principle of maximum likelihood and nonlinear fitting

statistics and data analysis (chapter 8)

Björn Malte Schäfer

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

May 27, 2016

outline

repetition

- fitting problem, linear and nonlinear models
- principle of maximum likelihood (and minimum χ^2)

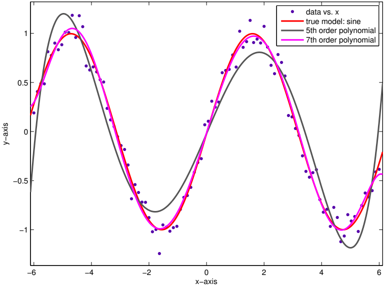
$$\mathcal{L} \propto \exp(-\chi^2/2) \quad \text{with} \quad \chi^2 = \sum_i \frac{(y_i - y)^2}{\sigma_i^2}$$

- Gauss-Markov theorem: correspondence of fits and statistical tests
- Γ-distribution for χ^2 for repeated experiments
- combination of likelihoods from different experiments

numerical exercise

assume n data points (x_i, y_i) as samples from a linear model $y(x) = ax + b$ with constant Gaussian error $\sigma_i = \sigma$. derive numerically the distribution $p(a, b)\mathrm{d}a\mathrm{d}b$ and from that the correlation coefficient r_{ab}. what's the transformation that diagonalises $p(a, b)$?

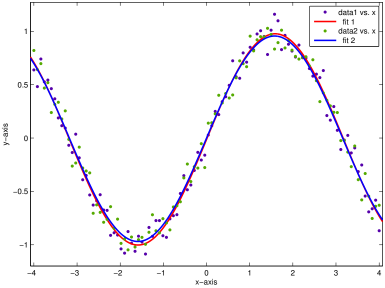
selecting the model...



fitting polynomials to a sine-wave

- choice of the model → lecture about Bayesian model selection
- in this lecture, we will assume that we fit data with the **correct** model

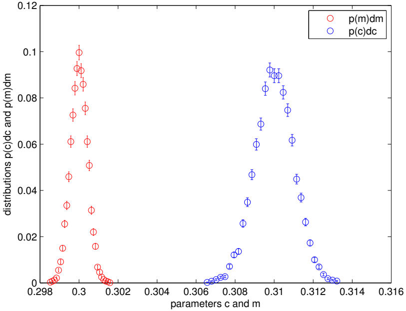
statistical properties of the likelihood



2 fits to the same model with new noise realisations

- 2 realisations of the noise → differences in the inferred parameters
- **what are the statistical properties of the likelihood, if the measurement is repeated?**

statistical properties of the likelihood



distribution of parameters

- fit of a **linear** model ($y(x) = mx + c$) → Gaussian likelihood
- uncertainty of parameters (width of likelihood) reflect noise
- on average, estimated parameters = true parameters (unbiased)

width of a Gaussian likelihood $\mathcal{L}$

- all information from a measurement about a model is contained in the likelihood $\mathcal{L}$
- for **simply shaped** likelihoods it should be sufficient to characterise our knowledge by stating the most probable estimates and their errors and covariances
- best estimate on μ: $\partial\mathcal{L}/\partial\mu = 0$ defines maximum μ^*
- error: width of the likelihood around μ^*, corresponds to curvature!
- Taylor-expansion of $L = \ln\mathcal{L} \rightarrow$ approximation with a Gaussian

$$L(\mu) = L(\mu^0) + \frac{\partial L}{\partial\mu}(\mu^0)(\mu - \mu^0) + \frac{1}{2}\frac{\partial^2 L}{\partial\mu^2}(\mu^0)(\mu - \mu^0)^2 + \dots$$

- exponentiate:

$$\mathcal{L} = \exp(L) \simeq \mathcal{L}_0 \exp\left(-\frac{1}{2}\frac{\partial^2 L}{\partial\mu^2}(\mu^0)(\mu - \mu^0)^2\right)$$

and identify $\sigma^2 \sim (-\partial^2 L/\partial\mu^2)^{-1}$, evaluated at μ^0

multivariate likelihoods

- more than one parameter: relation between the curvature and the inverse covariance
- Taylor-expansion of a multidimensional function:

$$L = L(\mu^0) + \sum_\alpha \frac{\partial L}{\partial \mu_\alpha}(\mu^0) + \frac{1}{2}\sum_{\alpha\beta}\frac{\partial^2 L}{\partial\mu_\alpha\partial\mu_\beta}(\mu-\mu^0)_\alpha(\mu-\mu^0)_\beta$$

  where the gradient vanishes due to the extremum
- we will see (shortly), that $C^{-1} = F$: the inverse covariance is the Fisher-matrix:

$$\mathcal{L}(\mu) = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}}\exp\left(-\frac{1}{2}(\mu-\mu^0)_\alpha(C^{-1})_{\alpha\beta}(\mu-\mu^0)_\beta\right)$$

- curvature of the likelihood: Fisher-matrix

$$F_{\alpha\beta} = \left\langle\frac{\partial^2 L}{\partial\mu_\alpha\partial\mu_\beta}\right\rangle$$

  width σ is given by $\sigma^2 \sim F^{-1}$

quadratic estimates: find the best parameters

- principle of maximum likelihood: find most **plausible** model parameters μ^*, which extremise the likelihood

$$\left.\frac{\partial\mathcal{L}}{\partial\mu_\alpha}\right|_{\mu=\mu^*} = 0$$

- remember: likelihood $\mathcal{L}(y_i|\mu)$ depends on data, it varies from measurement to measurement!
- finding the minimum: Taylor-expand $\mathcal{L}$ or $L = \ln\mathcal{L}$ around μ^0, which is a guessed value for the minimum μ^*

$$\left.\frac{\partial L}{\partial\mu_\alpha}\right|_{\mu} = \left.\frac{\partial L}{\partial\mu_\alpha}\right|_{\mu^0} + \sum_\beta \frac{\partial^2 L}{\partial\mu_\alpha\partial\mu_\beta}(\mu-\mu^0)_\beta + \ldots = 0$$

- apply iteratve Newton-Raphson scheme for finding the minimum:

$$(\mu-\mu^0)_\beta \simeq \sum_\alpha \left(\frac{\partial^2 L}{\partial\mu_\alpha\partial\mu_\beta}\right)^{-1}\frac{\partial L}{\partial\mu_\alpha} \rightarrow \mu_\beta = \mu_\beta^0 - \sum_\alpha \left(\frac{\partial^2 L}{\partial\mu_\alpha\partial\mu_\beta}\right)^{-1}\frac{\partial L}{\partial\mu_\alpha}$$

likelihood of a nonlinear fit

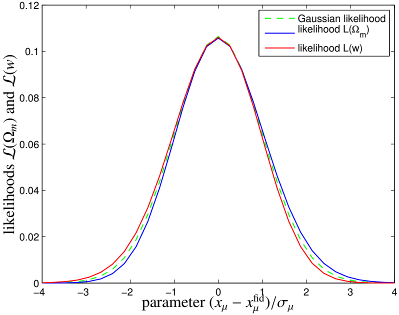
- nonlinear models $\rightarrow y = g(x)$ with a nonlinear function including parameters μ
  - χ^2 is not quadratic in the parameters μ
  - $\mathcal{L} \propto \exp(-\chi^2/2)$ is not Gaussian in μ
- **but:** strong measurements have very peaked liklihoods, and in the vicinity of μ^*, the model can be Taylor-expanded

$$\chi^2 = \sum_i \frac{1}{\sigma_i^2}(y_i - g(x_i,\mu))^2 \simeq \sum_i \frac{1}{\sigma_i^2}\left(y_i - \sum_\alpha \left.\frac{\partial g}{\partial \mu_\alpha}\right|_{\mu^*}(\mu-\mu^*)_\alpha \pm \ldots\right)^2$$

and χ^2 becomes then quadratic in μ

- close to the likelihood peak, everything looks Gaussian

example of a non-Gaussian likelihood



non-Gaussian and Gaussian likelihoods

- non-linearities in the model cause non-Gaussian shapes

Gaussian likelihoods: curvature

- make things easy: **Gaussian likelihood**

$$\mathcal{L}(y_i|\mu_\alpha) = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}}\exp\left(-\frac{1}{2}\sum_{ij} y_i (C^{-1})_{ij} y_j\right)$$

with covariance matrix C

- logarithmic likelihood L

$$L = \ln\mathcal{L} = \mathrm{const} - \frac{1}{2}\mathrm{tr}\ln C - \frac{1}{2}\sum_{ij} y_i (C^{1})_{ij} y_j$$

using $\ln\det C = \mathrm{tr}\ln C$

question

show that $\ln\det C = \mathrm{tr}\ln C$! (hint: principal axis transformation)

Gaussian likelihoods: curvature

- build first derivative of L wrt parameter μ_α

$$\frac{\partial L}{\partial \mu_\alpha} = -\frac{1}{2}\mathrm{tr}C^{-1}\frac{\partial C}{\partial \mu_\alpha} + \frac{1}{2}\vec{y}^t C^{-1}\frac{\partial C}{\partial \mu_\alpha}C^{-1}\vec{y}$$

with $\partial_\mu \ln C = C^{-1}\partial_\mu C$

- build second derivative of L wrt parameters μ_α and μ_β

$$\frac{\partial^2 L}{\partial \mu_\alpha \partial \mu_\beta} = \frac{1}{2}\mathrm{tr}C^{-1}\frac{\partial C}{\partial \mu_\alpha}C^{-1}\frac{\partial C}{\partial \mu_\beta} - \frac{1}{2}\mathrm{tr}C^{-1}\frac{\partial^2 C}{\partial \mu_\alpha \partial \mu_\beta} +$$

$$\frac{1}{2}\vec{y}^t C^{-1}\frac{\partial^2 C}{\partial \mu_\alpha \partial \mu_\beta}C^{-1}\vec{y} - \vec{y}^t C^{-1}\frac{\partial C}{\partial \mu_\alpha}C^{-1}\frac{\partial C}{\partial \mu_\beta}C^{-1}\vec{y}$$

- average $\langle\ldots\rangle$ over many measurements: for any matrix A

$$\langle \vec{y}^t A \vec{y}\rangle = \left\langle \sum_{ij} y_i A_{ij} y_j \right\rangle = \sum_{ij} A_{ij}\langle y_i y_j\rangle = \mathrm{tr}(AC)$$

Gaussian likelihoods: curvature

- average over many measurements replaces data y_i with the covariance matrix C
- Fisher matrix $F_{\alpha\beta}$: curvature of the likelihood surface

$$F_{\alpha\beta} = -\left\langle \frac{\partial \ln \mathcal{L}}{\partial\mu_\alpha\partial\mu_\beta} \right\rangle$$

- statistical errors (statistical uncertainties) σ_α^2 on the parameters μ_α follow from the inverse Fisher matrix
- **keep in mind**: large entries in F are good, they give small errors σ
- naturally, F is a positive definite and symmetric matrix, with real, positive eigenvalues

question

show that $\partial_\mu \ln C = C^{-1}\partial_\mu C$ and $\partial_\mu C^{-1} = -C^{-1}\partial_\mu C C^{-1}$!

Fisher-matrix $F_{\alpha\beta}$

- quantification of the **statistical errors** and **independence of parameters** of a fit
  - Gaussian noise σ_i
  - linear model $y(x)$
- Fisher matrix $F_{\alpha\beta}$: curvature of the likelihood surface

$$F_{\alpha\beta} = -\left\langle \frac{\partial \ln \mathcal{L}}{\partial\mu_\alpha \partial\mu_\beta} \right\rangle$$

- substitution of the second derivatives:

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left(C^{-1}\frac{\partial C}{\partial\mu_\alpha}C^{-1}\frac{\partial C}{\partial\mu_\beta}\right)$$

question

show that if data from two independent experiments are combined, the Fisher-matrices add!

Fisher-matrix: unbiased estimates

- if we estimate the best fit parameters: do they correspond to the **true** values used by Nature? at least for a Gaussian likelihood, with estimates following from a quadratic estimator, the answer is **yes!**
- estimates are **unbiased**: $\langle\mu\rangle = \mu^*$ (great news!)
- substitute Fisher matrix into the quadratic estimator

$$\langle\mu_\beta\rangle = \mu_\beta^0 + \frac{1}{2}(F^{-1})_{\alpha\beta}\left(\vec{y}^t C^{-1}\frac{\partial C}{\partial\mu_\alpha}C^{-1}\vec{y} - \mathrm{tr}C^{-1}\frac{\partial C}{\partial\mu_\alpha}\right)$$

- expand covariance C at initial guess μ^0

$$C(\mu) = C(\mu^0) + \frac{\partial C}{\partial\mu_\alpha}(\mu-\mu^0)_\alpha$$

- use $\langle y_i y_j\rangle = C_{ij}$

$$\langle\mu_\beta\rangle = \mu_\beta^0 + (\mu^* - \mu^0)_\gamma\underbrace{(F^{-1})_{\alpha\beta}F_{\alpha\gamma}}_{=\delta_{\beta\gamma}} = \mu_\beta^*$$

Cramer-Rao bounds σ_α

- on average, the estimate μ corresponds to the true value μ^*
- but what is the uncertainty when infering μ from data y_i?
- variance (just a single parameter!)

$$\sigma^2 \equiv \langle(\mu - \mu^*)^2\rangle = \langle\mu^2\rangle - (\mu^*)^2$$

use:

  - $\langle\vec{y}^t A\vec{y}\rangle = \mathrm{tr}AC$
  - Wick-theorem: $\langle y_i y_j y_k y_l\rangle = C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk}$
  - $\langle\vec{y}^t C^{-1}\partial_\mu C C^{-1}\vec{y} = \mathrm{tr}(C^{-1}\partial_\mu C)$

- finally: statistical error corresponds to the curvature of $\mathcal{L}$

$$\sigma_\alpha^2 = \langle(\mu - \mu^*)^2\rangle = \frac{1}{2}F^{-2}\mathrm{tr}\left(\frac{\partial \ln C}{\partial\mu}\frac{\partial \ln C}{\partial\mu}\right) = (F^{-1}) = \frac{1}{F}$$

Cramer-Rao bounds σ_α and correlations $r_{\alpha\beta}$

- for a multivariate likelihood, one distinguishes two types of error
    - marginalised errors: $\sigma_\alpha^2 = (F^{-1})_{\alpha\alpha}$ contains uncertainty in all other parameters
    - conditional errors: $\sigma_\alpha^2 = 1/F_{\alpha\alpha}$ assumes that all other parameters are perfectly known
- furthermore, the parameter might not independent: there might be compensating effects, and the parameters might show some degree of **degeneracy**
- quantified with the correlation coefficient $r_{\alpha\beta}$:

$$r_{\alpha\beta} = \frac{(F^{-1})_{\alpha\beta}}{\sqrt{(F^{-1})_{\alpha\alpha}(F^{-1})_{\beta\beta}}}$$

    - r close to $+1$: positively correlated
    - r very small: uncorrelated
    - r close to -1: anticorrelated

sensitivity

- what is important in an experiment to give small errors?
- obviously, large entries in $F_{\alpha\beta}$ give small uncertainties σ_α
- look at contributions to the Fisher matrix

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left(C^{-1}\frac{\partial C}{\partial\mu_\alpha}C^{-1}\frac{\partial C}{\partial\mu_\beta}\right) = \frac{1}{2}\mathrm{tr}\left(\frac{\partial\ln C}{\partial\mu_\alpha}\frac{\partial\ln C}{\partial\mu_\beta}\right) = \frac{1}{2}\mathrm{tr}\left(Q_\alpha Q_\beta\right)$$

- define **sensitivity** $Q_\alpha = \frac{\partial\ln C}{\partial\mu_\alpha}$
- good measurements have high Q_α:
  - small noise, C is small
  - depend strongly on a parameter, $\partial C/\partial\mu_\alpha$ is large
  - combine many measurements, sum over i

reparameterisation of the Fisher-matrix

- Fisher-matrix describes errors on parameters μ in a model $y(x)$
- what if you want to reexpress the errors in a different parameterisation for $y(x)$ with parameters τ?
- assume: there is an **invertible** mapping between μ and τ
- new Fisher matrix $F'_{\alpha\beta}$:

$$F'_{\alpha\beta} = \left\langle \frac{\partial^2}{\partial\tau_\alpha\partial\tau_\beta} \ln\mathcal{L} \right\rangle = \frac{\partial\mu_a}{\partial\tau_\alpha}\frac{\partial\mu_b}{\partial\tau_\beta} \left\langle \frac{\partial^2}{\partial\mu_a\partial\mu_b} \ln\mathcal{L} \right\rangle = J_{\alpha a}J_{\beta b}F_{ab}$$

with Jacobian matrices

$$J_{\alpha a} \equiv \frac{\partial\mu_a}{\partial\tau_\alpha}$$

fitting with the wrong model

- if there are systematic deviations present, the true model provides a bad fit!
- conversely, a fit would not correspond to the true model μ_t, and there are systematical errors in the infered parameters
- systematic bias $\delta \equiv \mu_w - \mu_t$
- write down χ^2-functionals

$$\chi_t^2 = \sum_i (y_i - y_t)^2 \quad \text{and} \quad \chi_w^2 = \sum_i (y_i - y_w)^2$$

- expand wrong χ_w^2 around the **true** best fitting model μ_t:

$$\chi_w^2 = \chi_w^2(\mu_t) + \sum_\alpha \frac{\partial \chi_w^2}{\partial \mu_\alpha}(\mu_t)\delta_\alpha + \frac{1}{2}\sum_{\alpha\beta} \frac{\partial^2 \chi_w^2}{\partial \mu_\alpha \partial \mu_\beta}(\mu_t)\delta_\alpha\delta_\beta$$

biases due to systematics

- average and find μ_w by extremisation

$$\underbrace{\left\langle\frac{\partial\chi_w^2}{\partial\mu_\alpha}\right\rangle_{\mu_t}}_{=a_\alpha} = -\underbrace{\sum_\beta\left\langle\frac{\partial^2\chi_w^2}{\partial\mu_\alpha\partial\mu_\beta}\right\rangle_{\mu_t}}_{G_{\alpha\beta}}\delta_\beta$$

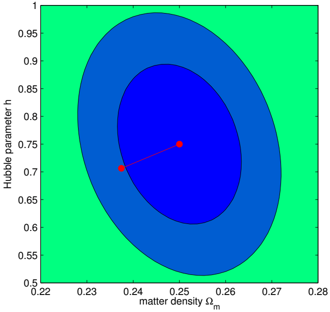
- solve for δ_β

$$\sum_\beta G_{\alpha\beta}\delta_\beta = a_\alpha \quad\rightarrow\quad \delta_\alpha = \sum_\beta (G^{-1})_{\alpha\beta}a_\beta$$

- systematic can be reduced if a strong prior is used

$$G_{\alpha\beta} \rightarrow G_{\alpha\beta} + F_{\alpha\beta}^{\mathrm{prior}}$$

example from cosmology: gravitational lensing



systematical and statistical errors

- uncorrected systematics bias the estimates

summary

- linear models and Gaussian noise provide Gaussian parameter likelihoods
- likelihood can be optimised using a quadratic estimator, e.g. Newton-Raphson
- estimates are unbiased
- variances are the smallest possible, Cramer-Rao errors
- systematics: model is incomplete, and parameter estimates are biased