Statistical Data Analysis: Lecture 14

1	Probability, Bayes' theorem
2	Random variables and probability densities
3	Expectation values, error propagation
4	Catalogue of pdfs
5	The Monte Carlo method
6	Statistical tests: general concepts
7	Test statistics, multivariate methods
8	Goodness-of-fit tests
9	Parameter estimation, maximum likelihood
10	More maximum likelihood
11	Method of least squares
12	Interval estimation, setting limits
13	Nuisance parameters, systematic uncertainties
→ 14	Examples of Bayesian approach

A typical fitting problem

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: $V_{ij}^{\text{stat}}, \; V_{ij}^{\text{sys}}$.

Predicted value: $\mu(x_i; \theta)$, expectation value $E[y_i] = \mu(x_i; \theta) + b_i$

control variable parameters bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta) \sim e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $$L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T V_{\text{stat}}^{-1}(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$$

$$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\vec{b}^T V_{\text{sys}}^{-1}\vec{b}\right]$$

$$\pi_\theta(\theta) \sim \text{const.}$$

Joint probability for all parameters

and use Bayes' theorem: $$p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$$

To get desired probability for θ, integrate (marginalize) over b:

$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in n 'equally valid' ways and
extract systematic error from 'spread' in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;

vary renormalization scale $(\mu/2 < Q < 2\mu\ ?)$

Can we incorporate the 'error on the error'?

(cf. G. D'Agostini 1999; Dose & von der Linden 1999)

Motivating a non-Gaussian prior $\pi_b(b)$

Suppose now the experiment is characterized by

$$y_i, \quad \sigma_i^{\mathrm{stat}}, \quad \sigma_i^{\mathrm{sys}}, \quad s_i, \quad i = 1, \ldots, n \ ,$$

where s_i is an (unreported) factor by which the systematic error is over/under-estimated.

Assume correct error for a Gaussian $\pi_b(b)$ would be $s_i\sigma_i^{\mathrm{sys}}$, so

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\mathrm{sys}}} \exp\left[-\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\mathrm{sys}})^2} \right] \pi_s(s_i) \, ds_i$$

Width of $\sigma_s(s_i)$ reflects 'error on the error'.

Error-on-error function $\pi_s(s)$

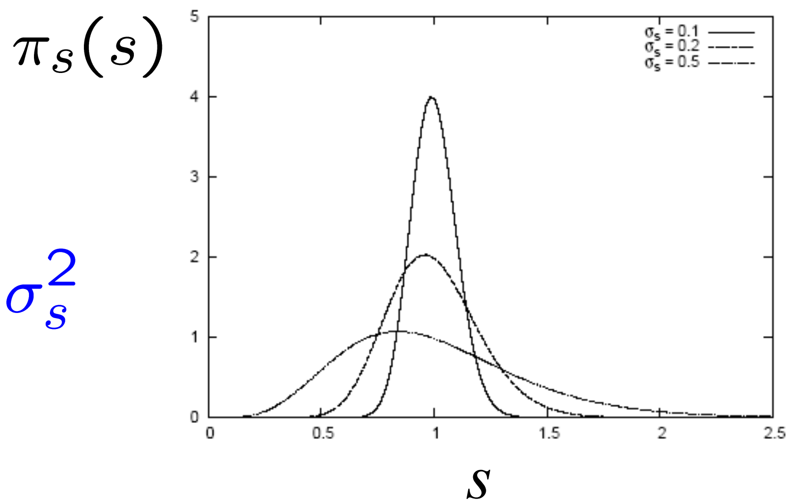
A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

$$\pi_s(s) = \frac{a(as)^{b-1}e^{-as}}{\Gamma(b)}$$

mean = b/a

variance = b/a^2

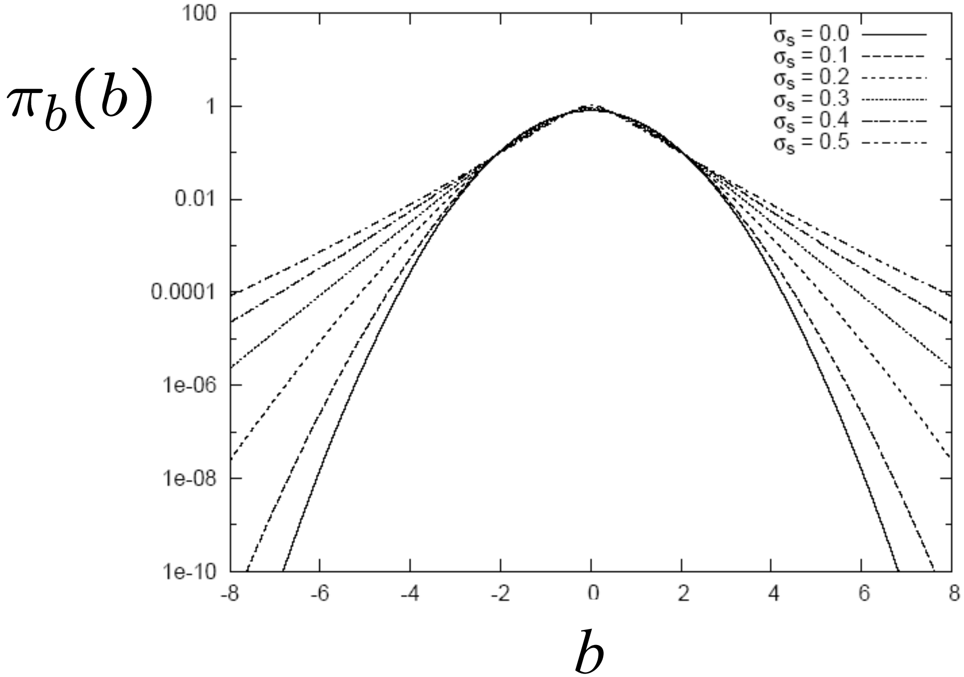
Want e.g. expectation value of 1 and adjustable standard Deviation σ_s , i.e., $a = b = 1/\sigma_s^2$



In fact if we took $\pi_s(s) \sim$ *inverse Gamma*, we could integrate $\pi_b(b)$ in closed form (cf. D'Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.

Prior for bias $\pi_b(b)$ now has longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\text{sys}})^2} \right] \pi_s(s_i)\, ds_i$$



Gaussian ($\sigma_s = 0$) $\quad P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

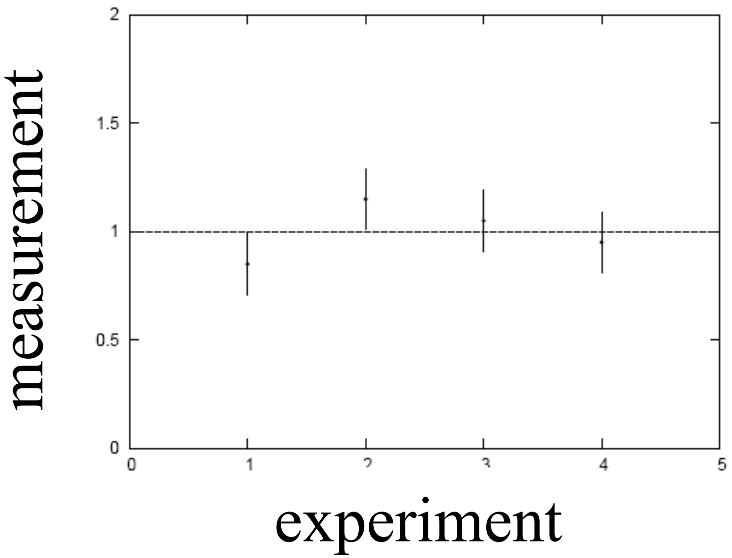
$\sigma_s = 0.5$ $\quad P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

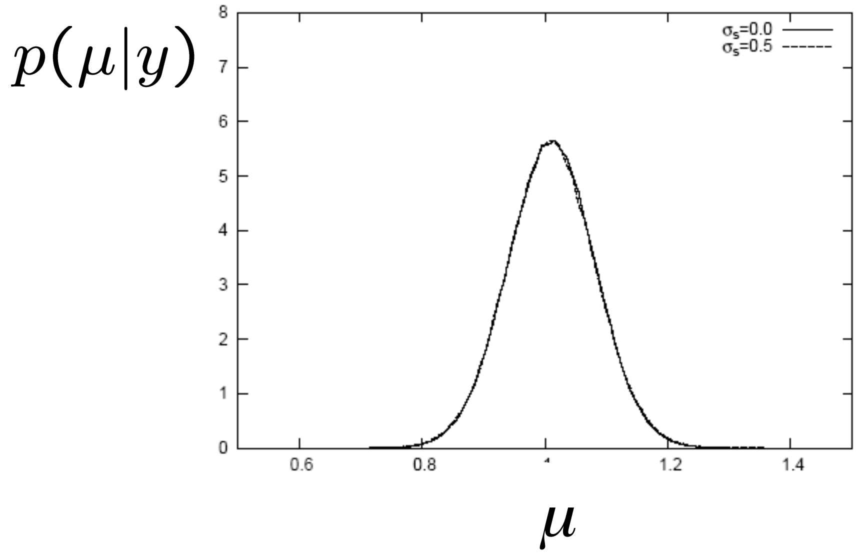
Suppose fit effectively averages four measurements.

Take $\sigma_{sys} = \sigma_{stat} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



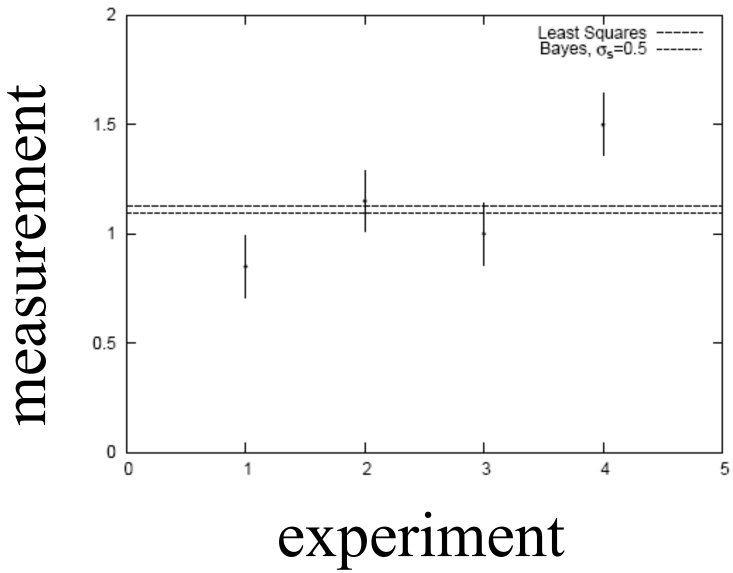
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$$\sigma_S = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$$

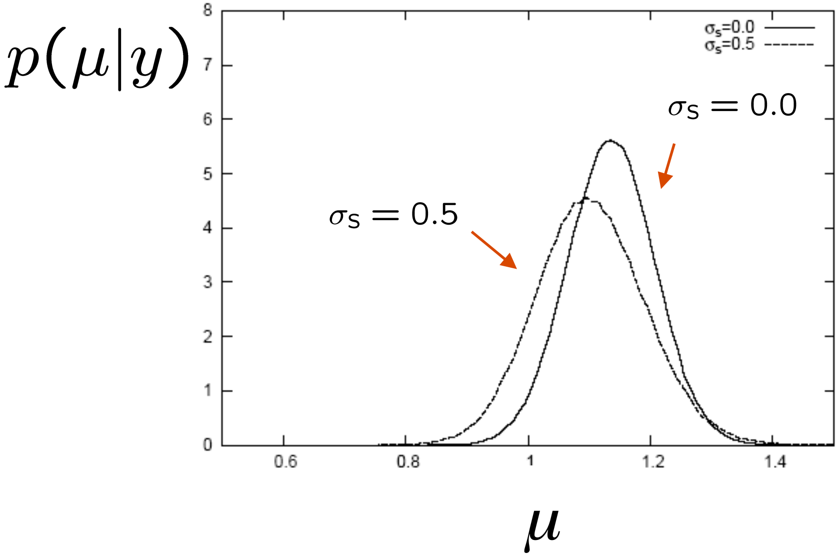
$$\sigma_S = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_S = 0.0 : \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_S = 0.5 : \quad \hat{\mu} = 1.093 \pm 0.089$$

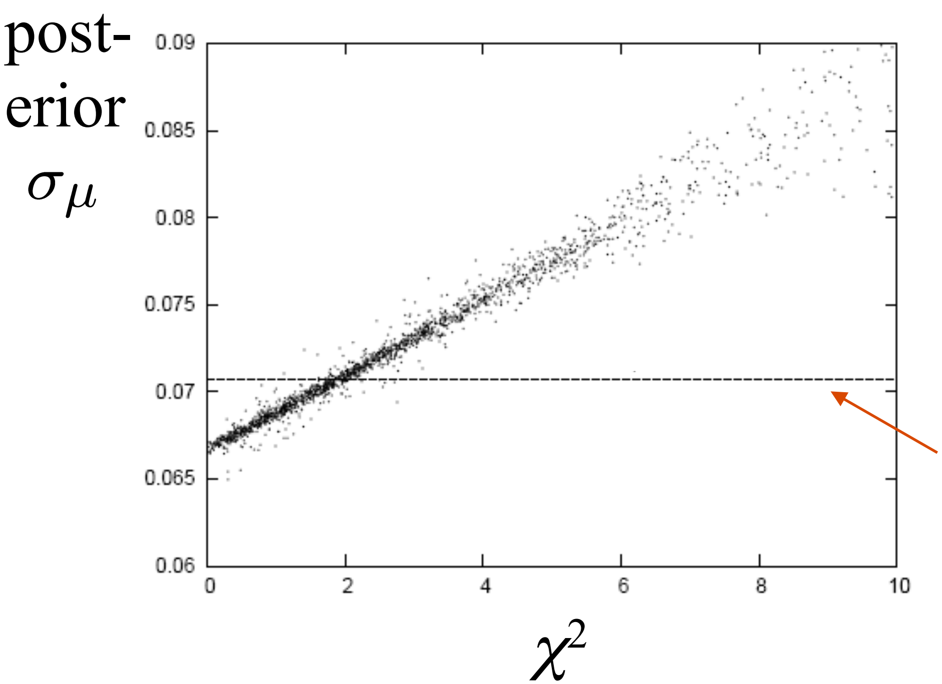
→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).



2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares

Is this workable in practice?

Should to generalize to include correlations

Prior on correlation coefficients ~ $\pi(\rho)$
(Myth: $\rho = 1$ is "conservative")

Can separate out different systematic for same measurement

Some will have small σ_s, others larger.

Remember the "if-then" nature of a Bayesian result:

We can (should) vary priors and see what effect this has on the conclusions.

Bayesian model selection ('discovery')

The probability of hypothesis H_0 relative to its complementary alternative H_1 is often given by the posterior odds:

no Higgs

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

Higgs

Bayes factor B_{01}

prior odds

The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_0 over H_1.

Interchangeably use $B_{10} = 1/B_{01}$

Assessing Bayes factors

One can use the Bayes factor much like a *p*-value (or Z value).

There is an "established" scale, analogous to HEP's 5σ rule:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Rewriting the Bayes factor

Suppose we have models H_i, $i = 0, 1, \ldots,$

each with a likelihood $p(x|H_i, \vec{\theta}_i)$

and a prior pdf for its internal parameters $\pi_i(\vec{\theta}_i)$

so that the full prior is $\pi(H_i, \vec{\theta}_i) = p_i \pi_i(\vec{\theta}_i)$

where $p_i = P(H_i)$ is the overall prior probability for H_i.

The Bayes factor comparing H_i and H_j can be written

$$B_{ij} = \frac{P(H_i|\vec{x})}{P(H_i)} \Big/ \frac{P(H_j|\vec{x})}{P(H_j)}$$

Bayes factors independent of $P(H_i)$

For B_{ij} we need the posterior probabilities marginalized over all of the internal parameters of the models:

$$P(H_i|\vec{x}) = \int P(H_i, \vec{\theta}_i|\vec{x})\, d\vec{\theta}_i$$

Use Bayes theorem

$$= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{P(x)}$$

So therefore the Bayes factor is

Ratio of marginal likelihoods

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i)\pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j)\pi_j(\vec{\theta}_j)\, d\vec{\theta}_j}$$

The prior probabilities $p_i = P(H_i)$ cancel.

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\, d\vec{\theta} \longleftarrow \text{'marginal likelihood'}$$

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
...

See e.g. Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\theta)$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\theta)$ fall off faster than $L(x|\theta)\pi(\theta)$

Note harmonic mean estimator is special case $f(\theta) = \pi(\theta)$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

Bayes factor computation discussion

Also tried method of parallel tempering; see note on course web page and also

Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Harmonic mean OK for very rough estimate.

I had trouble with all of the methods based on posterior sampling.

Importance sampling worked best, but may not scale well to higher dimensions.

Lots of discussion of this problem in the literature, e.g.,

Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

Wrapping up lecture 14

Bayesian methods are becoming increasingly important, especially thanks to computational methods like MCMC.

Allows incorporation of prior information not necessarily related to available measurements.

Requires specification of prior.

Model selection using Bayes factors

Often a computational challenge

Interpretation (arguably) more intuitive than p-value