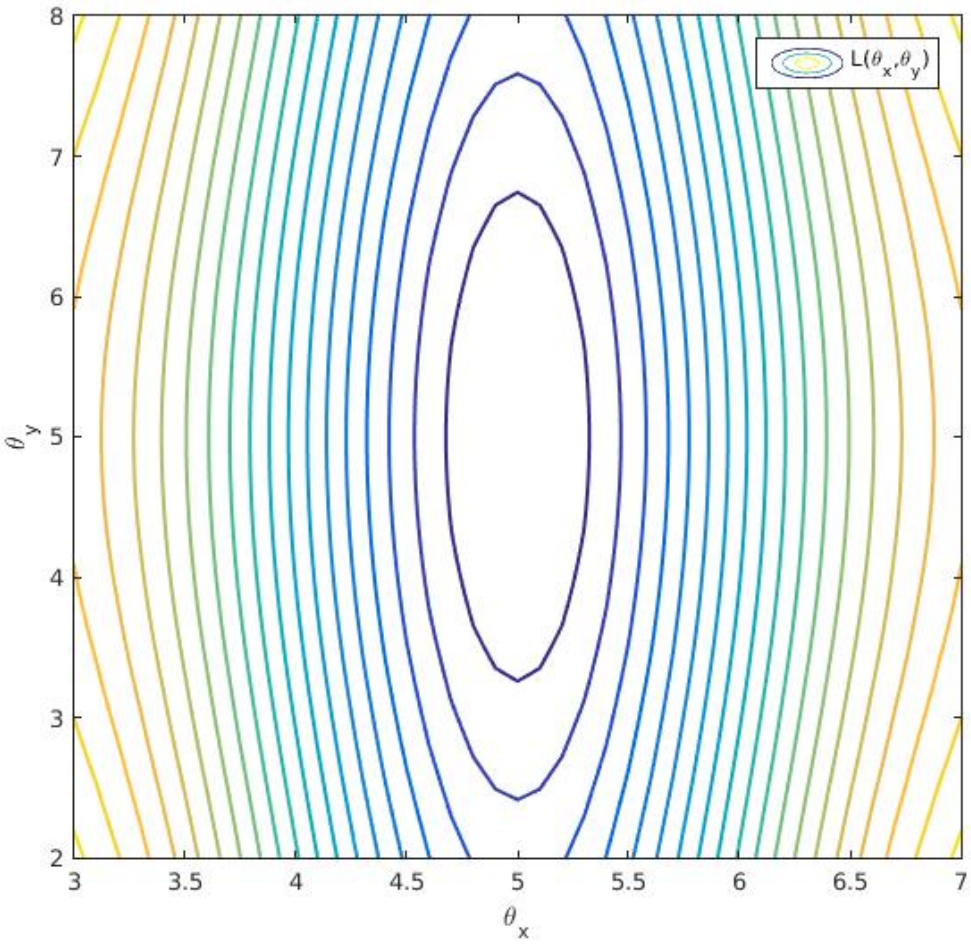

Figure 4.3: Maximum Likelihood Estimator in 2D

Cramér-Rao barrier

- In the following, all indices suppressed
- Consider score V:

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

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

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

Consider derivation with respect to a of

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

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

With Eq. (2) follows for 1. summant:

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

and thus:

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

$\left\langle -\frac{\partial^2}{\partial a^2} \mathcal{L}(x,a) \right\rangle$ called Fischer information matrix.

- Consider unbiased estimator $\hat{\theta}(x)$ for parameter a, i.e. $< \hat{\theta}(x) >= a$.

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

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

- Consider Cauchy-Schwarz inequality:

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

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

The Cramér-Rao barrier

- Curvature of Log-Likelihood determines estimator.

  Variance of estimator yields confidence interval.

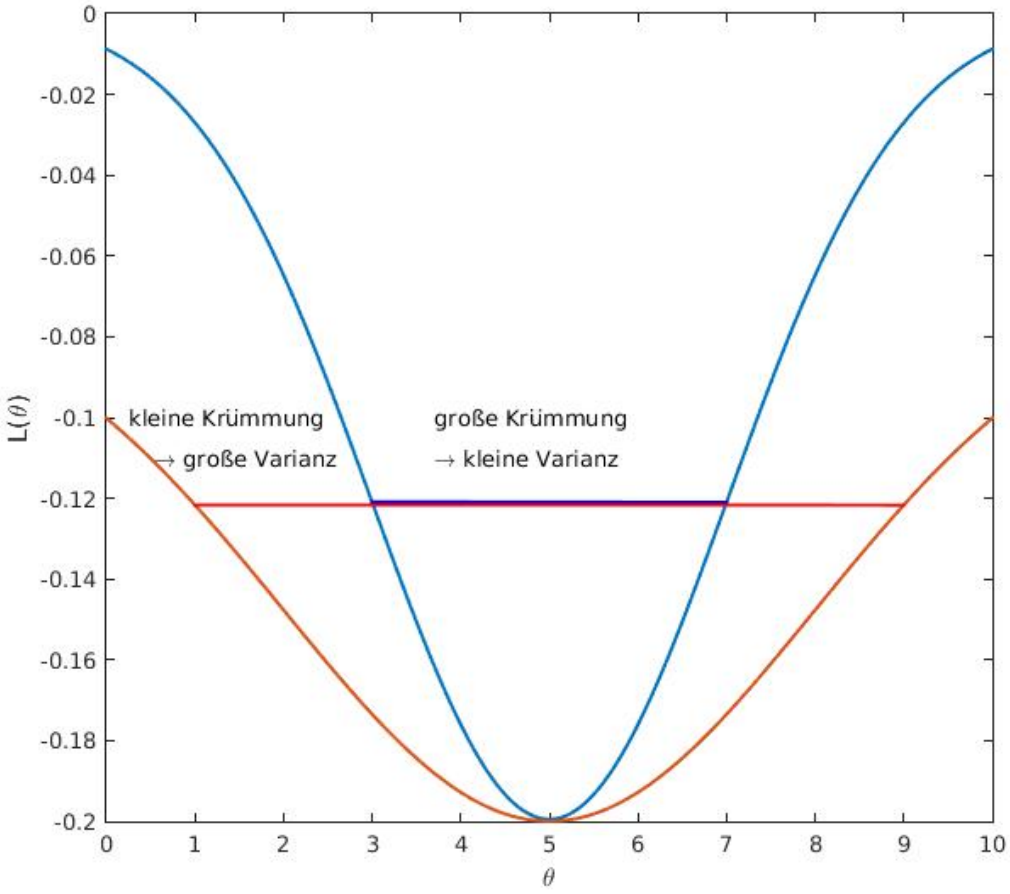

Figure 4.4: The Cramer-Rao-barrier for one parameter

- M.k.z.: Maximum Likelihood Estimator assumes lower limit, therefor

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

- Efficiency: Let $\hat{\Theta}_.$ be a non-MLE, then

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

MLE are top notch, retrieve the most information from the data.

Concrete examples:

- Normal distribution

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

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

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

  - Estimator for the mean

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

Therefor:

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

Variance of the estimator:

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

σ^2 dictates curvature of the Likelihood: Larger σ^2 yield smaller curvatures and thus larger variances of the estimators.

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

Typical $\frac{1}{\sqrt{N}}$-dependency

– Estimator for the variance

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

Remember Chap. 2.4: Unbiased estimator has $\frac{1}{N-1}$
MLE in general only asymptotically unbiased.
Calculation of $Var(\sigma^2)$: Home work

- Linear 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

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

Read from front to back: If one estimates based on (weighted) least squares, one has assumed a normal distribution

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

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

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

Further treated in chap. 10.1 and chap. 10.2

- Exponential distribution

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

4.2 Methods of Moments

- Likelihood sometimes difficult or impossible to calculate
- In those cases, Methods of Moments is an alternative
- Ansatz:

  Calculate moments μ_k ...

  - ... from the data: μ_k^{emp}
  - ... and from the model, parameterized theoretical moments $\mu_k^{theo}(\theta)$.
- Define estimator as:

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

  resp.

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

- As a rule

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

If the problem is linear in the parameters, the uncertainties are gaussian and considering first and second moments, it holds:

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

4. week

4.3 Bayesian approaches

Up until now frequentistic: There are true parameters
Bayesian world:

- Parameters are also random variables
- All probabilities are conditional probabilities

Conditional probability

$$p(A|B) = \frac{p(A,B)}{p(B)} \quad \text{probability for } A \text{ given } B$$

Consider dice: $A = \{1,2\}$, $B = \{1,2,3\}$

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

$$\begin{aligned} p(A|B) &= \frac{p(A,B)}{p(B)} \\ p(B|A) &= \frac{p(A,B)}{p(A)} \\ p(A|B) &= \frac{p(B|A)\,p(A)}{p(B)} \quad \text{Bayes theorem} \end{aligned}$$

With $A = \theta$ and $B = Daten$ and $p(Daten) = const$ follows

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

- The Likelihood $p(data|\theta)$ is decorated by the prior $p(\theta)$.
- The prior $p(\theta)$ is also a conditional probability, based on *prior*knowledge
- $p(\theta|data)$ is called a posteriori distribution
- Gives Maximum a posteriori (MAP) estimator and its distribution.

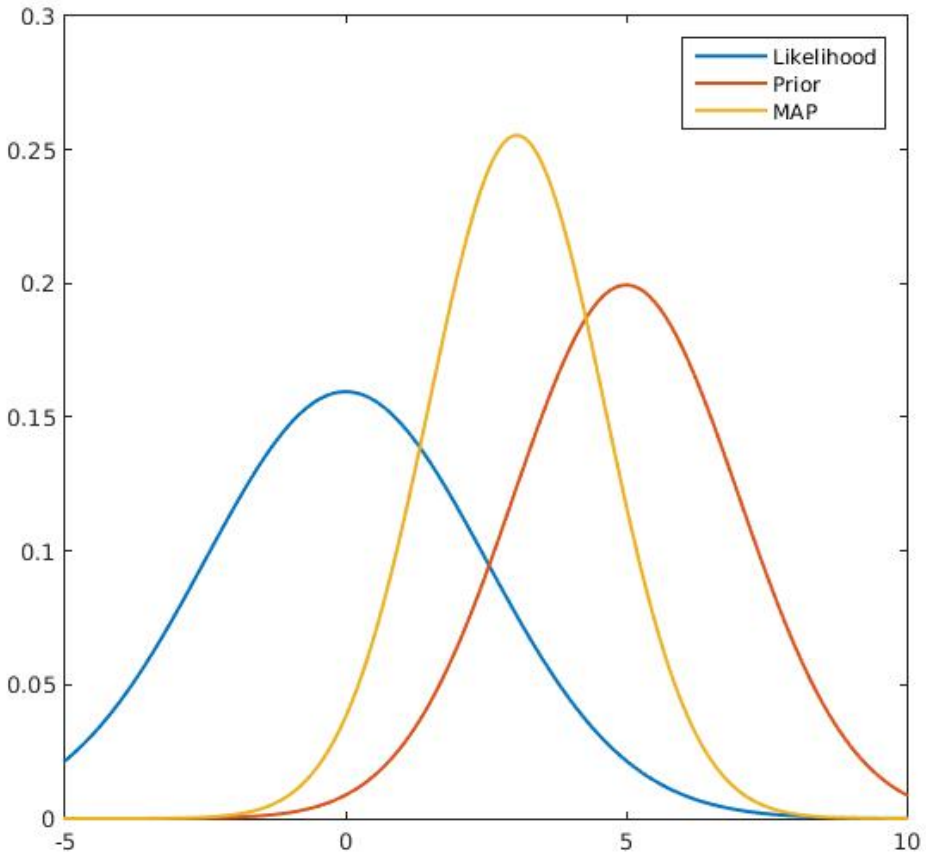

Figure 4.5: Influence of prior leads to bias, but smaller variance, but not in this Graph :-)

- Taking the logarithm of Eq. (3)

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

  Influence of Likelihood : $O(N)$, Influence of prior : $O(1)$,

  Asymptotically prior has no influence

- Problem: Prior usually unknown
- Gives (in frequentistic view) biased estimator in the infinite
- Advantage of Bayesian approach: Prior can introduce useful *prior* knowledge

  Accumulation of information through series of experiments, experimental design [38], empirical Bayes

Especially important in ill-posed inverse problems

- Simplest example:

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

  y is measured, x should be determined

- Is A singular or ill conditioned, i.e. almost singular, large

$$\text{Condition number} = \frac{\text{largest eigenvalue}}{\text{smallest eigenvalue}}$$

  x is estimated unbiased but estimator has huge variance and therefor large mean square error

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

- Prior can (strongly) reduce $Var(\hat{\theta})$ but leads to (small) Bias.

  Keyword: Regularization

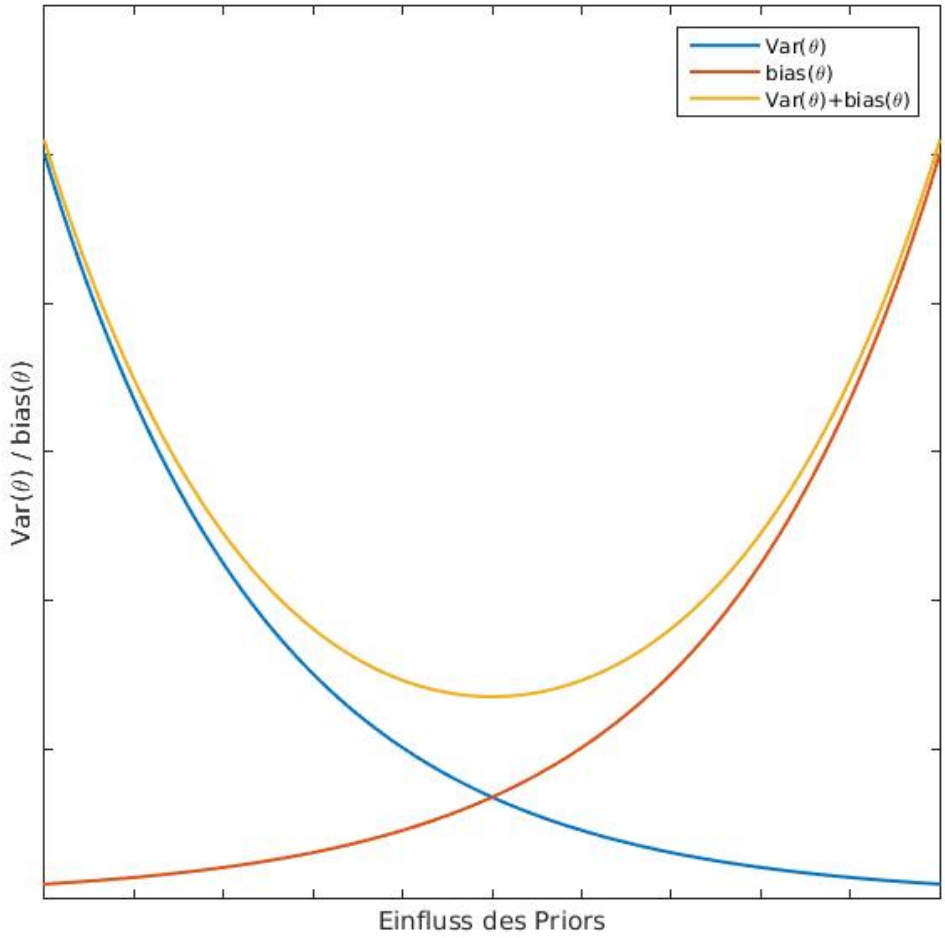

Figure 4.6: Behavior of bias and variance depending on the influence of the prior

Sensible priors:

- small $|a|$: $p(a) \propto e^{-|a|}$
- estimation of a function $f(x, a)$.
  Let $f(x, a)$ be smooth: $p(f(x, a)) \propto exp(-\frac{\partial^2}{\partial x^2} f(x, a))$

Calculation of $p(\theta|Daten)$

- Leads to complicated highly-dimensional integrals

- Monte Carlo Markov Chain Method [20, 29, 49], see Chap. 14
- Stochastic processes on the parameters
- Stationary density is the desired sample

4.4 Profile Likelihood

Confidence intervals ar based on the Fisher information matrix:

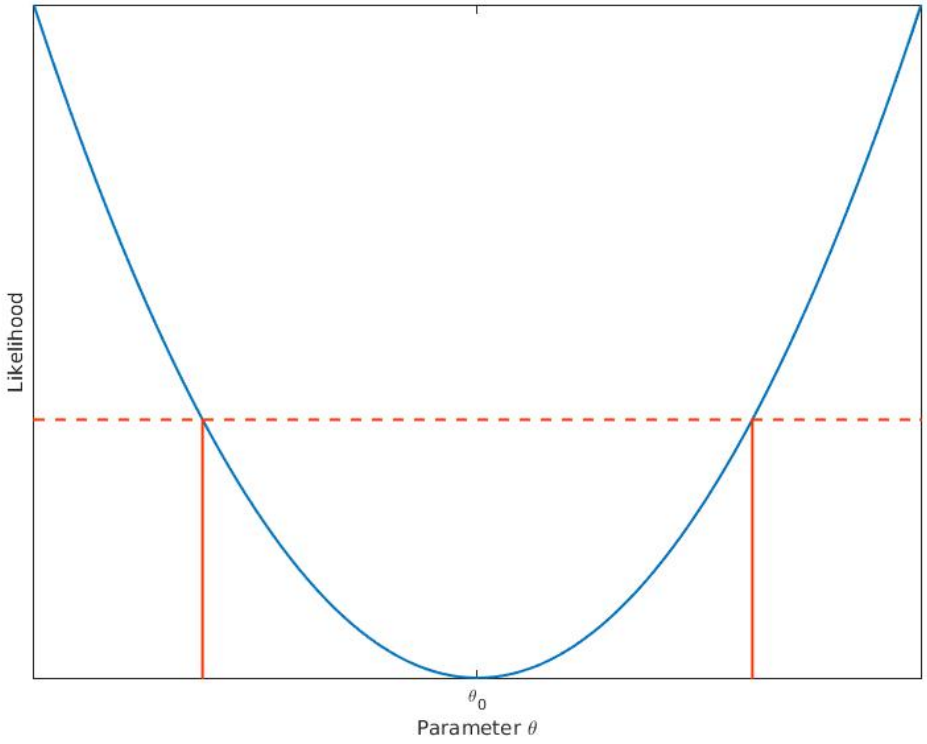

Figure 4.7: Profile Likelihood

- Strong assumptions on the asymptotic : quadratic approximation has to be true Only holds globally for linear models

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

  Otherwise only locally in the optimum

- If this holds , two possible statements:
  - Quadratic: Finite confidence intervals
  - Flat: Parameter not identified: Structural non-identifiability
    * Parameter can not be identified due to model structure
    * (Trivial) example
    $$y = (ab)x$$
    * (Highly) non-trivial examples i.e. in partially observed differential equations
- Not reparametrization invariant.

  By transforming a parameter, i.e. logarithm, confidence intervals do not change according to the transformation

Alternative: Profile Likelihood

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

Run along every parameter and optimize the others

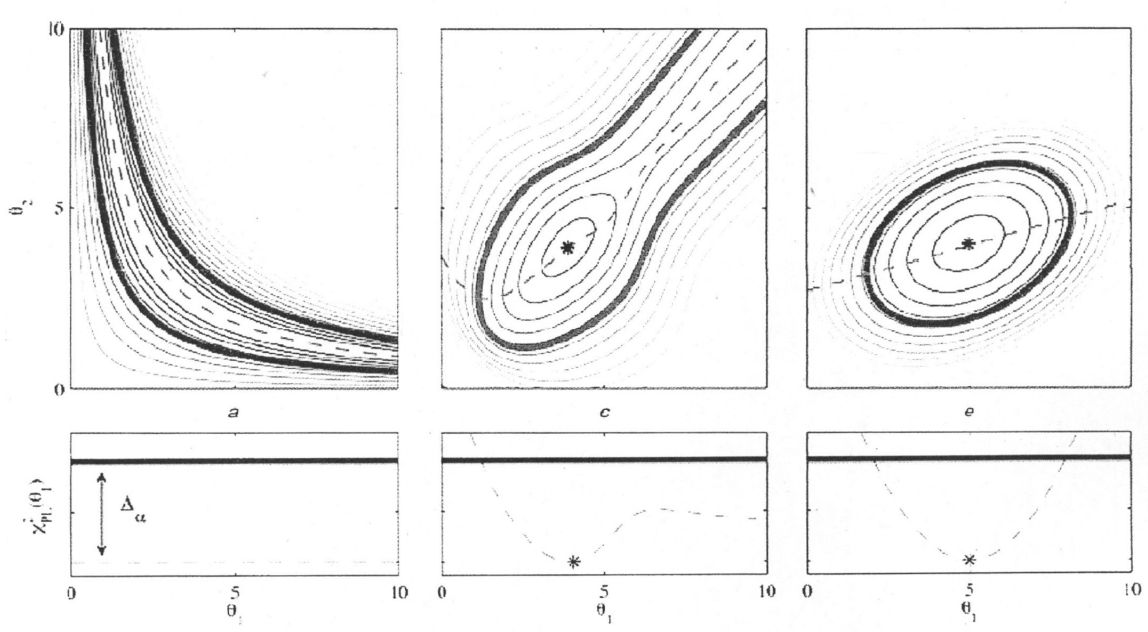

Figure 4.8: Profile Likelihood estimator,choice of confidence interval

Confidence interval given via:

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

Justification in chap. 5.2

Properties:

- Weaker asymptotic then Fisher information matrix based confidence intervals. Convexity of the Likelihood is sufficient.
- Reparametrization invariant
- Allows statements, if quadratic approximation is not valid
- Allows model reduction
- Allows experimental design

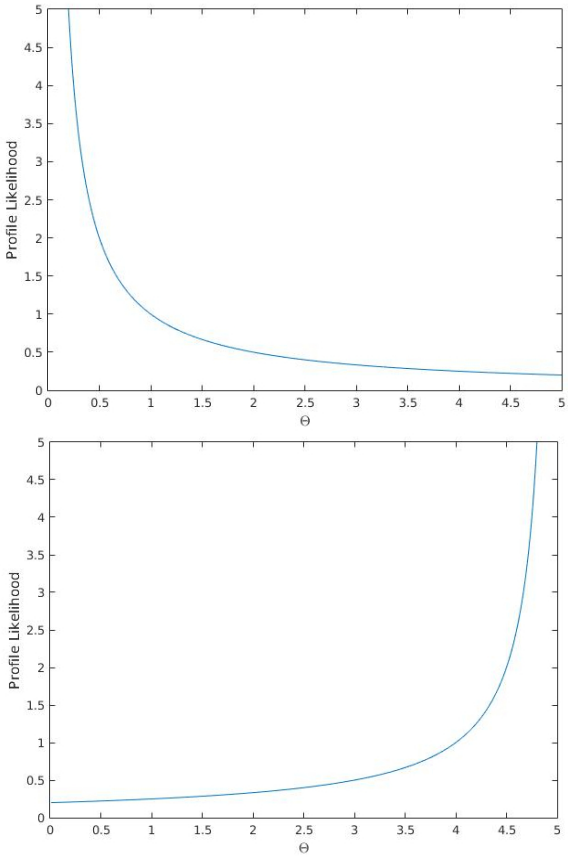

Figure 4.9: Possible courses lower and upper bound see exercise

- Allows definition of practical non-identifiability [54], i.e. problems which can be solved with additional data.

Lessons learned:

- Maximum likelihood Estimator is the best tool in the box
- Normal distributed error => MLE = weighted least squares
- Cramér-Rao barrier gives maximal possible accuracy
- Bayesian methods can consider prior information
- Profile Likelihood is highly informative alternative to asymptotic confidence intervals

5 Model selection

As a rule the true model is not known, but a finite number of candidate models

Two important cases

- Nested models
  - Let M_1 be a sub model with r_1 d.o.f. (θ_1) of
  - Higher model M_2 with r_2 d.o.f. (θ_2)
  - H_0: M_1 is an allowed simplification of M_2

  Easiest case:

  - M_1: 1. component of $\theta_1 = 42$
  - M_2: 1. component of $\theta_2 \subset R$
  - $r_1 = r_2 - 1$
- Non-nested models
  - M_1 and M_2 are competing for the explanation
  - M_1: $y = \sin ax$ vs. M_2: $y = \exp(bx) + cx^2$
  - M_1 and M_2 stand for different physics

Definition: Consistent Model selection method: For $N \to \infty$ the true model will be found with a probablity of 1 as long as it is part of the candidates

Occam's Razor: The simplest solution is usually the best.

All model selection methods

- Take into account the fact that a large model can always explain more
- Evaluate if the larger efforts are worth it

5.1 F-Test

Mother of all model selection tests:
Given

- Regression models, normal errors, least squares estimation problems
- Model M_1 with k_1 parameters, $\chi^2(M_1)$, d.o.f.: $N - k_1$
- Model M_2 with k_2 parameters, $\chi^2(M_1)$, d.o.f.: $N - k_2$
- Models nested, $k_2 > k_1$
- M_2 describes the data
- H_0: M_1 is an allowed simplification of M_2
- Unter H_0:

$$F = \frac{(\chi^2(M_1) - \chi^2(M_2))/(k_2 - k_1)}{\chi^2(M_2)/(N - k_2 - 1)}$$

  is F-distributed with $k_2 - k_1$ and $N - k_2 - 1$ d.o.f.

- Example
  - M_1: $y = a + bx$
  - M_1: $y = a + bx + cx^2$

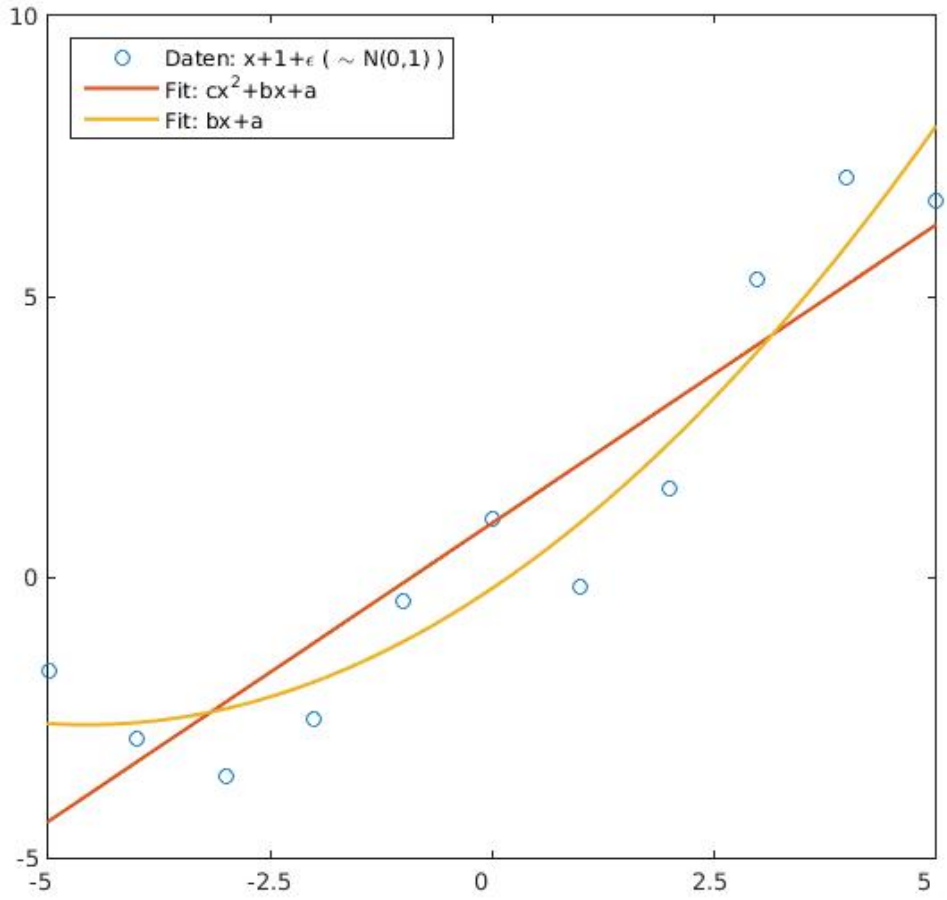

Figure 5.1: Noisy data of a linear course with fitted line and fitted parable

  - F-test measures amount of overfitting

- By reducing the significance level α with N, consistent selection method is obtained [4, 47]. Is H_0 true, it is not going to be rejected.

5.2 Likelihood Ratio Tests (LRT)

Best theory literature: [12]

Nomenclature:

- Given model M with parameter vectors $\theta \subset \mathbb{R}^r$.
- True parameter: θ_0
- Estimated parameter: $\hat{\theta}$
- $L = \mathcal{L}$

First LRT:

- H_0 : M is true
- H_1 : M is not true

Assumptions:

1. θ_0 does not reside on the edge of the parameter space.
2. The MLEs are asymptotically normal, e.g.:

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

   with

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

3. The model be identifiable, e.g. θ is uniquely determinable from the data, see identifiability in chap. 4.4.

Then holds asymptotically:

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

Difference of log-likelihoods is ratio of the likelihoods

proof (slight abuse of notation):

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

- 2. Term RHS = 0 because MLE.
- Neglecting terms of higher orders
- Σ^{-1} turns the correlations of the $\hat{\theta}$ out.
- Quadratic term become sums over r squared standard normal distributions $\Longrightarrow \chi^2_r$ -distribution
- Solve for $2(L(\hat{\theta}) - L(\theta_0))$
- Since $L(\theta_0)$ not known, more of theoretical interest but clarifies the principle.

Aber [72]:

- Estimate θ_0 from all data
- Many $\hat{\theta}$ from data fragments
- Test, if distributions $2(L(\hat{\theta}) - L(\theta_0))$ holds true

2. LRT: Given two models
Assumptions:

1. The models are nested, where M_2 is the sub model of M_1.
2. The higher model is correctly specified
3. The MLEs are asymptotically normal-distributed.
4. The true parameters do not lie on the edge of the parameter space
5. All parameters are identifiable under the null hypothesis.

H_0: M_1 is a valid simplification of M_2 Then holds asymptotically:

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

Proof:

- Analog to above
- Turning out the correlations

- Sum over squared standard normal distributions leads to χ^2-distributions

Comments

- Distribution of the LRTs follows from the asymptotic normality of the estimators
- LRT for regression case = F-Test, see e.g. [62]
- Consistent model selection method:
  For $N \to \infty$ and significance levels $\alpha \to 0$the true model will be found with probability = 1
- Related tests: Wald test, Lagrange-Multiplier test
- Profile Likelihood is LRT for one parameter therefor χ^2_1 distribution

5. week

In many models with growing complexity : Selection strategies for F-test, LRTs:

- Forward Selection
  - Test increasingly complicated models
  - Drawbacks:
    * False negative $\Longrightarrow$ Early stopping
    * There is for example no natural order in the non-linear
- Backward Selection
  - Starting from the most general model
  - Drawback
    * What is the most general model?
    * Existence of the Highest model
- Stepwise Selection

  After every forward step, perform a backward step.

  Is recommended.

5.2.1 Non standard test situations

"Non standard" means: Assumptions from above do not hold.

Most frequent case:

- Under H_0 parameter lies on the edge of the parameter space. [63, 73]
  Consequence: Estimator can not be normally distributed.
- Example: a scalar, parameter space $a \geq 0$
  - H_0: $a = 0$
  - H_1: $a > 0$

  Under H_0, instead of (asymptotic) normal distribution:
  - Potential negative values become 0
  - Potential positive values unchanged

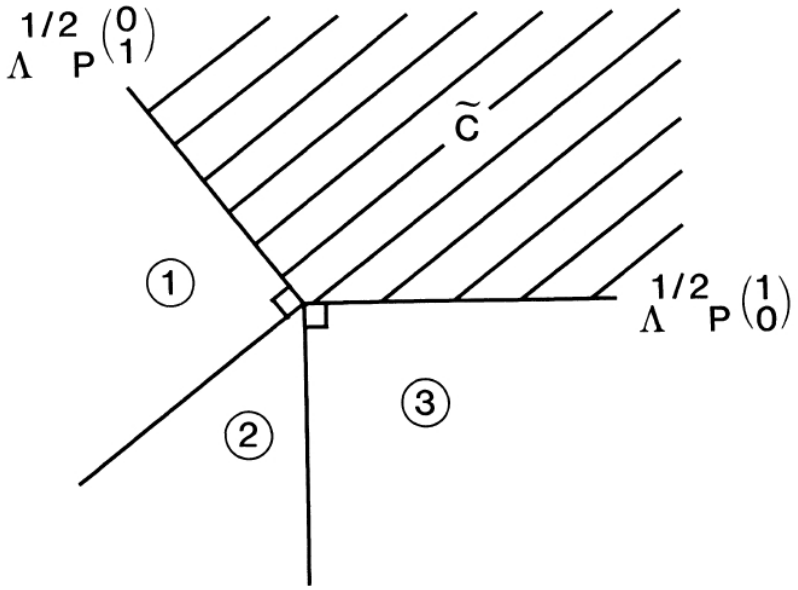

Figure 5.2: 2D-Parameter space: $\widetilde{C}$ shows the allowed parameter values under H_1. Under H_1 the parameter is localized around the origin. The asymptotic distribution of the LRT is a combination of χ^2_0, χ^2_1 and χ^2_2 with different probabilities dependent on the angle in $\widetilde{C}$. [63]

- Test statistic:

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

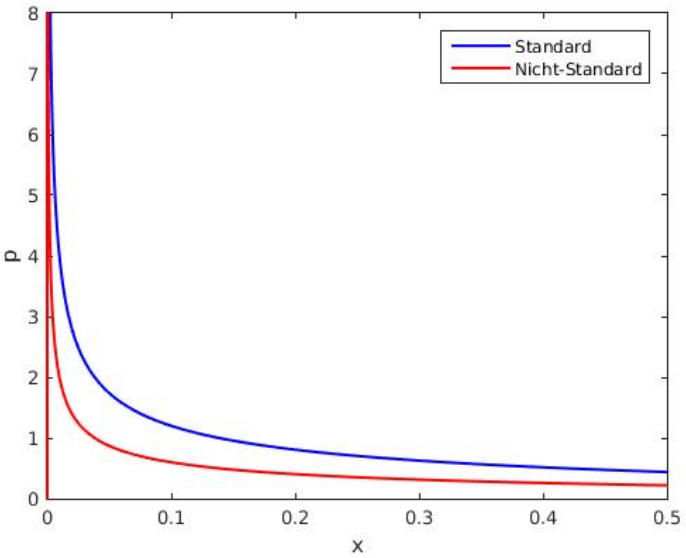

Figure 5.3: Normal vs. non standard

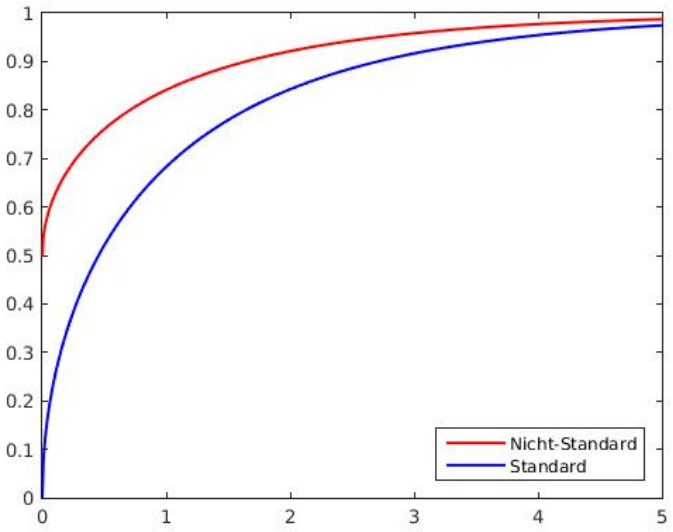

Figure 5.4: Cumulative distribution

- Important: If one does not consider "Parameter on the edge", Standard-LRT becomes conservative.

5.2.2 Non nested models [73]

If models are non nested one could take the higher model as a larger nest.
Prohibits itself in general since,

- Amount of identifiable parameters is limited, see chap. 4.4
- Then non standard situations would arise constantly

Simulative way out:

- Fit model 1 and 2 to the data, calculate the difference of their likelihoods
- Assume, model 1 is correct
- Simulate multiple data sets from model 1
- Fit both models to the simulated data
- Determine the distributions of the differences of their Likelihoods

- Check, if Originalfit-Likelihood difference is agreeable with simulated difference distributions
- Repeat for model 2
- Four possibilities
  - Both models will be rejected
  - Model 1 or model 2 will be rejected
  - No model can be rejected

5.3 Akaike Information Criterion (AIC)

Akaike himself called it An Information Criterion, AIC in [2] :-)
Original literature: [1, 2] nicely presented [37]

Principle:

- Unification of parameter estimation and model selection
- Based on the entropic measure, integrate and approximate.
- Formal analog to Cross-Validation [68]

Leads to:

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

For model selection, choose model with smallest AIC, no step wise procedure.

Comments:

- Popular because of it's simplicity.
- But: Consider nested models M_1 and M_2 with $\Delta k = 1$, in M_1 a parameter fixed

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

  Remember LRT:

Under H_0

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

Ergo: AIC is LRT with critical value α

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

In testtheoretical sense: 15.7 % error of 1. kind
Leads systematically to too complex models

- Not a consistent model selection method
- But good for finding models with high predictability
- Behavior for parameters on the edge and for non identifiability unclear
  Literature: [64]

5.4 Bayesian Information Criterion (BIC)

- Ingenues four page paper [61]
- Assumption: Weakest Bayesian priors and neglecting terms of higher order
- Yields

$$BIC = -2\log(Likelihood) + k\log(N), \quad k = dim(p)$$

considering amount of data

- Significance levels for a difference in the parameters [70]

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

- Choice of smallest BIC gives consistent model selection procedure
- Compare AIC vs. BIC, see [3, 7, 34, 45, 70]

Lessons learned:

- Model selection procedure evaluate higher explanation possibilities of more complex while considering the increasing number of parameters (Occam's Razor).
- F-test and Likelihood ratio test set scale, test statistics
- AIC and BIC simply order
- F-Test, LRT test and BIC are consistent model selection procedures
- AIC prefers systematically larger models than necessary

Part II
Numerics

There are two sorts of numerics:

- There is the one that one should understand, and...
- ... there is the one which one just has to know

6 Generation of random numbers

- Problem:

  How does one produce "random" numbers on a deterministic machine?

- Discussion : Detection of coincidence. Statistic hypothesis "5.6 is random" is not to be rejected .

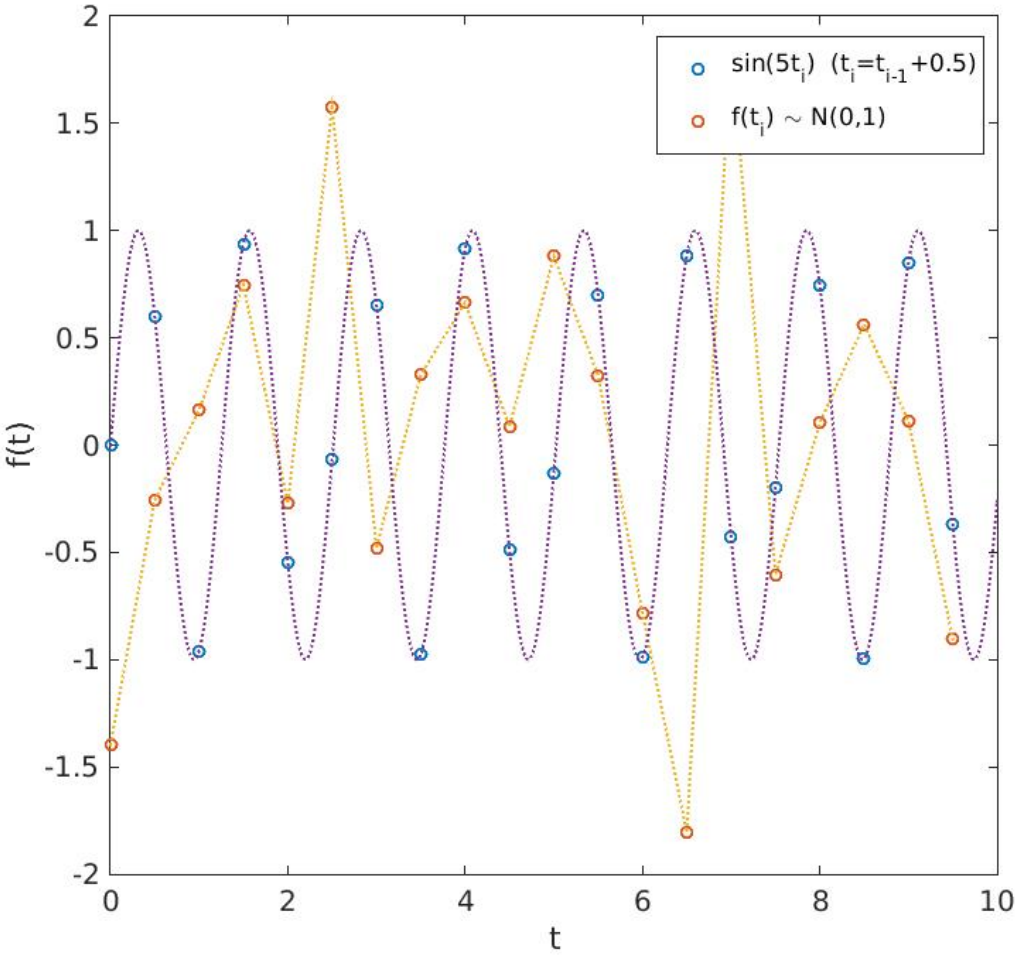

Figure 6.1: Sine series and white noise

- Coincidence = not predictable, de facto definition
- Solution :

  Chaotic dynamical systems show properties, which are not distinguishable from coincidence.

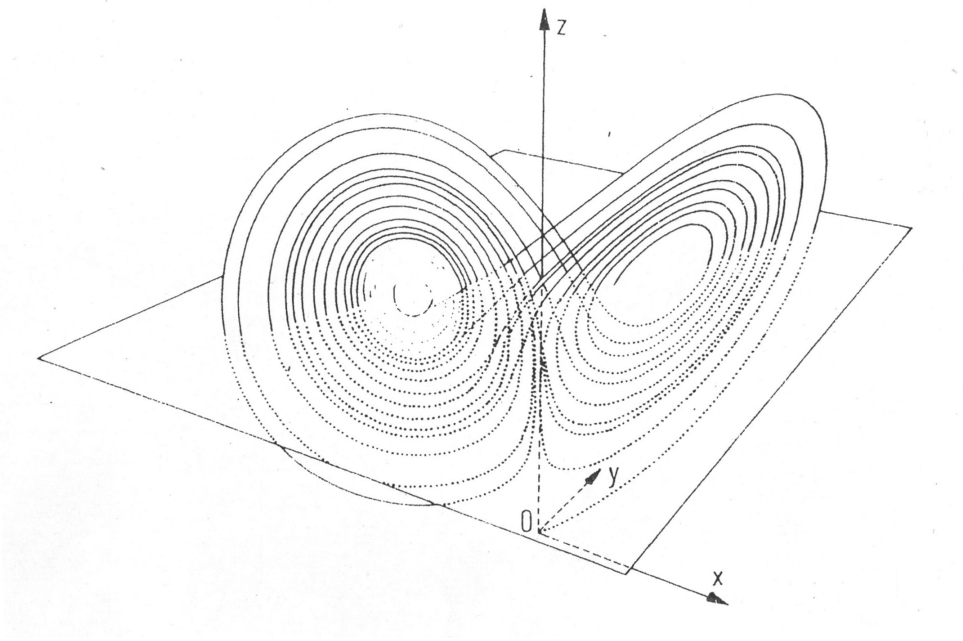

Figure 6.2: Lorenz 1963 'Deterministic Aperiodic Flow' [42]

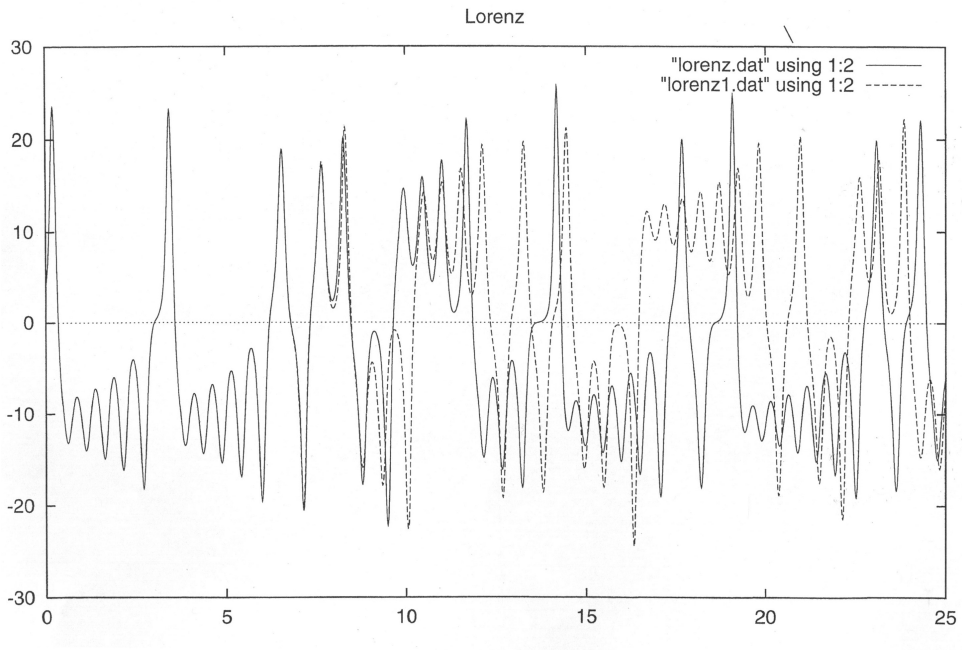

Figure 6.3: Lorenz 1963 'Deterministic Aperiodic Flow' [42]

- Figure Max(i+1)/Max(i)
- (Pseudo-)random number generator: Poincaré cut trough a high dimensional deterministic chaotic system.
- Similar values of $x(t)$ have very different values of $x(t+1)$.
- Replace by Dreieck (0.,1;1)
- All random generations are based on equally distributed random variables
- Replace by

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

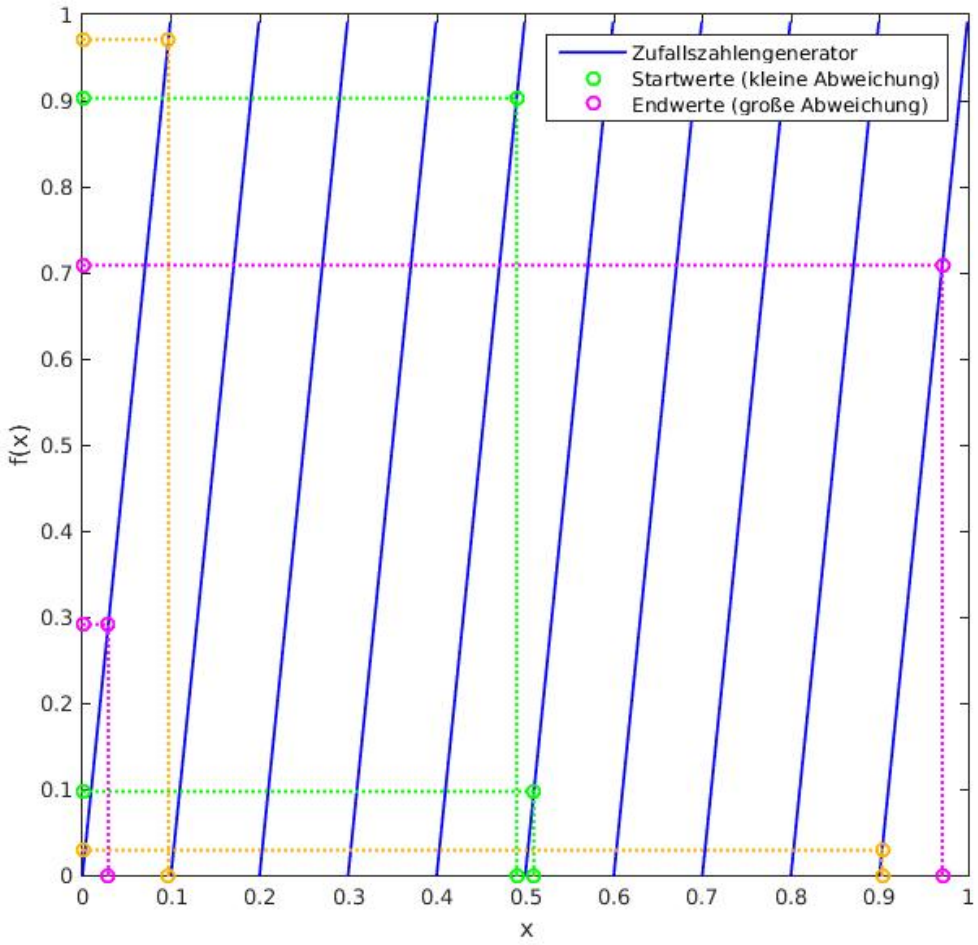

Figure 6.4: Principle of a random number generator

Transformation methods :

- Random variable X
- Form $Y(X)$
- It must apply

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

Ergo:

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

- As a rule:
  - X equally distributed
  - $Y(X)$ cleverly chosen
  - For distributions of Y it holds therefor

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

Examples :

- Exponentially distributed random variables

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

  - Let X be equally distributed
  - Choose $y(x) = -\log x$, $x = e^{-y}$
  - Yields:

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

- Standard normal distributed random variables

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

  - Here 2D transformation method, Box-Müller procedure

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

    $|.|$: Determinate of the Jacobi matrix
  - Let X be equally distributed

– Choose wisely:

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

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

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

– Gives 2 standard-normal distributed random numbers for 2 equally distributed ones.

– General normal distribution through shift about μ and scaling of σ.

- Cauchy

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

– Let x be equally distributed in $[-0.5, 0.5]$

– Then $y = \gamma\tan\pi x$ is Cauchy-distributed

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

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

– It also holds:

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

Remember: Ratios of random variables can be gnarly

Lessons learned:

- Random number generation in deterministic computers is based on non-linear dynamic
- Equal distribution is the mother of all random numbers
- The rest is generated for example through transformation method

7 Solution of linear equation systems

Given matrix A and vector b, find vektor x for:

$$Ax = b$$

Ubiquitous problem:

- Physics: Scattering experiments, Rheology
- Numeric, see Chap. 9 Optimization, Chap. 10 Non-linear modeling.
- b usually uncertain
$$Ax = b + \epsilon$$

Goal :

$$x = A^{-1}b, \text{ or somtimes: } \tilde{x} = \tilde{A}^{-1}b, \quad \tilde{A}^{-1} \text{ modified } A^{-1}$$

Problems and methods differ depending on the properties of the matrix A:

- A be $N \times N$ matrix (most important case)

  Good chance for unique solution.
  Possible problems:

  - Linear dependency on rows/columns of A
    - Matrix singular $\Longrightarrow$ No unique solution.
  - "Almost" linear dependency
    - Matrix ill-conditioned.
    - Let λ_i be the eigenvalues sorted in descending order:
      Condition number K:
      $$K = \frac{\lambda_1}{\lambda_N}$$
    - Large K: Uncertainty on b is going to be reinforced in solution x, see below.
  - N very large:
    - Rounding errors can cumulate.

- A be $M \times N$ matrix, $M < N$ (ar A be singular $N \times N$ matrix)

  - Under determined equation system.

- Solution not unique.
- Solution can become unique under additional assumption, see below.

- A be $M \times N$ matrix, $M > N$
  - Over determined system of equations.
  - Search for compromise which fulfills both equations as good as possible simultaneously .
  - For „as good as possible“in the sense of m.s.e. the unique solution is given by:

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

    $(A^T A)^{-1} A^T$ is called Pseudo-Inverse or also Moore-Penrose-Inverse. Treated in Chap. 10 Non linear modeling.

7.1 Gauß-Jordan - Elimination

A be $N \times N$ matrix, well conditioned.

- Basics:

  Formation of linear combinations of the system of equations does not change the solution.

- Idea:

  Bring system to an upper triangular form.

  Let:

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

  be the ith row of the system.

- Eliminate x_k in $E_{k+1}, \ldots, E_n$ through:

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

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

$a_{kk}^{(k)}$ is called Pivot element.

- Result:

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

- Solution x by Back substitution
  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:
  When $a_{kk}^{(k)}$ small, this leads to rounding errors in

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

- Solution:
  - "Equalize" the matrix A
  - Sort the rows beforehand, in a way that the numbers on the diagonal have increasing order.
  - Called Pivoting, detailed discussion in perturbation Chap. 4.5
- Complexity: $\mathcal{O}(N^3)$
- Drawback: Has to be recalculated for every b

7.2 Matrix decompositions

Matrix decompositions simplify life

7.2.1 LU decomposition

$N \times N$ matrix can be written as:

$$A = LU$$

with

- L: under triangular matrix (TM) with ones on the diagonals
- U: (arbitrary) upper TM.
- Crout's algorithm makes decomposition elegant (with pivoting).
- Stable for well conditioned matrix
- Complexity: $\mathcal{O}(N^3)$

Applications:

- Solution for $Ax = b$ by forward and backward substitution

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

  Decomposition only has to calculated only once for different b

- Calculation of A^{-1} by:
  for $(j = 1, \ldots, N)$

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

  Solutions $x_i^{(j)}$ given columns of A^{-1}.

- Economical calculation of the determinant through

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

  with $\mathcal{O}(N^3)$ instead of $\mathcal{O}(N!)$ in the definition-based calculation.

6. week

7.2.2 Cholesky decomposition

- Let A be symmetric and positive definite i.e. $vAv > 0 \; \forall \, v \neq 0$
- "Root" of the matrix:

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

- Covariance matrices fall under this category.

Generation of correlated Gaussian random variables:

- Covariance matrix, for $< x_i > = 0$

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

  Build:

$$BB^T = C$$

- Create uncorrelated RVs y_i and form correlated ones by:

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

- Proof:

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

7.2.3 Singular value decomposition (SVD)

A be $N \times N$ matrix, ill conditioned

- Condition number revisited:
  Condition number K:

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

  with $||\vec{y}||$ Euclidean norm, results in spectral norm $||A||$

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

- As always:

$$Ax = b$$

- Influence of errors Δb of b on the estimated $\hat{x} = x + \Delta x$:
  - Consider:

$$A(x + \Delta x) = b + \Delta b$$

- From
$$\Delta x = A^{-1}\Delta b$$
follows the estimation:
$$||\Delta x|| \leq ||A^{-1}||\,||\Delta b||$$
- For the relative errors $||\Delta x||/||x||$ follows with
$$||b|| = ||Ax|| \leq ||A||\,||x||, \quad \frac{1}{||x||} \leq \frac{||A||}{||b||}$$
all in all:
$$\frac{||\Delta x||}{||x||} \leq ||A||\,||A^{-1}||\frac{||\Delta b||}{||b||} = K(A)\frac{||\Delta b||}{||b||}$$
- Ergo, large K increase of the errors on b.
- $K = 10^6$ is disastrous in single precision.

- Singular-Value-Decomposition (SVD), Karhunen-Loéve-Transformation, main component analysis
  - Yields:
$$A = U\,[\mathrm{diag}(w_i)]\,V^T \quad ,$$
with
    * orthogonal U, $N \times N$ matrix
    * diagonal $N \times N$ matrix W with singular values $w_i \geq 0$, sign in U and V absorbed [2]
    * orthogonal V, $N \times N$ matrix
  - For the math, see Stoer, Bulirsch [67] Chap. 6.7
  - Inverse :
$$A^{-1} = V\,[\mathrm{diag}(1/w_i)]\,U^T$$
  - Solution of $Ax = b$
$$x = V\,[\mathrm{diag}(1/w_i)]\,U^T b \tag{4}$$

[2] If A is symmetric, the singular values are identical to the EV

- Advantage over Gauß-Jordan: b must not be known beforehand.
- Belongs to the 5 most important routines there are.
- Also works for $M \times N$ matrix, $M < N$.

Consider: A be $N \times N$ matrix, singular or ill conditioned

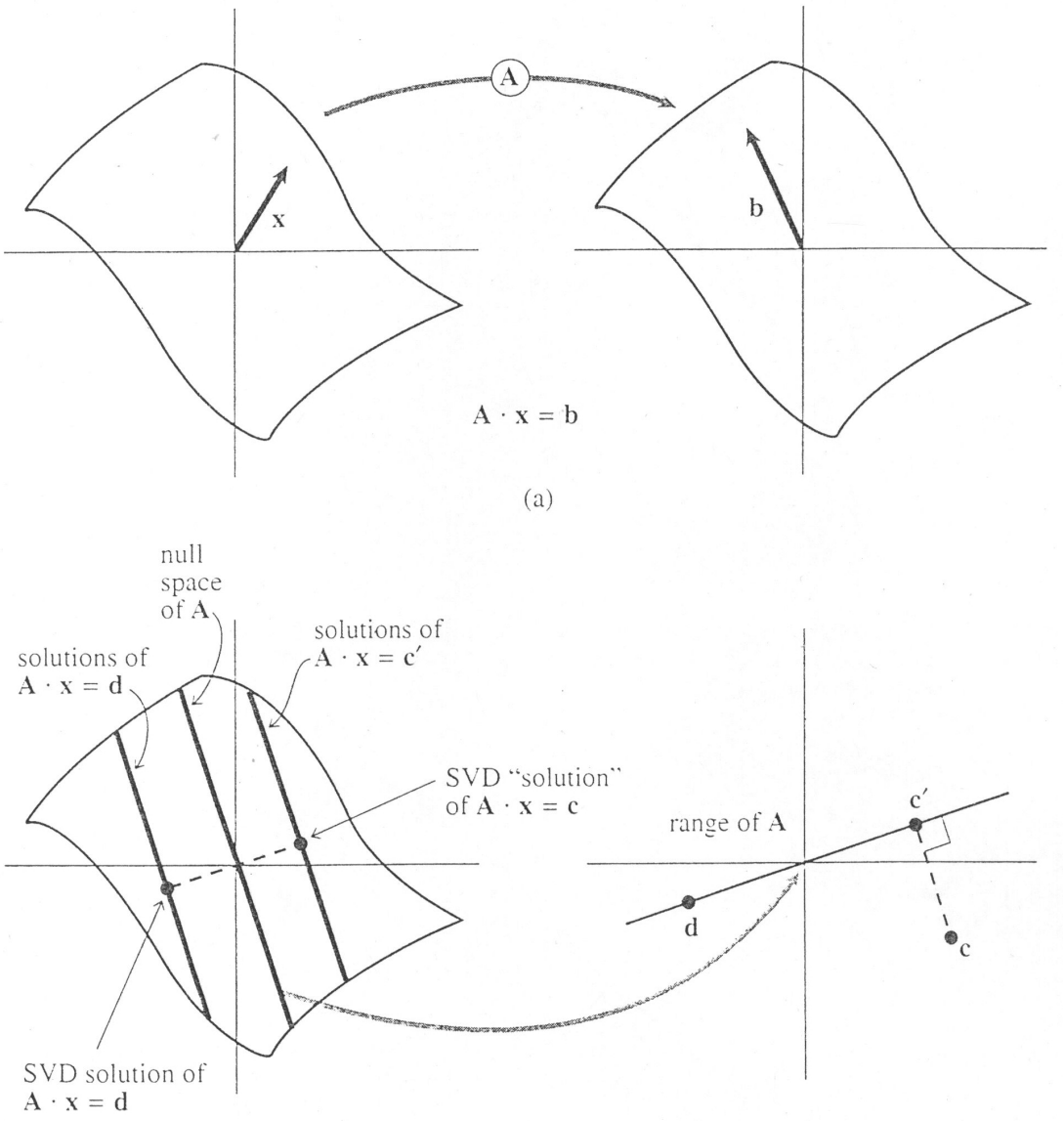

Figure 7.1: Singular value decomposition

- Graph:
  - The Eigen vectors of the 0- (or smaller) EV pose the problems in the inversion

  - Lead to large errors.
  - Solution: For small EV w_i, set in Eq. (4) $1/w_i = 0$ $(\infty = 0$:-))

- Mathematically:
  - x is estimated under the minimal norm.
    * A singular. The exist the core x_k with

      $$Ax_k = 0$$

      Range of A has dim $< N$
    * With x_{nk} not belonging to the core

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

    * Chosen solution : $x = x_{nk}$
  - Remember Bayesianisme, keyword: Regularization:
    * Additional information, to make the solution unique.
      Here :

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

    * Regulization entails:
      Reduction of the variance at the cost of a bias, see Excercises

- Minimum norm equivalent to:

  Search for solution for which holds:

  $$\text{Search } x, \text{which minimizes } r = ||Ax - b||^2$$

Remark : All treated algorithms have expense $\mathcal{O}(N^3)$

There are special methods for:

- Weakly occupied large matrices. Stoer/Bulirsch Kap. 8
  effort $\mathcal{O}(N)$ or $\mathcal{O}(N^2)$
- Inverse for "slightly changed" matrices, Recipes Chap. 2.7:
  - Sherman-Morrison equation

  - Woodbury equation

- Specially structured matrices, Recipes Chap. 2.8 :
  - Matrices with band structure (e.g. in finite element methods)
  - Vandermonde matrices $a_{ij} = \alpha_i^{j-1}$
  - Toeplitz matrices $a_{ij} = \alpha_{i-j}$
- Estimation of EV and Eigen vectors
  Recipes Chap. 11, Stoer/Bulirsch Chap. 6
  - Givens- and Householder reductions
  - $A = QR$ decompositions, Q orthogonal, R upper TM
  - Hessenberg form, populated from the first lower diagonal

Exercise:
Bias and variance in the solution of ill posed inverse problems.

Exercise:
Generation of correlated Gaussian random vectors

Lessons learned:

- Gauß-Jordan Elimination
- Different decompositions, which can simplify one's life:
  LU, Cholesky, SVD, ...
- SVD delivers minimal norm solution in ill posed problems

8 Zero point search

- Task: Given $f(x)$, estimate x_0, for:

$$f(x_0) = 0$$

- Usually only works iterative
- Important term:

  Order of convergence of iterative algorithms, also important for Chap. 9 Optimization and Chap. 10 Non linear modeling.

> Let $\epsilon(i)$ be the remaining uncertainty after i iterations. then the order of convergence γ is defined by:
>
> $$\lim_{i \to \infty} \epsilon(i+1) = \text{const } \epsilon(i)^{\gamma}$$

One dimensional case
Bisection

- Choose two points x_l and x_r, which enclose the zero point i.e. $f(x_l)f(x_r) < 0$
- Determine $x_{center} = \frac{x_r - x_l}{2}$
- Replace starting point with the same sign as $f(x_{center})$ by x_{center}.
- Iterate this until desired precision is reached.
- Evolution of uncertainty:

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

  thus linear order of convergence γ.

- Number n of necessary iterations for desired accuracy ϵ at initial uncertainty ϵ_0 :

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

- Globally convergent,but slow.

Secant method

- Requires sufficient linearity.

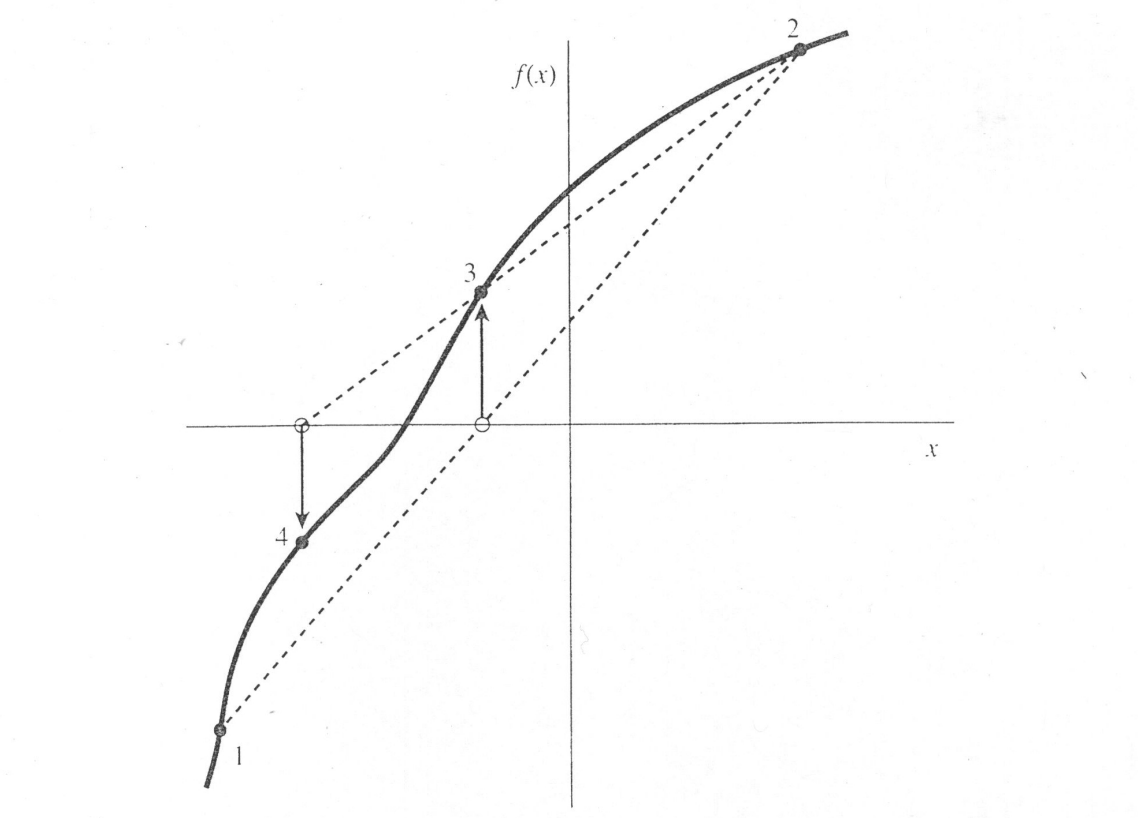

Figure 8.1: Secant method

- Iteration:

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

- It holds:

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

therefor super linear convergence γ

- Zero point not necessarily enclosed $\Longrightarrow$ secant method can diverge

Regula falsi

- Like secant method, but discard x_l or x_r depending on whether $f(x_l)f(x_{i+1}) > 0$ or $f(x_r)f(x_{i+1}) > 0$

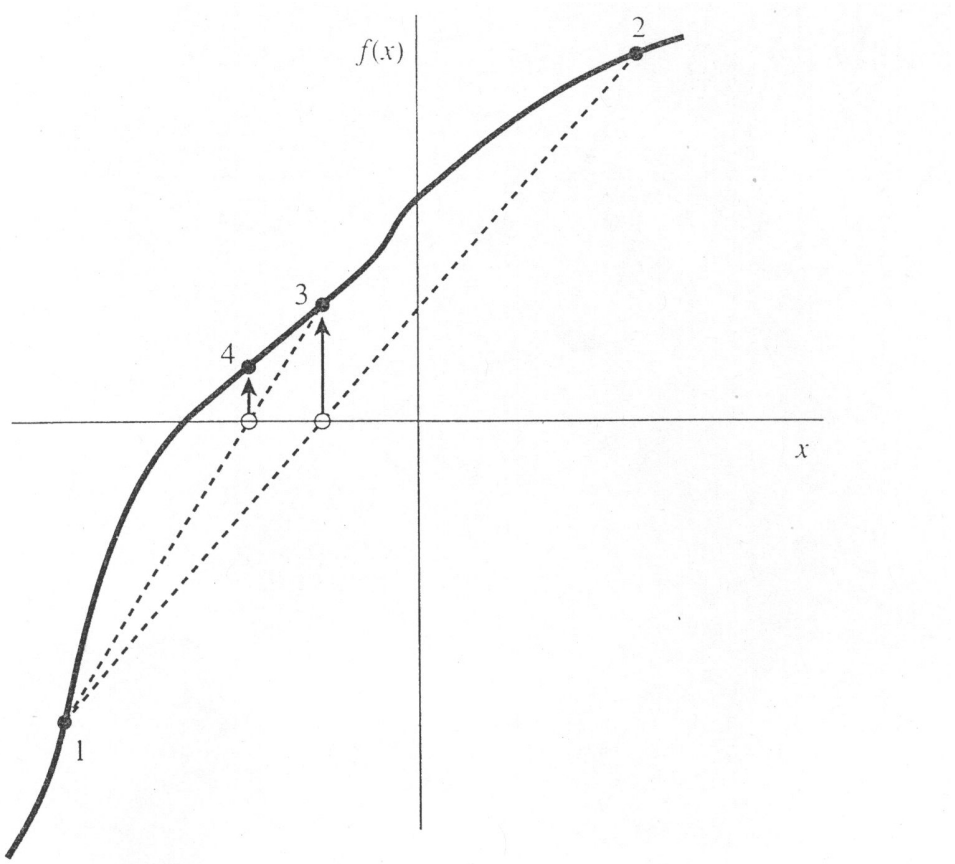

Figure 8.2: Regula-Falsi method

- Convergence order $\gamma \geq 1$, in general slower than secant method, but safe
- Secant method and regular falsi can be very slow in finite.

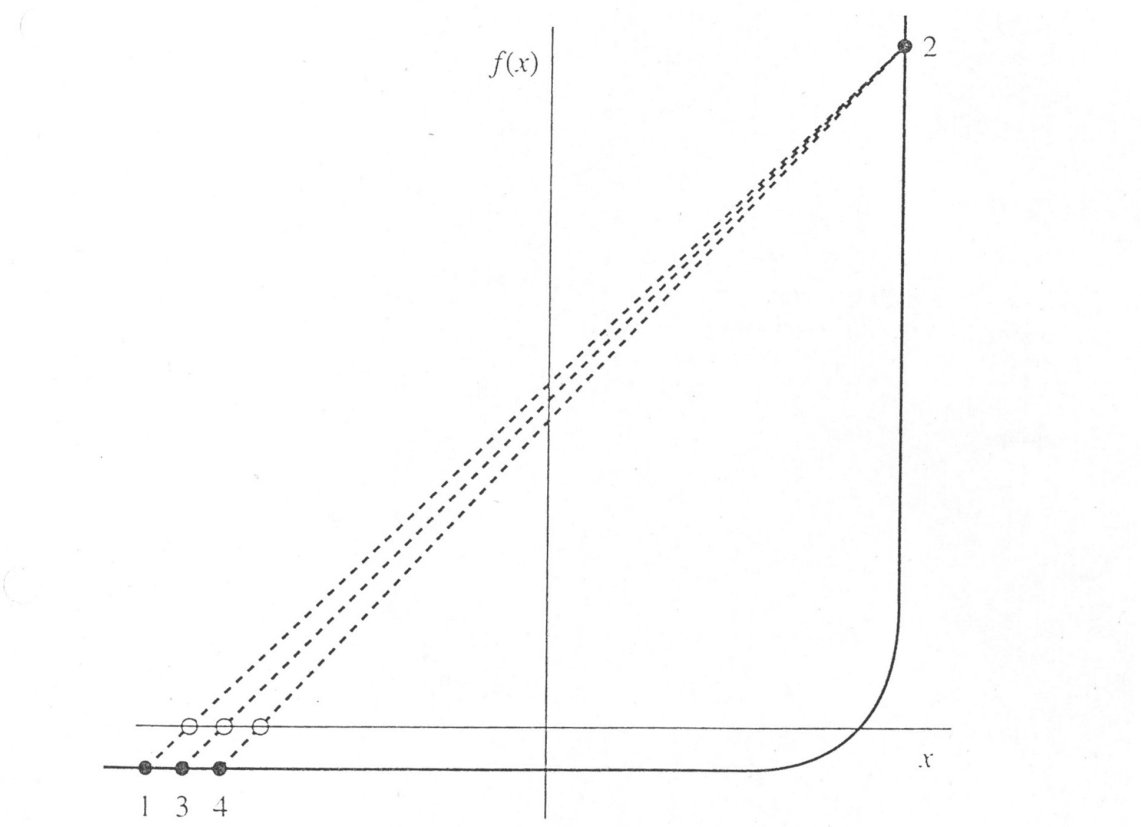

Figure 8.3: Example where secant and regula falsi method need many iterations

Newton-Raphson

- Uses and needs 1. derivation
- Idea: Taylor evolution:

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

- Close to the zero point $f(x_i + \delta) = 0$, $\delta^2 \ll 1$, everything well behaved, follows

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

- Determine order of convergence

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

  - Taylor evolution for $f(x_i)$, $f'(x_i)$ around zero point x_0 yields with all indices suppressed :

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

  - Introduce into

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

    yields:

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

    Expand:

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

  - With $\epsilon_i f''(x) \ll f'(x)$ follows:

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

    quadratic order of convergence

  - But only locally convergent

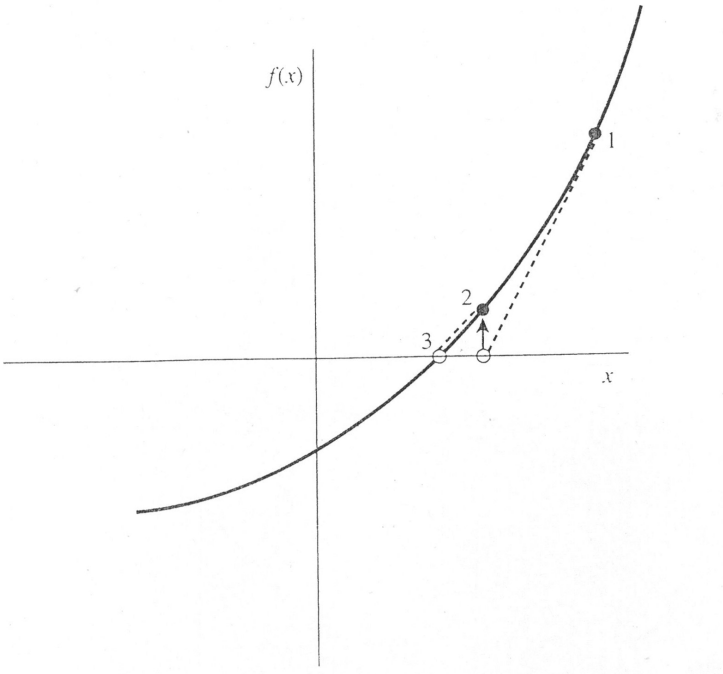

Figure 8.4: Newton-Raphson method converges

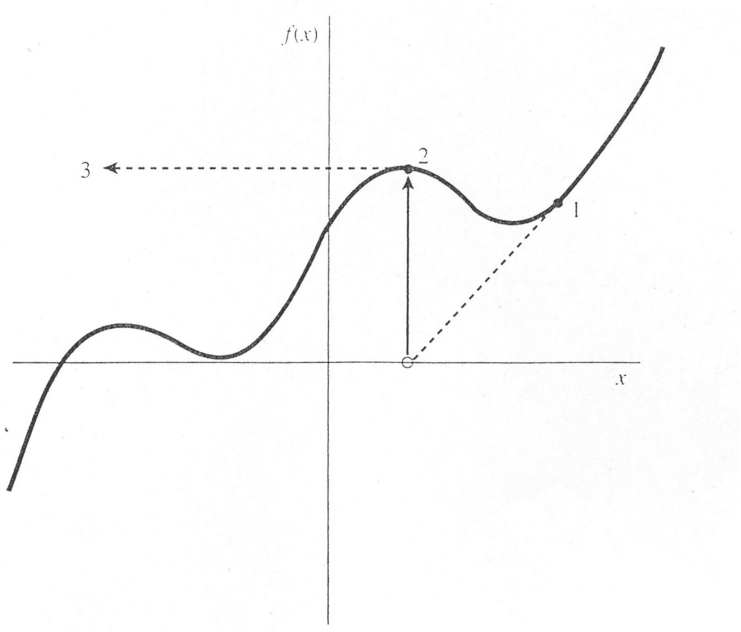

Figure 8.5: Newton-Raphson method divergences

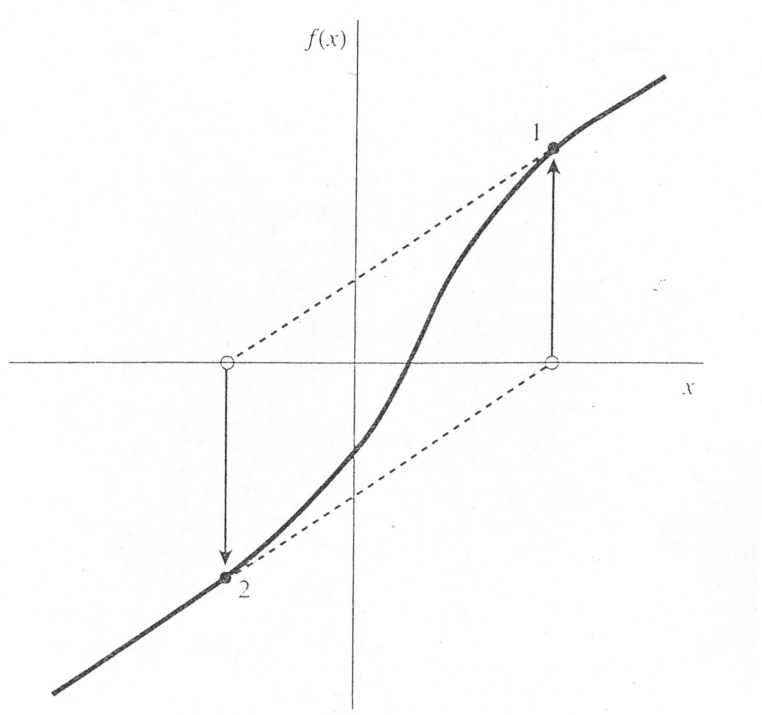

Figure 8.6: Newton-Raphson method unfortunate starting point

- Good for focusing: Start with bisection, then use Newton-Raphson

Schmankerl, Chaos theory revisited:
Find solution of

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

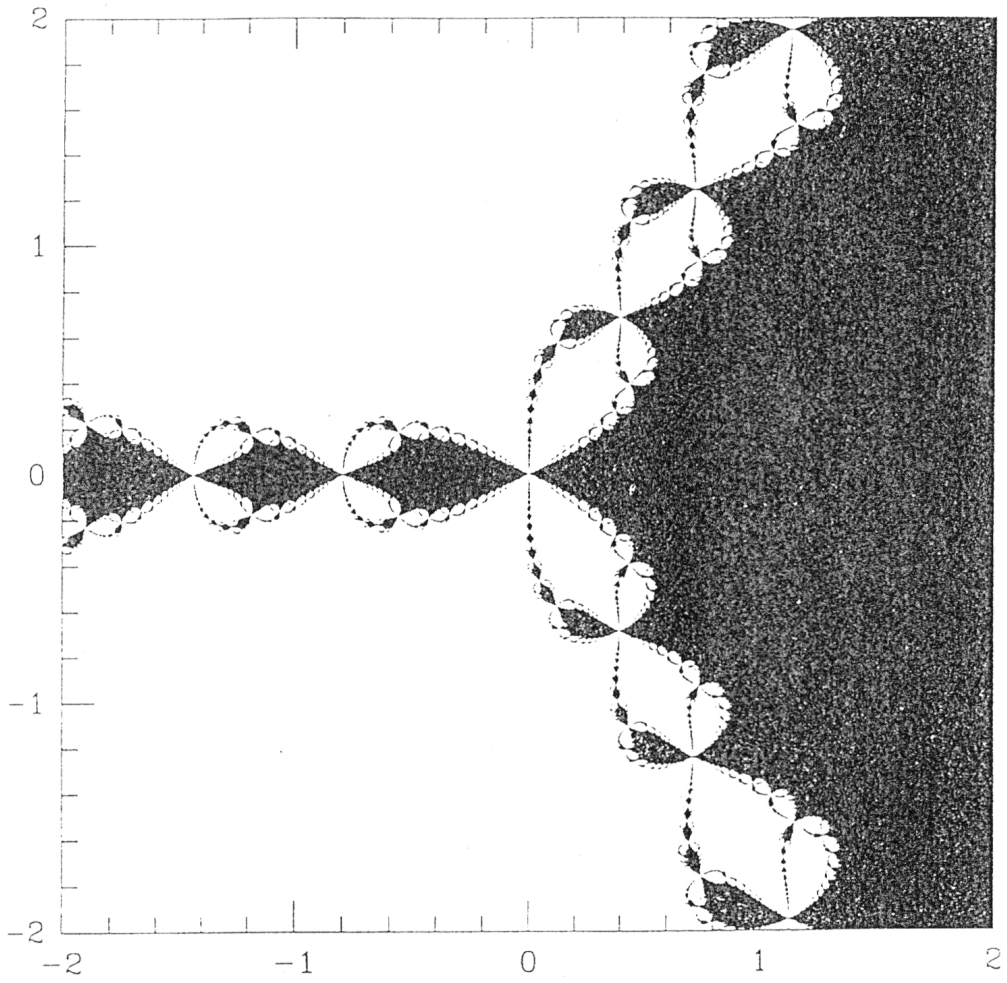

Figure 8.7: Fractal: In the black region, the Newton-Raphson method converges to $z = 1$.

Higher dimensional case

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

Hairy problem, e.g. number of solutions not clear a priori, see Recipes Chap. 9.7

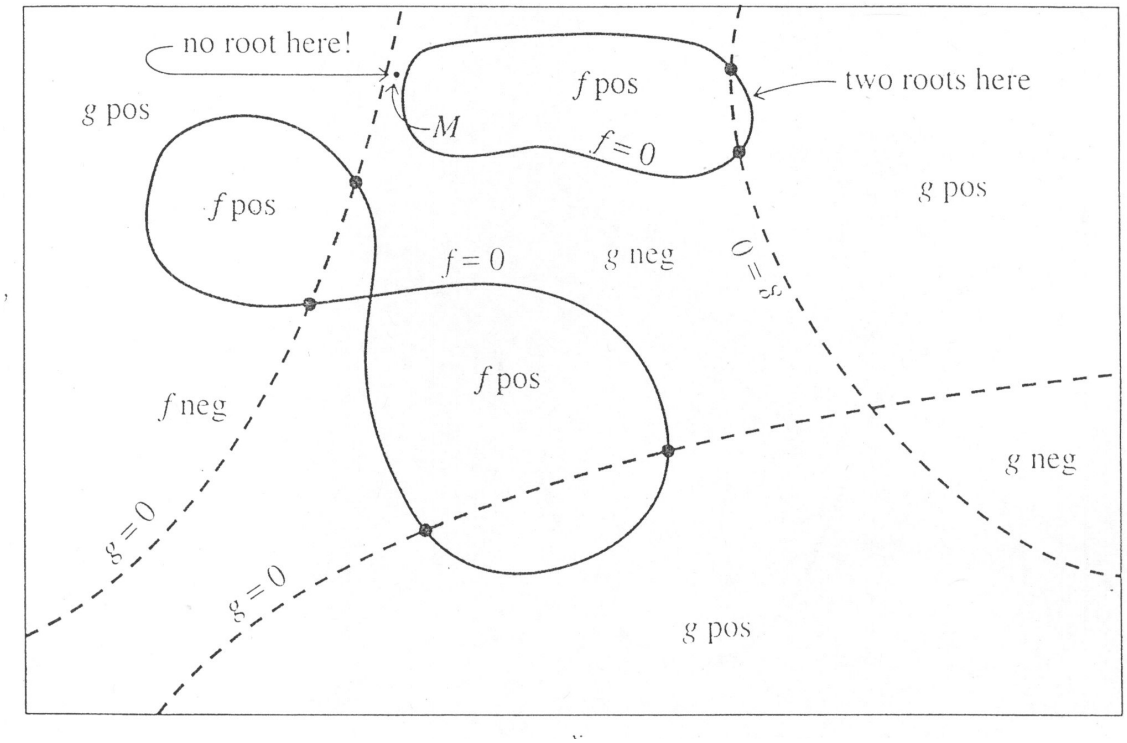

Figure 8.8: Solution for two nonlinear equations with two unknowns

Exercise:
Determination of the quantiles of the gaussian distribution

Lessons learned:

- In iterative algorithms: Order of convergence γ

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

- Bisection, secant method, regula falsi, Newton-Raphson.
- Trade-off: Order of convergence vs. convergence safety.

7.18

9 Optimization

- Literature: Recipes Chap. 10
- Task: Determine x, so :

$$f'(x) = 0, \quad f''(x) >< 0, \text{ as the case may be}$$

- Optimization encompasses minimization and maximization "one's f is the other's -f"
- Iterative algorithms

Differences of methods:

- 1 D vs. N D
- Derivative information available or not.
- Deterministic methods: Convergence against local optimum
- Stochastic methods: In principle global convergence

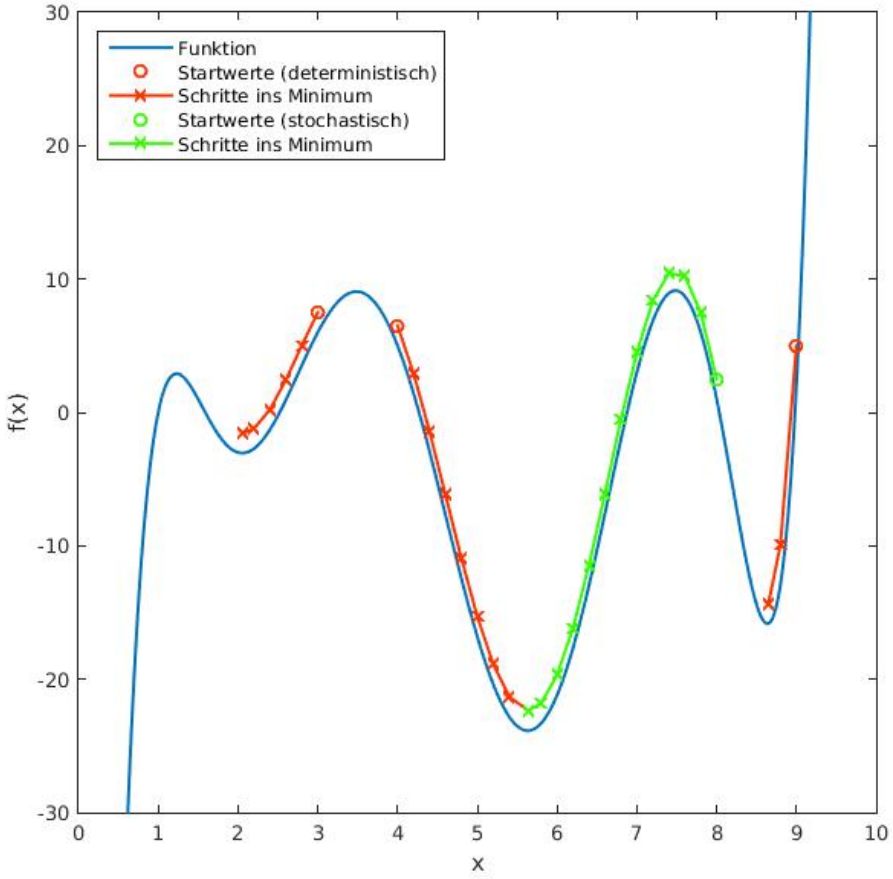

Figure 9.1: Difference between deterministic and stochastic methods

9.1 One dimensional case

Consider minimization

9.1.1 Bracketing, golden ratio search

Analog to bisection in Chap. 8 Zero point search
Consider:

- Zero point braketing needs 2 points

- Minima-bracketing needs 3 points (a, b, c).

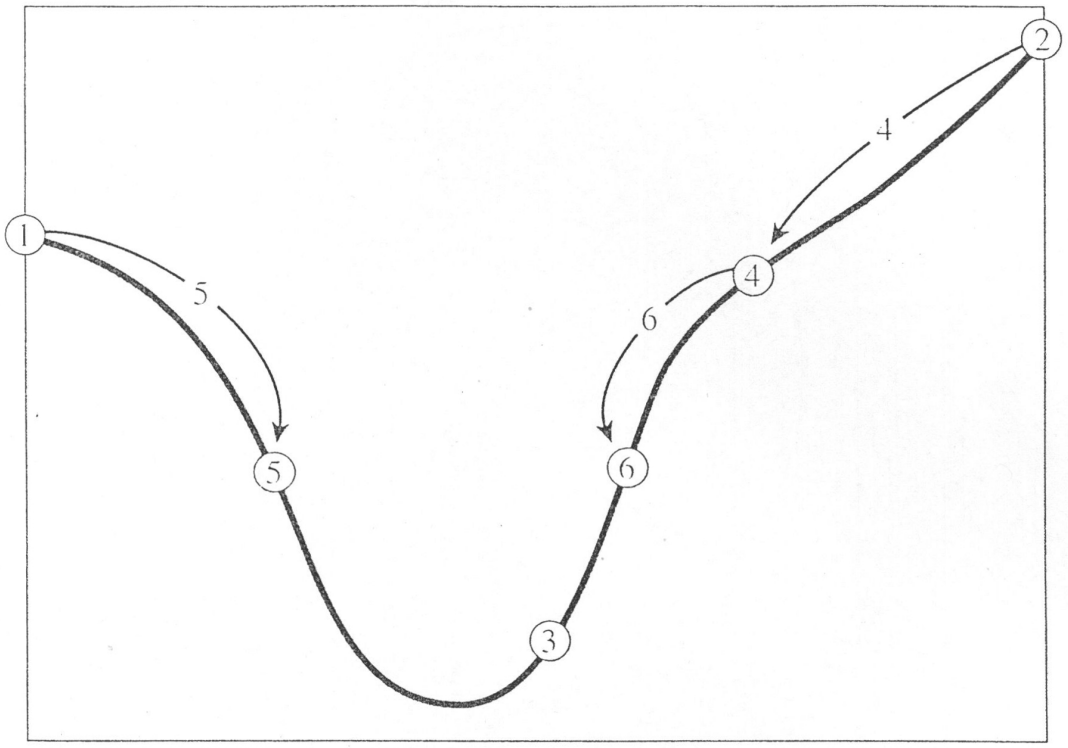

Figure 9.2: Minima-Bracketing

How does one search for a new in between point given (a, b, c) ?

- Let b be a fraction w on the way from a to c

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

- New point x be behind b by an additional fraction

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

Then the next braketing segment is:

– either $w + z$
– or $1 - w$

relative to the existing one.

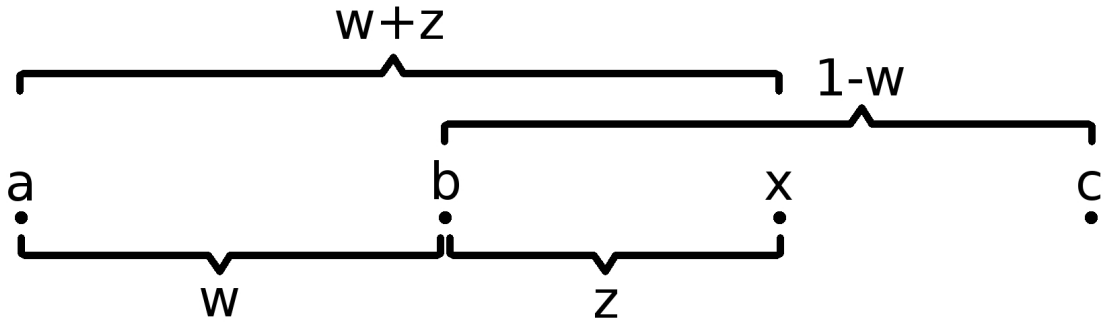

Figure 9.3: Scale invariance of the golden ratio

To minimize the worst case: Choose z in a way that potential next segments are equally large:

$$z = 1 - 2w \tag{5}$$

Per construction: x is symmetric to b in starting interval $|b - a| = |x - c| \Longrightarrow$ x lies in the longer segment

- Where lies the longer segment?Where does w come from?

  Assume w is as optimal as z should be

  Similarity of scale: x same portion of (b, c), if this was the longer segment, as b was in (a, c)

$$\begin{aligned} \frac{x-b}{c-b} &= \frac{b-a}{c-a} \\ \frac{x-b}{c-b}\frac{c-a}{c-a} &= \frac{b-a}{c-a} \\ \frac{z}{1-w} &= w \end{aligned} \tag{6}$$

  Eq. (5,6) together:

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

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

$$\frac{1 - w}{w} = \text{golden ratio}$$

- Starting with arbitrary points (a, b, c), the procedure convergences to the golden ratio
- Linear order of convergence

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

9.1.2 Parabolic interpolation, Brent's method

Analogously to Regula falsi.

- Regula falsi: Close to zero point, linear approximation is good
- Parabolic interpolation: Close to the optimum, quadratic approximation is good.

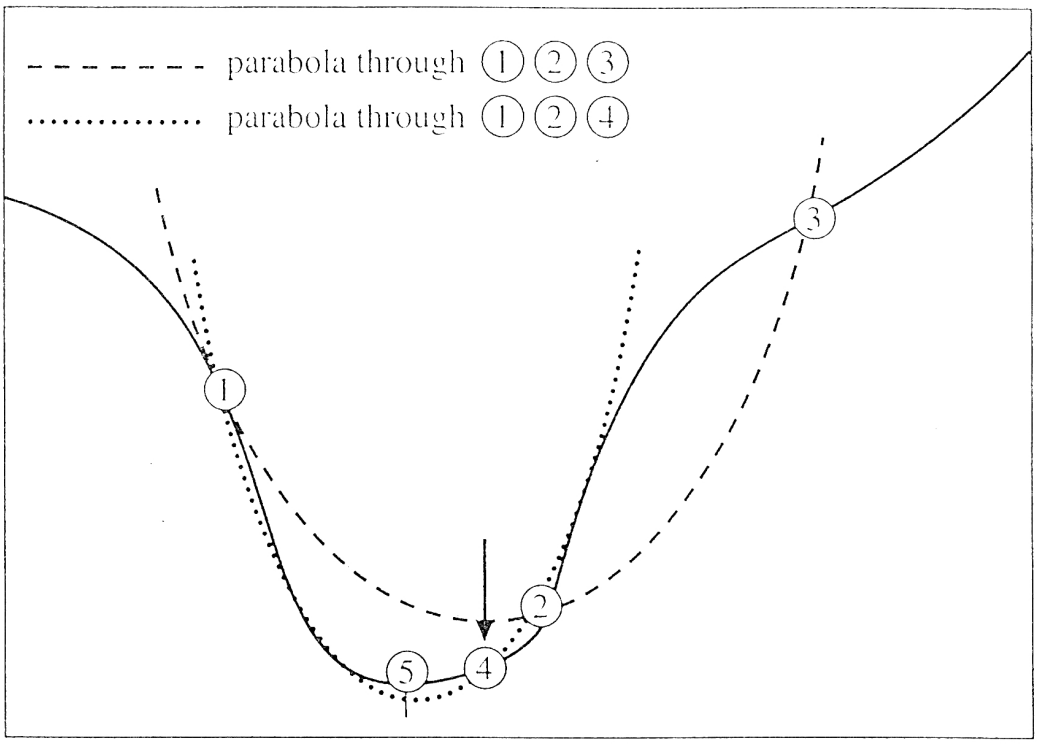

Figure 9.4: Convergence to a minimum through parabolic interpolation

Given (a, b, c) and $f(a), f(b), f(c)$, new point x through:

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

In 1D information by derivation usually unnecessary.

9.2 N-dimensional case

9.2.1 Only function evaluations

Naivest Ansatz

1. Choose starting point

2. Progress along one coordinate axis until minimum is reached
3. Repeat for all other coordinates
4. Go to 2.

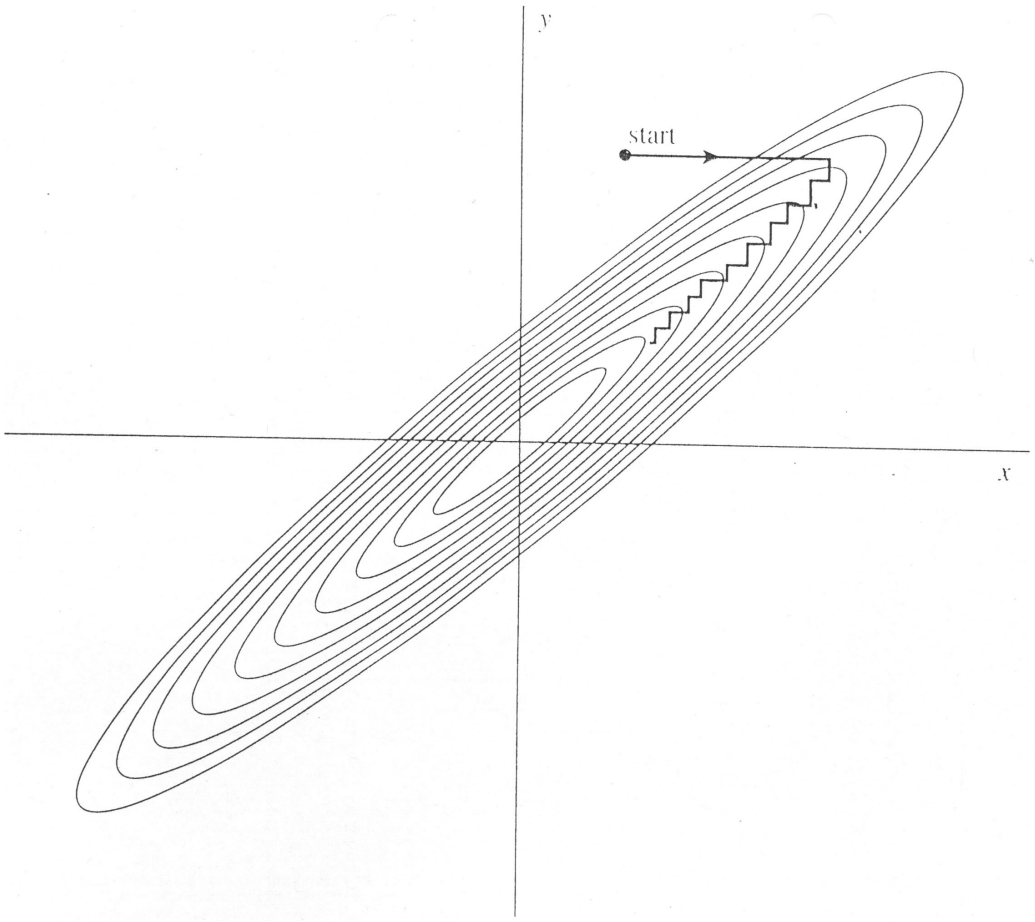

Figure 9.5: Successive minimization along the coordinate axis

This is very inefficient!

Powell's method
Based on `linmin()` :

- Given
  - Function $f(.)$ to be minimized
  - $\vec{P}$: Current point
  - $\vec{u}$: Direction of search
- Bracketiering minimum in direction $\vec{P} + \mu\vec{u}$.
- Find scalar λ, so $f(\vec{P} + \lambda\vec{u})$ minimal.
  1D - Problem, see above.
- Replace $\vec{P}$ by $\vec{P} + \lambda\vec{u}$.

Idea:
Try to find successive „good "directions of descend $\vec{u}_i$, $i = 1, \ldots, N$:

- Initialize: $\vec{u}_i = \vec{e}_i$, $i = 1, \ldots, N$
- Start position: $\vec{P}_0$
- For $i = 1, \ldots, N$: $\vec{P}_i =$`linmin`$(\vec{P}_{i-1}, \vec{u}_i)$
- For $i = 1, \ldots, N-1$: Replace $\vec{u}_i$ by $\vec{u}_{i+1}$
- Set $\vec{u}_N = \vec{P}_N - \vec{P}_0$, $\vec{P}_N - \vec{P}_0$: Average direction of success
- $\vec{P}_0 =$`linmin`$(\vec{P}_N, \vec{u}_N)$
- Iterate this.

Behavior of convergence:

- Quadratic approximation exact: Procedure after N iterations in optimum.
- Quadratic approximation good: Order of convergence quadratic.

7. week

9.2.2 Use of derivative information

Derivation must/should be known analytically. Approximation through e.g.

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

are difficult because

- Elimination in $f(x + \Delta x_i) - f(x)$
- complicated to control the "$\approx$"

Naivest idea: Steepest Descent

- Start position: $\vec{P}_0$
- Go from $\vec{P}_i$ to $\vec{P}_{i+1}$ by minimizing along the direction of $-\nabla f(\vec{P}_i)$
- Iterate until reaching goal

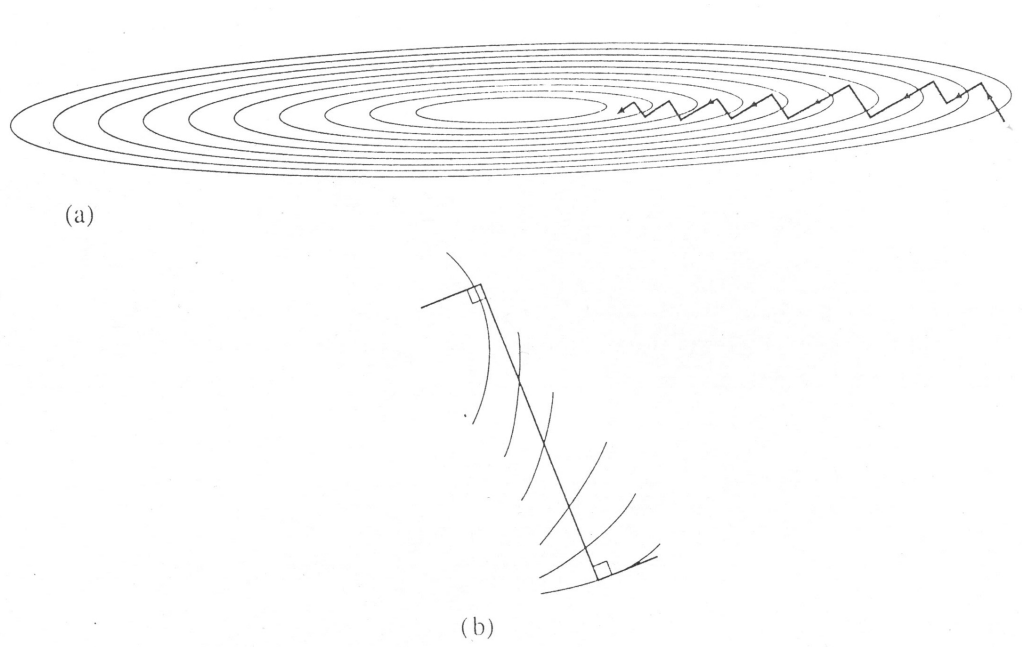

Figure 9.6: a) Steepest Descent method in long, narrow valley; b) Magnification of one step

DO NOT DO Steepest Descent

Reason:

- No consideration of curvature information
- Or: Wrong metric, comment ART.
  Steepest decent: Successive directions of search $\vec{u}_i$, $\vec{u}_{i+1}$ fulfill:

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

  Better :

$$0 = \langle \vec{u}_{i+1}\, A \vec{u}_i \rangle = \vec{u}_{i+1}^T\, A\, \vec{u}_i, \tag{7}$$

  with

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

  The direction in Eq. (7) is then called conjugated.

- Proof:
  - Let P be the origin of the coordinate system
    Taylor evolution:

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

    and with this

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

  - Change of $\nabla f(x)$ by movement of δx, shortly before convergence :

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

– If one has moved along the direction u_i to the minimum, new direction u_{i+1} should be maximally informative:

$$0 = u_{i+1}\delta(\nabla f(x)) = u_{i+1}Au_i$$

- Powell's method constructed conjugated directions
- Comment:

  When instead of a unique minimum, there is a long troth, then A is ill conditioned , remember chap. 7.2.3 SVD and Chap. 4.4 Non identifiability.

Variable metric or Quasi-Newton – procedure

- If close to minimum x_m, $\nabla f(x_m) = 0$, Taylor evolution around current point x_i:

  $$f(x_m) = f(x_i) + (x_0 - x_i)\nabla f(x_i) + \frac{1}{2}(x_0 - x_i)A(x_0 - x_i) + \ldots$$

  Derive:

  $$\nabla f(x_m) = \nabla f(x_i) + A(x_0 - x_i) \stackrel{!}{=} 0$$

  Straight to the goal with

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

  This dates back to Newton.

- But: Calculation of $A^{-1}(x)$ has effort $\mathcal{O}(N^3)$, remember Chap. 7
- Idea:

  During iterations collect information about the (local) Hesse matrix, preferably immediately A^{-1}.

The procedure:

1. Choose starting value x_0

2. Choose I_0, positive definite, symmetric
3. Go $x_{i+1} = x_i - I_i \nabla f(x_i)$
4. Use the DFP or BFGS updating formula ...

   RECIPES Eq. 10.7.8 and 10.7.9

- Go to 3.

Properties:

- Uses only gradient information
- It holds:
$$\lim_{i \to \infty} I_i = A^{-1}$$
- Complexity $\mathcal{O}(N^2)$
- Belongs to the 5 most important routines there are.

Conjugated gradient – procedure

- Generates iterative conjugated directions defined in Eq. (7).
- Does not construct the (inverse of the) Hesse matrix which is $\mathcal{O}(N^2)$ expensive.
- Method of choice for higher dimensions ($N > 100$)

Quasi-Newton and conjugated gradients – procedures converge quadratic,when close to the minimum.

In general:

- When to terminate the iteration?

  Termination criteria:

  i. Relative change of function value : $(f(x_i) - f(x_{i+1}))/f(x_i) < \epsilon_1$
  ii. Relative change of x : $|x_{i+1} - x_i|/|x_i| < \epsilon_2$

Recommendation: ii., because of "long troth".

- In all previous methods only convergence toward local optimum was guaranteed.

  Only cure: Try multiple starting values

9.2.3 Simulated annealing

Further literature:

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

All procedures up until now:

- Deterministic
- Target location given trough starting point
- Only convergence to local optimum

Probabilistic/statistic optimizer, here minimizer
Name giving:

- By annealing a liquid quickly, the formed crystal does not reach the global energy minimum but only a local one.
- There are many local minima, conflict: Near and far order.
- By annealing slowly, the global minimum is reached with high probability or at least approximately
- Reason: By slow annealing energy barriers can be surpassed with thermal energy (Boltzmann distribution).

Idea for numeric minimizer: May also run uphill sometimes

Procedure:

- Choose starting value x_0

- Produce random changes ϵ_i: $x_{i+1} = x_i + \epsilon_i$
- If $f(x_{i+1}) < f(x_i)$, accept x_{i+1}.
- If $f(x_{i+1}) > f(x_i)$, accept x_{i+1} with probability

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

  Remember: Boltzmann distribution
- Choose $T(i)$ large at beginning, let it go to 0 with increasing iterations

Problems:

- Choice of the annealing scheme $T = T(i)$, e.g. $T(i) \propto 1/i$
- Choice of the magnitude of the change ϵ_i, e.g. $< \epsilon_i^2 > \propto 1/i$
- Both need prior knowledge of the problem: No free lunch - theorem
- The prior knowledge corresponds to gradient and curvature information
- Does therefor not play a considerable role in „serious "applications

But: Can solve non polynomial (NP) hard problems in very good approximation.

Example: Traveling salesman problem $\mathcal{O}(N!)$

- N cities with coordinates (x_j, y_j)
- Look for tour through all cities which has the smallest length
- Configuration $conf$ is permutation with the numbers $j = 1, \ldots, N$
- Functional to be minimized: way length

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

- „Change "ϵ: Local changes of permutations.

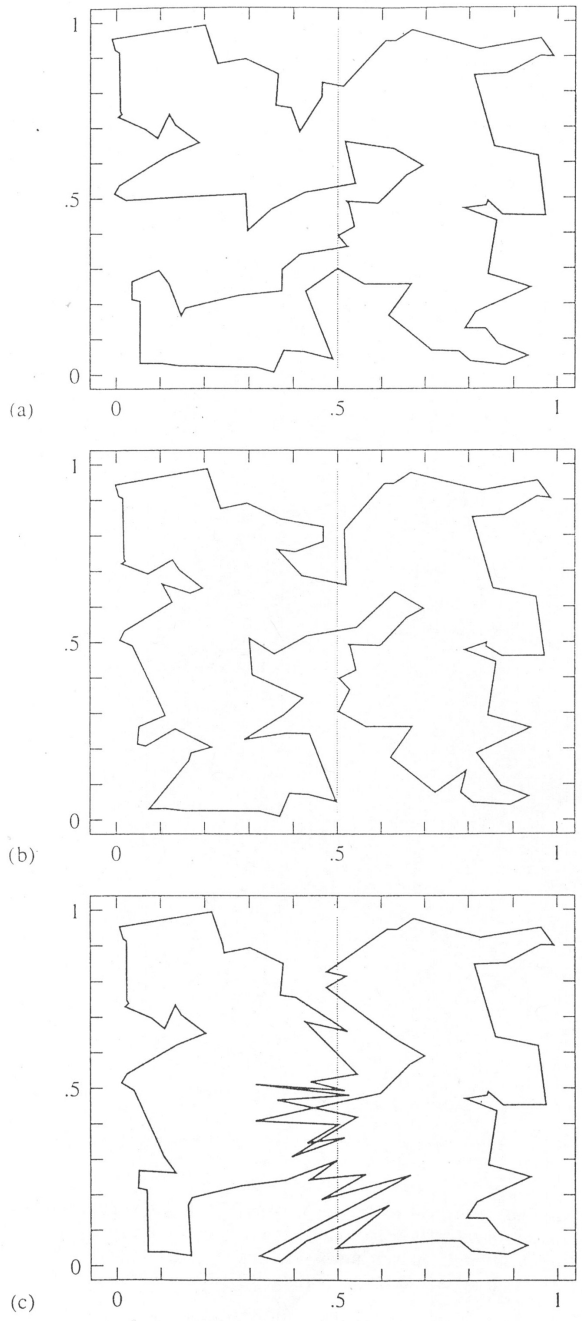

Figure 9.7: Traveling salesman a) no side conditions, b) fewest possible crossings of the river, c) most possible crossings of the river

- 1. Traveling salesman of history: Odysseus, 13 stations: 6.2×10^9 possibilities.

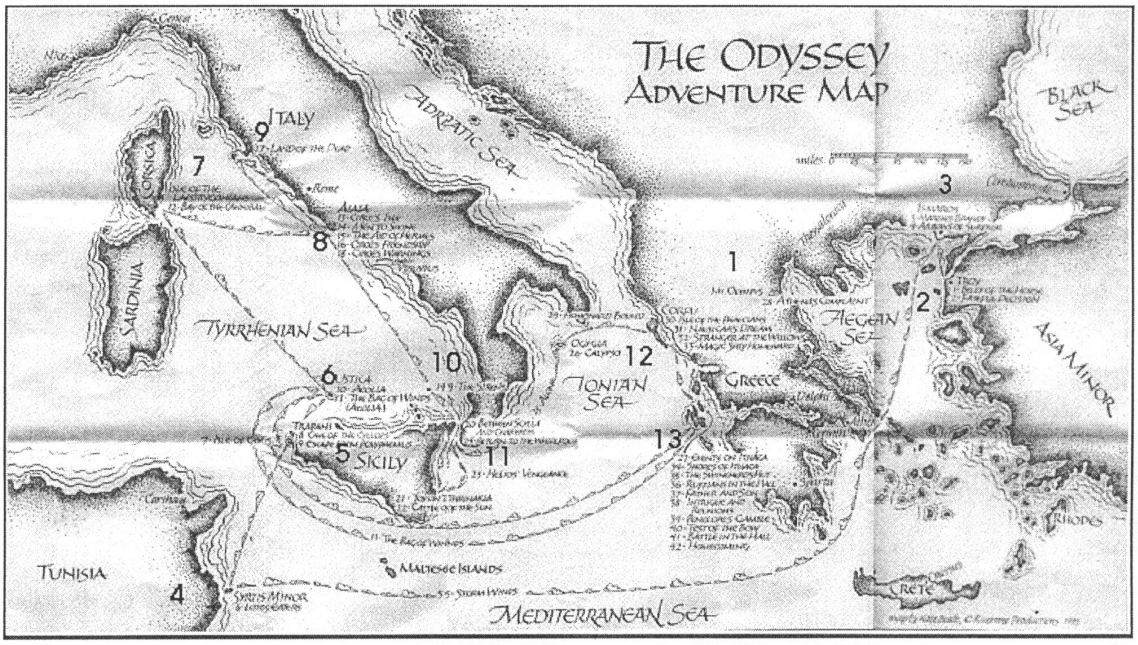

Figure 9.8: Odysseus' voyage route: His way home: 9000 km, shortest 6000 km

- Flexibility of the method:
  Expansion of the functional by a penalty term, remember, Chap. 4.3
  Example:
  - Assume: River divides the area.

  (i) Salesman is scared of crossing the river

  (ii) Salesman smuggler and wants to cross the river as often as possible

$$\mu_j = -1 \text{ for left of the river}, \quad \mu_j = +1 \text{ for right of the river}$$

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

  For

(i) $\lambda > 0$

(ii) $\lambda < 0$

see figure 9.7

Other stochastic optimizer:

- Evolutionary algorithms
- Genetic algorithm
- Particle swarm algorithm

Exercise:
Maximum entropy distribution for discrete distributions

Lessons learned:

- One dimensional: Golden ratio
- Higher dimensional: Steepest descend obvious, but not good
- Better: Incorporate curvature information: Quasi Newton
- Deterministic procedures: Locally convergent
- Stochastic procedures: In principle global convergence

8.18

10 Non linear modeling

Literature:

- *Numerical Recipes*, Chap. 15
- G.A.F. Seber and C.J. Wild. *Nonlinear Regression* [62] The classic
- G.J.S. Ross. *Nonlinear Estimation* [57] great book

Motivation:

- Chap. 9 optimization: General search of optima
- Here: Minimization of special functionals

Is

- $y(x) = y(x, a)$ a function parameterized with a, e.g. first principle equation with free parameters
- y_i, $i = 1, \ldots, N$: N measurements of the function $y(x, a)$ at points x_i
- Measurements in general with errors ϵ_i: $y_i = y(x_i, a) + \epsilon_i$, e.g. $\epsilon_i \sim N(0, \sigma_i^2)$
- Goal: estimating a based of N measurements (y_i, x_i)
- Putting it differently: modeling the connection (y_i, x_i) by $y(x, a)$
- Parameter estimation by minimization of:

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

  Remember: Weighted least square estimator is MLE for normal distributed errors

- If the model is correct, number of parameters k, it holds:

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

This allows goodness-of-fit test:

- H_0: The model is correct.
- H_1: The model is not correct.

Remember :

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

Under H_0: $\chi^2(\hat{a})$ for 99% confidence interval in the area

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

- Is $\chi^2(\hat{a})$ larger, the natural case:
  - Model wrong ?

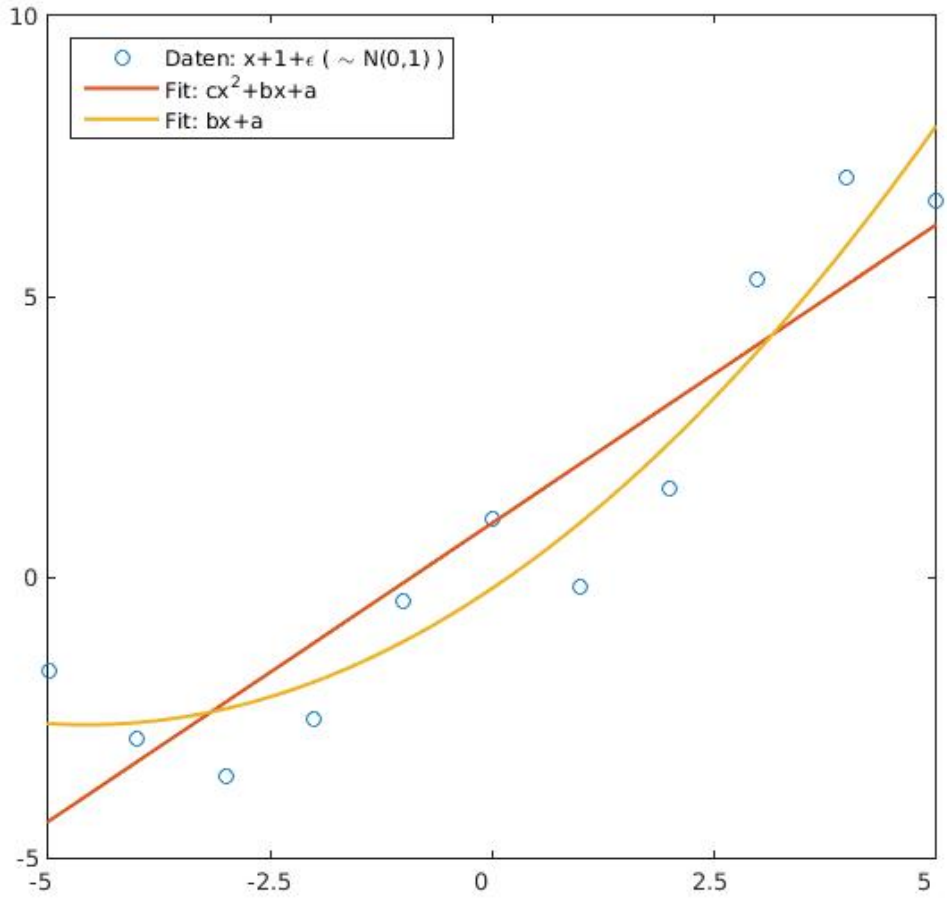

Figure 10.1: arg2

  - σ_i falsely too small?
  - Error not normal distributed ?

- Is $\chi^2(\hat{a})$ smaller, usually shouldn't happen:

  - σ_i is falsely too large?
  - Error not normal distributed?

10.1 Linear regression

Assumption: Gaussian errors:

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

Everything works analytically:

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

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} \stackrel{!}{=} 0 \tag{8}$$

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

With

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

it follows from Eq. (8, 9)

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

With determinant Δ:

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

follows:

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

Gaussian error propagation: Cramér-Rao barrier

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

By plugging in:

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

But: This is a 2D estimation problem:

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

with

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

Covariance σ_{ab}^2

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

σ_{ab}^2, resp. condition number of Σ says if estimator is dependent.

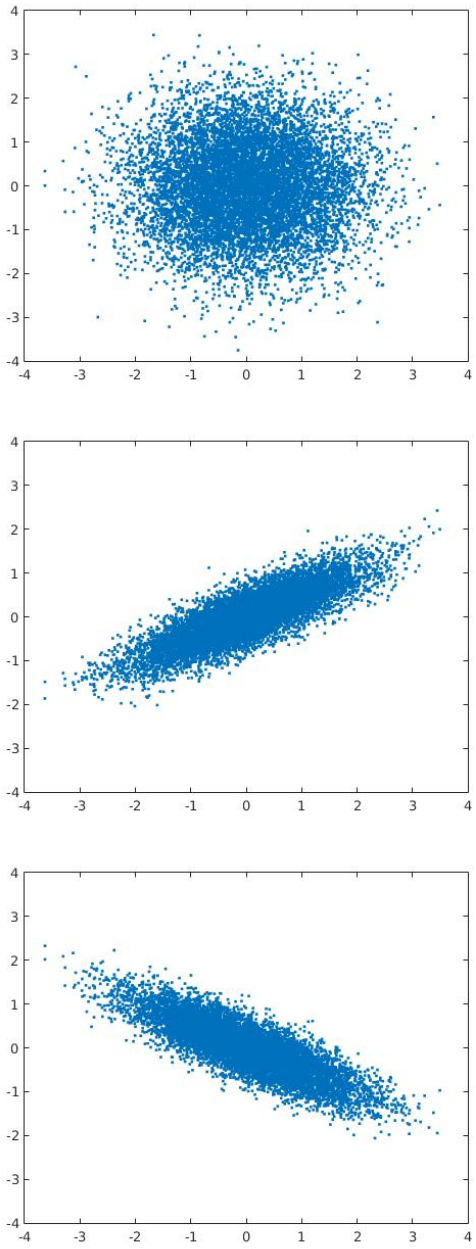

Figure 10.2: *2D*-normal-distributed-random-numbers with $C_1 = (0.71\ 0; 0\ 0.70)$, $C_2 = (0.78\ 0.39; 0.39\ 0.28)$, $C_3 = (0.79, -0.39;\ -0.39, 0.28)$

Condensed in: Correlations $r_{ab} \in [-1, 1]$ between estimation errors

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

Comment:
Often: Sums over many summants, can lead to rounding errors
Solution: Kahan-Summation [33]

Robust linear regression

Additional literature:

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

If the error distribution is:

- Non gaussian, χ^2 fitting is no longer MLE
- Symmetric, is χ^2 fitting bias-free, but has a larger variance

  Remember efficiency of an estimator:

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

  see exercise.
- Asymmetric, a bias can be produced.
- Slower decreasing than gaussian, fat-tailed, e.g. Cauchy, χ^2 is caught on these outliers.
- Solution: Robust procedures. Do not get caught on the outliers.

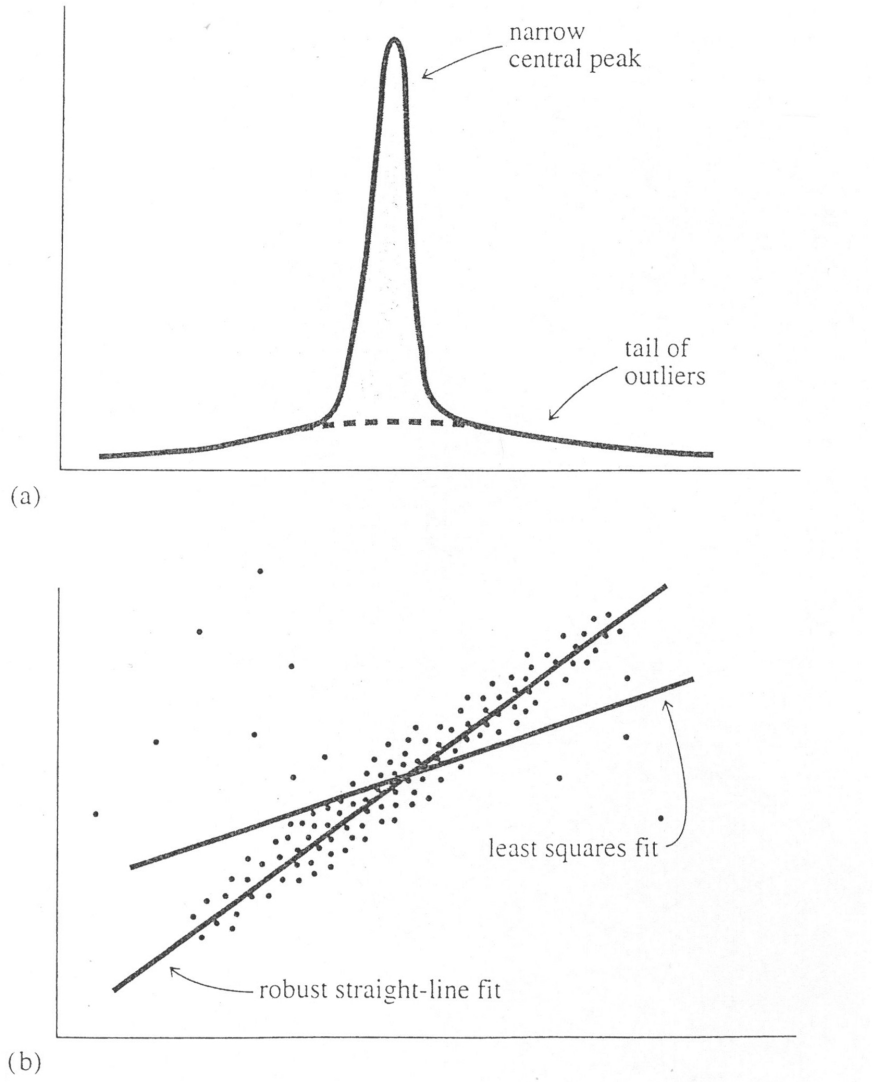

Figure 10.3: Examples for robust statistical methods: (a) One dimensional distribution with outliers. (b) Two dimensional distribution fitted to a line.

8. week

Remember:

- In the Gaussian case the Likelihood was:

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

and log Likelihood

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

Parameter estimation by setting the derivation to zero:

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

Discussion of factors:

- 1. Factor: Influence of data
- 2. Factor: Model specificity

- In general:

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

As a rule

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

- Define:

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

$\psi(.)$ is called Influence Function

- Yields, via generalization of Eq. (10), MLE condition

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

- Special case Gaussian for:

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

- Ergo: Influence of data increases with linear deviation.
- Therefor not robust.

Other distribution:

- Double exponential distribution

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

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

Ergo: Influence of data on the MLE only dependent on the sign.
Therefor significantly more robust!

- Example: Cauchy distribution

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

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

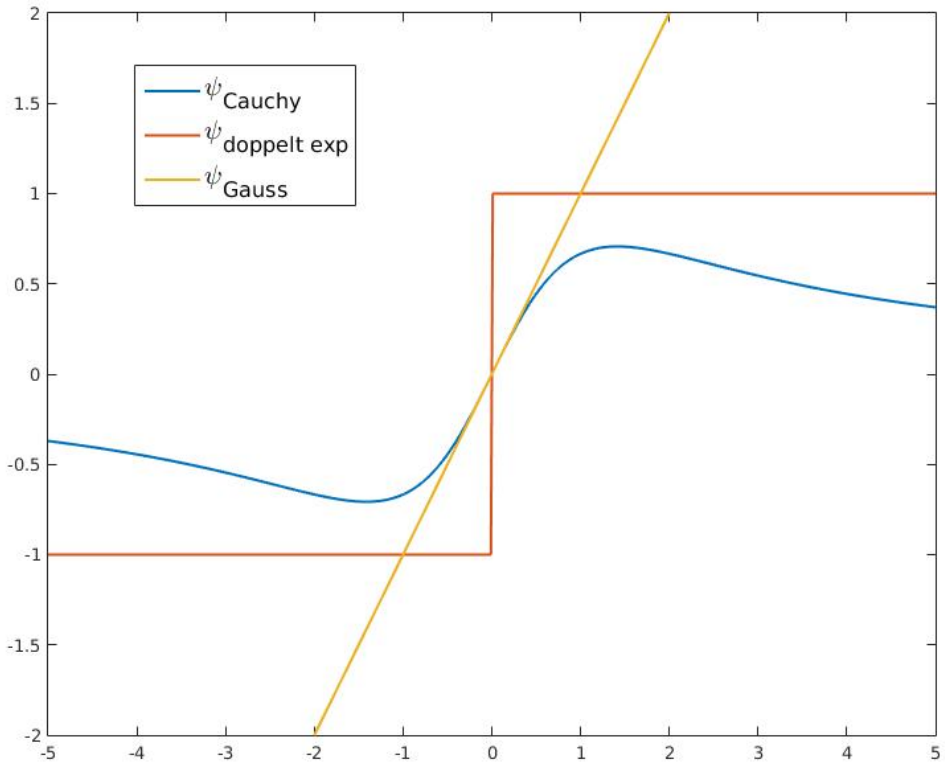

Figure 10.4: Influence Functions resulting from different distributions

Ergo: Influence of data on MLE decreases with higher deviation.

Therefor very robust!

- Turning the tables: Decreasing of influence by deviation can be used for construction of Influence functions for error models = "well-behaved" "+" "outliers"

  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

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

$c = 6.0$

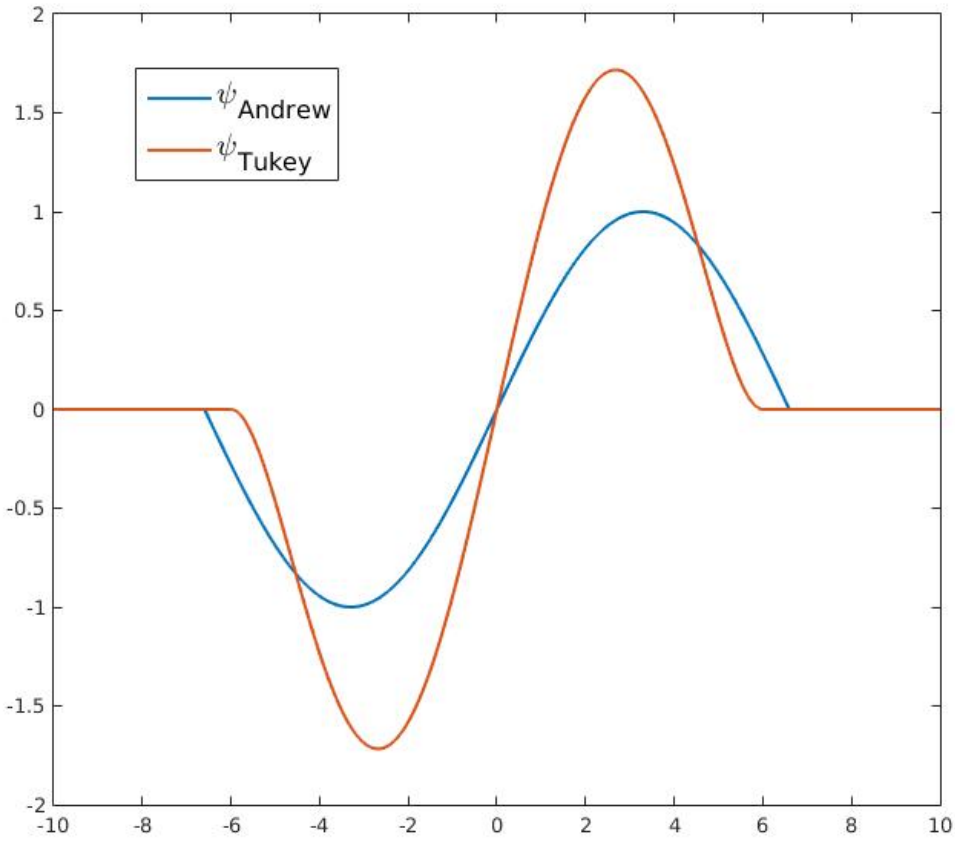

Figure 10.5: Constructed Influence Functions

Example for concrete calculation:

- Linear regression with double exponential errors

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

- Instead of χ^2, the log-likelihood is:

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

- Mental side calculation:

  Definition median:

  - Given N numbers $\{z_i\}$.
  - Sort.
  - If N uneven: $\text{med}\{z_i\} = z_M = z_{(N+1)/2}$
  - If N even: $\text{med}\{z_i\} = z_M = 0.5(z_{N/2+1} + z_{N/2})$

  Median z_M minimized :

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

  Proof:

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

With this:

- Iterative procedure
- Choose initial estimations $(\hat{a}_0, \hat{b}_0)$, e.g. from least squares estimator.
- For given $\hat{b}_j$

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

then analogously to Eq. (11), for given $\hat{a}_{j+1}$, follows $\hat{b}_{j+1}$ from:

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

Zero point search.

Becomes iterative with Bisection, see chap. 8 zero point search, solved.

- Iterate until desired precision

The saying "Since robust statistics are being used at CERN, no new particle was found", has been recently disproven

Exercise :
Sub-optimal behavior of the LS estimator in the case of non-gaussian distributed data

10.2 Non-linear regression

Simplest continuation from above:

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

or more general:

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

$X_k(x)$ Basis function, e.g. $\sin(\omega_k x)$

Model is

- linear in parameters,
- but has nonlinear basis functions.

Now:

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

Define:

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

- A is called Design matrix, is $(N \times M)$,
- It fixes: Which basis function is measured where.
- Mentioning optimal design, experimental design [52].

Maximum Likelihood estimator:

- Minimal condition for χ^2 :

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(x_i) \stackrel{!}{=} 0 \tag{12}$$

- With

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

  α is $(M \times M)$ matrix
  and

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

  and switching the sums in Eq. (12) follows:

$$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k \tag{13}$$

- The equations (12), resp. (13) are called Normal-equations
  and remind the form:

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

  in Chap. 7 Pseudo- or Moore-Penrose - inverses for this over determined equation system.

- This yields the point estimator.

Confidence intervals for parameters:

- Define:

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

- Consider:

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

- Remember error propagation:

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

  with

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

  follows:

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

- The term (.) is the even $\alpha = C^{-1}$, therefor:

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

  Thus C_{jk} yields the covariance between the estimation errors of a_j and a_k.

- Watch out:
  α, and therefor C, is independent of y_i.

With this: Optimal design

- Since $\alpha = A^T A$, the design is defining the errors.
- Optimal design: See linear regression on interval [-1,1], one can measure 4 times.

  Where should one measure, to get smallest possible errors ?
- There are different optimal criteria: A through D-optimal, ..., depending if trace, determinant or similar properties of the covariance matrix should become small .

Non-linear regression and SVD
Normally:

- (A lot) more data then parameters.
- The system Eq. (13) should be well solvable.

But:
If basis functions are not sufficiently independent $\Longrightarrow$
Problem badly conditioned no matter how much data is available.

- Consider monomes 1, x, x^2, x^3, ... with x equally distributed on interval [0,1] as basis functions.
- Then hold for A:
$$A_{lm} = \sum_{i=1}^{N} x_i^l x_i^m \propto \frac{1}{l+m+1}$$
- Remember Hilbert matrix, exercise Chap. 7 solving of linear equation systems
- At known density $p(x)$ polynomials orthogonal to that density can be used, rendering the procedure stable because A becomes diagonal
- Example $p(x) \sim$ equal distribution [-1,1]: Legendre-Polynomial
- There are recursive construction rules for polynomials orthogonal to empirical data [16].
- Recommendation: Use SVD, To check ill-conditioning and to treat it if necessary. The SVD generates these orthogonal polynomials.

10.3 Non-linear modeling

Reminder

- Linear regression: Linear in parameters and independent variable x
- Non-linear regression: Linear in parameters, non-linear in x

Now: Also non-linear in parameters, e.g.:

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

- Iterative procedure, similar to Chap. 9 optimization.
- Remember:
  Close to the optimum, the quadratic approximation is good, and Newton-step

$$a_{i+1} = a_i - A^{-1} \nabla f(a_i) \tag{14}$$

  leads to goal.

- In Chap. 9 optimization: A^{-1} unknown/expensive to determine
  - Quasi-Newton - procedure collects information about A^{-1} during iteration
  - Conjugated gradient approaches $< \delta a_{i+1} \, A \, \delta a_i >$.

10.3.1 Levenberg-Marquardt algorithm

Here one knows more:

- Functional:

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

- 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:

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

- Convention:

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

- If the fit is good, it holds for the second term of the Hesse-Matrix

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

  since the errors $\epsilon_i = (y_i - y(x_i, a))$ are uncorrelated.
  Therefor, define:

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

- With the above and with $\delta a_l = (a_{i+1} - a_i)_l$ Eq. (14) becomes

$$\sum_{l=1}^{M} \alpha_{kl}\delta a_l = \beta_k \tag{15}$$

- Notice : Steepest Descent reads:

$$\delta a_l = \text{const } \beta_l \tag{16}$$

- Idea Levenberg-Marquardt algorithm:
  - Far from the minimum, Newton step might be bad Eq. (15).
  - Perform gradient step Eq. (16). how to choose "const" ?
  - $\chi^2(a)$ dimensionless, dimension $[\beta_l]$ = dimension $[1/\delta a_l]$, consider Eq. (15)
    $\Longrightarrow$
    $1/\alpha_{ll}$ is scale candidate.
  - To be sure that the step is not too large, choose $\lambda \gg 1$ and set:

$$\delta a_l = \frac{1}{\lambda \alpha_{ll}}\beta_l \quad \text{or} \quad \lambda \alpha_{ll}\delta a_l = \beta_l \tag{17}$$

- Combine gradient step Eq. (17) and Newton step Eq. (15) by

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

  yields:

$$\sum_{l=1}^{M} \alpha'_{kl}\delta a_l = \beta_k \tag{18}$$

  Meaning:

  - If λ is large $\Longrightarrow$ α'_{kl} diagonal dominant $\Longrightarrow$ small gradient step
  - If $\lambda \to 0$, Hesse step

Procedure:

1. Choose starting estimation for a, calculate $\chi^2(a)$
2. choose small λ: $\lambda = 0.001$. Expresses hope
3. Solve Eq. (18) and calculate $\chi^2(a + \delta a)$
4. If $\chi^2(a + \delta a) \geq \chi^2(a)$, discard δa, choose $\lambda = 10\lambda$, go to 3
5. If $\chi^2(a + \delta a) < \chi^2(a)$, accept δa, choose $\lambda = 0.1\lambda$, go to 3.

Interpretation:
If Newton step

- good, more of them,
- bad, proceed with care with a gradient step.

Comments:

- Belongs to the 5 most important routines there are.
- Consider:
  - Equation (18) can be ill conditioned
  - The tub again.

- Solve with SVD.

- Termination criteria :
  - If only small changes in χ^2, problem "tub"
  - Better, if $\lambda > 10^5$, corresponds to no change in a anymore.

After convergence:

- Asymptotic covariance matrix of the errors in the estimated parameters

$$C = \alpha^{-1} = \left\{ \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i, a)}{\partial a_k} \frac{\partial y(x_i, a)}{\partial a_l} \right] \right\}^{-1} \tag{19}$$

- Alternative to Levenberg-Marquardt: Trust-Region Approach

9. week

Exercise:
Non linear modeling and model tests

10.3.2 Monte Carlo confidence intervals

- Remember Chap. 2.4. The standard deviation of the parameter estimator yields -generally only asymptotically- confidence intervals for the true parameters, i.e. the true value lies with 95% confidence in

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

- The covariance matrix in Eq. (19) for the errors of estimated parameters only holds asymptotically

- An alternative: Profile likelihood, see Chap. 4.4

- What everyone would prefer:

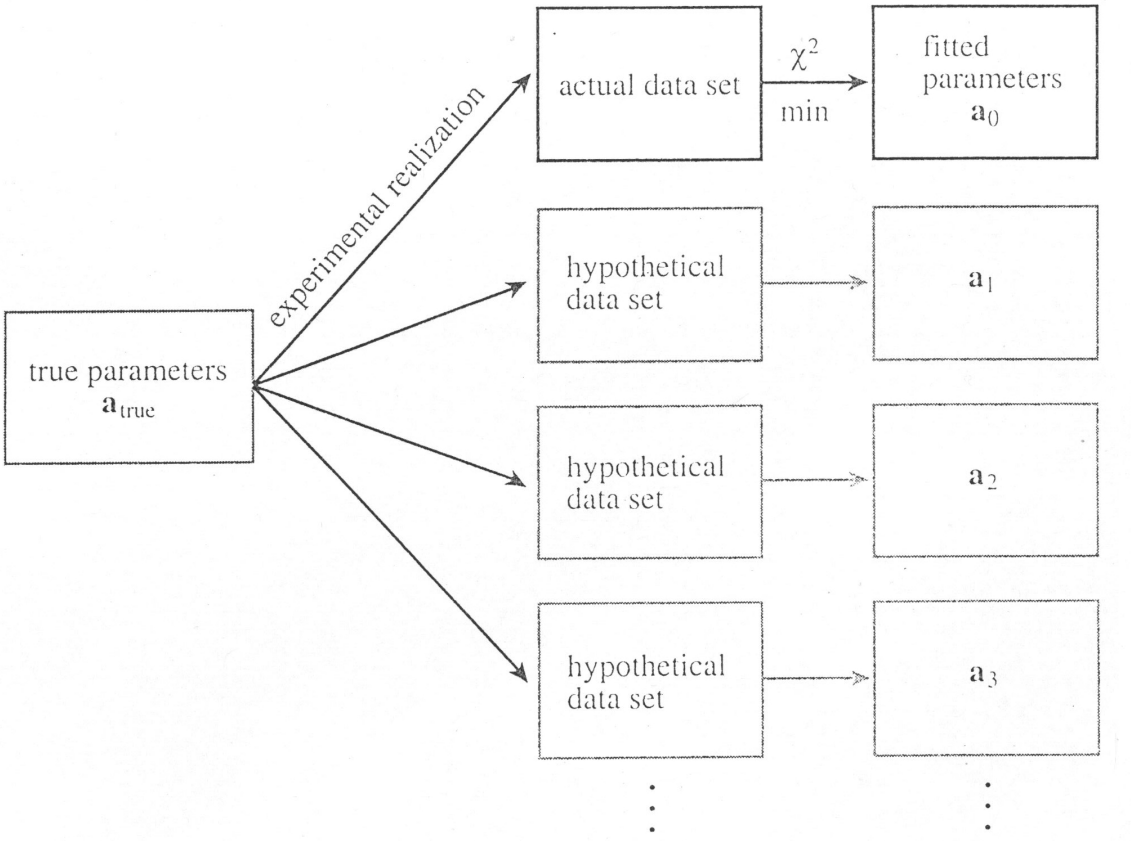

Figure 10.6: A statistic universe of data sets for an underlying model

- Would yield the complete distribution of the estimated parameter.
- Is however not available (and data splitting does not help.).
- A statement about the true value is needed based on a (finite) data set.

Asymptotic confidence intervals, Chap. 4.1

- Asymptotically it holds under mild conditions

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

with

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

Yields confidence intervals for the parameter

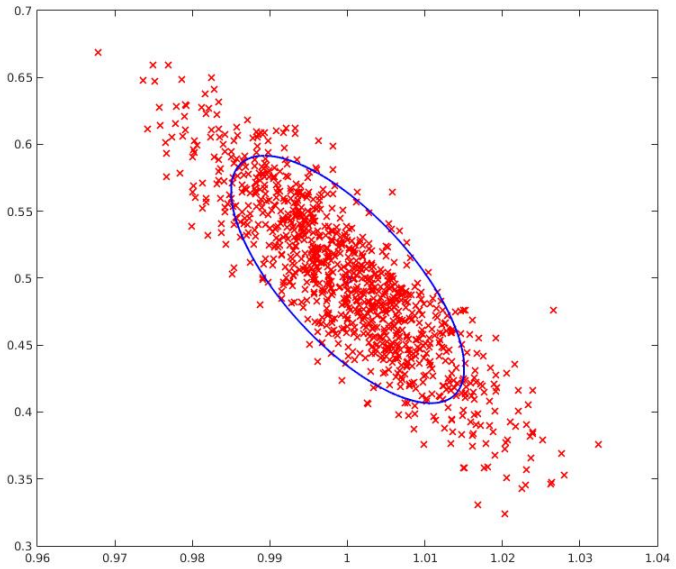

Figure 10.7

- For regression this is analogous to error propagation:

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

- In non-linear modeling it only holds asymptotically .

An ansatz in the finite region: χ^2-Contour confidence interval

- Confidence region by Iso-log-likelihood contours
- Determination by variation of the parameters around the estimated ones.

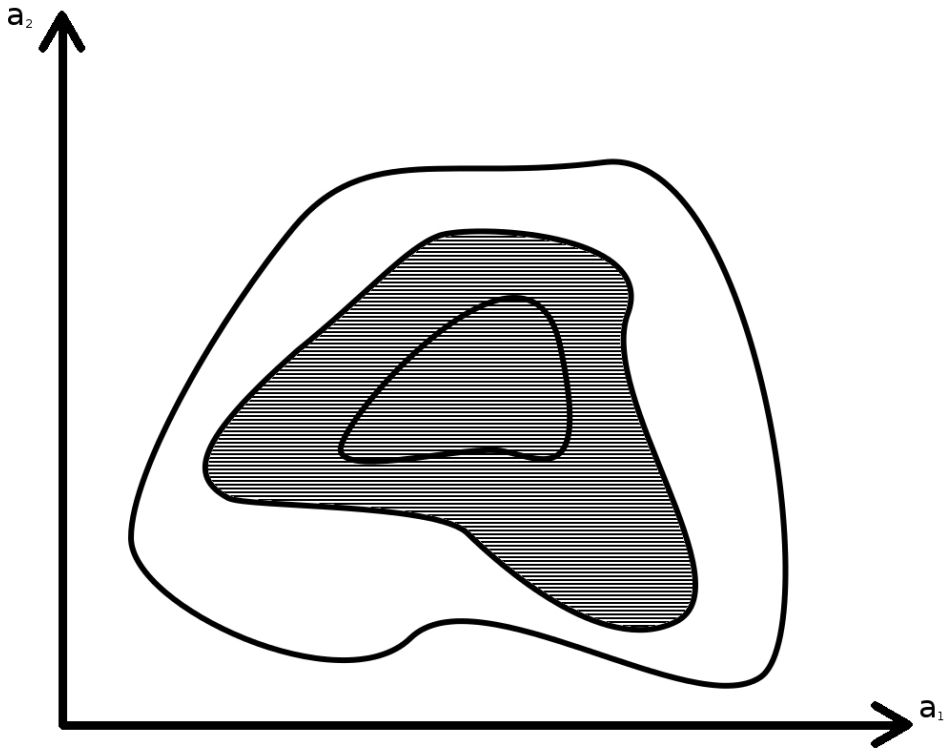

Figure 10.8: Confidence regions by Iso-log-likelihood contours

Fundamental Alternative in the Finite: Monte Carlo confidence interval

(i) Parametric bootstrap

- Bootstrap: To pull oneself out of the swamp by the own boots © Münchhausen
- Estimate parameter $\hat{a}$
- Produce new data sets with
  - Parameters $\hat{a}$
  - new errors under parametric assumptions to their distributions
- determine confidence region from distribution of the estimated $\hat{a}_i$.

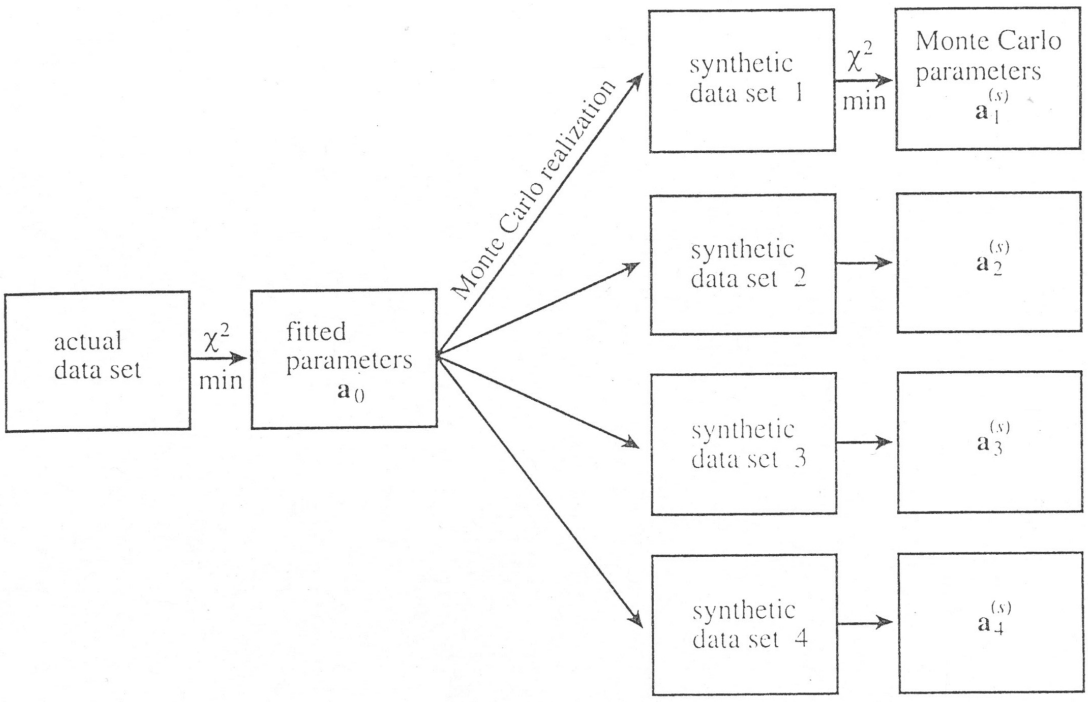

Figure 10.9: Monte-Carlo simulation of an experiment

(ii) Non parametric bootstrap [14, 43]

- Generate "new" data sets through pull with put back out of the original data (incl. their errors)
- ≈ 32 % of the data will be replaced.
- Some data points will appear more than oce
- Fit parameters.
- Confidence regions from distributions of the fits
- Correctness of the method: Deep.
- Takes empiric distribution of the errors into account

One example for (i), [62], p. 110 ff

- Consider the model from exercise sheet 5:

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

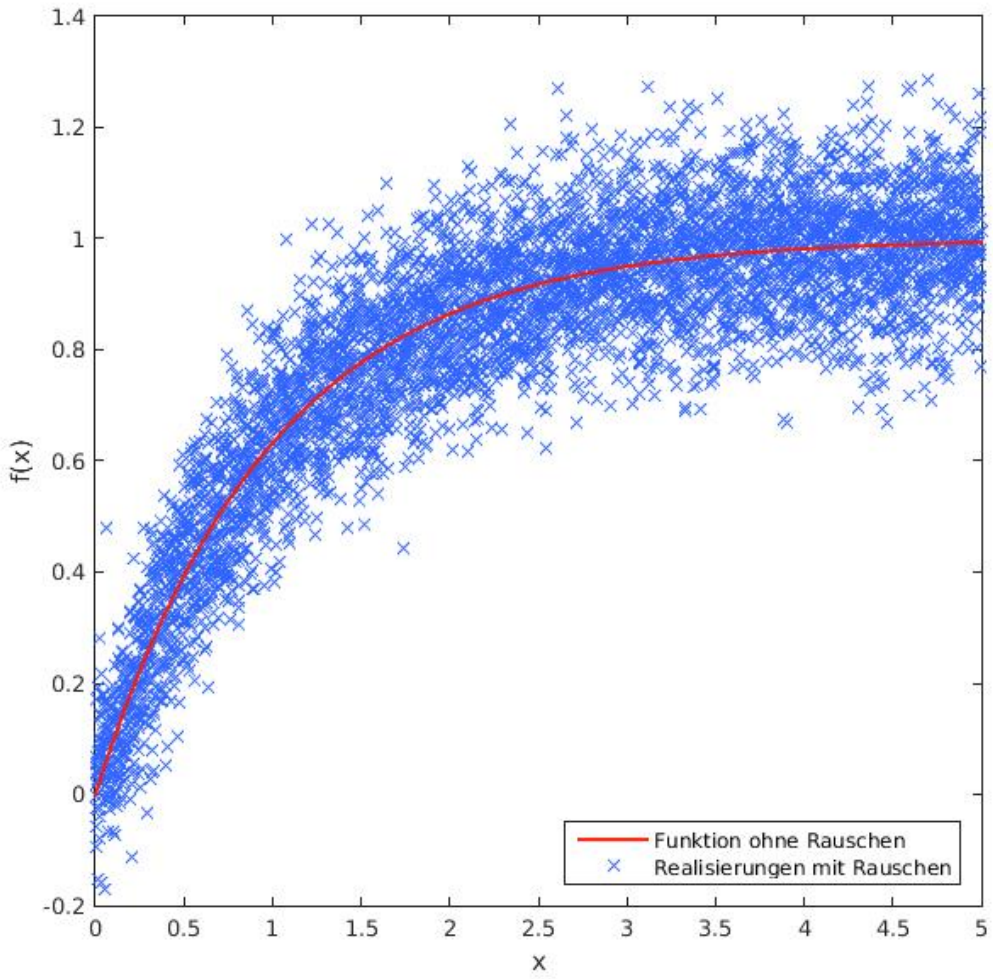

Figure 10.10: Plot of the function with $\sigma = 0.1$

- This function has 2 characteristics:
  - Slope at 0: $\frac{dy}{dx}|_{x=0} = \beta\gamma$
  - Saturation for $x \to \infty$: β
- If one chooses $x \in [0, 1]$, it follows:

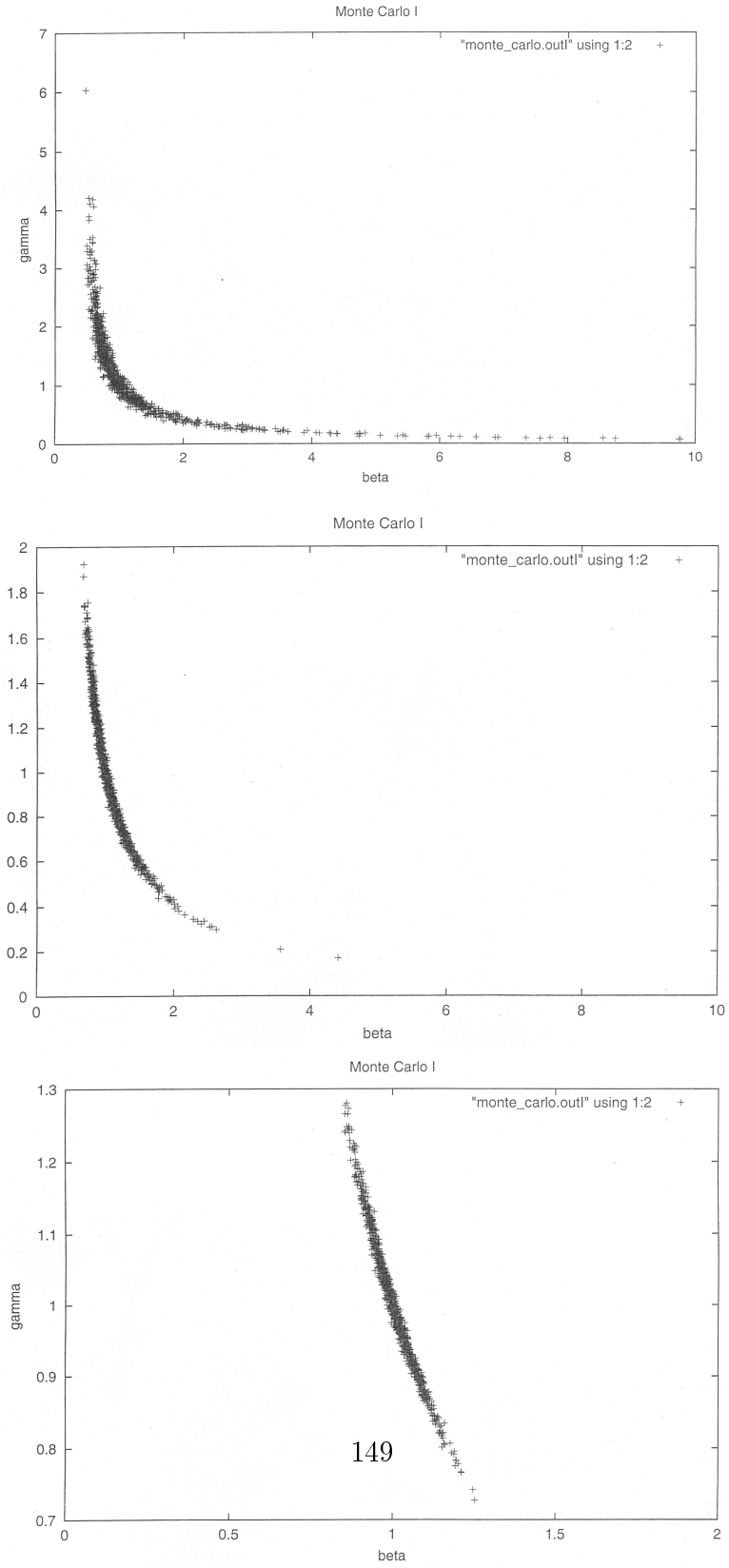

Figure 10.11: $x \in [0,1]$, $n_1 = 10, n_2 = 100, n_3 = 1000$

- If one chooses $x \in [0, 5]$, it follows:

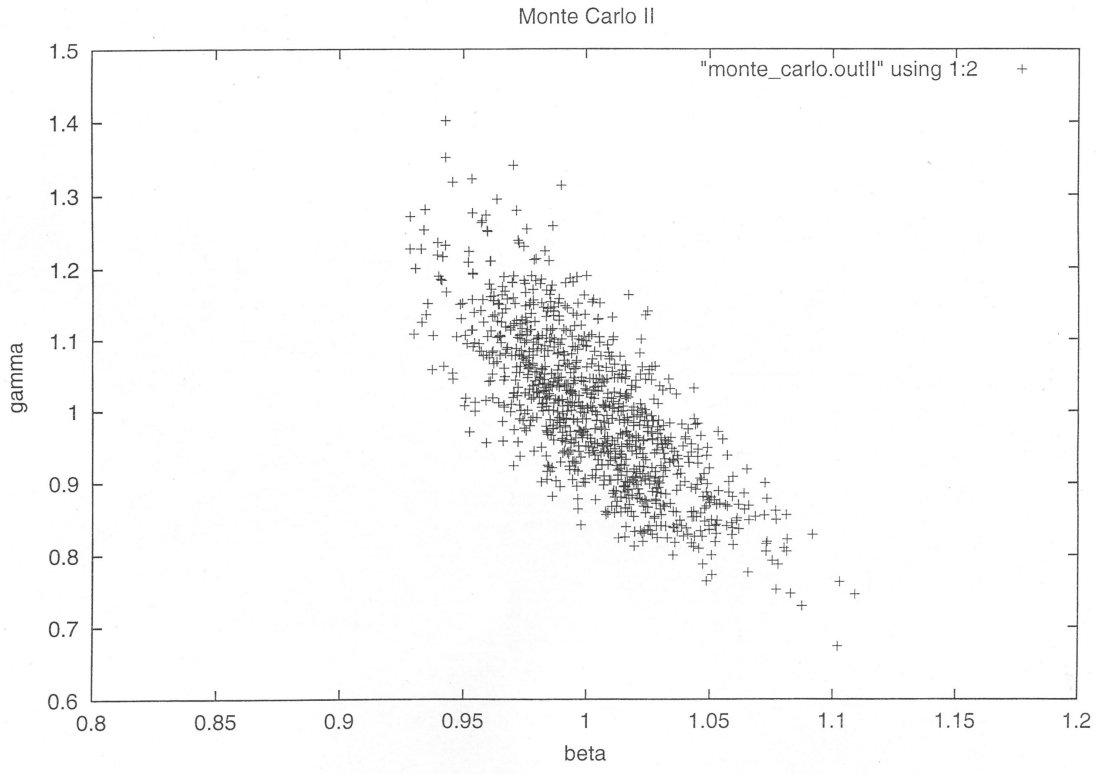

Figure 10.12: $x \in [0, 5]$, $n = 10$

Explanation, $\Rightarrow$ experimental design [52].

Lessons learned:

- Linear regression and robust estimators for non-gaussian distributions
- Non linear regression
- Non linear modeling, Levenberg-Marquardt algorithm
- Confidence intervals

11 Integration of differential equations

11.1 Ordinary differential equations (ODE)

Literature:

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

Task:

- Given a dynamical system:

$$\dot{\vec{x}} = \vec{f}(\vec{x}), \quad \text{Initial value} : \vec{x}(t_0)$$

- Find trajectory $\vec{x}(t)$, $t > t_0$, which matches the true trajectory up to controllable error.

  Nomenclature:

$$\frac{d}{dt} = \dot{}, \quad \frac{d}{dx} = ', \text{ Consider: } \ddot{x} = \dot{f}(x) = f'(x)\dot{x} = f'(x)f(x) \tag{20}$$

11.1.1 Explicit procedure

Basic idea :

- Integration step : h
- Taylor evolution :

$$x_{t+h} = x_t + \dot{x}_t h + \frac{1}{2}\ddot{x}_t h^2 + \frac{1}{6}x_t^{(3)} h^3 + \mathcal{O}(h^4) \tag{21}$$

  $\dot{x}_t$ given by $f(x_t)$, but one does not want to compute $x_t^{(n)}$.

- Abort after first order: Euler method:

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

  "First order procedures"

- Idea: Higher order through smart function evaluation.
  - Consider:

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

  - With Eq. (20) second order term cancels itself in Eq. (21) and one obtains a second order procedure (Midpoint Method).

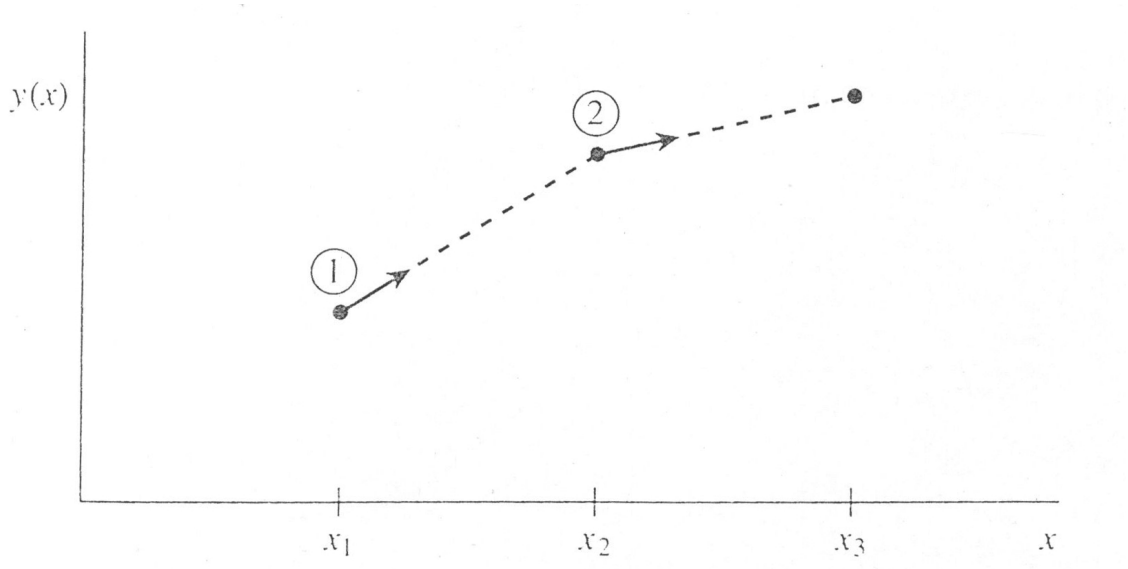

Figure 11.1: Euler method. simplest and least precise method to integrate an ODE

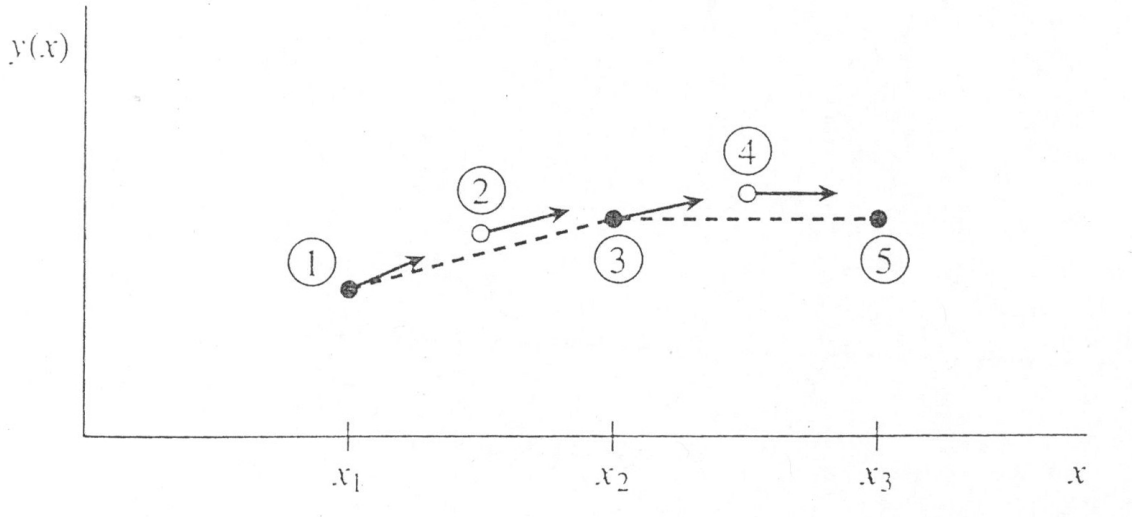

Figure 11.2: Midpoint method. Second order method

– This thread can be continued
In general:

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

Specially :

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

is called 4. order Runge-Kutta (1895)

- "Explicit", because x_{t+h} is given explicitly given by values form earlier time points.

- Belongs to the 5 most important routines there are.
- In general:

  A 4. order Runge-Kutta step with h is more precise than 2 Midpoint steps with $h/2$ is more precise than 4 Euler steps with $h/4$.

Step length control
"up to a controllable error"
Approximation error is a function of $f(.)$, step length should be adapted.

- Idea 1:

  Step Doubling: Integrate ODE with 4. order Runge-Kutta with

  (i) Step length h: Result : $x_1(t+h)$
  (ii) Two steps with $h/2$: Result : $x_2(t+h)$

  The difference:

  $$\Delta = x_2 - x_1$$

  estimates the approximation error, of order $\mathcal{O}(h^5)$.

- Idea 2:

  Embedded Runge-Kutta: Integrate ODE with

  (i) 5. order Runge-Kutta result : $x_5(t+h)$
  (ii) 4. order Runge-Kutta result : $x_4(t+h)$
  without extra work (=embedded)

  The difference:

  $$\Delta = x_5 - x_4$$

  estimates the approximation error, of order $\mathcal{O}(h^5)$

Practical procedure:

- Choose h and desired precision Δ_g

- Determine the to h belonging error Δ
- Consider Δ scales with h^5.
- Choose desired h_g after:

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

Choose desired precision Δ_g

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

Richardson extrapolation, Stoer-Bulirsch method
Idea:

- The error Δ is a function of h with $\Delta(0) = 0$.
- Determine $\Delta(h_i)$, $h_i = h_0/i$ and extrapolate to $\Delta(0)$.

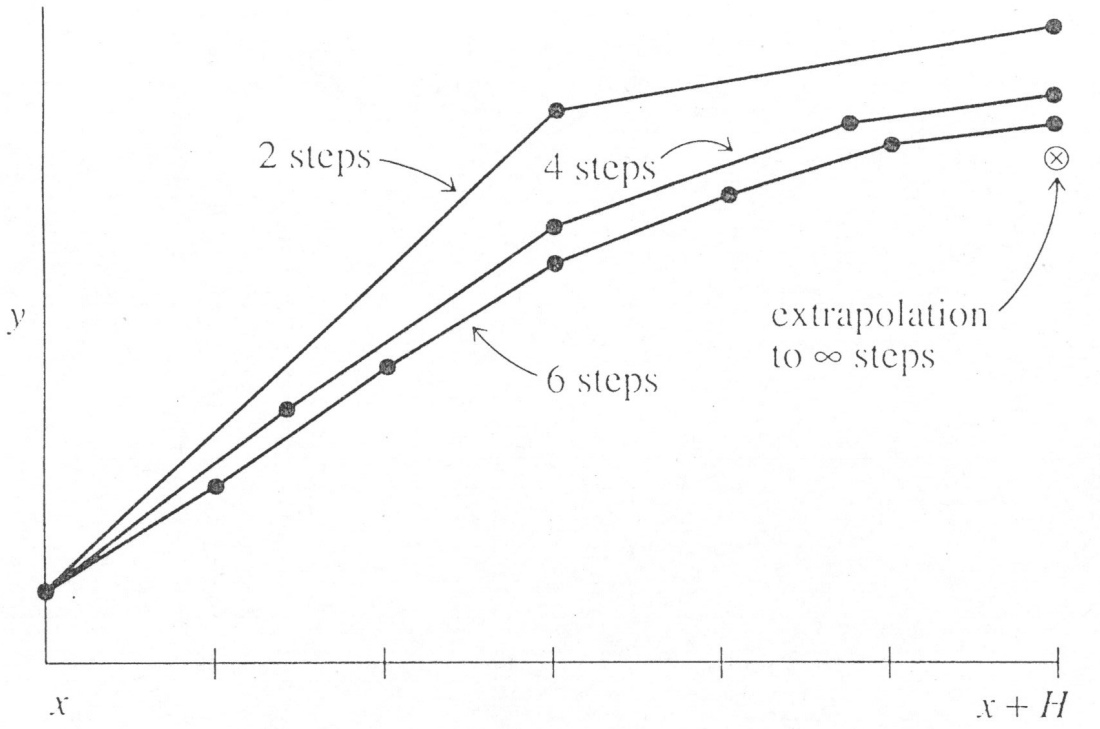

Figure 11.3: Richardson extrapolation, like in the Stoer-Burlisch method used

- Extrapolation yields also estimation error of the extrapolation.

11.1.2 Implicit procedure

The problem:

- Consider the ODE system:

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

  with $\lambda_i > 0$.

- The general solution is:

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

- By integrating the equations with the Euler method, the numerical trajectories are:

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

  Those converge only if: $|1 - h\lambda_1| < 1$, $|1 - h\lambda_2| < 1$

- Let $\lambda_2 \gg \lambda_1$, then
  - Component $C_2 e^{-\lambda_2 t}$ can be neglected for the solution ,
  - But step length is given by λ_2.
- Systems of this kind are called stiff. Step length control converges to $h = 0$.
- Above argument also holds for Runge-Kutta and Stoer-Bulirsch.

The solution:
Implicit method

- Consider 1D case:

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

  The explicit (or forward-) Euler method is:

$$x_{t+h} = x_t + \dot{x}_t h = (1 - ch)x_t \tag{22}$$

  Remember: "Explicit", because x_{t+h} here explicit given by x_t.

- Method is unstable, when $h > 2/c$, then $|x_t| \to \infty$ for $t \to \infty$.

- Eq. (22) based on:

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

it also holds (implicit differentiation):

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

This leads to:

$$x_{t+h} = x_t + \dot{x}_{t+h}h = x_t - cx_{t+h}h \Longleftrightarrow x_{t+h} = \frac{x_t}{1+ch} \tag{23}$$

an implicit method, because x_{t+h} is present on both sides of the equation.

- This is stable for all h, for linear systems there are for $h \to \infty$ even the correct asymptotic solution.
- Above argument also holds for non linear systems.
  - For explicit method: Stability only for

$$h < \frac{2}{\lambda_{\max}}, \quad \lambda_{\max} \text{ largest Eigenvalue of the Jacobi matrix of} f(.)$$

  - Implicit method: always stable.
- Not all systems are linear :-)

For

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

implicit differentiation reads:

$$x_{t+h} = x_t + f(x_{t+h})h \tag{24}$$

A self-consistent equation

Trying linerisation, remember Newton step from Chap. 9 optimization :

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

Sorting yields:

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

- Hope:
  h small enough for this to be a sufficiently good solution for Eq. (24).
- Remember:
  Every iteration needs a matrix inversion.

There are generalizations for:

- Runge-Kutta 4. order: Rosenbrock method
- Stoer-Bulirsch extrapolation: Bader-Deuflhard method

11.1.3 Integration of Hamiltonian systems

Recap Hamiltonian systems

- Exist for a d-dimensional Hamiltonian system $d/2$ conserved variables, the system is integrable
- Then dynamic is equivalent to a torus.
- If the system is integrable, one concentrates on angular variables and only needs to evaluate sine functions.

Otherwise:

- For Hamiltonian systems:

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

  has to fulfill the flux representation f_H^t :

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

and the theorem of Louiville, ,meaning:

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

- Such algorithms are called symplectic integrators see [10, 15].
- Idea: After every step one projects back to the allowed energy shell.

Exercise:
Integration of the van der Pol oscillator 10. week

11.2 Partial Differential Equation

This chapter was provided by Daniel Lill.

PDEs are differential equations in multiple variables, for example the diffusion equation:

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

with diffusion constant D.

They are omnipresent in physics:

- Wave equation
- Maxwell equation
- Schrödinger equation

General :

- Like with ODEs: Discretization. Here: More dimensional grid.
- An exact solution needs appropriate boundary conditions.
  Two important classes:

– Initial value problems, for example wave equation
  Every time step can be calculated one after the other.
– Boundary value problem, for example Poisson equation
  Simultaneous solution on entire grid

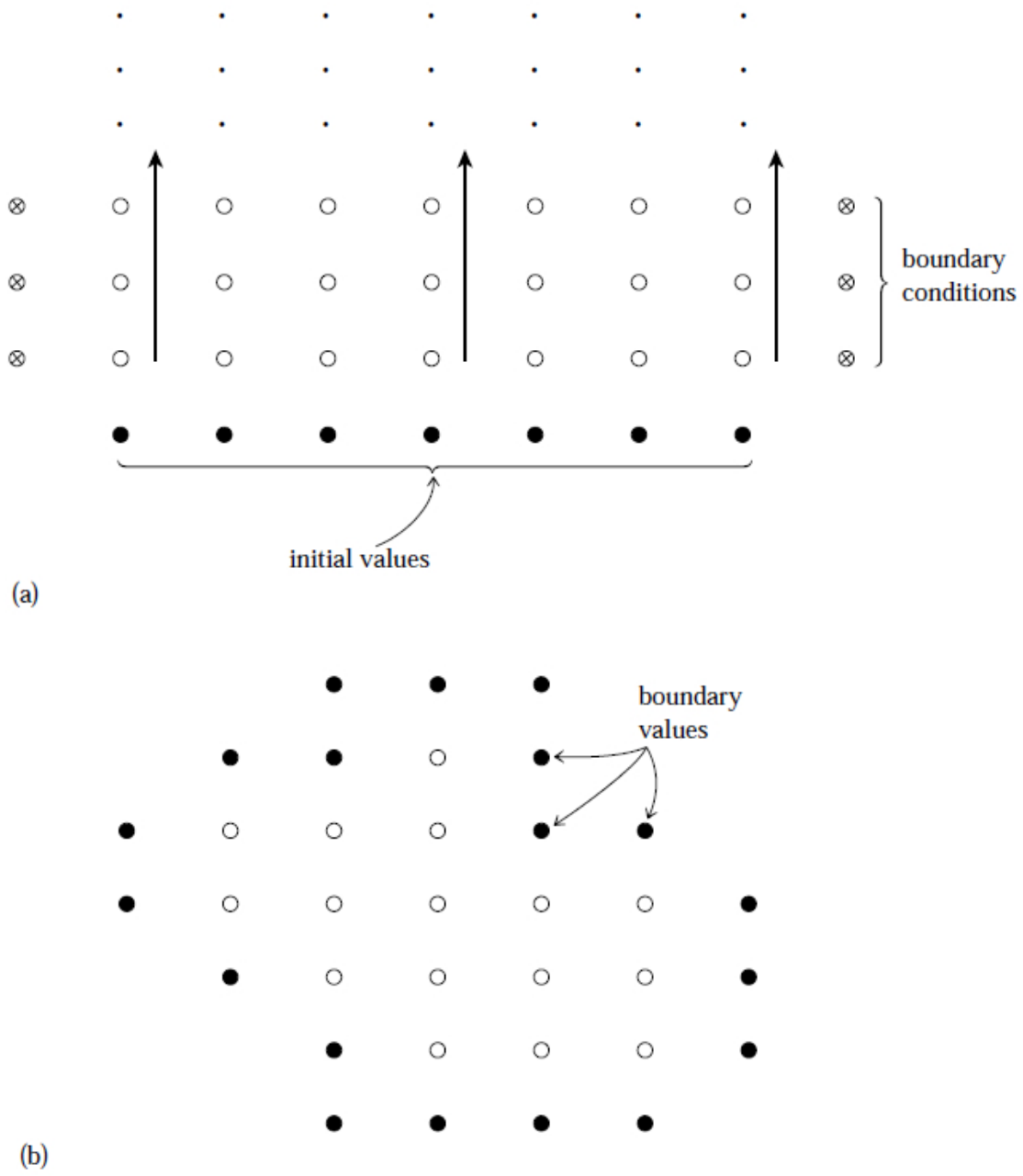

Figure 11.4: On the difference between initial- and boundary value problems , from Numerical Recipes, 3. ed

11.2.1 Initial Value Problem on the Example of the One-Dimensional Diffusion Equation

The diffusion equation in one dimension:

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

with diffusion constant D and the boundary condition $u(x,0) = f(x)$

Finite differences

- First derivative

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

- Second derivative

  Taylor evolution

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

  Addition of the equations with "+" and "−"

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

  gives an approximation of the first derivative:

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

FTCS differences scheme

- FTCS = Forward Time Centered Space differences scheme on x-t-grid:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D\,\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) \tag{25}$$

  with $u_j^n = u(j\Delta x, n\Delta t)$ with $j = 1, \ldots, J$ and $n = 1, \ldots, T/\Delta t$

- Boundary conditions are: $u_j^0 = f_j$ and z.B. $u_0^n = u_{J+1}^n = 0$

- Main question: Is this algorithm stable?

  An algorithm is called stable if it is insensitive towards rounding errors.

von Neumann stability analysis:

- Let $u_j^n = N_j^n + \epsilon_j^n$
  - u_j^n the exact solution of the differential equation
  - N_j^n the solution with rounding errors
  - $(-)\epsilon_j^n$ the rounding error
- Consider: Due to linearity the rounding error ϵ_j^n also solves the PDE.

  The rounding error thus has the same growth characteristics the solution itself.
- Taking a look at the separation ansatz

$$u_j^n = T_n X_j$$

  with this Eq. (25)

$$T_{n+1}X_j - T_n X_j = s\, T_n(X_{j+1} - 2X_j + X_{j-1})$$

  with

$$s = \frac{D\Delta t}{\Delta x^2}$$

- Divide by T_n and X_j and sorting yields:

$$\frac{T_{n+1}}{T_n} = 1 - s\left(2 - \frac{(X_{j+1} + X_{j-1})}{X_j}\right)$$

  Left side only depends on n, Right side only depends on $j \Longrightarrow$ both sides need to be constant.

$$\frac{T_{n+1}}{T_n} = g \quad \Longrightarrow \quad T_n = T_0 g^n$$

  Growth factor g

$$1 - s\left(2 - \frac{X_{j+1} + X_{j-1}}{X_j}\right) = g \quad \Longrightarrow \quad g = 1 - 2s(1 - \cos(k\Delta x))$$

  Ergo: Stable if $|g| < 1$, thus $s < \frac{1}{2}$.

- Strong restriction for the step size Δt, which goes $\propto \Delta x^2$.

  For ODEs a small Δt is sufficient, here additional assumption is required.

Implicit differences scheme BTCS

- The implicit differences scheme BTCS (Backward Time Centered Space)

$$u_j^{n+1} - u_j^n = s(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) + \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) \tag{26}$$

  has growth factor

$$g = \frac{1}{1 + 4s\sin^2(k\Delta x/2)}$$

  and is thus stable for all s.

- For $\Delta t \to \infty$ the equation is in equilibrium:

$$\partial_{xx} u = 0$$

  and the solution is, the same as with implicit ODEs solvers for linear systems, asymptotically correct.

Crank-Nicolson scheme

- Crank-Nicolson scheme: Mean between FTCS- and BTCS scheme.

$$u_j^{n+1} - u_j^n = \frac{s}{2}(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} + u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)$$

  growth factor

$$g = \frac{1 - s(1 - \cos k\Delta x)}{1 + s(1 - \cos k\Delta x)}$$

- Stable for al s
- Cut time error is $\mathcal{O}(\Delta t^2)$.

For non-linear PDE von Neumann stability analysis only yields necessary but not always sufficient stability conditions.

11.2.2 Boundary Value Problem

Example: Laplace equation:

$$\triangle u(\vec{x}) = 0 \text{ für } \vec{x} \in V$$

given with boundary conditions for $u(\partial V)$ or $\frac{\partial u}{\partial n}(\partial V)$.

- Finite differences for 2D Laplace equation:

$$\frac{u_{j+1,i} - 2u_{j,i} + u_{j-1,i}}{\Delta x^2} + \frac{u_{j,i+1} - 2u_{j,i} + u_{j,i-1}}{\Delta y^2} + \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta y^2) = 0$$

  Sorting yields:

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1})$$

  u_{ij} is then the mean of its nearest neighbors.

- Naive iterating: Jacobi iteration

$$u_{i,j}^{(n+1)} = \frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)})$$

- Faster is the Gauß-Seidel procedure:
  - Start on the bottom left and calculate the values in the first row from left to right
  - Use the new values already for the next points

$$u_{i,j}^{(n+1)} = \frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)})$$

- Even faster is the successive over-relaxation procedure:

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \omega(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)} - 4u_{i,j}^{(n)}) \tag{27}$$

  with cleverly chosen ω.

11.2.3 Method of Finite Elements

- Instead of PDE approximation by finite differences ...
- Approximation of the solution through linear combination of basis functions

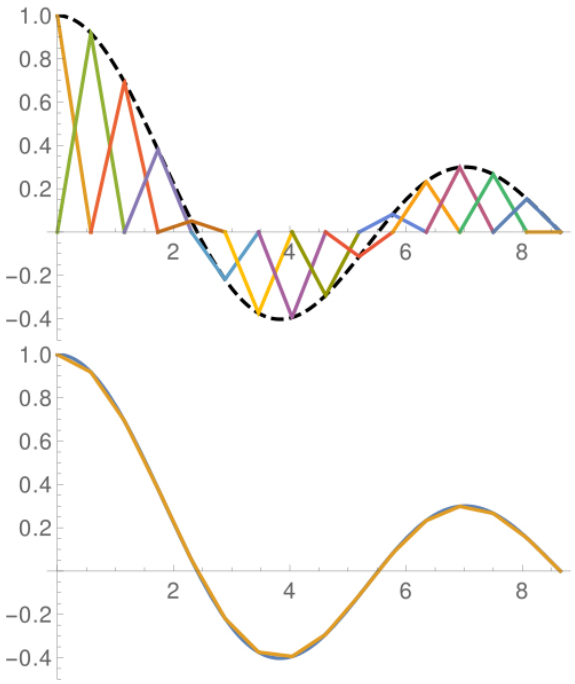

Figure 11.5: The Bessel function is approximated by the linear combination of the colored triangle functions.

Example:

- Poisson equation in 1 D

$$u''(x) = -\rho, \qquad x \in [0,1]$$

Boundary condition

$$u(0) = u(1) = 0$$

Formulation of the problem in its weak form with respect to the basis functions v_i:

$\forall v_i$ mit $v_i(0) = v_i(1) = 0$ holds:

$$\int_0^1 -\rho v_i(x)dx = \int_0^1 u''(x)v_i(x)dx = u'(x)v_i(x)|_0^1 - \int_0^1 u'(x)v_i'(x)dx \qquad (28)$$

The first term on the right disappears due to the boundary conditions of v_i.

- Divide the region $[0, 1]$ into smaller intervals $[x_i, x_{i+1}]$ and link each point x_i to a triangle function:

$$v_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & x \in [x_i, x_{i+1}] \\ 0 & \text{else} \end{cases}$$

The function $u(x)$ is expressed as a linear combination of these v_i

$$u(x) = \sum_i a_i v_i(x) \qquad (29)$$

Instead of infinite dimensional now finite dimensional

- Eq. (28) becomes with Eq. (29):

$$\forall j: \qquad \int_0^1 \rho v_j(x)dx = \sum_i a_i \int_0^1 v_i'(x)v_j'(x)dx$$

A linear system of equations.

- With

$$M_{ij} = \int_0^1 v_i'(x)v_j'(x)dx$$

and

$$w_j = \int_0^1 \rho v_j(x)dx$$

follows

$$\sum_i M_{ij} a_i = w_j \quad \Longrightarrow \vec{a} = M^{-1}\vec{w}$$

Consider: The matrix M is only sparse, therefor inversion is quick.

In figure 11.6 the 1D Poisson problem is shown with $\rho = -2$ and with two different grids. The parabola is well displayed in both, but on the right hand side the grid was chose more coarse grain towards the edges.

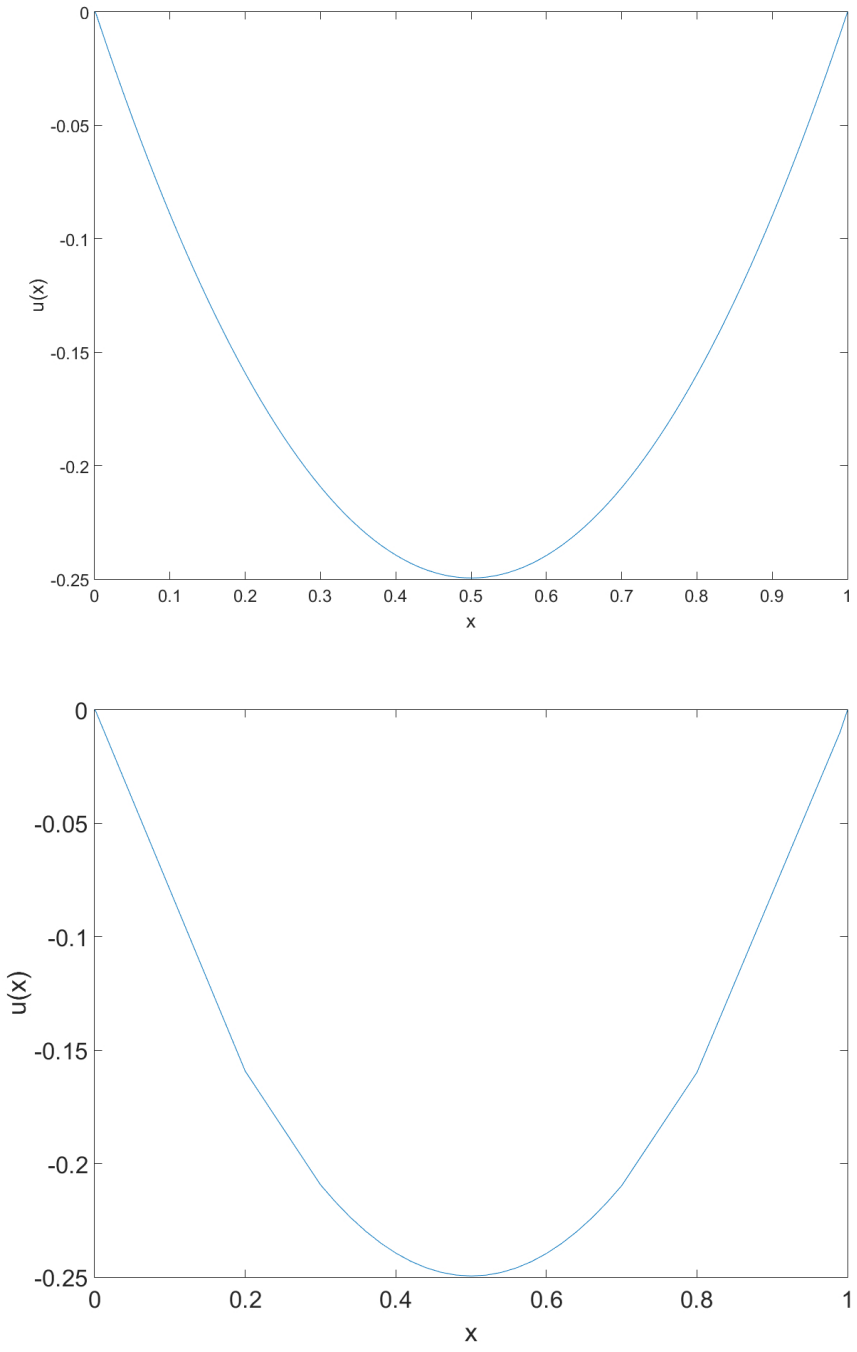

Figure 11.6: The 1-d Poisson problem a) with fine grid b) with coarse grid on the edges and with fine grid towards the minimum of the parabola

Comments to the method of finite elements

- Generalization towards higher dimensions: In 2D the area is triangulated. To every point i multiple triangles j will then be linked, to which a basis function $v_j(x, y)$ belongs. Analogous for higher dimensions.
- Grid can be adjusted flexibly to the geometry of the problem. Example: Crash simulations, where the grid of the crumple zone is finer than that of rear of the vehicle .
- Different basis function, for example polynomials, are also possible.
- Spectral methods work similarly:
  - The function $u(.)$ is evolved into a finite amount of basis functions
  - But: Basis functions have full support, for example Fourier series
  - Cut at the determined frequency

11.3 Stochastic differential equations

Literature:

- P.E. Kloeden, E. Platen. *Numerical Solution of Stochastic Differential Equations* [31], mathematically extensive
- P.E. Kloeden, E. Platen, H. Schurz. *The Numerical Solution of SDE through Computer Experiments* [32], with simulation software
- B. Øksendal. *Stochastic Differential Equations* [48], good book
- J. Honerkamp. *Stochastic Dynamical Systems* [25] Chap. 10, condensed display for physicists

Stochastic differential equation (SDE), physicists definition, Langevin equation

$$\dot{x} = f(x, \epsilon) = a(x) + b(x)\epsilon, \qquad \epsilon \sim N(0, 1)$$

- $a(x)$: Deterministic part : Drift-term
- $b(x)\epsilon$: Stochastic part: Diffusion-term
- ϵ: Dynamic noise
- Fundamental problem : $\dot{x}$ and x not smooth
- Mathematical-definition

$$dx = a(x)dt + b(x)dW \tag{30}$$

  more to this below

Why stochastic DEs ?

- Modeling of outside influences on open (deterministic) systems.

  Classic example: Brownian motion:

$$x(t) = x(t-1) + \sigma\epsilon(t), \quad \epsilon(t) \sim N(0, 1)$$

  Time scale separation between slow pollen and fast moving water particles

  Physical interpretation

$$\begin{aligned} x(t) &= x(t-\Delta t) + \sigma\epsilon(t) \\ \frac{x(t)-x(t-\Delta t)}{\Delta t} &= \frac{\sigma\epsilon(t)}{\Delta t} \\ \lim_{\Delta t\to 0}: \quad \dot{x} &= \tilde{\epsilon} \end{aligned}$$

Velocity is white noise with 0 mean, we are going to think about $\tilde{\epsilon}$ further down.

- Modeling of complicated parts in a deterministic system.
- In fact always needed in non Hamiltonian dissipative systems because of the Fluctuation-dissipation theorem: Where there is friction, there is stochastic behaviour in dynamics [40].
- In Hamiltonian systems noise leads to divergence.

Meaning term $b(x)\epsilon$:

- State dependent variance
- Parametric noise:
$$\dot{x} = -(c+\epsilon)x = -cx - \epsilon x$$
Noisy parameter

Integration of SDEs
Instead of a taylor evolution in Eq. (21) different methods for integration of deterministic DE from Chap. 11 can be read as approximations of integrals:

$$\begin{aligned} \dot{x} &= f(x) \\ &\Longleftrightarrow \\ x_{t+h} &= x_t + \int_t^{t+h} f(x_{t'})dt' \end{aligned}$$

- Explicit Euler method: $\int_t^{t+h} f(x_t')dt' \approx f(x_t)h$
- Implicit Euler method: $\int_t^{t+h} f(x_t')dt' \approx f(x_{t+h})h$

- Runge-Kutta: Integral evaluation on multiple points

For SDEs this only works over integral interpretation.

$$x_{t+h} = x_t + \int_t^{t+h} f(x_{t'}, \epsilon_{t'})dt' = x_t + \int_t^{t+h} (a(x_{t'}) + b(x_{t'})\epsilon_{t'})dt'$$

Consider easiest example: Linear damped stochastic driven system

$$\begin{aligned} \dot{x} &= -\alpha x + \sigma\epsilon \\ x_{t+h} &= x_t + \int_t^{t+h} -\alpha x_{t'}dt' + \sigma \int_t^{t+h} \epsilon_{t'}dt' \end{aligned}$$

But what is an integral over $\epsilon_{t'}$?

- Consider:

$$\int_t^{t+h} \epsilon_{t'}dt'$$

  Does not make sense in neither Riemann nor Lebesgue way.

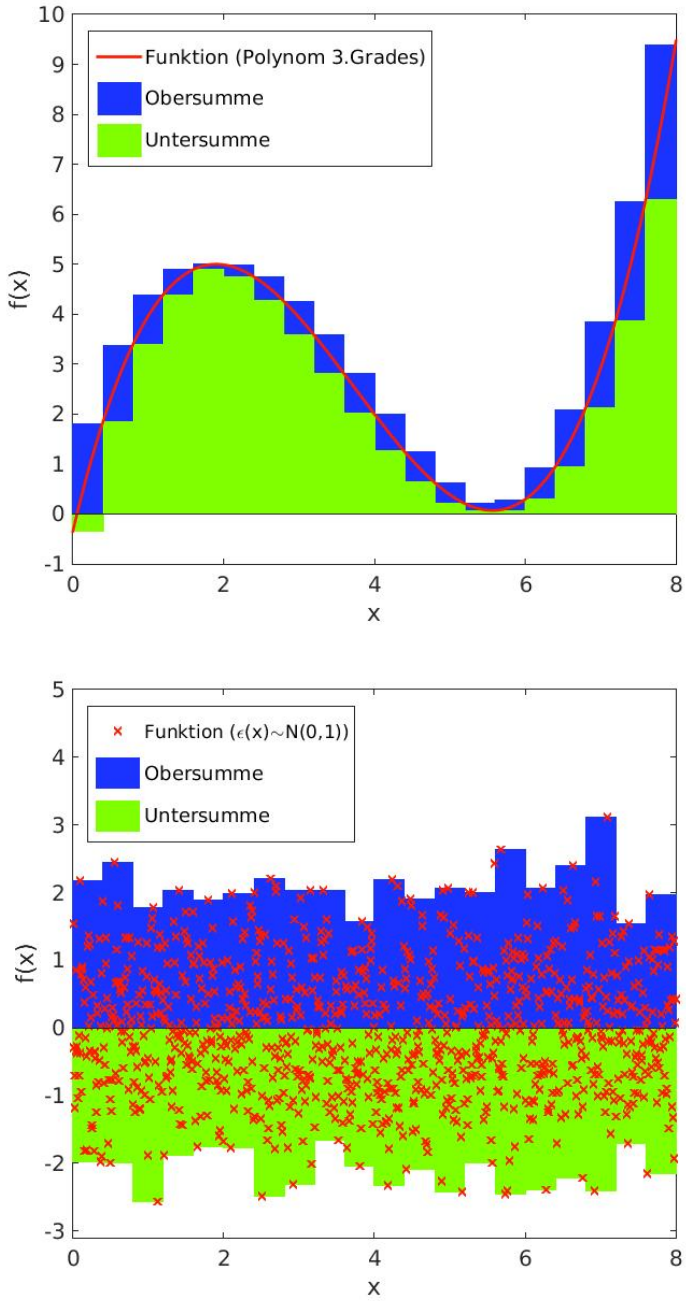

Figure 11.7: Over- and undersumms in a polynomial and a stochastic function

- Observation:

Result of the integral is Brownian motion

- Brownian motion in discrete time ($\Delta t = 1$) is:

$$x(t) = x(t-1) + \sigma\epsilon(t) \qquad x(0) = 0, \quad \epsilon(t) \sim N(0,1)$$

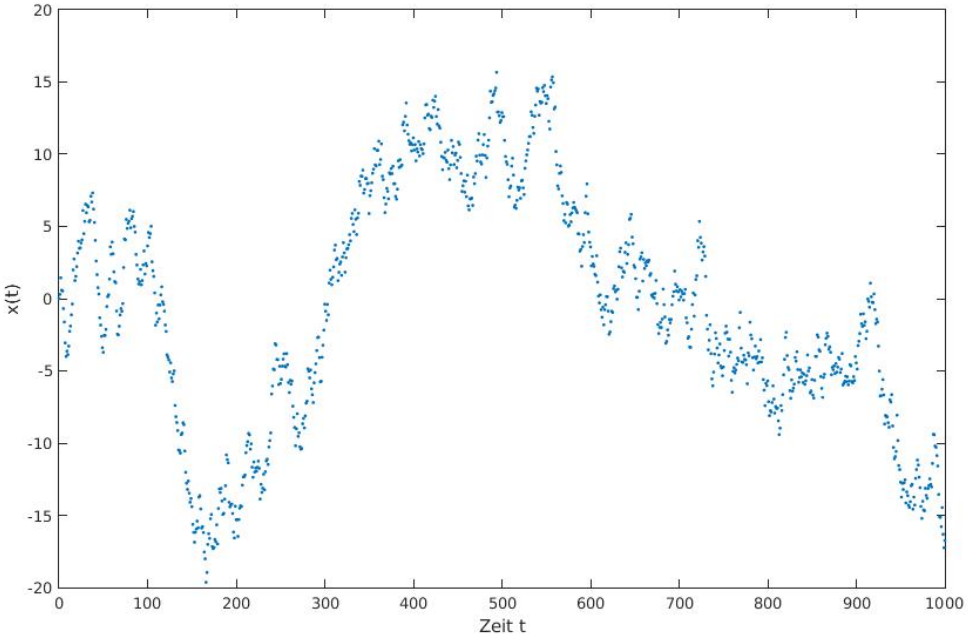

Figure 11.8: Brownian motion

- Put into one another:

$$\begin{aligned} x(t-1) &= x(t-2) + \sigma\epsilon(t-1) \\ x(t) &= \sigma \sum_{t'=1}^{t-1} \epsilon(t') \end{aligned}$$

- Since variance additive it holds:

$$\langle x^2(t) \rangle = \sigma^2 t, \quad \langle x(t) \rangle = 0 \tag{31}$$

$x(t)$ is Gaussian random variable with mean 0 standard deviation $\sigma\sqrt{t}$

- DEFINE:

$$\boxed{\int_t^{t+h} \epsilon_{t'} := \sqrt{h}\epsilon_t}$$

- Remarks: Mathematicians turn it around:
  1. Define time continuous Brownian motion through Eq. (31), Wiener process
  2. Define "ϵ" as increments, i.e. additions to the Wiener process, dW in Eq. (30)

For connoisseurs to self study : Ito and Stratonovich integral

- For additive noise identical
- For multiplicative noise different

With this, Euler method for $\dot{x} = a(x) + b(x)\epsilon$

$$x_{t+h} = x_t + a(x_t)h + b(x_t)\epsilon_t\sqrt{h} + \mathcal{O}(h)$$

- Higher order in general very difficult since appearance of very complicated statistical integrals, see [25].
- Euler causes: Integration time step in general $\ll$ natural sampling timestep, see [71] especially for choice of integration time step.

Exercise:
Integration of the stochastic van der Pol oscillator 11. week

11.4 Gillespie algorithm

Literatur:

- Original [21]
- See also: [19, 44, 53]
- Critical examination of fundamentals and interpretation [75]

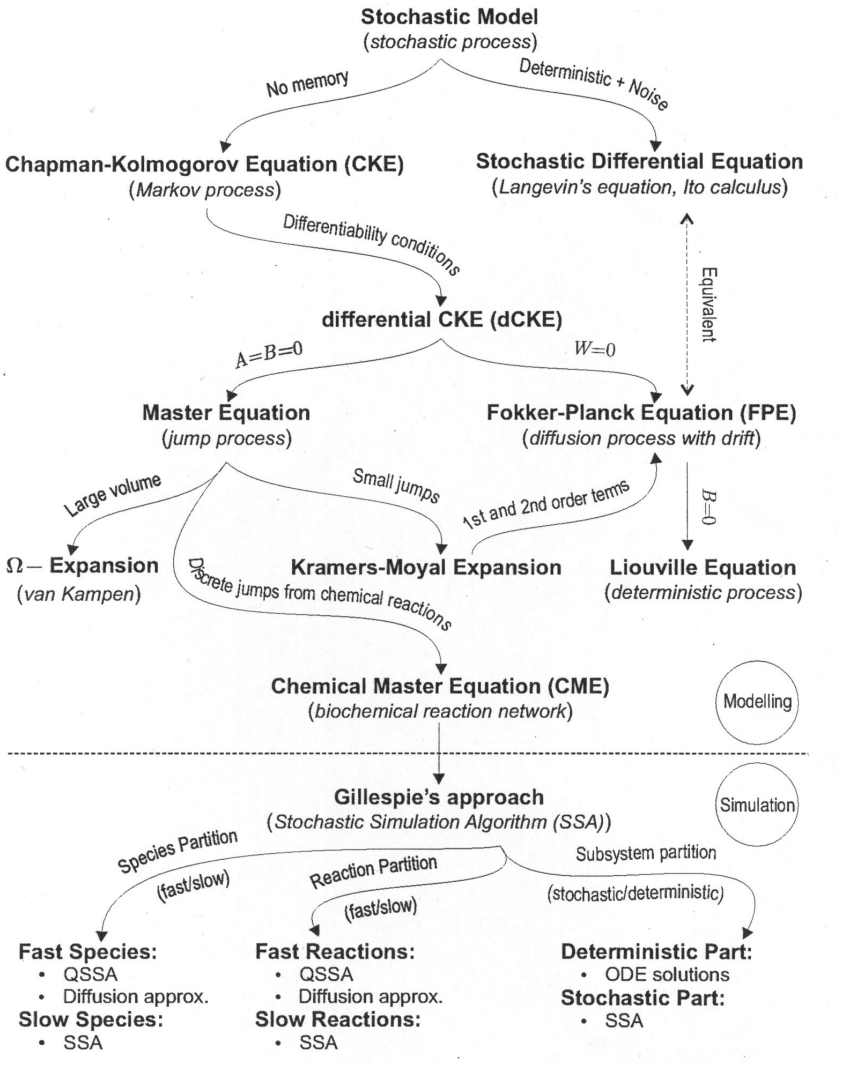

Figure 11.9: Overview of statistical models

All dynamics are discrete

- Population dynamics of animals
- Chemical reactions between molecules
- Banking traffic
- Occupation number formalism in quantum mechanics, a and $a^\dagger$
- DE are limit case

Consider chemical reactions in the following

Let S be a species and

$$P_n(t) = Prob(\#S(t) = n \text{ at timepoint t})$$

Consider:

- Propensity $a_i(.)$: Probability per time unit for change of state
- Influx to $P_n(t)$

  $$n - \nu_i \overset{a_i(n-\nu_i)}{\longrightarrow} n$$

  with $a_i(n - \nu_i)$ rate of change of ν_i, given the state was in $n - \nu_i$
- Outflow of $P_n(t)$

  $$n \overset{a_i(n)}{\longrightarrow} n + \nu_i$$

  with $a_i(n)$ the rate of change of ν_i, given state was in n

Then Chemical Master equation:

$$\dot{P}_n = \sum_{i=1}^{N} a_i(n - \nu_i)P_{n-\nu_i} - a_i(n)P_n$$

Usually

- More than one species: $P(S_1, S_2, \ldots, S_K)$
- Multiple possible reactions $R_1, R_2, \ldots, R_M$

- Not solvable analytically.

Gillespie algorithm: Instead of analytical solution

- Simulate many trajectories
- Determine results by averaging or consideration of distributions
- It can be shown: Gillespie algorithm produces the correct distributions

Gillespie algorithm answers:

- When will the next reaction take place?
- Which will it be?

Central property: Reaction probability function $P(i, \tau)$
$P(i, \tau)d\tau$: Probability for reaction R_i in Interval $(t + \tau, t + \tau + d\tau)$, given system in state $S(t)$

$$P(i, \tau)d\tau = P_0(\tau)\, P_i(d\tau) \tag{32}$$

with

- $P_i(d\tau) = a_i d\tau$: Probability for reaction R_i to happen in interval $(t+\tau, t+\tau+d\tau)$.
- $P_0(\tau)$: Probability of no reaction happening in interval $(t, t + \tau)$ given state $S(t)$

The probability of any reaction happening in interval $d\tau$ is:

$$\sum_{i=1}^{M} a_i d\tau$$

- Define:

$$a^* = \sum_{i=1}^{M} a_i$$

  Probability for no reaction in interval $d\tau$: $1 - a^* d\tau$.

- Therefor

$$P_0(\tau + d\tau) = P_0(\tau)(1 - a^* d\tau)$$

  Yields differential equation

$$\dot{P}_0 = -a^* P_0, \text{with solution } P_0(\tau) = e^{-a^*\tau}$$

  $P_0(0) = 1$ o.k.

- Taken together with Eq. (32):

$$P(i, \tau) = a_i e^{-a^*\tau}$$

Central questions:

- Which reaction is the next one?
- When is it going to happen?

When ?
Summation over all reactions

$$\bar{P}(\tau) = \sum_{i=1}^{M} P(i, \tau) = a^* e^{-a^*\tau}$$

$\bar{P}(\tau)d\tau$: Probability for any next reaction in the interval $(t + \tau, t + \tau + d\tau)$

Which reaction?
Given a reaction happens in interval $(t + \tau, t + \tau + d\tau)$, the conditional probability

$$\tilde{P}(i|\tau) = \frac{P(i, \tau)}{\bar{P}(\tau)} = \frac{a_i e^{-a^*\tau}}{a^* e^{-a^*\tau}} = \frac{a_i}{a^*}$$

gives the probability of it being reaction i.

On the way to the algorithm:

- When ?

- The cumulative distribution $F(t)$ for $\bar{P}(\tau)$ reads:

$$F(t) = \int_0^t \bar{P}(\tau)d\tau = a^* \int_0^t e^{-a^*\tau}d\tau = 1 - e^{-a^*t}$$

- Let r_1 be an equally distributed random number in interval $[0, 1]$
- If one chooses t in a way that $F(t) = r_1$, the probability density of t is that of $\bar{P}(\tau)$
- With this one gets t by

$$t = F^{-1}(r_1) = \frac{1}{a^*} \log\left(\frac{1}{1 - r_1}\right)$$

- Since r_1 has same equal distribution than $1 - r_1$, it holds for the random time variable of time τ of the next reaction :

$$t = F^{-1}(r_1) = \frac{1}{a^*} \log\left(\frac{1}{r_1}\right) = -\frac{1}{a^*} \ln r_1$$

- Which one ?
  - Let r_2 be an equally distributed random number in interval $[0, 1]$
  - Which reaction takes place is determined by

$$\sum_{i=1}^{j-1} a_i \leq r_2 a^* < \sum_{i=1}^{j} a_i$$

Determination of the propensities a_i

- $c_i\, dt$: Probability that a given single reaction R_i occurs in the next time step dt.
- h_i: Number of combinations of reactants
- $a_i\, dt = h_i c_i\, dt$: Probability of reaction R_i in the next time step.
- Examples

Reaction R_i	c_i	h_i
$S_1 \xrightarrow{k} \dots$	k	$\#S_1$
$S_1 + S_2 \xrightarrow{k} \dots$	k/V	$\#S_1 \cdot \#S_2$
$2S_1 \xrightarrow{k} \dots$	$2k/V$	$\frac{1}{2}\#S_1 \cdot (\#S_1 - 1) = \binom{\#S_1}{2}$

Gillespie algorithm:

1. Initialization
   - Set $t = 0$
   - Choose number of molecules $\#S_i(0)$
2. Calculate propensities
   - $a_i\,dt = h_i c_i\,dt$: Probability of reaction R_j in next time step
   - Calculate $a^* = \sum_{i=1}^{M} a_i$
3. Draw two equal distributed random numbers r_1, r_2
   - Determine $\tau = -\frac{1}{a^*} \log r_1$
   - Determine j so that
   $$\sum_{i=1}^{j-1} a_i \leq r_2 a^* < \sum_{i=1}^{j} a_i$$
4. Update
   - Update the number of molecules according to the reaction scheme
   - Set $t = t + \tau$
   - Go to point 2.

Exercise:
Gillespie algorithm

Lessons learned:

- Runge-Kutta integrators für ODEs through clever function evaluations.
- Stiff systems need implicit integrators
- Stochastic differential equations, characteristic $\sqrt{h}$
- Partial DGLs, coupling of δx and δt in explicit methods
- Gillespie algorithm for the chemical master equation

12 Non-parametric estimators

12.1 Non-parametric density estimators

Literature:

- B.W. Silverman. *Density Estimation* [65] The bible

Exercise:

- Given N realizations x_i of a random variable X with density $\rho_X(x)$, estimate the density.
- Parametric density estimator
  - For standard deviations like Gaussian, exponential or χ^2_r estimate parameters of the distributions by comparison with the moments.
  - Alternative: Fit to the cumulative distribution of the data
- Non-parametric density estimators don't assume a parametric distribution

Naivest access: Histogram

- Split x axis into bins of width h starting from anchor point x_0:

$$bin_m = [x_0 + mh, x_0 + (m+1)h], \qquad m \in \mathbb{Z}$$

- Estimate $\rho(x)$ by

$$\hat{\rho}(x, x_0, h) = \frac{1}{Nh}(\text{Number of } x_i \text{ in } bin_m) \tag{33}$$

- Problem 1: Ho to choose anchor point x_0?
- Problem 2: How to choose h?

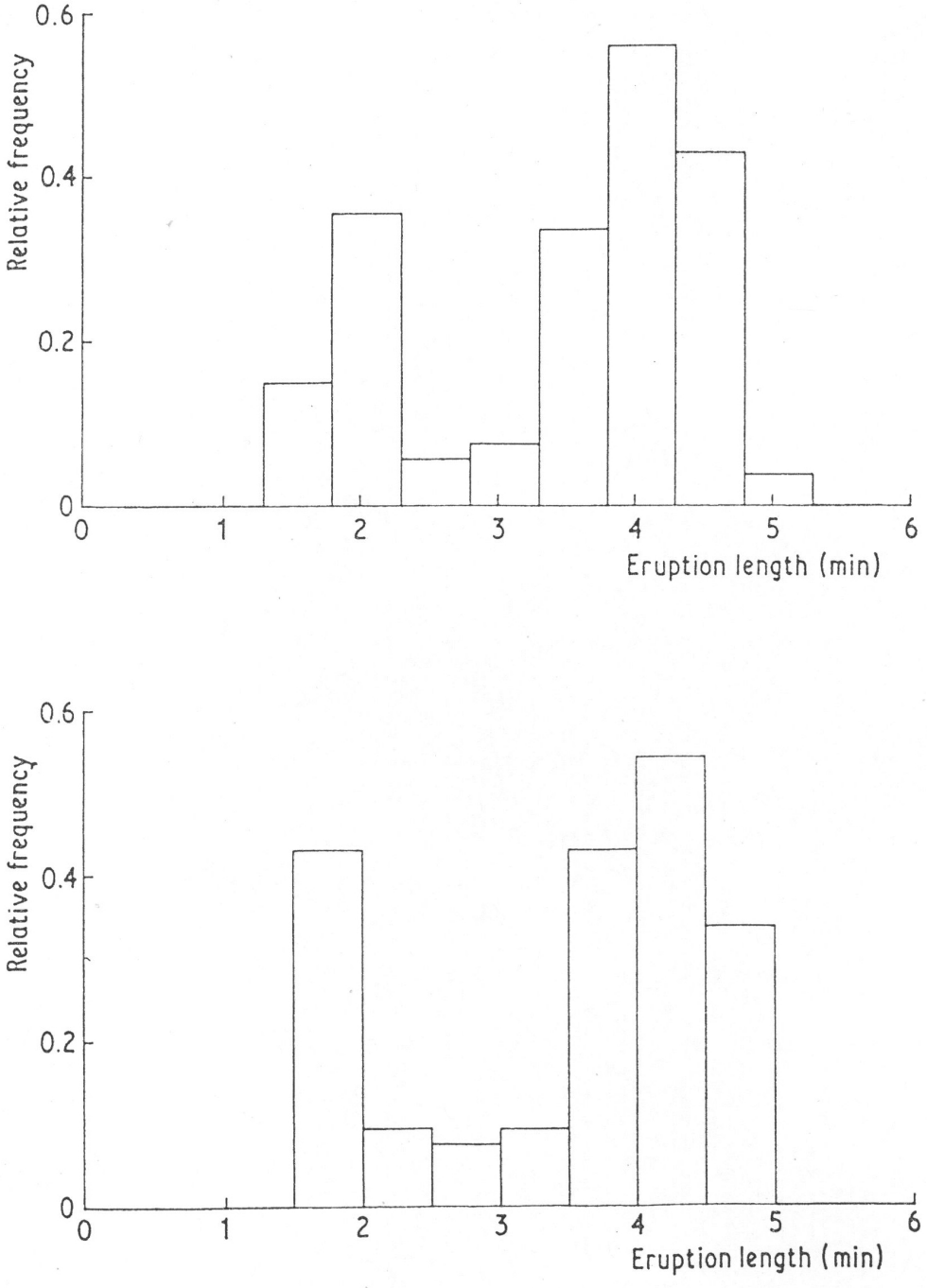

Figure 12.1: Histograms of the eruption length of the Old Faithful Geyser

Naivest access:

- Replace Eq. (33) by

$$\hat{\rho}(x) = \frac{1}{2Nh}(\text{number of } x_i \in [x-h, x+h])$$

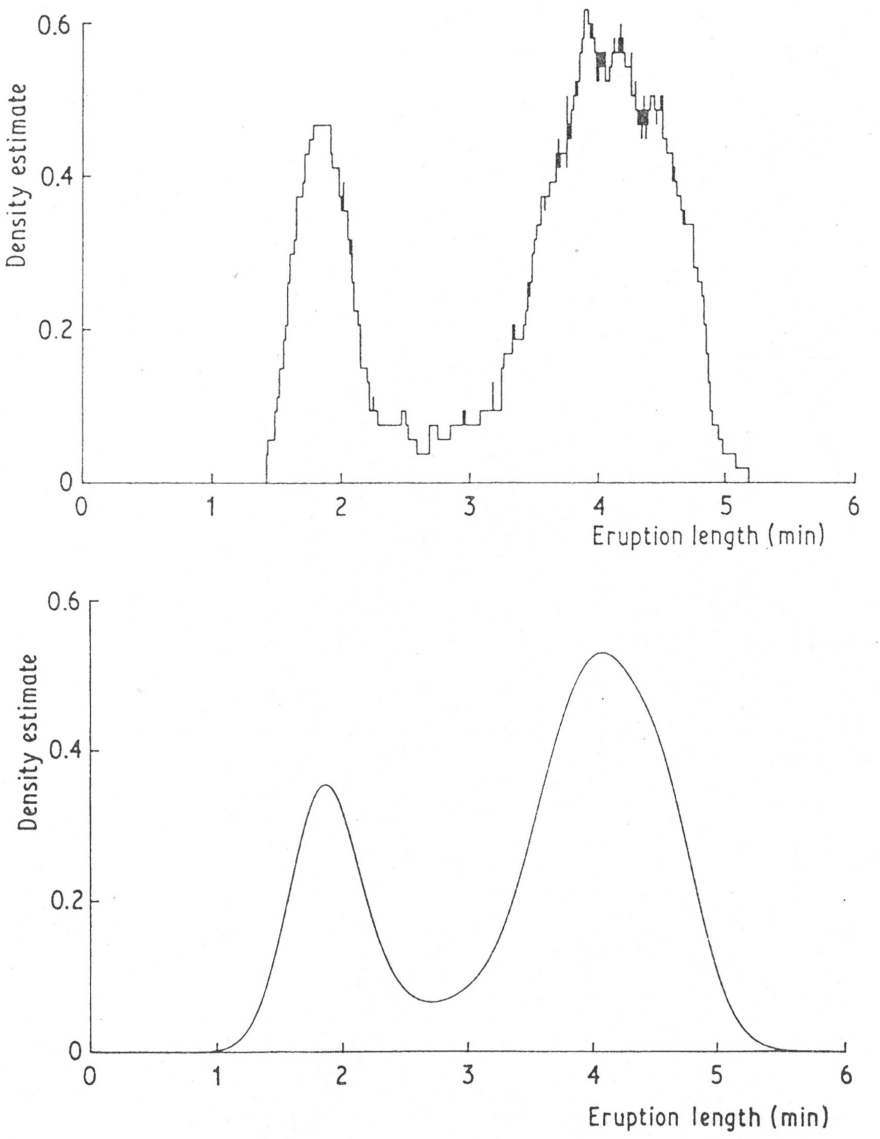

Figure 12.2: Core estimator for the data of Old Faithful Geyser

Core estimator (fixed size):

- Consider that the naive estimator can be expressed via:

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{else} \end{cases}$$

$$\hat{\rho}(x) = \frac{1}{Nh} \sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right)$$

- Idea: Instead of a rectangular box $w(x)$ choose a smooth function $K(x)$ which fulfills

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

  and which is positive for now.

$$\hat{\rho}_K(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

- Problem 2 : h ? stays
- explain "fixed size"

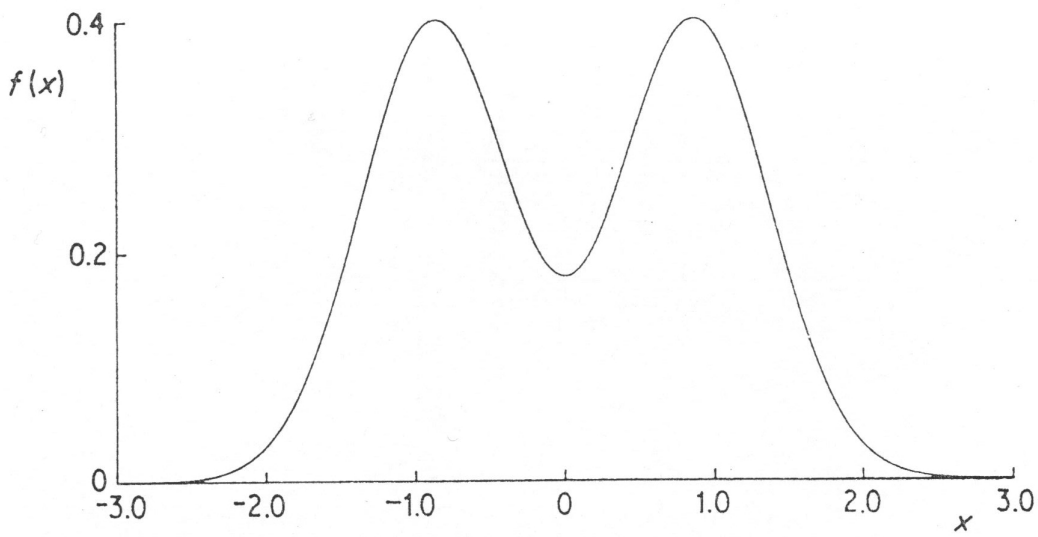

Figure 12.3: True density of the data

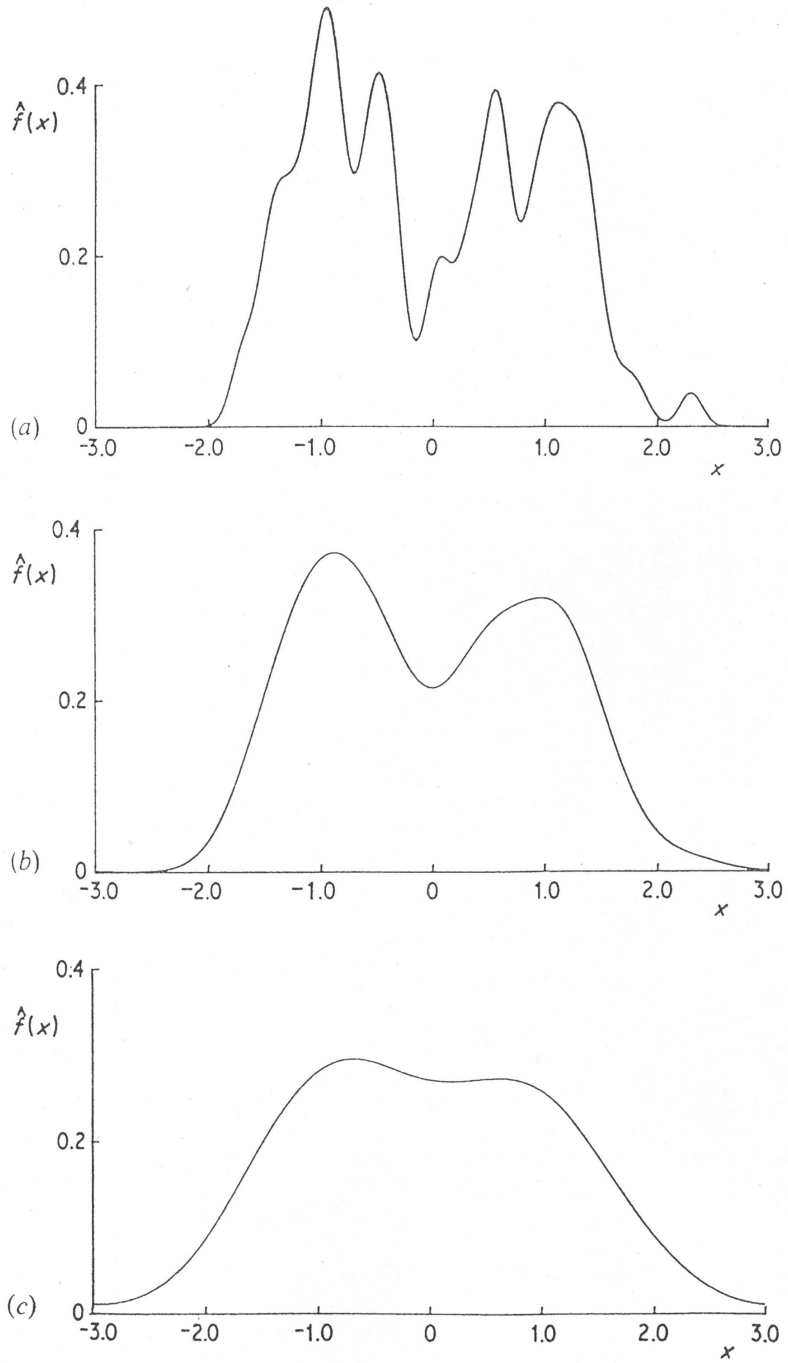

Figure 12.4: Estimated density of the data for 200 simulated data points with (a) h=0.1; (b) h=0.3; (c) h=0.6

Nearest Neighbor method (fixed mass) :

- Idea:

  Where there are many points,choose small h

- Choose: Integer k

- Let $d(x, x_i)$ be the distance between x and x_i

  Sort $d(x, x_i)$ by increasing order: $d_1(x), d_2(x), \ldots, d_N(x)$

- and define the "k-th nearest neighbor" estimator:

$$\hat{\rho}_{NN}(x) = \frac{k}{2Nd_k(x)}$$

  Illustrate equations. When solved $k = 2d_k(x)N\rho(x)$ is the expected amount.

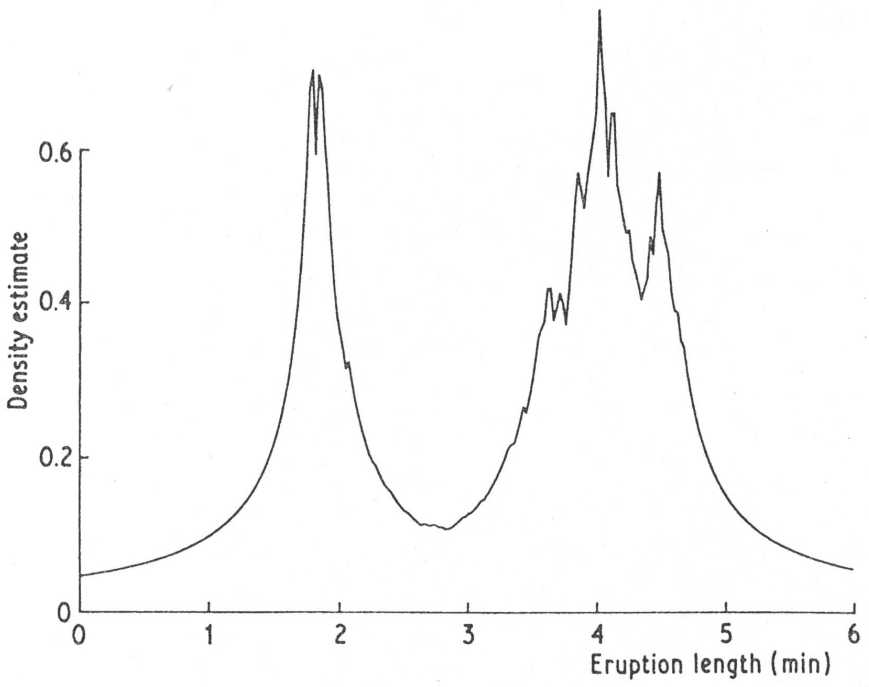

Figure 12.5: Nearest neighbor estimation for the data of the Old Faithful Geyser

- or more general:

$$\hat{\rho}_{NN}(x) = \frac{1}{N d_k(x)} \sum_{i=1}^{N} K\left(\frac{x - x_i}{d_k(x)}\right)$$

- Instead of problem 2, now problem 2a: Choice of k ?
- Explain "fixed mass"

Mathematization (for core estimator, analogous for NN estimator)

Assumptions:

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2 K(t)dt = k_2 \neq 0,$$

Calculation of the bias:

- Expectation value of the estimator

$$\langle \hat{\rho}(x) \rangle = \frac{1}{Nh} \sum_{i=1}^{N} \left\langle K\left(\frac{x - x_i}{h}\right) \right\rangle = \frac{1}{h} \int K\left(\frac{x - y}{h}\right) \rho(y) dy$$

$$\begin{aligned} bias(x) &= \langle \hat{\rho}(x) \rangle - \rho(x) \\ &= \frac{1}{h} \int K\left(\frac{x - y}{h}\right) \rho(y) dy - \rho(x) \end{aligned}$$

- Transformation of variables: $y = x - ht$ and $\int K(t)dt = 1$:

$$\begin{aligned} bias(x) &= \int K(t)\rho(x - ht)dt - \rho(x) \\ &= \int K(t)(\rho(x - ht) - \rho(x))dt \end{aligned}$$

- Taylor evolution:

$$\rho(x - ht) = \rho(x) - ht\rho'(x) + \frac{1}{2}h^2 t^2 \rho''(x) + \ldots$$

$$\begin{aligned} bias(x) &= -h\rho'(x) \int tK(t)dt + \frac{1}{2}h^2 \rho''(x) \int t^2 K(t)dt + \ldots \\ &= \frac{1}{2}h^2 \rho''(x) k_2 + \mathcal{O}(h^3) \end{aligned}$$

Observation:

- Bias does not depend on N.
- Only on $\rho''(x)$ & h
- Illustrate

Analogous calculation for the variance yields:

$$Var(\hat{\rho}(x)) = \frac{1}{Nh}\rho(x)\int K(t)^2 dt$$

- Variance depends on $\rho(x)$, N and h
  Link to counting processes

Consistent estimator in the limit:

- $h \to 0$
- $Nh \to \infty$
- Ergo: h slower towards 0 than N towards ∞

Optimal core

- 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

$$MISE(\hat{\rho}) = \int MSE(\hat{\rho}(x))dx$$

- Minimization of the MISE with respect to h:

$$MISE = \frac{1}{4}h^4 k_2^2 \int \rho''(x)^2 dx + \frac{1}{Nh}\int K(t)^2 dt \tag{34}$$

  yields:

$$h_{opt} = k_2^{-2/5}\left(\int K(t)^2 dt\right)^2 \left(\int \rho''(x)dx\right)^{1/5} N^{-1/5} \tag{35}$$

  - Optimally h has to scale with $h \propto N^{-1/5}$,
  - Prefactor sadly contains curvature of the true density.

- Introduction of Eq. (35) in Eq. (34) yields:

$$MISE = \frac{5}{4} C(K) \left(\int \rho''(x) dx \right)^{1/5} N^{-4/5}$$

with

$$C(K) = k_2^{2/5} \left(\int K(t)^2 dt \right)^{4/5}$$

Under assumption from above for $K(t)$ this is minimized Epanechnikow core

$$K_{Ep}(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} t^2\right) & \text{if } -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{sonst} \end{cases}$$

- Efficiency of core K :

$$Eff(K) = C(K_{Ep})/C(K)$$

Core	K(t)	Efficiency
Triangle	$1 - \|t\|$ für $\|t\| < 1$	0.986
Gaussian	trivial	0.951
Rectangle	$1/2$ für $\|t\| < 1$	0.930

Conclusion:

- Rectangle is bad
- Gaussian does not have a finite carrier, also bad
- Triangle is o.k.

Choice of h: Cross-validation

- Idea:
  - Assuming one has one additional observation x_{N+1}

  - Then the log likelihood would be: $\mathcal{L}(h) = \log \hat{\rho}_h(x_{N+1})$, and it could be maximized with respect to h.
  - Sadly this is not available so:

- Define the "leave-one-out" estimator:

$$\hat{\rho}_h^{-i}(x_i) := \frac{1}{(N-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right)$$

  and the cross-validation function $CV(h)$:

$$CV(h) := \frac{1}{N} \sum_{i=1}^{N} \log \hat{\rho}_h^{-i}(x_i)$$

- Determine "optimal" h trough maxed $CV(h)$.

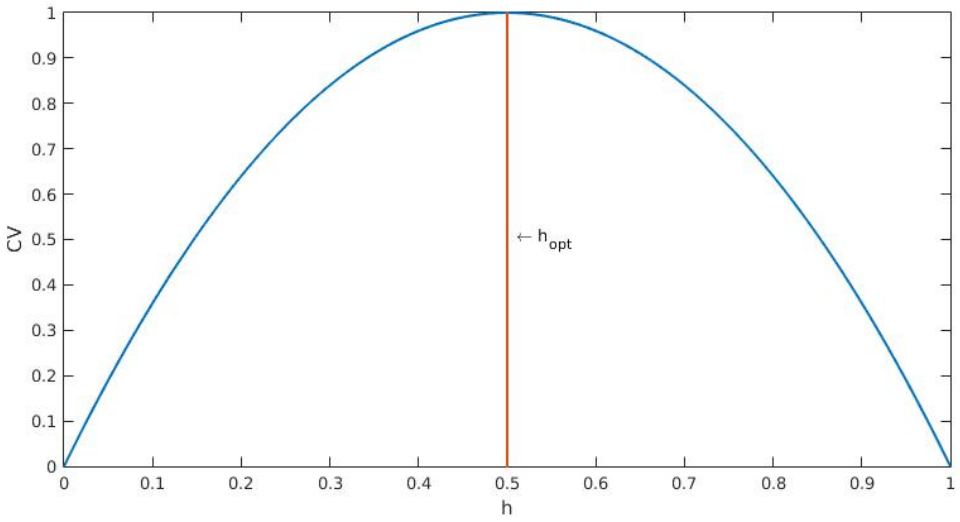

Figure 12.6: Desired behavior of $CV(h)$

- There are many other heuristic ideas and they all have their problems.

12.2 Non-Parametric Regression

Literature:

- W. Härdle. *Applied Nonparametric Regression* [23]

The setting:

- Task:

  Given N realizations of the model

  $$y = m(x) + \epsilon, \qquad ''m()'', \text{ because this is the } \textit{mean} \text{ of } y \text{ ist}, \quad \epsilon \sim N(0, \sigma^2)$$

  Estimate $m(x)$ non-parametric, i.e. without assumption of a parameterized model like in Chap. 6, based on measurements (y_i, x_i).

- Ansatz, once again core estimator:

  $$\hat{m}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i$$

  or Nadaraya-Watson core estimator:

  $$\hat{m}(x) = \frac{\sum_{i=1}^{N} K((x - x_i)/h)}{\sum_{i=1}^{N} K((x - x_i)/h)} y_i$$

  due to normalization.

- With

  $$\begin{aligned} c_K &= \int K^2(u)du \\ d_K &= \int u^2 K^2(u)du \end{aligned}$$

  it holds for Mean Square Error analogous to above:

  $$MSE(x) = \frac{\sigma^2 c_K}{Nh} + h^4 d_K^2 \frac{m''(x)^2}{4}$$

- Same as with the density core estimator:

  Width of the core controls trade-off between bias and variance

- Once again consistent estimator for $h \to 0$, $Nh \to \infty$
- For density core estimator positive core were natural, this is not not necessary anymore. See below.
- For the choice of h, see exercise.

Equivalence of Core Estimator and Local Non-Linear Regression

- Consider square core[3]:

$$K(u) = \begin{cases} \frac{1}{2h} & \text{if } |u| < h \\ 0 & \text{else} \end{cases}$$

  Consider for fixed x:

$$\frac{1}{N} \min_{a,b} \sum_{i=1}^{N} K(x - x_i)(y_i - a - b(x - x_i)^2)^2$$

  the local parabola fit to the interval determined by the uniform core.

  Result will be:

$$\hat{m}(x) = \hat{a}$$

- The normal equations (remember Chap. 10.2) are

$$\frac{\partial}{\partial a}: \qquad \frac{1}{N} \sum_i K(x - x_i)(y_i - \hat{a} - \hat{b}(x - x_i)^2) = 0$$

$$\frac{\partial}{\partial b}: \qquad \frac{1}{N} \sum_i K(x - x_i)(y_i - \hat{a} - \hat{b}(x - x_i)^2)(x_i - x)^2 = 0$$

- Define

$$\tilde{y}(x) := \sum_i K(x - x_i) y_i$$

[3]Unusual definition to avoid constantly dividing by h

Assume x_i is equal distributed and consider:

$$\frac{1}{N}\sum_i K(x-x_i) \approx 1$$

- Approximate

$$\frac{1}{N}\sum_i K(x-x_i)(x-x_i)^2 \approx \int_{-\infty}^{\infty} K(x-u)(x-u)^2 du = \int_{-1/2h}^{1/2h} (x-u)^2 du = h^3/3$$

Analog

$$\frac{1}{N}\sum_i K(x-x_i)(x-x_i)^4 \approx \int_{-\infty}^{\infty} K(x-u)(x-u)^4 du = \int_{-1/2h}^{1/2h} (x-u)^4 du = h^5/5$$

- With

$$A = \frac{1}{N}\sum_i K(x-x_i)(x-x_i)^2 y_i$$

The normal equations are thus:

$$\begin{aligned} 0 &= \tilde{y} - \hat{a} - \frac{h^3}{3}\hat{b} \\ 0 &= A - \frac{h^3}{3}\hat{a} - \frac{h^5}{5}\hat{b} \end{aligned}$$

Leads for $\hat{a}$ to:

$$0 = 3h^2\tilde{y} - 5A + \frac{4}{3}h^2\hat{a}$$

- Introducing everything:

$$\hat{a} = \frac{3}{4N}\sum_i K(x-x_i)\left(3 - 5\left(\frac{(x-x_i}{h}\right)^2\right) y_i$$

- Sharp observation shows:

$$\hat{m}(x) = \hat{a} = \frac{1}{N} \sum_i K^*(x - x_i) y_i$$

with

$$K^*(u) = \begin{cases} 3/8(3 - 5(u/h)^2 & \text{if } |u| < h \\ 0 & \text{else} \end{cases}$$

a parabolic core.

- On the other hand: Parabolic core corresponds to local parabola fit

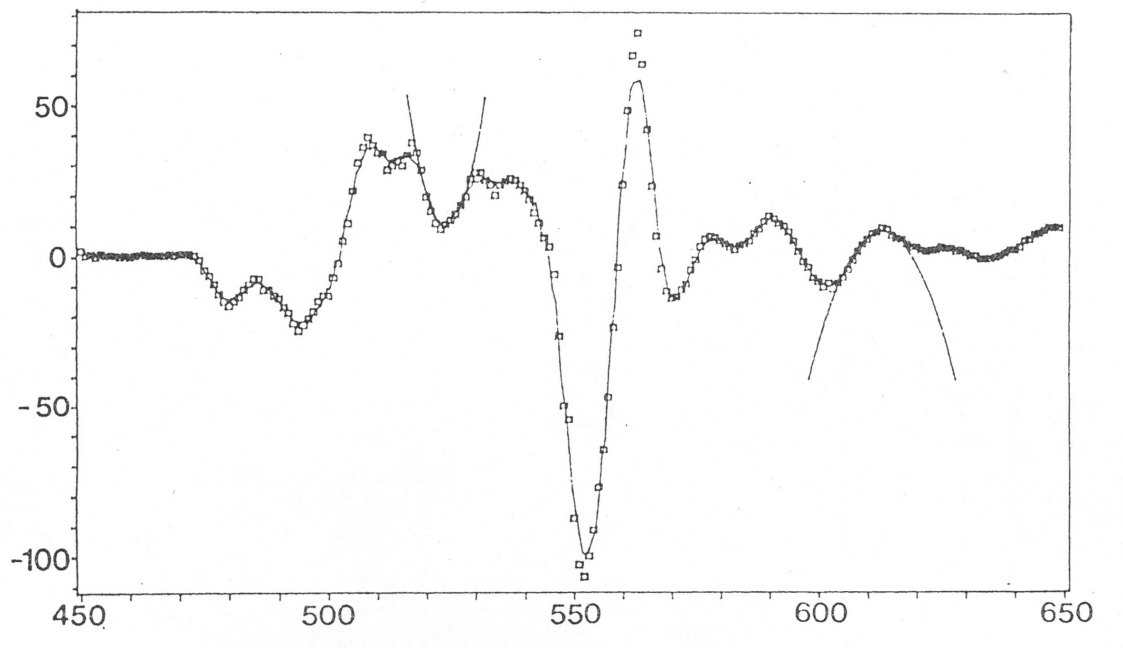

Figure 12.7: Local parbola fits in comparison to the core estimator

- For different core with higher orders
- Remember: Non-parametric regression = Parametric with many parameters

Savitzky-Golay - Filter

- Idea: Turning it around

  Determination of core coefficients from polynomial fit

  Let the data be equidistant, $\Delta x = 1$.

$$\hat{m}(x_i) = \sum_{j=-h}^{h} c_j y_{i+j}$$

- Choose c_j so that it corresponds to a polynomial fit with

$$y_i = a_0 + a_1 i + a_2 i^2 + \ldots a_M i^M$$

  to the data $y = (y_{i-h}, \ldots, y_{i+h})$ Then analogous to above:

$$\hat{m}(x_i) = \hat{a}_0$$

- Remember Chap. 6 Non-linear regression

  The design matrix A is:

$$A_{il} = i^l$$

  and the normal equations lead to:

$$A^T A a = A^T y \text{ oder } a = (A^T A)^{-1} A^T y$$

  In practice: Coefficients a are linear in the data.

- Therefor c_j is a_0, if y is replaced by unity vectors e_j:

$$c_j = \{(A^T A)^{-1} A^T e_j\}_0 = \sum_{m=0}^{M} \left\{(A^T A)^{-1}\right\}_{0m} j^m$$

  For M=2, h=2, the coefficients are:
  $-.0086, 0.343, 0.486, 0.343, -.0086$ and are not positive.

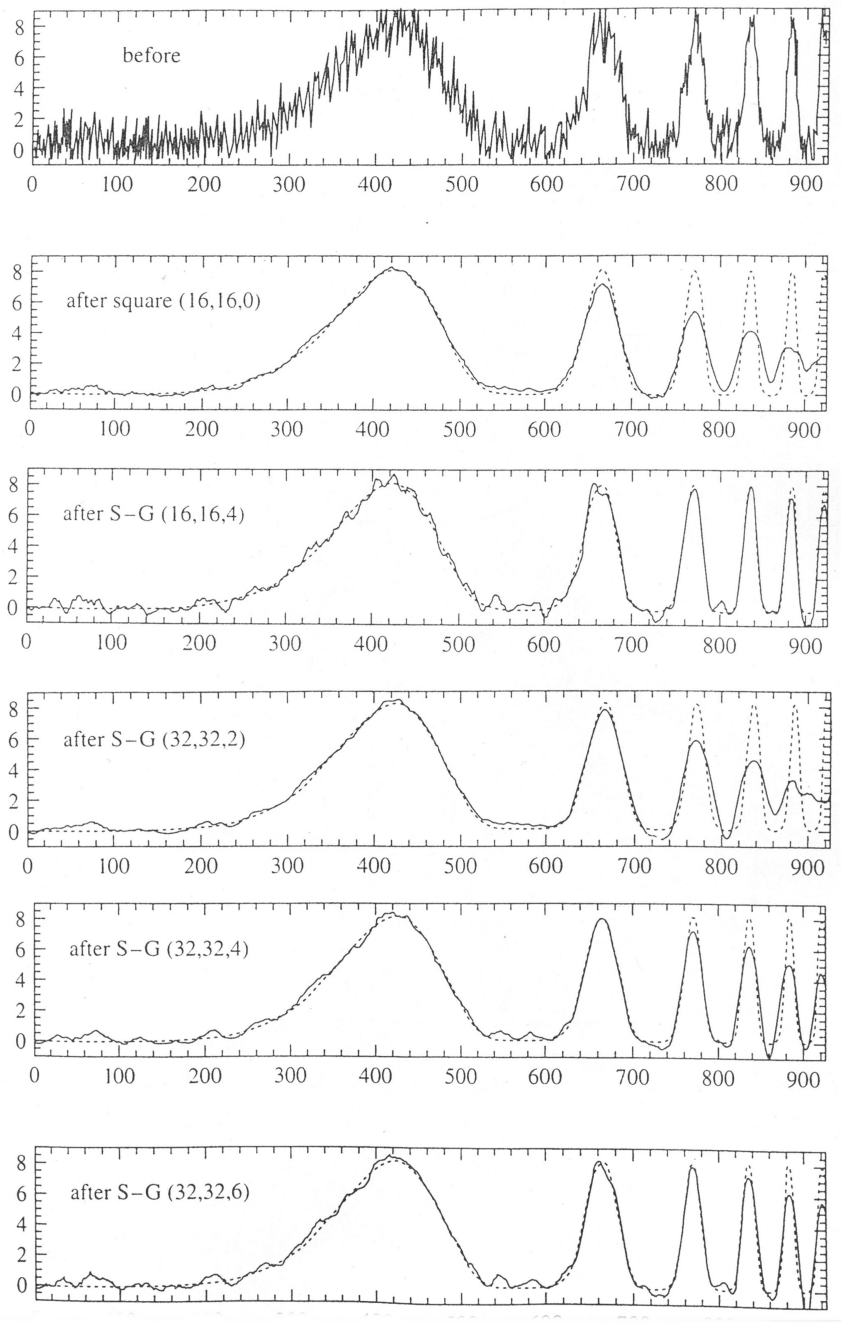

Figure 12.8: (a) Noisy data; (b) Fitted without S-G filter, 16 points left and right; (c) with S-G filter of grade 4, 16 points left and right; (d) with S-G filter of grade 2, 32 points left and right; (e) with S-G filter of grade 2, 32 points left and right; (f) with S-G filter of grade 6, 32 points left and right.

Estimation of derivatives
Repeating the same process for a_1 yields an estimation for the first derivative and so on.

Spline smoothing

- Adapting a function $g(x)$ with many degrees of freedom, for example higher order polynomial, using least squares

$$a = argmin \sum_i (y_i - g(x_i, a))^2$$

  to the data y_i, then $g(x, a)$ will interpolate the data and will be very variable locally.

- Idea:

  Require a certain smoothness of $g(x, a)$.

  Smoothness can be estimated via:

$$\int (g''(x, a))^2 dx$$

- Remember regularization Chap. 4.3 and define

$$S_\lambda(g) = \sum_{i=1}^{N} (y_i - g(x_i, a))^2 + \lambda \int (g''(x, a))^2 dx$$

- Consider
  - $\lambda = 0$: Interpolation
  - $\lambda = \infty$: Linear regression
- Minimization of $S_\lambda(g, a)$ over all double differentiable functions has an exact solution:

  $\hat{m}_\lambda(x)$ is:
  - Cubic polynomial between consecutive x_i values.
  - Continuous at the x_i values.

- First and second derivative continuous, third derivative not continuous.
- Second derivative = 0 at x_1 and x_N.

- Is called Spline: "a slat of wood, metal, etc" (Oxford dictionary)
  If bent, this is very smooth.
- If the error on the data is know, λ can be fixed.
- Can be formulated as core estimator (not pretty).

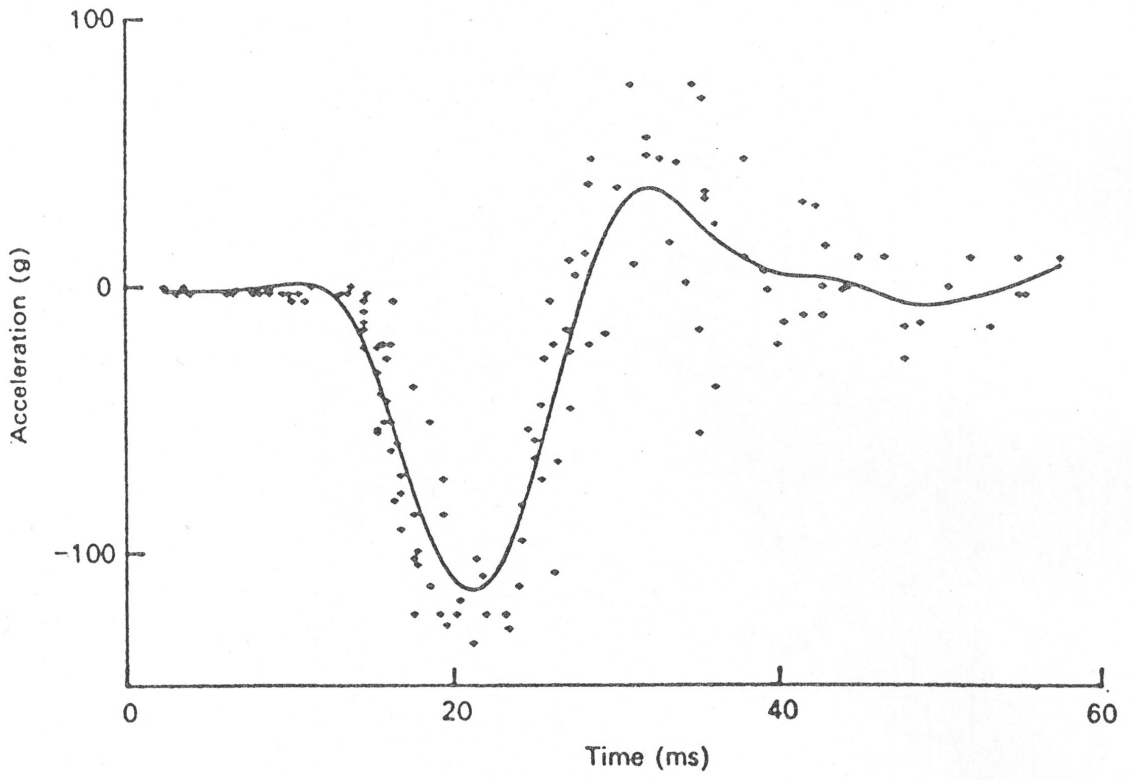

Figure 12.9: Spline smoothing of a data set

Robust smoothing
If the errors are not Gaussian, the Median filter:

$$\hat{m}_M(x) = \text{med}\{y_i\}, \quad \{y_i | x_i \in [x-h, x+h]\}$$

can be of use, for example in noise suppression in black and whit pictures.

The curse of high dimensions
Distributing N points equidistantly in d-dimensional unitcubes $[0,1]^D$, the distance $dist_{NN}$ between two points is:

$$dist_{NN} = N^{-1/D}$$

Example $N = 10000$:

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 dimensions every point has 2.5 neighbors in every directions, so realistically none.

Exercise: Crossvalidation

12. week

Lessons learned:

- Non-parametric density estimation: Core estimator and nearest neighbor estimator
- Bias and variance of the estimators
- Optimal core and optimal h
- Non-parametric regression = parametric with many parameters
- Savitzky-Golay - Filter & Splines

13 Spectral analysis

Literature:

- M.B. Priestley *Spectral analysis and time series* [51].
  The mathematical classic No.1
- P.J. Brockwell, R.A. Davis *Time Series: Theory and Methods* [9].
  The mathematical classic No.2
- J. Honerkamp *Stochastic Dynamical Systems* [25] Chap. 13.3
  Condensed version for physicists

Definition Auto-covariance function (ACF):
Let $x(t)$ be a stationary process with $\langle x(t) \rangle = 0$, then the auto-covariance function is:

$$ACF(\tau) = \langle x(t)x(t+\tau) \rangle$$

Definition spectrum:

$$S(\omega) = \int e^{-i\omega\tau} ACF(\tau) = \langle |f(\omega)|^2 \rangle$$

with

$$f(\omega) = \int e^{-i\omega t} x(t)$$

The Fourier transformation orthogonal (all eigenvalues= 1), meaning:

$$\int S(\omega) d\omega = \mathrm{Var}(x(t))$$

Spectrum is "variance per frequency" representation of the process.

Time-discrete process:
Consider:

$$x(i) = ax(i-1) + \sigma\epsilon(i), \qquad 0 < a < 1, \quad \epsilon(i) \sim N(0,1)$$

- If $\sigma = 0$

$$x(i) = x(0)e^{-i/\tau}$$

  a relaxator with $\tau = -1/\log a$

- $\sigma \neq 0$: Process will constantly be brought out of equilibrium around 0 due to noise

  Physically: Stochastic driven relaxator

- Process is called Auto-regressive process of order 1, AR[1].

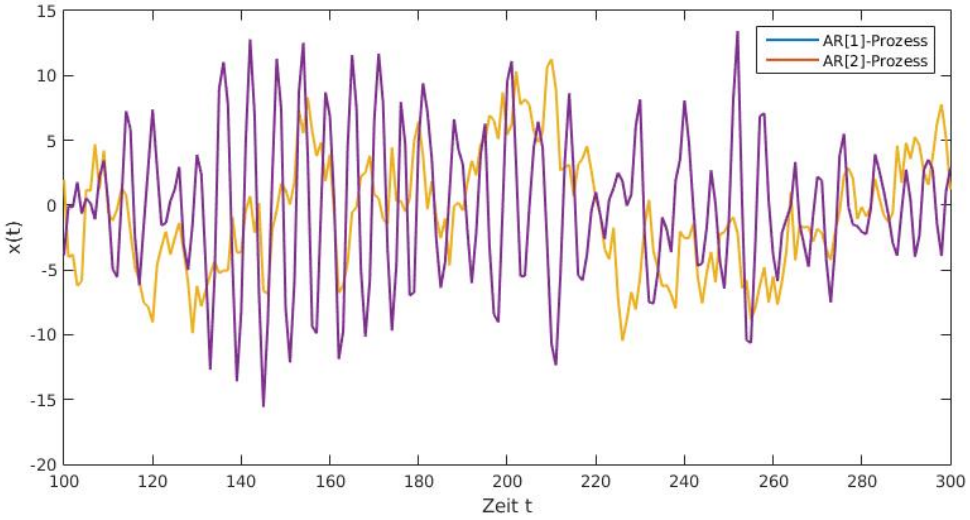

Figure 13.1: Realizations of linear stochastic processes of orders 1 and 2.

- AR[2] process:

$$x(i) = a_1 x(i-1) + a_2 x(i-2) + \epsilon(i)$$

  yields with:

$$\begin{aligned} a_1 &= 2\cos(2\pi/T)e^{-1/\tau} \\ a_2 &= -e^{-2/\tau} \end{aligned}$$

  a stochastic driven damped oscillator with period T and relaxation time τ.

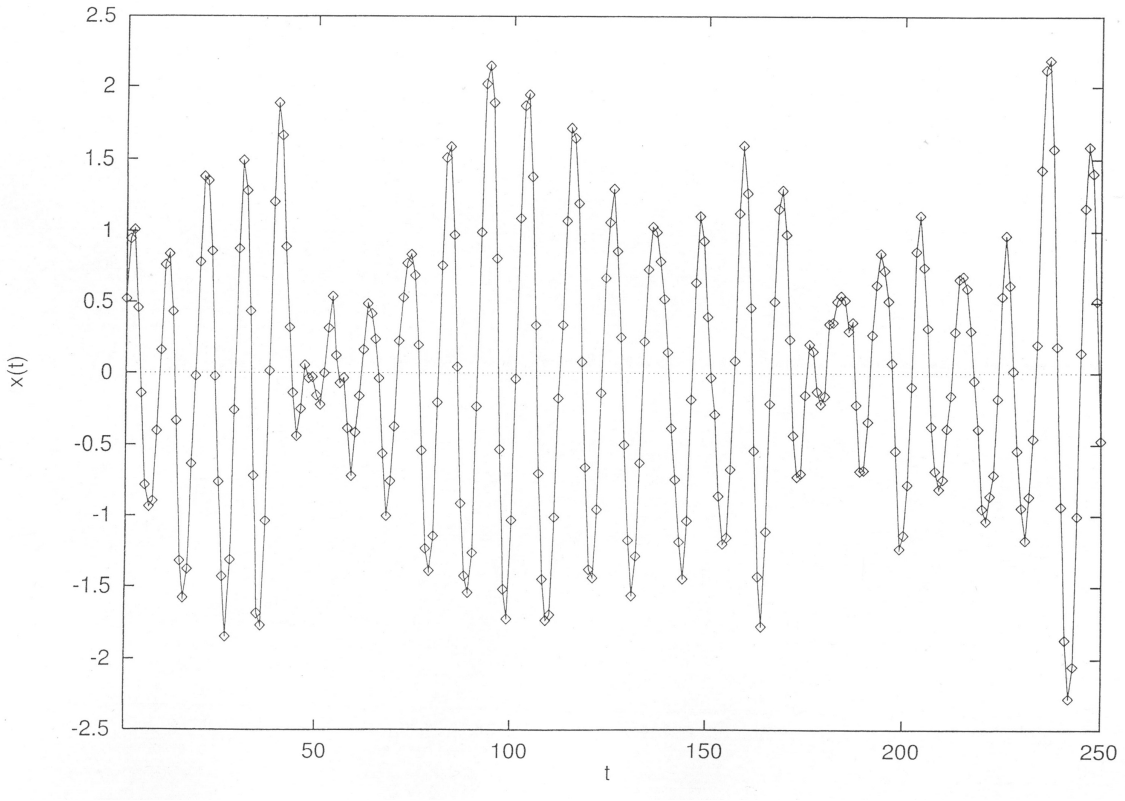

Figure 13.2: Realization of a linear stochastic process of order 2

13.1 Spectra of AR[p] Processes

- Define Backshift-Operator:

$$B(x(t)) = x(t-1)$$

- Let

$$f(\omega) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} e^{-i\omega t} x(t)$$

(Normalization will be left out from now on) then

$$\sum_{t=1}^{N} e^{-i\omega t} B(x(t)) = \sum_{t=1}^{N} e^{-i\omega t} x(t-1) = e^{-i\omega} \sum_{t=1}^{N} e^{-i\omega t} x(t) = e^{-i\omega} f(\omega)$$

In general:

$$\sum_{t=1}^{N} e^{-i\omega t} B^d(x(t)) = e^{-id\omega} f(\omega)$$

- AR[p] process:

$$x(t) - \sum_{j=1}^{p} a_j B^j(x(t)) = \epsilon(t)$$

- Fourier transformation:

$$f(\omega)(1 - \sum_{j=1}^{p} a_j e^{-ij\omega}) = \tilde{\epsilon}$$

- Spectrum:

$$S(\omega) = \langle |f(\omega)|^2 \rangle = \frac{1}{2\pi} \frac{\sigma^2}{|1 - \sum_{j=1}^{p} a_j e^{-ij\omega}|^2}$$

Important:
Spectrum of AR[p]-process is smooth.

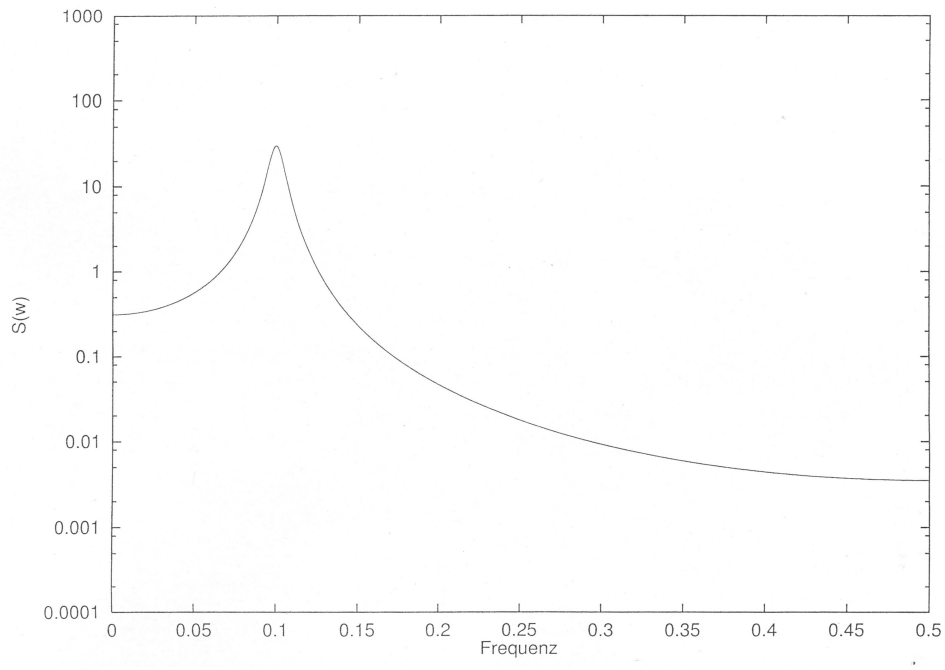

Figure 13.3: Spectrum of a linear stochastic process of order 2

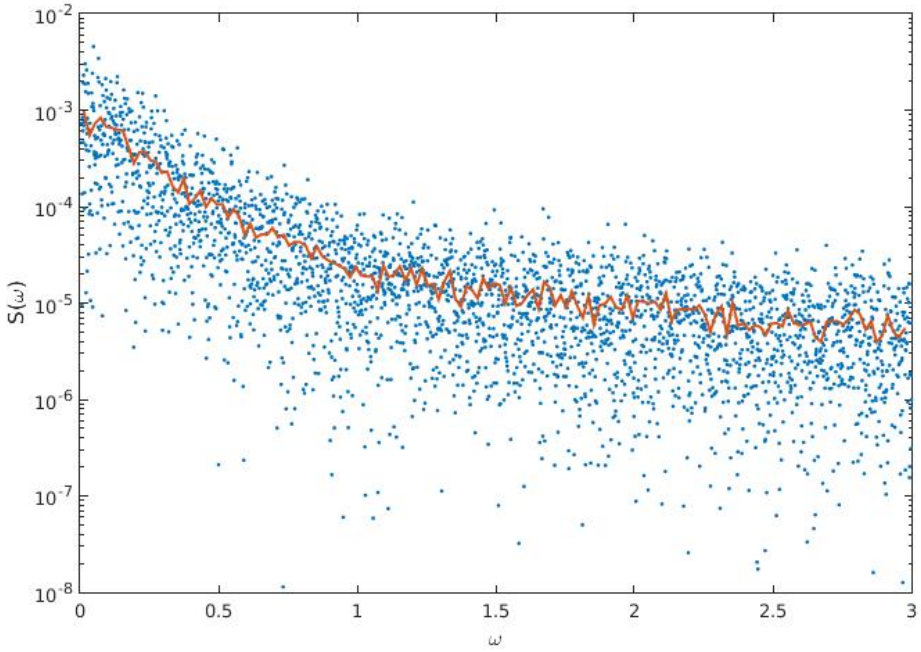

Figure 13.4: Periodogram and estimated spectrum of a linear stochastic process of order 1

Smoothness also holds

- for all non-linear stochastic processes.
- in reality for all chaotic processes.

In general: Always, if the ACF dissociates, i.e. the process is forgetful, mixing.

13.2 Fast Fourier Transform (FFT)

Cooley & Tukey, 1965 [11].

- The calculation of the Fourier transform

$$f(\omega_k) = \sum_{t=0}^{N-1} e^{-i\omega_k t} x(t)$$

for all Fourier frequencies

$$\omega_k = \frac{2\pi k}{N}, \quad k = -N/2 \ldots, 0, \ldots, N/2$$

has complexity $\mathcal{O}(N^2)$.

- For $x(t)$ being real:

$$f(\omega_k) = f^*(-\omega_k)$$

degrees of freedom have to be counted.

- Divide and Conquer - strategy
  Let $N = 2^n$

$$\begin{aligned} f(\omega_k) = f_k &= \sum_{t=0}^{N-1} e^{i\omega_k t} x(t) \\ &= \sum_{t=0}^{N/2-1} e^{-i\omega_k (2t)} x(2t) + \sum_{t=0}^{N/2-1} e^{-i\omega_k (2t+1)} x(2t+1) \\ &= \sum_{t=0}^{N/2-1} e^{-i2\omega_k t} x(2t) + e^{i\omega_k} \sum_{t=0}^{N/2+1} e^{-i2\omega_k t} x(2t+1) \\ &= f_k^e + e^{i\omega_k} f_k^o \quad e \text{ like even} \quad o \text{ like odd} \end{aligned} \tag{36}$$

- f_k^e and f_k^o periodic in k with period $N/2$
- For f_k^e and f_k^o the decomposition can be repeated.
  Yields : f_k^{ee}, f_k^{eo}, f_k^{oe} and f_k^{oo}.
  Effective FT length $N/4$ each, the rest is periodic.
- Iterate this, until length of the Fourier transform = 1.
- But

$$\sum_{t=0}^{0} e^{i\omega t} x(0) = x(0)$$

This means there are representations:

$$f_k^{eooeeeo..oe} = f^{eooeeeo..oe} = x(t) \quad \forall\, t \tag{37}$$

does not depend on k, since periodic in k with period 1.

- Length of the chain *eooeeeo..oe*: $\log_2 N$
- Now main point: Bitreversal:
  - Which sequence of e's and o's belong to which t
  - Turning the order of the e's and o's around
  - Replace the sequence *eo..oeeeooe* with $e = 0$ and $o = 1$
  - Gives the binary representation for every t.
  - even/odd decomposition bitreversed constructs binary representation from the bottom up
  - Example 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	00

  - Example 8:

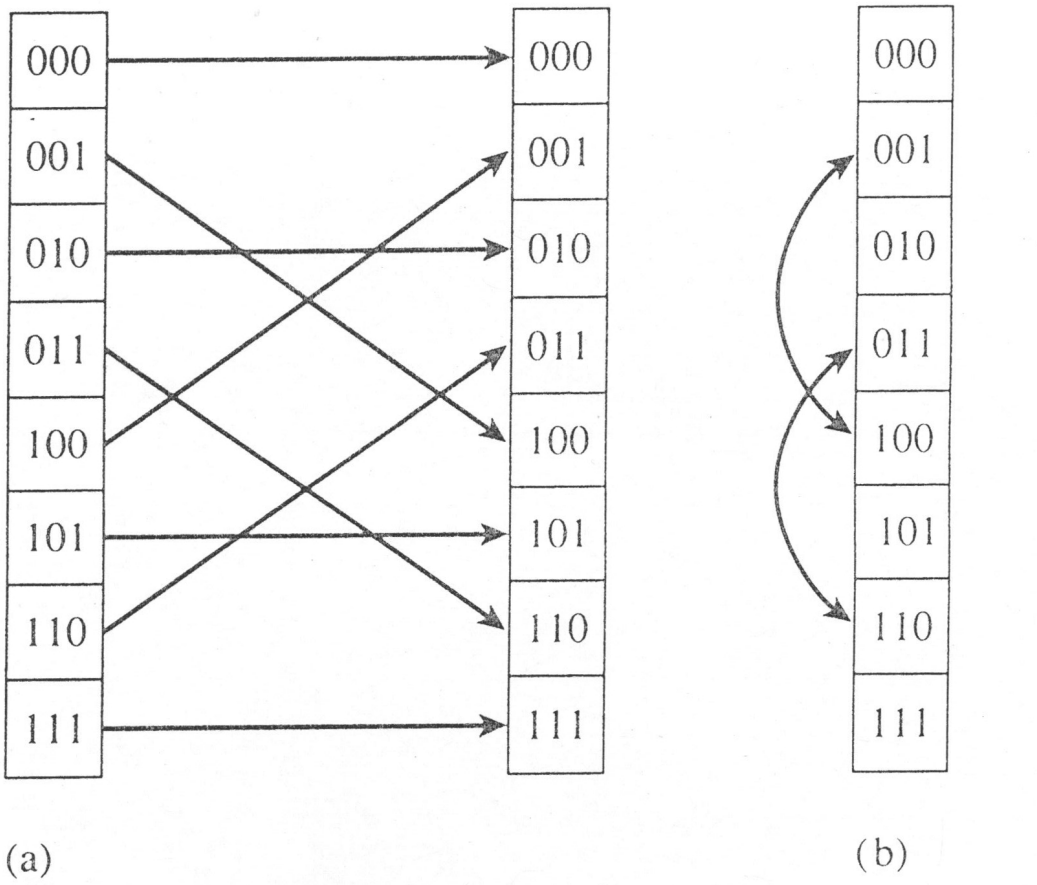

Figure 13.5: Rearrangement of an array by bitreversal, (a) between two arrays and (b) in one array

  - Consider: Only needs pairwise switches

- Starting point for the inversion of Eq. (37) using $\log N$ applications of Eq. (36).
- Accounting quite simple:
  - Sort data in bitreversed order: Single point transformation
  - Combine neighboring points
    - Two point transformation

    - Example $N = 64$, $n = 6$

$$f_k^{eooee} = f^{eooeee} + e^{i\omega_k} f^{eooeeo}$$

    - Result needs 2 points space
    - Complexity: $\mathcal{O}(N)$
  - Combine neighboring point pairs
    - 4 point transformation
    - Needs 4 points space
    - Complexity: $\mathcal{O}(N)$
  - Combine neighboring quadruples
  - and so on

- Starting from N data points $\log N$ times Eq. (36), yields effort $\mathcal{O}(N \log N)$
- FFTs also exist for $N = 2^n\, 3^k\, 5^l$
- The same divide and conquer approach is also applicable in other situations.

13.3 Spectral Analysis of Time-Discrete Processes

An estimator $\hat{\Theta}_N$ based on N data points is called consistent, if it holds:

$$\lim_{N\to\infty} (\hat{\Theta}_N - \Theta) \to_{prob} 0$$

i.e. bias and variance run with N towards 0.

Periodogramm of white noise

- Let $x(t) = \epsilon(t) \sim N(0, \sigma^2)$
- Then $f(\omega_k)$ is

$$f(\omega_k) = \frac{1}{N} \sum_{t=1}^{N} e^{-i\omega_k t} \epsilon(t)$$

- With

$$|e^{-i\omega_k t}| = 1 \text{ and } \langle \epsilon(t_i)\epsilon(t_j)\rangle = \delta_{ij}$$

follows:

$$f(\omega_k) \sim N_C(0, \sigma^2)$$

  - Independent of ω_k (hence "white" noise)
  - With independent real and imaginary portions (because $\sin(\omega_k t)$ and $\cos(\omega_k t)$ are orthogonal)
  -

$$\langle f(\omega_k), f(\omega_l)\rangle = \sigma^2 \delta_{kl}$$

$f(\omega_k)$ independently complex normal distributed.

- Spectrum was

$$S(\omega) = \langle |f(\omega)|^2 \rangle$$

- $|f(\omega)|^2$ has special name: Periodogram

$$Per(\omega_k) = |f(\omega)|^2$$

- Since

$$Per(\omega_k) = |f(\omega_k)|^2 = (\mathrm{Re}(f(\omega_k)))^2 + (\mathrm{Im}(f(\omega_k)))^2$$

it holds for $x(t) = \epsilon(t)$:

$$Per(\omega_k) \sim \chi_2^2$$

For non-white (in general nonlinear) processes the central limit theorem is of help, and it holds in general (with correct prefactors):

$$Per(\omega_k) \sim \frac{1}{2} S(\omega_k)\chi_2^2, \qquad \omega_k \neq 0, \pi$$

independent of N. (For $\omega_k = 0, \pi$: $Per(\omega_k) \sim S(\omega_k)\chi_1^2$, since only $\cos(\omega_k t)$ contributes).

- Since
$$\langle \chi^2_2 \rangle = 2, \qquad Var(\chi^2_2) = 4 \qquad SD(\chi^2_2) = 2$$
the periodogram is an unbiased estimator,
- But : Standard deviation of the periodogram is independent from N (and equal to the expectation value)

Thus the periodogram is not a consistent estimator for the spectrum!

- Increasing amounts of data:

  Instead of smaller variances for the estimator one obtains better resolution in the frequency space.

Central:
Because the (true) spectrum is smooth, spectra can be estimated, by smoothing the periodogram:

$$\hat{S}(\omega_k) = \sum_{l=-h}^{h} W_l Per(\omega_{k+l})$$

This yields with $N \to \infty$ $h \to \infty$, and $h/N \to 0$ a consistent estimator.

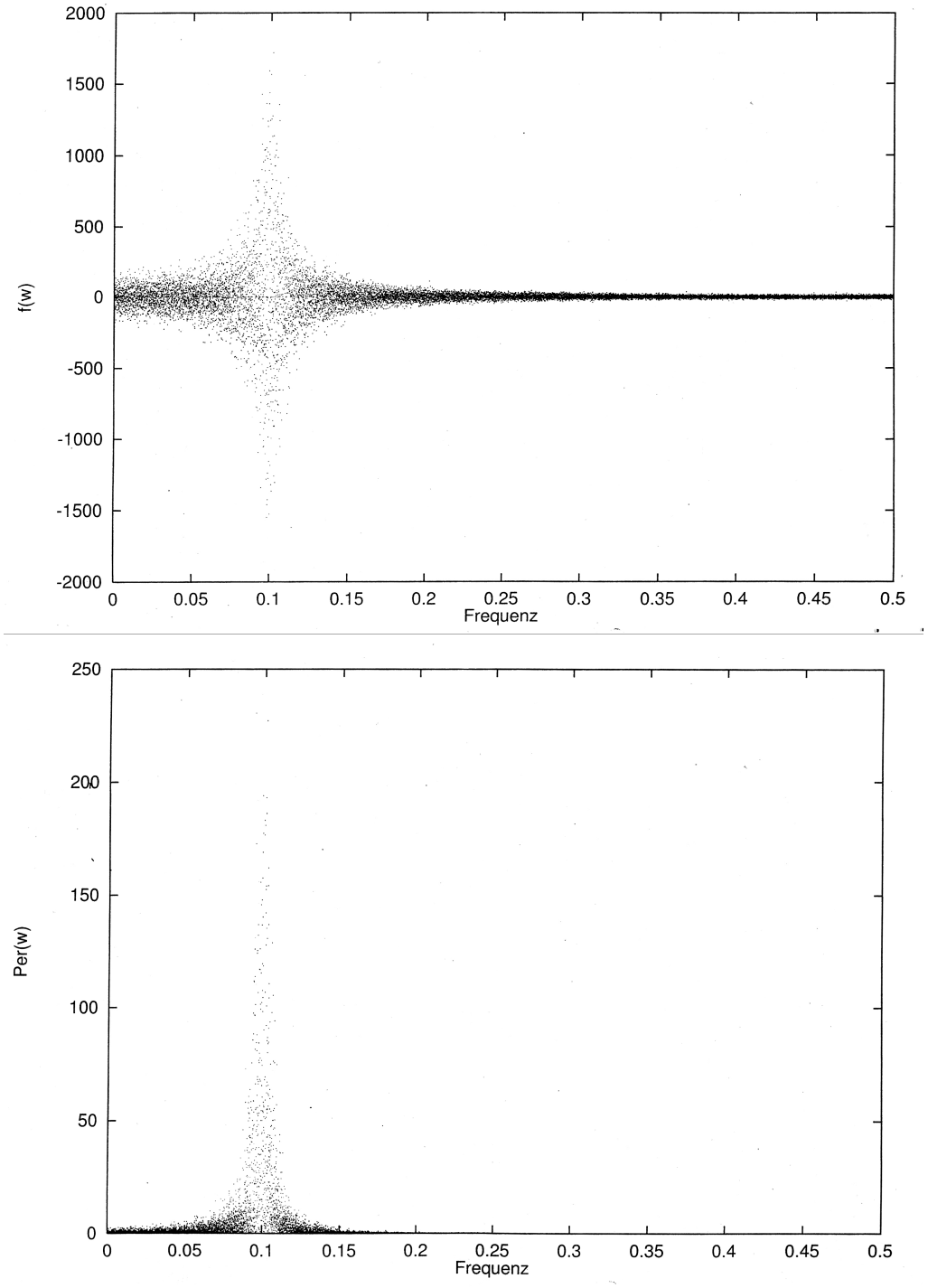

Figure 13.6: Linear stochastic process of order 2

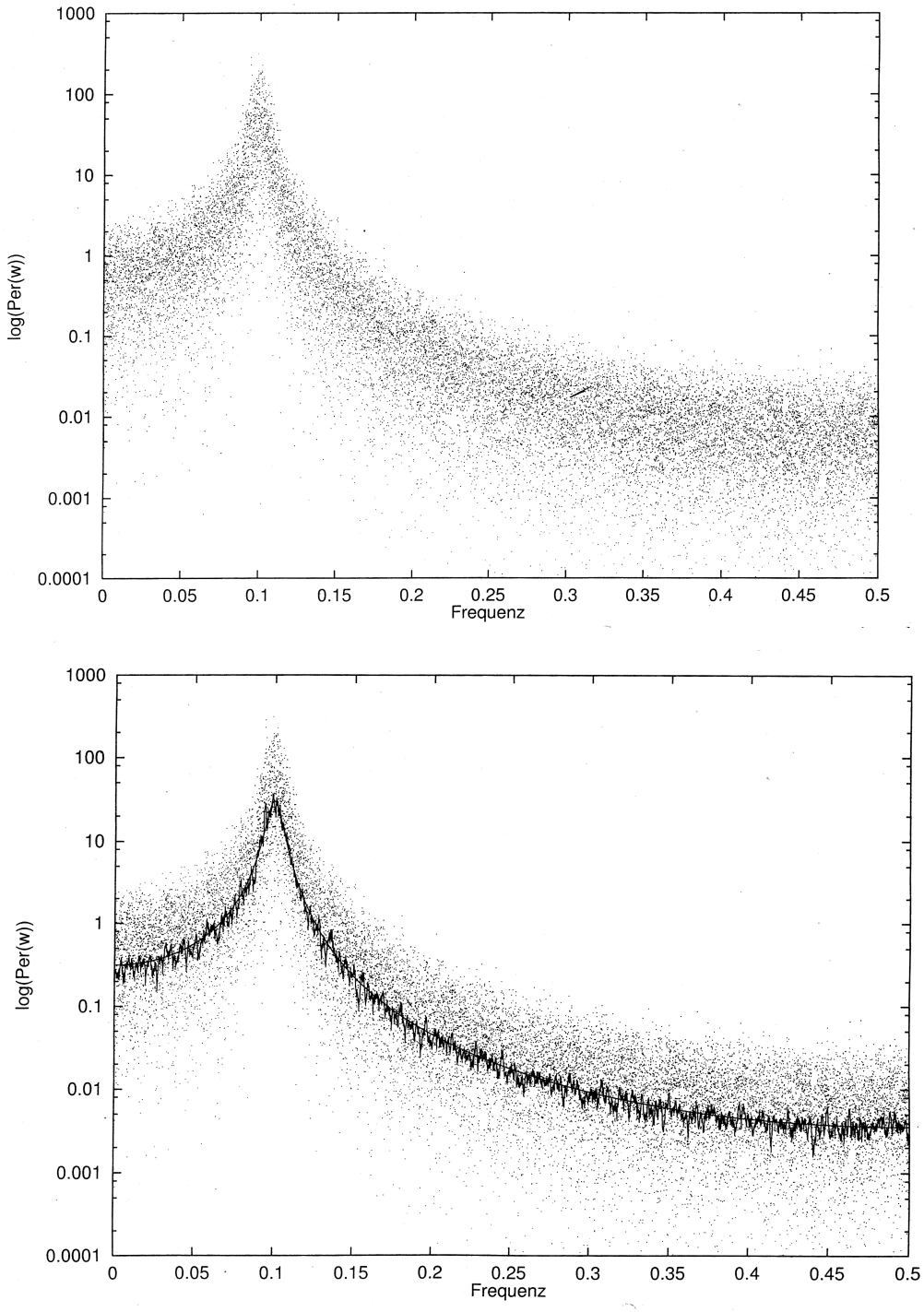

Figure 13.7: Linear stochastic process of order 2

Different methods:

- Cut time course in L pieces and take the mean of their periodograms

  Let $M = N/L$

$$Per_l(\omega_k) = |\sum_{t=1}^{M} e^{-i\omega_k t} x((l-1)M + t)|^2$$

$$\hat{S}(\omega_k) = \frac{1}{L} \sum_{l=1}^{L} Per_l(\omega_k)$$

- ACF windows, Remember QM: Folding in frequency space is multiplication in time space and vice versa.

$$\hat{S}(\omega_k) = \sum_{\tau=1}^{N} w(\tau) e^{-i\omega_k \tau} ACF(\tau)$$

  $w(\tau) = 0$ für $\tau > \tau_{max}$. $\tau_{max} \propto 1/h$.
  Method of choice before the invention of FFt.

In case of

- linear processes the Fourier components stay independent
- non-linear processes correlations will arise.
- See next semester for details

Comparison Fourier series vs. Fourier transformation (FT) (stochastic process)
For example for saw tooth:

$$y = x \text{ for } -\pi < x < \pi, \quad \text{and periodically continued}$$

holds:

$$y = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \dots\right)$$

The "periodogram (=spectrum)" is thus:

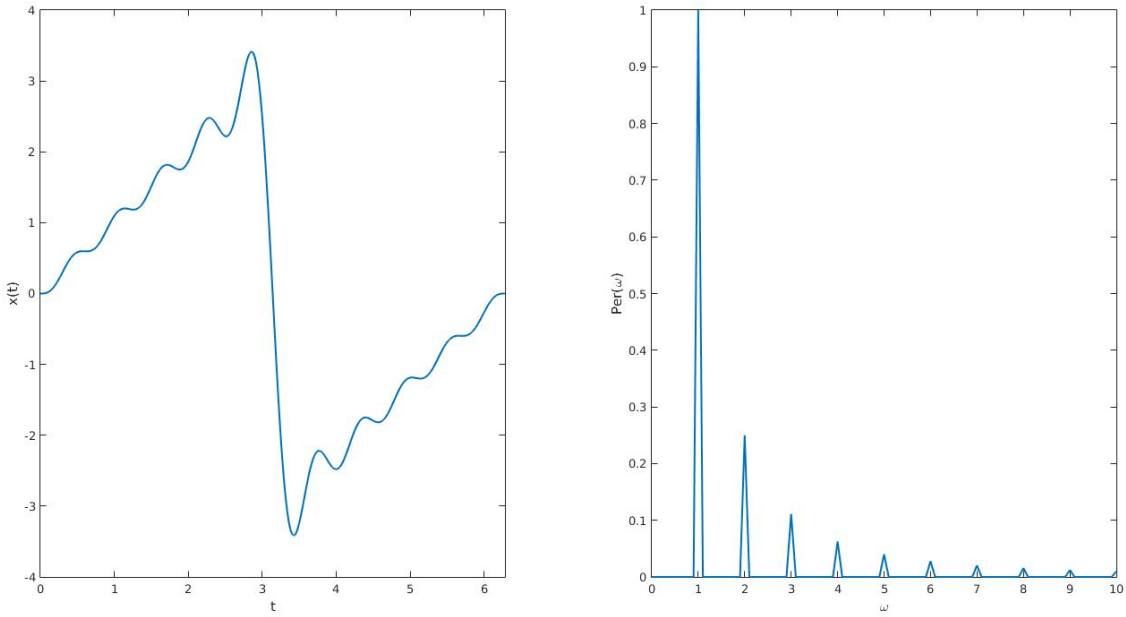

Figure 13.8: Periodogram of a saw tooth curve

Consider van der Pol oscillator:

$$\ddot{x} = \mu(1 - x^2)\dot{x} - x$$

Cubic non-linearity, perturbation theory, higher harmonics for (2i+1) fold of the fundamental frequency.

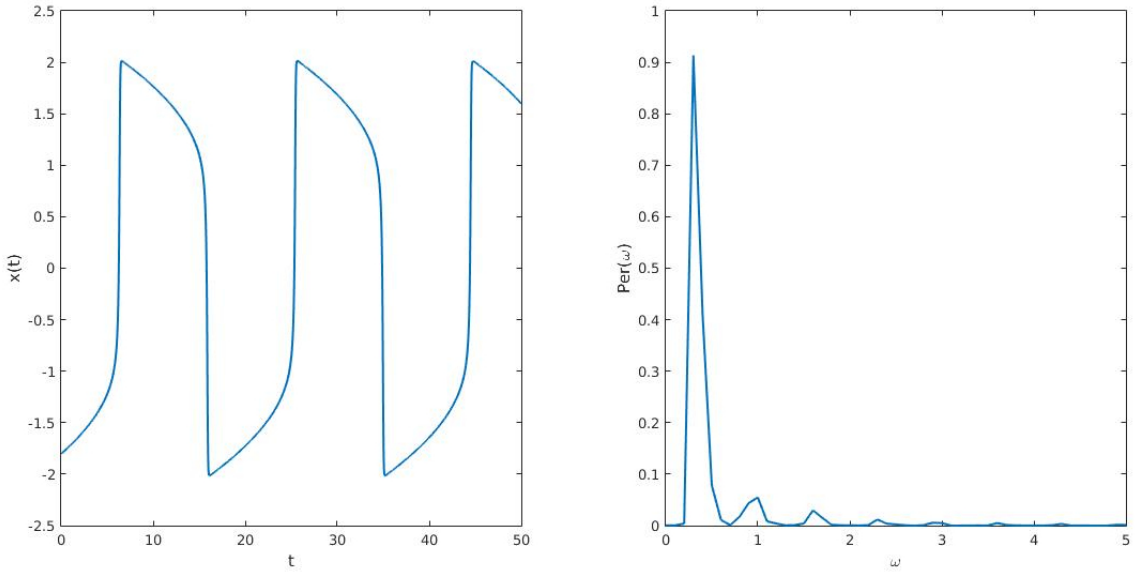

Figure 13.9: Periodogram of the van-der-Pol oscillator

Leakage and Tapern:
The FT sees the tie course $x(t), t = 1, \ldots T$ as a segment of an infinitely long series $y(t), t \in Z$:

$$x(t) = w_u(t)y(t), \quad w_u(t) = \begin{cases} 1 & \text{if } 1 \leq t \leq T \\ 0 & \text{else} \end{cases}$$

Effect Leakage:

- Multiplication in time space is folding in frequency space
- Spectral estimation "blurred"
- Mass is transported from peaks to valleys.
- Is worst for $w_u(t)$.

Treatment:

- Choose $w(t)$, with softer time course, for example Bartlett window

$$w_B(t) = \begin{cases} 1 - \left| \frac{t - \frac{1}{2}T}{\frac{1}{2}T} \right| & \text{if } 1 \leq t \leq T \\ 0 & \text{else} \end{cases}$$

- This is called Tapern.

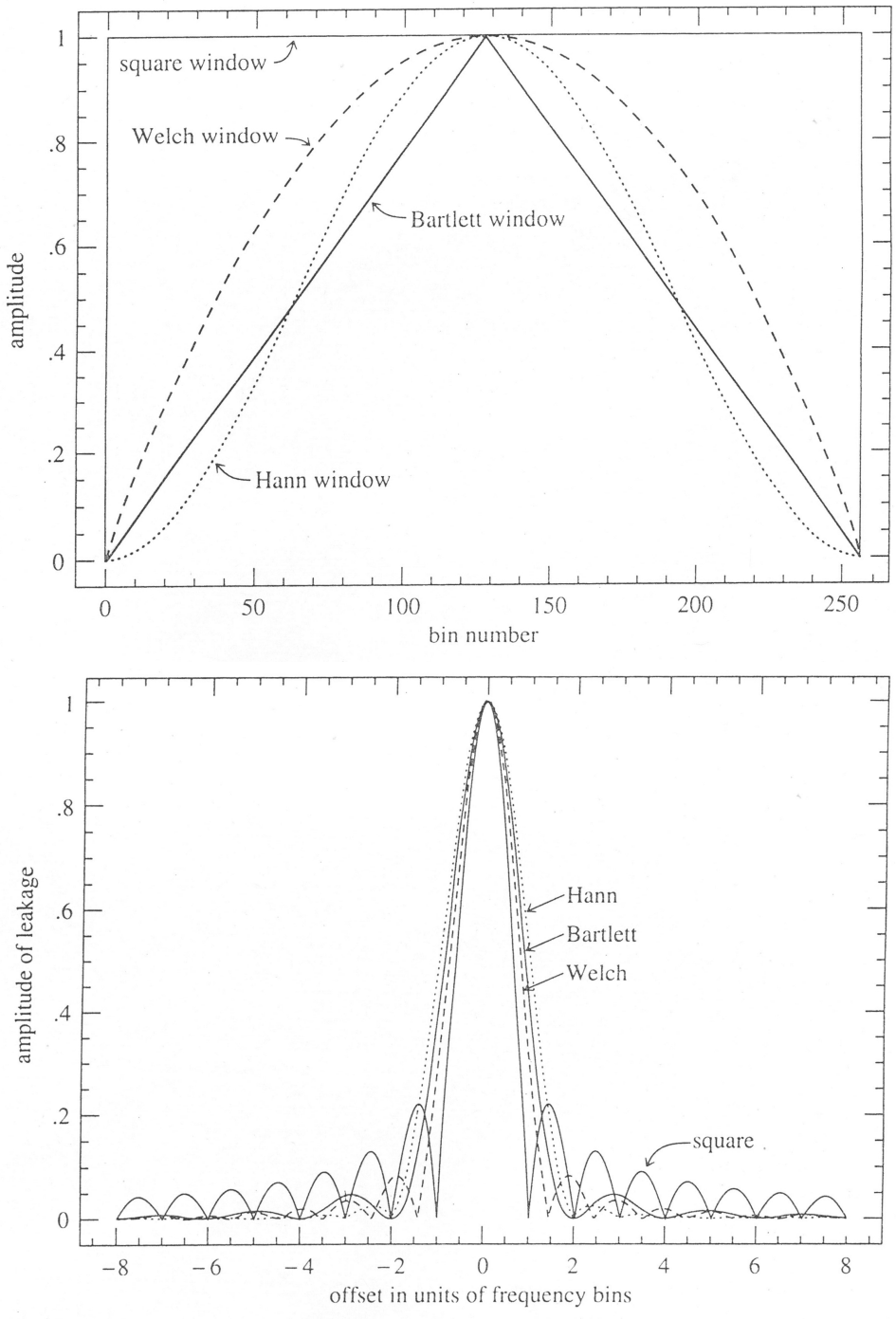
square window
Welch window
Bartlett window
Hann window
amplitude
bin number
0
50
100
150
200
250
.2
.4
.6
.8
1
amplitude of leakage
offset in units of frequency bins
Hann
Bartlett
Welch
square
−8
−6
−4
−2
2
4
6
8

Figure 13.10

The resulting spectrum is to be normalized with

$$g = \frac{T}{\sum_{t=1}^{T} w^2(t)}.$$

Exercise:
Simulation and spectral estimation for AR[2] process

Lessons learned:

- Fast-Fourier Transformation
- χ^2_2-distribution of the periodogram of mixing processes
- Consistent Estimator for the spectrum

14 Markov Chain Monte Carlo Procedure

Literature:

- 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 discussion [29]

Bayesian Ansatz (biased version):

- There are no "true" parameters.
- Parameters are random variables.
- Every probability is a conditional probability.
- Prior knowledge is the condition.

Bayes theorem:
From

$$p(a,b) = p(a|b)p(b) = p(b|a)p(a)$$

follows

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)}$$

allows "shoveling" of $p(b|a)$ to $p(a|b)$.
Let b be the parameters, a be the data, then the MLE idea was: Reading $p(a|b)$ as a function of b.

But for Bayesians $p(b|a)$ makes sense. $p(b)$ represents the prior knowledge. $p(a)$ is constant and therefor neglected.

$$p(b|a) \propto p(a|b)p(b) = \text{ Likelihood } \times \text{ Prior knowledge}$$

If the error model is Gaussian and the prior knowledge, or prior, infers that the norm of b is rather small, for example:

$$p(b) \propto e^{-\lambda b^2}$$

the taking the logarithm yields:

$$p(b|a) \propto \sum_{i=1}^{N} \frac{(a_i - a(x_i, b))^2}{\sigma_i^2} + \lambda b^2$$

the minimum norm regularization of the SVD from Chap. 7 Solutions of linear equation systems.

Gibbs Sampler
The equation

$$p(b|a) \propto p(a|b)p(b)$$

gives the possibility, to estimate the parameters of a model in a Baysian context.
Problem: The high dimensional integrals.

Gleichung

Way out: The Gibbs sampler
It can be shown: Pulling single parameters works.

ZEICHNUNG Schema

Convergence: Let 2 processes run in parallel, if Intravariance = Intervariance, then it converges.

Choice of the prior

- If the prior does not change the type of the distribution class of the Likelihood it is called a conjugate prior. This makes a lot of things easier.
- A prior with a very broad distribution is called uninformative.
- In the case of an uninformative prior, the whole thing is MLE and only a matter of integration technique.

15 Classification

Literature:

- 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 Discriminant Analysis
Mahalanobis distance

Clustering

Kohonen map
Optimization of the trans information [41]
MDS and projection pursuit
Literatur:

- J.W. Sammon *A nonlinear mapping for data structure analysis* [60]
- P.J. Huber *Projection Pursuit* [27]

Exercise:
Given a high dimensional data set, look for structures.
See also:
ISOMAP [69]
LLE [58]
What is missing
On essentials :

- Integration calculation, Recipes Chap. 4 and 7.6, Stoer Chap. 3
- Stochastic approximation [30, 56], The great flood, Thresholding

References

[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.