Statistics and Probability

Bibliography

- Bayesian Data Analysis, Carlin, Stern and Rubin, CHAPMAN & HAA/CRC
- Bayesian Reasoning in Data Analysis, Giulio D'Agnostini, World Scientific.
- ICIC Data Analysis Workshop 2016, Alan Heavens Lectures.
- MACSS 2016, 2017, 2019 Lecture notes.

Statistics

- Descriptive
- Inferential:
  - Infer something from data set.
  - Test hypothesis.
  - Select a model or take decisions.

In cosmology and astrophysics, as well as in many areas, most of the problems consist of having a set of data from which we want to INFER something.

* **Infer some parameter values.**

  What is the value the parameters involved in the LCDM paradigm?

  What are the values of parameters involved in a model of COVID-19 positive cases.

* **Test an hypothesis.**

  Is the CMB consistent with a scale free initial power spectrum of fluctuations, and with a gaussian distribution?

  Is COVID-19 evolution in a exponential or a decreasing state.

* **Select a model.**

  ¿Is General Relativity the correct and final theory, or modified theories works better?

  Is a SI, SIR, SEIR, or else a better model to describe COVID-19

Probability & Inference

Typical answer

- "The ratio of the number of favorable cases to the number of all cases"
- " The ratio of the number of times the event occurs in a test series to the total number of trials in the series"
- Frequentist approach

A subjective definition

- A formal definition would be: "The quality, state, or degree of something being supported by evidence strong enough make it likely though not certain to be true"
- A simple definition: "A measure of the degree of belief that an event will occur"
- Bayesian approach.

Probability Rules

$$0 < p(x) < 1$$

Probability of event x happens is coherent

$$p(x) + p(\sim x) = 1$$

Probability that event "x" happen, and probability of event x do not happen are complementary.

$$p(x, y) = p(x|y)p(y)$$

Product rule

$$p(x) = \sum_i p(x, y_i)$$

Probability that event "x" happen, given that y happened: Marginalization

$$p(x) = \int p(x, y)dy$$

In the continuous limit we change the sum by an integral.

BAYES THEOREM arise from the these rules.

Probability function (discrete variable)

To each possible value of x we associate a degree of belief.

$$f(x) = p(X = x)$$

f(x) must satisfy the Probability rules.

Define the Cumulative distribution function ,

$$F(x_k) \equiv P(\leq x_k) = \sum_{xi \leq x_k} f(x_i)$$

CDF

with properties:

$$F(-\infty) = 0$$

$$F(\infty) = 1,$$

Also define de mean, or expected value.

$$\mu = \bar{x} = E(x) = \sum_i x_i f(x_i)$$

In general:

$$E(g(x)) = \sum_i g(x_i) f(x_i) \qquad E(aX + b) = aE(X) + b$$

- The standard deviation and Variance .

$$\sigma^2 \equiv Var(X) = \bar{X^2} - \bar{X}^2 \qquad \sigma = \sqrt(\sigma^2)$$

$$Var(aX + b) = a^2 Var(X),$$

- The mode, or the most probable value (for unimodal functions)

$$\left(\frac{df(x)}{dx}\right)_{x_m} = 0$$

Probability density function (continus variable)

The degree of believe of each value is quantified by the probability density function, pdf. $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')dx' \qquad \text{CDF}$$

$$E(X) = \int_{\infty}^{\infty} x f(x)dx$$

$$E(g(X)) = \int_{\infty}^{\infty} g(x) f(x)dx$$

$$\sigma^2 \equiv Var(X) = \bar{X}^2 - \bar{X}^2$$

Centra Limit Theorem

- The mean and variance of a linear combination of random variables is given by:

$$Y = \sum_{i=1}^{n} c_i X_i \qquad \sigma_Y^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

- CLT: The distribution of a linear combination Y will be approximately normal if the variables X_i are independent and \sigma_Y^2 is much larger than any single component c_i^2 \sigma_i^2 from a non-normally distributed X_i.

Parameter Estimation. Level 0

- Least square method.

$$a = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i\, y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$
$$= \frac{\bar{y}\left(\sum_{i=1}^n x_i^2\right) - \bar{x} \sum_{i=1}^n x_i\, y_i}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}$$
$$b = \frac{n \sum_{i=1}^n x_i\, y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$
$$= \frac{\left(\sum_{i=1}^n x_i\, y_i\right) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}$$

Where this comes from?

Minimize the residual

$$R^2 \equiv \sum [y_i - f(x_i, a_1, a_2, ..., a_n)]^2$$

assuming:

- A linear function f=ax+b.

- Errors are Gaussian and uncorrelated.

Minimization implies:

$$\frac{\partial R^2}{\partial a_i} = 0$$

Parameter Estimation, Level 1

χ^2 Minimization

$$\frac{\partial \chi^2}{\partial \theta} = 0$$

$$\chi^2 = \sum (y_i - y(x_i, \theta))^2 / \sigma_{y_i}^2$$

θ : free parameters

σ_{y_i} : variance on y_i

Homework: show explicitly that the linear least square method is derived from the minimization of the chi-square when the model is a straight line.

Parameter Estimation and optimization

$$\chi^2 = \sum (y_i - y(x_i, \theta))^2 / \sigma_{y_i}^2 \qquad \frac{\partial \chi^2}{\partial \theta} = 0$$

θ : free parameters

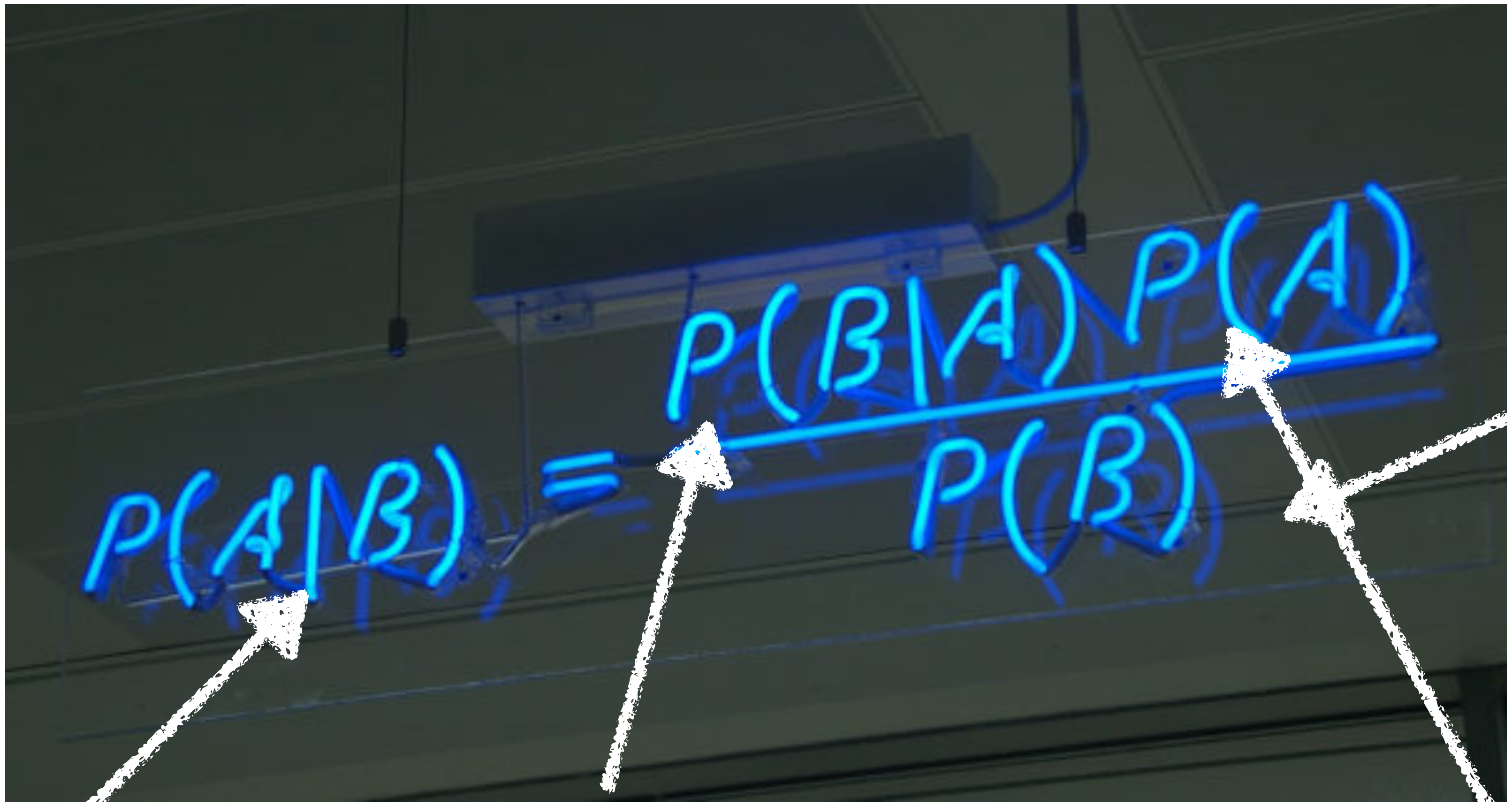
Chisq minimization becomes difficult (sometimes imposible) when the number of parameters increases...

Least square, and minimum Chi^2 methods are just special cases of Statistical Inference. This Chi^2 is a gaussian distribution if data points are independent, and errors are also Gaussian.

Bayes Theorem

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Derive Bayes theorem from p(x,y)=p(y,x)



Evidence: The probability of de data, in all possibilities

Posterior

Degree of believe that some truth proposition event B implies that A is also truth.

Likelihood

Degree of believe to which some truth proposition A, event, implies that proposition B, event, is also true.

Prior

The degree of our a priori believe of proposition A, event, based on previous knowledge.

A question in cosmology would be: Given the observed CMB data with current experiments (A), what is the probability that the matter density of the Universe is between 0.9 and 1.1 (B).

$P(0.9 < \Omega_m < 1.1|Data)$?

Example

Suppose you have the following results for an allergy test

1) It gives a positive result with probability 0.8, when patients have the allergy
2) It gives a false positive with probability 0.1.

If you make yourself the test, and result positive, what is the probability that you actually have the allergy?

P(A): probability of having the allergy

P(T): probability of test being true

What is the question in terms of probability?

We want p(A|T)

Example

Suppose you have the following results for an allergy test

1) It gives a positive result with probability 0.8, when patients have the allergy
2) It gives a false positive with probability 0.1

If you make yourself the test, and result positive, what is the probability that you actually have the allergy?

P(A): probability of having the allergy

P(T): probability of test being true

We want p(A|T)

Do we have all the information to respond it?

Example

Suppose you have the following results for an allergy test

1) It gives a positive result with probability 0.8, when patients have the allergy Sensitivity
2) It gives a false positive with probability 0.1 Specificity
3) The probability of having the allergy in the population is 0.01

If you make yourself the test, and result positive, what is the probability that you actually have the allergy?

We want $p(A|T)$

P(A): probability of having the allergy

P(T): probability of test being true

Do we have all the information to respond it?

Example (Solution)

We want p(A|T)

We know p(T|A)=0.8, p(T|~A)=0.1   , P(A)=0.01

Bayes Theorem:

$$p(A|T) = \frac{p(T|A)p(A)}{p(T)}$$

$$p(A|T) = \frac{p(T|A)p(A)}{p(T,A) + p(T,\sim A)}$$

$$p(A|T) = \frac{p(T|A)p(A)}{p(T|A)p(A) + p(T,\sim A)p(\sim A)}$$

$$p(A|T) = \frac{0.8 * 0.01}{0.8 * 0.01 + 0.1 * 0.99} = 0.075$$

Example: Efficiency of COVID-19 rapid tests.

- - The information we have about rapid COVID tests is that:
  - They have a 60% of sensitivity.
  - If you show no symptoms, but you make yourself a test and it results positive. What is the probability that you actually have COVID-19?.

Tarea: Las catafixias de Chabelo
(adaptado de The Monty Hall problem)

Hay 3 puertas. Detrás de una hay un premio. Supongamos que eliges la puerta 2. Chabelo abre la puerta 1 y allí no hay un premio. ¿Te quedas con tu elección original o cambias a la puerta 3?

Argumenta tu respuesta usando el Teorema de Bayes

Likelihood

- The probability , under the assumption of a model/theory, to observe the data as was actually obtained.

$$\mathcal{L} \longrightarrow P(\text{Data}, \text{Model})$$

- For data that can be thought as samples of a sequence of normal random variables that have a mean and a variance, the likelihood is Gaussian

$$\mathcal{L} \propto \prod_i^n \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right)$$

mu will be the expected mean given our model

- A minimization of the Chi-square correspond to the maximization of the likelihood.

Gaussian Likelihood

$$-\ln(\mathcal{L}(\vec{x},\vec{y}|\vec{\theta})) \propto \frac{1}{2}\sum_i \left(\frac{(y_i - \lambda(x_i,\vec{\theta}))^2}{\sigma_i^2}\right)$$

In terms of data points and parameters. Lambda is our model for y_i

For a Gaussian Likelihood , the minimization of the Chisq is equivalent to the maximization of the Likelihood. i.e. the best fit and the most likely model coincide.

Homework: Write down a Poisson likelihood, and identify which parameter(s) correspond to the data and which to the model.

$$L \propto \Pi_i \frac{exp^{-\lambda_i}\lambda_i^{d_i}}{d_i}$$

Bayes Theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

Probability of our parameters given the data

Probability of data given the Model

Our state of knowledge before the experiment

Probability of getting the data we observe.

For parameter inference we can omit the evidence, since is a normalization factor of the posterior. However it is crucial for model selection, it is computed as:

$$p(x) = \sum_k p(x|\theta_k)p(\theta_k) \quad \text{or} \quad p(x) = \int d\theta p(x|\theta)p(\theta)$$

Bayes Theorem

$$p(\theta, x) \propto p(x|\theta)p(\theta)$$

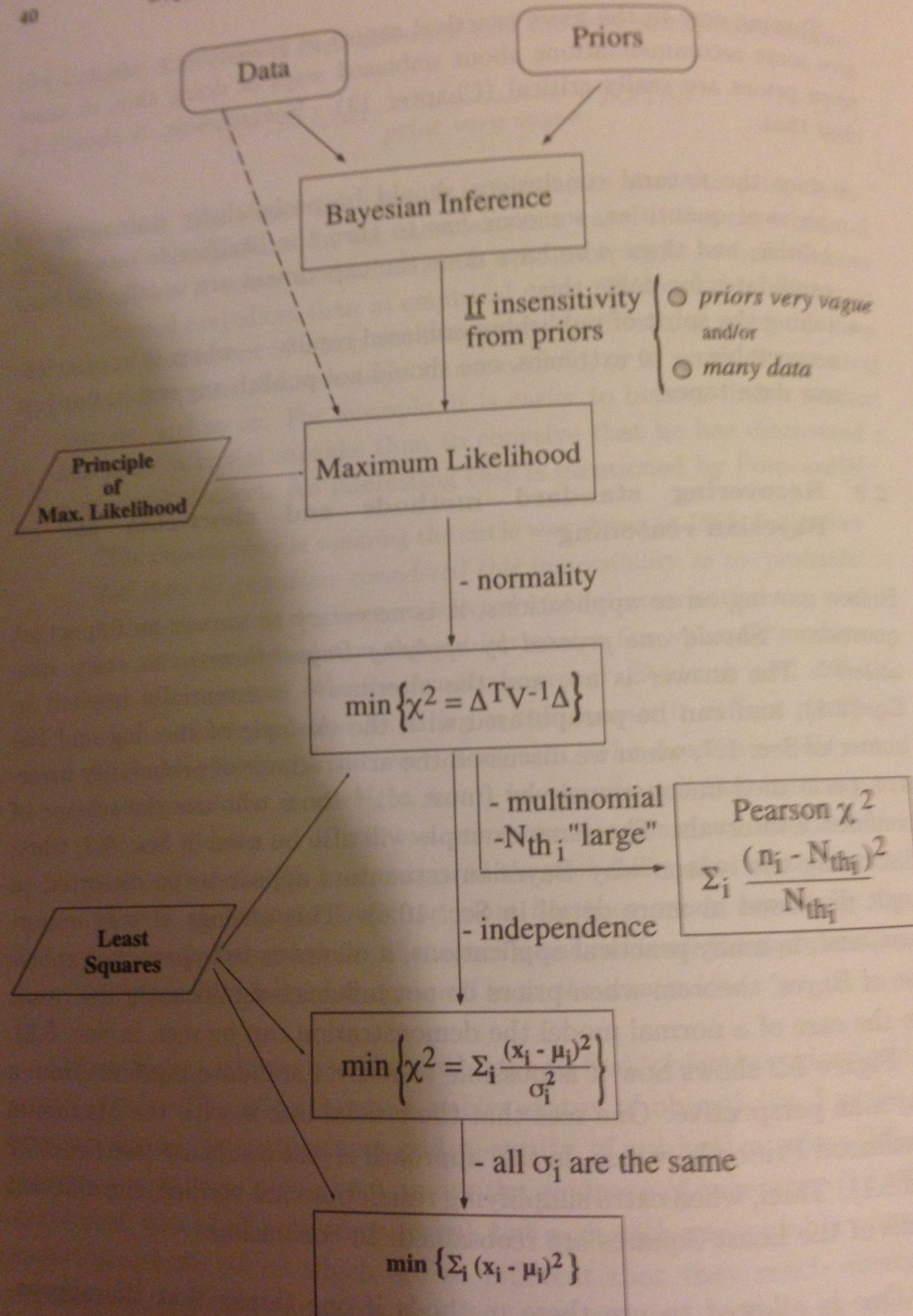
- Here we are not only worry about maximizing the likelihood, but moreover to find the posterior probability distribution, taking into account our priors.

- If Likelihood is Gaussian and data is very good, priors should be irrelevant and the posterior should also be Gaussian. Maximum Likelihood and Bayes Theorem should give the same answer. In real life this does not always happen.

- How we choose the prior? Well, this can be a flat prior if we do not know much about the parameters before the experiment. Or use the result of a previous experiment as our prior. The posterior of an experiment, can be prior for a new experiment.

The prior

- A common choice when you have no informations is to use flat priors within a given range.
- But you can use Gaussian, or another probability distribution if you have more information.
- For parameter inference is usual not worry about normalization factors in the prior, likelihood of posterior, but if we want to do model comparison, then it is more important.


Data
Priors
Bayesian Inference
If insensitivity from priors
priors very vague
and/or
many data
Principle of Max. Likelihood
Maximum Likelihood
- normality
$\min\left\{\chi^2 = \Delta^T V^{-1} \Delta\right\}$
- multinomial
-N_{th_i} "large"
Pearson χ^2
$\Sigma_i \frac{(n_i - N_{th_i})^2}{N_{th_i}}$
- independence
Least Squares
$\min\left\{\chi^2 = \Sigma_i \frac{(x_i - \mu_i)^2}{\sigma_i^2}\right\}$
- all σ_i are the same
$\min\left\{\Sigma_i (x_i - \mu_i)^2\right\}$

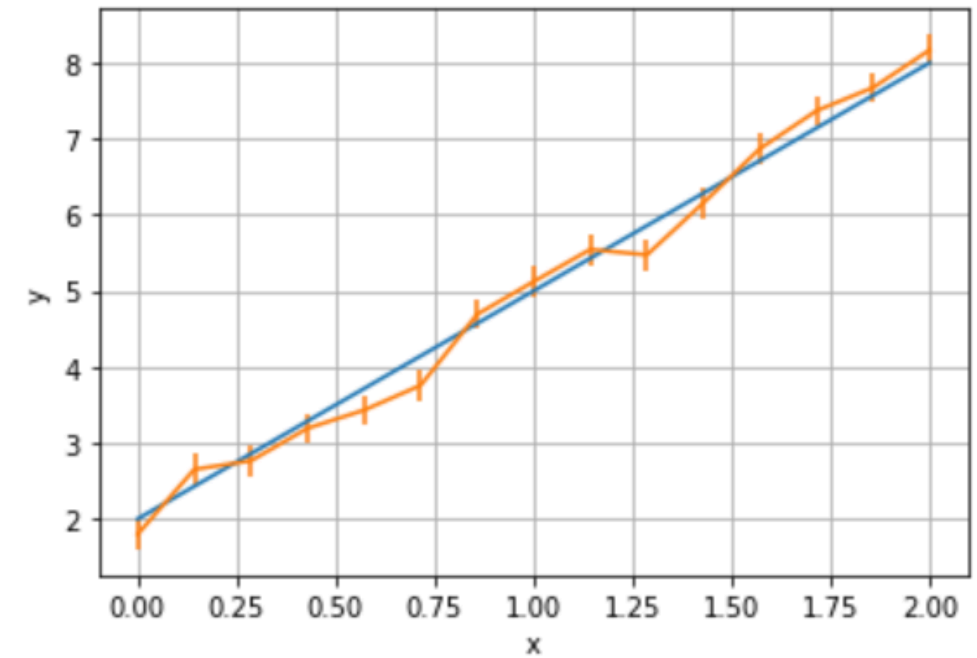
Encontrar a y b que mejor ajustan los datos

Encontrar a y b que mayor probabilidad tienen de de ser lo parámetros en mi modelo, dados los datos.

Inferencia ——>

Datos: x, y, yerr

Modelo $\mu = ax + b$



- Mínimización del χ^2

$$\chi^2 = \sum_i \frac{y_i - \lambda(x_i, \vec{\theta})}{\sigma_i^2}$$

$\lambda(x, \vec{\theta})$ ➡ $\vec{\theta} = (a, b)$

$\sigma_i = y_{err,i}$

- Minimización analítica
- Minimización Numérica

- Maximización del likelihood

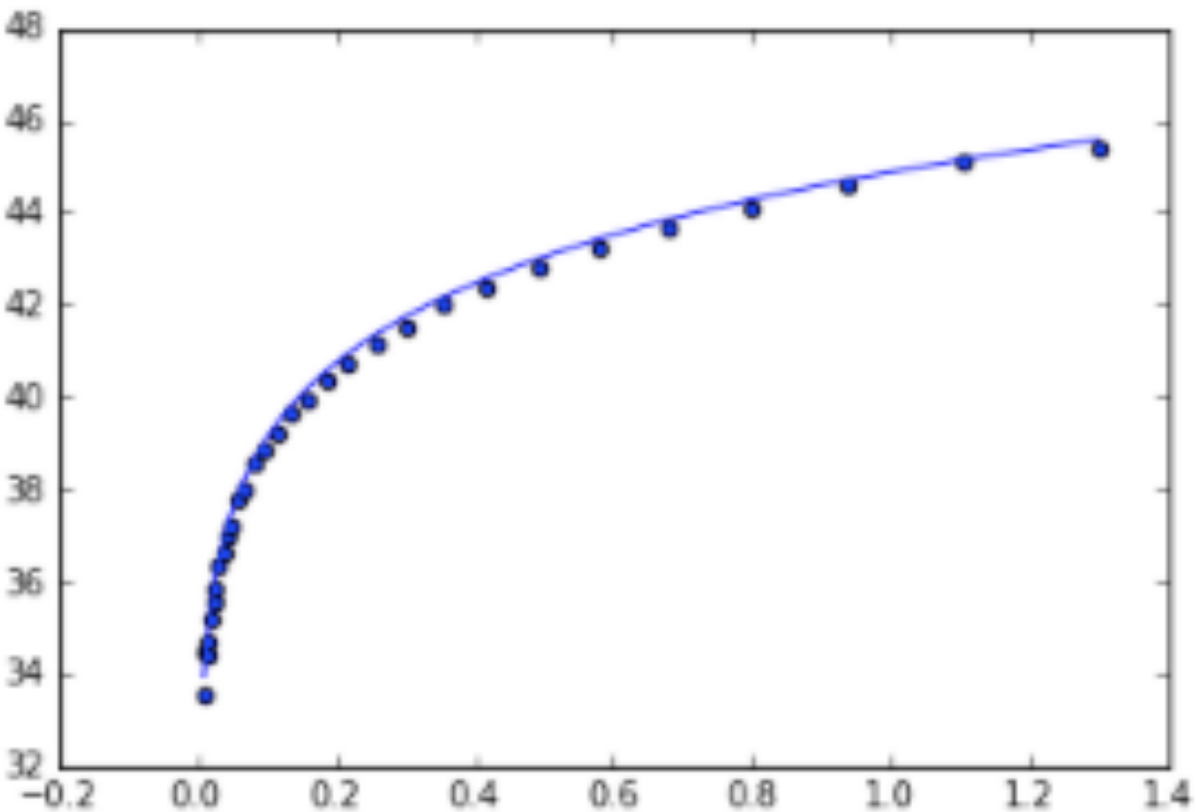
$$\mathcal{L} \propto \prod_i^n \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_i^2}\right)$$

- Maximización analítica
- Maximización Numérica
- Maximización Métodos
- Montecarlo

- Maximización del Posterior

$$P(\lambda(x, \theta) \,|\, x) \propto L(x \,|\, \lambda(x, \theta)) * Pr(\theta)$$

Ej. Find cosmological parameters with SuperNova Data (you'll work a simpler example)



$$\mu = 25 - 5Log_{10}(H_0/100) + 5Log_{10}(D_L/Mpc)$$

$$D_L = \frac{(1+z)c}{H_0\sqrt{|1-\Omega|}} S_k(r), \text{ donde,}$$

$$r(z) = \sqrt{|1-\Omega|}\int_0^z \frac{dz'}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda+(1-\Omega)(1+z')^2}} \text{Mpc}$$

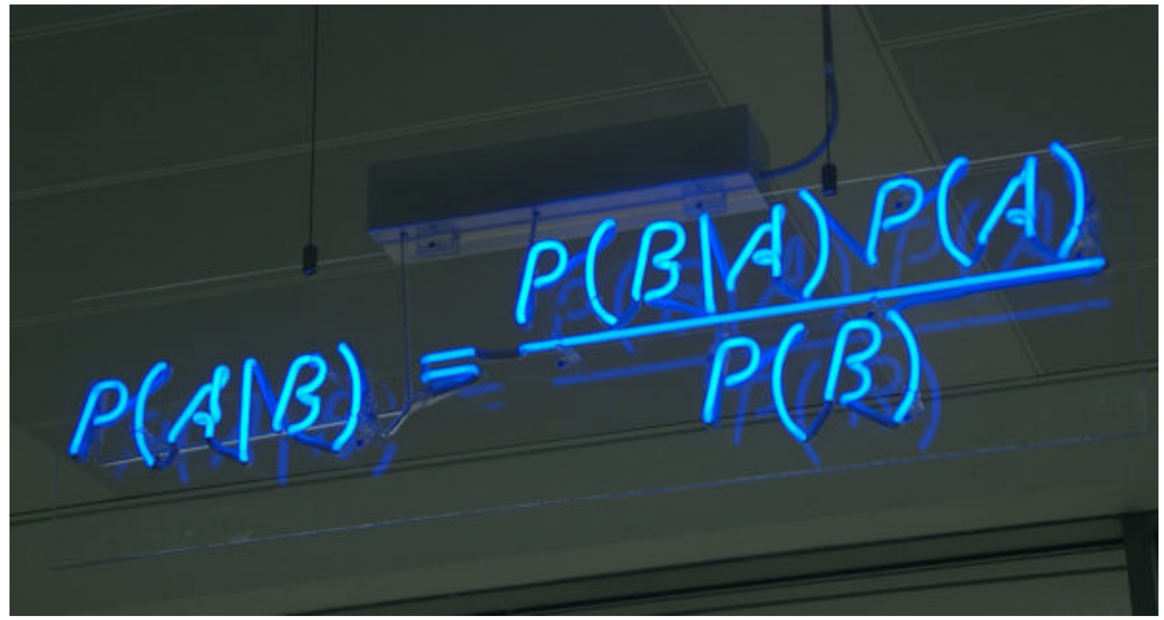
Approximate Solution

$$D_L = \frac{c}{H_0}(1+z)[\eta(1,\Omega_m) - \eta(1/(1+z),\Omega_m)]$$

$$\eta(a,\Omega_m) = 2\sqrt{s^3+1}\,[a^{-4} - 0.1540s\,a^{-3} + 0.4304s^2\,a^{-2} + 0.19097s^3a^{-1} + 0.066941s^4]^{-1/8}$$

$$s^3 = (1-\Omega_m)/\Omega_m$$

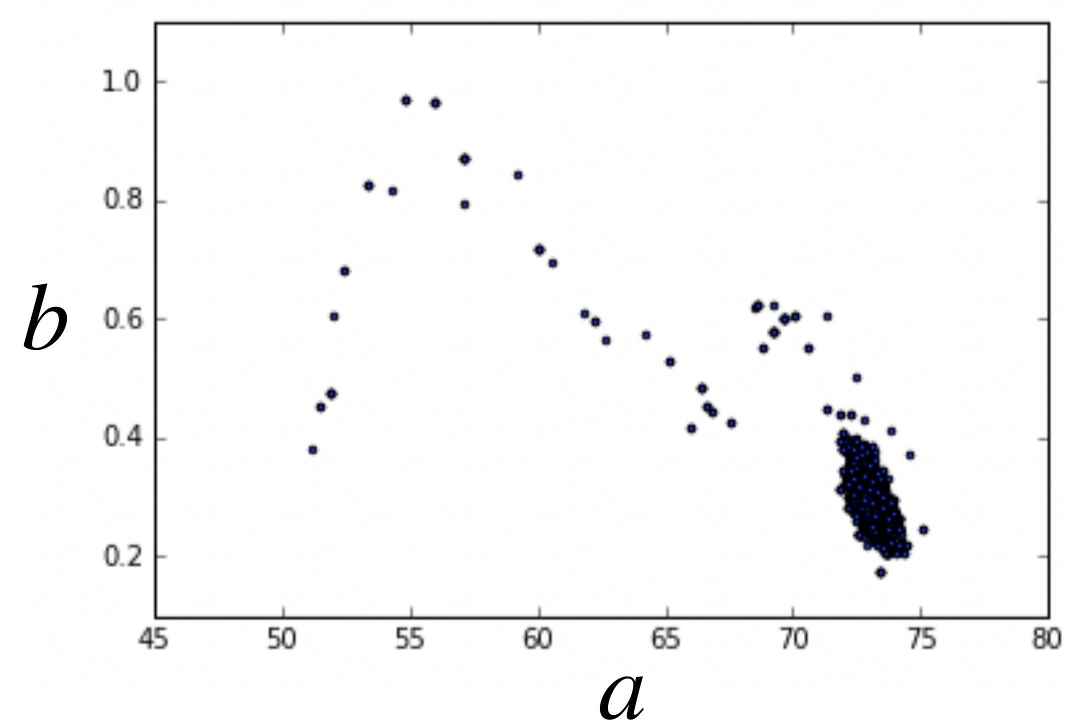
Monte Carlo Markov Chain for Bayesian inference



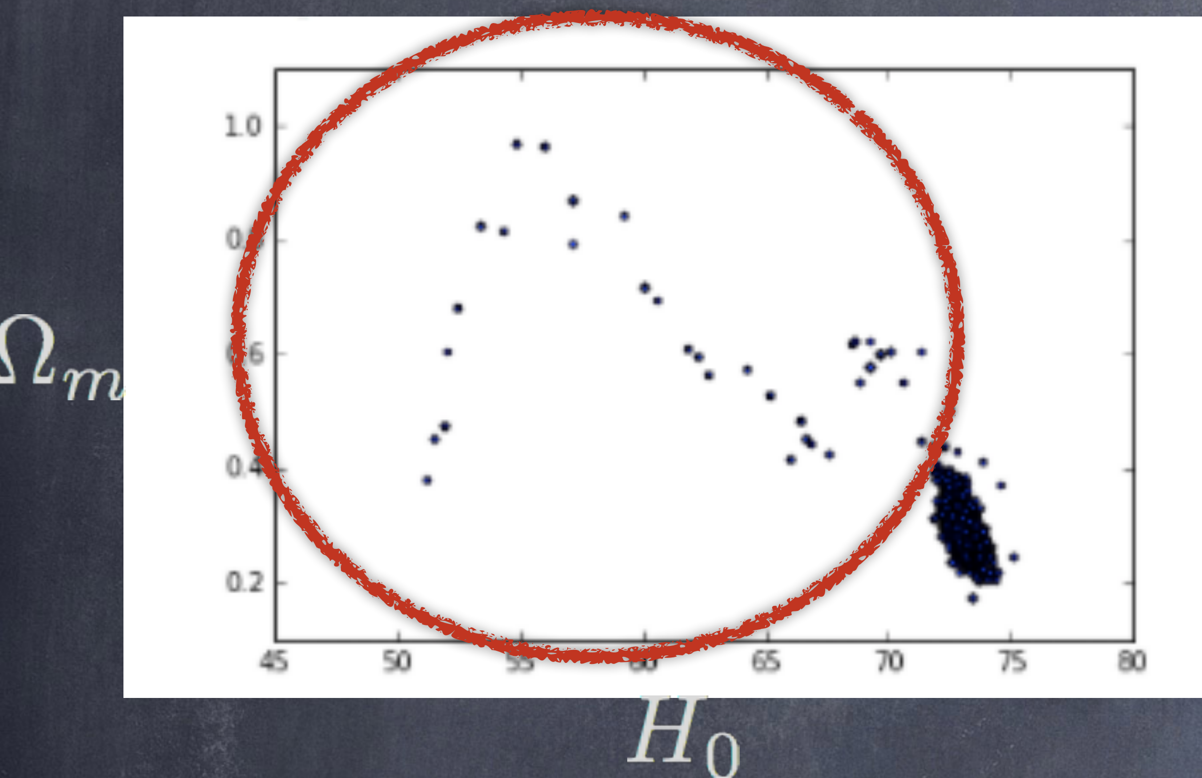
$$P(\theta \mid D) \propto L(D \mid \theta) * P(\theta)$$

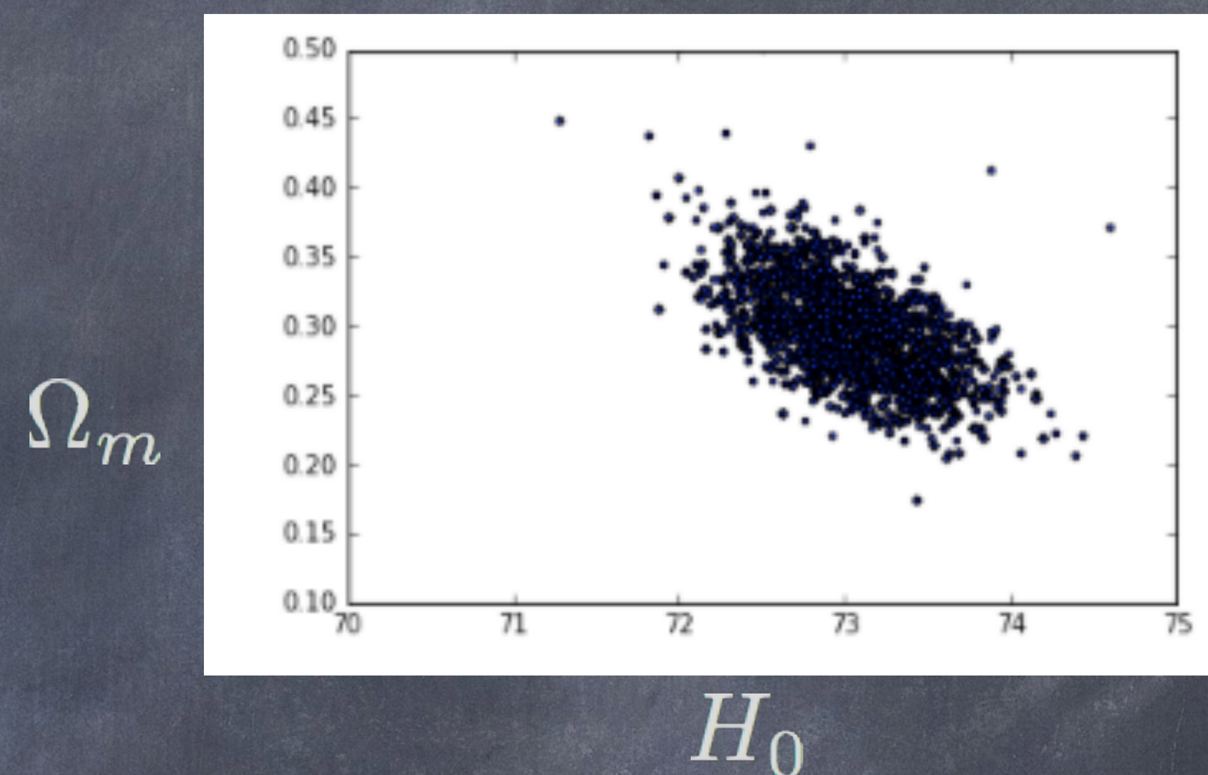
$$\ln(P(\theta \mid D)) \propto \ln(L(D \mid \theta)) + \ln(P(\theta))$$

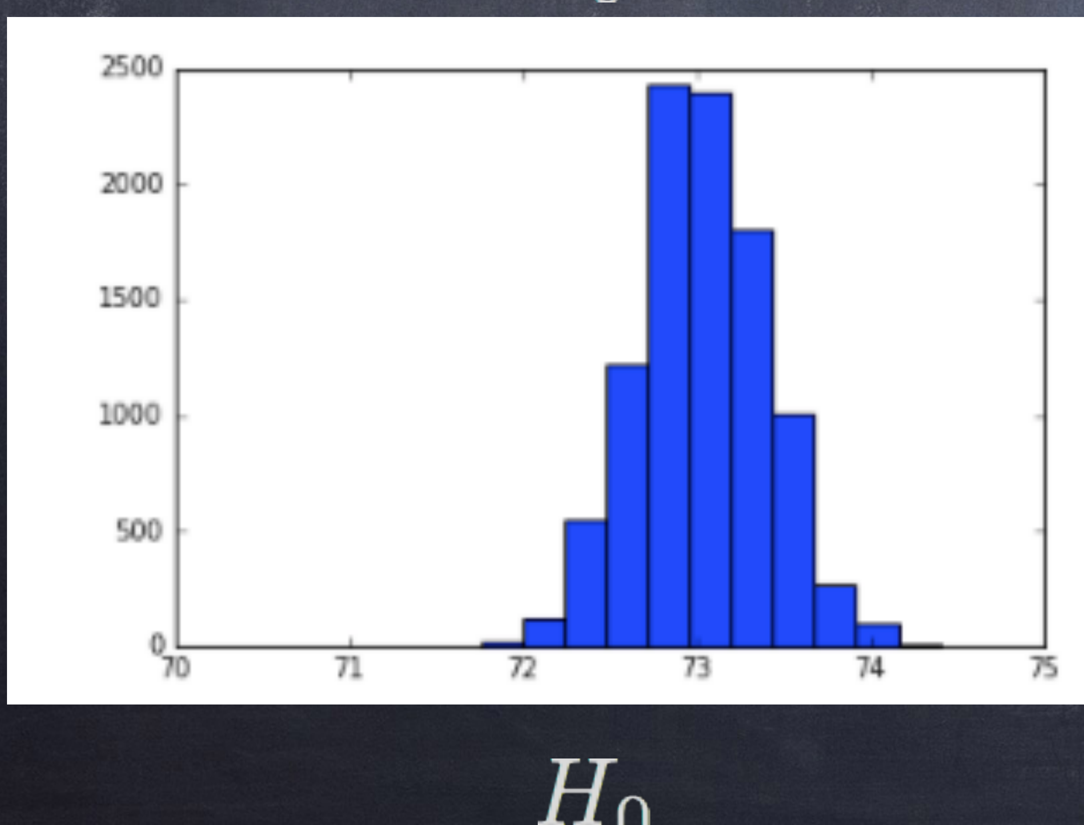
Posterior and Likelihood depends on your model assumptions as well

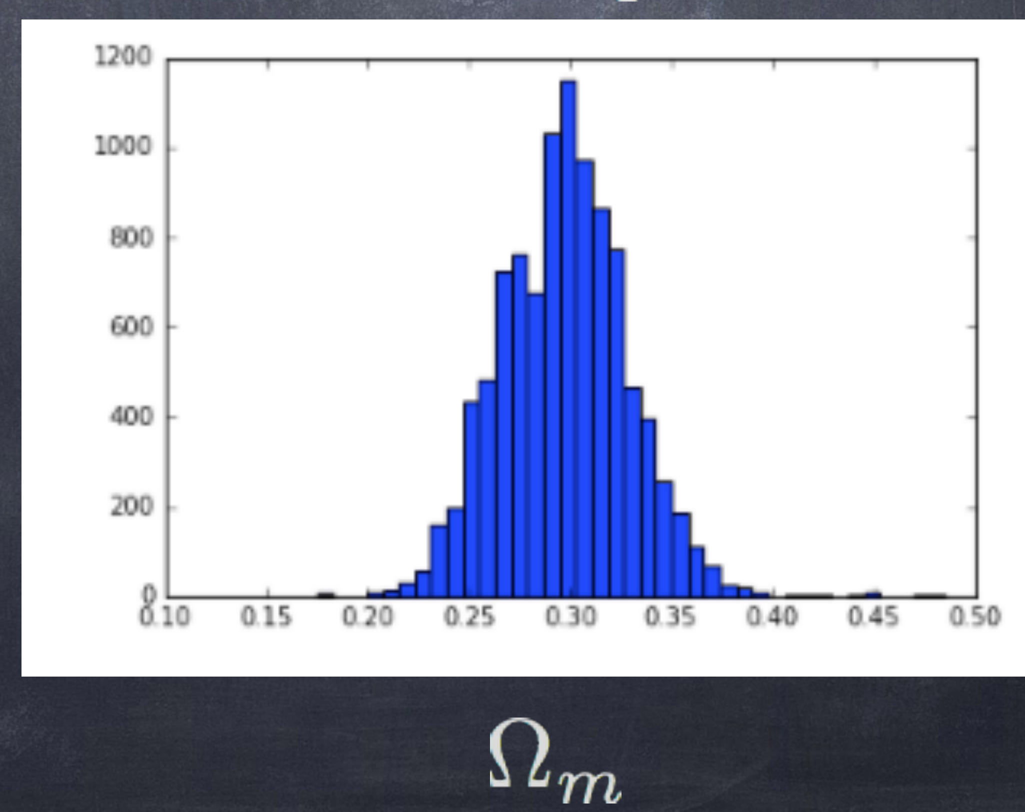


Walker

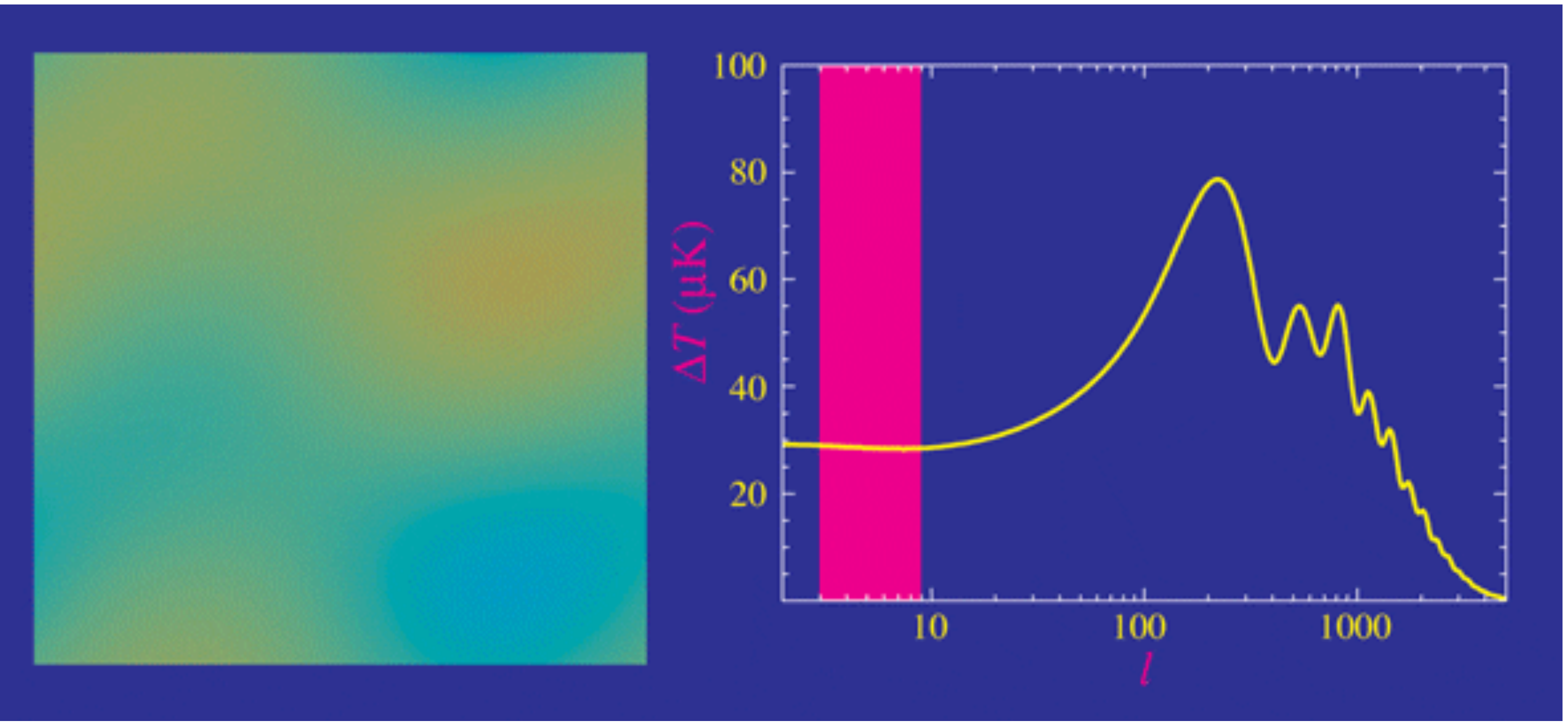








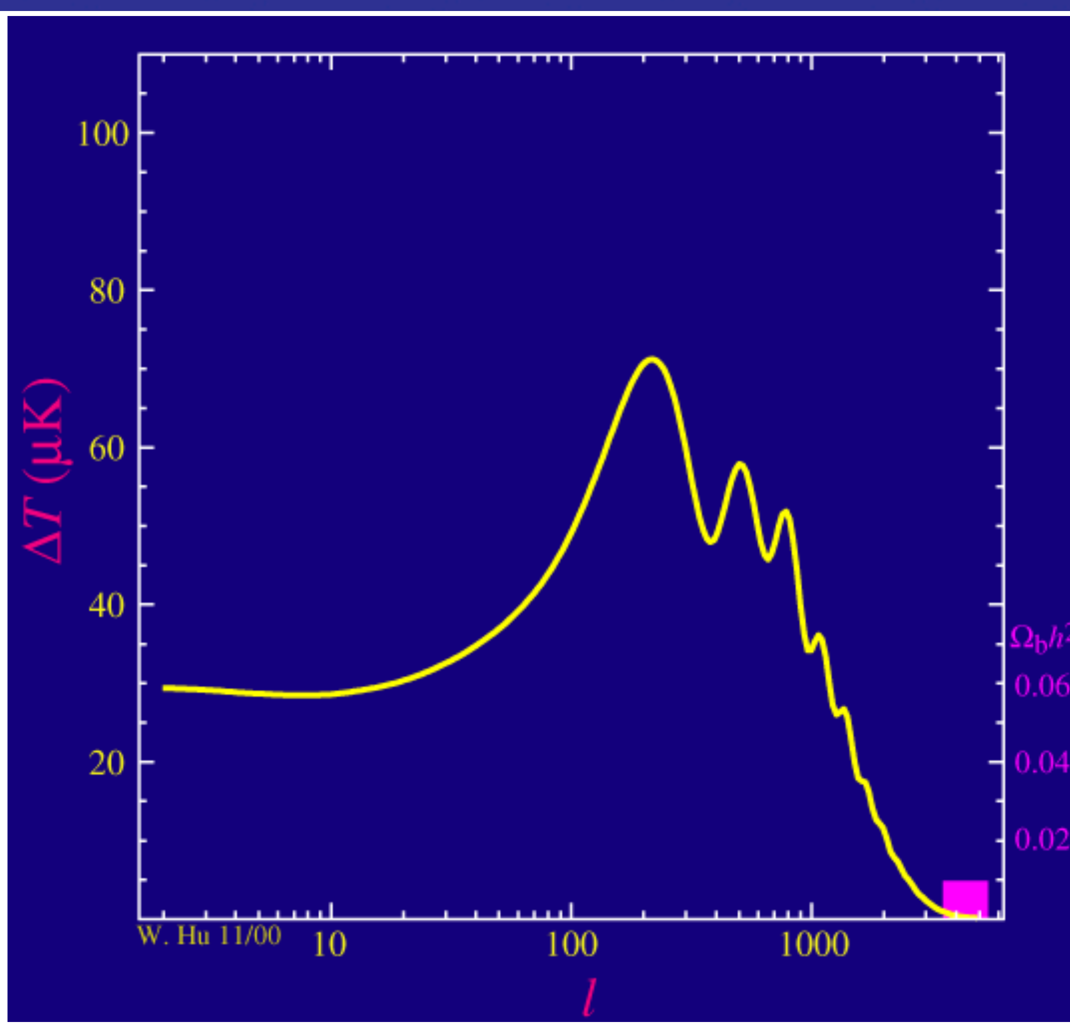
Ej. CMB DATA and Cosmological model

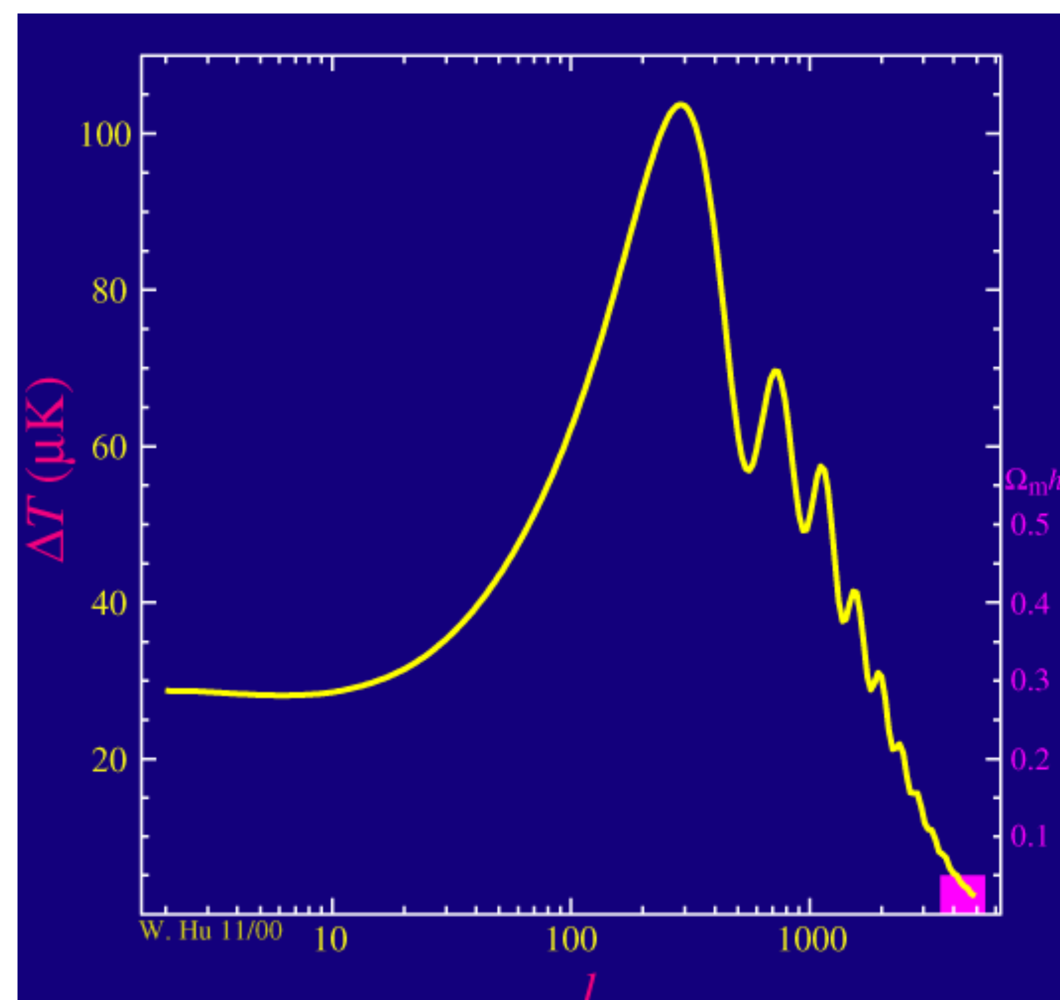


6 cosmological parameters

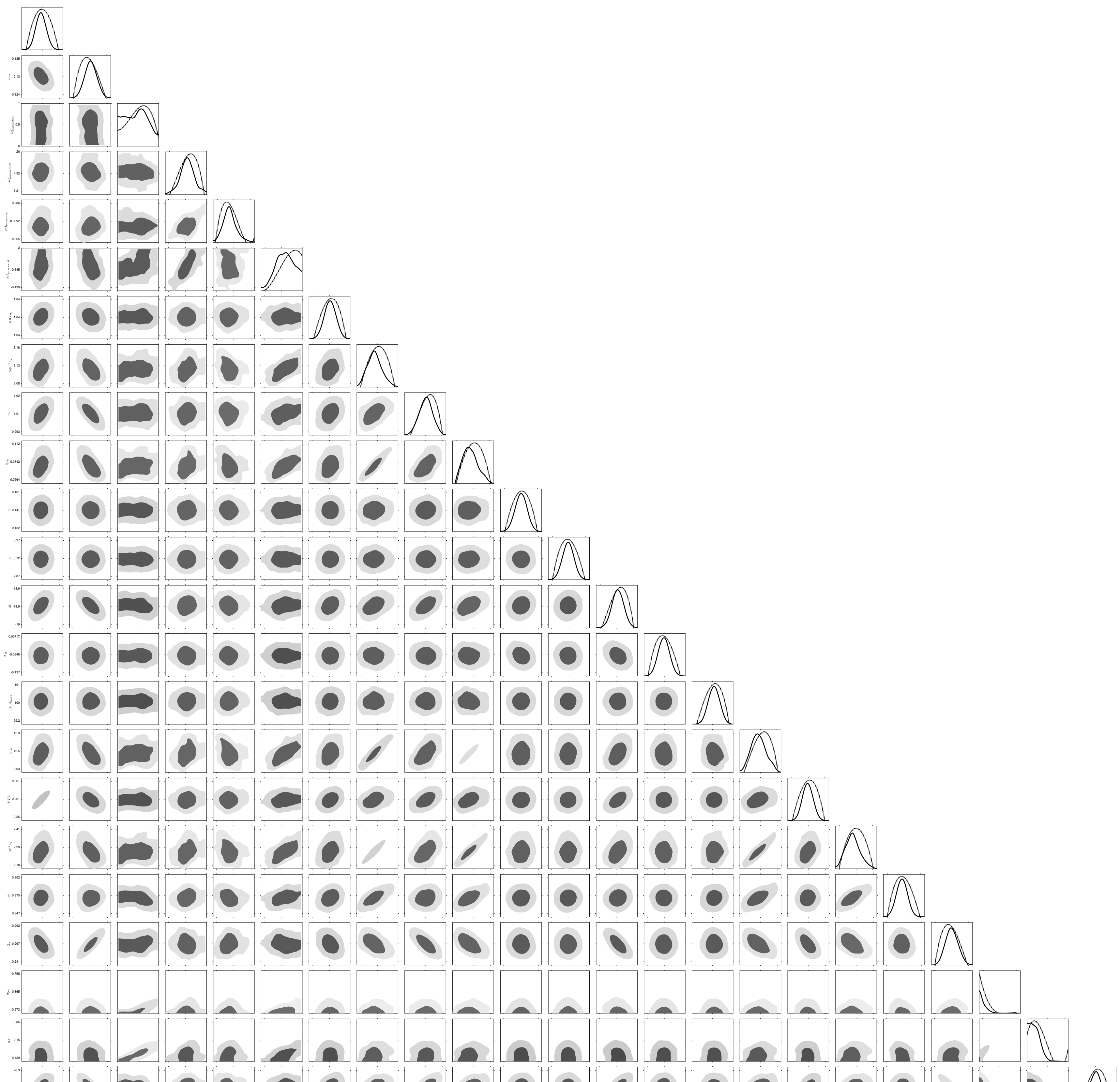
+

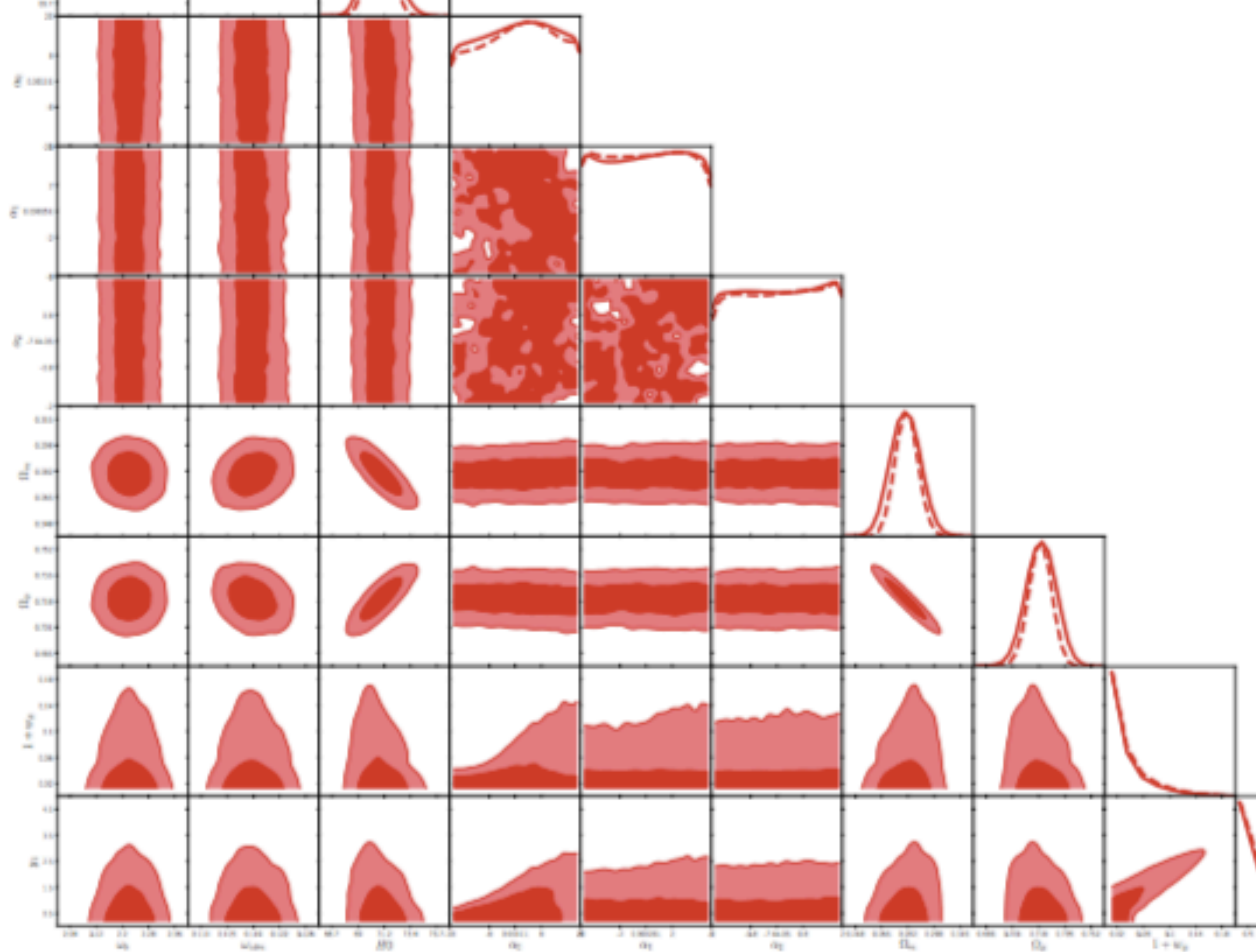
many more nuisance parameters





Ej. CMB DATA and Cosmological model





MCMC metropolis algorithm.

- Define an starting point for the parameters. Compute its posterior, and reserve it.
- Draw a sample parameter from a normal distribution centered in the starting point, and with some $\sigma 0$ dispersion.
- If the Posterior of the new sample, P(new) is higher than the intial one, P(old), we accept the sample and save it. New—>old.
- If the Posterior of the new sample is lower than the old one: draw a random number between 0-1, p(new)/p(old) is larger than such number then we accept it (New—>old), not otherwise. (Old remains). Always save the value of the parameters, regardless it was repeated.
- Draw a new sample and start again....
- After many steps, look at the resultant distribution (the chains) of parameters, i.e., the Posterior...
-

Ej: Implement it in python and make sure it works for the example of the straight line, but do it in a way it is easy to change data, model, likelihood and prior.