Imperial College London

Probability: More Examples & Concepts

Andrew Jaffe

September 2013

ICIC

Road map

- Examples
  - Gaussian Linear models
  - Poisson statistics
- Confidence intervals
- Hierarchical Models
  - Nuisance parameters
  - Sufficient and nearly-sufficient statistics
- Model comparison: Model likelihood/Bayesian Evidence

References

- 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"
- MacKay, *Information theory, Inference & Learning Algorithms*
- Jaynes, *Probability Theory: the Logic of Science*
  - And other refs at http://bayes.wustl.edu
- Hobson et al, *Bayesian Methods in Cosmology*
- Sivia, *Data Analysis: A Bayesian Tutorial*

ICIC

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.

ICIC

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 $p(d_i|n_i s I) = \delta(d_i - n_i - s)$
  - 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 unifom dist'n in $(-\Sigma_1, \Sigma_2) \to (-\infty, \infty)$

ICIC

→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 $P(\sigma|\mathrm{I}) \propto 1/\sigma$
  - marginalized $P(s|\mathrm{I}) \propto [s - 2s \langle d \rangle + \langle d^2 \rangle]^{-1/2}$
  - (very broad distribution!)

ICIC

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

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

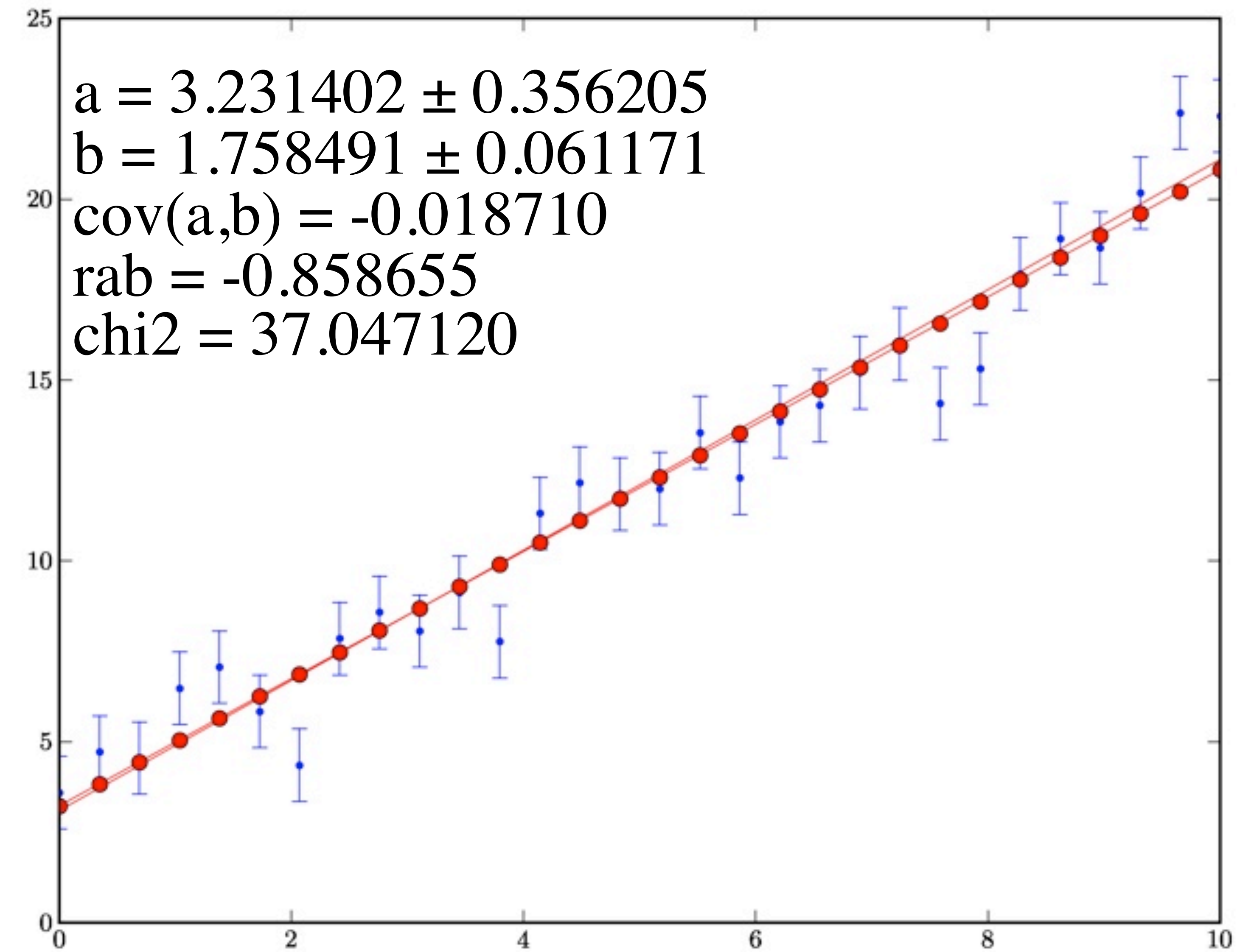
  - Multivariate gaussian in *d*
  - Linear in (a,b): also has form of a multivariate gaussian in (a,b)
    - but not a distribution in (a,b) until you apply Bayes' theorem and add a prior
  - Maximized at the value of the "least squares" est. for (a,b), with the same numerical values for the errors (&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

marginals?

ICIC

Inference from a Gaussian: Straight-line fitting

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



ICIC

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:

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

ICIC

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 $e^{-\chi^2_{\max}}$
- 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

ICIC

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)

Chi-squared

- The exponential factor of a Gaussian is always of the form $\exp(-\chi^2/2)$
- Likelihood: $\chi^2 = \sum(\mathrm{data}_i - \mathrm{model}_i)^2/\sigma_i^2$
- For fixed model, χ^2 has χ^2 distribution for $\nu = N_{\mathrm{data}} - N_{\mathrm{parameters}}$ "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)

ICIC

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

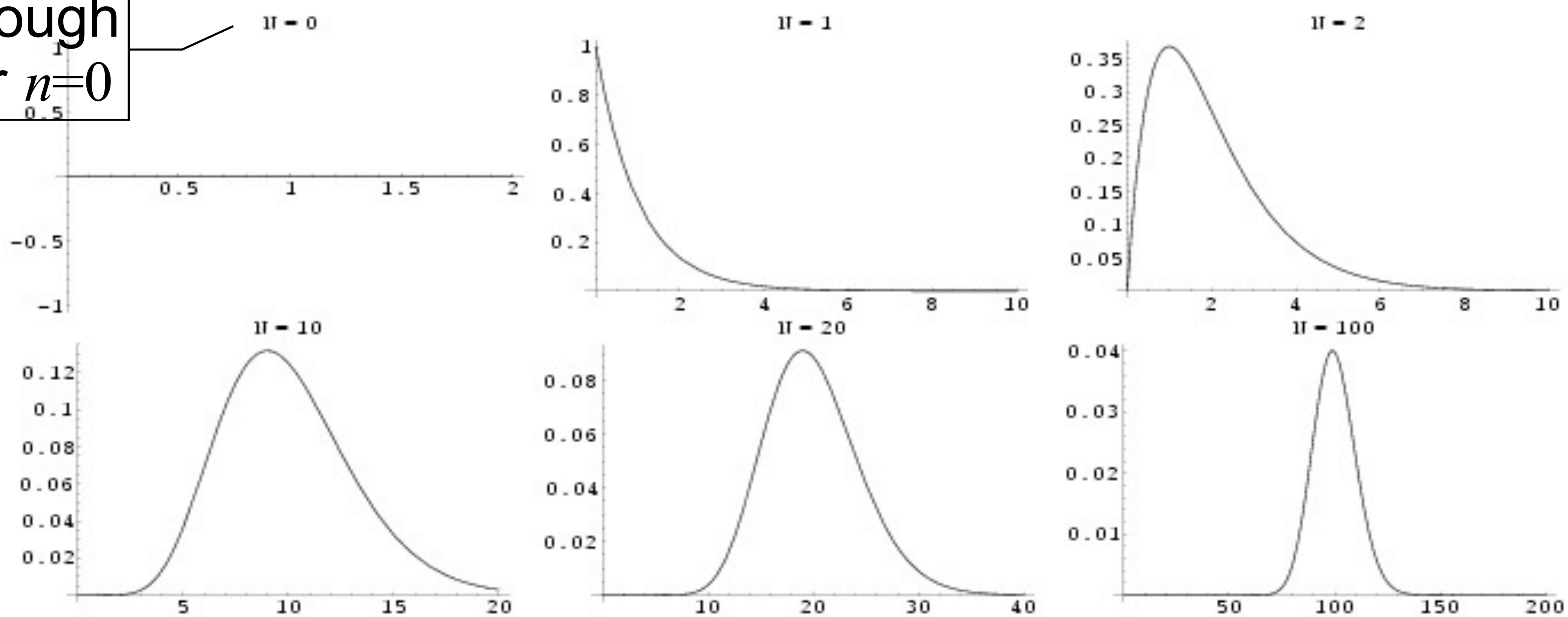
$$P(r|nI) = \frac{e^{-r} r^{n-1}}{(n-1)!}$$

- nb. (n-1) comes from $p(r|\mathrm{I})dr \propto 1/r$

ICIC

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

ICIC

Poisson rates

- Complications [see Loredo articles and optional problems]
  - 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): $s=s(t)$

ICIC

Credible Intervals

- The posterior contains the full inference from the data and our priors
- Sometimes, this can be a bit unwieldy.
- Traditionally, we compress this down into "credible intervals" (cf. frequentist "confidence intervals")
- A 100α % credible interval (a,b) is defined s.t.

$$P(x \in [x_-, x_+] | d, I) = \int_{x_-}^{x_+} P(x|d, I)\, dx = \alpha$$

  - We typically pick traditional values of α such as 68%, 95%, 99% (1, 2, 3σ)
  - if the mean is $\bar{x} = \int x P(x|dI)\, dx$
    this is often reported as $x = \bar{x} \pm (x_+ - x_-)$

ICIC

Confidence Intervals

- A 100α % confidence interval (a,b) is defined s.t. a fraction α of all realizations contain the correct value.
- Doesn't depend on the prior. But depends on the distribution of possible experimental results (i.e., the likelihood, considered as a function of the data, not the theoretical parameters) — results that didn't arise!
  - We typically pick traditional values of α such as 68%, 95%, 99% (1, 2, 3σ)
  - if the mean is $\bar{x} = \int x P(x|dI)\, dx$
    this is often reported as $x = \bar{x} \pm (x_+ - x_-)$
  - because this looks the same as a credible interval (and for problems like the Gaussian is numerically identical), there is occasionally ***confusion***...

ICIC

Confidence Intervals in Practice

- Neyman-Pearson approach
- Especially complicated when the possible parameter region has boundaries
- Feldman & Cousins, "Unified approach to the classical statistical analysis of small signals", *PRD*57, 7, 1998
- For data d and CL f, find $[x_-, x_+]$ s.t. $P(d \in [x_-, x_+] \mid \mu) = f$
- See also, Daniel's discussion of p-values...

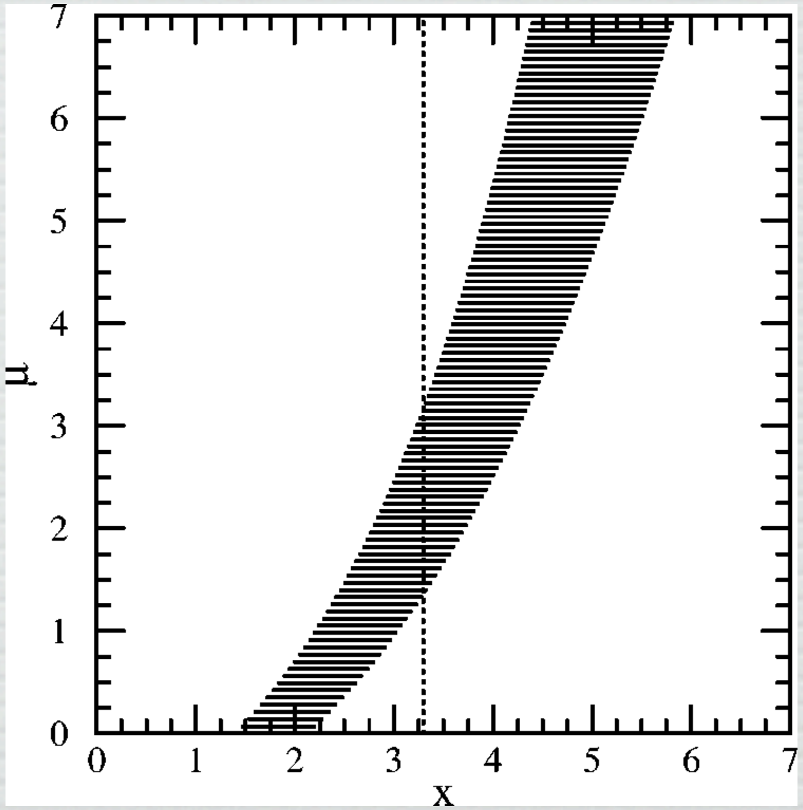


FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2] | \mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0. The confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.

Nuisance parameters

- We can sometimes separate our parameter space into those parameters that we "care about" and those we don't.
  - E.G.,
    - detector characteristics
    - phenomenological parameters for non-physical models
- We call these "nuisance parameters" and very often marginalize over them.
- *Beware*: if the posterior for the nuisance parameter is complicated, marginalization may be dangerous

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!)

ICIC

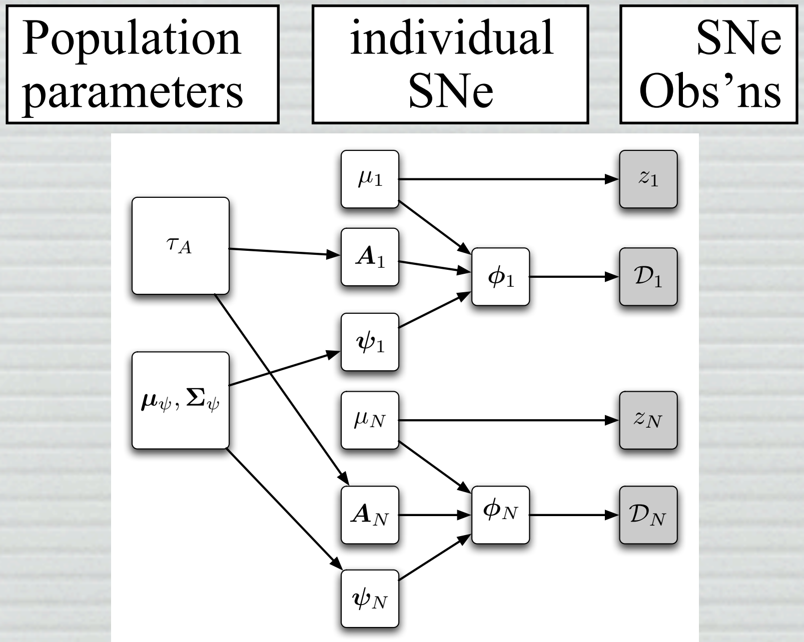
Hierarchical Models

- "Data reduction" vs "Data Analysis" (vs "Science"?)
- Describe inference from data as a series of levels:
  - parameters describing:
    - the instrument
      - e.g., gain, noise properties
    - individual observations
      - e.g., supernova brightness at a particular epoch; galaxy shape for weak lensing
    - the whole survey
      - e.g., mean (unstretched) supernova light curves; luminosity functions
    - the "scientific content" of the data
      - e.g., Hubble diagram; lensing power spectrum
    - the cosmological or astrophysical goals of the survey
      - e.g., $\Omega_{m,}$ etc
  - Some parameters need external priors (e.g., instrumental)
  - Some parameters get priors from the next level in the hierarchy
    - e.g., the prior for the Hubble diagram depends on the prior for the cosmological parameters

ICIC

Hierarchical Models

- Linear Models with errors in both dimensions
  - e.g., Kelly, B. C. (2007). "Some Aspects of Measurement Error in Linear Regression of Astronomical Data", *ApJ* 665:1489, 2007, arXiv:0705.2774v1
  - Unlike 1-d errors, need full model for generating data
    - $x = \xi + n_x$; $y = \eta + n_y$; $\eta = \eta(\xi; \theta)$ (e.g., $\eta = \alpha\xi + \beta$)
    - actual independent variable $\xi \sim p(\xi \mid \psi, I)$
    - actual dependent variable ("signal") $\eta \sim p(\eta \mid \xi, \theta, I)$
    - observed data $x, y \sim p(x, y \mid \eta, \xi, I)$
    - no analytic solution even for simple models! (see Daniel's discussion tomorrow)
- Models as Directed Acyclic Graphs
  - e.g., Mandel et al, "Type IA Supernova Light Curve Inference", *ApJ* 704:629, 2009, arXiv:0908.0536



Sufficient Statistics

- Sometimes, the likelihood only depends on a [simple] function of the data, a "statistic", S(D)
  - P(D | theory) dD = P(S(D)| theory) dS
    - trivial if you can invert to get D(S), but can be true in other cases
  - e.g., when estimating the mean of *iid* Gaussian data, the likelihood only depends on $\sum_i d_i/n$ and *n*.
    - (independent of the prior)
    - i.e., the sufficient statistic is what we're interested in
  - This is especially nice in the context of *hierarchical models* as we can consider each step as *data compression*
    - Will see this in more detail tomorrow with the CMB
  - Sometimes this is only approximately true
    - e.g., an *estimate of* the power spectrum $\hat{C}_\ell$ (even with errors) contains most but not all information about the underlying field
      - not to be confused with the full likelihood P(data| C_ℓ)

ICIC

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 $P(I_i \mid \text{data})$

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:

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

ICIC

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:

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

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

so

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

— just the normalization!

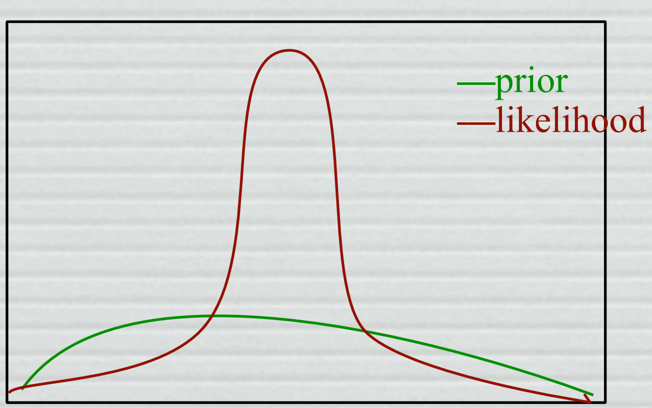
Model likelihood
(sometimes called "evidence")

ICIC

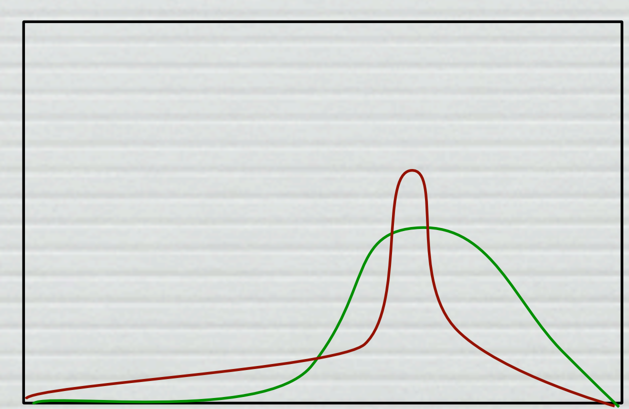
Model Comparison

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

- model probability ∝ average likelihood, weighted by prior
- automatic penalty for more complicated models (≡ more parameter 'volume')
- recall for the linear model, normalization $\propto e^{-\chi^2_{\max}}$ gives factor ~ $|N|^{1/2}$ ∝ volume of error ellipsoid



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



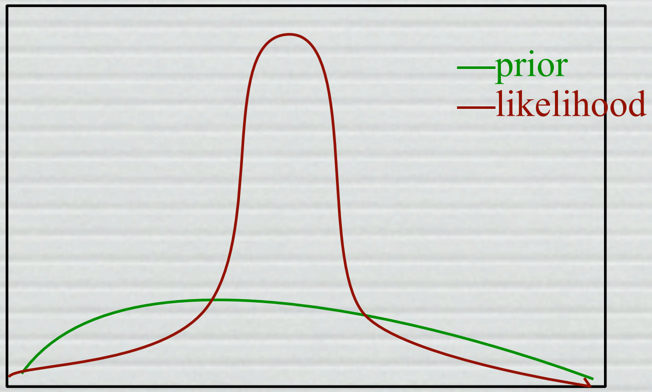
Ockham's Razor

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

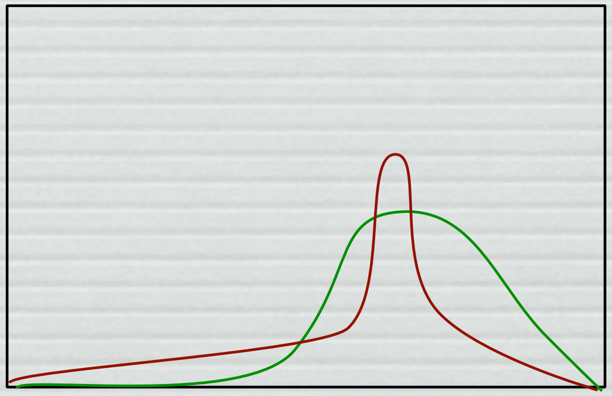
$$\simeq\; P(i|I)P_{\max}(D|\theta_i I_0 I_i)\,\frac{\text{posterior volume}}{\text{prior volume}}$$

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

Favors simpler model
"Ockham Factor"



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



ICIC

Ockham's Razor

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

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

- Linear, Gaussian models:
  - $P_{\max}(D|\theta_i I) = \frac{1}{|2\pi M|^{1/2}} e^{-\chi^2_{\min}/2}$
  - volume $\propto |M|^{1/2} = \sigma_1\sigma_2\sigma_3 \ldots \sigma_N$ for correlation matrix M
  - must have *proper* prior distributions (finite $|M|$) for this to make sense

ICIC

Model comparison and parameter priors $P(\theta_i|I)$

- Now, all *priors must be normalized*
- Model likelihoods must converge:

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

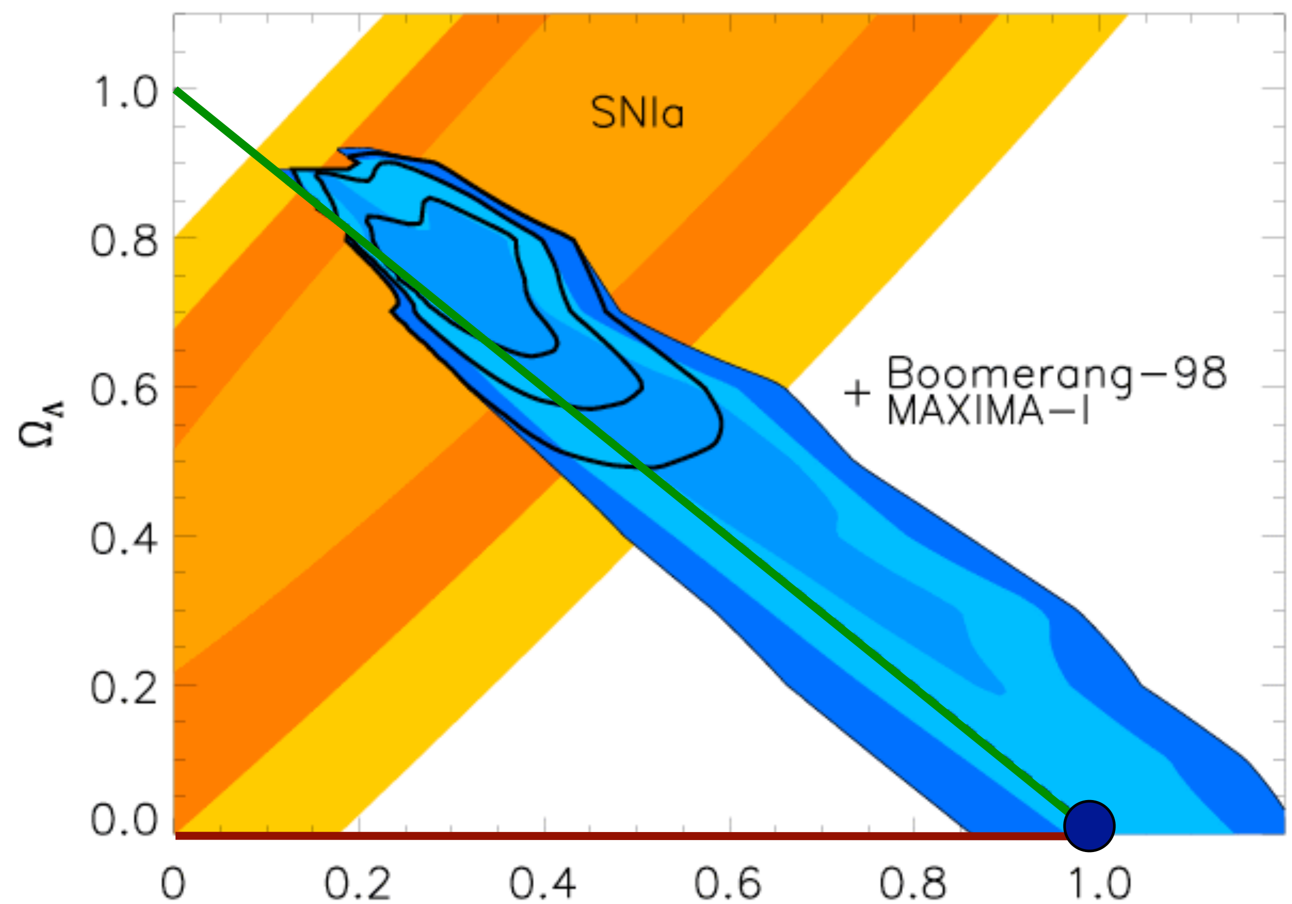
- e.g., linear models
- This is a very serious restriction in some cases.
  - Note that the posterior for a parameter may — and usually does — exist in the limit of an infinitely-wide prior, but in general the evidence does not:

$$\begin{aligned} P(i|DI) &\simeq P(i|I) P_{\max}(D|\theta_i I_0 I_i) \frac{\text{posterior volume}}{\text{prior volume}} \\ &\to 0 \end{aligned}$$

ICIC

Application: is the Universe flat?

- *nested models:*
  - old std CDM: $\Omega_\Lambda=0$, $\Omega_m=1$
  - flat: $\Omega_\Lambda+\Omega_m=1$
  - $\Omega_\Lambda=0$, $0\le\Omega_m\le1$
  - $0\le\Omega_m\le1$, $0\le\Omega_\Lambda\le1$



- 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!)

ICIC

Conclusions

- Gaussian linear models are equivalent to "generalized least squares"
- Hierarchical Models can describe the full solution for a general scientific problem from data gathering to science exploitation
  - there is very often a lot of data compression along the way, in the form of sufficient (or nearly sufficient) statsitics
- The model likelihood (aka Bayesian Evidence) is a tool for comparing well-specified models (but there is no real "alternative-free" test in the Bayesian formalism).

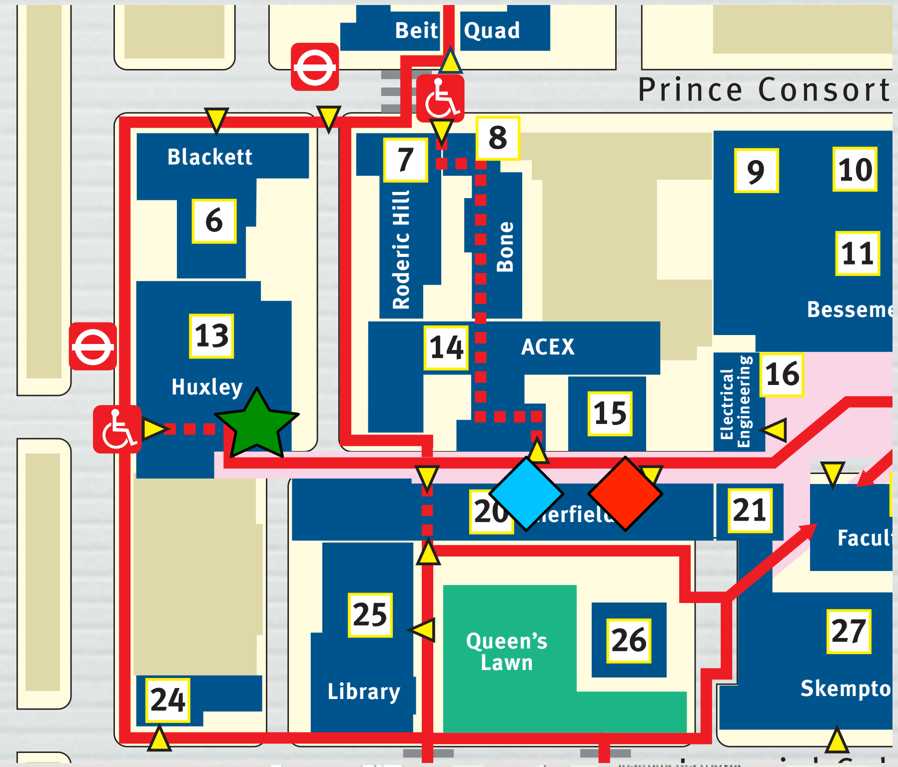
ICIC

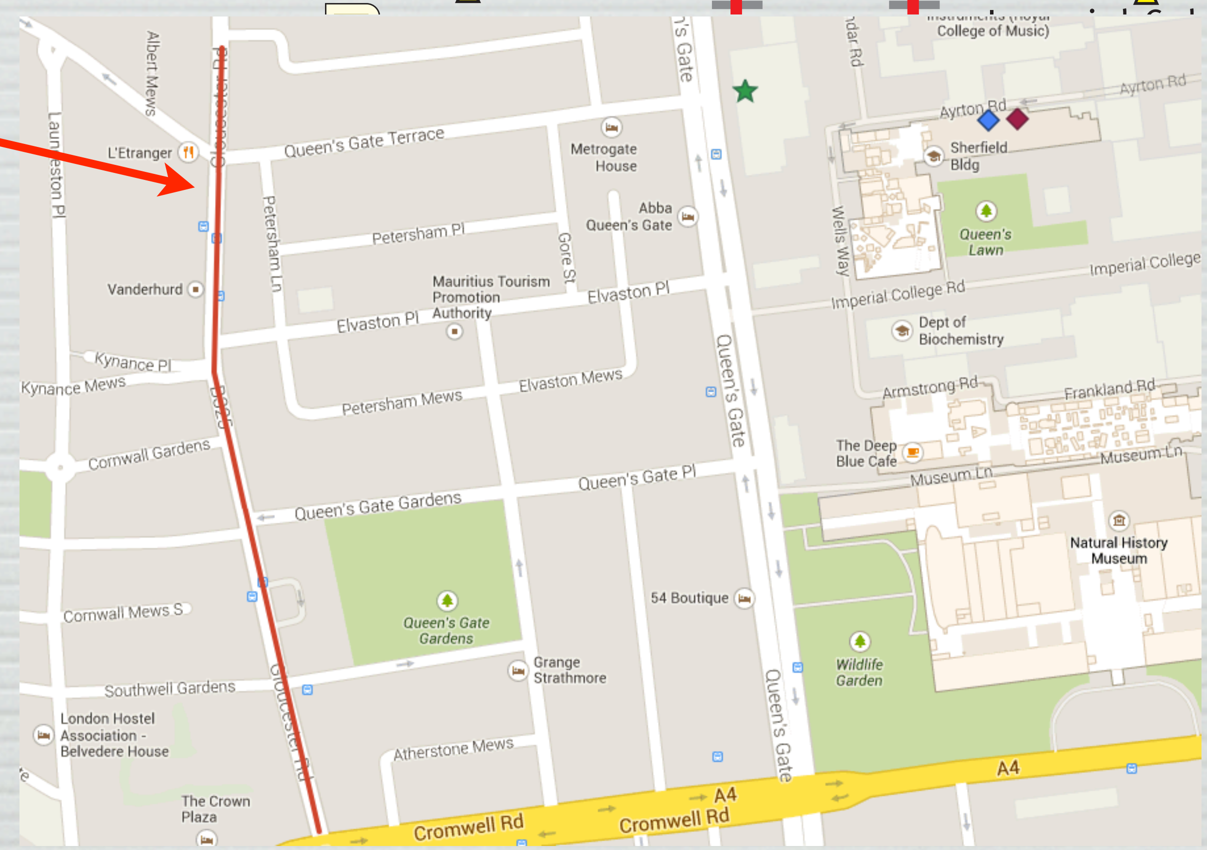
Lunchtime logistics

- On campus: SCR ◆ & JCR ◆ — go out main walkway from here ★ (Huxley 311). Other cafeterias are available.

- Off campus: Gloucester Road

- After lunch: please sit in alternate rows for the problem session (so we can reach you, not to avoid working together!)





ICIC