Monte-Carlo Markov chain methods

statistics and data analysis (chapter 09)

Björn Malte Schäfer

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

June 20, 2016

outline

likelihood

- comparison between data y_i and model $y(x_i)$ at positions x_i with noise σ_i:

$$\mathcal{L} \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2\right) = \exp\left(-\frac{\chi^2}{2}\right) \tag{1}$$

with the χ^2-functional,

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2 \tag{2}$$

- likelihood $\mathcal{L}$: probability of obtaining the data y_i under the model $y(x)$
- likelihood depends on the model $y(x)$ and its parameters μ
- curvature $F_{\mu\mu}$ of $\ln\mathcal{L} = -\chi^2/2$ determines the inverse error, which depends on
  - magnitude of the errors σ_i
  - scaling of the model with parameters μ
  - location x_i of the data

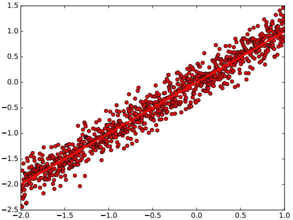
linear and nonlinear models

- the likelihood depends on the model parameters μ **through** the model $y(x)$:

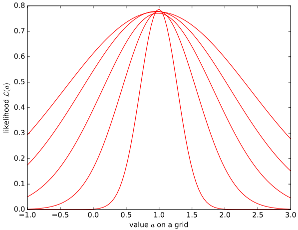
$$\mathcal{L}(\mu) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2\right) \tag{3}$$

- shape of the likelihood
  - linear (polynomial models): $y(x) \propto \mu \rightarrow \mathcal{L}$ is Gaussian
  - nonlinear dependence of $y(x)$ on $\mu \rightarrow$ non-Gaussian $\mathcal{L}$
- linear models: quantification of the posterior with the Fisher-matrix, suitable quantification of errors and covariance from inverse F
- Gaussian likelihood: use the relation between curvature $F_{\mu\mu}$ and confidence interval σ_μ^2
- nonlinear models: not possible. evaluation of the likelihood on a grid in μ, complicated functional shape of $\mathcal{L}(\mu)$

fit of a linear model



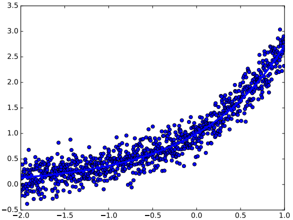
fit to artificial data



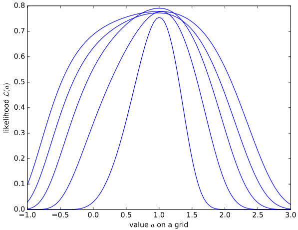
likelihood, different σ_i

- fit of a linear model $y(x) = \mu x$ with parameter μ
- Gaussian $\mathcal{L}$, width increases with increasing σ_i

fit of a nonlinear model



fit to artificial data



likelihood, different σ_i

- fit of a nonlinear model $y(x) = \exp(\mu x)$ with parameter μ
- non-Gaussian $\mathcal{L}$, width increases with increasing σ_i

evaluations of likelihoods

- it is always possible to evaluate $\mathcal{L}(\mu)$ on a grid
- but with many parameters, this might not be efficient:
  - n_{grid} resolution points in d dimensions
  - $n\mathrm{grid}^d$ points in total
- it is possible to sample from the likelihood $\mathcal{L}(\mu)$ at a very efficiently
- Monte-Carlo Markov-chains (MCMC)
  - typically, and MCMC chain needs much less than n_{grid}^d evaluations, rather just $n_{\mathrm{grid}}^{d/2}$ evaluations
- we'll encounter only the most basic MCMC-sampling techniques

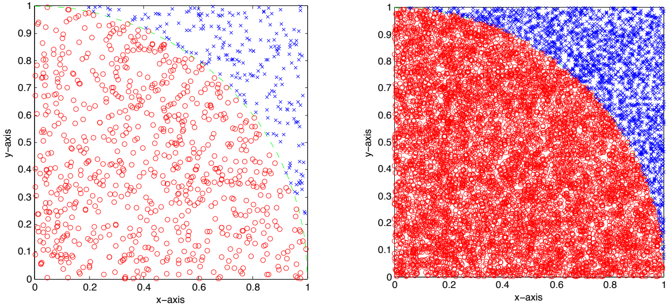
Monte-Carlo integration

- use randomness to your advantage in Monte-Carlo integrations
- particularly useful for
  - many dimensions
  - complicated integration boundaries

for professional use

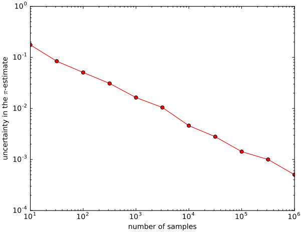
CUBA library, www.feynarts.de

Monte-Carlo integration

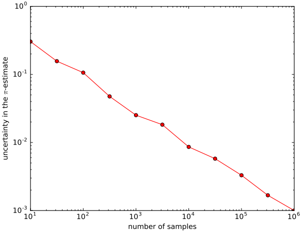


10^3 **samples** 10^4 **samples**

Monte-Carlo integration



precision 2d



precision 3d

- scaling of accuracy $\propto \sqrt{n_{\mathrm{sample}}}$
- same in 2d and 3d

barometric formula: fluid mechanics point of view

- solve for the density ρ of a fluid at rest inside a grav. potential Φ
- continuity and Navier-Stokes-equations:

$$\partial_t \rho = -\nabla(\rho\vec{\upsilon}) \tag{4}$$

$$\partial_t \vec{\upsilon} + (\vec{\upsilon}\cdot\nabla)\vec{\upsilon} = -\frac{\nabla p}{\rho} - \nabla\Phi + \mu\Delta\vec{\upsilon} \tag{5}$$

- apply stationarity condition $\vec{\upsilon} = 0$ and $\partial_t\vec{\upsilon} = 0$:

$$\partial_t \rho = 0 \tag{6}$$

$$0 = -\frac{\nabla p}{\rho} - \nabla\Phi \tag{7}$$

  density is stationary as well!
- assume an equation of state $p = \rho$ of an ideal gas:

$$\frac{\nabla p}{\rho} = \frac{\nabla \rho}{\rho} = \nabla \ln \rho = -\nabla\Phi \tag{8}$$

barometric formula: fluid mechancis point of view

- solved by an exponential:

$$\rho \propto \exp(-\Phi) \quad \text{solves} \quad \nabla \ln \rho = -\nabla\Phi \tag{9}$$

- for a homogeneous gravitational field: $\Phi \propto h$, if $\nabla\Phi$ = const:

$$\rho \propto \exp(-h) \tag{10}$$

with height h: barometric formula

question

what would be different for a polytropic equation of state $p \propto \rho^\alpha$?

barometric formula: statistical physics point of view

- the above derivation used the fluid mechanical equations
- but you can equivalently argue with statistics! (and it's even simpler)
  - a particle's energy fluctuates in thermal equilibrium
  - that's due to contact with a heat bath which keeps the temperature fixed
  - a particle can borrow a certain energy $\Delta\epsilon$ at temperature T with the probability
$$p = \exp\left(-\frac{\Delta\epsilon}{kT}\right) \tag{11}$$
    with the Boltzmann-constant k
  - the particle uses the energy $\Delta\epsilon$ for rising in the gravitational field to reach the height $h \propto \Delta\epsilon$
  - p is at the same time the fraction of particles that reaches the height h, so it's proportional to density
- therefore, $\rho \propto \exp(-h)$, like before

Markov-processes

idea:

if we think of $\chi^2(\mu)$ as a potential with μ as a position and could set up a particle doing thermal motion along μ, it would reflect in the distribution of its position the likelihood $\mathcal{L}(\mu)$

- hop $\mu_n \rightarrow \mu_{n+1}$ should always happen at the same probability
- hop $\mu_n \rightarrow \mu_{n+1}$ depends only on μ_n, but not on μ_{n-1} and earlier positions
- all probabilities to reach μ_{n+1} add up to one
- ergodicity: a sequence $\{\mu_n\}$ samples $\mathcal{L}(\mu)$

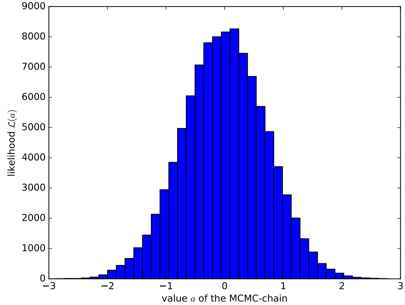
Metropolis-Hastings algorithm

- start at initial position μ
- draw a new value μ'
- and evaluate $\Delta\chi^2 = \chi^2(\mu') - \chi^2(\mu)$
- if $\Delta\chi^2 < 0$ (better fit, downhill), keep μ'
- if $\Delta\chi^2 > 0$ (worse fit, uphill), decide if you keep μ':
  draw a number a uniformly from the interval $[0\ldots1]$
  - if $a < \exp(-\Delta\chi^2)$: keep μ'
  - if $a > \exp(-\Delta\chi^2)$: reject μ' and keep μ instead
- repeat, starting with the new μ

Metropolis-Hastings algorithm

produces a sequence of samples for μ with follow the distribution $\mathcal{L}(\mu) = \exp(-\chi^2/2)$, and you need to supply $\chi^2(\mu)$

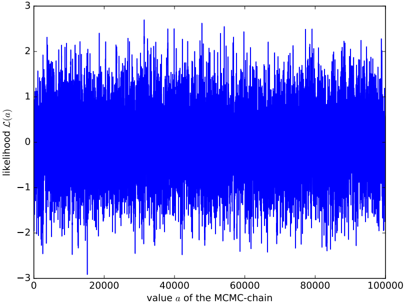
Metropolis-Hastings-algorithm



sample distribution of an MCMC-chain moving on a parabola

- simple test case: set $\chi^2(\mu) = \mu^2$ as a parabola
- likelihood $\mathcal{L}(\mu) \propto \exp(-\chi^2/2)$ is a Gaussian

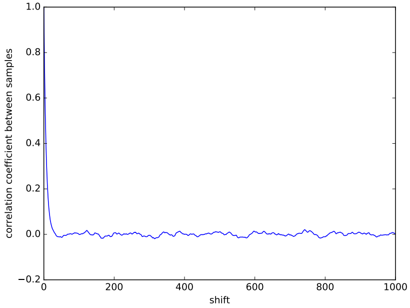
Metropolis-Hastings-algorithm



time sequence of samples of an MCMC-chain

- sequence of samples μ from the MH-algorithm
- it is very unlikely to move far from the minimum of the parabola: requires multiple, unlikely steps

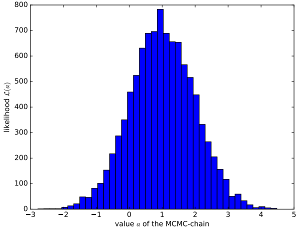
Metropolis-Hastings-algorithm



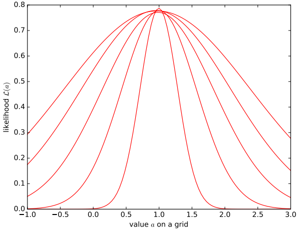
correlations of samples along the MCMC-chain

- Markov-processes: sampling depends on previous samplings
- samplings are correlated: to be measured by correlations $\langle\mu(n)\mu(n+\Delta n)\rangle$ for a shift Δn

fit of a linear model



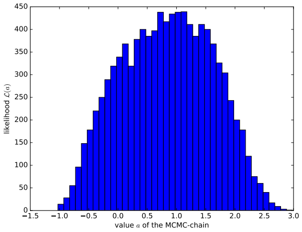
MCMC-sampling of $\mathcal{L}$



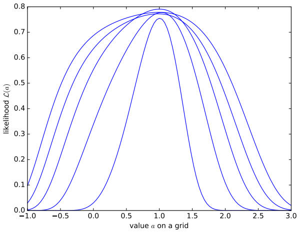
analytic $\mathcal{L}$

- MCMC-sampling recovers Gaussian shape typical for linear models

fit of a nonlinear model



MCMC-sampling of $\mathcal{L}$



analytic $\mathcal{L}$

- MCMC recovers non-Gaussian shape of the likelihood for nonlinear models

effective linear models

- nonlinear models $\rightarrow y(x)$ 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

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

  and χ^2 becomes then quadratic in μ
- close to the likelihood peak, everything looks Gaussian
- if the data is good and very constraining, the assumption of a Gaussian likelihood is good

Markov-processes and detailed balance

- detailed balance: Boltzmann-distribution in equilibrium
- ratio of jumps into μ and out of μ must be identical:

$$p(\mu)P(\mu \to \nu) = p(\nu)P(\nu \to \mu) \tag{12}$$

  with distributions $p(\mu)$ and transition probabilities $p(\mu \to \nu)$
- detailed balance condition: actually **Bayes' law**, if you interpret the transition $p(\mu \to \nu) = p(\nu|\mu)$
- if detailed balance is fulfilled, the distributions don't change: each point μ loses probability to the point ν at the same rate as the inverse process
- rewrite:

$$\frac{p(\mu)}{p(\nu)} = \frac{p(\nu \to \mu)}{p(\mu \to \nu)} \tag{13}$$

detailed balance

- separate transition probability $p(\mu \to \nu)$ into
  - a proposal distribution $g(\mu \to \nu)$, and
  - an acceptance distribution $A(\mu \to \nu)$,

  with $p(\mu \to \nu) = g(\mu \to \nu) \times A(\mu \to \nu)$
- insert:

$$\frac{A(\mu \to \nu)}{A(\nu \to \mu)} = \frac{p(\nu)}{p(\mu)} \times \frac{g(\nu \to \mu)}{g(\mu \to \nu)} \tag{14}$$

- Metropolis' choice:

$$A(\mu \to \nu) = \max\left(1, \frac{p(\nu)}{p(\mu)} \times \frac{g(\nu \to \mu)}{g(\mu \to \nu)}\right) \tag{15}$$

  can be shown to sample from a stationary distribution in a unique way
- often, one chooses a symmetric proposal, $g(\mu \to \nu) = g(\nu \to \mu)$

summary

- linear models have Gaussian, nonlinear models non-Gaussian likelihoods
- likelihood evaluation for arbitrary models on a grid in parameter space, if no special functional form is expected
- non-Gaussian likelihoods: relation between curvature and confidence intervals not present
- difficult for many parameters: use MCMC instead
- MCMC generates samples of $\mathcal{L}$ given χ^2
- advantage for many dimensional parameter spaces
- Metropolis-Hastings algorithm for generating samples