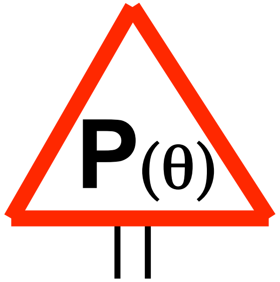
Statistical tools for cosmology

4th UniverseNet School
Lecce 2010

Plan

- What will not be discuss
- The likelihood function $\mathcal{L}$
- Which bad treatments do frequentist adepts impose to $\mathcal{L}$?
- What do Bayesian illusionists cook up out of $\mathcal{L}$?
- Monte Carlo likelihood sampling methods
- Parameter forecasts

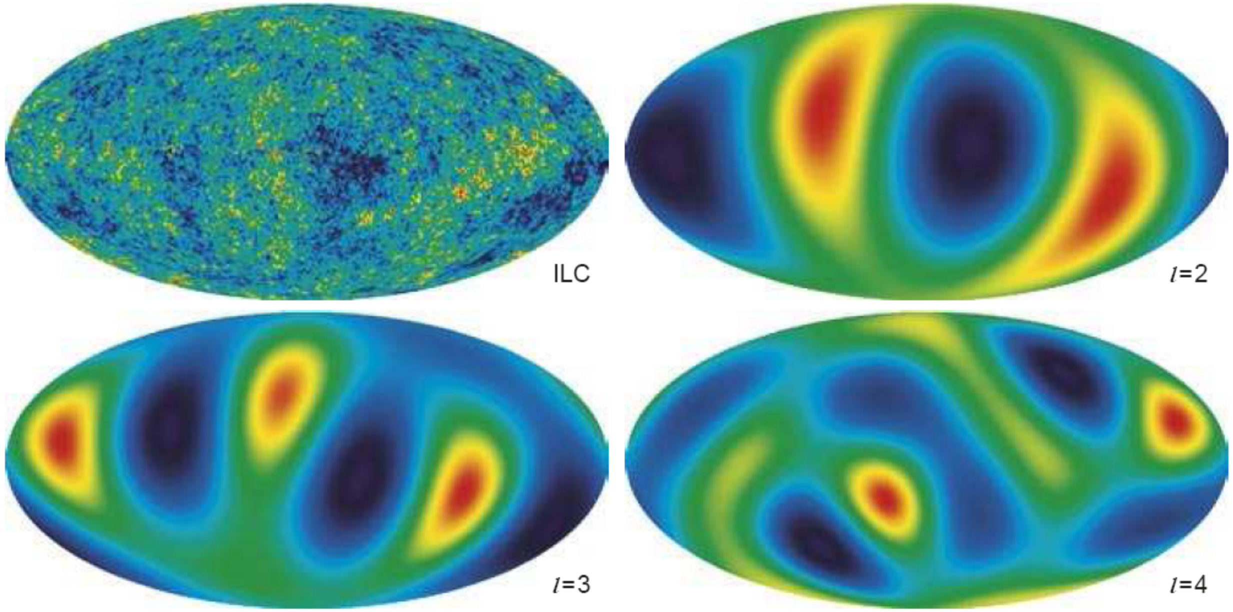
Statistics are dangerous



- Can provide silly answers to consistent questions

- Can provide (appearently) consistent answers to silly questions

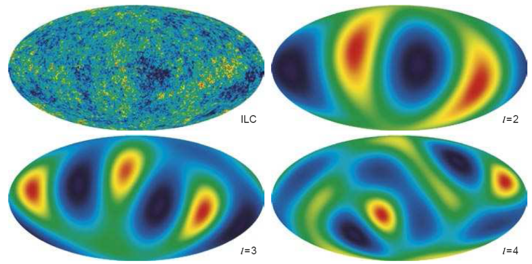
Loosely-defined questions



- "why does the large-scale universe looks like it looks like?"

Loosely-defined questions

- Identify a strange property of the data
- Compute many realizations of given theory, and find that only ε% of them have this strange property… conclude that theory is wrong…



- Mathematically dangerous because need to trust absolute likelihood, even in tails

- Conceptually dangerous because "strangeness" cannot be quantified. *If theory correct, probability that "such a strange thing can happen" should be significant, but probability that "this strange thing happens" can be small.*

discussion in Hamann, Shafieloo & Souradeep 2009

Loosely-defined questions

- Claims based on anthropic arguments combined with some probability calculation…
- E.g.: if $\rho_\Lambda^{1/4}$ > [a few] x 10^{-3}eV, no star formation, no life !

  so $\rho_\Lambda^{1/4}$ in the range from 0 to few 10^{-3}eV … and the fact that we measure a value as large as ~10^{-3}eV has a probability of order one.

- Assumes all values of $\rho_\Lambda^{1/4}$ a priori equiprobable: "flat prior on $\rho_\Lambda^{1/4}$". Why not on on ρ_Λ? Or on $\ln[\rho_\Lambda]$?
- Answer depends entirely on prior, which we cannot decide.
- Will not face this questions… only interested in parameter inference, eventually in model selection …

The likelihood function

- Question:
  - " If one assumes a theoretical model and some instrumental characteristics, what is the probability of a given data set? "

Likelihood function: $\boxed{\mathcal{L}\left(D|M(\theta_i)\right)}$

The likelihood function

- Function of $D = (x_1^{\mathrm{obs}}, ..., x_n^{\mathrm{obs}})$ depending on:
  - experimental noise
  - theoretical predictions:
    - deterministic theory: set of numbers

$$(x_1^{\mathrm{th}}, ..., x_n^{\mathrm{th}})|_{\{\theta_i\}}$$

    - stochastic theory: probability distribution

$$\mathcal{P}|_{\{\theta_i\}}(x_1^{\mathrm{th}}, ..., x_n^{\mathrm{th}})$$

The likelihood function

- Ex: Gaussian theory and noise:
  - Independent points:

$$\mathcal{L}(\{x_j^{\rm obs}\}) \propto \Pi_j \frac{1}{\sqrt{\sigma_{\rm inst}^2 + \sigma_{\rm th}^2(\theta_i)}} \exp\left(\frac{(x_j^{\rm obs} - \bar{x}_j^{\rm th}(\theta_i))^2}{2(\sigma_{\rm inst}^2 + \sigma_{\rm th}^2(\theta_i))} \right)$$

$$\Rightarrow -2 \ln \mathcal{L}(\{x_j^{\rm obs}\}) = {\rm cste} \; + \; \sum_j \frac{(x_j^{\rm obs} - \bar{x}_j^{\rm th}(\theta_i))^2}{\sigma_{\rm inst}^2 + \sigma_{\rm th}^2(\theta_i)} \equiv {\rm cste} \; + \; \chi^2$$

The likelihood function

- Ex: Gaussian theory and noise:
  - Correlated measurements:

$$\chi^2(\{x_j^{\rm obs}\}) = \Sigma_{j,k}\ (x_j^{\rm obs} - \bar{x}_j^{\rm th}(\theta_i))\ \mathbf{C}_{jk}^{-1}\ (x_k^{\rm obs} - \bar{x}_k^{\rm th}(\theta_i))$$

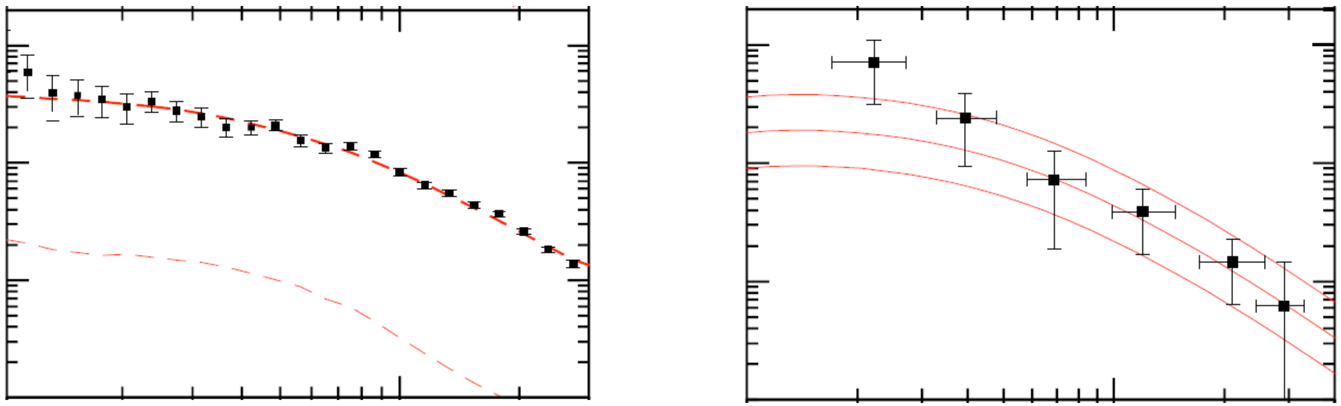
data covariance matrix

(frequently, measurement in various bins are correlated by instrument, data processing, binning)

The likelihood function

- Ex: Gaussian theory and noise:
  - Correlated measurements:

SDSS $P_{gv}(k)$
Tegmark et al. 03

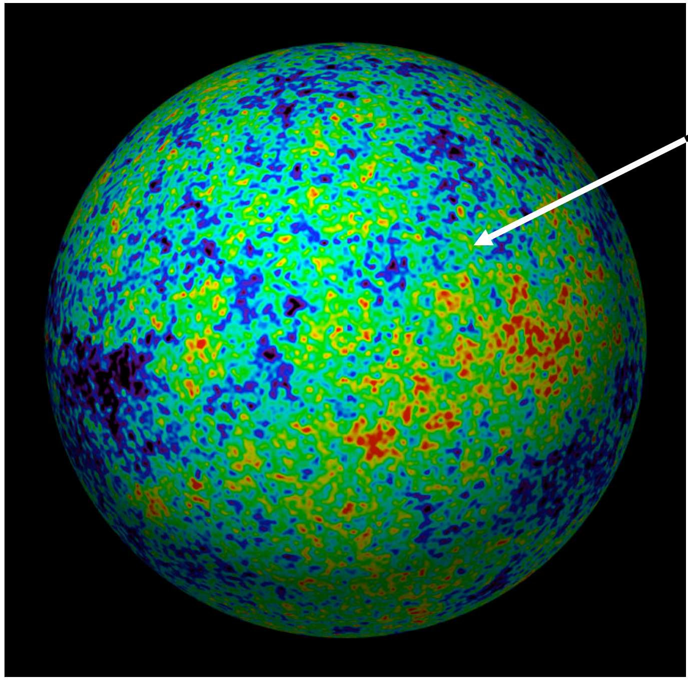


Correlated   Decorrelated

(frequently, measurement in various bins are correlated by instrument, data processing, binning)

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, ideal experiment:



$$\frac{\delta T}{T}(\hat{n}) \text{ for each pixel, } \theta \sim \pi / l_{\rm max}$$

$$\Downarrow$$

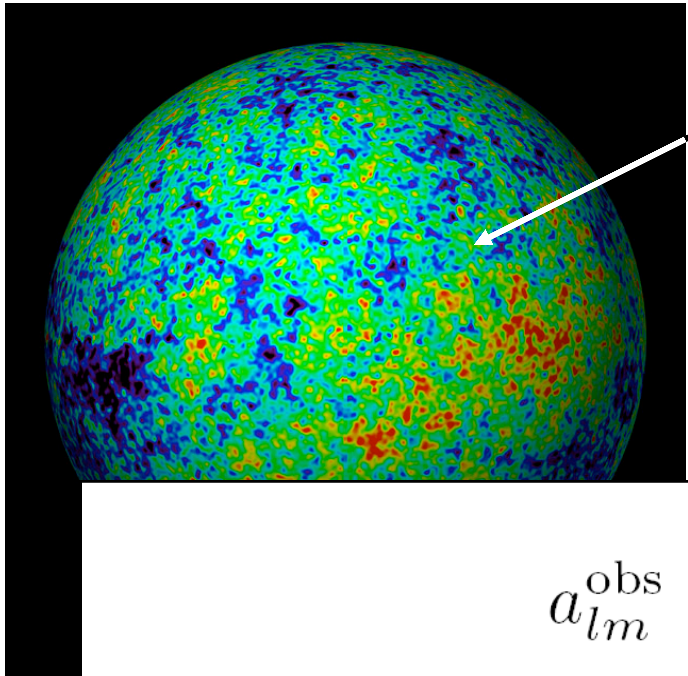
$$a_{lm}^{\rm obs} = \ldots \int d\hat{n} Y_{lm}(\hat{n}) \frac{\delta T}{T}(\hat{n})$$

$$\Downarrow$$

$$C_l^{\rm obs} = \sum_{m=-l}^{l} \frac{|a_{lm}^{\rm obs}|^2}{2l+1}$$

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, ideal experiment:



$\frac{\delta T}{T}(\hat{n})$ for each pixel, $\theta \sim \pi / l_{\max}$

$$\Downarrow$$

$$a_{lm}^{\rm obs} = \dots \int d\hat{n} Y_{lm}(\hat{n}) \frac{\delta T}{T}(\hat{n})$$

$$\Downarrow$$

$$C_l^{\rm obs} = \sum^{l} \frac{|a_{lm}^{\rm obs}|^2}{2l+1}$$

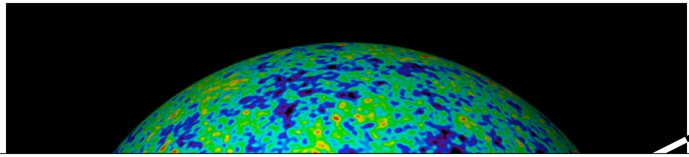
$$a_{lm}^{\rm obs} = a_{lm}^{\rm th} + a_{lm}^{\rm noise}$$

gaussian, variance $C_l^{\rm th}$

gaussian, uncorrelated, variance $N_l^{\rm exp}$ given by exp. sensitivity and resolution

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, ideal experiment:



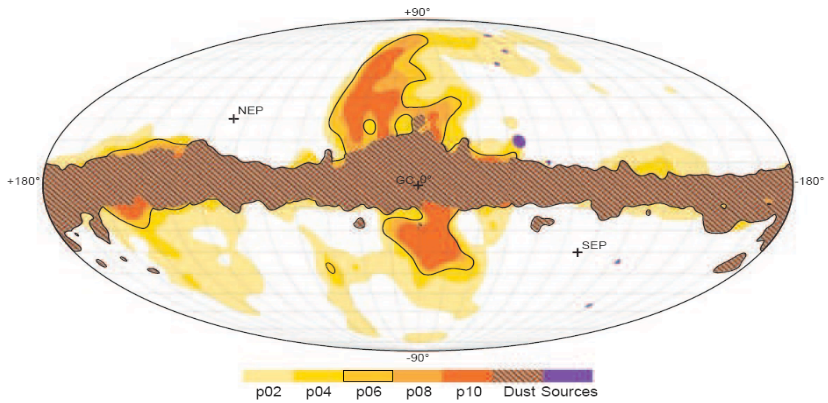
$\frac{\delta T}{T}(\hat{n})$ for each pixel, $\theta \sim \pi / l_{\max}$

$$\mathcal{L}(\{a_{lm}^{\rm obs}\}) \propto \Pi_{l,m} \frac{1}{\sqrt{C_l^{\rm th} + N_l^{\rm exp}}} \exp\left[- \frac{|a_{lm}^{\rm obs}|^2}{2(C_l^{\rm th} + N_l^{\rm exp})} \right]$$

$$\mathcal{L}(\{C_l^{\rm obs}\}) \propto \Pi_l \left(\frac{C_l^{\rm obs}}{C_l^{\rm th} + N_l^{\rm exp}} \right)^{l-\frac{1}{2}} \exp\left[- \frac{(2l+1) C_l^{\rm obs}}{2(C_l^{\rm th} + N_l^{\rm exp})} \right]$$

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, real experiment:
  - a_{lm} 's are correlated by sky cut



  - noise is not isotropic

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, real experiment:
  - Many other effects inducing correlations/ distorsions:
    - instrument (beam shape, calibration, baseline drift…)
    - data processing (time-ordered data $\Rightarrow$ map $\Rightarrow$ a_{lm}, C_l)
    - Foreground removal (point-like sources, etc.)

  - $\left.\begin{array}{l}\text{when possible, analytical modelling} \\ \text{otherwise, Monte Carlo reconstruction}\end{array}\right\}$ (complicated) likelihood

The likelihood function

- Ex: CMB temperature, gaussian fluctuations, real experiment:
  - Final likelihood = approximation (precision vs. computability)

For WMAP7 $\begin{cases} \text{large } l\text{'s : correlated gaussian } C_l\text{'s} \\ \quad \text{(non-trivial analytical approx. to cov. mat.)} \\ \text{small } l\text{'s : improved each time} \end{cases}$

(since WMAP3: correlated gaussian pixels)

The likelihood function

- Summary and message:
  - For CMB & LSS data, likelihood is:
    - Non-gaussian, involving correlations
    - Difficult to estimate
    - Always approximate

    - Should be use with great care and not "over-intepreted" or "over-trusted"…

Frequentist approach

- $\mathcal{L}\left(D|M(\theta_i)\right)$ = only relevant quantity, everything derives from it

"given a model and over many possible observations, probability that nature choose a particular parameter set $\{\theta_i\}$ proportional to $\mathcal{L}$ "

(seen now as a function of θ_i's)

Frequentist approach

- Based on intuition, not on theorems

- Maximum of $\mathcal{L}(D|M(\theta_i))$ gives goodness-of-fit of model and best-fit parameters

- For each θ_i, range in which

$$\boxed{\mathcal{L}(D|M(\theta_i)) > \text{threshold}}$$

= allowed range at given confidence level

Frequentist approach

- Goodness-of-fit: frequentist's rule of thumbs:
  - Given n = # of data points, m = # of params,

    $Q(\chi^2|n\text{-}m)$ = 1 - cumulative distr. func. of $\chi^2_{(n-m)}$

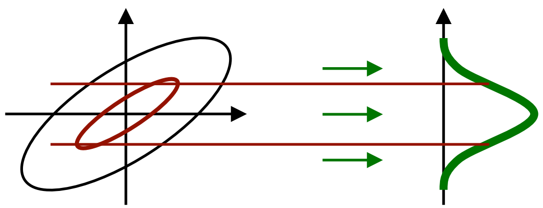
    = probability of obtaining a better fit
  - Assumes that $\chi^2(\{x_i^{\rm obs}\}) = \sum_i \frac{(x_i^{\rm obs} - \bar{x}_i^{\rm th})^2}{\sigma_i^2}$

    $$\text{with } \forall i \;\; \mathcal{P}(x_i^{\rm obs}) \propto \exp\left(- \frac{(x_i^{\rm obs} - \bar{x}_i^{\rm th})^2}{2\sigma_i^2} \right)$$
  - INAPROPRIATE in most cosmological contexts: should compute ratio:

    $$\frac{\int_{\mathcal{L}>\mathcal{L}_{\rm obs}} dD \mathcal{L}\left(D|M(\theta_i)\right)}{\int dD \mathcal{L}\left(D|M(\theta_i)\right)}$$ (relies on likelihood tails)

Frequentist approach

- Confidence limits: frequentist's rule of thumbs:
  - Confidence limit for θ_i = range in which at least one model is found with $\chi^2 - \chi^2_{min} < 1$ (68%CL), < 4 (95%CL), etc.
  - Based on assumption that $\mathcal{L}\left(D|M(\theta_i)\right)$ is a multivariate gaussian w.r.t. $\{\theta_i\}$ (Fisher matrix approximation)



  - Should compute $\Delta\chi^2$ such that: $$\frac{\int_{-2\ln\mathcal{L} < \chi^2} d\theta_i \mathcal{L}\left(D|M(\theta_i)\right)}{\int d\theta_i \mathcal{L}\left(D|M(\theta_i)\right)} = \mathrm{C.L.}$$

Bayesian approach

- Existence of a space of possible models with a measure of probability
  - "invert the likelihood" to get probability of model given the data:

$$\mathcal{P}(A\&B) = \mathcal{P}(A)\mathcal{P}(B|A) = \mathcal{P}(B)\mathcal{P}(A|B)$$

$$\Downarrow$$

$$\mathcal{P}(B|A) = \frac{\mathcal{P}(B)\ \mathcal{P}(A|B)}{\mathcal{P}(A)}$$

Bayesian approach

- Existence of a space of possible models with a measure of probability

  - Bayes theorem:

likelihood

prior probability

$$\mathcal{P}(\{\theta_i\}|M,D) = \frac{\mathcal{L}\,(D|M(\{\theta_i\}))\;\;\Pi(\{\theta_i\})}{\mathcal{P}(D|M)}$$

posterior probability

evidence

Bayesian approach

- Existence of a space of possible models with a measure of probability
  - Bayes theorem:

likelihood

prior probability

$$\mathcal{P}(\{\theta_i\}|M, D) = \frac{\mathcal{L}(D|M(\{\theta_i\}))\ \Pi(\{\theta_i\})}{\mathcal{P}(D|M)}$$

posterior probability

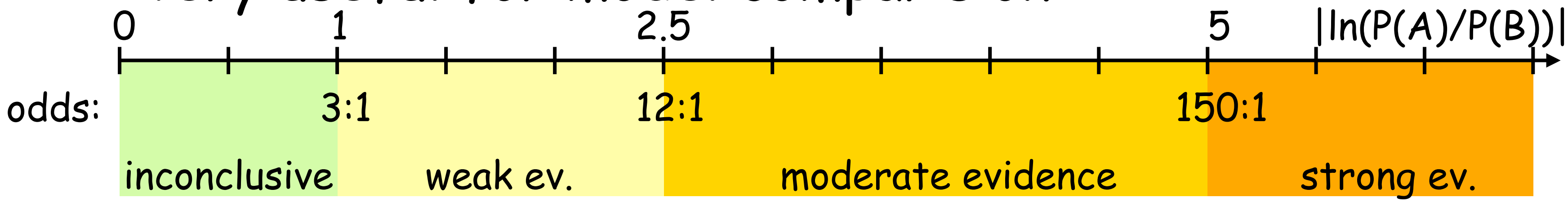
evidence

- normalization implies :

$$\mathcal{P}(D|M) = \int d^N\theta_i\ \ \mathcal{L}(D|M(\{\theta_i\}))\ \Pi(\{\theta_i\})$$ = goodness-of-fit

Bayesian approach

- The evidence:
  - Quantitative implementation of Occam's razor
  - very useful for model comparision:



  - Computation of P(A) involved …

    (thermodynamical integration, see Beltran, Slosar, Garcia-Bellido, JL, Liddle 05; see also NULTINEST approach of Feroz, Hobson, Bridges 08)

  - … but ln(P(A)/P(B)) is not if models are nested:

if A = sub-case of B with θ_1=a,

$$P(A)/P(B) = P_B(\theta_1=a|B,D) \,/\, \Pi_B(\theta_1=a)$$

Bayesian approach

- The evidence:
  - to be compared with approximate estimators for model selection:
    - N = # data points, k = # free parameters
    - $\chi^2_{min}=-2 \ln \mathbf{L}_{max}$

Tables e.g. in WMAP papers

$$\Delta(AIC) = \Delta \chi^2_{min} + 2\, \Delta\, k$$

(Akaike: frequentist)

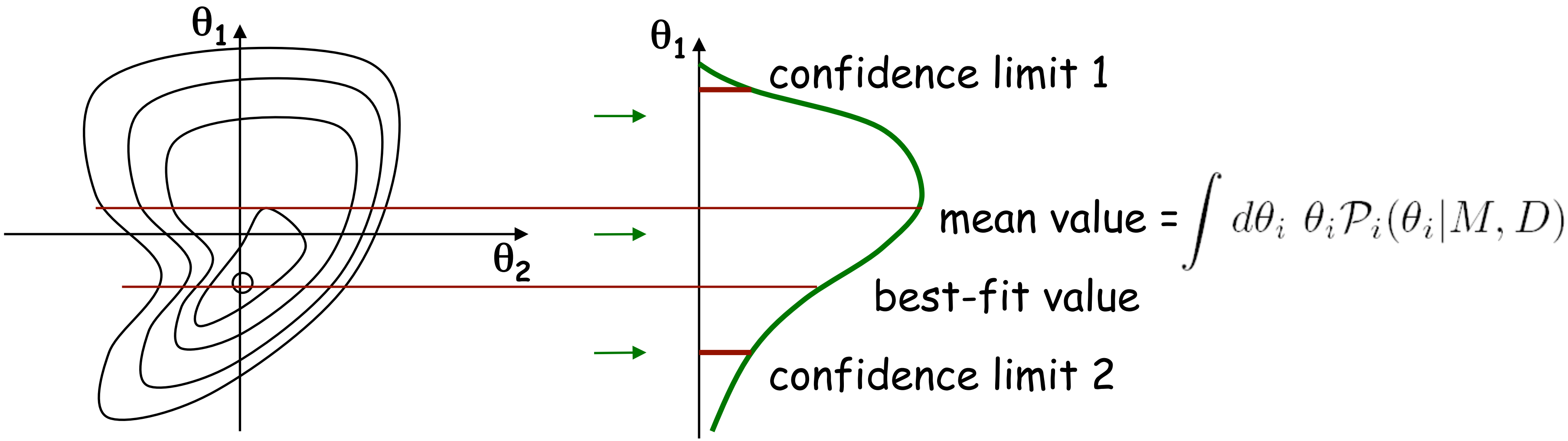
$$\Delta(BIC) = \Delta \chi^2_{min} + 2\, \Delta\, (k \ln N)$$

(approximation to Bayesian evidence)

Bayesian approach

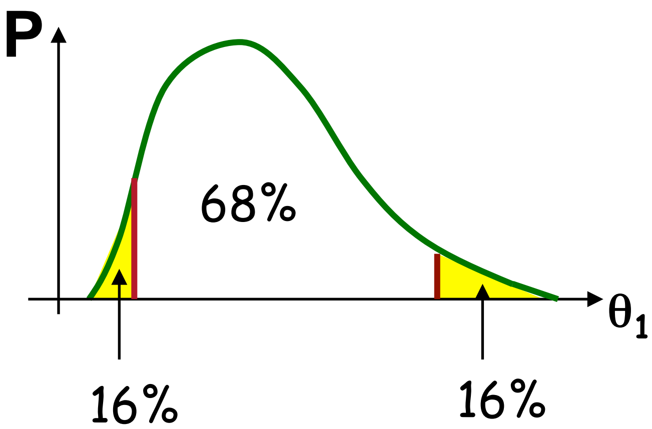
- Marginalization and C. L.:

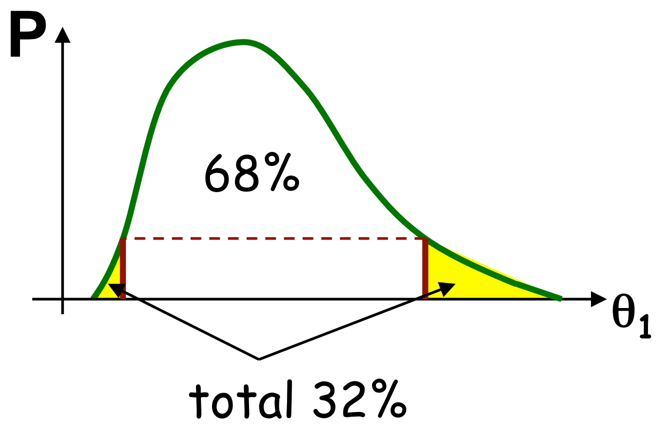
$$\forall i \ \ \mathcal{P}_i(\theta_i|M,D) = \int d^{N-1}\theta_{j\neq i} \ \ \mathcal{P}(\{\theta_j\}|M,D)$$



Bayesian approach

- C.L. definition not unique:



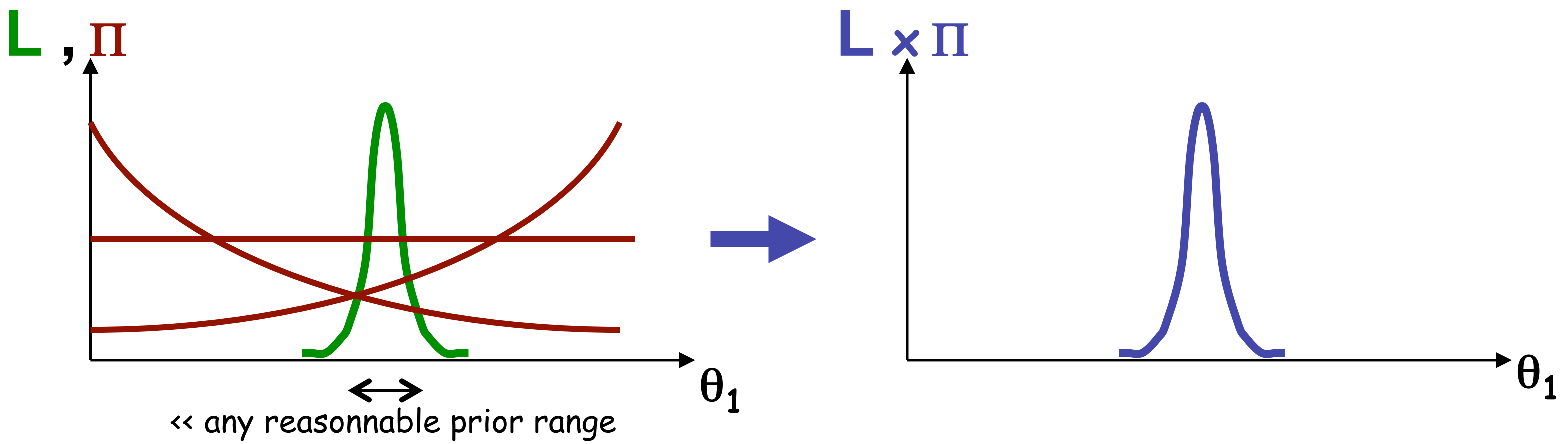


COSMOMC (Getdist.f90)
Lewis & Bridle

modified Getdist.f90 of
Hamann, Hannestad, Raffelt, Wong 2007

Bayesian approach

- Why particle-physics-frequentist-formated people are often perplex:
  - C.L.'s are prior-dependent (as well as means, evidence, …), unless parameter strongly constrained by data



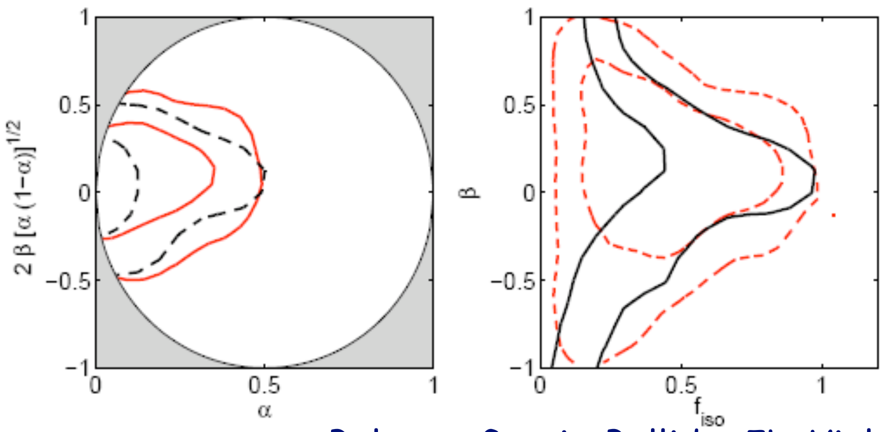
Prior ambiguity for all "un-necessary parameters": tensors, isocurvature fraction, neutrino mass, extra rel. d.o.f., etc…

Bayesian approach

- Why particle-physics-frequentist-formated people are often perplex:
  - C.L.'s are prior-dependent (as well as means, evidence, …), unless parameter strongly constrained by data

Example of adiadabatic + CDM isocurvature model :

Flat prior on
"isocurvature fraction in C_l,
Cross-correlation fraction in C_l" ?
or
"isocurvature-to-adiabatic amplitude ratio, cosine of correlation angle" ?



Beltran, Garcia-Bellido, JL, Viel 05

Bayesian approach

- Why particle-physics-frequentist-formated people are often perplex:
  - C.L.'s are prior-dependent (as well as means, evidence, …), unless parameter strongly constrained by data

Example of adiadabatic + CDM isocurvature model :

Bayes factor ln(P(A)/P(B)) changes by factor 2…

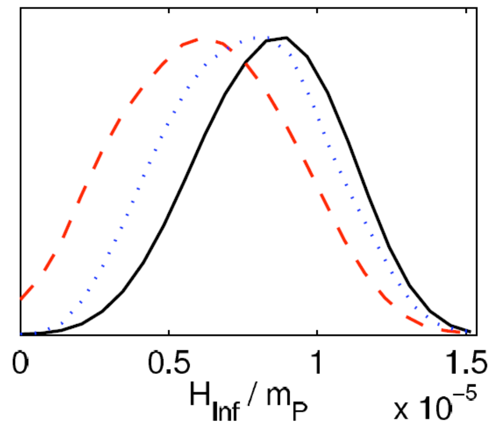
Beltran, Garcia-Bellido, JL, Liddle, Slosar 05; Trotta 05

Bayesian approach

- Why particle-physics-frequentist-formated people are often perplex:
  - C.L.'s are prior-dependent (as well as means, evidence, ...), unless parameter strongly constrained by data

Example of H_{infl} non-zero, prior-dependent mean when take flat priors on (A,r,n), or on HSR parameters, or directly on H_{infl}, or ln[H_{infl}] !!!

... while likelihood peaks in $r=H_{inf}=0$!

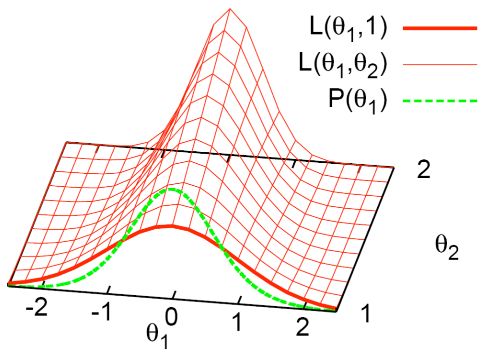


Consequence of likelihood being strongly non-gaussian w.r.t. un-necessary parameter

Hamann, Krauss, Valkenburg 08

Bayesian approach

- Why particle-physics-frequentist-formated people are often perplex:
  - C.L.'s are prior-dependent (as well as means, evidence, …), unless parameter strongly constrained by data
  - Everything can happen:
    - Posterior probability of best-fit can be poor
    - Likelihood of model built from means $\{\theta_i\}$ can be low
    - Adding MORE data can make the bounds WEAKER (if datasets disagree)
    - Adding EXTRA free parameters can make bounds STRONGER



Comparision

- If $\mathcal{L}\left(D|M(\theta_i)\right) \longrightarrow$ multi-variate gaussian function of $\{\theta_i\}$:
  Frequentist best-fits, C.L.'s $\longrightarrow$ Bayesian means, C.L.'s
  - Belief that "data improving, parameters better constrained, debate on statistics will close"…
  - NO! The "frontier" will move but there will always be a frontier…

Example of papers comparing two approaches with same models/datasets:
- Reid, Verde, Jimenez, Mena 0910.0008
- Boyarsky, JL, Ruchayskiy, Viel 0812.0010
- …

Comparision

- If $\mathcal{L}(D|M(\theta_i)) \longrightarrow$ multi-variate gaussian function of $\{\theta_i\}$:
  Frequentist best-fits, C.L.'s $\longrightarrow$ Bayesian means, C.L.'s

  – Belief that "data improving, parameters better constrained, debate on statistics will close"…

  – NO! The "frontier" will move but there will always be a frontier…

  – Bayesian supporters say future will be Bayesian because it is a better defined framework

  – … or because it is computationally much more tractable …

Monte Carlo methods

- Old (< 2003) approach to parameter extraction:
  - Sample power spectra OR likelihood in a grid in parameter space
  - Use marginalization (Bayesian) or maximization (frequentist) algorithms
  - N params: typically 10^N evaluations (weeks…)
    - (plus, if frequentist: interpolation problems + 10xN maximizations)

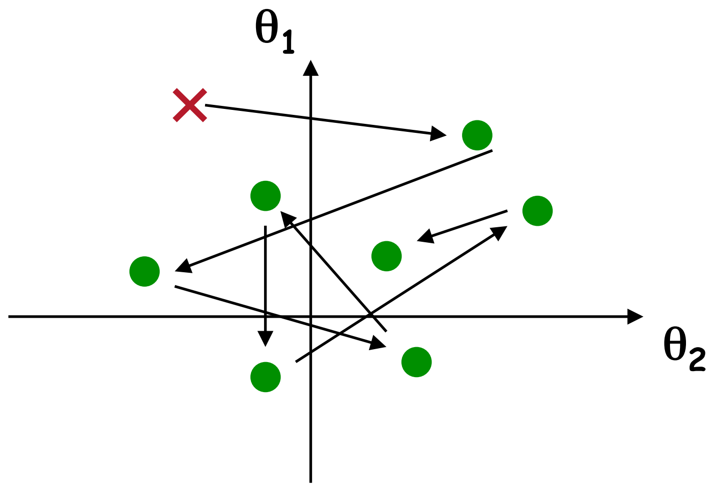
- CosmoMC (Lewis & Bridle 2002): Monte Carlo Markhov Chains (MCMC) with Metropolis-Hastings algorithm, evaluations $\propto$ N

Monte Carlo methods

- Old (< 2003) approach to parameter extraction:
  - Sample power spectra OR likelihood in a grid in parameter space
  - Use marginalization (Bayesian) or maximization (frequentist) algorithms
  - N params: typically 10^N evaluations (weeks…)
    - (plus, if frequentist: interpolation problems + 10xN maximizations)
- CosmoMC (Lewis & Bridle 2002): Monte Carlo Markhov Chains (MCMC) with Metropolis-Hastings algorithm, evaluations $\propto$ N
- Other methods: nested sampling, importance sampling, …

Monte Carlo Makhov Chains

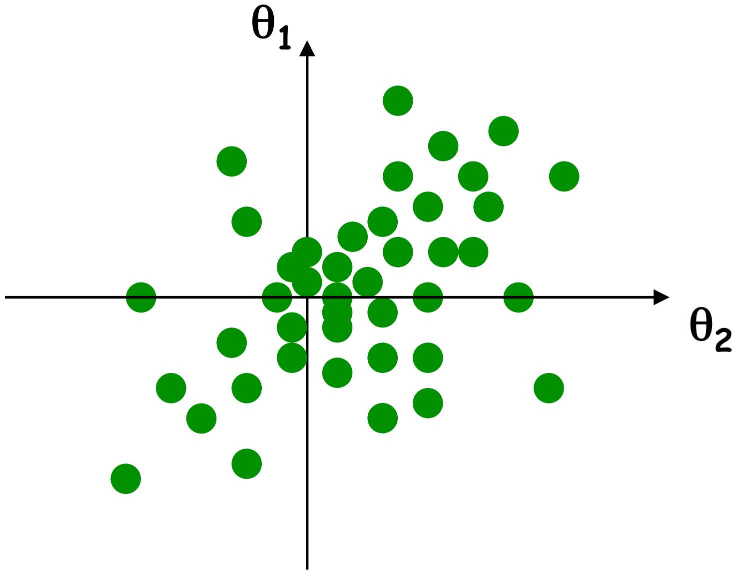
Principle:



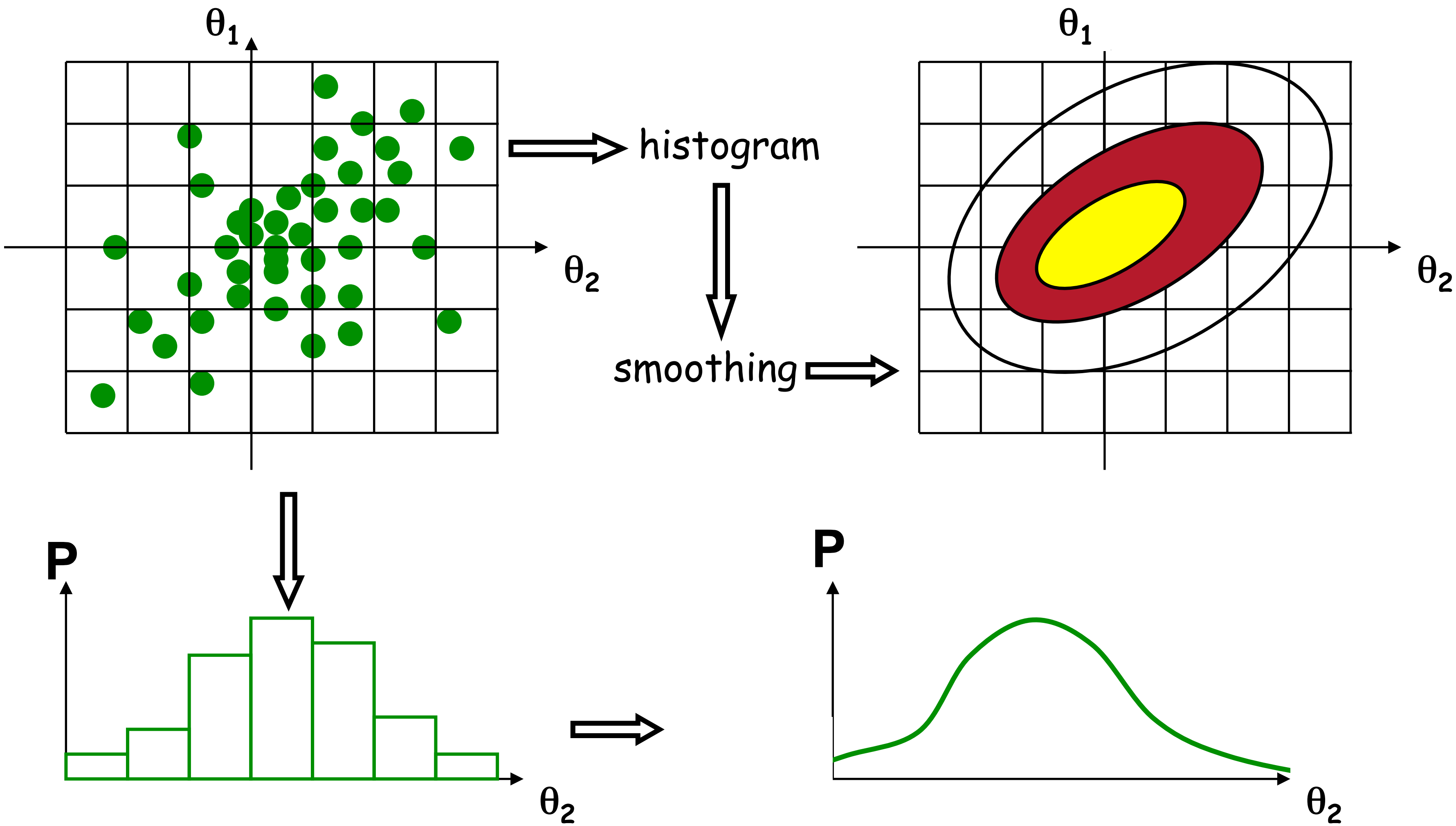
- Each *possible* next point is chosen randomly;
- $\mathcal{L}\left(D|M(\{\theta_i\})\right)\;\;\Pi(\{\theta_i\})$ is evaluated at this possible new point;
- choice to go there or not is governed by a probability dictated by the "Metropolis-Hastings" algorithm:

$$\lim_{N\to\infty} n(\{\theta_i\}) \propto \mathcal{L}\left(D|M(\{\theta_i\})\right) \times \Pi(\{\theta_i\})$$

Monte Carlo Makhov Chains



Monte Carlo Makhov Chains



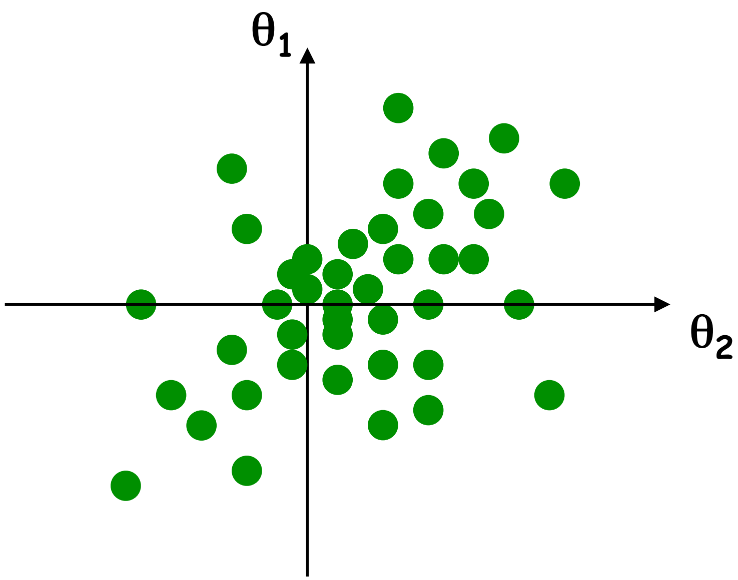
Monte Carlo Makhov Chains

- Convergence issue:
  - Several possible convergence tests, CosmoMC comes with many of them:
    - Ex: for each basis vector in param space,

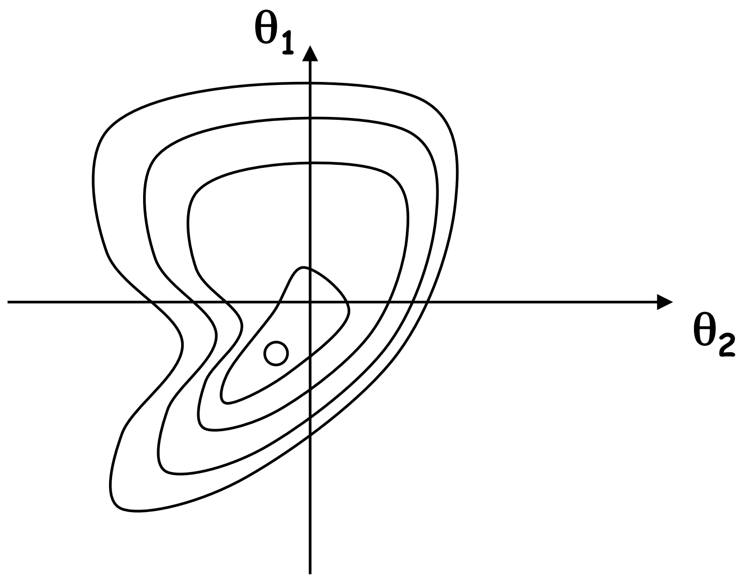
      R-1= (variance of chain means / mean of chain variances – 1) : proves that chains (or chain subsamples) agree with each other, but does not mean that they have converge
  - Known problem for bimodal distribution…
  - … but even nicely behaved distribution can be tricky (if a parameter has non-gaussian probability or participates to a degeneracy)

Monte Carlo methods

- Nested sampling :

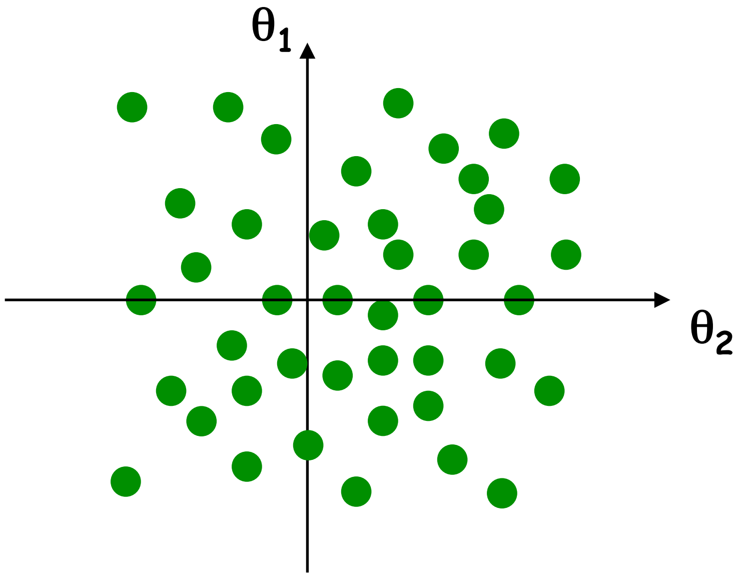


vs



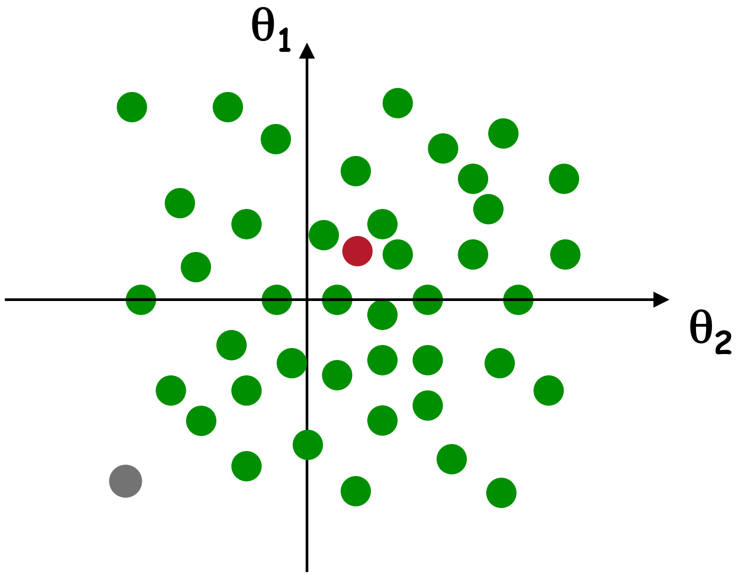
Monte Carlo methods

- Nested sampling :



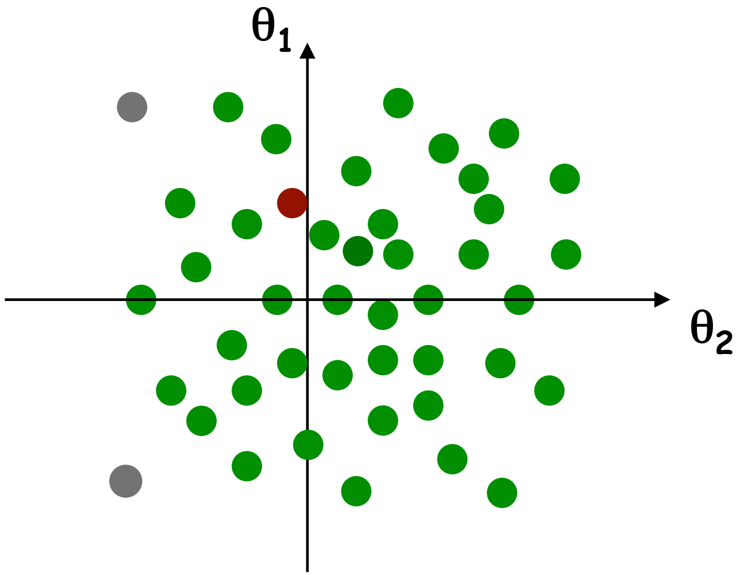
Monte Carlo methods

- Nested sampling :



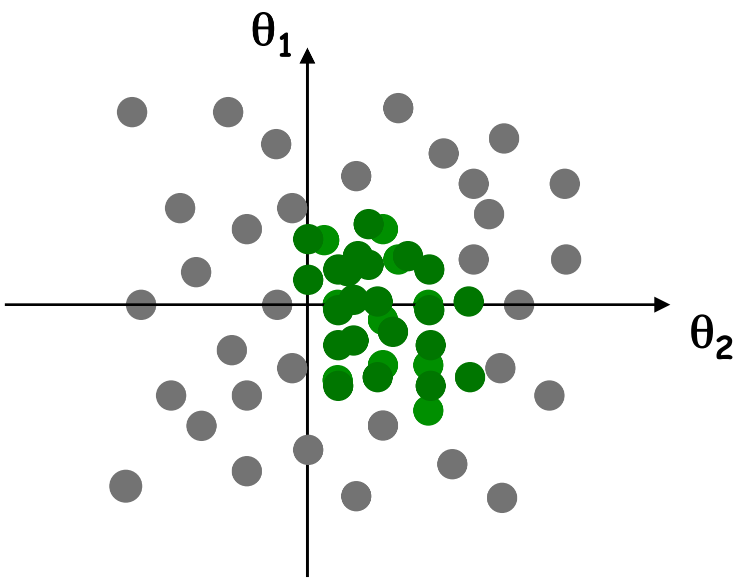
Monte Carlo methods

- Nested sampling :



Monte Carlo methods

- Nested sampling :



- At each step, enveloppe of remaining point = estimate of isolikelihood contour with $\mathcal{L}$=$\mathcal{L}$(last point eliminated)

- Collection of many isolikelihoods: knowledge of $\mathcal{L}$, and hence of evidence and posteriors

- Various algorithms for finding new points; some of them adapted to the case of multimodal likelihoods or banana-shaped denegeracies (MULINEST by Feroz, Hobson, Bridges, 08)

- Less easy to parallelize than MCMC, but more efficient when curving denegeracies

Summary of robustness of confidence limits

- When quoting/making use of C.L., beware of:
  - Uncertainties related to data:
    - Systematic errors
    - Approximations to true likelihood
  - Ambiguities related to methodology:
    - Priors, underlying model
  - Uncertainties in parameter extraction method:
    - MCMC convergence
    - Chains binning

- … C.L. on "non-necessary parameters" should only be regarded as rough estimates… in that case comparing Bayesian/frequentist is healthy!!

Parameter forecasts

- Fisher matrix analysis : $F_{ij} = [d^2 \ln\mathcal{L}/d\theta_i d\theta_j]_{max}$
  - Assume instrument sensitivity
  - Assume best-fit (fiducial) model
  - Approximate likelihood as gaussian wrt θ_i around best-fit
  - Compute $dC_l/d\theta_i\,|_{max}$ and infer F_{ij}
  - Infer $\Delta\theta_i$ from simple algebra (inversion of F_{ij})

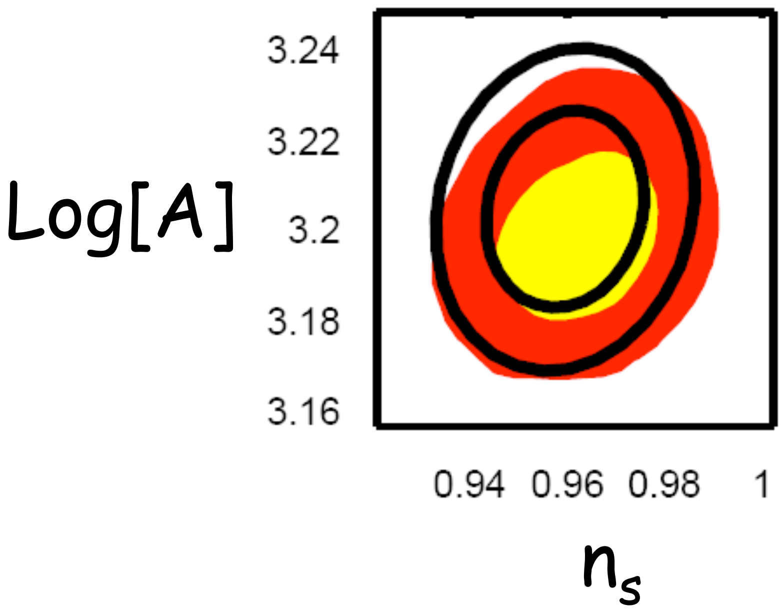
  – Problem 1: gaussian approximation can fail significantly (curving degeneracies, hard bounds)
  – Problem 2: unless C_l = linear function of θ_i, numerical estimate of $dC_l/d\theta_i$ depends on step-size

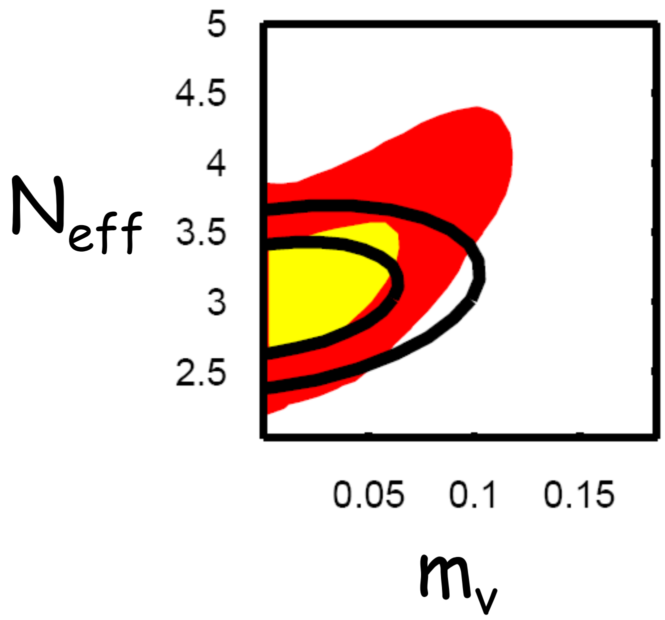
Hu, Eisenstein & Tegmark 98

- mock data + full MCMC parameter extraction

Parameter forecasts

- ex: forecast for Planck with lensing extraction
  (Perotto, JL, Tu, Hannestad, Wong 2006)





- Forecast for DE sound speed varying in range [0:1]
  (Ballesteros & JL 2010)

Bibliography

- R. Trotta, 0803.4089 [astro-ph]
- D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms.* Cambrdige University Press, 2003
- G. J. Feldman and R. D. Cousins, *A Unified approach to the classical statistical analysis of small signals,* physics/9711021
- D. Karlen, *Credibility of confidence intervals, . Prepared for Conference on* Advanced Statistical Techniques in Particle Physics, Durham, England, 18-22. Mar 2002.
- A. Lewis, S. Bridle, astro-ph/0205436
- F. Feroz, M. Hobson, M. Bridges, 0809.3437 [astro-ph]