linear regression, likelihood $\mathcal{L}$ and χ^2

statistics and data analysis (chapter 7)

Björn Malte Schäfer

Graduate School for Fundamental Physics
Fakultät für Physik und Astronomie, Universität Heidelberg

May 27, 2016

outline

1. **data fitting**
2. **likelihood**
3. **properties of $\mathcal{L}$**
4. **distribution of χ^2**
5. **combining measurements**

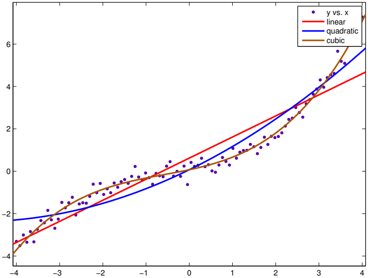
repetition

- formulation of hypotheses
- statistical tests: acceptance and rejection
- likelihood and likelihood ratios
- confidence levels and error types
- t-test, F-test and Kolmogorov-Smirnov test
- χ^2-distribution and student-t distribution

numerical exercise

repeat the numerical experiment of fitting, but this time try to fit an exponential $y(x) = \exp(-\lambda x)$ to artifical data (y_i, x_i) with some amount of Gaussian noise. what can you say about the distribution $p(\lambda)\mathrm{d}\lambda$?

data fitting - motivation



data points (x_i, y_i), polynomial models $y(x)$

- data (x_i, y_i) with errors σ_i, polynomial model $y(x)$
- best model?
  $\rightarrow$ linear inversion problem formulated with the moments!

data fitting - general case

- correspondence: t-test $\leftrightarrow$ fitting a straight horizontal line
- measurements x_i, drawn from a Gaussian probability distribution

$$p(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1}$$

  with parameters μ and σ
- why not simply $\mu \simeq \sum_i^N x_i/N$? correct **estimate**, but how likely is a different value μ' given the data x_i?
- **wanted**: distribution of μ given the observed data
  - infer the most likely value
  - peakiness of the distribution around the most likely value, errors
  - good alternative values μ' (multimodal distribution)

data fitting - questions

- consider fitting as a statistical test
  - define likelihood
  - optimised likelihood $\rightarrow$ principle of max. likelihood
- what is the best way to maximise the likelihood $\rightarrow$ Levenberg-Marquardt-algorithm
- what is the distribution of the model parameters if the measurement is repeated? $\rightarrow$ Γ distribution of the χ^2-functional for a Gaussian likelihood
- what are the properties of this posterior distribution $\rightarrow$ Fisher information, Cramer-Rao bounds
- how can measurements from different experiments be combined? $\rightarrow$ statistical independence of likelihoods
- what happens if data is fitted with the wrong model? $\rightarrow$ bias
- what if there are two competing models? what complexity is needed for describing the data? $\rightarrow$ (Bayesian) model selection

analogies to the t-test

1. t-test: estimate mean from a set of Gaussian random numbers
   $\rightarrow$ advantage: probability distribution for the estimated mean
2. fitting of a horizontal line: minimisation of the χ^2-functional gives arithmetic mean
   $\rightarrow$ advantage: linear problem, computation of σ_i-weighted moments

fuse both ideas:

derive the probability distribution of model parameters in the fitting of an arbitrary, nonlinear function

- linear models (polynomials) are a direct, exactly solvable generalisation to the t-test
- nonlinear models will involve a numerical extremisation scheme
- probability distribution of the model parameters depends on the choice of the likelihood

definition of a likelihood

- likelihood $\mathcal{L}$: probability of finding **past events** in a given model
- how probable was it to get a value x_i in the random process for an **assumed model** μ with fixed Gaussian dispersion σ?

$$\mathcal{L}(x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \tag{2}$$

- likelihood for N **independent** events

$$\mathcal{L}(x_1 \ldots x_N|\mu) = \prod_i^N \mathcal{L}(x_i|\mu) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left(-\frac{1}{2\sigma^2}\sum_i^N (x_i-\mu)^2\right) \tag{3}$$

- varying μ can optimise $\mathcal{L}$ and give the **best estimate** of μ
- definition of the $\chi^2(\mu)$-functional from the logarithmic likelihood $L = -\ln\mathcal{L}$

$$\mathcal{L}(x_1 \ldots x_N|\mu) \propto \exp\left(-\frac{\chi^2}{2}\right) \quad \text{with} \quad \chi^2 = \frac{1}{\sigma^2}\sum_i^N (x_i-\mu)^2 \tag{4}$$

likelihood as inverse conditional probability

- conditional probability: probability of A conditional on B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \rightarrow \quad P(B|A) = \frac{P(A|B)}{P(A)} P(B) \tag{5}$$

  results in **Bayes' law**

- identify:
  1. $P(B) = P(\mu)$: **prior** distribution of μ, before experiment was carried out
  2. $P(B|A) = P(\mu|x_i)$: **posterior** distribution of μ, after the measurement
  3. $P(A|B)/P(A) = \mathcal{L}(x_i|\mu)$ is the **likelihood** to obtain the measurement x_i for a model value μ

- likelihood function $\mathcal{L} : \mu \rightarrow P(x_i|\mu)/P(x_i)$
- likelihood gives the **increase of knowledge** on μ from the data x_i

question

derive Bayes' law from the definition of conditional probability

principle of maximum likelihood

- estimate model parameters μ by maximising $\mathcal{L}(x_i|\mu)$

$$\left.\frac{\partial\mathcal{L}}{\partial\mu}\right|_{\mu=\hat{\mu}} = 0 \tag{6}$$

  which defines the **maximum likelihood estimate** $\hat{\mu}$ of the parameter μ
- interpretation: given the correct model ($\mu = \hat{\mu}$) the probability of having obtained the measurement x_i must be largest
- if $\langle\hat{\mu}\rangle = \mu$, $\mathcal{L}$ is called **unbiased**, else $\langle\hat{\mu}\rangle = \mu + b$ with the **bias** b
- more convenient to look at $L = -\ln\mathcal{L}$ (logarithm is monotone)
- often, $\mathcal{L}$ is approximately Gaussian: σ_μ is the error in μ

question

show that $\hat{\mu} = \sum_i^N x_i/N$ and $\hat{\sigma} = \sum_i^N (x_i - \hat{\mu})^2/N$ using $\partial\ln\mathcal{L}/\partial\mu = 0$ for a Gaussian likelihood $\mathcal{L}(x_i|\mu,\sigma)$

principle of maximum likelihood

truth in science

we **suspect** the true parameters in a fixed physical model to be the ones that could have produced the data with the highest possible probability... please stand back and realise what a weak statement this is!

what about

- if Nature deals out very untypical values in a measurement?
- if the distribution is such that the most probable value is not the average value if repeated over many experiments?
- if the distribution is such that it has infinite variance (like the Cauchy distribution): what errors do the parameters have?
- if there are multiple maxima of the likelihood-function? which one is the right one?

Gauss-Markov-theorem

take one step back

in linear models (polynomials), the χ^2-minimisation gave a sensible result. is there any theoretical justification for that?

- justification of the χ^2-approach: **Gauss-Markov-theorem**
  - expecation value of the residuals is zero
  - variances are equal (unity)
  - covariances between the residuals are zero (uncorrelated data)
  - and the model depends **linearly** on parameters (polynomials)
- **least-squares**, i.e. χ^2-minimisation gives
  - **best** estimate of the model parameters (smallest variance)
  - **unbiased** estimates
  - inversion of a **linear** problem yields estimates
- still true for Gaussian likelihoods in nonlinear models. correspondence: $\mathcal{L}(x_i|\mu) = \exp(-\chi^2/2)$

repeating the experiment: distribution of χ^2

- generalisation of the χ^2-functional to measure the distance between data y_i and model prediction $y(x_i)$ if data is taken at points x_i with the individual error σ_i:

$$\chi^2 = \sum_i^N \left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2 \quad \rightarrow \quad \mathcal{L} \propto \exp\left(-\frac{\chi^2}{2}\right) \tag{7}$$

  where the model $y(x)$ has one or more parameters, which we're trying to measure

- distribution of the χ^2-functional if experiment is repeated

$$\chi^2 = \sum_i^N \left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2 \tag{8}$$

→ definition of an **Γ-distribution** for the χ^2

Γ-distribution for χ^2

- distribution of a sum of Gaussian random numbers with zero mean and unit variance $\rightarrow$ Γ-distribution
  1. distribution of a sum of random numbers
  2. distribution of the square of a random number
- first step: distribution of a sum of random numbers $\rightarrow$ 2 variants
- variant 1:
  - use cumulative distribution and redefine integration

$$s = x_1 + x_2 \rightarrow P(s \leq S) = \int_{x_1+x_2 \leq S} \mathrm{d}x_1 \mathrm{d}x_2 \, p(x_1)p(x_2) \tag{9}$$

  - substitute $u = x_1 + x_2$ and $v = x_2$: **convolution**

$$P(s \leq S) = \int_{-\infty}^{s} \mathrm{d}u \int_{-\infty}^{+\infty} \mathrm{d}v \, p(u-v)p(v) = \int_{-\infty}^{s} \mathrm{d}u \, p * p(u) \tag{10}$$

    which can be generalised to $N \geq 2$ by induction
  - and finally

$$p(x_1 + x_2) = \frac{\partial}{\partial s} P(s \leq S) = p * p(s) \tag{11}$$

Γ-distribution for χ^2

- variant 2:
  - characteristic function and properties of the exponential

$$\phi_{x_1+x_2}(t) = \langle\exp(-\mathrm{i}(x_1+x_2)t)\rangle = \langle\exp(-\mathrm{i}x_1 t)\rangle\langle\exp(-\mathrm{i}x_2 t)\rangle = \phi_{x_1}(t)\phi_{x_2}(t) \tag{12}$$

  - transform back to real space: product becomes convolution

$$p(s = x_1 + x_2) = p * p(s) \tag{13}$$

  - Gaussian: complete separability $\rightarrow$
    - sum of Gaussian random numbers is exactly Gaussian distributed
    - Gaussians give always Gaussians under convolution, with quadratic adding of the width: $\sigma^2 = \sigma_1^2 + \sigma_2^2$ ($\rightarrow$ importance of finite width for CLT)

question

show that convolutions in real space are products in Fourier-space

Γ-distribution for χ^2

- second step: distribution of the square $y = x^2$: use cumulative distribution:

$$P(y < Y) = P(x < \sqrt{y}) = \int_{-\infty}^{\sqrt{y}} \mathrm{d}x\, p(x) \tag{14}$$

- transform probability density:

$$p_y(y) = p_x(\sqrt{y})\frac{\mathrm{d}\sqrt{y}}{\mathrm{d}y} \quad \rightarrow \quad p_y(y) = \frac{\partial}{\partial y}P(y < Y) = \frac{1}{2\sqrt{y}}p_x(\sqrt{y}) \tag{15}$$

- for a Gaussian $p_x(x)$ with $\mu = 0$ and $\sigma = 1$:

$$p(y) = \frac{1}{2\sqrt{2\pi}}\frac{1}{\sqrt{y}}\exp(-y/2) \tag{16}$$

Γ-distribution for χ^2

- Γ-distribution

$$f_{\alpha,\nu} = \frac{\alpha^\nu}{\Gamma(\nu)} x^{\nu-1} \exp(-\alpha x) \tag{17}$$

- normalised with the Γ-function (generalisation of the factorial)

$$\Gamma(\nu) = \int_0^\infty \mathrm{d}x\, x^{\nu-1} \exp(-x) \tag{18}$$

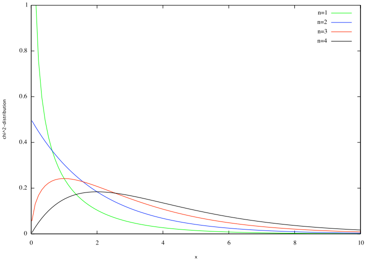
- convolution relation of Γ-distributions

$$f_{\alpha,\nu} * f_{\alpha,\mu} = f_{\alpha,\mu+\nu} \tag{19}$$

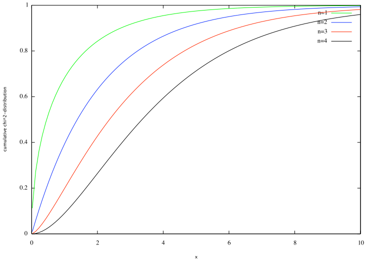
- Γ-distribution for χ^2-functional, for N data points

$$p_N(s) = f_{\frac{1}{2},\frac{N}{2}}(s) = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})} s^{\frac{N}{2}-1} \exp(-s/2) \quad \text{with} \quad s = \sum_i^N x_i^2 \tag{20}$$

visual impression of Γ-distributions



Γ-distribution for χ^2

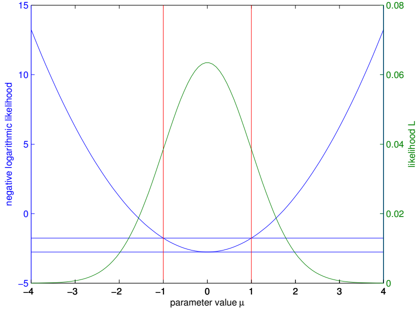


cumulative Γ-distribution

- equivalence:
  - maximisation of $\mathcal{L} = \exp(-\chi^2/2)/(\sqrt{2\pi})^N$
  - minimisation of $L = -\ln\mathcal{L} = \chi^2 - \frac{N}{2}\ln(2\pi)$

because the logarithm is monotonic

error bounds from a χ^2-fit



likelihood $\mathcal{L}$ and logarithmic likelihood L for $N = 3$

- CLT: likelihood $\mathcal{L}$ assumes a Gaussian shape for large N
- error bounds: $\Delta\chi^2 = n$ from the best fit point $\rightarrow$ $n\sigma$-bounds

combining likelihoods

- Bayes' law can be used for combining two measurements
- first measurement: x_i

$$p(\mu|x_i) = \frac{p(x_i|\mu)}{p(x_i)} p(\mu) = \mathcal{L}(x_i|\mu) p(\mu) \tag{21}$$

- second measurement: y_i

$$p(\mu|y_i) = \frac{p(y_i|\mu)}{p(y_i)} p(\mu) = \mathcal{L}(y_i|\mu) p(\mu) \tag{22}$$

- posterior of the first measurement is the prior of the second:

$$p(\mu|x_i, y_i) = \mathcal{L}(x_i|\mu)\mathcal{L}(x_i|\mu) p(\mu) = \mathcal{L}(x_i, y_i|\mu) p(\mu) \tag{23}$$

- independent likelihoods multiply:

$$\mathcal{L}(x_i, y_i|\mu) \equiv \mathcal{L}(x_i|\mu)\mathcal{L}(x_i|\mu) \tag{24}$$

combining likelihoods

- if each of the individual likelihoods is peaked, the combined $\mathcal{L}$ is more strongly peaked
- per induction, arbitrarily many likelihoods (from independent measurements) can be combined
- for a Gaussian likelihood $\mathcal{L} \propto \exp(-\chi^2/2)$:

$$\chi^2 \propto \ln \mathcal{L} = \ln(\mathcal{L}_1 \mathcal{L}_2) = \ln \mathcal{L}_1 + \ln \mathcal{L}_2 \propto \chi_1^2 + \chi_2^2 \tag{25}$$

question

what's the prior for the first ever measurement in a new field?

question

why is it not possible to multiply the posteriors as independent probabilities?

reduced χ^2 - consistency of a fit to data

- define the **reduced** χ^2:

$$\chi_r^2 = \frac{\chi^2}{\nu} \quad \text{with} \quad \nu = N - n_{\mathrm{param}}(+1) \tag{26}$$

  for N data points and n_{param} model parameters
- idea: normalise χ^2 for number of data points and model complexity
- for **Gaussian** statistics:
  - $\chi_r^2 \gg 1$: bad fit
  - $\chi_r^2 \simeq 1$: fit ok!
  - $\chi_r^2 \ll 1$: overfitting, fit too good
- but again: what we would like to have is a way of measuring **if a model is good**, or if we should rather start using a different model $\rightarrow$ Bayesian model selection

watch out

Gaussian statistics: 32% of points are outside the error bars!

p-value

- comparison between data and model gave a certain value for χ^2_{fit}
- if the fit is mediocre → have we just been unlucky with the data?
- what would be the probability of obtaining data **more extreme** (and therefore providing worse fits) than the data observed?
- for Gaussian likelihoods

$$\mathcal{L} \propto \exp\left(-\frac{\chi^2}{2}\right)$$

- compute the p-value: integral over the unlikely region of the likelihood, for a Gaussian likelihood: error-function $\mathrm{erf}(\chi^2_{\mathrm{fit}})$

$$p = \int_{\chi^2 > \chi^2_{\mathrm{fit}}} \mathcal{L}(\chi^2)$$

p-value is *not*

probability of the null-hypothesis being true, nor size of the test α!

correlated data points

- likelihood for N **independent** events

$$\mathcal{L}(x_1 \ldots x_N|\mu) = \prod_i^N \mathcal{L}(x_i|\mu) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left(-\frac{1}{2\sigma^2}\sum_i^N (x_i-\mu)^2\right) \tag{27}$$

- likelihood for N data points with correlations: introduce **data covariance**

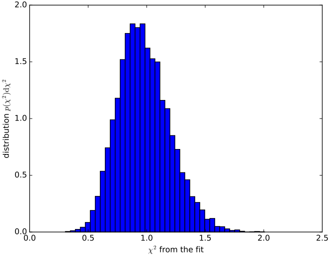
$$\mathcal{L}(\{x_i\}|\mu) = \frac{1}{\sqrt{(2\pi)^N \det C}} \exp\left(-\frac{1}{2}(x_i-\mu)C^{-1}_{ij}(x_j-\mu)\right) \tag{28}$$

which replaces the error σ_i

- the multivariate Gaussian distribution suggests as a χ^2 the expression

$$\chi^2 = \sum_{ij}(x_i-\mu)C^{-1}_{ij}(x_j-\mu) \tag{29}$$

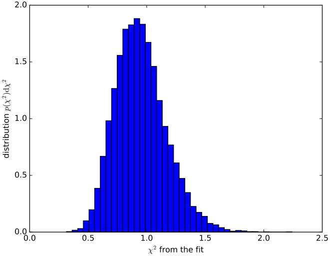
χ^2-distribution for the mean



distribution $p(\chi^2)\mathrm{d}\chi^2$ for the mean

- mean of data $y_i = \mathrm{const}$ with Gaussian noise
- estimate $\bar{y} = \langle y_i \rangle$, with $\chi^2 = \sum_i (y_i/\sigma)^2$

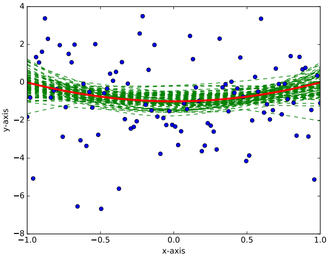
χ^2-distribution for fitting a straight line



distribution $p(\chi^2)\mathrm{d}\chi^2$ for a linear model (straight line)

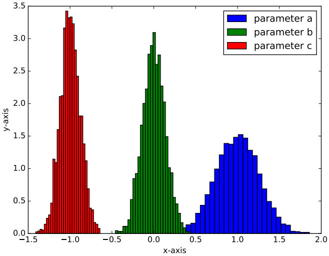
- linear model $y = ax + b$, with Gaussian noise
- generalises in an obvious way to fitting polynomials

outlook from the χ^2-distribution to errors



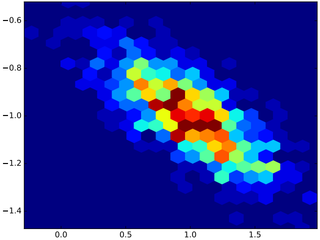
- polynomial model with Gaussian noise: linear inversion problem
- repetition of the experiment yields another noise realisation, and a different fit

outlook from the χ^2-distribution to errors



- each fit depends on the particular noise realisation
- repetition of the experiments gives a distribution of model parameters

outlook from the χ^2-distribution to errors



- and the model parameters are not independent
- they can be traded for each other: degeneracy

finding the minimum χ^2

- fit can be done exactly by matrix inversion for linear (polynomial) models
- nonlinear models: the fit can only be found by minimising χ^2 numerically:
  1. Gauss-Newton algorithm
  2. Newton-Raphson algorithm
  3. Levenberg-Marquardt algorithm (this is the one you want)
- but actually, this is rarely done in practice because of direct evaluation of likelihoods

summary

- fitting a nonlinear model to data with Gaussian errors according to the principle of maximum likelihood requires the minimisation of the χ^2-functional. in this case $\mathcal{L} \propto \exp(-\chi^2/2)$
- if the measurement is repeated (new realisation of the noise), the χ^2-values are distributed according to a Γ-distribution
- likelihood describes a distribution of infered model parameters from data, and is symmetric in data and model for the case of a Gaussian likelihood
- likelihood is combined in Bayes' law with the prior to form the posterior distribution of model parameters, which describes the knowledge one has about a model after carrying out the experiment

next steps:

derivation of the parameter distribution, forecasts of statistical errors, Fisher-information, and systematical errors