Cosmological Observations

Statistical Inference

step 6 - **Constrain the values of the parameters of the model**

Cosmological datasets are usually large, noisy, and with systematics → the problem to solve is not a straightforward system of equations relating precise values of an observable with a cosmological function of the parameters.

To solve the problem requires the use of random variables and the calculation of probabilities.

The standard way to **estimate the values of the model free parameters** (the cosmological parameters and the nuisance parameters) in cosmological analyses is through Bayesian inference.

Additional Bibliography:

- David MacKay - “Information theory, Inference and Learning Algorithms” (chapter 29), CUP 2003

- Hobson, Jaffe, Liddle, Mukherjee and Parkinson - “Bayesian methods in Cosmology” includes chapters on MCMC (Lewis, Bridle 2002) and model selection (Mukherjee, Parkinson, Liddle 2005)

- Amendola and Tsujikawa - “Dark Energy” (chapter 13) includes sections on Fisher matrix and analytical marginalizations

- Dan Coe - “Fisher matrices - a quick-start guide and software” 2009 arXiv:0906.4123

Forward Probability (the frequentist approach)

vs.

Inverse Probability (the Bayesian approach)

The goal is to compute the probability distribution of the data given the fixed and true value of the parameters. **Data are random variables** with a **probability density function** (pdf). Their probability corresponds to the **frequency** with which its values occur in repetitions of the experiment.

A **statistic** is computed from the data and a pdf is derived for the statistic. From values obtained for a statistic with a known pdf, a rejection level may be assigned to a **hypothesis** (a parameter value).

There is no probability distribution of the parameter values, they are absolute quantities.

Here **the vector of parameters is a random variable**, and has a probability function.
They are unobserved variables.

We want to compute that probability: a **conditional probability given the data**.

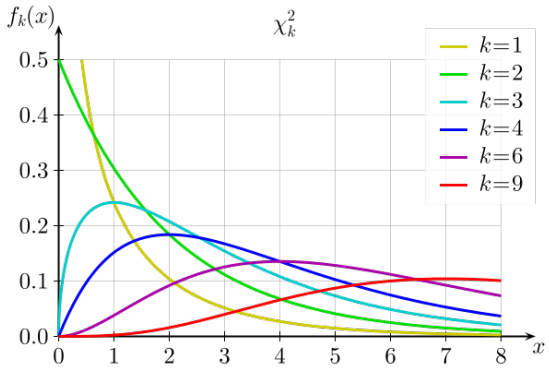
Data are also random variables and a **joint probability** may be defined: P(m,d)

Note:
d = data (the estimated physical property)
m = model (the values of the parameters)

Example : using the chi-square statistic in the frequentist approach

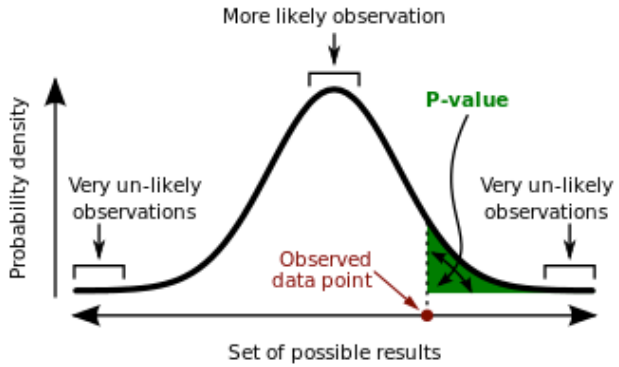
The chi-square is an example of a statistic

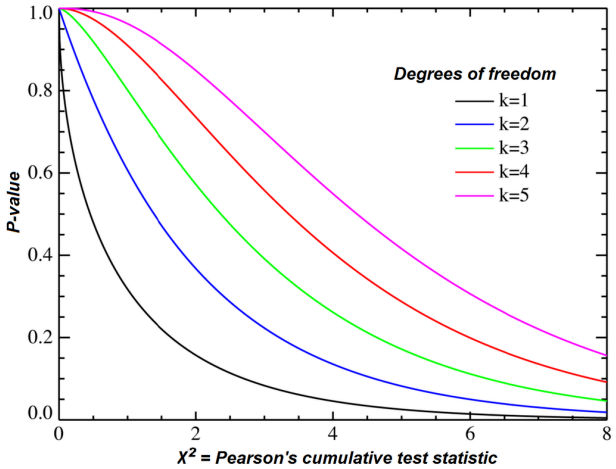
$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$



with known pdf (also called the chi-squared distribution)

Computing the chi-square value and knowing the pdf we can compute the p-value → if p_value < threshold then the data rejects the hypothesis (the fact of the parameter value being the assumed one)





Example : using the chi-square statistic in the Bayesian approach

The conditional probability of the *data given the model* is a Gaussian in the chi-square statistic.

Through Bayes theorem this implies that the conditional probability of the *model given the data* is well sampled by the values of that Gaussian.

Both methods use the chi-squared statistic, but in *frequentist hypothesis testing* the crucial information is the chi-squared distribution, while in *Bayesian parameter inference* the crucial information is the theoretical d(m) expression.

The joint probability may be written in terms of the Probability of m conditional to d (the probability of m to equal m_i, given that the data equals d_i), and the intrinsic probability of the data to be equal to d_i :

$$P(m,d) = P(m|d)\, P(d)$$

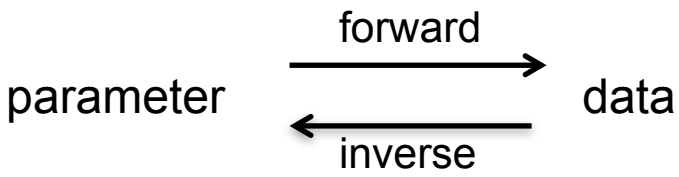
or also the other way around :

$$P(m,d) = P(d|m)\, P(m)$$

So, in Bayesian Inference we consider 2 spaces:

the **data space** (where random variables d live) $P(d|m)$

the **parameter space** (where random variables m live) $P(m|d)$



The two spaces are related through **Bayes theorem**, which we can obtain by equating the two expressions:

$$P(m|d) = \frac{P(d|m)\,P(m)}{P(d)}$$

P(m|d) is the probability of the parameter values given the data. It is a distribution in the parameter space.
It is known as the **posterior** distribution → this is what we want to get.

P(m) is the probability of the parameter values independently of these data → it can be something we know beforehand from another experiment, or from some intrinsic property of the model, or it may just be flat (no special restriction on that parameter).
It is known as the **prior**.

P(d|m) is the probability of getting the measured data given the parameter values. It is a distribution in the data space.

Remember, in inverse probability we do not want to study the properties of the data space but rather we want to use P(d|m) to compute/infer P(m|d).

Now, it is reasonable to assume that if the probability of the observed data, given a parameter value, is low (high), then that value is unlikely (likely) to occur.

For these two reasons P(d|m) is named the likelihood of the *parameters,* L(m), even though it is a quantity in the *data* space.

P(d) is the probability of the data independently of the parameter values.
It may be obtained from the joint probability by integrating over the full range of parameter values, i.e., **marginalizing** over *all* parameters.

$$P(d) = \int_m P(d|m)\, P(m)$$

Being independent of m, it is a normalization constant for P(m|d), in Bayes theorem.

Note that it is independent of the parameter values, but not on the modeling, and its value may be used as a criteria for model comparison.
For this reason it is known as the **evidence**.

Note that for any **parametrization/theory/model** (e.g.: Ω_m , Ω_Λ),
(e.g. Ω_m , Ω_Λ , Ω_v) the whole universe of possible **models/parameter values** has a total probability of 1.

Thus when working within one case, P(m|d) may be renormalized to 1 and the evidence is not needed. But when comparing two cases, the absolute value has valuable information: the highest absolute value is the preferred case, hence the name → there is highest evidence for that case.

The data space

We know many things about the data space:

- we have a sample of the distribution there (the measured data)

- we know **moments of the distribution** - the mean, the variance-covariances (either from computing the average and dispersion of the measured sample, or computing from theory e.g. d(m))

For most practical applications, data is large (even one measurement of SN magnitude involves a large number of photons) and the central limit theorem tells us that the full distribution (for which we just have a sample) must be a Gaussian.

$$P(d|m) = L(m) = \frac{1}{(2\pi)^n |C|^{1/2}} \exp\left[-\frac{1}{2}(\bar{d} - d(m))C^{-1}(\bar{d} - d(m))^T\right]$$

ex: 2D

$$P(d|m) = A \exp\left[-\frac{1}{2}\frac{1}{(1-\rho^2)}\left(\frac{\bar{d}_1 - d_1(m)}{\sigma_1^2}\right)^2 + \left(\frac{\bar{d}_2 - d_2(m)}{\sigma_2^2}\right)^2 - 2\rho\frac{(\bar{d}_1 - d_1(m))(\bar{d}_2 - d_2(m))}{\sigma_1\sigma_2}\right]$$

data is 2D $\quad [C] : 2\times 2 = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ Covariance matrix

Correlation matrix is the covariance normalized by each sigma, such as to leave only the correlation information but not the absolute values

$$= \begin{bmatrix} \frac{\sigma_{11}}{\sigma_{11}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{\sigma_{22}}{\sigma_{22}} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Notation: 2 indices mean a square quantity

σ_{11} is σ^2

σ_1 is $\sqrt{\sigma^2} = \sigma$.

Correlation $\rho = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} = \dfrac{\sigma_{12}}{\sigma_1 \sigma_2}$

Note that each data point d_i is usually obtained from many measurements.

It is a random variable with mean $\bar{d}_i$ and variance σ_{ii} (std error $\sqrt{\sigma_{ii}} = \sigma_i$)

Ex: SN $d_i = \mu_{z_i}$

CMB $d_i = C_{\ell_i}$

The full $\underline{d}$ vector may have very large dimension.

Ex: SN in 100 redshift bins

$\mu_{z=0.1}$, $\mu_{z=0.15}$, ... $[C] = 100 \times 100$

CMB in 1000 ℓ-modes $[C] = 1000 \times 1000$

$C_1, C_2 \ldots C_{100} \ldots C_{1000}$

the number of correlations ℓ is: $\frac{n^2 - n}{2} = \frac{n(n-1)}{2}$

(n^2: total number; n: diagonal; 2: symmetry)

So for SN we write:

$$P(d|m) = A \exp\left\{-\frac{1}{2}\left[\mu_{obs} - (5\log_{10} D_L(m) + 25)\right]^T C^{-1}\left[\mu_{obs} - (5\log_{10} D_L(m) + 25]\right]\right\}$$

For N redshift bins C is a $N \times N$ matrix, with N variances and $\frac{N(N-1)}{2}$ covariances

and μ and D_L are N-dimensional vectors

The parameter space

From Bayes theorem we can compute the posterior from the likelihood, if we know the prior and if we renormalize the evidence.

Notice that a Gaussian likelihood does not necessarily lead to a Gaussian posterior (even in the case of a flat prior), because changing from one space to the other involves an inversion d(m) → m(d)

Only in the case that the response of the observable is linear in the parameter values, will the posterior also be a Gaussian.

→ this is the case for geometrical probes D(z;m), i.e., for SN, BAO, but not for structure formation probes P(k;m).

Assuming Gaussian,

$$P(m|d) = A \exp\left[-\frac{1}{2}\left(\bar{m} - m(d)\right)^T C^{-1} \left(\bar{m} - m(d)\right)\right]$$

Now, the dimension of C is the dimension of the parameter space

$[C] = p \times p \rightarrow$ number of parameters, not related with number of redshift bins or ℓ-modes.

Usually it is much lower than the dimension of the data space.

In 2D it has also the general form

$$C = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where 1, 2 are for example Ω_m, Ω_Λ

and not z_1, z_2

Analogous to the data space, we could say that one value of the data vector will imply one value of the parameter vector m, which compared to the true (or fiducial) parameter vector $\bar{m}$, will give a value of $P(m|d)$, i.e., the likelihood of the data.

We usually do not say "the likelihood of the data" because the data is the quantity measured ("known") and not the model.

To compute $P(m|d)$ exactly and analytically we would need to write

$$m(d)$$

This is usually impossible to do.

A possible way forward is to notice that $\frac{\partial^2 \ln -P(m|d)}{\partial m^2}\Big|_{m_0} = C^{-1}_{pp}$

and use Bayes theorem to relate $\frac{\partial^2 \ln -P(m|d)}{\partial m^2}\Big|_{m_0}$ with $\frac{\partial^2 \ln -P(d|m)}{\partial d^2}\Big|_{d} = C^{-1}_{NN}$

To estimate the parameter values and their uncertainties from data, we need to find their distribution in the parameters' space, i.e., the posterior distribution (or its moments, since in practice we do not need the full distribution).

There are two general ways of doing this:

- **Sampling the distribution** - we can get a sample of the posterior distribution by direct computation of the likelihood on a **grid**, or by using stochastic methods (**Monte Carlo**)

- **Fisher matrix** - we can compute a lower limit for the second-order moments of the posterior distribution (i.e., the variance of the parameters) in a deterministic way. However we cannot compute the first-order moment in a similar way (i.e, the values of the parameters).

Sampling the distribution: in a deterministic way

Grid

Since the likelihood is proportional to the posterior distribution that we want to find, a direct way to sample it is to compute its values at some points in the parameter space:

compute the likelihood in a **grid** - a hypercube of likelihoods.

This does not give us a sample distributed as the posterior, but just give us some values of that function.

Disadvantages:

- the resolution of the grid may be too low to make contours,
- will waste time computing in low likelihood places,
- the number of required points increase fast with dimension of the grid

Maximization means to reduce a dimension by fixing it in the grid.

Marginalization means to reduce a dimension by summing along it on the grid.

Analytical marginalization (over nuisance parameters)

One way to decrease the dimension of the problem, making it possible to compute a grid of lower dimension is to marginalize the likelihood in advance,

i.e., to integrate the likelihood dependence on one or more parameters, obtaining a new likelihood with less dimensions. This should mainly be done to parameters we are not interested in.

Let us consider a data vector x_i (for example the distance modulus measurements at various redshifts, with associated error bars σ_i), and the theoretical vector m_i (for example the distance modulus computed for the same redshifts, which is a function of the values of the cosmological parameters).

The Gaussian likelihood of a theoretical model given the data vector is:

$$L(p) = N \exp\left[-\frac{1}{2}\sum_z \frac{(\bar{d}_z - d_z(p))^2}{\sigma_z^2}\right]$$

Marginalization with an additive bias

Now, consider that there is a systematic effect contributing to the distance modulus in an additive way, parameterized by a parameter α.

Therefore the theoretical prediction, now including that effect, is:

$d_i \rightarrow d_i + \alpha$

This means that the theoretical model that will be applied to fit the data gets an extra parameter: $d(p_1, \ldots, p_n) + \alpha$

We need to estimate the cosmological parameters (p_i) in the presence of α, i.e., allowing for all possible values of α.

Instead of building a N+1 dimension grid (and since we are not interesting in estimating α, but only in including its impact on the estimation of p_i), we can marginalize a priori over all possible values of α.

Let us then marginalize over a generic additive parameter:

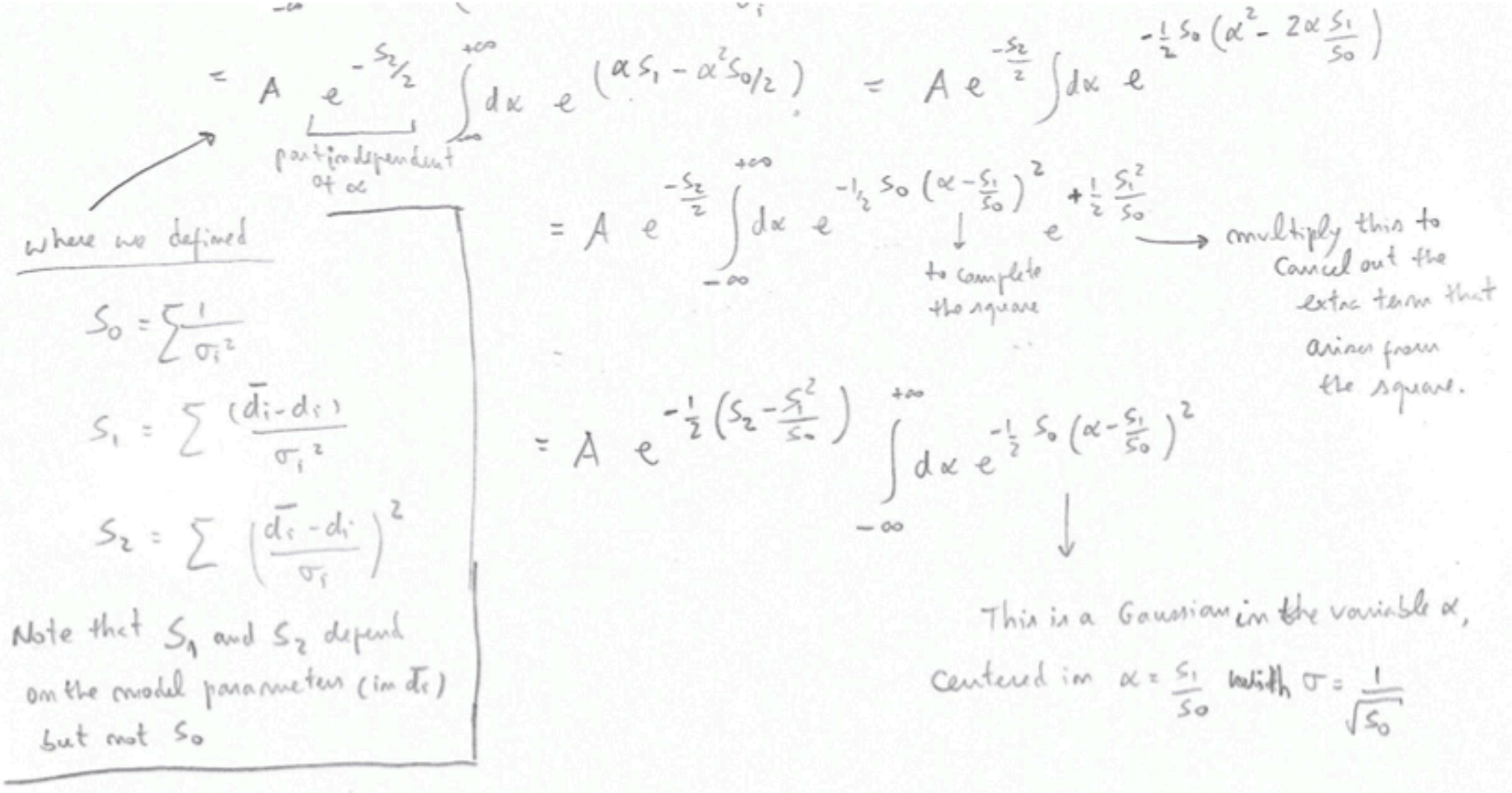
notation
- observed : $\bar{d}_i$
- theoretical : $d_i(m) = d_i + \alpha$

To marginalize is to get the constraint on the parameter (sub)vector p, assuming all possible values of α, i.e., to integrate the likelihood over $\underline{\alpha}$, obtaining a new expression for the likelihood.

So, $$L(m) = A \exp\left[-\frac{1}{2}\sum_i \frac{(\bar{d}_i - d_i - \alpha)^2}{\sigma_i^2}\right]$$

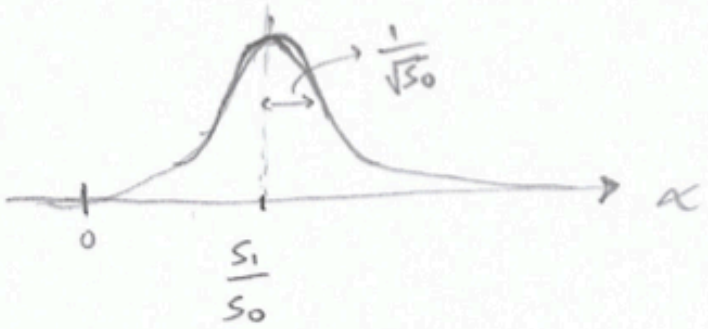
Marginalize over α $\rightarrow$ $$L_{New}(m) = A \int d\alpha \; L(m) =$$

$$= A \int_{-\infty}^{+\infty} d\alpha \; \exp\left(-\frac{1}{2} \sum_i \frac{(\bar{d}_i - d_i)^2 + \alpha^2 - 2\alpha(\bar{d}_i - d_i)}{\sigma_i^2}\right) =$$



Notice that the result of this integral only depends on the width of the Gaussian ($S_0^{-1/2}$) and not on its central point S_1/S_0.

So it is just a constant, i.e., it is independent on the cosmological parameters contained in S_1 and S_2 .



The integral $[-\infty, +\infty]$ of a Gaussian does not depend on the central point, only on the width:

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(x-x_0)^2} dx = \sqrt{(2\pi)^m \det C} = 2\pi\sigma = \frac{2\pi}{\sqrt{S_0}}$$

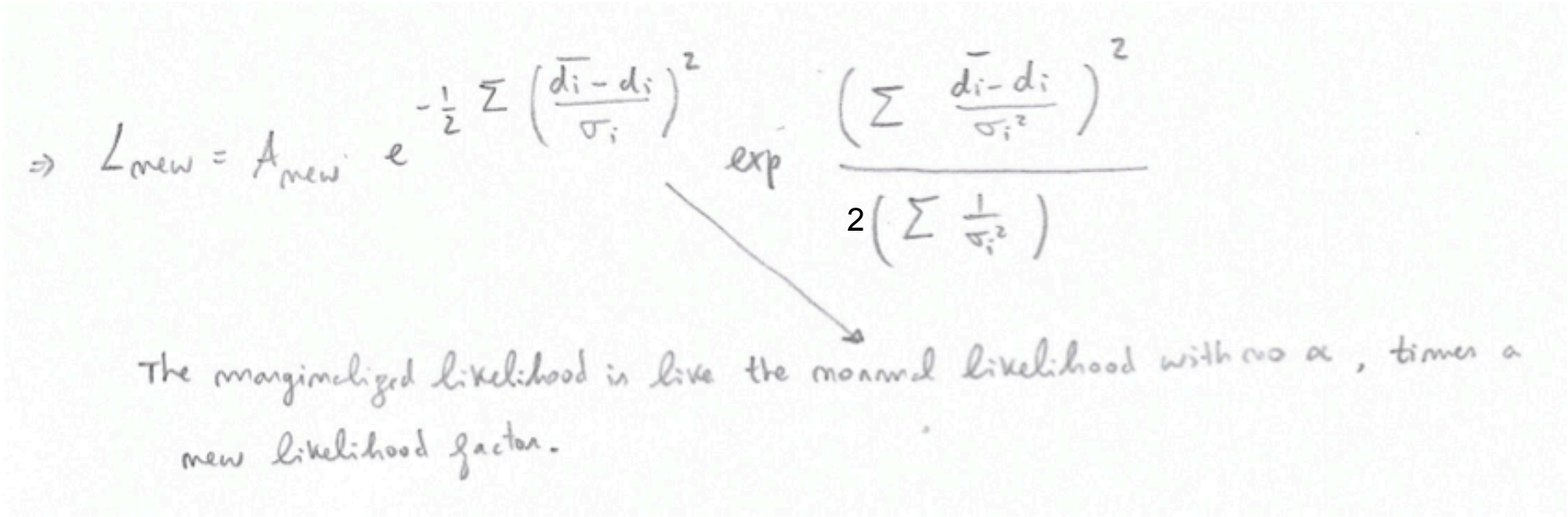
→ For 1 parameter (α) $m=1$, $\det C = \sigma = \frac{1}{\sqrt{S_0}}$

$\Rightarrow$ The new likelihood (marginalized over α) is:

$$L_{new} = A_{new}\ e^{-\frac{1}{2}\left(S_2 - \frac{S_1^2}{S_0}\right)}$$

A_{new} → new constant of normalization

$$A_{new} = \frac{1}{\sqrt{(2\pi)^m S_0 \det C}}$$



The values of this likelihood on the points of a grid in the n-dimensional space of the cosmological parameters ($p_1 \ldots p_{n)}$, are identical to the ones that would be obtained by first computing the original likelihood on the points of a grid in the (n+1)-dim space of the cosmological parameters ($p_1 \ldots p_n$, α) , and then summing up all the likelihood values along the α dimension on each p (dim n) point → so this method only requires an n-dimension grid, instead of an (n+1)-dimension one.

Marginalization with a multiplicative bias

If the systematic effect contributes to the distance modulus in a multiplicative way, parameterized by an α parameter, the (n+1)-dim likelihood is:

$$L(p,\alpha) = N \exp\left[-\frac{1}{2}\sum_z \frac{(\bar{d}_z - \alpha d_z(p))^2}{\sigma_z^2}\right]$$

Marginalizing over α, (using the same approach as in the previous calculation) the n-dim likelihood becomes:

$$L_2(\theta) = N_2 \exp\left[-\frac{1}{2}\left(\ln S_{02} - \frac{S_{11}^2}{S_{02}}\right)\right]$$

where $S_{ab} = \sum_z \frac{\bar{d}_z^a d_z(p)^b}{\sigma_z^2}$

Note that the result depends on the way the effect is included in the modelling.

If the multiplicative bias parameter is applied to correct the data (instead of being included in the theoretical modelling), the (n+1)-dim likelihood is written as

$$L(p,\alpha) = N \exp\left[-\frac{1}{2}\sum_z \frac{(\alpha \bar{d}_z - d_z(p))^2}{\sigma_i^2}\right]$$

In this case, after marginalizing over α the resulting n-dim likelihood obtained is different.

It is given by

$$L_1(\theta) = N_1 \exp\left[-\frac{1}{2}\left(S_{02} - \frac{S_{11}^2}{S_{20}}\right)\right]$$

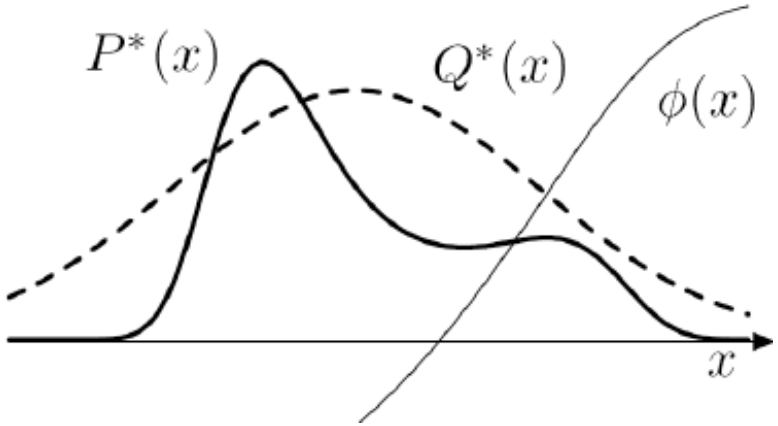
Sampling the distribution: in a stochastic way

It a way is to get a **sample of the distribution**, i.e., a group of points in the parameter space in the same proportion as in the full distribution - and not just to know the values of the probability at certain points of the space (as with the grid).

Importance Sampling

One possible way to do this is to sample from a known distribution (Q), that we assume will be similar to our target distribution (P).

(for example Q may be a smooth approximation of P)



We need to define weights to get a true sample of P from a sample of Q.

Introduce weights

$$w_r = \frac{P^*(\mathbf{x}^{(r)})}{Q^*(\mathbf{x}^{(r)})}$$

This is impossible if we know nothing about P.

But if we suspect it may be similar to Q then this method is very useful, because we do not need to generate any points for P.

We just need to get the Q points and change their weights - for example computing the likelihood of those points (with our data).

The ratio between likelihood and their probability value under Q will be the new weight → the Q sample is changed into a P sample.

The advantage is that we did not need to compute likelihoods in points of a grid (which might be a bad coverage of the P sample) but on the sample points of Q (which are a better coverage of the P sample).

Monte Carlo Markov chain (MCMC)

The Markov method is related with importance sampling in that we sample from an auxiliary distribution Q.

But now Q does not need to be similar to P.

We start sampling one point of P and center Q on that point.

Then we sample from Q - but Q depends on the current position in space.

Q is not important - it may change from point to point.

This method builds correlated samples: each point depends on the previous one - this is the definition of a **Markovian process**.

This works if it fulfills the following properties:

It must be **irreducible** → there is a non-zero probability of reaching any model from any starting model.

For example, if the target distribution has several local maxima, it may happen that the chain cannot pass from one of those regions to another. In this case it may converge to different distributions, depending on the starting point of the chain.

It must be **aperiodic** → it must not oscillate between different sets of models in a periodic movement.

It must be **invariant** → once the chain follows the target distribution, all subsequent iterations will also have that same distribution.

The most used algorithm of Markov chain Monte Carlo (MCMC), has these three properties. It is called **Metropolis-Hastings**:

Given a point m, a candidate new point m' is generated from Q(m,m').
The point is accepted to be part of the sample with a certain probability:

$$\alpha(m, m') = \min\left(1, \frac{P(m')Q(m'|m)}{P(m)Q(m|m')}\right)$$

P are the likelihoods of the two points.
If Q are symmetric distributions → Q(m'|m)=Q(m|m')

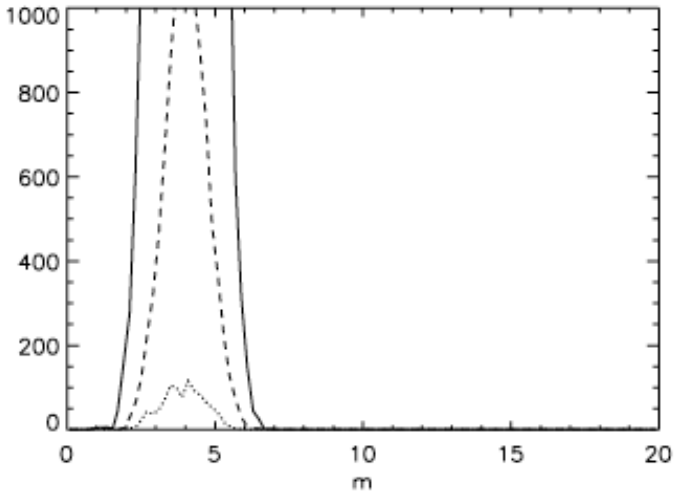
In summary:

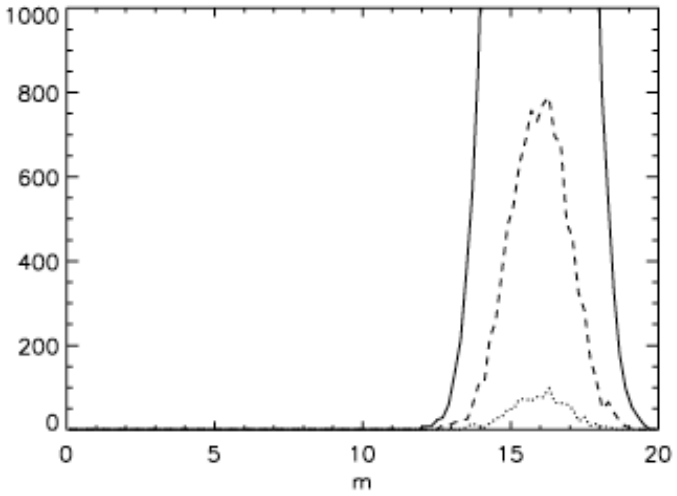
If P(m') > P(m) → m' becomes a new point of the sample
If P(m') < P(m) → m' may or may not become a new point, with a probability P(m')/P(m). The better it is, the better chance to be accepted.

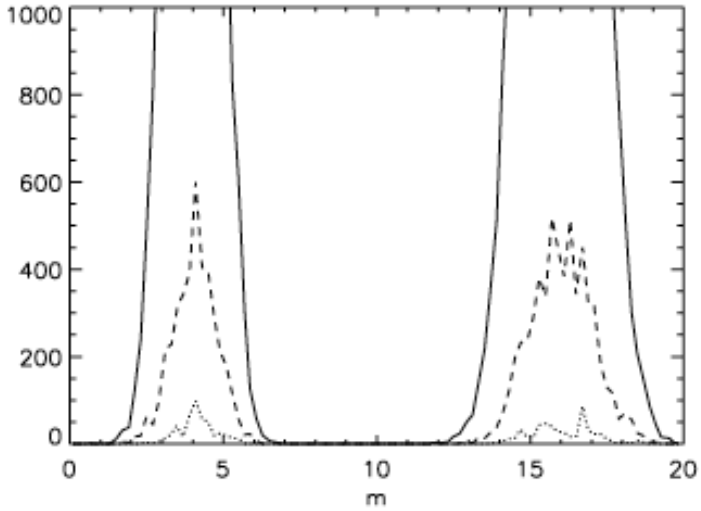
When m' is not accepted, the chain stays at m → the weight of m in the sample increases.

Properties of MCMC

Starting point







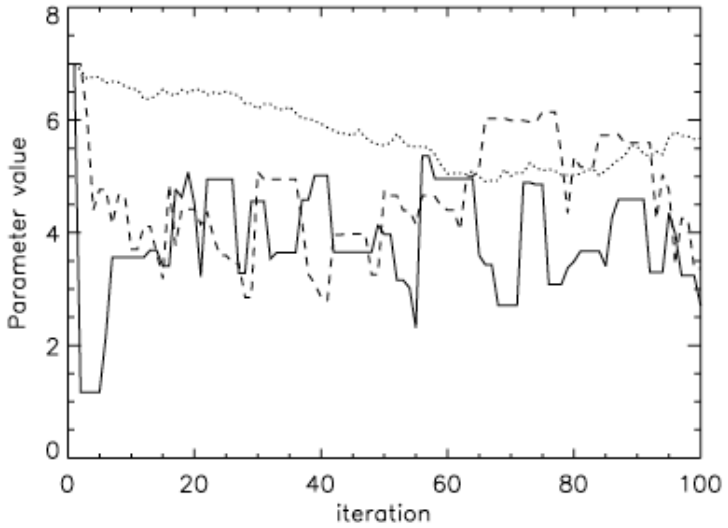
Sampling a binomial distribution
with different choices of starting points.

Step (scale of Q)

In comparison with the typical scale of P:

if too large → once the chain gets into a high posterior region, most of the subsequent proposed models will be in regions of lower posterior, and are likely to be rejected.

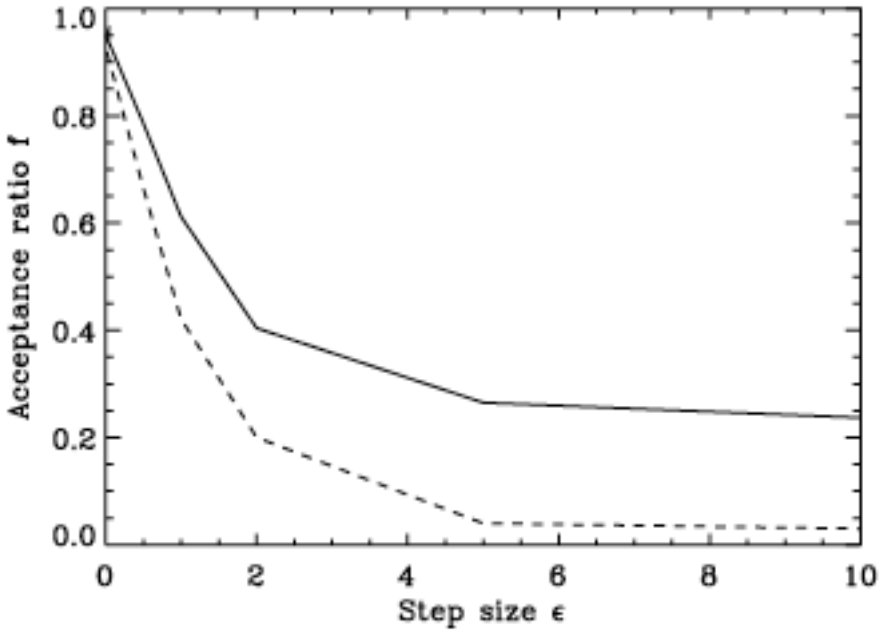
if too small → acceptance rate will be larger and the chain will move frequently. However, it will move in small steps, taking a long time to probe all space and being virtually non-irreducible.



increasing steps:

dot
dash
solid

Acceptance rate



Each point in the line is made from a full chain.

Optimal acceptance rate

0.3 - 0.5

Optimal step size

2 x σ_parameter

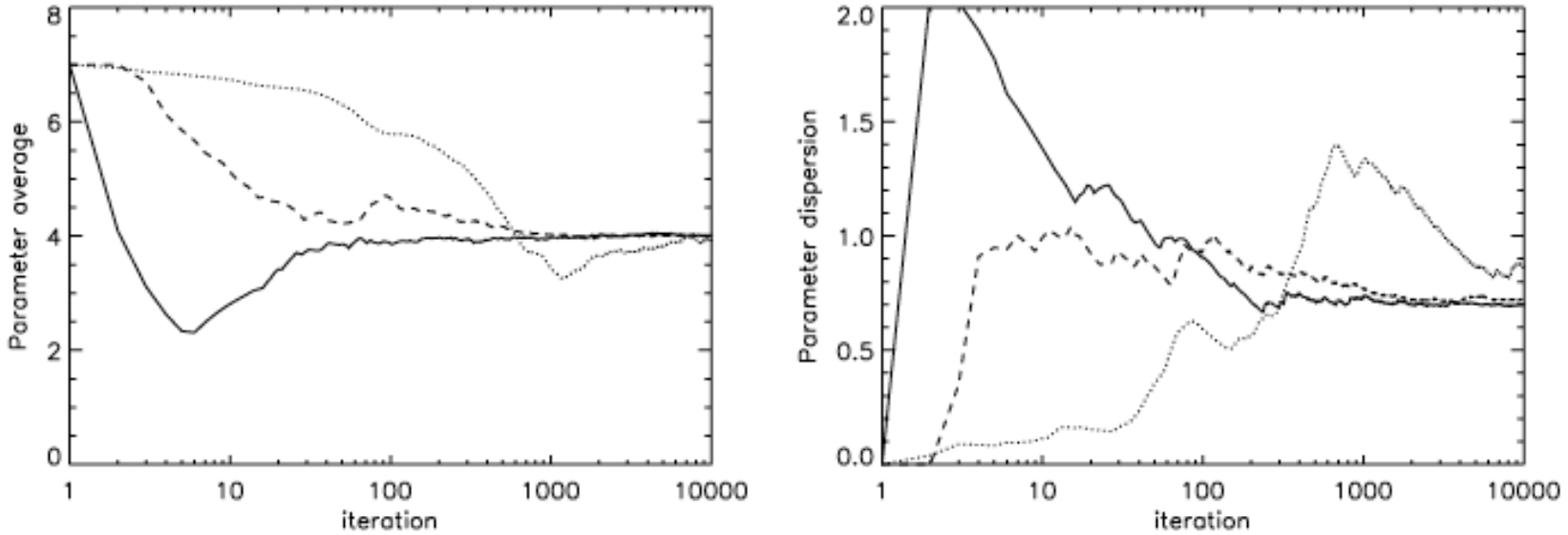
Q may have different scales on different directions.

Q may be chosen to be aligned with degeneracy directions for larger efficiency.

Convergence

How to assess convergence?

Convergence is related with the amount of time needed for the chain to start sampling from the target.



Comparing chains: the one with small step is the least efficient.

The part of the chain built before convergence need to be removed: the **burn-in** (it may be a large fraction of the chain).

How to quantify convergence? the **Gelman-Rubin** convergence test

$$R = \frac{1}{W}\left(B + \frac{W\,I}{I-1}\right)$$

$$W(p) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{I-1}\sum_{i=I/2}^{I}[p_i(j) - \bar{p}\,(j)]^2$$

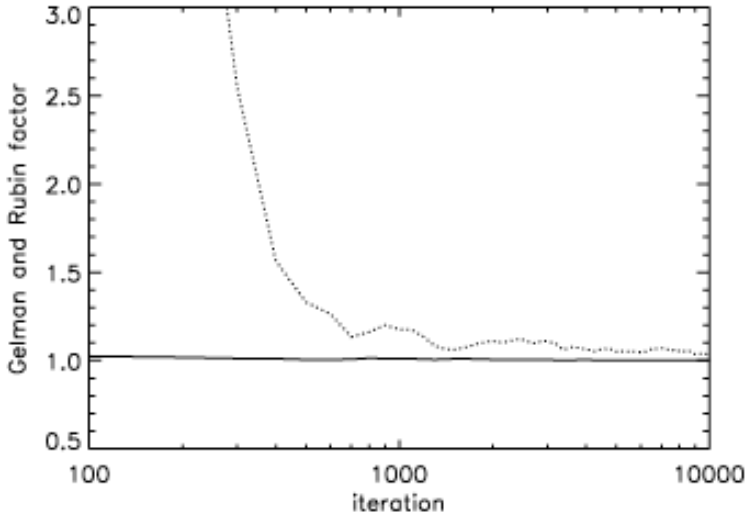
within-chain dispersion

$$B(p) = \frac{1}{J-1}\sum_{i=1}^{J}[\bar{p}\,(j) - \bar{p}]^2$$

between-chain dispersion

Convergence may require a long time (e.g. order 10^6 points)

For a large number of parameters (N > 4), MCMC is usually faster than grid computation



Correlation

The resulting chain is correlated → samples are not independent

We can compute the **correlation between points as function of separation in the chain:** (values of a parameter p in positions i and i+j)

variance = $< (p_i - p_0) (p_{i+j} - p_0) > = <p_i p_{i+j} - p_i p_0 - p_0 p_{i+j} + p_0 p_0 >$

since $< p_i > = < p_{i+j} > = p_0$ → variance = $< p_i p_{i+j} > - p_0^2$

covariance = $< (p_i - p_0) >^2 = < p_i p_i > - p_0^2$

The correlation is the variance normalized by the covariance (and it has a value < 1) :

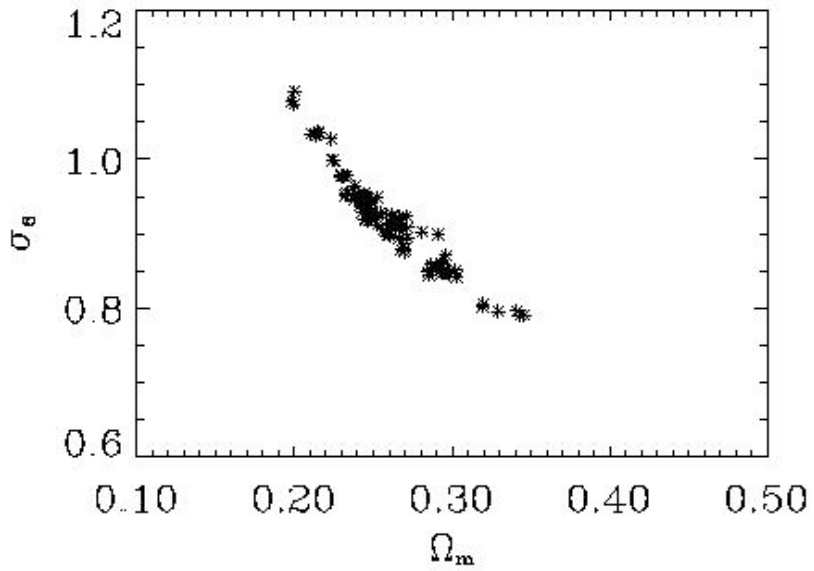
$$c_j = \frac{\langle \tau_i \tau_{i+j} \rangle - \langle \tau_i \rangle^2}{\langle \tau_i^2 \rangle - \langle \tau_i \rangle^2}$$

Reduce the correlation length of the chain : for each point of the chain remove the j subsequent points such that the c_j correlation is larger than a certain threshold (e.g. $c_j > 0.5$) → **thin-out** the chain

Output sample

The resulting chain - converged, with burn-in removed, and thinned-out - is a sample of the posterior in parameter space, P(m|d).

A plot of the cloud of points directly shows the probability density of the sample P(m|d).



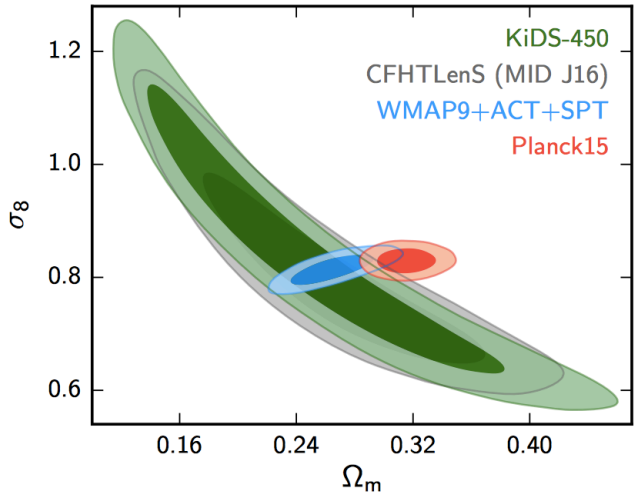
Parameter constraints

Having obtained a converged representative sample of values of P(m|d), the values of the likelihood are no longer needed to compute parameter constraints (they were only needed to build the chain).

We can compute **averages**, **dispersions** and **correlations** for all parameters directly from the chain → they are moments of the P(m|d) distribution.

Notice that the results for each parameter are already marginalized over all the other parameters.

We can also draw **contour plots**: **iso-probability contours, enclosing a given fraction of the total probability**



The resulting P(m|d) is not necessarily Gaussian → contours are not necessarily ellipses.

Fisher Information Matrix

Assume the posterior is a Gaussian distribution in the parameters' space, centred in the best-fit m_0
This method allows us to compute the covariance of the posterior distribution from the curvature matrix of the likelihood.

$$P(m|d) \propto \exp\left[-\frac{1}{2}(m_0 - m)C^{-1}(m_0 - m)^T\right]$$

Since it is a Gaussian distribution, the covariance matrix in the parameters space (that represents the uncertainty of the parameters estimation) is computed from

$$C_{ij}^{-1}(m_0) = \left(\frac{\partial^2(-\ln P)}{\partial p_i p_j}\right)_{|m_0}$$

where $m = m_0$ is the parameter value with maximum probability, i.e., the peak of the posterior.
This is also the mean value, since it is a Gaussian (symmetric) distribution.

Inserting this in Bayes theorem we can write

$$C_{ij}^{-1}(m_0) = F_{ij} + \left(\frac{\partial^2(-\ln \text{Prior})}{\partial p_i p_j} \right)_{|m_0}$$

i.e., the inverse covariance matrix of the posterior is the second-order derivative of the ln(prior) (which is zero for flat priors) + the Fisher matrix,

The **Fisher matrix** is defined as the second-order derivative of the chi-square function (the ln(likelihood)) with respect to the parameters, averaged over all parameter values.

For a Gaussian or any symmetric distribution this is just the second-order derivative taken at the peak, known as the **curvature matrix** (also known as the Hessian), i.e.,

$$F_{ij} = \left\langle \frac{\partial^2 \mathcal{L}}{\partial p_i p_j} \right\rangle \approx \left(\frac{\partial^2 \mathcal{L}}{\partial p_i p_j} \right)_{|m_0} \qquad \mathcal{L} = -\ln L$$

F_{ij} is an N_p x N_p matrix, where N_p is the number of parameters.

Naturally, the parameter values at the peak (the vector m_0) are not known.

But the goal of this method is to find the uncertainty on the parameters
(called the **credible intervals** in the Bayesian approach - also called the **confidence intervals** a name from the hypothesis testing in the frequentist approach)
and not the actual parameter values (the peak of the distribution - called the **best-fit** values).

So any reasonable value may be chosen for m_0 - this is called the **fiducial value**.
The result we are looking for is F_{ij}, the inverse covariance matrix in the parameters' space → it will give us **the uncertainty of the parameters' values around the fiducial value.**

For this reason, this method is only used to make **forecasts** of the precision of achievable with future data, while with real data we want to find out not only the uncertainty of the estimates but the actual predictions for the parameters → likelihood sampling methods are used with real data (e.g. MCMC).

Now, even if the posterior distribution is not Gaussian in reality, it is still useful to consider the Fisher matrix method, because the **Rao-Cramer inequality** states that:

the parameters confidence intervals obtained from a Fisher matrix analysis are a lower limit of the true ones.

This result may be derived from the **Cauchy-Schwarz inequality**,

$$O_L(f^2)\, O_L(g^2) \geq [O_L(fg)]^2$$

(O_L stands for linear operation and *f* and *g* are general functions).

If we choose the linear operator to be the **expectation value**, i.e. the ensemble average <> :

$O_L(g^2) = E\,[(m-m_0)^2]$, i.e., the variance (E denotes the expectation value <>)

$O_L(f^2) = E\,[(d\ln L/dm)^2]$

the **Cauchy-Schwartz inequality** becomes: $\mathrm{Var}(m) \geq \dfrac{(E[(m-m_0)\partial \ln L/\partial m])^2}{E[(\partial \ln L/\partial m)^2]}.$

The numerator is

$$(E[(m)\partial \ln L/\partial m] - m_0\, E[\partial \ln L/\partial m])^2 = \left(\int m \frac{\partial \ln L}{\partial m} L\right)^2 = \left(\frac{\partial E(m)}{\partial m}\right)^2 \geq 0.$$

while the denominator is:

$$\int \left(\frac{\partial \ln L}{\partial m}\right)^2 L = \int \frac{\partial \ln L}{\partial m} \frac{\partial L}{\partial m} = E\left[\frac{\partial^2 \mathcal{L}}{\partial m^2}\right] = F.$$

which proves the Rao-Cramer inequality: **Var (m) >= F^{-1}**

In other words, the elements of the covariance matrix in the parameter space are larger than the elements of F^{-1}

i.e., ***the inverse Fisher matrix is a lower limit of the covariance matrix.***

Notice that **large values of Fisher matrix mean small uncertainties on the parameters values**
(this is consistent with the Fisher matrix being the curvature matrix → large curvature in the likelihood means a peaked distribution → small sigma).

Computing the Fisher matrix in practice

The Fisher matrix approximation is very useful because it is very fast to compute, being basically the derivative of the cosmological function with respect to the cosmological parameters.

Let us start from the log-likelihood:

$$-\ln L \propto \frac{1}{2}\left(\mu_{obs}(z) - \mu_{th}(z, \Omega_i)\right)^T C_{zz}^{-1} \left(\mu_{obs}(z') - \mu_{th}(z', \Omega_i)\right)$$

here written considering SN data $\mu_{obs}(z)$ and the corresponding true value μ_{th} that is function of the cosmological parameters: μ_{th} (z; Ω); $C_{zz'}$ is the covariance matrix of the data.

Now, remember that the Fisher matrix is computed from the second-order derivatives of the log-likelihood computed at the fiducial value, i.e., at the peak of the distribution, i.e., at $\mu_{obs} = \mu_{th}$ (fiducial parameters).

This allows us to simplify the computation. Consider one element of the X^2 sum and a diagonal covariance with elements σ^2. The derivation is:

$$\frac{\partial^2 (\mu_{obs}-\mu_{th})^2 \frac{1}{\sigma^2}}{\partial\Omega_a \partial\Omega_b}\Bigg|_{\mu_{obs}=\mu_{th}} = \frac{\partial}{\partial\Omega_b}\left[\frac{2(\mu_{obs}-\mu_{th})}{\sigma^2}\frac{\partial\mu_{th}}{\partial\Omega_a}\right] =$$

$$= \frac{2}{\sigma^2}\frac{\partial}{\partial\Omega_b}\left[(\mu_{obs}-\mu_{th})\right]\frac{\partial\mu_{th}}{\partial\Omega_a}\Bigg|_{\mu_{obs}=\mu_{th}} + \frac{2}{\sigma^2}\underbrace{(\mu_{obs}-\mu_{th})}_{\to 0}\frac{\partial^2\mu_{th}}{\partial\Omega_a\partial\Omega_b}\Bigg|_{\mu_{obs}=\mu_{th}}$$

$$= \frac{2}{\sigma^2}\frac{\partial\mu_{th}}{\partial\Omega_a}\frac{\partial\mu_{th}}{\partial\Omega_b}$$

So in practice we just need to compute the first derivative of the cosmological function μ_{th} with respect to the cosmological parameters, due to the condition $\mu_{obs} = \mu_{th}$.

This result is valid if the covariance matrix does not depend on the cosmological parameters. If this is not the case the covariance (i.e. $1/\sigma^2$ in this example) also needs to be differentiated). Usually that dependence is weaker, and the above formula is a good approximation.

So the general practical formula of the Fisher matrix is:

$$F_{ab} = \left(\frac{\partial \mu_{th}(z;\Omega)}{\partial \Omega_a} \right)^T C^{-1}_{zz'} \frac{\partial \mu_{th}(z';\Omega)}{\partial \Omega_b}$$

(notice that the factor 2 cancels out with the ½ of the X^2)

We just need to **compute the derivatives of μ with respect to all cosmological parameters.** For each parameter we will have a vector (a discretized function of z), that contracted with the covariance matrix will produce a number for each pair of cosmological parameters.

In other words, the Fisher matrix has dimension of N_p x N_p and is the sum of the products of two derivatives over the redshift range, normalized by the variances:

$$F_{ab} = \sum_z \left(\frac{\partial \mu_{th}(z)}{\partial \Omega_a} \frac{\partial \mu_{th}(z)}{\partial \Omega_b} \right) \frac{1}{\sigma_z^2}$$

(written here for the case of a diagonal covariance).

The diagonal terms of the Fisher matrix (for a diagonal covariance matrix) are just:

$$F_{\Omega_m \Omega_m} = \sum_z \left(\frac{\partial \mu_{th}(z)}{\partial \Omega_m} \right)^2 \frac{1}{\sigma_z^2}$$

If the **derivative of the cosmological function** with respect to a certain parameter is larger than with respect to another one, it means that the cosmological function is more sensitive to the first one → the corresponding component of the Fisher matrix is larger → the corresponding F^{-1} value is smaller → the uncertainty on the first parameter is smaller than on the second one.

But what about the absolute value of the uncertainty? We saw it is smaller, but is it small? That depends on the data covariance matrix that equally affects the derivatives with respect to all parameters:
If **the data errors** are small → the derivatives are divided by a small number → the components of the Fisher matrix are larger → the corresponding F^{-1} value are smaller → the parameters are estimated with smaller uncertainties.

In summary:

The inverse of the Fisher matrix is a covariance matrix in the parameters space.

The square root of its diagonal gives the error bars on the estimated parameters. As in the data space, if the parameters are correlated, the full Fisher matrix is needed to quantify the errors.

The error associated with the estimate of a cosmological parameter depends on two factors:

- the sensitivity of the cosmological function to the parameter (the derivatives)

- the precision and accuracy of the data (the data covariance matrix)

Finding the credible intervals in the parameters space (the **contours**)

The computation of the Fisher matrix we just did is exact, regardless of the posterior being a Gaussian or not.

Now, to plot the contours in the parameters' space we will consider the approximation that the posterior is a Gaussian and consider a Taylor expansion of the log-posterior in the parameters space:

$$\mathcal{L}(m) - \mathcal{L}(m_0) = 0 + \quad (p_i - p_{i0}) F_{ij} (p_j - p_{j0}) + \mathcal{O}(\Delta_p^3)$$

Accuracy of the method:

- The expansion shows explicitly that the Fisher matrix method gives a lower limit for the parameters' variance. The result is only exact if higher-order derivatives are zero (which happens for a Gaussian, which is fully described by only two moments).

- Moreover, the result of this method is not accurate (even in the case of a Gaussian posterior) if the fiducial value chosen is not at the peak of the distribution.

The first term of the Taylor expansion, $(\partial L/\partial m)|m_0$ is zero, since the derivative is taken at the maximum of the likelihood (the peak).

This equation, to second order, is a quadratic equation in the variables Δp_i, with the center of the coordinates in m_0.

Note the Fisher matrix is **semi-definite positive** by construction from the derivatives of the likelihood (also from Cauchy-Schwartz).

$$\frac{\partial^2 \mathcal{L}}{\partial p_1^2} \frac{\partial^2 \mathcal{L}}{\partial p_2^2} \geq 2 \frac{\partial^2 \mathcal{L}}{\partial p_1 \partial p_2}$$

(i.e., the correlation coefficients are smaller than 1)

→ **the points of constant Δ*L* define a (hyper)ellipse**.

A value of $\Delta L = L_0 - L(p_i)$ gives a contour level, or (n-sigma) confidence interval, that connects all points p_i in the parameter space that have the same likelihood.

1D "contours": (1 parameter)

In a 1D normalized Gaussian posterior distribution, consider the parameter values p_{min} and p_{max} (respectively to the left and the right of the peak at p_0) such that

$$\ln L(p_0) - \ln L(p_{min}) = \ln L(p_0) - \ln L(p_{max}) = 1$$

i.e., the log-likelihood of those two points differs $\Delta L=1$ from the log-likelihood of the peak. This is called the **1-sigma** level

For 1-sigma the quadratic equation is simply:

$$\Delta L = L_0 - L(p_i) = 1 \rightarrow F\,(p-p_0)^2 = 1 \rightarrow (p-p_0) = \mathrm{sqrt}(1/F)$$

→ **The 1-sigma error is sqrt(1/F)**

Incidently, if we compute the integral of the normalized Gaussian from p_{min} to p_{max} the result is 0.683, meaning that the volume enclosed by the contour $\Delta L=1$ contains 68.3% of the total probability.

Other probability levels are usually defined:

$\Delta L=4$ is **2-sigma** → contains 95.4% of the total probability
$\Delta L=9$ is **3-sigma** → contains 99.7% of the total probability

2D contours: (2 parameters)

Integrating a 2D normalized Gaussian, we find that the 68.3%, 95.4% and 99.7% values correspond to different likelihood levels than in a 1D Gaussian.

The levels are now $\Delta L = 2.3, 6.2, 11.8$, respectively.
Nevertheless, they are still called 1, 2 and 3-sigma levels.
(For example, in 2D, $\Delta L=1$ only encloses 40% of the probability).

The quadratic equation for a fixed ΔL is

$\Delta L = F_{xx} (x-x_0)^2 + 2F_{xy} (x-x_0)(y-y_0) + F_{yy} (y-y_0)^2$ → **this defines an ellipse.**

So iso-probability contours in the Fisher matrix method are ellipses. Larger ellipses correspond to larger probability volumes

In this way, the components of the inverse Fisher matrix give us directly:

sig^2_{xx} - variance of parameter x
sig^2_{xy} - covariance (correlation of x and y)
sig^2_{yy} - variance of parameter y

If F is diagonal there is no correlation and the matrix axes are along the parameter axes x and y.

3D contours: (3 parameters)

The (hyper-)ellipse equation for a fixed ΔL is

$$\Delta L = F_{xx} (x\_x_0)^2 + 2F_{xy} (x-x_0)(y-y_0) + F_{yy} (y-y_0)^2 + 2F_{xz} (x-x_0)(z-z_0) + 2F_{yz} (y-y_0)(z-z_0) + F_{zz} (z-z_0)^2$$

The components of the inverse Fisher matrix give us directly:

sig^2_{xx}- variance of parameter x
sig^2_{xy} - covariance (correlation of x and y)
sig^2_{yy} - variance of parameter y
sig^2_{xz} - covariance between x and z
sig^2_{yz} - covariance between y and z
sig^2_{zz} - variance of parameter z

But how can we plot ellipses (2D contours) in this case?

There are two ways to plot the ellipses on each of the 3 planes (x,y), (x,z), (y,z).

Consider the contour in the (x,y) plane. The two ways are:

Maximizing: one of the parameters is kept fixed at the maximum (in the x,y case, we fix $z = z_0$).

This corresponds to a 2D slice through the 3D hyper-ellipse.

In pratice → remove z line and column from F → use this reduced F to plot the contour (x,y) or invert it to read the uncertainties directly on the new F^{-1}

Marginalizing: integrating over the full range of the 3rd parameter.

This corresponds to projecting the hyper-ellipse on a 2D plane.

Integrating the likelihood will remove the dependence on the third parameter from the multivariate Gaussian, obtaining a Gaussian without that parameter that can be differentiated to get a (reduced) Fisher matrix.

In practice → remove z-axis line and column from the covariance F^{-1}, obtaining a reduced F^{-1} → use it to read directly the uncertainties or invert it to insert in the ellipse equation and plot the contour.

Notice that marginalizing results in a larger ellipse than maximizing.

The uncertainty volume can also be reduced by neglecting the axis that have small variance → Principal Components Analysis

Figure-of-Merit

The area of a 1-sigma ellipse is:

$$\pi\, a\, b = 2.3\, \pi \,/\, \mathrm{sqrt}(\det F)$$

The square-root of the determinant of the 2D Fisher matrix is proportional to the inverse of the area of the ellipse.

The **Figure-of-Merit (FoM)** is defined as

$$\mathrm{FoM} = \mathrm{sqrt}(\det F)$$

The FoM of the ellipse in the w_0, w_a plane is used to quantify the constraining power of cosmological surveys: it is the called the **dark energy FoM**

Notice that it is also possible to marginalize on a reduced interval instead of integrating to infinity.
This is equivalent to introduce a **prior**, restricting the interval of a given parameter. In this case the prior contribution needs to be added to the Fisher matrix → this should result in larger error bars than the maximization but smaller than the full marginalization.