Bayesian model comparison with applications

Johannes Bergström

Department of Theoretical Physics, KTH Royal Institute of Technology

July 16, 2013

Outline

Physics – how to do it?

- Experiment and observe – compare with predictions of models
- No perfect experiments – always noise/uncertainties, limited resources/sensitivity/range
- Logically deducing the true model doesn't work
- All we can say is if a model is plausible description of data or not
- But how to determine this?

Important information

If you really don't like statistics ..... you can stop listening now

Principle of Bayesian inference

Bayesian inference in a nutshell

- Assess hypotheses/models by calculating their plausibilities, conditioned on some known and/or presumed information.

Cox's Theorem (1946)

- The unique calculus of plausibility is probability theory (using some requirements incl. comparability, consistency)
- Unique extension of deductive logic incorporating uncertainty
- truth $\rightarrow$ 1, falsehood $\rightarrow$ 0

Probability interpretations: what is distributed in $\Pr(X)$?

Bayesian probability

- Describes uncertainty
- Defined as plausibility
- Probability distributed over different propositions X
- X is not distributed nor random

Frequentist probability

- Describes "randomness"
- Defined as long-run relative frequency of event
- X is distributed – a random variable

1. Foundations
2. Bayesian inference
3. Examples and applications

Bayesian inference – updating probabilities

Updating probabilities

- Models $H_1 \ldots H_r$, data $\mathbf{D}$. Bayes' theorem:

$$\Pr(H_i|\mathbf{D}) = \frac{\Pr(\mathbf{D}|H_i)\Pr(H_i)}{\Pr(\mathbf{D})}$$

- $\Pr(H_i)$ – prior probability
- $\Pr(H_i|\mathbf{D})$ – posterior probability
- $\Pr(\mathbf{D}|H_i) = \mathcal{L}(H_i)$ – likelihood of H_i

$$\begin{aligned}\frac{\Pr(H_i|\mathbf{D})}{\Pr(H_j|\mathbf{D})} &= \frac{\mathcal{L}(H_i)}{\mathcal{L}(H_j)}\frac{\Pr(H_i)}{\Pr(H_j)} \\ \text{Posterior odds} &= \text{Bayes factor} \cdot \text{Prior odds}\end{aligned}$$

- Usually Prior odds $= 1$

Calculate either

- Bayes factor/posterior odds
- In addition assume precisely one of the $H_i's$ correct $\Rightarrow$ finite $\Pr(H_i|\mathbf{D})$

Model likelihood or evidence

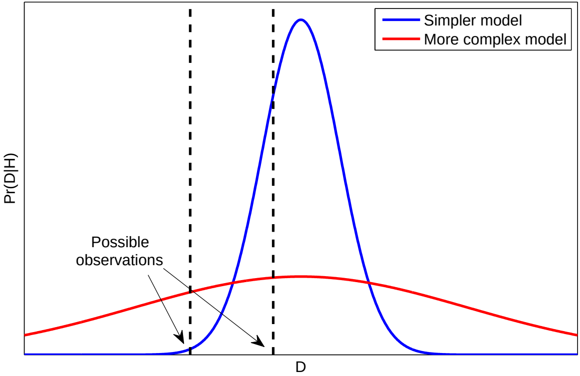
- Models usually have free parameters $\mathbf{\Theta}$
- Likelihood for model – evidence –

$$\begin{aligned}\mathcal{L}(H) = \Pr(\mathbf{D}|H) &= \int \Pr(\mathbf{D}|\mathbf{\Theta}, H) \Pr(\mathbf{\Theta}|H)\mathrm{d}^N\mathbf{\Theta} = \int \mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathrm{d}^N\mathbf{\Theta} \\ \text{Model likelihood} &= \text{Average likelihood of model parameters}\end{aligned}$$

- $\pi(\mathbf{\Theta})$ – Prior distribution – plausibility of parameters assuming model correct
- Evidence balances quality of fit vs. model complexity – can favour simpler model
- All probabilities conditioned on relevant background information (models, experimental setup, . . .)

Occam's razor

- Evidence = probability with which model predicted observed data
- Occam's razor – "simple" ≡ predictive
- Complex models compatible with large variety of data – predict less



Jeffreys scale

- Scale of interpretation easily calibrated: Jeffreys scale

$\lvert\log(\text{odds})\rvert$	odds	$\Pr(H_1 \lvert \mathbf{D})$	Interpretation
< 1.0	$\lesssim 3:1$	$\lesssim 0.75$	Inconclusive
1.0	$\simeq 3:1$	$\simeq 0.75$	Weak evidence
2.5	$\simeq 12:1$	$\simeq 0.92$	Moderate evidence
5.0	$\simeq 150:1$	$\simeq 0.993$	Strong evidence

Priors

- Must specify priors on all model parameters – not invariant under general reparametrizations
- Important part of Bayesian analysis – consider carefully
- Uniform prior in the variable you happen to be writing your equations in (signal rate, x-section) often bad choice
- Improper prior always bad choice
- Evaluate sensitivity to prior choice

Parameter inference

Parameter inference – posterior distribution

- Assuming model H correct, infer its parameters

$$\Pr(\mathbf{\Theta}|\mathbf{D}, H) = \frac{\Pr(\mathbf{D}|\mathbf{\Theta}, H)\Pr(\mathbf{\Theta}|H)}{\Pr(\mathbf{D}|H)} = \frac{\mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})}{\mathcal{L}(H)}$$

- Posterior of subsets of parameter by integrating over other parameters
- Posterior not enough to test/compare any model(s), claim discoveries – by definition

Comparing models using posterior

- Compare nested model with $\eta = \eta_0$ using

$$\frac{\mathcal{L}(\eta = \eta_0)}{\mathcal{L}(\eta \neq \eta_0)} = \frac{\Pr(\eta_0|\mathbf{D}, H)}{\pi(\eta_0|H)} = \frac{\text{Posterior at } \eta_0}{\text{Prior at } \eta_0} \quad \text{(Savage-Dickey density ratio)}$$

Frequentist model evaluation: P-values

P-values

- P-value $\equiv$ probability of obtaining equal or more extreme data than the observed assuming H_0
- Extreme $\equiv$ large value of test statistic (χ^2, profile likelihood, ...)
- Converted into "No. of σ's" using Gaussian CDF: $S = \phi^{-1}(1 - p)$

P-values are not   See also D'Agostini, 1112.3620

- Probability H_0 correct
- Probability data is "just a fluctuation"
- Probability of incorrectly rejecting H_0
- Type-1 error rate α (0.05, 0.01...)
- Interpretation needs uniform scale – not really possible

Model comparison in particle physics

In particle physics

- Use to compare ("test") different models
- Testing existence of "new physics"
- Discovery is primary – precise parameter values describing new physics often secondary

Possible applications

- $\theta_{13} = 0$ vs. $\theta_{13} > 0$
- CP-violation vs. CP-conservation
- Normal vs. inverted ordering
- Maximal vs. nonmaximal θ_{23}
- Evidence of effects of neutrino mass: $0\nu\beta\beta$, β-decay, cosmology.
- Theoretical models of lepton mass, flavour, DM, ...
- ...

1. Foundations
2. Bayesian inference
3. Examples and applications

Leptonic mixing angle θ_{13} – flashback to fall 2011

Question

Is $\theta_{13} = 0$ or not?

Profile likelihood ratio   Schwetz, Tórtola, Valle, 1108.1376

$$\frac{\mathcal{L}(\theta_{13}^{\max})}{\mathcal{L}(\theta_{13}=0)} \simeq 150 \quad (\Delta\chi^2 \simeq 10) \quad \Rightarrow \quad p \simeq 1.5 \cdot 10^{-3}$$

Model comparison   Bergström, 1205.4404

- Compare model $\theta_{13} > 0$ ($\in [0, \pi/2]$) with model $\theta_{13} = 0$
- Compact parameter space $\Rightarrow$ robust results
- Approx $\mathcal{L}(\theta_{13}) \propto \mathcal{L}_{\text{profile}}(\theta_{13}) \Rightarrow$

$$\frac{\mathcal{L}(\theta_{13} > 0)}{\mathcal{L}(\theta_{13} = 0)} \simeq 3$$

- Barely weak preference for $\theta_{13} > 0$

Assign 0.5 prior $\Rightarrow \Pr(\theta_{13} = 0|\mathbf{D}) \simeq 0.25$

Leptonic mixing angle θ_{23} – today

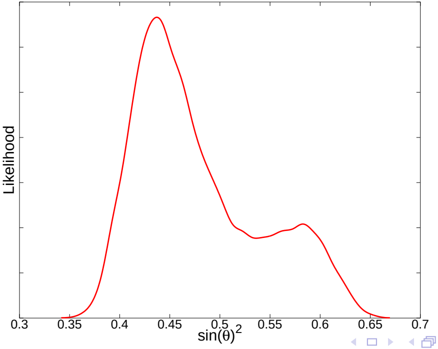
Question

θ_{23} is large, but is θ_{23} maximal ($\pi/4$) or not?

Profile likelihood (for NO)

νfit v1.1: www.nu-fit.org, 1209.3023 (Gonzalez-Garcia, Maltoni, Salvado, Schwetz)

$$\frac{\mathcal{L}(\theta_{23}^{\max})}{\mathcal{L}(\theta_{23} = \pi/4)} \simeq 2.5 \quad (\Delta\chi^2 \simeq 1.8) \quad \Rightarrow \quad p \simeq 0.18$$



Leptonic mixing angle θ_{23} – today

Model comparison

- Use $\mathcal{L}(s_{23}^2) \propto\!\!\!\!\sim \mathcal{L}_{\text{profile}}(s_{23}^2)$ and $\pi(s_{23}^2) = 1$
- Compare model likelihoods

$$\frac{\mathcal{L}(\theta_{23} \neq \pi/2)}{\mathcal{L}(\theta_{23} = \pi/4)} \simeq 0.3$$

- Maximal mixing preferred by data (weakly)
- Model with maximal θ_{23} (slightly) better than non-maximal model

$$\text{Assign 0.5 prior} \quad \Rightarrow \quad \Pr(\theta_{23} = \pi/4|\mathbf{D}) \simeq 0.75$$

- Octant comparison

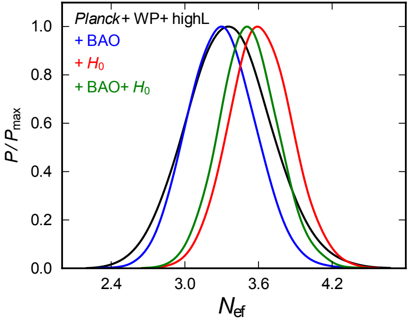
$$\frac{\mathcal{L}(\theta_{23} < \pi/4)}{\mathcal{L}(\theta_{23} > \pi/4)} \simeq 2$$

Future prospects

- Strong evidence for maximal mixing requires uncertainty on s_{23}^2 of roughly 0.002 (0.02 for moderate)

Neutrino parameters and cosmology

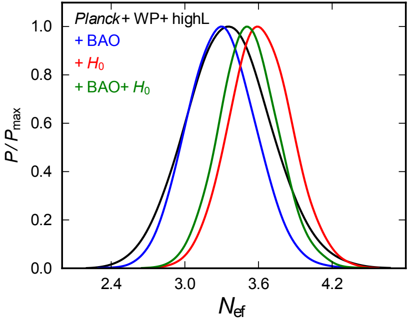
- Cosmological data sensitive to N_{eff} Planck collaboration, 1303.5076



- How much evidence is there against $N_{\text{eff}} = 3.046$?

Neutrino parameters and cosmology

- Cosmological data sensitive to $N_{\rm eff}$ Planck collaboration, 1303.5076



- How much evidence is there against $N_{\rm eff} = 3.046$?
- Answer: cannot say – information is missing
- Posterior obtained assuming $N_{\rm eff} \neq 3.046$
- Model comparison

$$\frac{\mathcal{L}(N_{\rm eff} = 3.046)}{\mathcal{L}(N_{\rm eff} \neq 3.046)} = \frac{\text{Posterior at 3.046}}{\text{Prior at 3.046}}$$

Results, $N_{\rm eff} < 10$

Verde, Feeney, Mortlock, Peiris, 1307.2904

Taking $N_{\rm eff} < 10 \Rightarrow$

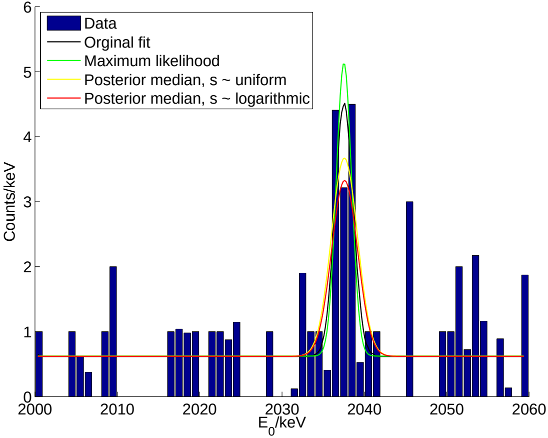
- With H_0 – no evidence of additional $N_{\rm eff}$
- Without H_0 – weak evidence against additional $N_{\rm eff}$
- No evidence of additional $N_{\rm eff}$ pre-Planck too

Signal discovery in spectra

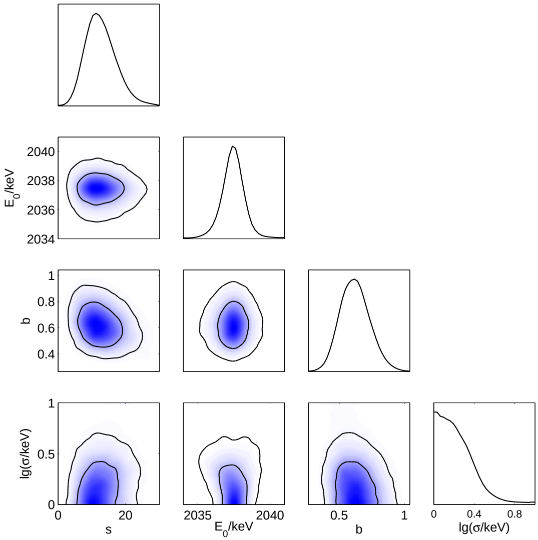
Bergström, 1212.4484; Caldwell, Kröniger, physics/0608249

Question

- Is there a signal?



Estimate signal strength



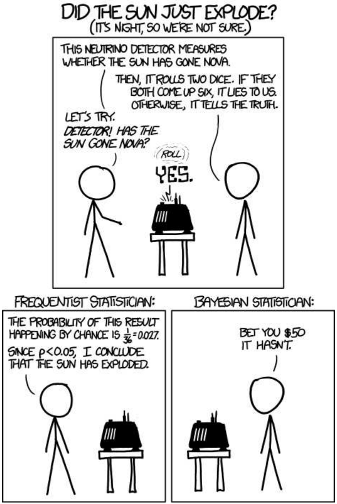
Signal discovery

- Compare evidences of $s + b$ model with b-only model
- No need for distributions of test statistic
- Do need prior on signal rate
- Automatic compensation for LEE $\propto$ signal/spectrum widths

Summary, conclusions

- Bayesian inference rocks!!!
- Consider your priors carefully
- Don't just estimate parameters of a fixed model – compare models too

Thanks for listening!



http://www.xkcd.com/1132/

Extra slides

Extra slides

Analysing Beyond the Standard Model models

BSM models

- Many BSM models have large – unconstrained – parameter spaces

Theorists' favourite method – random scans

- Generate many points in parameter space
- Accept points which pass "cuts" (e.g., at 2σ)
- Draw conclusions form distribution of points and/or the fraction of accepted points

Warning

- No statistical/probabilistic measure attached to density of points
- No statistical/probabilistic interpretation of results possible
- But sometimes rough approximation of Bayesian analysis (reinvented?)

U(1) flavour models – lepton sector

Work with L. Merlo and D. Meloni

The models

- Charged lepton masses (as quarks) are hierarchical
- Mixing seem less so – but is hierarchy or anarchy preferred?
- U(1) symmetry ⇒ obtain lepton masses and mixing "naturally" by suppressing charged lepton and neutrino mass matrix elements by ϵ^{n_i}

Parameters

- $\epsilon < 1$ – flavon VEV/cutoff scale
- n_i – 4 integer charges of lepton doublets/singlets
- 30 additional "order one" parameters and phases in Yukawa/mass matrix

Data

- m_e/m_μ, m_e/m_τ, leptonic mixing parameters, $\Delta m^2_{21}/\Delta m^2_{31}$

Analysing U(1) models

χ^2-analysis

- $\Delta\chi^2(\epsilon,\ \text{charges}) = 0$ – all charges and ϵ can fit data equally well
- Theorists' response: So what?!?
- Most of these values are unnatural – require large cancellations – hence implausible

Bayesian analysis

- Consistently incorporated in Bayesian analysis through priors on $\mathcal{O}(1)$ parameters
- Fix charges $\Rightarrow$ nice Gaussian posteriors of ϵ
- Compare charge assignments using model comparison
- Fit charges as free parameters simultaneously
- Compare "Anarchy" in neutrino sector (doublet charges $= 0$) with "Hierarchy" probabilistically $\Rightarrow$ some preference for Hierarchy

Neutrinoless double beta decay Bergström, 1212.4484

Neutrinoless double beta decay

- Majorana neutrinos can mediate $0\nu\beta\beta$
- Signal strength $s \propto |\text{Nuclear matrix element}|^2 |m_{ee}|^2$
- $m_{ee} = \sum_i m_i U_{ei}^2$

Fitting data

- Requires prior on m_{ee} – not uniform
- NME calculations uncertain – unconstrained by data
- NME uncertainties cannot be included in likelihood – but in prior

Compatibility of parameter constraints of ≥ 2 data sets

- A model comparison question – compare “data compatible” with “data incompatible”

Prior on m_{ee} – posterior using oscillation + β-decay