Imperial College London

Statistics for Cosmology

Andrew Jaffe
Imperial College

TIARA Winter School
AS/NTU, Taipei
February 2014

ICIC

Imperial College London

[Mostly Bayesian] Statistics for Cosmology

Andrew Jaffe
Imperial College

TIARA Winter School
AS/NTU, Taipei
February 2014

ICIC Imperial Centre for Inference in Cosmology

Topics

- Probability theory
- Simple Examples:
- Model-fitting
  - Gaussian models — linear fitting & χ^2
  - Poisson fitting (rates)
- Probability for theorists
  - there are random statistical processes in the early Universe
- Bayesian model comparison
  - How can we tell if our parameterization is good?
- Link with tomorrow: CMB data analysis
  - Hierarchical models

References

- Hobson, Jaffe et al (eds.), *Bayesian Methods in Cosmology*
- Loredo's *Bayesian Inference in the Physical Sciences*:
  - `http://astrosun.tn.cornell.edu/staff/loredo/bayes`
  - "The Promise of Bayesian Inference for Astrophysics" & "From Laplace to SN 1987a"
- Jaynes, *Probability Theory: the Logic of Science*
  - And other refs at `http://bayes.wustl.edu`
- MacKay, *Information theory, Inference & Learning Algorithms*
- Sivia & Skilling, *Data Analysis: A Bayesian Tutorial*

What is probability?

- Frequentist view: p describes the relative *frequency of outcomes* in infinitely long trials
- Bayesian view: p expresses our *degree of belief*
- Bayesian view is what we seem to want from experiments: e.g. *given the Planck data, what is the probability that the density parameter of the Universe is between 0.9 and 1.1?*
- Cosmology is in good shape for inference because we have decent model(s) with parameters – well-posed problem

ICIC

Probability

- $P(A|B)$ = probability of "A" given "B"
  - Probabilities measure "degrees of belief"
    - $P=1$ —certainty [true]
    - $P=0$ —impossibility [false]
  - A, B are propositions
    - "Socrates is a man", "All men are mortal", "the Hubble constant is between 61 and 66 km/s/Mpc"
  - *All* probabilities are conditional (on knowledge/ belief)

Probability

- $P(A|B)$
  - Trivial extension to "probability density"
    - $p(x|I)\,dx = P(\text{"}x \text{ lies in } [x, x+dx]\text{"}|I)$
      - Value of $p(x)$ is not a probability!
  - Only consistent extension to Aristotelian logic for non-certain cases (0/1) [Cox]
    - other derivations:
      - Kolmogorov – axiomatic, de Finetti – consistency
    - Bernoulli, Laplace, Keynes, Popper, ...

p(x|y) is not the same as p(y|x)

- x = female, y=pregnant
- p(y|x) = 0.03
- p(x|y) = 1

ICIC

Laws of Probability

- $P(AB|I) = P(A|BI)\ P(B|I)$
  - "AB" = "A and B"
- $P(A + B|I) = P(A|I) + P(B|I) - P(AB|I)$
  - "A+B" = "A or B"
  - nb. $P(A + B|I) = P(A|I) + P(B|I)$ iff $P(AB|I)=0$
- $P(A|I) + P(\sim A|I) = 1$ [$\sim A$ = "not A"]
- $\int dx\ p(x|I) = 1$
  $\int dx\ p(xy|I) = p(y|I)$

Moments of distributions

- Define *Expectation*

$$E[f(x)] = \langle f(x) \rangle = \int dx\, f(x)\, p(x|I)$$

- mean $\mu = \langle x \rangle = \int dx\, x\, p(x|I)$
- variance $\mathrm{var}(x) = \langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$
- In general, *n*th moment $\langle x^n \rangle$,
  central moment $\langle (x-\mu)^n \rangle$
- Can also define cumulants, $\kappa_n = \mathrm{Cum}_n[x]$:
  - expectations of nth-order polynomials
  - add for independent "random variables"
    - $\mathrm{Cum}_n[x+y] = \mathrm{Cum}_n[x] + \mathrm{Cum}_n[y]$
  - mean, variance, skewness, kurtosis,...

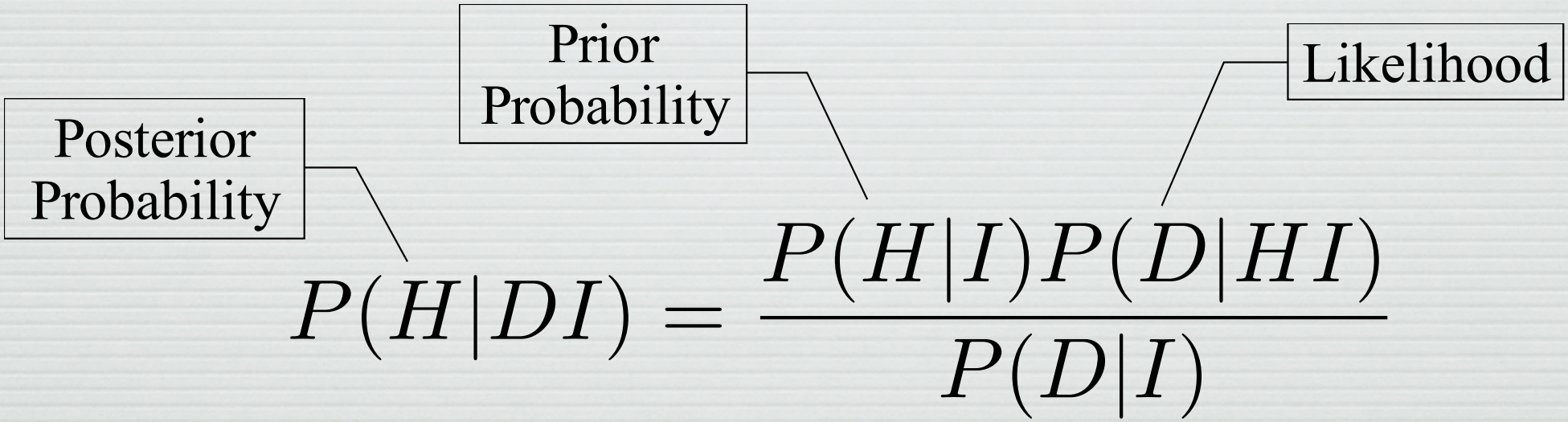
Bayes' Theorem

- Product rule:
  $P(DH|I) = P(D|HI)\,P(H|I)$
  $= P(H|DI)\,P(D|I)$

$$P(H|DI) = \frac{P(H|I)P(D|HI)}{P(D|I)}$$

- H = hypothesis
  D = data
  I = other "background" information
- Model for learning (H) from experience (D) within some context (I)

Bayes' Theorem



- Denominator doesn't depend on H, so
  - $P(H|DI) \propto P(H|I)\ P(D|HI)$ is sufficient
- $P(D|I) = \sum_H P(H|I)\ P(D|HI)$ [normalization]

Bayes' Theorem and Inference

- If we accept *p* as a degree of belief, then what we often want to determine is*

$$p(\theta|x)$$

θ: model parameter(s), *x*: the data

To compute it, use Bayes' theorem $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$

Note that these probabilities are all conditional on a) prior information I, b) a model M

$$p(\theta|x) = p(\theta|x, I, M) \text{ or } p(\theta|x\, I\, M)$$

*This is RULE 1: start by writing down what it is you want to know

RULE 2: There is no RULE n, n>1

ICIC

Bayes' Theorem

$$P(\theta|DI)\ d\theta = \frac{P(\theta|I)P(D|\theta I)}{\int d\theta'\ P(\theta'|I)P(D|\theta' I)} d\theta$$

- Theory parameterized by (continuous) θ:
  - Use probability densities

- Marginalization

$$P(\theta|DI) = \int d\varphi\ P(\theta\varphi|DI)$$

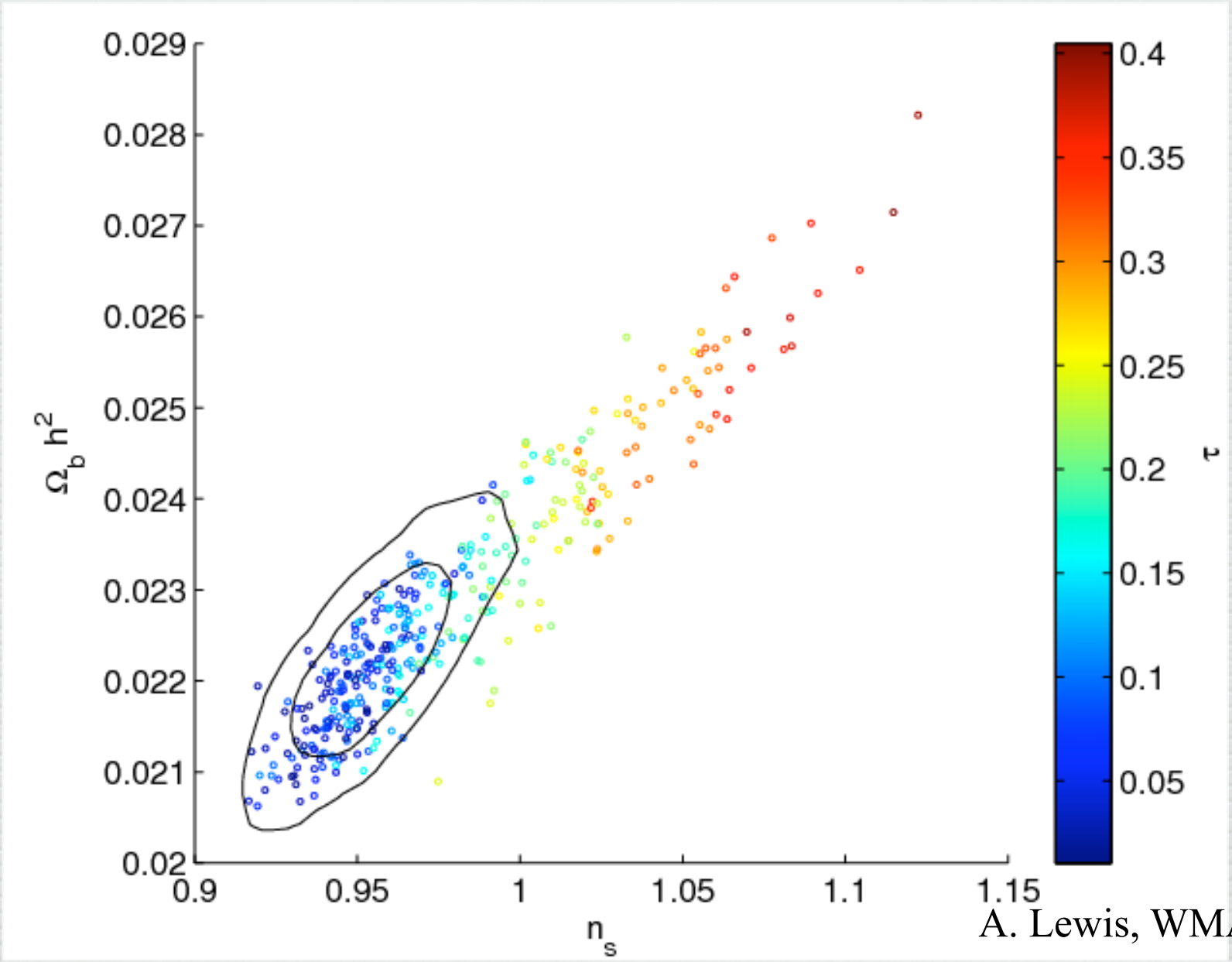
  - φ: "nuisance" parameter
    - e.g., Background level, unknown noise, etc.
    - (but a nuisance in one context is signal in another!)

Bayes' Theorem

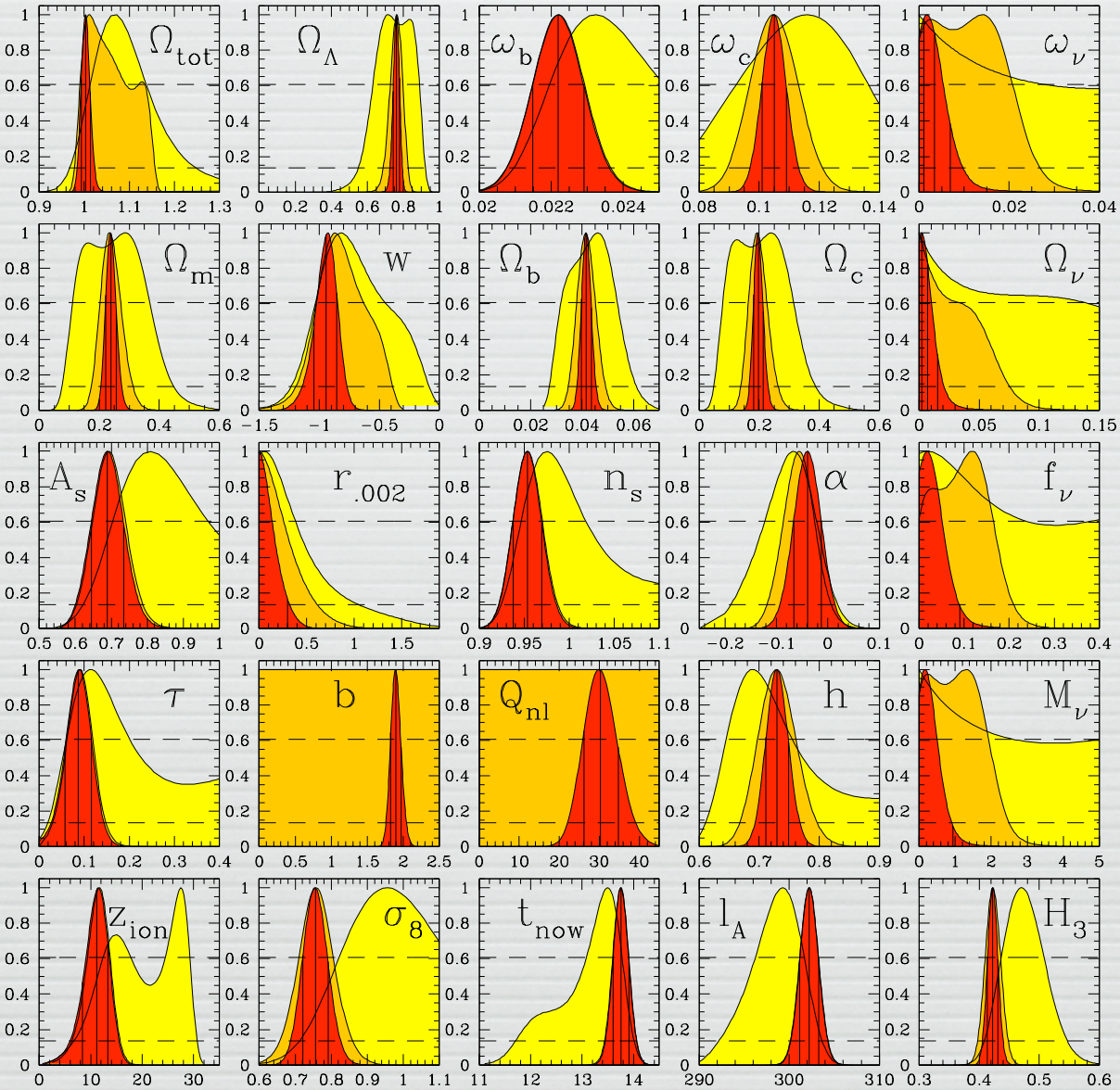
$$P(\theta|DI)\ d\theta = \frac{P(\theta|I)P(D|\theta I)}{\int d\theta'\ P(\theta'|I)P(D|\theta' I)} d\theta$$

- Posterior contains full "inference from data"
- Can *sometimes* be summarized by moments, peaks, integrals, etc.
  - maximum posterior
  - 68% enclosed probability levels
  - mean and variance

Cosmology Example



A. Lewis, WMAP 3yr



Probabilities depend on data **and** model

SDSS+WMAP: Tegmark et al 2006

FIG. 12: Constraints on key individual cosmological quantities using WMAP1 (yellow/light grey distributions), WMAP3 (narrower orange/grey distributions) and including SDSS LRG information (red/dark grey distributions). If the orange/grey is completely hidden behind the red/dark grey, the LRGs thus add no information. Each distribution shown has been marginalized over all other quantities in the "vanilla" class of models parametrized by $(\Omega_\Lambda, \omega_b, \omega_c, A_s, n_s, \tau, b, Q_{\rm nl})$. The parameter measurements and error bars quoted in the tables correspond to the median and the central 68% of the distributions, indicated by three vertical lines for the WMAP3+SDSS case above. When the distribution peaks near zero (like for r), we instead quote an upper limit at the 95th percentile (single short vertical line). The horizontal dashed lines indicate $e^{-x^2/2}$ for $x = 1$ and 2, respectively, so if the distribution were Gaussian, its intersections with these lines would correspond to 1σ and 2σ limits, respectively.

Estimating the parameter(s)

- Commonly the mode is used (the peak of the posterior)
- Mode = *Maximum Likelihood Estimator*, if the priors are uniform
- The *posterior mean* may also be quoted, but beware
- Ranges containing x% of the posterior probability of the parameter are called *credibility intervals* (or *Bayesian confidence intervals*)

ICIC

Assigning probabilities

- Probabilities *per se* don't exist in nature
  - Descriptions of phenomena for which we have only limited information.
- The principal of indifference: if I don't have *any* reason to prefer H_1 to H_2, then $P(H_1|I) = P(H_2|I)$
  - coin flipping: $P(\text{heads}|I) = P(\text{tails}|I) = 0.5$
  - colored balls from an urn: $P(\text{red}|n_{\text{red}}\ N_{\text{tot}}\ I) = n_{\text{red}}/N_{\text{tot}}$
  - generalize to $P(\text{"M star"}|I)$, $P(\text{"Giant Elliptical"}|I)$
- If you know the initial conditions — positions of balls in the urn or state of the coin, or other galaxy properties — the probabilities you assign are different than if you don't

Frequentist vs. Bayesian statistics

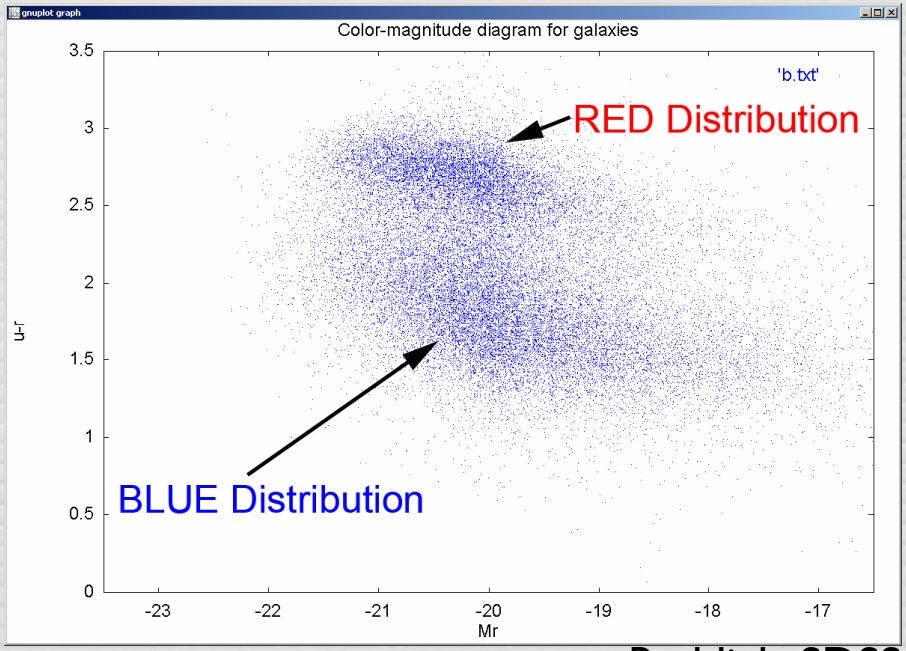
- What does $H_0 = 63 \pm 3$ km/s/Mpc mean?
- Bayesian:
  - the posterior distribution has 68% of its integral within 60-66.
  - The posterior can be used as a prior on a new application of Bayes' theorem.
- Frequentist:
  - Performing the same procedure will cover the real value within the limits 68% of the time.
  - Then what?

Frequentist vs. Bayesian statistics

- In the **Frequentist** (aka **Orthodox**) interpretation, probability is *only* defined as asymptotic ratios of long-term frequencies. Hard to justify when no such long-term makes sense (*e.g.*, cosmology!)
- Frequentist results can depend on likelihood of "unobserved data" (*i.e.*, often don't obey the "likelihood principle")
- Conversely, the Bayesian approach is accused of subjectivity (*i.e.*, non-scientific!)
  - (but at least it's controlled subjectivity: need to explain the assumptions)

Probability and frequency

- But in the absence of other information, you can and do use "number counts" to estimate underlying distributions (but beware of selection effects).
  - detailed way to do this — density estimation — notoriously difficult!
  - Easier with parameterised models. E.G.,
    - Gaussian
    - Schechter function for luminosity distribution



Raddick, SDSS

Bayesian/Frequentist Correspondence

- Why do both methods seem to work?
- frequentist mean ~ likelihood maximum
  frequentist variance ~ likelihood curvature
- Correspondence is *exact* for
  - linear gaussian models (mapmaking)
  - variance estimation with no correlations and "iid" noise — simple version of C_l problem
    - e.g., all sky, uniform noise
    - likelihood only function of d_{lm}^2
    - breaks down in realistic case of correlations, finite sky, varying noise
  - "asymptotic limit"
    - ~ high l iff noise correlations not "too strong"
- But we still want to bootstrap from point estimates to the full likelihood function

Likelihoods & Priors

- Hard part is deciding how to parameterize the theory:
  - The more strongly the data depends on the theory, the easier to test
    - How do we deal with P(house prices|sunspot activity)??
    - When the data are uninformative, the prior dominates
  - Conversely, the parameterization doesn't have to represent *physics* at all
    - can test *correlation* in the absence of *causation*

Likelihood functions P(data|theory, I)

- *a.k.a.* "sampling distributions":
  - Probability of getting the actual observed data given the theory we are trying to test
  - Hypergeometric distribution and relatives (binomial, multinomial) when sampling from a population with distinct properties (*number of giant ellipticals in a cluster*)
  - Poisson distribution for *n* counts sampled from some rate *r (photons from a source)*
  - Gaussian/Normal distribution, arises from sum of small, unknown effects (signal+noise)

Prior probabilities

- Informative: value of $T_{\rm cmb}$, $m_{\rm electron}$, etc.
- Uninformative:
  - Uniform distribution for "location parameters"
    - $P(x|I)\,dx = dx/(x_{\rm max}\text{-}x_{\rm min})$ if $x_{\rm min} < x < x_{\rm max}$
  - Uniform in log for 'scale parameters'
    - $P(\sigma|I) \propto d\sigma/\sigma = d\ln\sigma$ if $\sigma_{\rm min} < \sigma < \sigma_{\rm max}$
  - Maximum entropy
    - (Gaussian/Normal distribution)
- Can be tested (Bayesian model comparison)

Priors

- Depend on parameterization!

$$p(f|I)\,df = p(x|I)\,dx = \left[p(x|I)\left(\frac{df}{dx}\right)^{-1}\right]_{x=x(f)} df$$

  - Jacobian: $f=f(x)$
  - (can also be derived from product rule)
- *The better the data, the less important the prior*
- Be careful of "improper" priors
  - Ratio of limits ≠ limit of ratios!

Poisson rates

- Likelihood: probability of observing *n* counts if the rate is *r*

$$P(n|rI) = \frac{e^{-r} r^n}{n!}$$

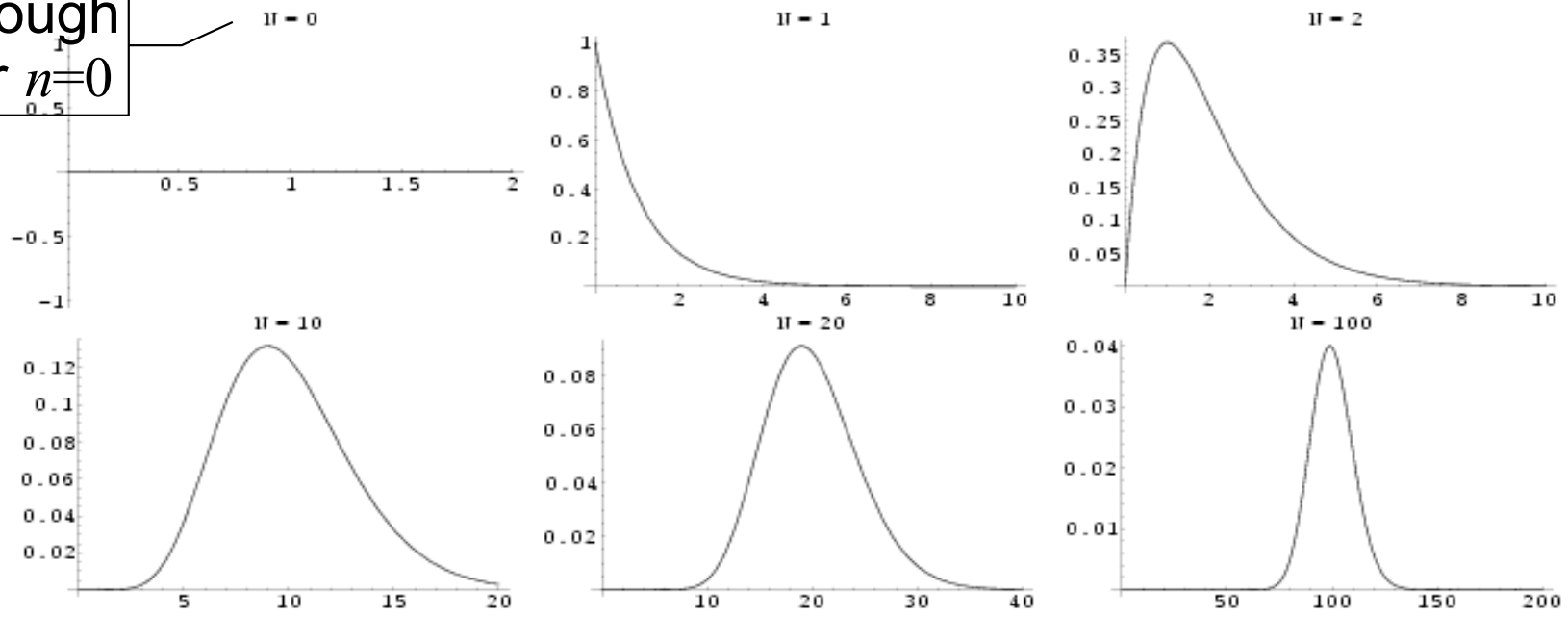
- Posterior: probability that rate is *r* given *n* counts

$$P(r|nI) = \frac{e^{-r} r^{n-1}}{(n-1)!}$$

- nb. $(n\text{-}1)$ comes from $p(r|\mathrm{I})\, dr \propto dr/r$
  - Uniform in log — scale parameter

Inferences for a Poisson rate

Not enough info for $n=0$



Infer: $r = n \pm \sqrt{n}$ (mean±√variance)

Note "asymptotic gaussianity" for large N

Poisson rates

- Complications [see Loredo articles]
  - Backgrounds: $n = b + s$
    - *Can solve for/marginalize over known or unknown* b
    - e.g., n_b counts from time T_b spent observing background rate b, n_s from T_s spent observing $(s+b)$
    - (e.g., Loredo)
  - Spatial or temporal variation in the signal (or background): $s=s(t)$

The Gaussian Distribution

$$P(x|\mu\sigma I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

- Moments: $\langle x \rangle = \mu \qquad \langle (x-\mu)^2 \rangle = \sigma^2$
  - all higher cumulants $\kappa_n = 0$
- Central Limit Theorem
  - Arises very often: sum of many independent "random variables" tends to Gaussian
  - Additive noise is often well-described as Gaussian
- Maximum Entropy
  - Bayesian interpretation: if you know only the mean and variance, Gaussian is the "least informative" consistent distribution.

Inference from a Gaussian: Averaging

- Consider *data* = *signal* + *noise*,
- $d_i = s + n_i$
  - Noise, n_i, has zero mean, known variance σ^2
    - Assign a Gaussian to $(d_i - s)$
      - Alternately: keep n_i as a parameter and marginalize over it with $p(d_i|n_i\, s\, I) = \delta(d_i\text{-}n_i\text{-}s)$
  - Prior for s (*i*.e., a and b)?
    - To be careful of limits, use Gaussian with width Σ, take $\Sigma \to \infty$ at end of calculation
      - Same answer with [improper] unifom dist'n in $(-\Sigma_1,\Sigma_2) \to (-\infty,\infty)$
        - $P(s|\mathrm{I}) = \mathrm{const}$

→calculation

Inference from a Gaussian: Averaging

- Posterior:

$$P(s|dI) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{1}{2}\frac{(s-\bar{d})^2}{\sigma_b^2}\right]$$

  - best estimate of signal is average ± stdev:
    - $s = \bar{d} \pm \sigma_b = \bar{d} \pm \sigma/\sqrt{N}$
  - What if we don't know σ? try Jefferys $P(\sigma|I) \propto 1/\sigma$
    - marginalized $P(s|I) \propto [s - 2s\langle d\rangle + \langle d^2\rangle]^{-1/2}$
    - (very broad distribution!)

Inference from a Gaussian: Straight-line fitting

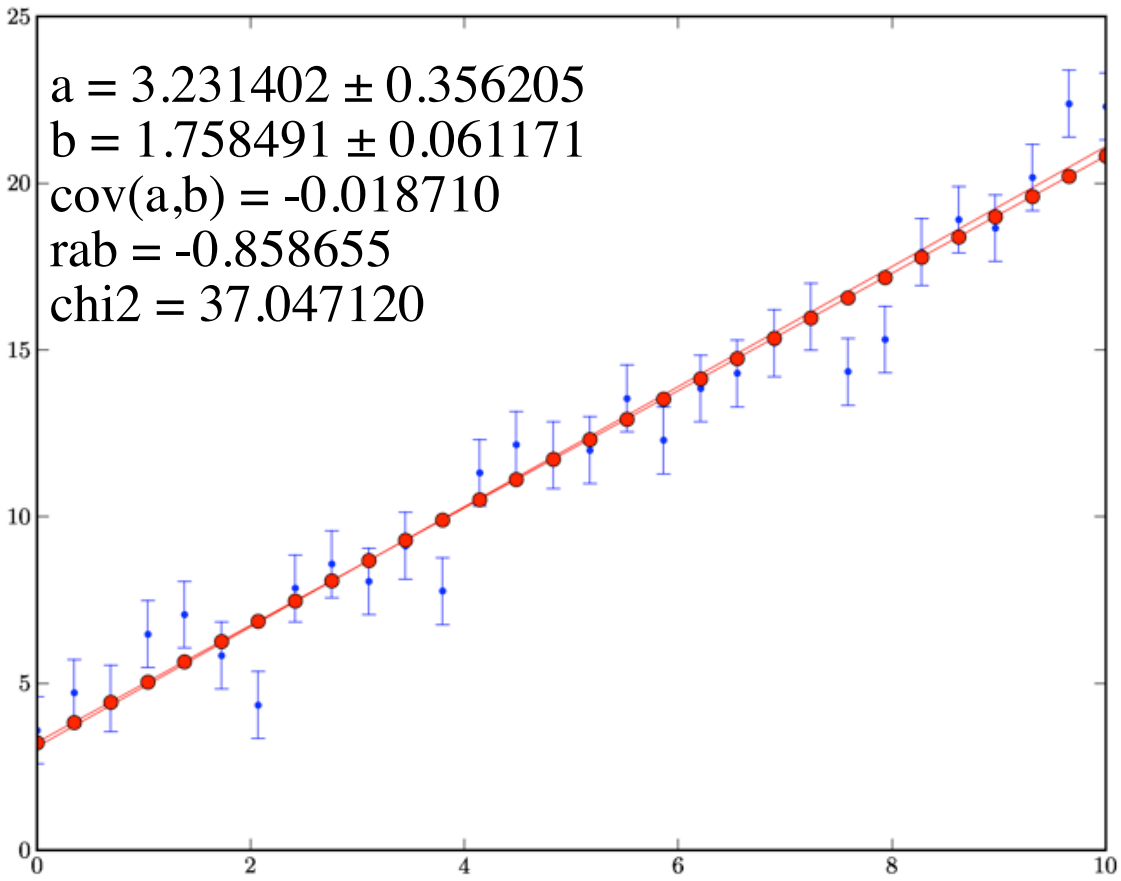
- Now consider *data* = *signal* + *noise*, where signal depends linearly on time:
  - $d_i = at_i + b + n_i$, with "iid" gaussian noise $\langle n_i \rangle = 0$; $\langle n_i^2 \rangle = \sigma^2$
- Likelihood function is

$$P(d|a,b,I) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(d-at_i-b)^2}{\sigma^2}\right]$$

  - Multivariate gaussian in *d*
  - Linear in (a,b): also has form of a multivariate gaussian in (a,b)
    - but not a distribution in (a,b) until you apply Bayes' theorem and add a prior
  - Maximized at the value of the "least squares" est. for (a,b), with the same numerical values for the errors (& covariance)
    - (but, recall, with a very different interpretation of those errors)

Inference from a Gaussian: Straight-line fitting

- This means that for these problems you can just use usual canned routines...



General linear models (I)

- Consider $d(t_i) = \sum_p x_p f_p(t_i) + n_i$
  i.e., a sum of known functions with unknown amplitudes, plus noise — want to estimate a_p
  - e.g., linear fit: $f_0(t)=1, f_1(t)=t$
- assume zero-mean Gaussian noise, possibly correlated: $\langle n \rangle = 0$, $\langle n_i\, n_j \rangle = \mathbf{N}_{ij}$
  - typically, noise is stationary (isotropic): $\mathbf{N}_{ij} = N(t_i - t_j)$
- rewrite in matrix-vector form:

$$d_i = \sum_p A_{ip} x_p + n_i \qquad \text{with } A_{ip} = f_p(t_i)$$

- Likelihood:

$$P(d_i | x_p I) = \frac{1}{|2\pi N|^{1/2}} \exp\left[-\frac{1}{2}(d - Ax)^T N^{-1} (d - Ax)\right]$$

General linear models (II)

$$d_i = \sum_p A_{ip} x_p + n_i \qquad \text{with } A_{ip} = f_p(t_i)$$

complete the square

- Can rewrite the likelihood as

$$P(d_i|x_p I) \propto \exp\left[-\frac{1}{2}(d - A\bar{x})^T N^{-1}(d - A\bar{x})\right] \times \exp\left[-\frac{1}{2}(x - \bar{x})^T C^{-1}(x - \bar{x})\right]$$

$$\propto \underbrace{\exp\left[-\frac{1}{2}(d - AWd)^T N^{-1}(d - AWd)\right]}_{\textit{depends on data, not params}} \times \underbrace{\exp\left[-\frac{1}{2}(x - Wd)^T C^{-1}(x - Wd)\right]}_{\textit{depends on data and params}}$$

- with $W = (A^T N^{-1} A)^{-1} A^T N^{-1}$ and $C = (A^T N^{-1} A)^{-1}$
- Parameter-independent factor is just $e^{-\chi^2_{\rm max}}$
- Parameter-dependent factor shows that likelihood is multivariate Gaussian with mean

$$\bar{x} = Wx = (A^T N^{-1} A)^{-1} A^T N^{-1} d$$

and variance C

General linear models (III)

- In limit of an infinitely wide uniform (or Gaussian) prior on x:

$$P(x_p|dI) = \frac{1}{|2\pi C|^{1/2}} \exp\left[-\frac{1}{2}(x - Wd)^T C^{-1}(x - Wd)\right]$$

  nb. normalization cancels out $e^{-\chi^2_{\rm max}}$

- Covariance matrix $\langle \delta x_p \, \delta x_q \rangle = C_{pq}$ gives error $\sigma_p^2 = C_{pp}$ if we *marginalize* all other parameters.
- Inverse covariance gives error $\sigma_p^2 = 1/C^{-1}_{pp}$ if we *fix* other parameters
  - nb. marginalization doesn't move mean (max) values *for this case*
  - cf. Fisher matrix $F \Leftrightarrow C^{-1}$
- Aside: with a finite Gaussian prior on x, can derive the *Wiener filter*, as well as power-spectrum estimation formalism (see tomorrow's lecture on the CMB)

Correlation

- Linear model posterior: *multivariate Gaussian*

$$P(x_i|\mu_i M_{ij} I) = \frac{1}{|2\pi M|^{1/2}} \exp\left[-\frac{1}{2}(x_i-\mu_i)M_{ij}^{-1}(x_j-\mu_j)\right]$$

$$\langle x_i \rangle = \mu_i \qquad \langle (x_i-\mu_i)(x_j-\mu_j)\rangle = M_{ij}$$

- Correlation: non-zero off-diag elements of M_{ij}
  - measured by $\rho_{ij}=M_{ij}/\sqrt{(M_{ii}M_{jj})}$

Careful: correlation vs covariance vs dependence

Chi-squared

- The exponential factor of a Gaussian is always of the form $\exp(-\chi^2/2)$
- Likelihood: $\chi^2=\sum(\text{data}_i - \text{model}_i)^2/\sigma_i^2$
- For fixed model, χ^2 has χ^2 distribution for $\nu=N_{\text{data}}-N_{\text{parameters}}$ "degrees of freedom"
  - peaks at $\chi^2=\nu\pm\sqrt{(2\nu)}$
- model may be bad if χ^2 is too big
  - *or* too small ("overfitting" — too many parameters)
- (frequentist argument, but good rule of thumb)

Bivariate gaussians and covariance

- Covariance matrix given by

$$M = \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix}$$

- Correlation coefficient: $-1 \le \rho \le 1$
  - uncorrelated and independent if Gaussian and $\rho = 0$

can always "rotate" to uncorrelated parameters: eigenvectors of M (not just 2d)

- Diagonal entries give *marginalized* variances
- Error ellipse

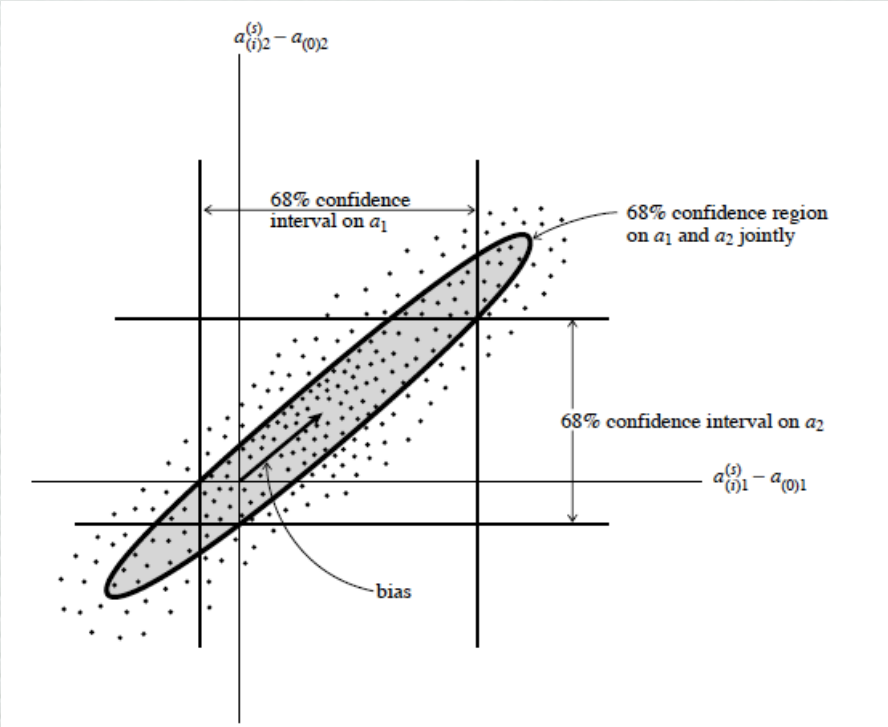


Figure 15.6.3. Confidence intervals in 1 and 2 dimensions. The same fraction of measured points (here 68%) lies (i) between the two vertical lines, (ii) between the two horizontal lines, (iii) within the ellipse.

Numerical Recipes

Errors: the Gaussian approximation

- If we assume uniform priors, then the posterior is proportional to the likelihood.

If further, we assume that the likelihood is single-moded (one peak at θ_0) , we can make a Taylor expansion of lnL:

$$\ln L(x;\theta) = \ln L(x;\theta_0) + \tfrac{1}{2}(\theta_\alpha - \theta_{0\alpha})\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta}(\theta_\beta - \theta_{0\beta}) + \ldots$$

$$L(x;\theta) = L_0 \exp\left[-\tfrac{1}{2}(\theta_\alpha - \theta_{0\alpha})H_{\alpha\beta}(\theta_\beta - \theta_{0\beta}) + \ldots\right]$$

where the Hessian matrix is defined by these equations. Comparing this with a gaussian, the *conditional error* (keeping all other parameters fixed) is

$$\sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}$$

Marginalising over all other parameters gives the *marginal error*

$$\sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}$$

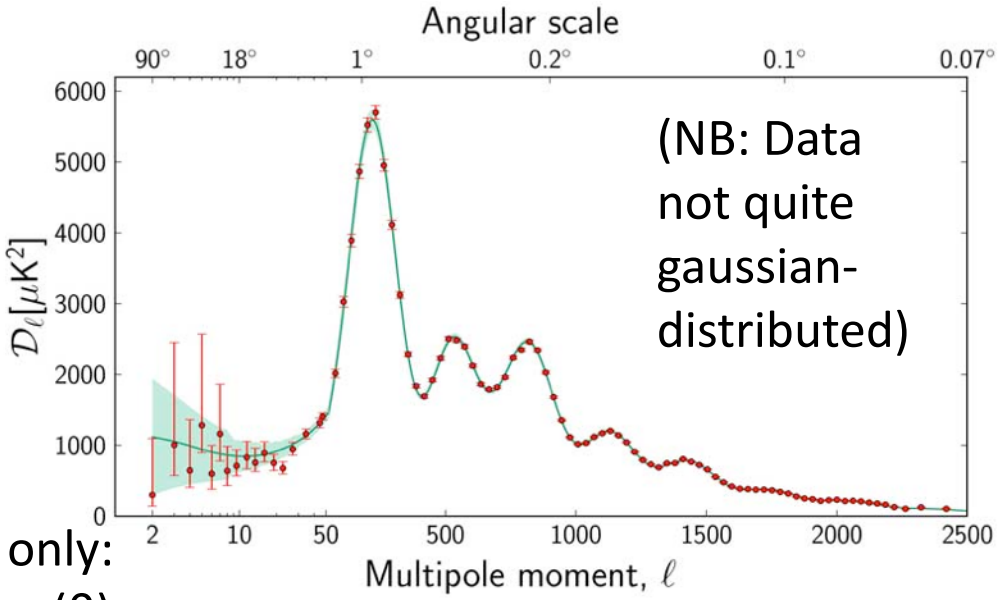
ICIC

Fisher Matrices

- Useful for forecasting errors, and experimental design
- The likelihood depends on the data collected. Can we estimate the errors before we do the experiment?
- With some assumptions, yes, using the Fisher matrix

$$F_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle$$

- F ~ (correlation matrix)$^{-1}$

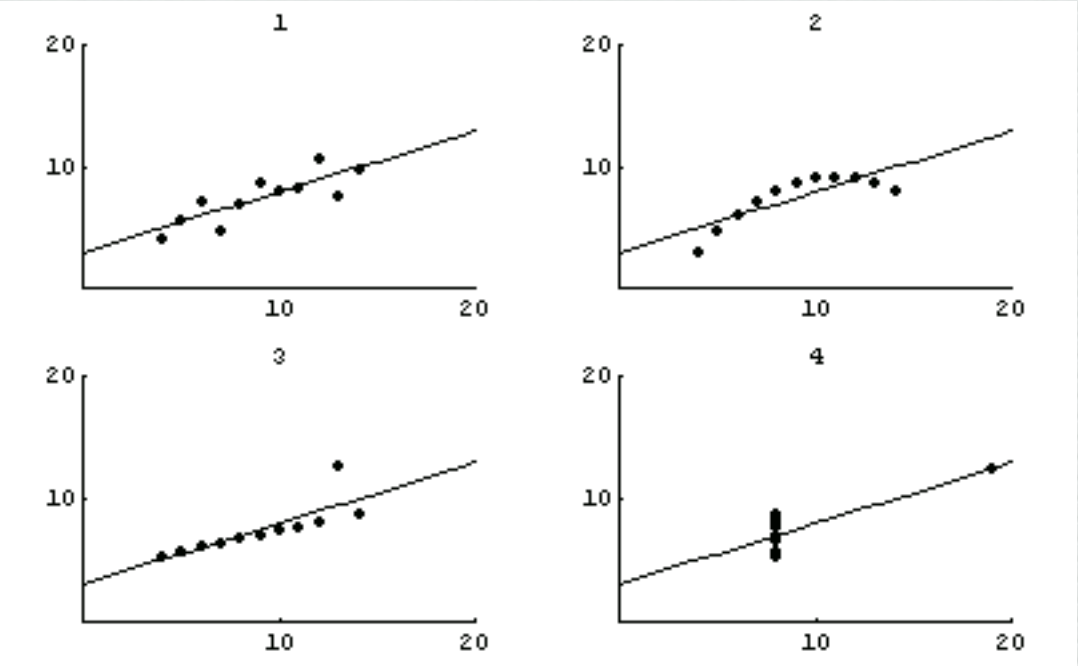


For gaussian data, we need to know only:
1. The expectation value of the data, μ(θ)
2. The covariance matrix of the data, C(θ)

ICIC

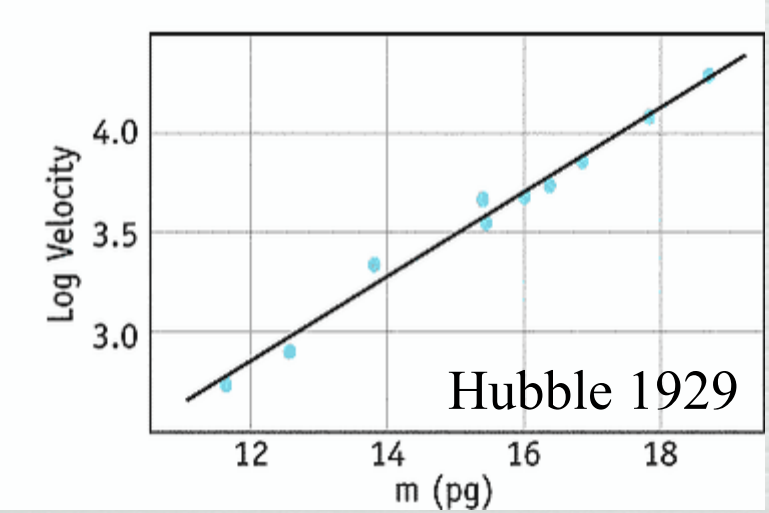
Beware

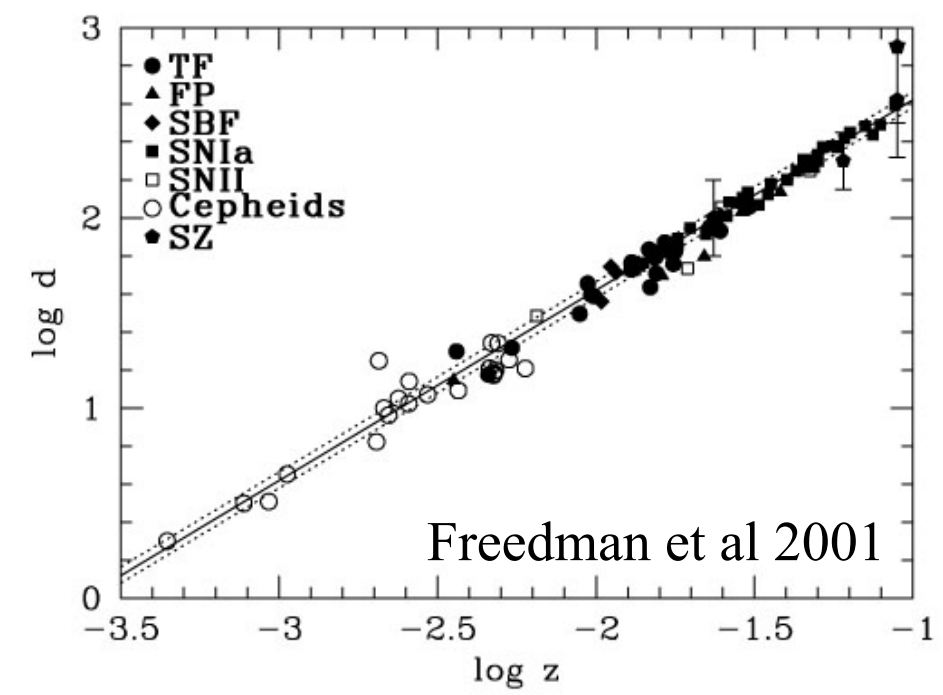
- Your fit is only as good as your model
  - (Anscombe, c/o Wolfram)



Beware

- Your fit is only as good as your model
  - (Anscombe, c/o Wolfram)
  - Not just an academic point:





Selection effects

- Not all correlation is physical!

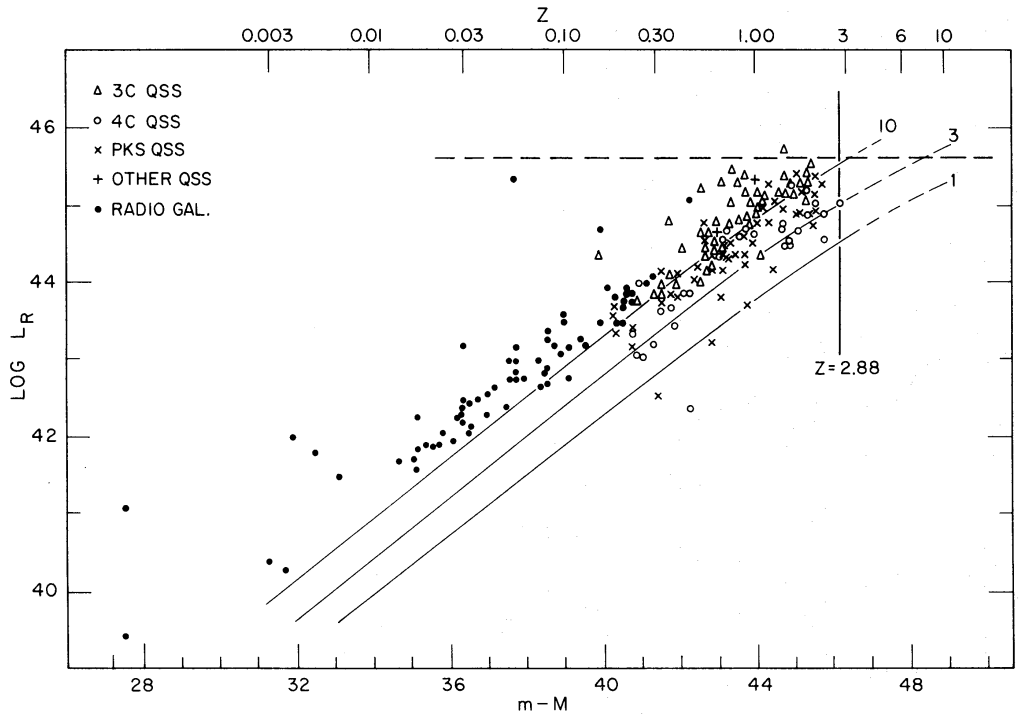


FIG. 7.—The radio power plotted against redshift for radio galaxies from table 2 and quasars from tables 3 and 4. Quasars are *triangles* if from the 3C catalog, *open circles* if 4C, *crosses* if Parkes, and *vertical crosses* if others. The radio galaxies are *dots*. Lines of apparent flux densities are shown at 10, 3, and 1 flux units at 178 MHz. No part of the area between the f = 3. f.u. line and the upper-envelope line at 4×10^{45} ergs s^{-1} should be denied the observer by selection effects if the 4C catalog is used. The lack of redshifts larger than $z \simeq 2.8$ in the area to the right of the vertical line suggests that such redshifts do not occur.

Sandage 1972

Gaussian Processes

- Simplifications:
  - a linear combination of Gaussian-distributed variables is also Gaussian-distributed
  - cf. the central limit thm: a linear combination of [lots of, almost] any variables is Gaussian-distributed
- Extensions/complications
  - What if we don't know σ? try Jefferys $P(\sigma|\mathrm{I})d\sigma \propto d\sigma/\sigma$
  - Variable noise: allow $\sigma=\sigma_i$
  - Correlated noise: correlation matrix $\langle n_i n_j \rangle = N_{ij}$
  - Generalized Least-squares: Curve fitting, CMB Mapmaking
    - $d_i = A_{im} s_m + n_i$
      - determine s_m for known A_{im}
      - Equivalent to χ^2 minimization for "uninformative" prior on s

Power spectra: Stationarity and Isotropy

- Let's assume a "random process" as our model for the *underlying signal*.
  - e.g., fractional overdensity $\delta(\boldsymbol{x}) \equiv \delta\varrho(\mathbf{x})/\bar{\varrho}$
  - The signal is the function $\delta(\boldsymbol{x})$ — an infinite-dimensional vector
    - correlation matrices/tensors ⇒ multivariate functions
  - mean $\langle\delta\rangle = 0$, variance $\langle\delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle$
  - statistical isotropy: $\langle\delta(\boldsymbol{x})\delta(\boldsymbol{y})\rangle = \xi(|\boldsymbol{x}\text{-}\boldsymbol{y}|)$
    - depends only on distance between points
    - go to Fourier domain — $\langle\tilde{\boldsymbol{\delta}}(\boldsymbol{k})\tilde{\boldsymbol{\delta}}(\boldsymbol{q})\rangle=(2\pi)^3\boldsymbol{\delta}_{\mathrm{Dirac}}(\boldsymbol{k}+\boldsymbol{q})\mathrm{P}(k)$
      - i.e., equivalent to a diagonal matrix with same entry for all $k=|\mathbf{k}|$

Random processes in the early Universe

- The distribution of matter (and curvature) seems to be consistent* with small fluctuations produced by a statistically isotropic Gaussian random process, evolving under interactions due to known laws of gravity and particle physics
  - E.G., from a weakly-coupled scalar field in an almost-de Sitter Universe — ***inflation***

- *What does it mean to be "consistent with a distribution" — any field is a *possible* distribution with infinitesimal (measure zero) probability

Cosmological power spectra

- Power spectrum gives the variance:

$$Pr\left(\tilde{\delta}_{\mathbf{k}}|P(k)\right) = \frac{1}{\sqrt{2\pi P(k)}} \exp\left(-\frac{1}{2}\frac{\left|\tilde{\delta}_{\mathbf{k}}\right|^2}{P(k)}\right)$$

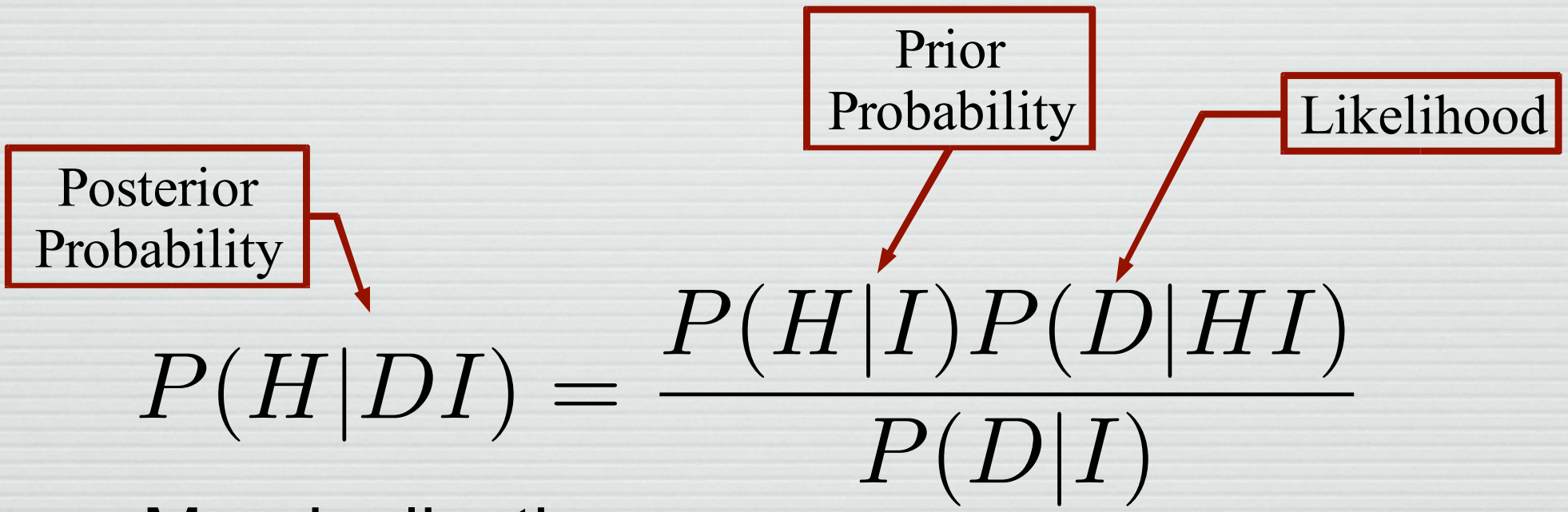
- Spectrum will itself depend upon cosmological parameters: $P(k) = P(k; \Omega_m, \Omega_\Lambda, H_0, \ldots)$
- If linear evolution $\rho \Rightarrow a_{\ell m}$ (CMB fluctuation)

$$Pr(a_{\ell m}|C_\ell) = \frac{1}{\sqrt{2\pi C_\ell}} \exp\left(-\frac{1}{2}\frac{|a_{\ell m}|^2}{C_\ell}\right)$$

- CMB power spectrum $\left\langle a^*_{\ell m} a_{\ell' m'} \right\rangle = \delta_{\ell\ell'}\delta_{mm'}\, C_\ell$

Recap

- Bayes' Theorem:



- Marginalization:

$$P(\theta|DI) = \int d\varphi \, P(\theta\varphi|DI)$$

Bayesian Model Comparison

- Until now, given a model, measure its parameters
- Move "up" a level: choose between models
  - Deuterium line or interloper?
  - Flat universe or curved?
  - Dark Energy or cosmological constant?
  - Is a given star/galaxy a member of a cluster or a superposition?
  - Dark matter or MOND?
  - (nb. not just between two)
- But really, just apply the same machinery

Bayesian Model Comparison

- How do we tell if our model (choice of parameters, θ) is a good description of the data?
- Need to specify alternatives: can choose amongst models (but no pure "goodness-of-fit" test)
- Let the prior information be $I = I_0 (I_1 + I_2 + ...)$
  - common information (I_0) and a choice between Model 1(I_1), Model 2 (I_2), ...
  - Now, use Bayes' thm to get P(I_i | data)

Bayesian Model Comparison

- Full set of parameters are then
  - *i*: choose between models
  - θ_i: parameters for each model
    - (can be different for each model – and different numbers of parameters per model)
- Joint likelihood for model *i* and its parameters:

$$P(i\theta_i|DI) \propto P(i|I)P(\theta_i|I_0I)P(D|\theta_iI_0I)$$

Bayes' theorem and model comparison

- Marginalize over parameters θ_i:

$$P(i\theta_i|DI) \propto P(i|I)P(\theta_i|I_0 I)P(D|\theta_i I_0 I)$$

but recall usual Bayes' thm:

$$P(\theta|DI)\,d\theta = \frac{P(\theta|I)P(D|\theta I)}{\int d\theta'\; P(\theta'|I)P(D|\theta' I)}d\theta$$

so

$$\propto \frac{P(\theta|I)P(D|\theta I)}{P(D|I)}d\theta$$

- $P(i|DI) \propto P(i|I)P(D|II_i)$ — just the normalization!

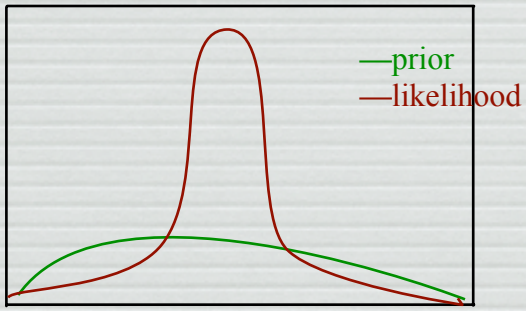
Model likelihood (sometimes called "evidence")

Model Comparison

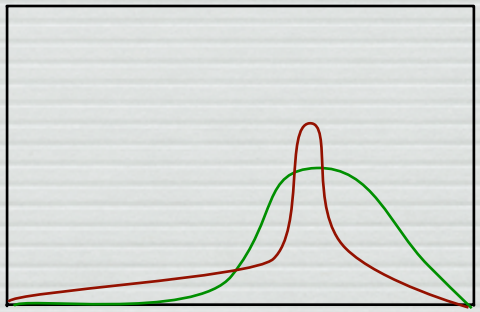
- model probability $\propto$ average likelihood, weighted by prior
- automatic penalty for more complicated models (≡ more parameter 'volume')

$$P(i|DI) \propto P(i|I)P(D|II_i)$$

$$= P(i|I)\int d\theta_i P(\theta_i|II_i)P(D|\theta_i I_i I)$$



likelihood strongly-peaked compared to prior, but better "best fit"



Ockham's Razor

$$P(i|DI) = P(i|I) \int d\theta_i P(\theta_i|II_i) P(D|\theta_i I_i I)$$

$$\simeq P(i|I) P_{\max}(D|\theta_i I_0 I_i) \frac{\text{posterior volume}}{\text{prior volume}}$$

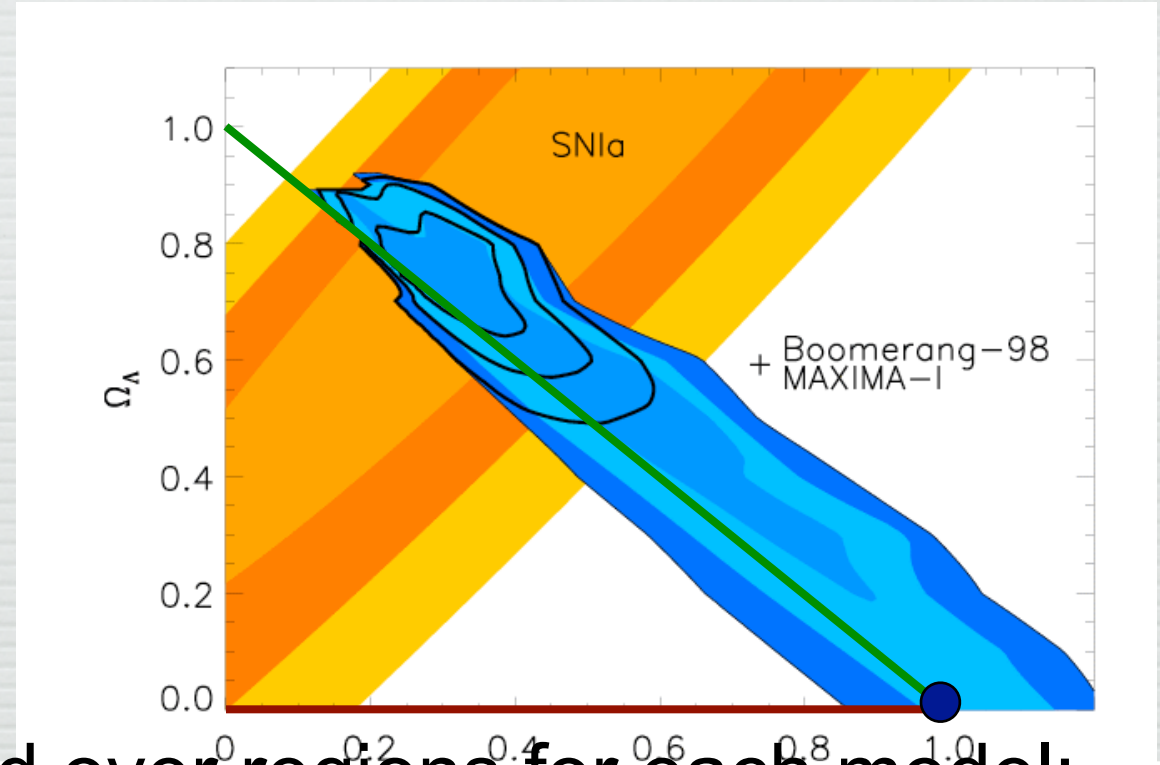
favors better-fitting model
(often, more complicated one)

Favors simpler model
"Ockham Factor"

- must have *proper prior distributions* for this to make sense
- Can't go to "uninformative limit"

Application: is the Universe flat?

- *nested models*:
  - old std CDM: $\Omega_\Lambda=0$, $\Omega_m=1$
  - flat: $\Omega_\Lambda+\Omega_m=1$
  - $\Omega_\Lambda=0$, $0\leq\Omega_m\leq1$
  - $0\leq\Omega_m\leq1$, $0\leq\Omega_\Lambda\leq1$



  - Integrate likelihood over regions for each model:
    - CMB alone prefers both std CDM & flat
    - CMB+SNe prefers flat
      - (would really prefer $\Omega_\Lambda=0.7$, $\Omega_m=0.3$, but that's not an *a priori* model that would occur to us!)

Sampling from the posterior

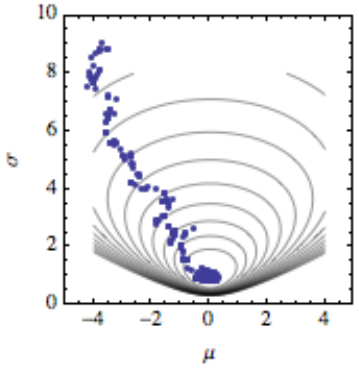
- Infeasible to directly explore P($\boldsymbol{\theta}$|data) for many parameters $\boldsymbol{\theta}$
  - e.g., even the 6-parameter base LCDM model would require ~$100^6 = 10^{12}$ evaluations for 100 grid points in each direction…
- Instead, *generate samples* $\boldsymbol{\theta}_i$ from the distribution.
  - Easy to evaluate moments (means, variances)

$$\langle \theta \rangle = \frac{1}{N} \sum_i \theta_i \quad \text{or, more generally} \quad \langle f(\theta) \rangle = \frac{1}{N} \sum_i f(\theta_i)$$

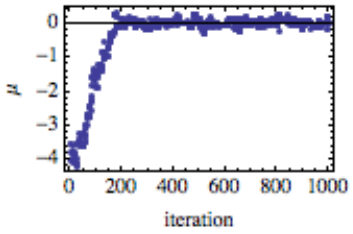
MCMC

- Generate samples from posterior P(x)
- Most methods require being able to generate samples from some simpler distribution
- e.g., Markov Chain Monte Carlo
  - Start with proposal distribution $Q(x^*|x)$: probability of proposing point x^* if starting at point x
  - often $Q(x|y) = Q(|x-y|)$ (Metropolis)
    - Metropolis Algorithm:
      - given point $x^{(i)}$, generate x^* from $Q(x^*|x^{(i)})$
      - accept x^* as $x^{(i+1)}$ with probability min[1,P(x^*)/P($x^{(i)}$)];
      - otherwise $x^{(i+1)} = x^{(i)}$
      - repeat…

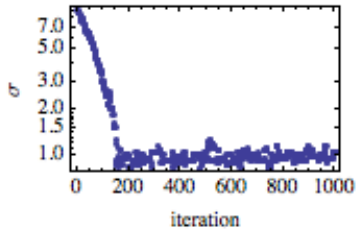
prop Sig[μ] = 0.2
prop Sig[σ] = 0.2
acceptance = 27.6%
μ = 0.0208 ± 0.0988
σ = 0.985 ± 0.074



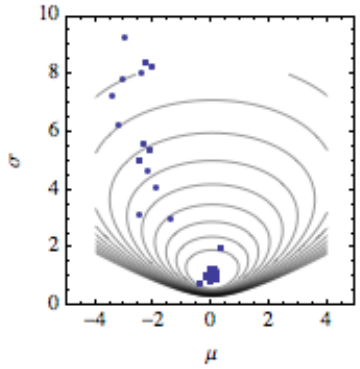
μ trace



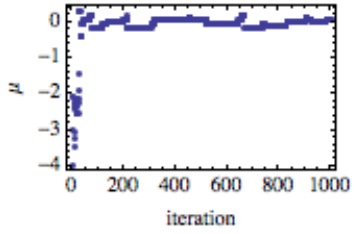
σ trace



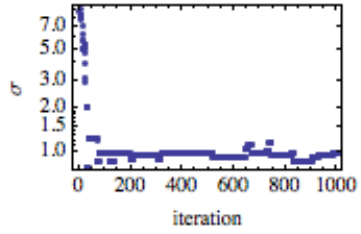
prop Sig[μ] = 0.8
prop Sig[σ] = 0.8
acceptance = 4.3%
μ = −0.0193 ± 0.0887
σ = 0.974 ± 0.0513



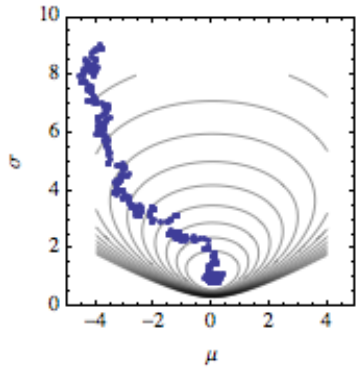
μ trace



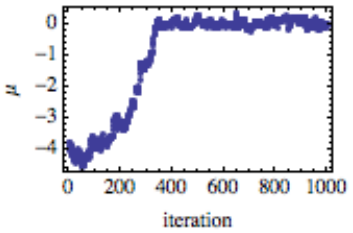
σ trace



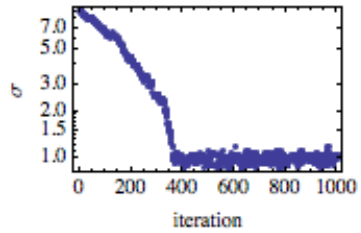
prop Sig[μ] = 0.1
prop Sig[σ] = 0.1
acceptance = 54.%
μ = −0.128 ± 0.497
σ = 1.2 ± 0.575



μ trace

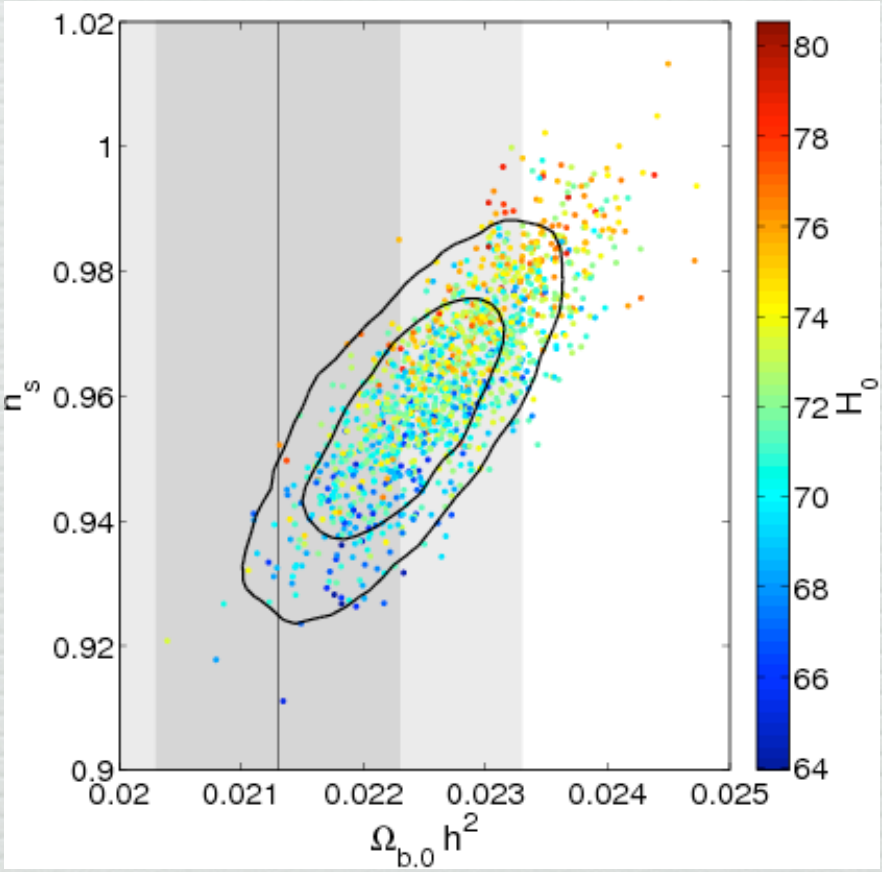


σ trace



Monte Carlo methods for the CMB

- Markov Chain Monte Carlo: A. Lewis' CosmoMC
  - coupled with fast deterministic calculation of power spectrum as fn of cosmological parameters
  - e.g. CMBFAST, CAMB, CLASS
  - Other techniques
    - e.g., Skilling's "nested sampling" which also allows fast calc'n of model likelihoods ("evidence")



Case Study: CMB Data analysis

- data

  $d_t = A_{tp}\, T_p + n_t$

- $A_{tp} = 1$ when observing pixel p at time t

  0 otherwise

- T_p = underlying (CMB?) temperature field

- n_t = noise

  $\langle n_t\, n_{t'} \rangle = N(t\text{-}t') \Rightarrow$ stationary, Gaussian noise

CMB Data analysis

- $d=AT+n$ is the data; what is the theory?
- Bayes' theorem lets you ask any question you want, and calculate $P(\theta|d)$
  - $\theta = T_p \rightarrow$ the CMB map
  - $\theta = C_\ell \rightarrow$ the CMB power spectrum

$$\langle T_p T_{p'} \rangle = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell B_\ell^2 P_\ell(\hat{x}_p \cdot \hat{x}_{p'})$$

  - $\theta = \{\Omega_m, \Omega_\Lambda, \Omega_b, \Omega_{\rm tot}, H_0, n_s, \sigma_8, ...\} \rightarrow$ cosmological parameters
  - in principle, separate questions to be asked of the data; in practice, can ask them in sequence

Conclusions

- [Bayesian] probability gives a framework for learning from data
  - sharpening distributions: prior ⇒ posterior
- Probability measures degrees of belief
  - "contextual" rather than "subjective": $P(\mathrm{H}|\mathrm{D}\mathbf{I})$
  - with same background information **I**, different agents will agree on probability assignment
- Need models:
  - from theoretical quantities and instrumental parameters to observables (signal and noise)
  - probabilistic and/or deterministic
- Write down what you know!

Likelihood Function

- $d_t = A_{tp} T_p + n_t$
  - $\langle n_t \, n_{t'} \rangle = N_{tt'} = N(t\text{-}t')$ [Fourier Tr. of $N(f)$]
  - stationary, Gaussian noise:
    - assign Gaussian Likelihood:

$$P(d|TI) = \frac{1}{|2\pi N|^{1/2}} \exp\left[-\frac{1}{2}(d - AT)^T N^{-1} (d - AT)\right]$$

this is a “generalized linear model” or “generalized least squares” problem!

Making CMB Maps

- what is $P(T_p|d_t I)$?
  - assign uniform $P(T|I)$, complete the squares, &c

$$P(T_p|d_t I) = \frac{1}{|2\pi C_N|^{1/2}} \exp\left[-\frac{1}{2}(T-\bar{T})^T C_N^{-1}(T-\bar{T})\right]$$

$$C_N = \left(A^T N^{-1} A\right)^{-1}$$

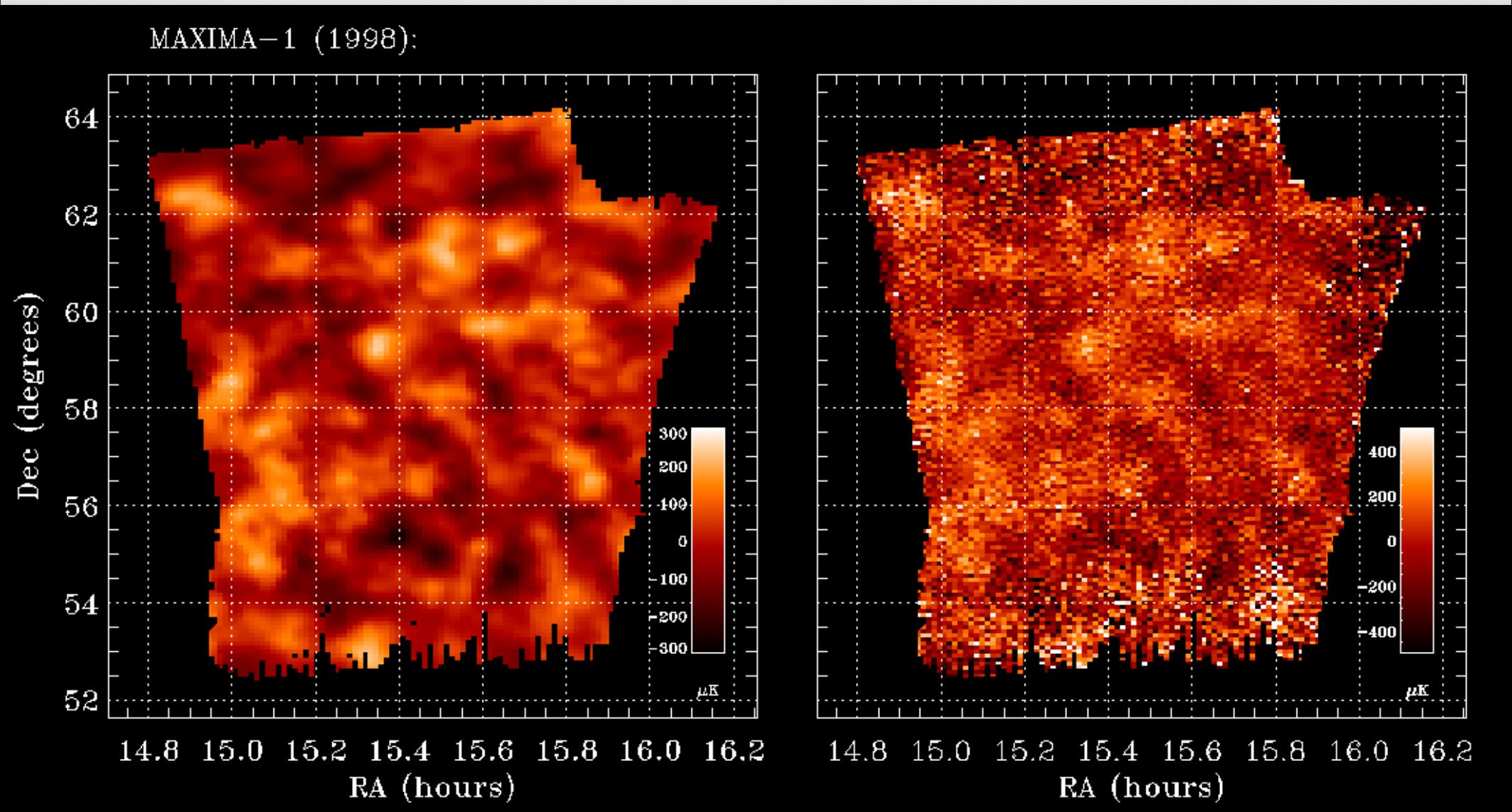
$$\bar{T} = C_N A^T N^{-1} d$$

"Sum of weights"

"Weighted average of data"

  - Multivariate gaussian – least squares solution
    - $O(N_{pix}^2)$ for map, $O(N_{pix}^3)$ for map+C_N
  - Aside: can assign Gaussian prior to T
    - Wiener filter

MAXIMA-I


MAXIMA−1 (1998):
Dec (degrees)
64
62
60
58
56
54
52
14.8 15.0 15.2 15.4 15.6 15.8 16.0 16.2
RA (hours)
300
200
100
0
−100
−200
−300
μK
14.8 15.0 15.2 15.4 15.6 15.8 16.0 16.2
RA (hours)
400
200
0
−200
−400
μK

From maps to power spectra

- Max-likelihood Map is a "sufficient statistic"
  - Likelihood only depends on data through ML map:

$$\bar{T} = C_T A^T N^{-1} d$$

    - Can ignore full data vector d (billions of points) in favor of map (millions)
- Now:

$$\begin{aligned} P(C_\ell|dI) &= \int dT\ P(C_\ell T|dI) \\ &= \int dT\ P(C_\ell|\bar{T}I)\ P(T|C_\ell \bar{T} I) \\ &\propto \int dT\ P(C_\ell|I) P(T|C_\ell I) P(\bar{T}|TI) \end{aligned}$$

C_ℓ prior

T prior given C_ℓ: Gaussian w/ $\langle \mathrm{TT} \rangle \Leftrightarrow C_\ell$

Likelihood

Spectrum Estimation: Gaussianity and non-Gaussianity

- Signal prior, $\mathrm{P}(T|C_\ell\, I)$:
  - Assign the Gaussian distribution $a_{\ell m} \sim \mathrm{N}[0,\, C_\ell]$
    - € $\langle |a_{\ell m}|^2 \rangle = C_\ell \rightarrow \mathrm{C}_{T,pp'} = \langle T_p\, T_{p'} \rangle$
  - normal/Gaussian distribution is MaxEnt
    - "objective" [?] prior if variance doesn't depend on *m*
      - *isotropy*
  - But how could we write the distribution if we do want to estimate higher moments?
    - Open problem...

From Maps to C_ℓ

- map + Gaussian noise & signal:
- $$P(\bar{T}|C_\ell I) = \frac{1}{|2\pi M|^{1/2}} \exp\left(-\frac{1}{2}\bar{T}^T M^{-1}\bar{T}\right)$$

  with covariance matrix
  - $M_{pp'} = C_{T,pp'}(C_\ell) + C_{N,pp'}$
- complicated function of C_ℓ
  - Iterative, "quadratic" algorithms (e.g., Newton-Raphson)
  - Matrix manipulations: $O(\#_{pixel})^3$ operations
    - $\#_{pixel}$ > 100,000 - 10^6 - 10^7 soon

From C_ℓ to cosmology

- Calculate & characterize posterior probability over some space of cosmological models and imposed priors
  - $P(\theta|DI) = \int dC_\ell \, P(\theta|I) \, P(C_\ell|\theta I) \, P(C_\ell|DI)$

    $= P(\theta|I) \, P(C_\ell[\theta] \mid DI)$

    $= P(\theta|I) \, P(C_\ell[\theta] \mid C_\ell^{\rm est}, \sigma_\ell, \text{shape}, I)$

    Variance

    ML est.
  - Complicated likelihood function back to haunt us
    - Not a Gaussian in C_ℓ (no "sufficient statstics"—simple $\chi 2$ inappropriate)
    - instead, can use (e.g.) offset lognormal approx.
      - $P(C_\ell[\theta] \mid C_\ell^{\rm est}, \sigma_\ell, \text{shape}) \sim N[\ln(C_\ell+x_\ell), \sigma_\ell/(C_\ell+x_\ell)]$

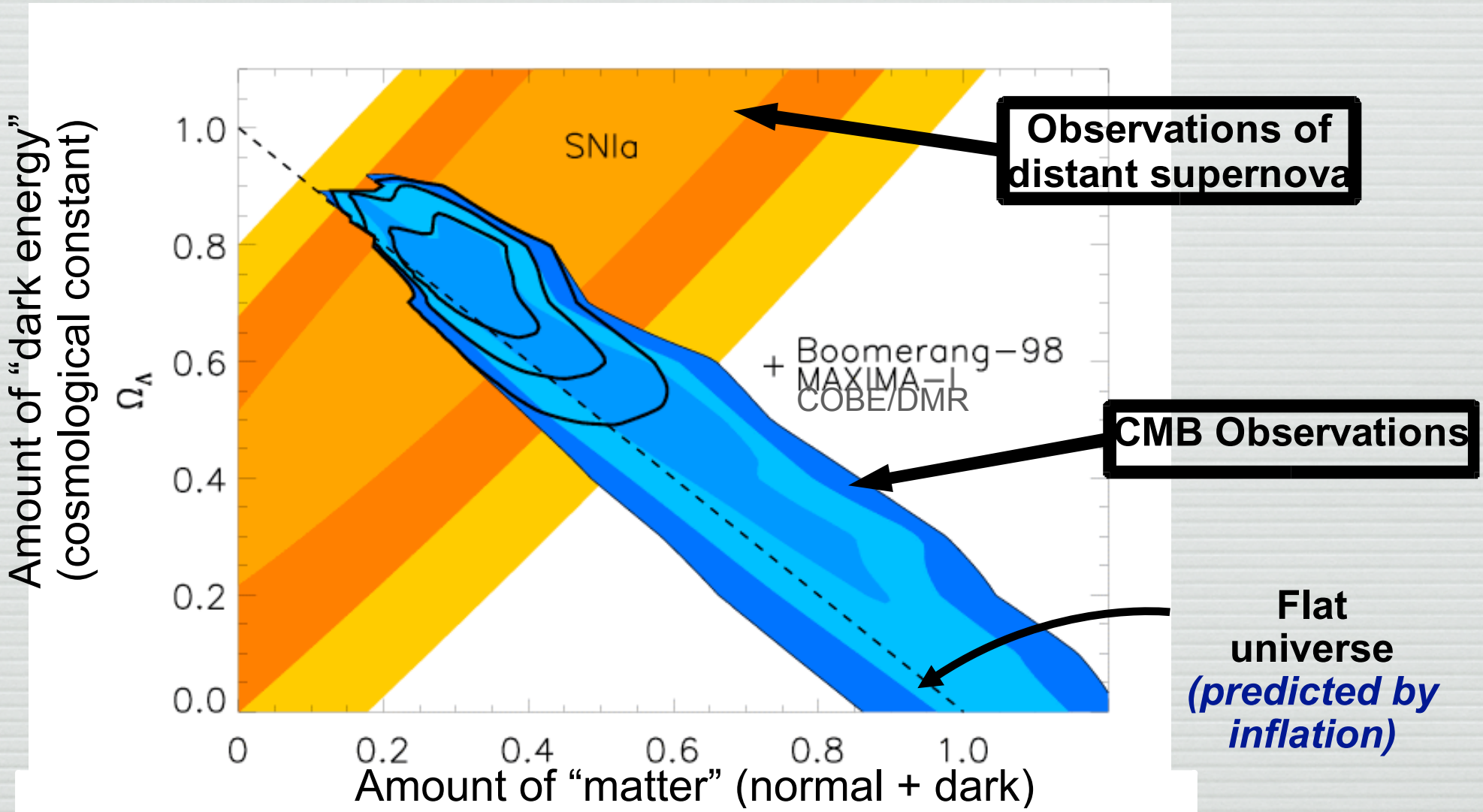
From C_ℓ to cosmology
From C_ℓ to cosmology

- Cosmological Parameters from $C\ell$=$C\ell[\theta]$
  - $\theta = \{\Omega_m, \Omega_\Lambda, \Omega_b h^0, \Omega_{\rm cdm} h^2, \tau, n_s, \sigma_8\}$
- Calculate & characterize posterior probability over some space of cosmological models and imposed priors
  - nuisance parameters: experimental beam, calibration
  - degeneracies: same C_ℓ from different parameters e.g., sensitive to $1\text{-}\Omega_m\text{-}\Omega_\Lambda$ rather than $\{\Omega_m, \Omega_\Lambda\}$

Cosmology from C_ℓ

- Calculate (approx.) likelihood over model grid
  - MAXIMA, BOOMERANG, CBI
- or sample from distributions using Markov-Chain Monte Carlo (MCMC)
  - dominated by $C_l[\theta]$ calculation [CMBFAST, Seljak & Zaldariagga; CAMB, Lewis et al]
- Marginalize over experimental parameters and subsets of cosmological parameters
- Priors ...dominate some results

The Constituents & Geometry of the Universe



Hierarchical Models

- The CMB problem turns out to be a hierarchical model:
  - ask progressively more complicated questions of the data, with (approximately) no dependence on the details of previous results
    - timestream →map →power spectrum →cosmology
    - also amounts to a "radical compression" of the data
      - (millions of samples →a few cosmological parameters)
    - model for scientific inference in many circumstances
    - but depends on details and assumptions (*e.g.*, more complicated if there are non-cosmological contaminants in the "CMB" data)

Hierarchical Models

- Paradigm for most measurements
  - get raw data from the instrument
    - which may actually involve some on-board averaging, implicitly using Gaussian model from last time
  - Data reduction to get spectra, images, etc.
  - Interpret these to extract spectral lines, object properties, etc.
  - Combine individual measurements (lines, objects)
  - Put measurements into scientific context
  - ...
- in principle, "sufficient statistics" at each step
  - in practice often "assume Gaussianity"

Homework

- Any obvious application of Bayesian methods to your work?
- Mock data & detailed problem sheet at http://astro.imperial.ac.uk/~jaffe
  - (available tomorrow)
  - linear fit on subset of data and whole

Key Points

- Probabilities are about information
- Bayes' Theorem:
  - posterior $\propto$ prior $\times$ likelihood
  - when in doubt, write down all the probabilities
- Linear models, Gaussian errors: χ^2 minimization
- Hierarchical models (CMB)
- Model comparison