Regression, Markov Chain Monte Carlo (MCMC), Bayesian statistics & Applications

Dr. Veselina Kalinova

Institute for Sustainable Hydrogen Economy
Forschungszentrum Jülich

Outline

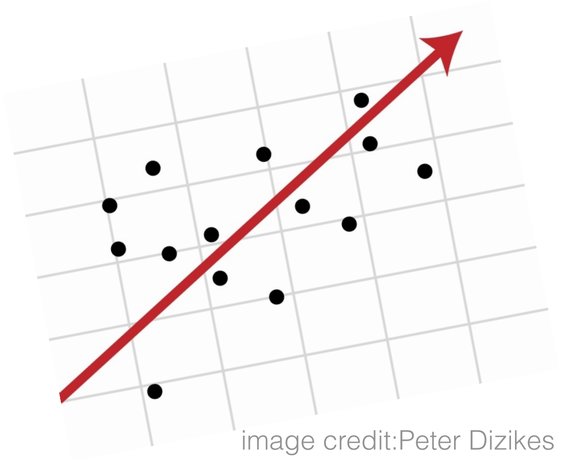


image credit:Peter Dizikes

- Regression Analysis
- Monte Carlo & Markov Chain
- Bayes' theorem
- Maximum Likelihood Function
- Bayesian generalisation
- The Metropolis algorithm
- Applications

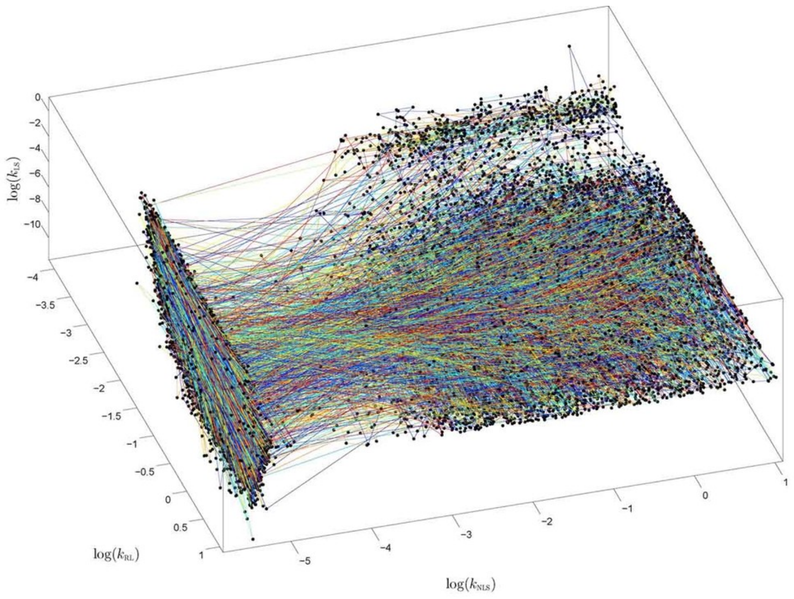


image credit: Pratyush Sinha

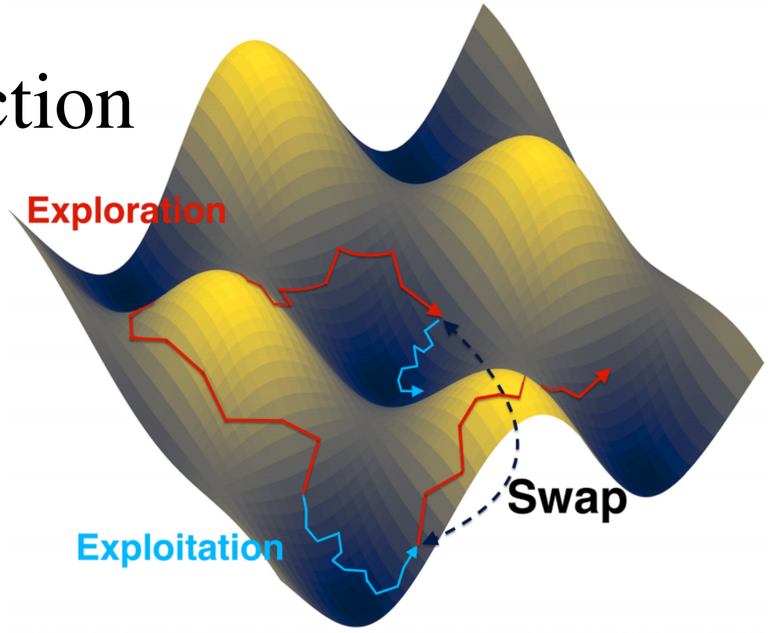
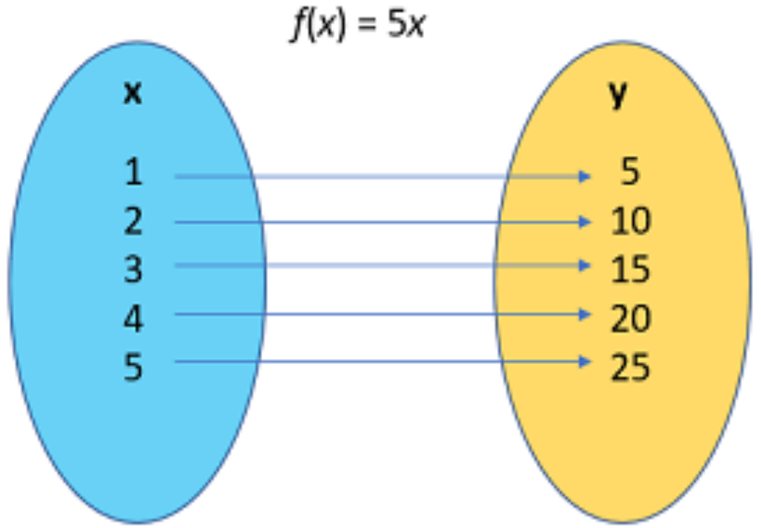


image credit: https://paperswithcode.com/methods/category/markov-chain-monte-carlo

Overview

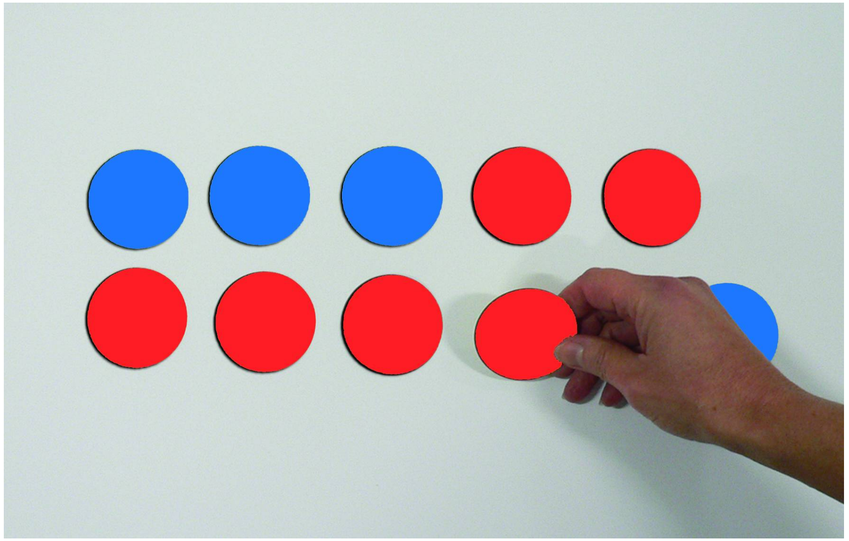
Supervised Machine Learning (SML): having input variables (x) and output variables (y), adopt an algorithm to understand the mapping function: y=f(x).



source: https://www.statisticshowto.com/mapping-diagram-for-functions/

Goal of SML: to find a function f(X), which can predict well the new output values (Y') for a given set of input values (X').

- Regression problem: the output value is a real value, i.e., "height" or "weight"



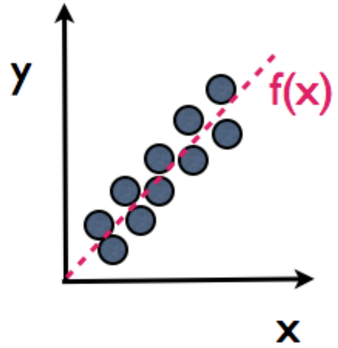
source: https://www.autopresseducation.co.uk/shop/magnetic-counters-strip-10-counters-3/

• Classification problem: the output value is a category, i.e., "red" or "blue"

Regression Analysis

2.1 Linear regression

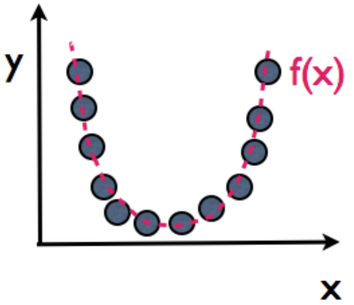
Example: a) line-fitting



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1}$$

where i=1,...,n is the particular observation; β_0 and β_1 - linear parameters, ϵ_i - error.

b) parabola-fitting



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{2}$$

where i=1,...,n is the particular observation; β_0, β_1 and β_2 - linear parameters, ϵ_i - error (still linear regression, because the coefficients β_0, β_1 and β_2 are linear, although there is a quadratic expression of x_i).

2.2 Ordinary least squares

We define the residual e_i as the difference between the value of the dependent variables, predicted by the model y_i^{MOD} and the true value of the dependent variables, y_i^{OBS}, i.e.,

$$e_i = y_i^{OBS} - y_i^{MOD} \tag{3}$$

One way to estimate the residuals is through "ordinary least squares" method, which minimaze the *sum of the squared residuals (SSE)*:

$$SSE = \sum_{i=1} n e_i^2 \tag{4}$$

The the mean square error of regression is calculated as

$$\sigma_\epsilon^2 = SSE/dof, \tag{5}$$

where $dof = (n - p)$ with n-number of observations, and p-number of parameters or $dof = (n - p - 1)$ if intersept is used.

2.3 Best fit of function

The best fit of a function is defined by the value of the "chi-square":

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i^{OBS} - y_i^{MOD}]^2}{\epsilon_{y_i}^2}, \tag{6}$$

where our data/observations are presented by y_i^{OBS} with error estimation ϵ_{y_i}, and model function y_i^{MOD}.

It is aslo necessary to know the number of degrees of freedom of our model ν when we derive the χ^2 , where for n data points and p fit parameters, the number of degrees of freedom is $\nu = n - p$. Therefore, we define a reduced chi-square χ^2_ν as a chi-square per degree of freedom ν:

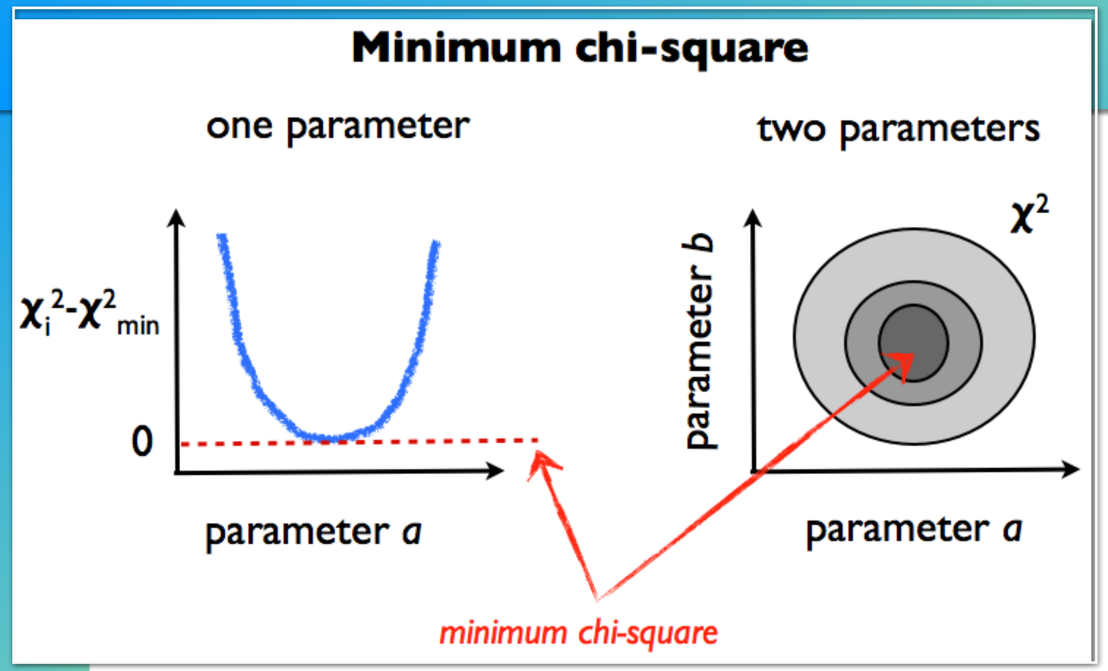
Regression

$$\chi^2_\nu = \chi^2/\nu, \tag{7}$$

where $\nu = (n - m)$ with n - number of measurements, and p - number of fitted parameters.

- $\chi^2_\nu < 1 \rightarrow$ over-fitting of the data
- $\chi^2_\nu > 1 \rightarrow$ poor model fit
- $\chi^2_\nu \simeq 1 \rightarrow$ good match between data and model in accordance with the data error

Regression Analysis



2.4 The minimum chi-squared method

- The optimum model is the satisfactory fit with several degrees of freedom, and corresponds to the minimisation of the function (left panel)
- Often used in astronomy when there is not realistic estimations of the data uncertainties

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i^2}, \tag{8}$$

where O_i - observed value, and E_i - expected value.

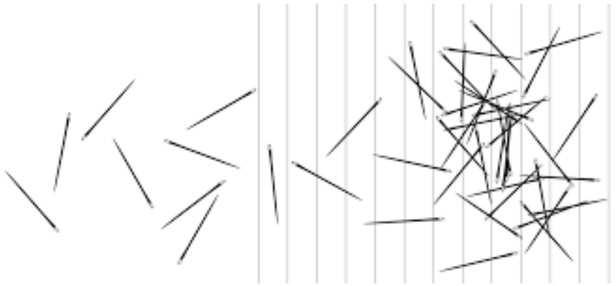
The minimum chi-square method aims to find the best fit of a function

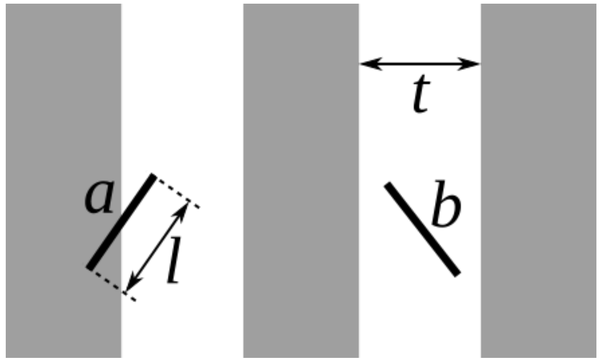
Left panel: for one parameter, Right panel: for two parameters

If we have the chi-squared for two parameters, the best fit of the model can be represented as a contour plot (see right panel).

Markov Chain Monte Carlo

3.1 Monte Carlo method (MC):



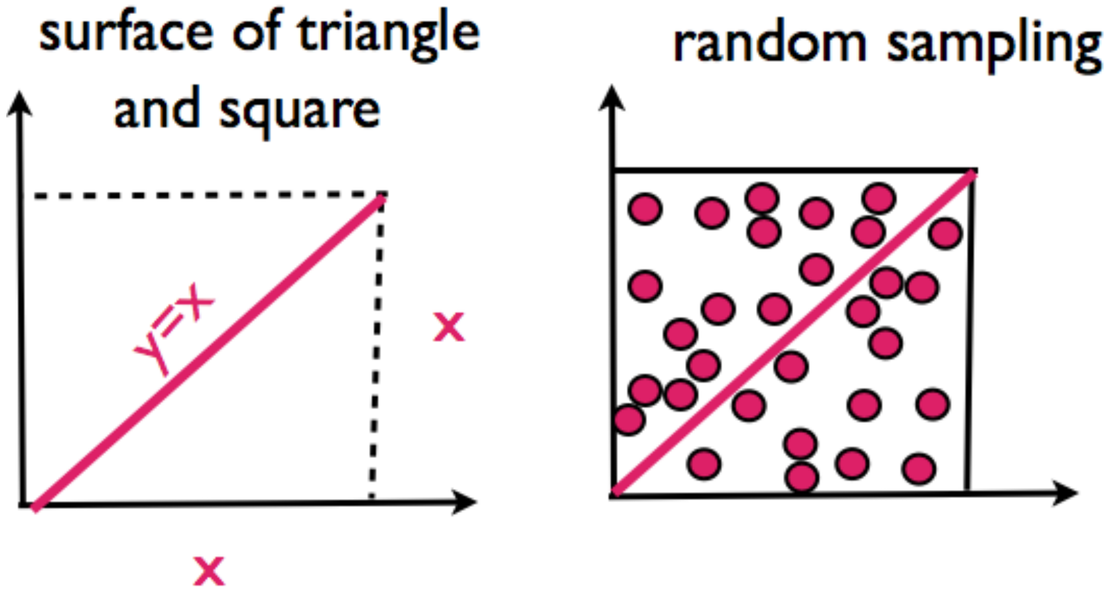


source: Wikipedia

- Definition:
  "MC methods are computational algorithms that rely on repeated random sampling to obtain numerical results, i.e., using randomness to solve problems that might be deterministic in principle".

- History of MC:
  First ideas: G. Buffon (the "needle experiment", 18th century) and E. Fermi (neutron diffusion, 1930 year)
  Modern application: Stanislaw Ulam and John von Neumann (1940), working on nuclear weapons projects at the Los Alamos National Laboratory
  The name "Monte Carlo": chosen as a secret name for the nuclear weapons projects of Ulam and Neumann; it is named after the casino "Monte Carlo" in Monaco, where the Ulam's uncle used to play gamble. MC reflects the randomness of the sampling in obtaining results.

Markov Chain Monte Carlo



Steps of Monte Carlo:

a) define a domain of possible inputs

b) generate inputs randomly from a probability function over the domain

c) perform deterministic computation on the inputs (one input - one out- put)

d) aggregate (compile) the results

- Perfoming Monte Carlo simulation

  Area of the triangle, $A_t = \frac{1}{2}x^2$
  Area of the square, $A_{box} = x^2$

  Therefore, $\frac{1}{2} = \frac{A_t}{A_{box}} \Rightarrow A_{box} = \frac{1}{2}A_t$

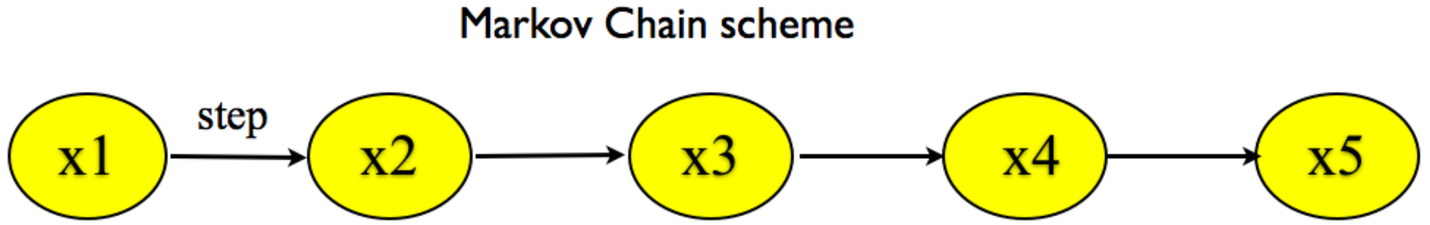
We can define the ratio between any figure inside the square box by random sampling of values.

$\frac{16}{30} \sim \frac{1}{2}$ by counting the randomly seeded points

$\frac{1}{2} \sim \frac{counts\,in\,triangle}{counts\,in\,box}$

Monte Carlo algorithm is random - more random points we take, better approximation we will get for the area of the triangle (or for any other area imbeded in the square) !

Markov Chain



- **First idea**: Andrey Markov, 1877

- **Definition**: If a sequence of numbers follows the graphical model in Fig. 4, it is a "Markov Chain". That is, the probability p_i for a given value x_i for each step "i":

$$p(x_5|x_4, x_3, x_2, x_1) = p(x_5|x_4) \tag{9}$$

The probability of a certain state being reached depends only on the previous state of the chain!

- **A descrete example for Markov Chain**:
We construct the Transition matrix **T** of the Markov Chain based on the probabilities between the different state X_i, where i is the number of the chain state:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

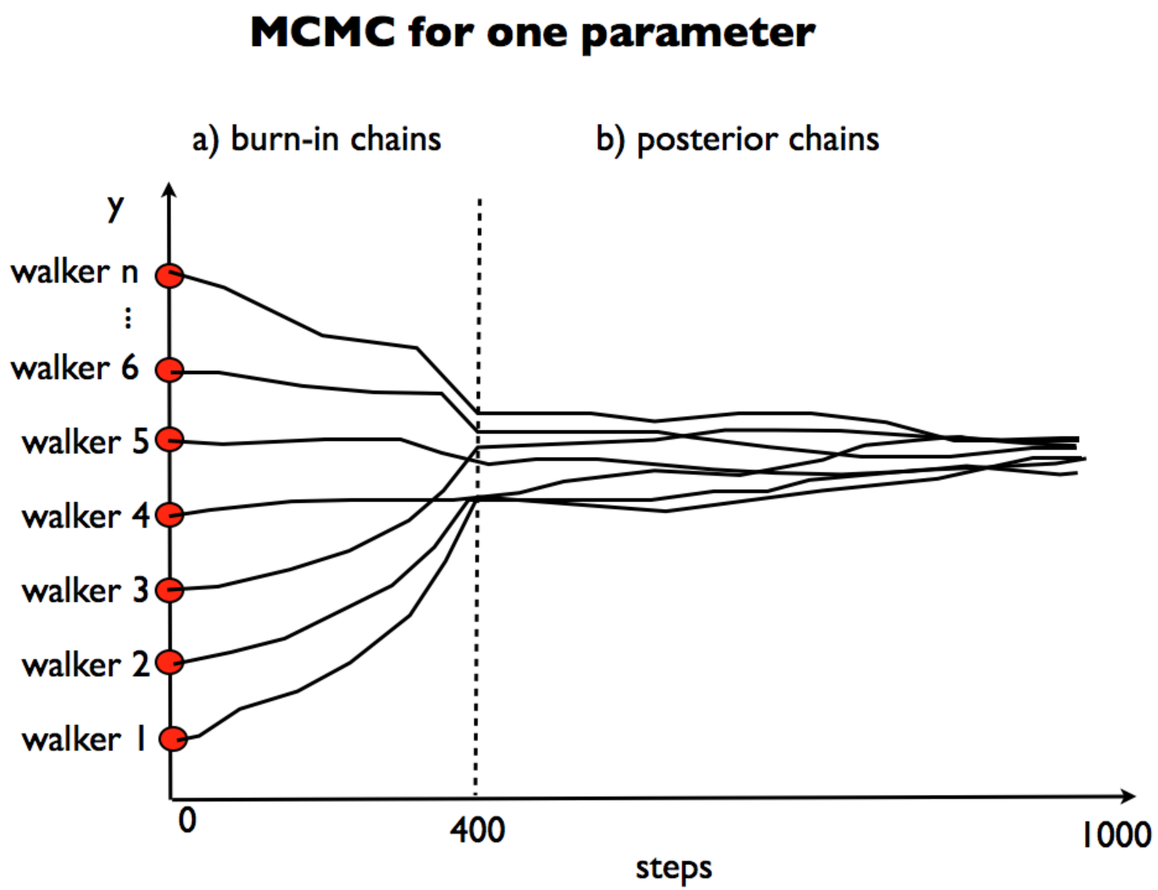
Let's take a starting point X0 with initial probabilities X0 = (0.5, 0.2, 0.3). The next step X1 will evolve as X1 = X0 x **T** = (0.18, 0.64, 0.18)

Additional two conditions have to be applied in the evolution of the system, the chains have to be:

a) Irreducible - for every state Xi, there is a positive probability of moving to any other state.

b) Aperiodic - the chain must not get trapped in cycles.

Phases of Markov Chain

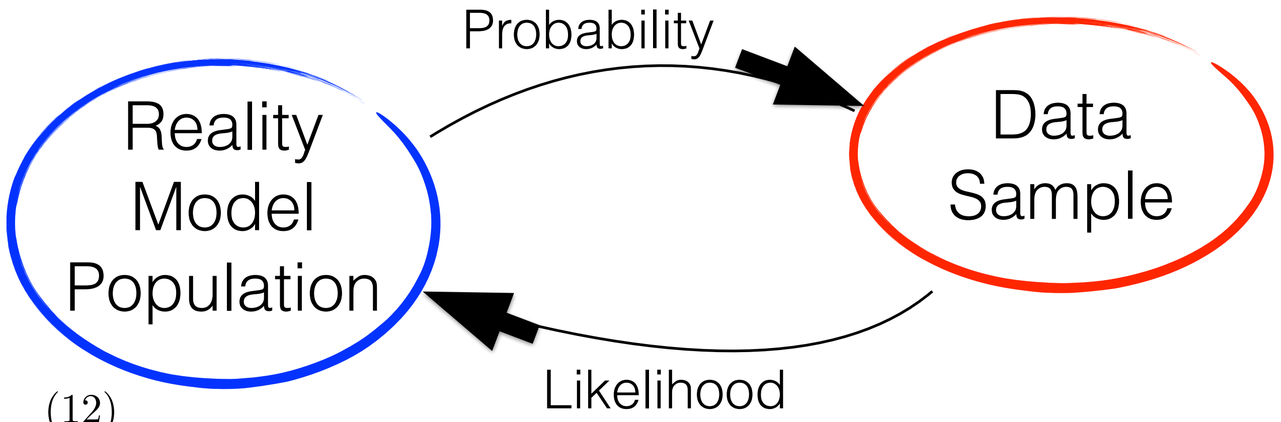


- **”burn-in” chain** - throwing some initial steps from our sampling, which are not relevant and not close to the converging phase (e.g., we will remove some stuck walkers or remove ”bad” starting point, which may over-sample regions with very low probabilities)

- **“posterior” chain** - the distribution of unknown quantities treated as a random variables conditional on the evidence obtain from an experiment, i.e. this is the chain after the burn-in phase, where the solution settles in an equilibrium distribution (the walkers oscillate around a certain value)

Maximum Likelihood Function

Given training data set: $x_1, x_2, ..., x_n$
Given probability function: $\mathrm{P}(x_1, x_2, ..., x_n; \theta)$
Asked: Find the maximum likelihood estimated of θ



$$\mathcal{L}\infty(\theta|x_1, x_2, ..., x_n) = P(x_1|\theta)\, P(x_2|\theta) \ldots P(x_n|\theta) \tag{12}$$

Or in short, the likelihood is expressed as the product of the individual probabilities for a given θ,

$$\mathcal{L}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{m} P(x_i|\theta) \tag{13}$$

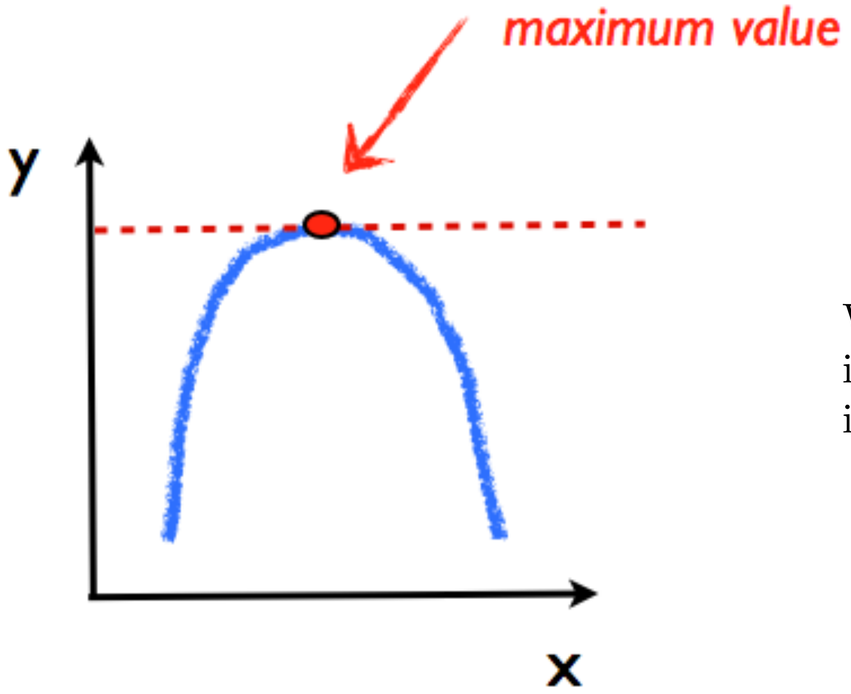
log-likelihood function: the maximisation of $\mathcal{L}$ is difficult to calculate as a product of different probabilities; and we find instead the logarithmic function of the likelihood, where this product turns to a sum:

$$\ln \mathcal{L}(\theta|x_1, x_2, ..., x_n) = \sum_{i=1}^{m} \ln P(x_i|\theta) \tag{14}$$

A likelihood function (often simply called the likelihood) measures how well a statistical model explains observed data by calculating the probability of seeing that data under different parameter values of the model

It is constructed from the joint probability distribution of the random variable that (presumably) generated the observations

Maximise and verify log-probability function



a) maximaze log-probability function

We need to find the maximum value of the log-probability function, corresponding to the optimum best-fit value of the function for a given parameter θ. This is exactly the derivative of the $\ln \mathcal{L}$ with repect to θ made equal to zero:

$$\frac{\partial \ln \mathcal{L}(\theta|x_1, x_2, ..., x_n)}{\partial \theta} = 0 \tag{15}$$

b) verify log-probability function (see Fig. 7)

To find the global maximum of the log-probability function,

$$\frac{\partial^2 \ln \mathcal{L}(\theta|x_1, x_2, ..., x_n)}{\partial \theta} < 0 \tag{16}$$

Example

Let's find the likelihood function of data represented by a line in the form $y = f(x) = mx + b$, where any reason for the data to deviate from a linear relation is an added offset in the y-direction. The error y_i was drawn from a Gaussian distribution with a *zero mean* and *known variance* σ_y^2.

In this model, given an independent position x_i, an uncertainty σ_{y_i}, a slope m, an intercept b, the frequence distribution p is:

$$p(y_i|x_i, \sigma_{y_i}, m, b) = \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} e^{-\frac{|y_i - mx_i - b|^2}{2\sigma_{y_i}^2}}. \tag{23}$$

Therefore, the likelihood will be expressed as:

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i|x_i, \sigma_{y_i}, m, b) \Rightarrow \tag{24}$$

$$\ln \mathcal{L} = K - \sum_{i=1}^{N} \frac{|y_i - mx_i - b|^2}{2\sigma_{y_i}^2} = K - \frac{1}{2}\chi^2, \tag{25}$$

where K is some constant.

Thus, the likelihood maximization is identical to χ^2 minimization !

Bayesian generalisation

The Bayesian generalization of the frequency distribution p, described in equation 23, have the following expression:

$$p(m, b|\{y_i\}_{i=1}^N, I) = \frac{p(\{y_i\}_{i=1}^N|m, b, I)\, p(m, b|I)}{p(\{y_i\}_{i=1}^N|I)}, \tag{26}$$

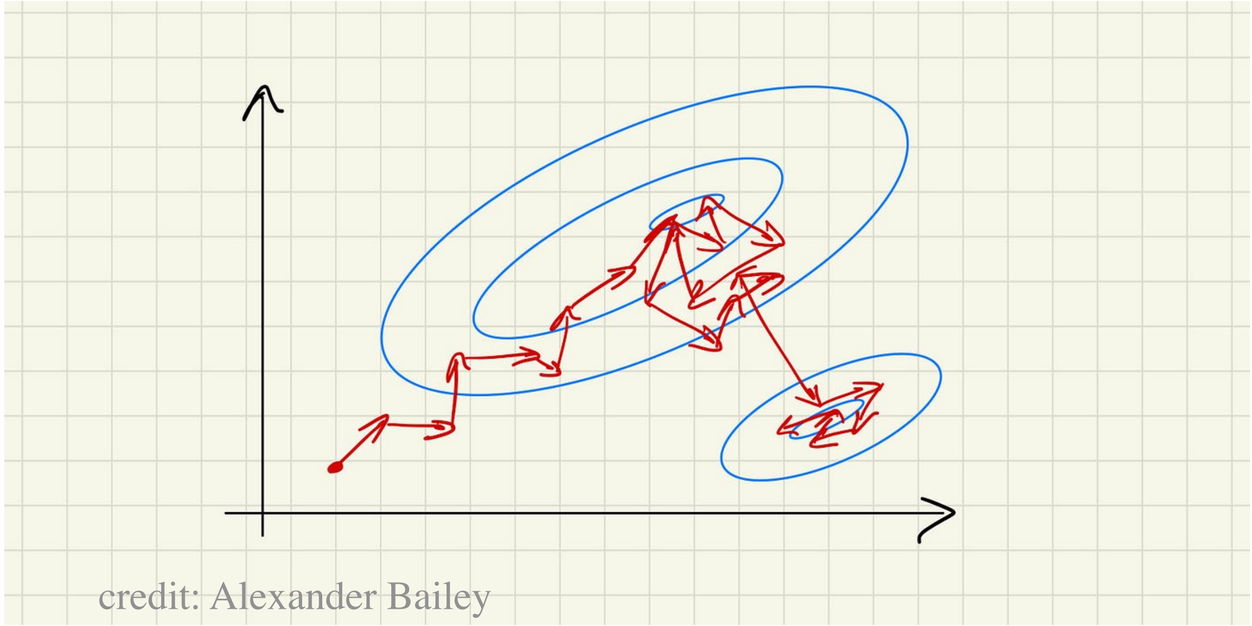
where m, b – model parameters
$\{y_i\}_{i=1}^N$ – short-hand for all of the data y_i
I– short-hand for all the prior knowledge of the x_i and σ_{y_i}.

Further, we can read the contributors in equation 26 as the following:
$p(m, b|\{y_i\}_{i=1}^N, I) \rightarrow$ *Posterior* distribution
$p(\{y_i\}_{i=1}^N|m, b, I) \rightarrow$ *Likelihood* $\mathcal{L}$ distribution
$p(m, b|I) \rightarrow$ *Prior* distribution
$p(\{y_i\}_{i=1}^N|I) \rightarrow$ *Normalization* constant

Or,

$$Posterior = \frac{Likelihood \times Prior}{Normalization} \tag{27}$$

The Metropolis-Hastlings algorithm


credit: Alexander Bailey

- First ideas (in the modern time):
  The algorithm is originally invented by Nicholas Metropolis (1915-1999), but generalised by Wilfred Hastings (1930-2016), called Metropolis-Hastings algorithm.

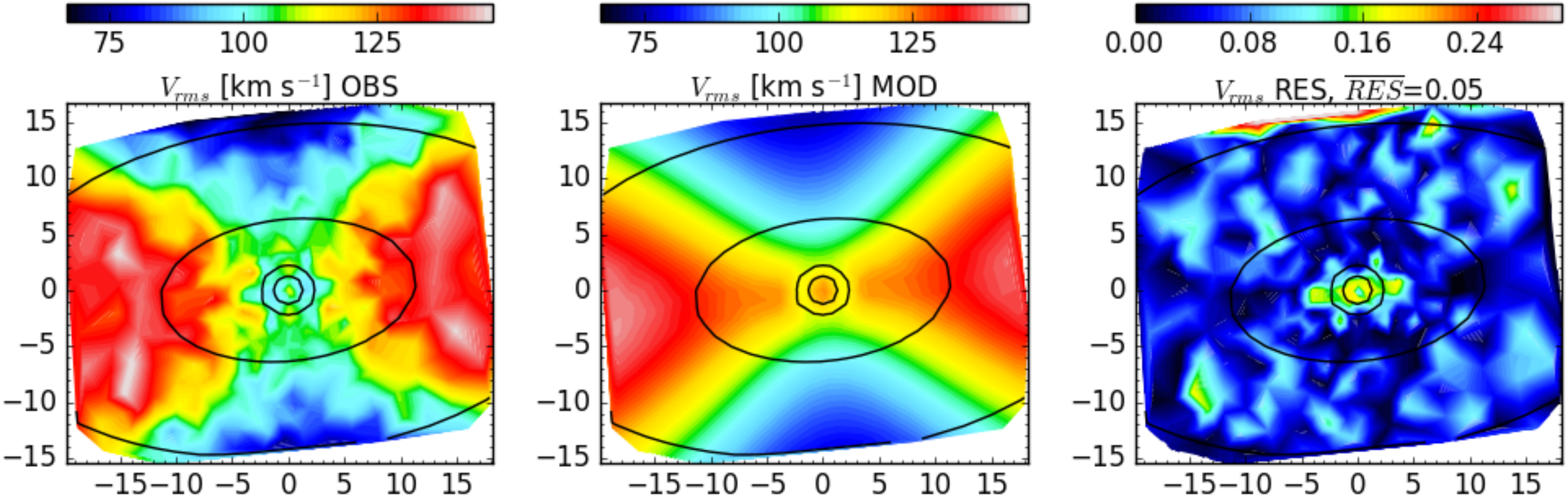
Basic assumptions in the algorithm

- assumes a symmetric random walk for the proposed distribution, i.e., $q(x|y) = q(y|x)$
- the posterior distribution is approximated to the Bayesian probabilities since the dominator term in equation 27 is difficult to calculate in practice, but at the same time possible to ignore due to its normalisation nature
- the walkers are keep jumping to explore the parameter space even if they already found the local minimum
- to improve the efficiency in the MCMC algorithm, the burn-in chains are removed

Application

Fitting galaxy dynamical models

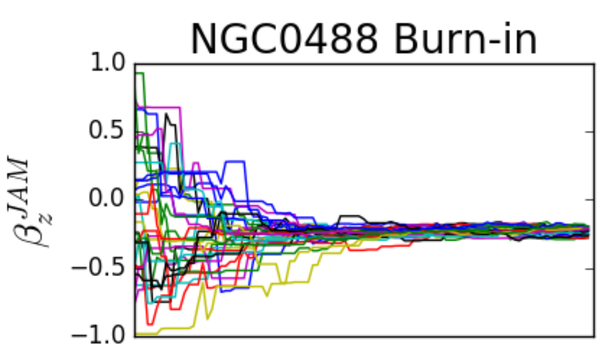
second velocity moment maps and residual, including both velocity and velocity dispersion of the galaxy

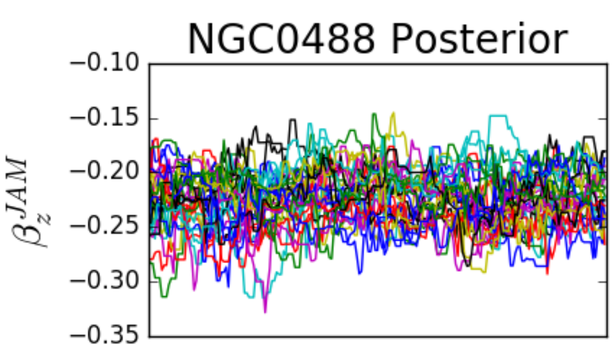


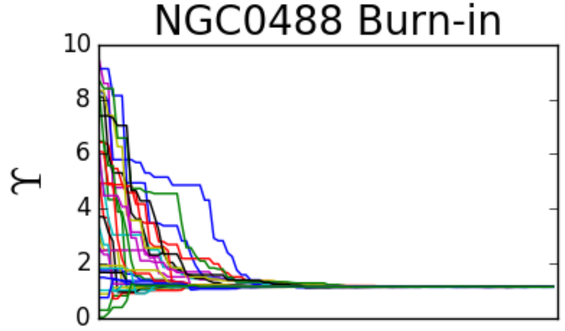
credit: Kalinova et al, MNRAS, 2017a

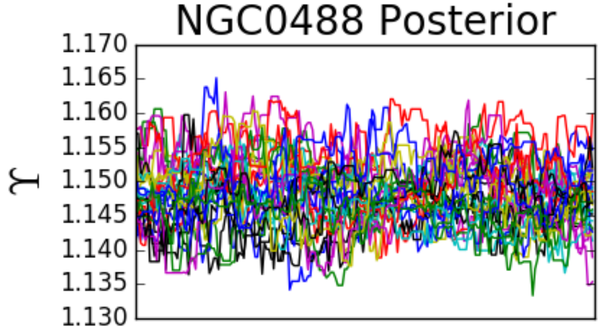
Application

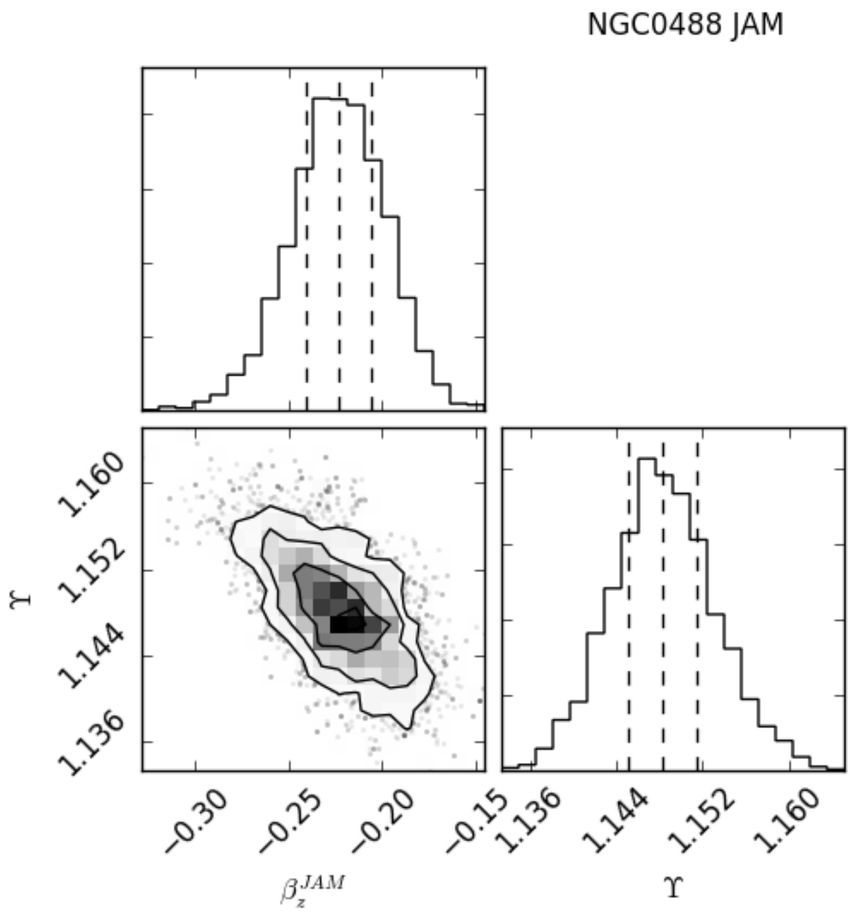
Example of MCMC+Bayesian approach









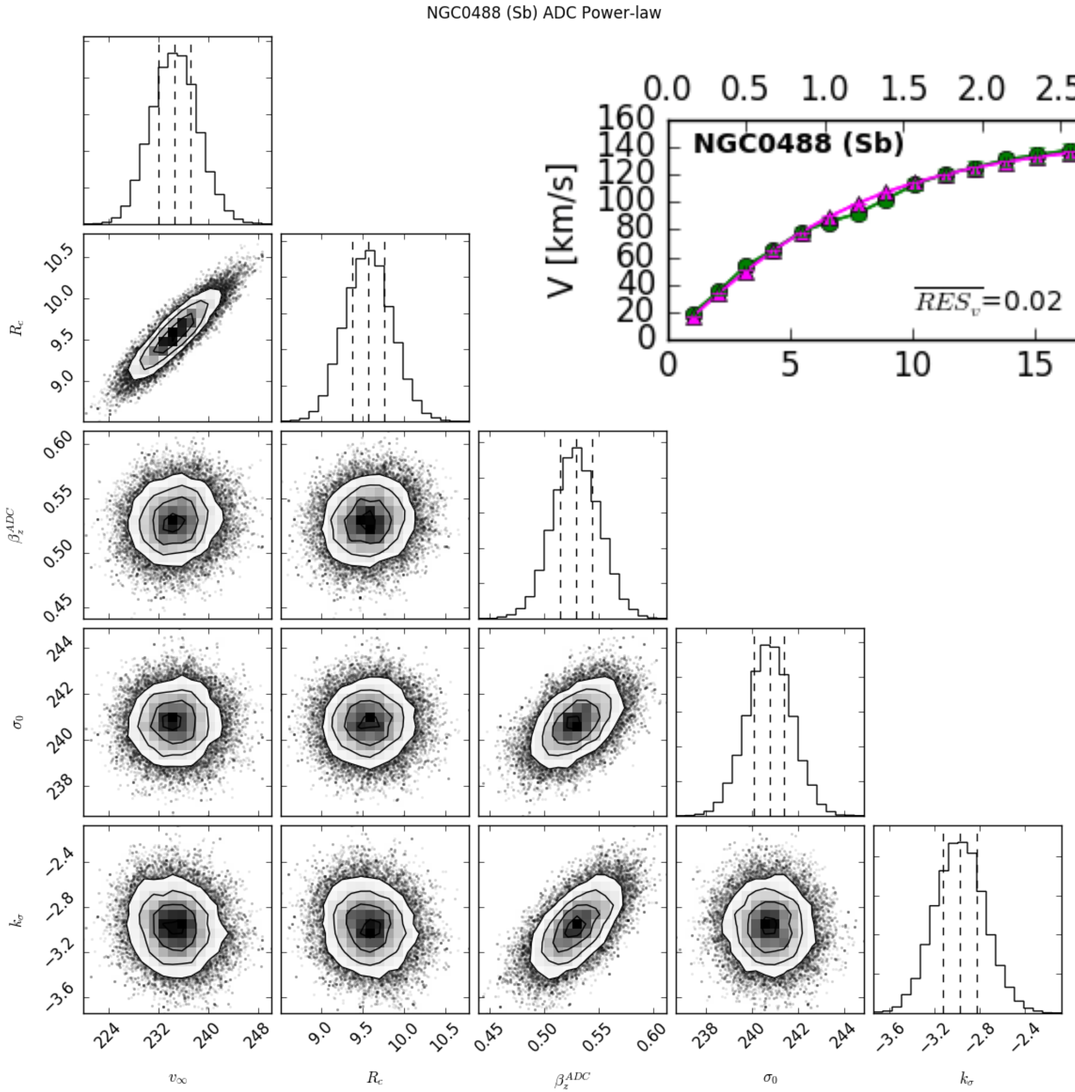


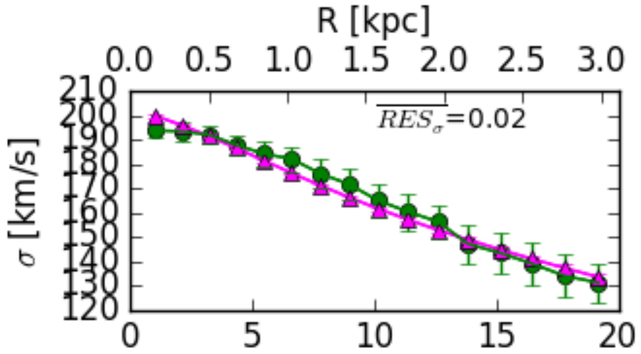
credit: Kalinova et al, MNRAS, 2017a

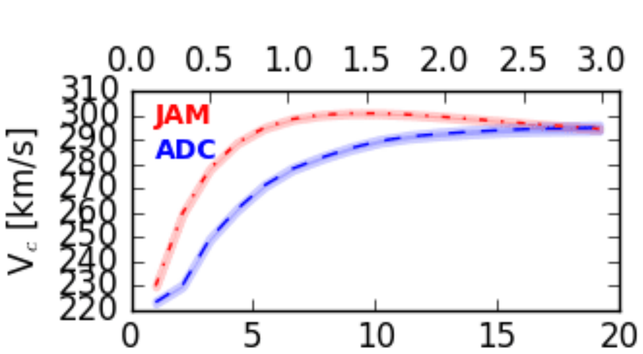
JÜLICH Forschungszentrum

Application

Obtaining velocity curves from two dynamical models







- MCMC+Bayesian method provides the opportunity to fit simultaneously multiple parameters and search for their global minimum of probability function
- robust estimation of errors for each parameter

credit: Kalinova et al, MNRAS, 2017a

Summary

- Markov Chain Monte Carlo is powerful method for models with several fitting parameters
- The likelihood maximisation function is identical to chi-squared minimisation
- Bayesian generalisation help to narrow down the probability functions of the walkers
- Metropolis-Hastlings method: walkers are keep jumping to explore the parameter space even if they already found the local minimum

Summary

- Markov Chain Monte Carlo is powerful method for models with several fitting parameters
- The likelihood maximisation fu[illegible] to chi-squared minimisation
- Bayesian generalisat[illegible] narrow down the probability functions of the walkers
- Metropolis-Hastlings method: walkers are keep jumping to explore the parameter space even if they already found the local minimum

THANK YOU!