PhyStat-ν 2019

A Statistical Summary

David A. van Dyk

Statistics Section, Department of Mathematics, Imperial College London

CERN, January 2019

Imperial College London

Disclaimer

Statisticians have "discussions" (of talks) rather than "summaries" (of conferences).

This is an incomplete discussion of some of the topics that I found interesting.

Other interesting topics.... but my time is limited.

I'm sure I'm missing important contributions!

Correct me if I mischaracterize your work!!

Imperial College London

....and forgive me if I stand on my soap box a bit....

Statistical Framework for Discovery

Hypothesis Testing

H_0: The null hypothesis (e.g., no CP-violoation, $\delta = 0$)

H_1: The alternative hypothesis (e.g., CP-violation)

- Without further evidence, H_0 is presumed true.
- "Deciding" on H_1 means scientific discovery: new physics.

Other Model Hypothesis Tests

Selection: May be no presumed model. (e.g., normal/inverted hierarchy)

Multiple: More than two models. (STEFANO G)

Checking: Is data consistent with model? H_1 not required.

... a.k.a., "goodness of fit"

Different statistical approaches, e.g., Frequentist, Bayesian.

Imperial College London

Model Fitting vs Model Selection/Checking

Model Fitting

- Specify one model, fit parameters, estimate uncertainty.
- Frequency and Bayesian tend to agree. (JIM B, BOB C)
- Choice of prior distribution is often not critical.
- Some *"model selection"* tasks can be accomplished via model fitting and computing
  - confidence or credible intervals.
  - probability of region in parameter space.
  - formal model selection is often easier in these cases

    The one-sided tests that Jim B discussed.
- Computation is often much easier too!

  (e.g., Bayes Factors / Monte Carlo p-values, aka toys)

Perfect Storm: Model fitting is both much more challenging and scientifically higher profile.

Imperial College London

Frequentist or Bayesian?

Do you have to choose??

- Bayes prescribes method — Frequency evaluates method.
- Frequency evaluation of Bayesian methods.
  I like intervals to include true values... at least once in a while.
  Rob, Glen over lunch: marginalize/profile nuisance parameters
- Model fitting: often little difference in fits and errors.
  ... but not always, see below.
- Why not control detection error
  and assess probability of new physics?
- Why throw away half of your tool box?

I'm impressed with the openness of neutrino researchers to both Bayesian and Frequency based methods.

- Several Bayesian procedures *(Alex H, Stefano G, Matteo A, Glen C)*
- My experience with cosmologists & particle physicists.

Imperial College London

Outline

Imperial College London

Outline

Imperial College London

A modular approach to statistics

Wouter on "Model Building for Systematic Uncertainties"

- Separate model building from analysis *... in his case via RooFit*
- Modular approach to set generative model / likelihood.

Tom: what would be needed to deploy RooFit in neutrino experiments?

This can be taken further: *... two examples: EM and FC*

1. *Modelling:* Identify scientific goal, derive likelihood, priors
2. *Deriving Statistical Methods*
   - estimates, intervals, model checking and selection
   - minimize χ^2, likelihood-based, Bayes, others?
3. *Evaluating Statistical Methods* *(What are the operating characteristics?)*
   - frequency based: coverage, error rates, mean square error
   - Bayesian: complete summary of information?
4. *Computation:* "what to compute" vs."how to compute it."

And Always Report Your Likelihood!

Imperial College London

That's What We Want to See

Likelihood function from paper

all the analysis categories. The complete likelihood is given in Eq. (1):

$$\mathcal{L}(\text{data}|\Delta\vec{\sigma}^{\text{fid}}, \vec{n}_{\text{bkg}}, \vec{\theta}_{\text{S}}, \vec{\theta}_{\text{B}}) =$$

$$\prod_{i=1}^{n_{\text{cat}}} \prod_{j=1}^{n_b} \prod_{l=1}^{n_{m_{\gamma\gamma}}} \left(\frac{\sum_{k=1}^{n_b} \Delta\sigma_k^{\text{fid}} K_k^{ij}(\vec{\theta}_{\text{S}}) S_k^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{S}}) L + n_{\text{OOA}}^{ij} S_{\text{OOA}}^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{S}}) + n_{\text{bkg}}^{ij} B^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{B}})}{n_{\text{sig}}^{ij} + n_{\text{bkg}}^{ij}} \right)^{n_{\text{ev}}^{lij}}$$

$$\text{Pois}(n_{\text{ev}}^{ij}|n_{\text{sig}}^{ij} + n_{\text{bkg}}^{ij}) \text{Pdf}(\vec{\theta}_{\text{S}}) \text{Pdf}(\vec{\theta}_{\text{B}}),$$

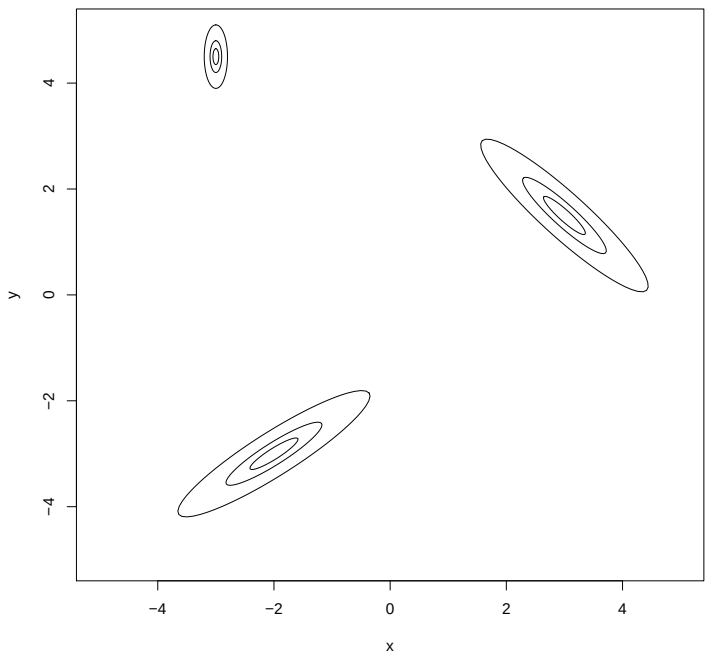
(STEFAN SCHMITT)

Imperial College London

Combining Analyses / Global Fits

Nicholas on Atlas/CMS Combing Procedures

- Combine likelihood functions, *... same logic for unfolding!*
  - not estimates, χ^2 values, or p-values! (LOUIS W/ P-VALUES)
- Careful diagnostics to be sure the individual fits are consistent before combining.



- Alvaro on Global Analysis of Reactor Anti-Neutrino Data: "Goodness of fit is meaningless because the data are totally incompatible."

What would it mean to average the MLEs of these likelihoods or combine there χ^2 values?

Imperial College London

Systematics Errors

Lots of discussion of Systematics!

(Alex H, Chao Z, Wouter V, Nicholas W, Constantinos A, Thierry L)

Bob: *"When you discover a new dimension/particle, can you convince the world you understand the systematics well enough to back up your claim?"*

Constantinos on Systematic Errors in Neutrino Experiments

Sometimes systematics can be well quantified... but there are

- "Highly non-trivial uncertainties in systematics"
- Modeled with a "Mix of theory, empirical models, extrapolation, and guesses"
- "Many key uncertainties are not reweighted – tend to be ignored"

Uncertainty in the Systematics!

Imperial College London

Systematics Errors

Glen on Uncertainty in Error Parameters (Systematics)

- Reported variance is unbiased estimate of true variance
- Use gamma distribution to quantify relative uncertainty.
- Profile out the true (unknown) variance in closed form.
- Use Bartlett Correction to improve asymptotics.

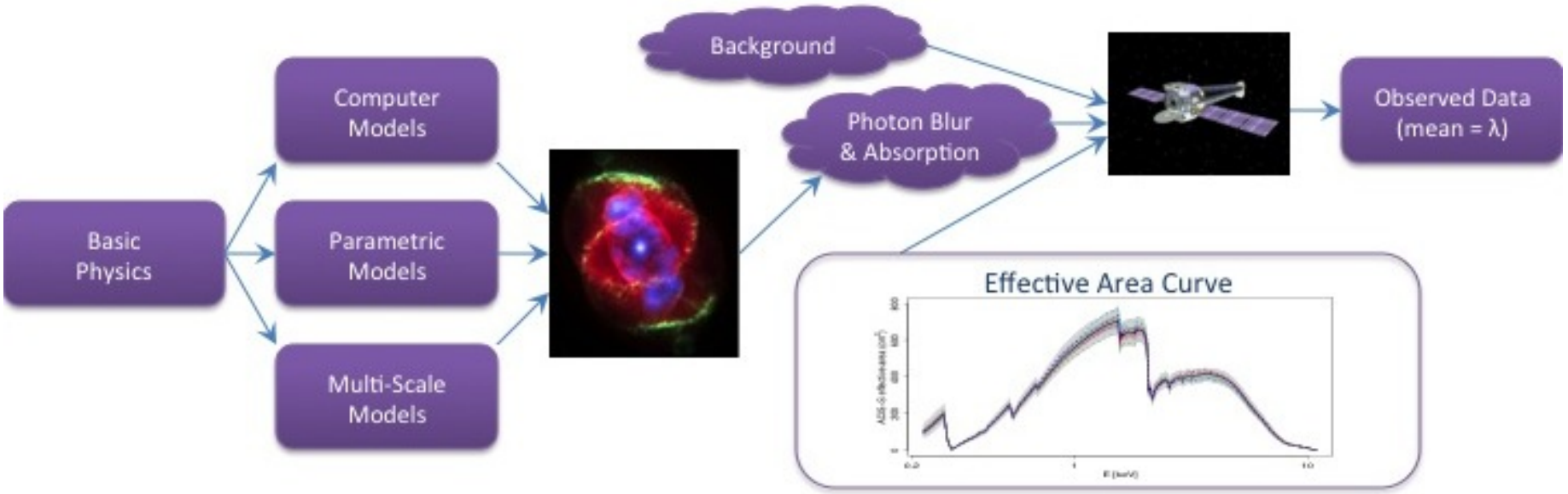
.... see also Anthony D's discussion of Higher-Order Likelihood Inference.

Meanwhile Alain warns: *"It is not recommended (i.e. should be forbidden really) to fit some data with a convenient but arbitrary or unsure or model-dependent function (i.e. fit looks good) and act as if the error matrix of the fit represents the uncertainty on the fit data. It does not, – and this can go very wrong!"*

... even for systematics?

Imperial College London

Calibration of X-ray Detectors



- Embed physics models into multi-level statistical models.
- Must account for complexities of data generation.
- Effective area: instrument sensitivity as function of energy.
- PCA to derive low-dimensional prior on eff. area, *A*.
- Similar for smearing – embed unfolding in unified analysis.

Imperial College London

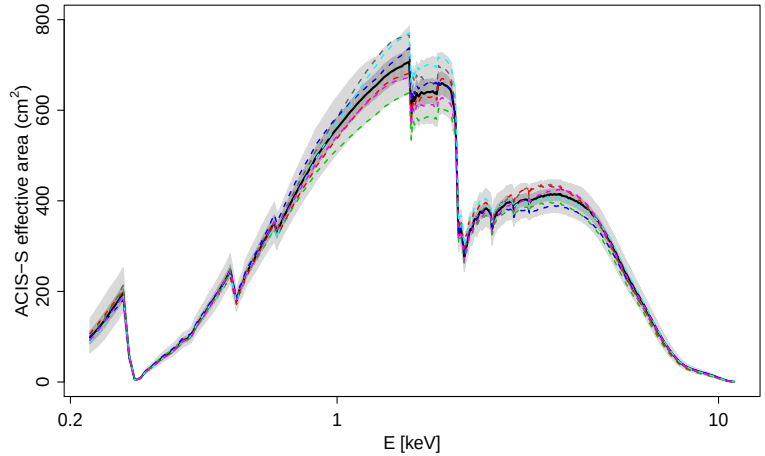
Accounting for Uncertainty in Effective Area

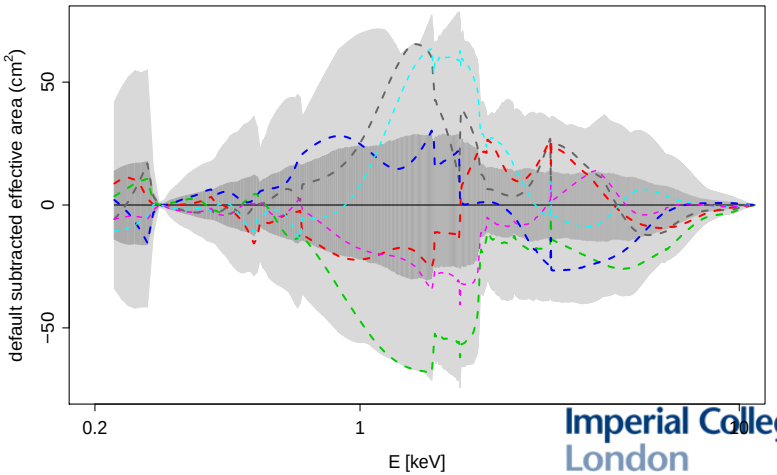
- Introduce a Bayesian approach to ***reduce*** *prior assumptions.*
- Proceed by averaging the

  standard model, $p(\theta|A, Y)$, over uncertainty in A, $\pi(A)$:

$$\pi(\theta|Y) = \int p(\theta|A, Y)\pi(A)dA.$$

- Use PCA summary of calibration sample to derive prior for A.
- No parametric models needed!

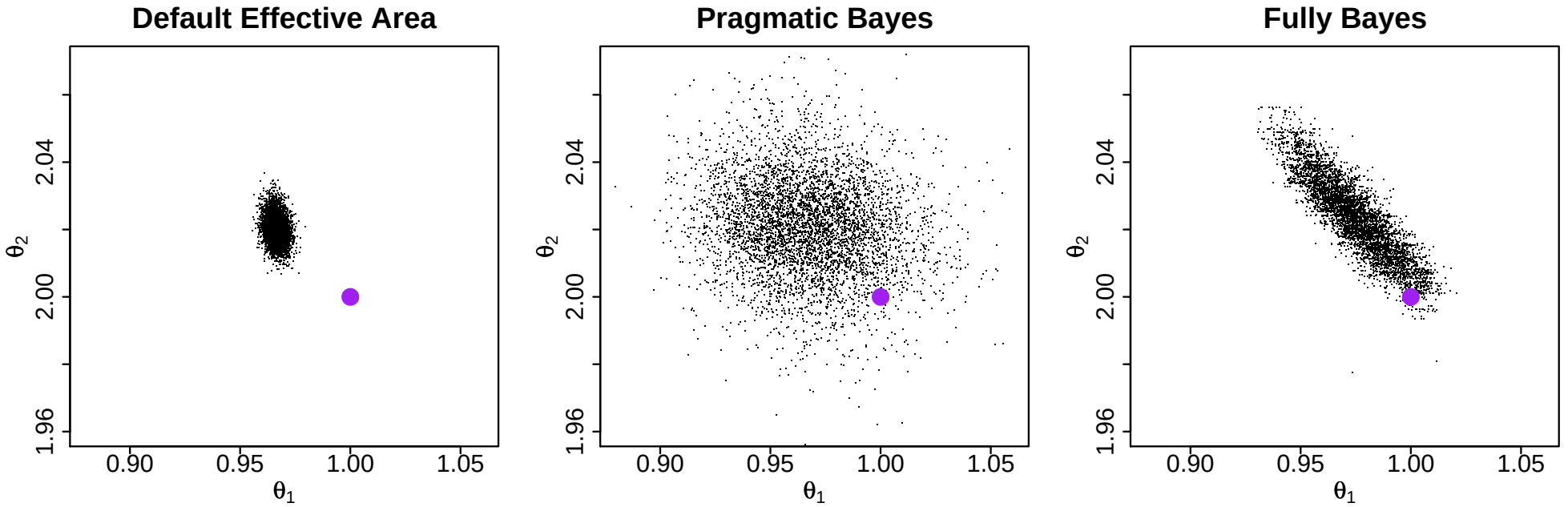




Xu, DvD et al (2014) Astrophysical J **794**, 97.
Chen, DvD et al (2019+) JASA, to appear, arXiv:1711.09429

Imperial College London

Sampling From the Full Posterior



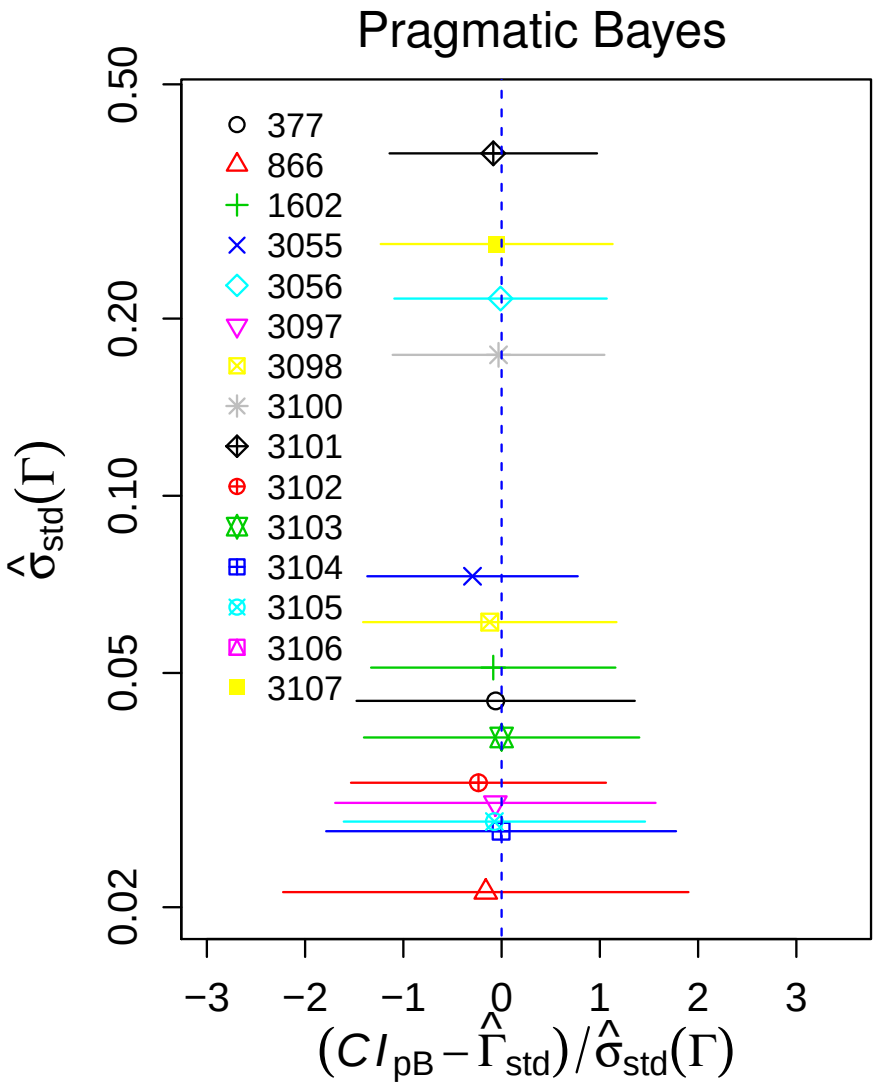
Spectral Model (purple bullet = truth):

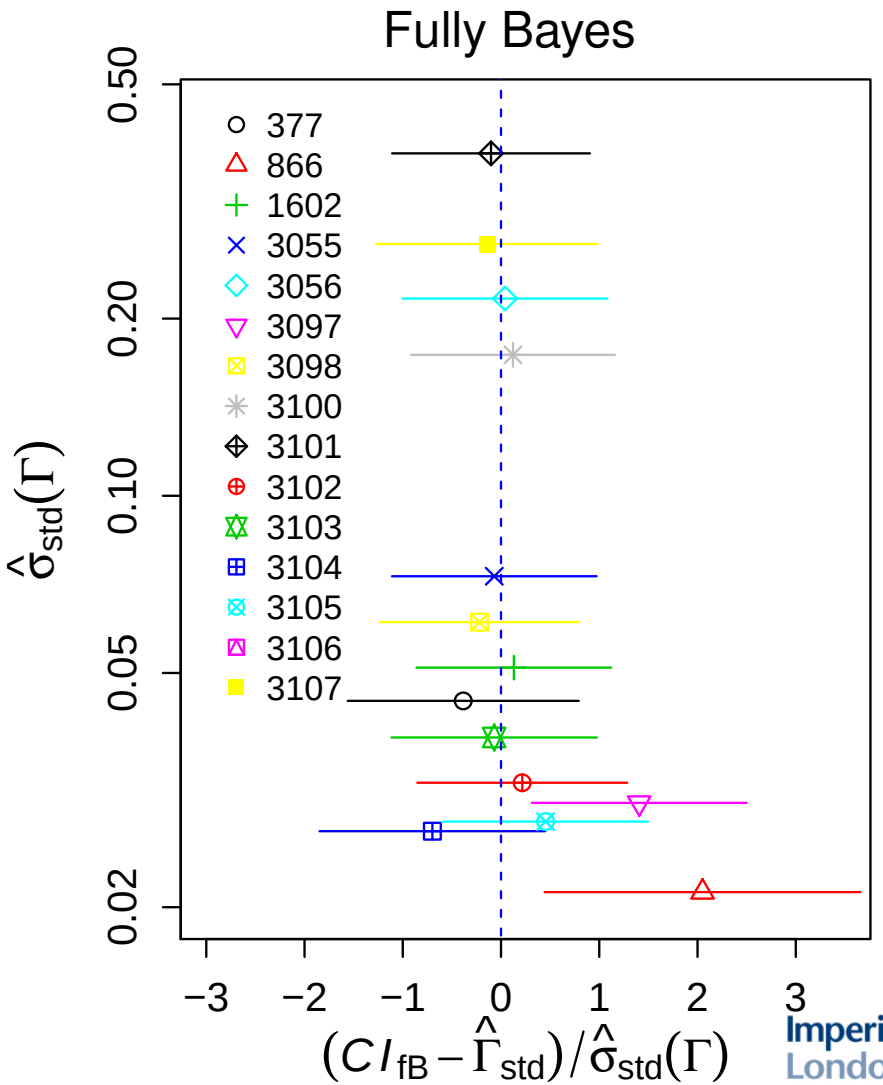
$$f(E_j) = \theta_1 E_j^{-\theta_2}$$

Pragmatic Bayes is clearly better than standard method, but a Fully Bayesian Method is the ultimate goal.

Imperial College London

How it Works on a Sample of Radio-Loud Quasars





Imperial College London

Unfolding and Deconvolution

Nice overview by Mikael K, Stefan S, Phillip R, Stephen D, Xin Q

- X is a smeared / blurred version of "ideal" data Y.
- Suppose $Y \sim f$ so that $X \sim f \circ K$
- Unfolding ignores f and models Y as Multinomial(p), fits p by ML or Penalized ML.

... stopping EM early is an outdated strategy.

Better Strategy *... but what if there are complex errors in K?*

- Compare $f \circ K$ directly with X.
- No need for regularization (f will provide it automatically).
- Or comparre $f_1 \circ K$ with $f_2 \circ K$.

If smeared data can't distinguish models.... like mass hierarchy!

- Avoid background subtraction. (STEPHEN D CAN DO!)
- LIRA: Bayesian, estimate regularization on the fly!

Weak structure in f – Esch, DvD et al (2004), Astrophysical J, **610**, *1213*

Imperial College London

Bayesian vs Frequentist – Large Sample Asymptotics

Frequentist justification of likelihood based methods:

under certain conditions...

1. $\hat{\theta}_{\text{MLE}}$ is an *asymptotically* unbiased estimator of θ
2. The sampling variance of $\hat{\theta}_{\text{MLE}}$ goes to zero *as* $n \to \infty$.
3. (standardized) $\hat{\theta}_{\text{MLE}}$ *converges* in distribution to normal.

Bayesian estimates enjoy the same asymptotic properties!

if prior assigns positive probability to a neighborhood of θ

- Large sample asymptotics are primary justification for likelihood-based methods.
- Bayesian methods enjoy alternative (small sample) justification.

Imperial College London

When to worry

If your analyses are based on asymptotic frequency properties,

- your data being Gaussian is not enough.

You need to watch for warning signs....

- strange (non-convex?) contours (HIMMEL)
- MLE/MAP on boundary of parameter space
- confidence intervals contain non-physical values

If asymptotics don't apply investigate frequency properties via Monte Carlo!

... or base inference on small sample justification of Bayesian analyses.

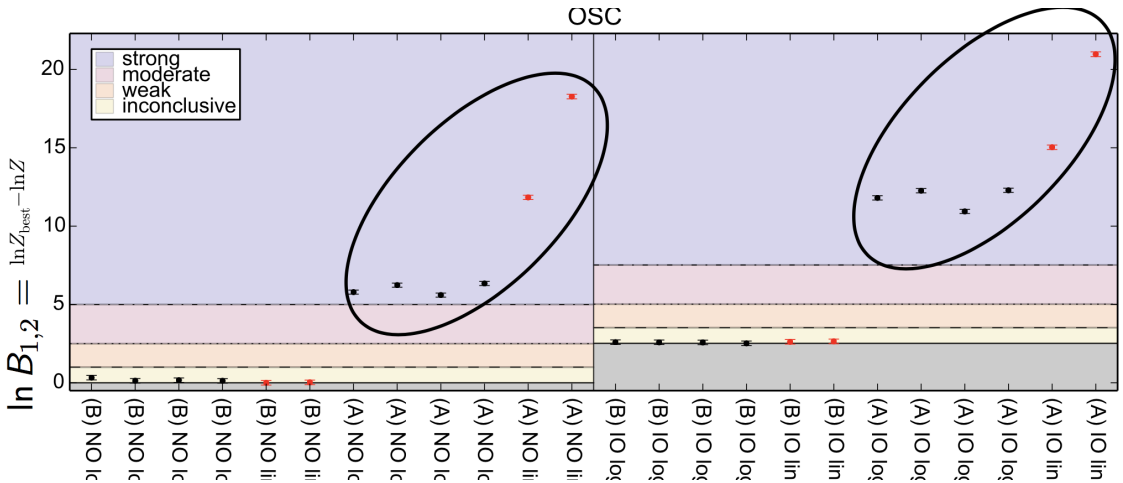
Imperial College London

Sensitivity to Choice of Prior

Prior sensitivity is concern *(even for Bayesians!)* (ALEX, STEFANO, MATTEO)

Despite relative insensitivity with parameter estimation – Jim B

Stefano on Normal vs Inverted Hierarchy:



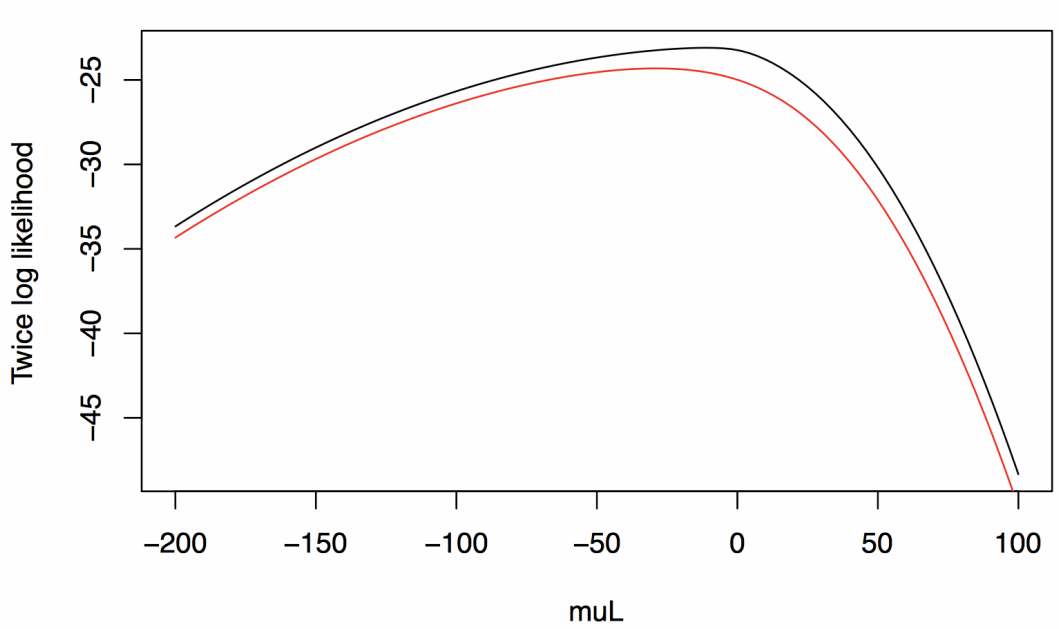
Why the difference? Not typically sensitive to prior (JIM)

1. Choice of parameterization? (m_1, m_2, m_3) *vs* $(m_{\min}, \Delta m^2_{21}, |\Delta m^2_{31}|)$
2. Choice of prior? *MLE and posterior are invariant, but not MAP or mean*
3. Including different constraints or external information?

Imperial College London

But then again, maybe the prior is influential...

Anthony D computed profile likelihood:



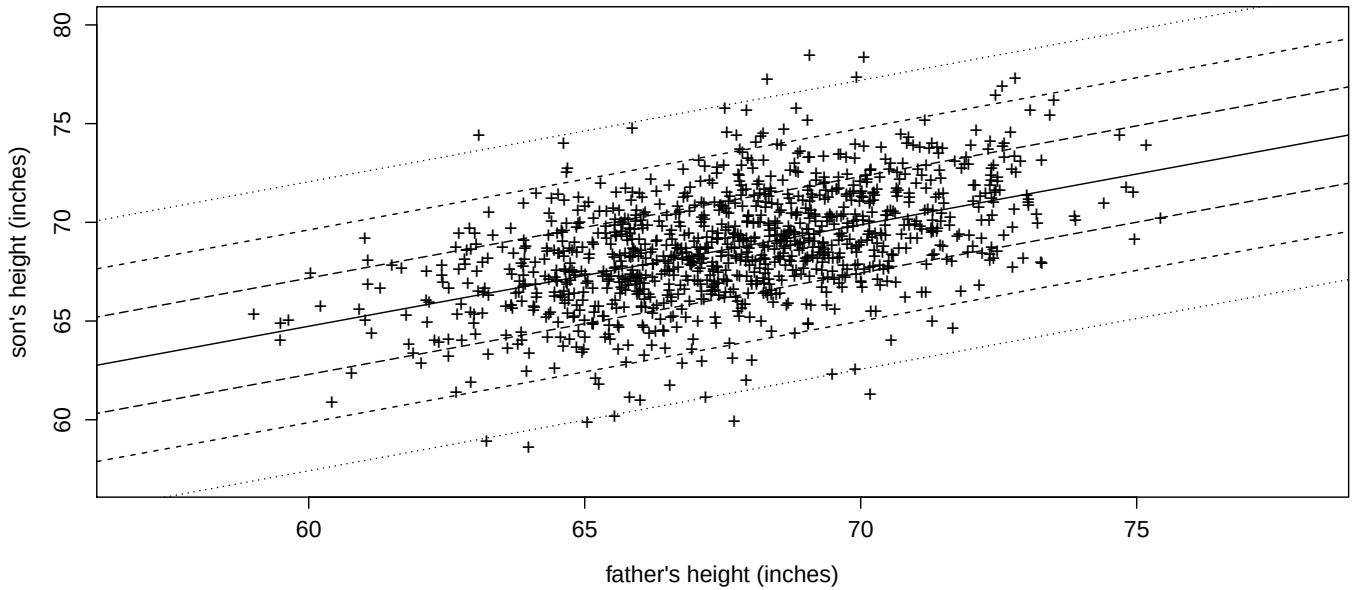
$2\ell_{\mathrm{p}}(\mu_L)$ for normal hierarchy (black) and inverted hierarchy (red)

Anthony: "Need more and different data"

Imperial College London

When Prior Matters – Simple Example

Suppose we are interested in correlation, ρ, of the heights of father's and their adult sons.



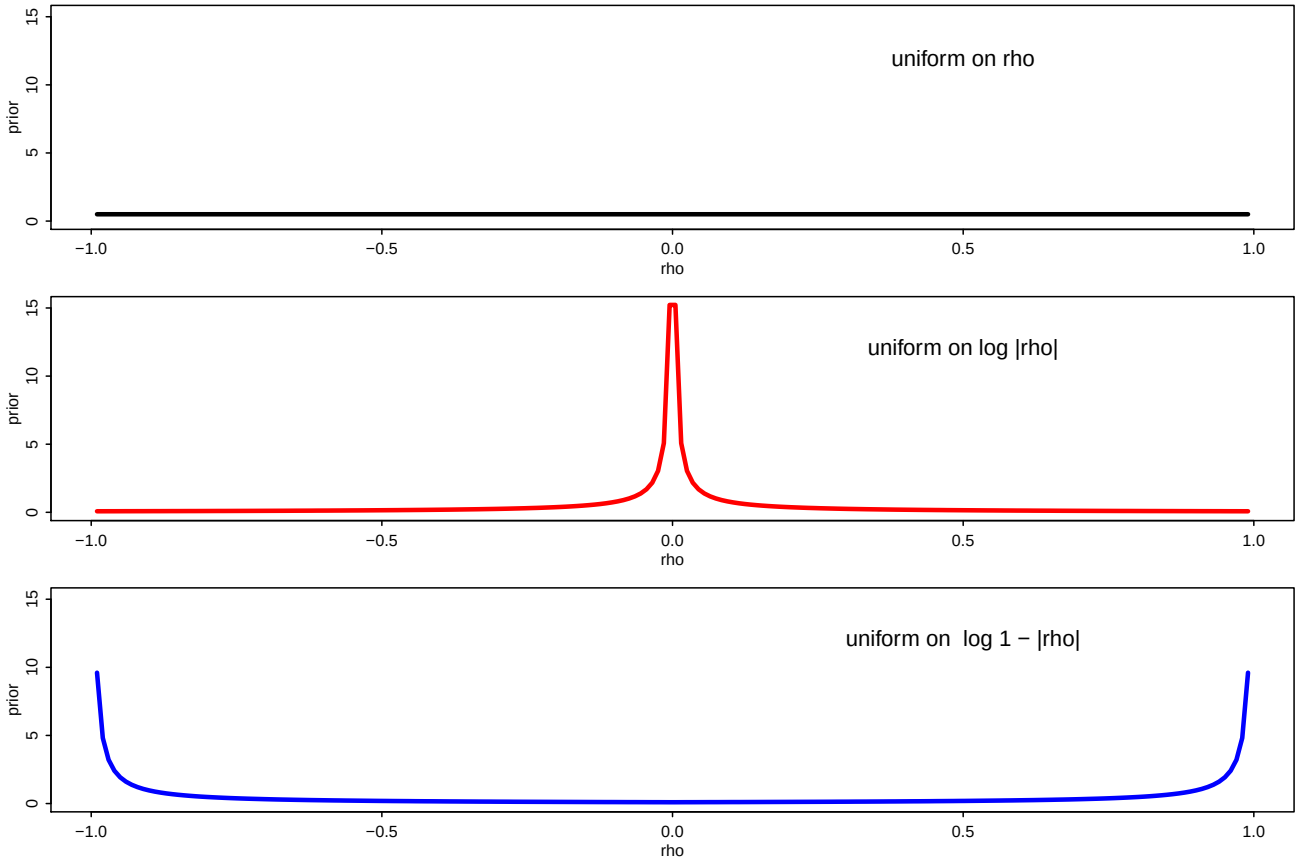
Model with bivariate Gaussian distribution

Imperial College London

When Prior Matters – Simple Example

What if we lost the coupling between fathers and sons?

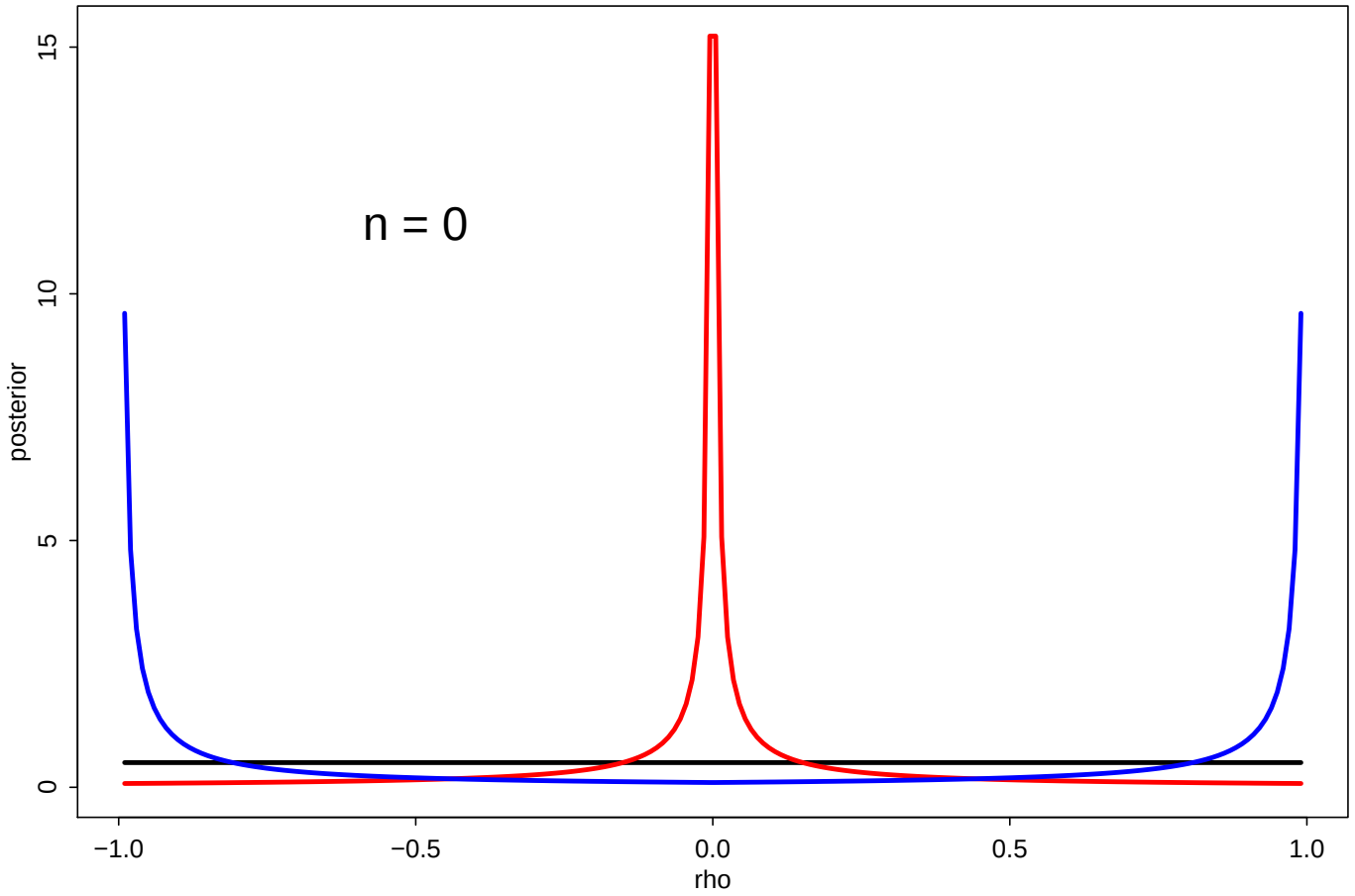
- Easy to estimate means and standard deviations
- For correlation, ρ, prior will matter – consider three choices.



Imperial College London

Posterior for ρ with $n = 0$

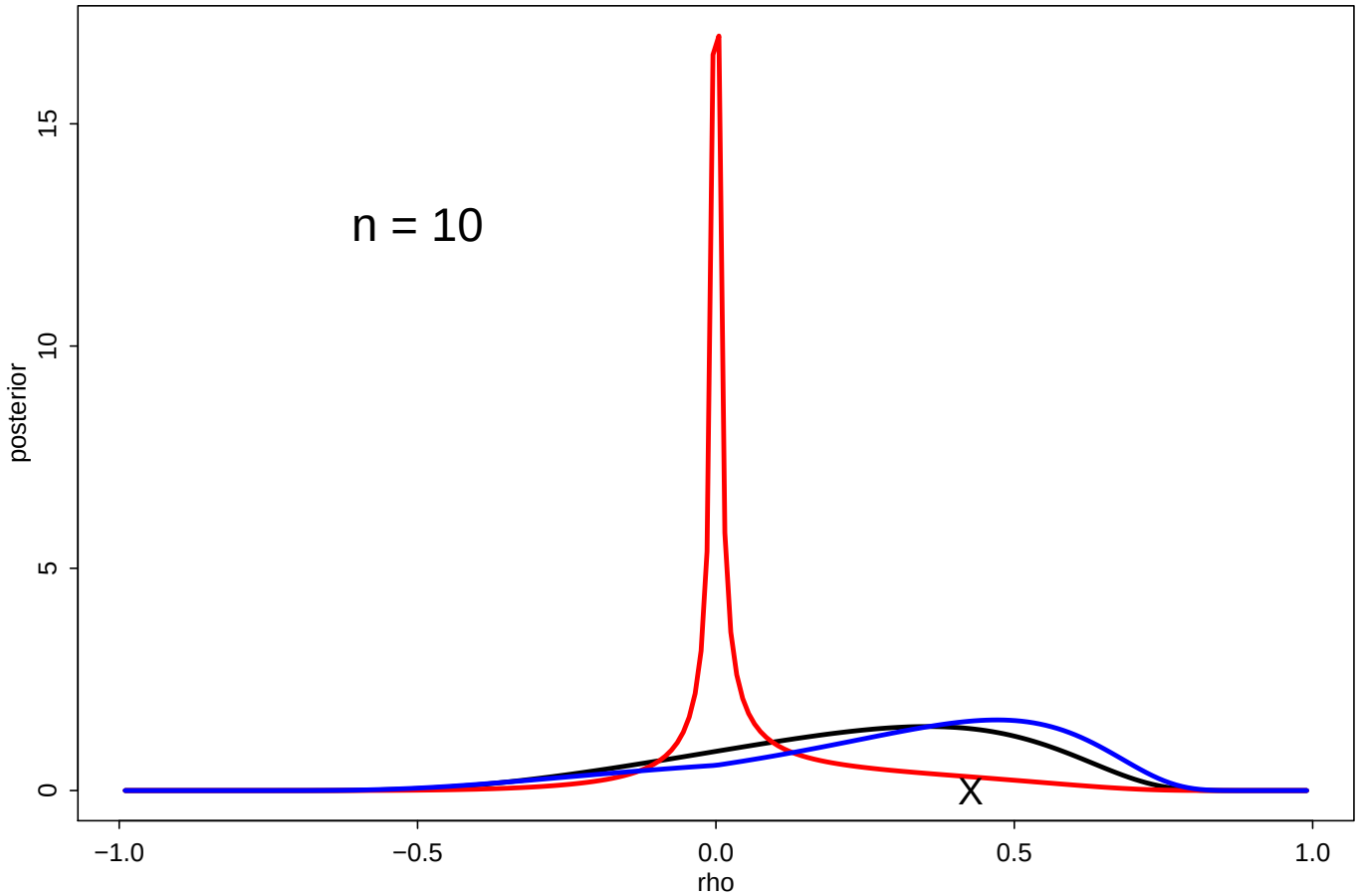
$n = 0$, no bivariate pairs



Imperial College London

Posterior for ρ with $n = 10$

$n = 10$, the number of bivariate pairs

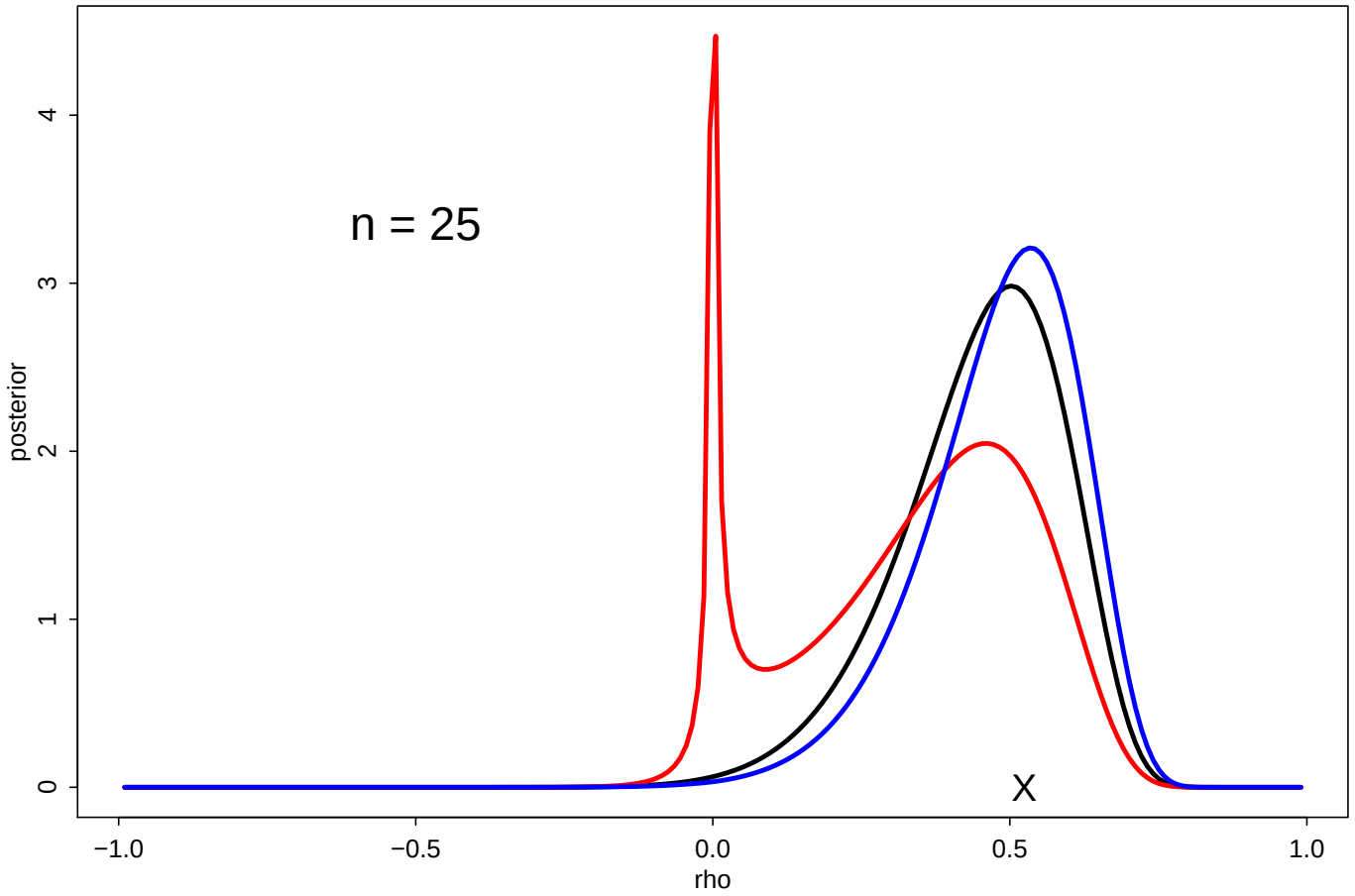


If prior is overly influential, may have bigger problems.

Imperial College London

Posterior for ρ with $n = 25$

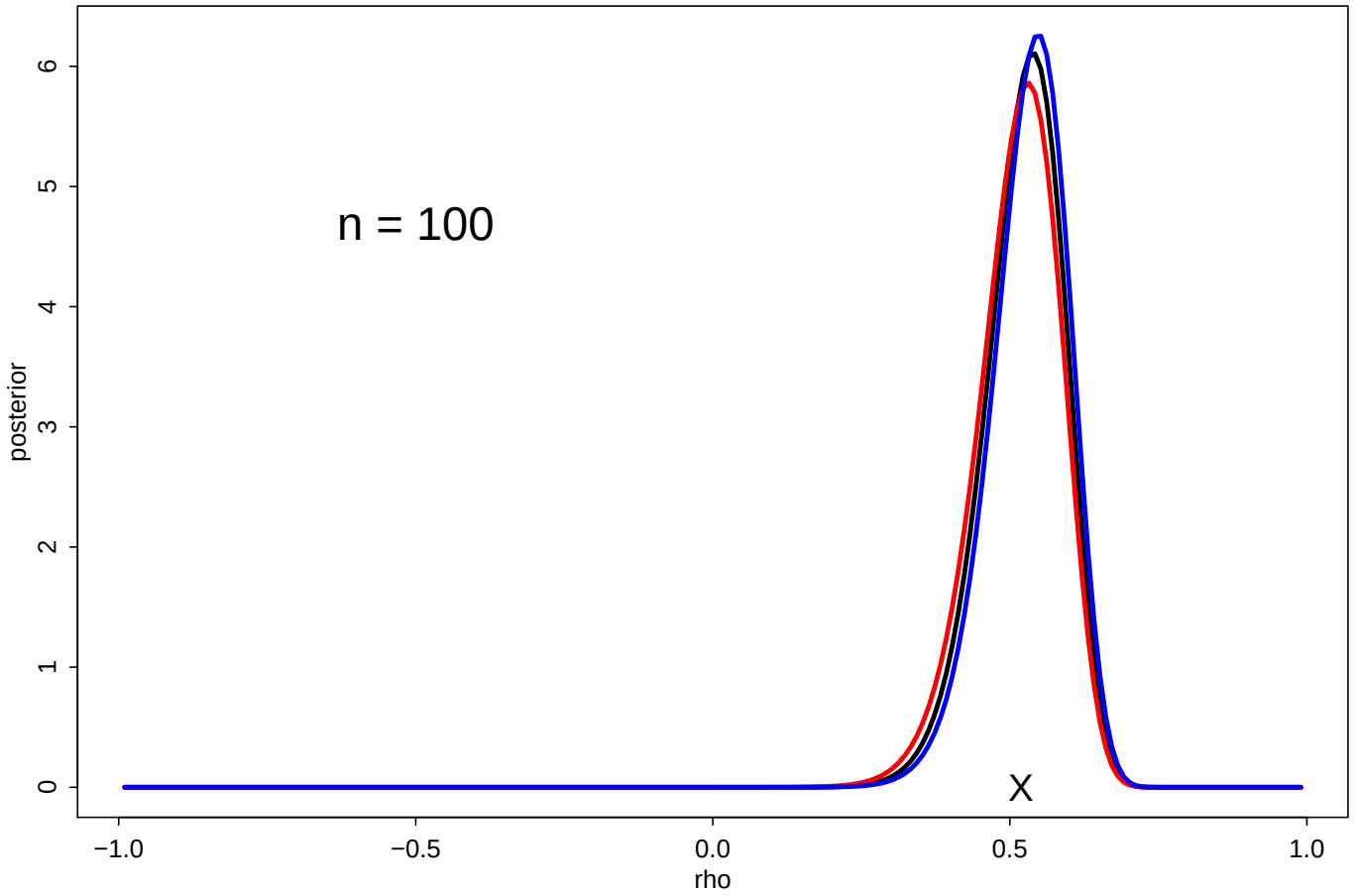
$n = 25$, the number of bivariate pairs



Imperial College London

Posterior for ρ with $n = 100$

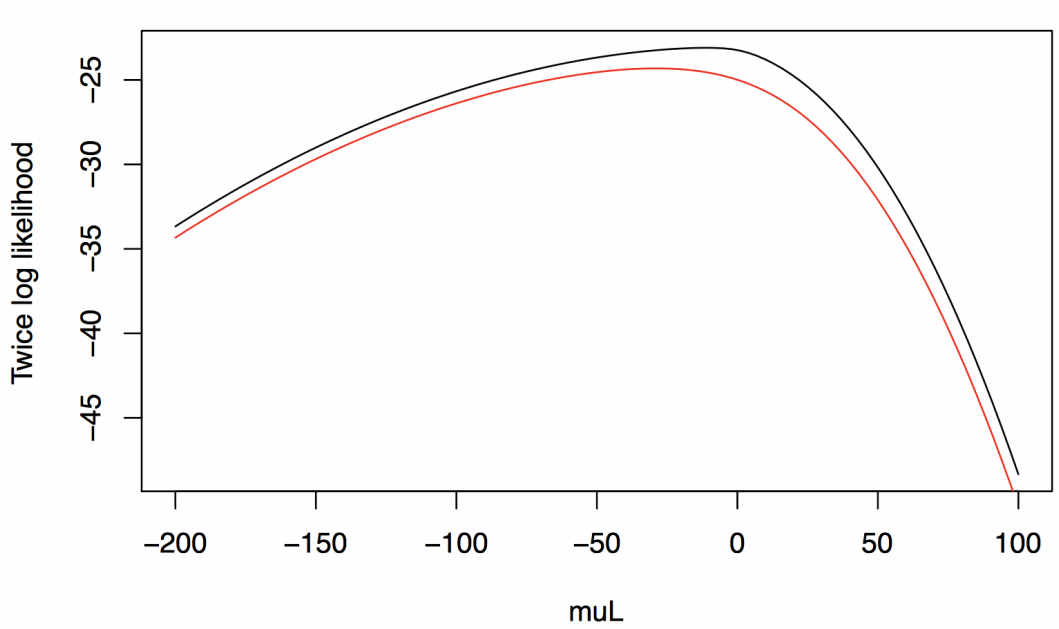
$n = 100$, the number of bivariate pairs



Imperial College London

Back to Mass Hierarchy

Anthony D computed profile likelihood:



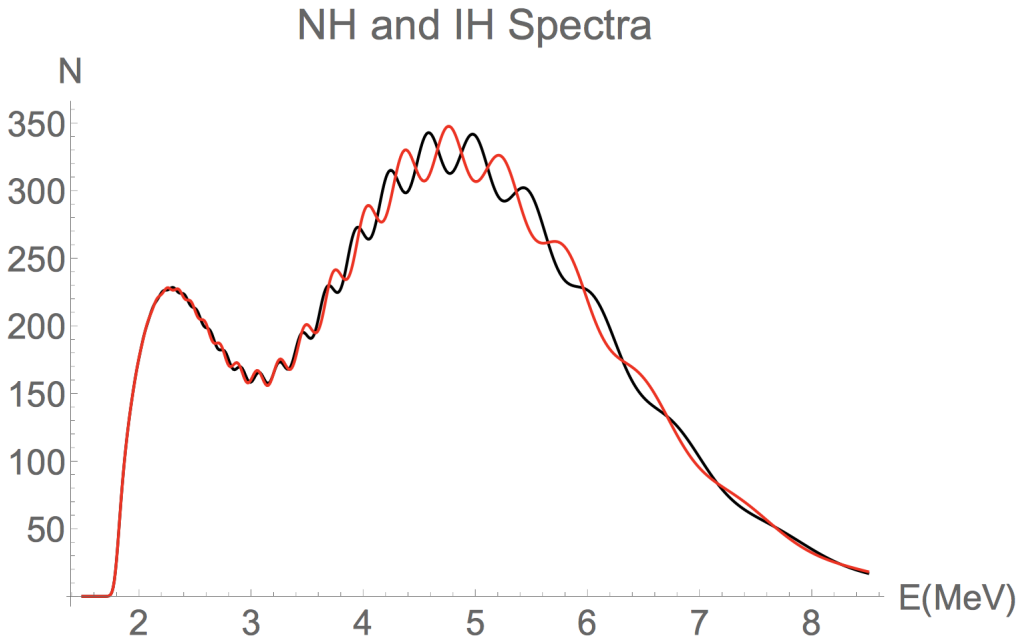
$2\ell_{\mathrm{p}}(\mu_L)$ for normal hierarchy (black) and inverted hierarchy (red)

Anthony: "Need more and different data"

Imperial College London

Back to Mass Hierarchy

But there appears to be more to the data: (EMILIO C)



Expected spectra for normal and inverted hierarchy at 53km, finite energy resolution

Emilio argues that distinguishing NH from IH is difficult, but looks like more than three Gaussian variates.

Imperial College London

When The Prior Matters – Strategies

Avoid confusing "uniform" with "uninformative"

"The statisticians will kill me [for using uniform prior]" – Stefano

- What does "uniform" mean? Depends on choice of scale.
- On a unit hypercube of dimension D,

$$\Pr(\text{within } \epsilon \text{ of boundary}) = 1 - (1 - 2\epsilon)^D.$$

With $\epsilon = 0.01$:

- $D = 10$ gives $\Pr(\text{within } \epsilon \text{ of boundary}) = 18\%$
- $D = 35$ gives $\Pr(\text{within } \epsilon \text{ of boundary}) = 50\%$

... the curse of dimensionality (CHAD S)

Reference Prior: Maximize (some measure of, e.g, KL, Hellinger) expected discrepancy between prior and posterior distributions. (JIM)

Sensitivity Analysis: Try multiple priors — and sample sizes.

Best to use subjective prior, derive a reference prior, or – best of all – get more data!

Imperial College London

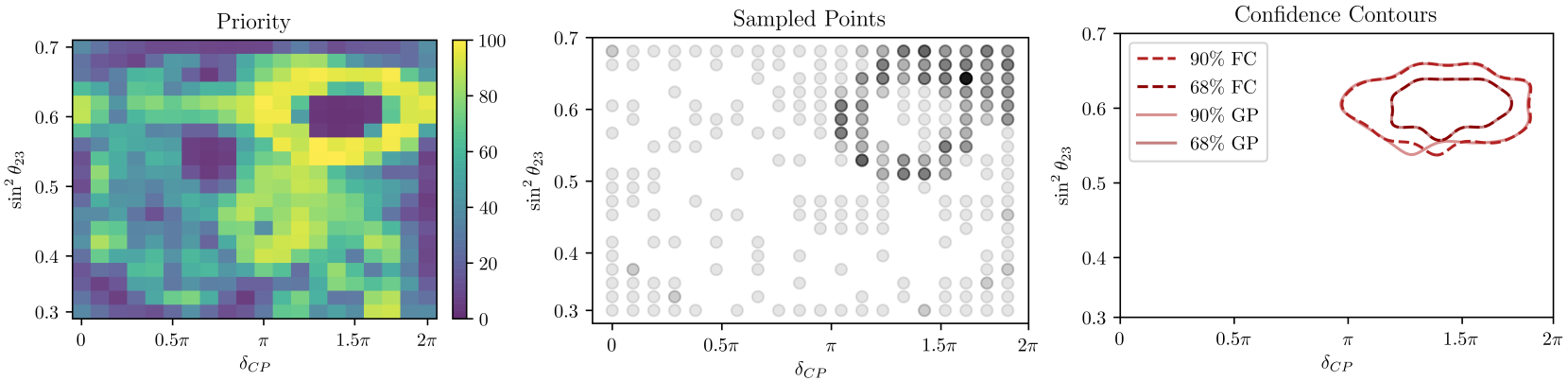
Feldman Cousins

Feldman-Cousins remains popular as ever
(ALEX H, NICHOLAS W, MATTEO A, CHAO Z, BANNANJE N)

Bob C provided a nice review and history.

Bannanje on Efficient Inference with Gaussian Processes

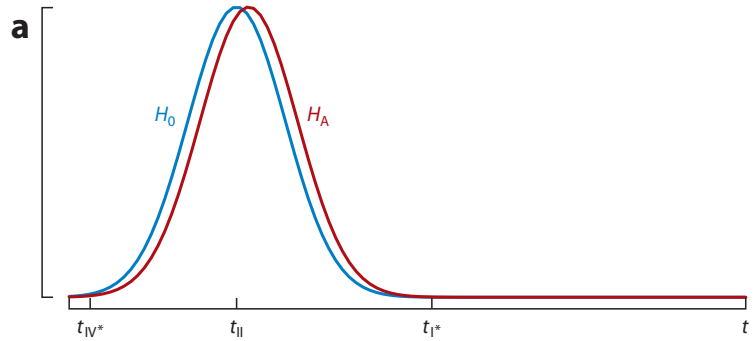
- Use Gaussian Process to interpolate and save computing



Can this overcome the computation cost of FC??

Imperial College London

CL_S – Avoiding Exclusion Under an Insensitive Test



(CHRISTOPHE B, CHAO Z)

- Exclusion is unwarranted
- Do not exclude H_1 if both $1 - p_0$ and $1 - p_1$ are small

Read (2000) suggested excluding H_1 only if

$$\mathrm{CL}_S = \frac{1-p_1}{1-p_0} = \frac{\Pr(T < t_{\mathrm{obs}}|H_1)}{\Pr(T < t_{\mathrm{obs}}|H_0)} \leq \alpha.$$

Exclude H_1 if $T < t_{\mathrm{obs}}$ much less likely under H_1 than under H_0

Bob C: Better to report both p-values.

DvD: Three parameter sets: no sensitivity, excluded, not excluded.

Imperial College London

Outline

Imperial College London

The Problem with P-values

The misuse of P-values:

- Replace data with "data as extreme of more extreme"
  – not particularly conservative.
- Often mistakenly interpreted as $\Pr(H_0)$, but:
  - Cannot be calibrated vis-vis $Pr(H_0)$.
  - **Do not measure relative likelihood of hypotheses.**
  - Can vastly overstate evidence for H_1. (JIM)
- May depend on bits of H_0 that are of no interest.
- Single filter for publication / judging quality of research.
- Cherry-picking results based on p-value / publication bias.

Reviewers, Editors, and Readers want a simple black-and-white rule: $p < 0.05$, or $> 5\sigma$.

But statistics is about quantifying uncertainty, not expressing certainty.

Imperial College London

5σ Discovery Threshold

5σ is required for "discovery" (Louis)

- High profile false discoveries led to conservative threshold
- Treat Bump Location as known (multiple-testing) (DvD)
- "What would you have done had you had different data"
- **Calibration, systematic errors, and model misspecification**
- Of course **cranking down α does not address these issues**

"In particle physics, this criterion has become a convention ... but should not be interpreted literally [1] *."*

Bob: *"Two* 3.5σ *results are better than one* 5σ *result."*
DvD: *"Calibrated* 3.5σ *result is better than uncalibrated* 5σ*."*
Louis: *"Extraordinary claims require extraordinary evidence."*
Thamaso *"Can we agree not to quote more than* 5σ *??."*

Imperial College London

The Problem with Priors

Bayesian methods have challenges of their own. (BOB C)

$$\text{Bayes Factor} = \frac{p_0(y)}{p_1(y)} \text{ with } p_i(y) = \int p_i(y|\theta)p_i(\theta)d\theta.$$

$$\Pr(H_0 \mid y) = \frac{p_0(y)\pi_0}{p_0(y)\pi_0 + p_1(y)\pi_1} = \frac{\pi_0}{\pi_0 + \pi_1(\text{Bayes Factor})^{-1}}$$

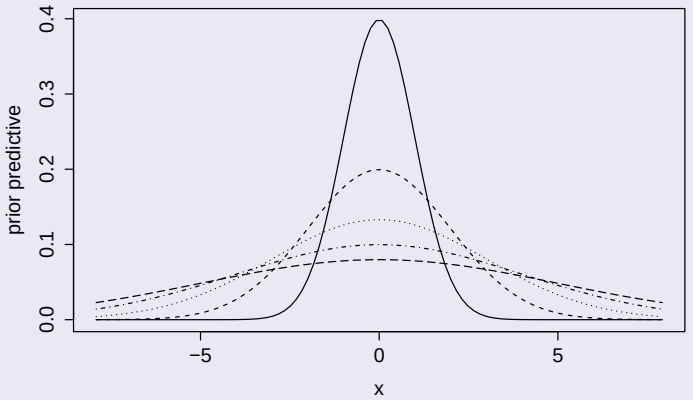
Example:

Likelihood: $y \sim \mathrm{N}(\mu, 1)$

Test: $\mu = 0$ vs $\mu \neq 0$

Prior Dist'n: $\mu \sim \mathrm{N}(0, \tau^2)$

Prior Pred.: $y \sim \mathrm{N}(0, 1 + \tau^2)$



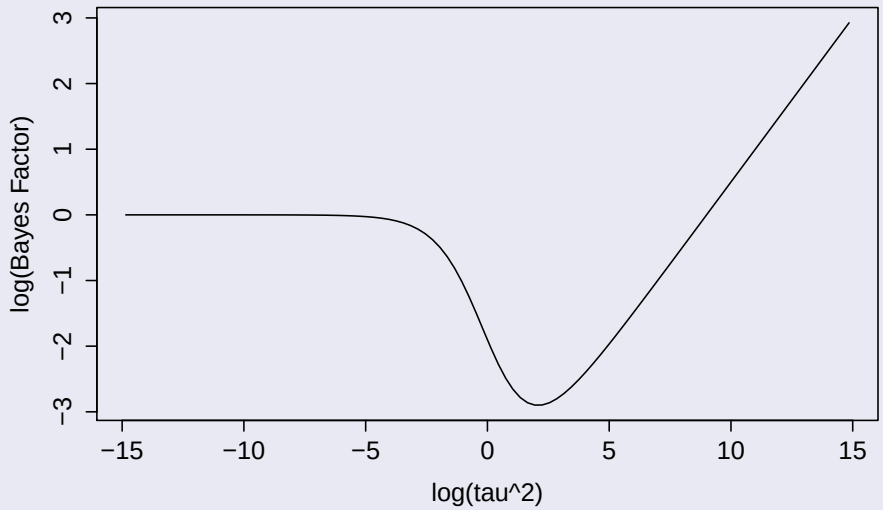
Value of $p_1(y)$ depends on τ^2!

Choice of Prior Matters!

Bayes Factor (JIM B, BOB C, ALEX H, STEFANO G)

$$H_0: \quad y \sim \mathrm{N}(0,1).$$
$$H_A: \quad y \sim \mathrm{N}(0, 1+\tau^2).$$

- Observe $y = 3$
- log(Bayes Factor)



Must think hard about choice of prior and report!

Imperial College London

Bayes Factors and Likelihood Ratios

Likelihood Ratio optimizes parameters, whereas Bayes Factor marginalizes.

$$\text{Likelihood Ratio} = \frac{\max_{\theta_0} p_0(y \mid \theta_0)}{\max_{\theta_1} p_1(y \mid \theta_1)} \neq \text{Bayes Factor} = \frac{\int p_0(y \mid \theta_0)\, p(\theta_0)\, \mathrm{d}\theta_0}{\int p_1(y \mid \theta_1)\, p(\theta_1)\, \mathrm{d}\theta_1}$$

....unless there are no parameters under either model.

A Bayesian Occam's Razor (Bob C, Stefano G)

- Suppose $p(\theta_i)$ are both essentially flat over range where corresponding likelihoods are non-negligible.

$$\text{Bayes Factor} = \frac{\int p_0(y \mid \theta_0)\, p(\theta_0)\, \mathrm{d}\theta_0}{\int p_1(y \mid \theta_1)\, p(\theta_1)\, \mathrm{d}\theta_1} \approx \frac{p(\hat{\theta}_0) \int p_0(y \mid \theta_0)\, \mathrm{d}\theta_0}{p(\hat{\theta}_1) \int p_1(y \mid \theta_1)\, \mathrm{d}\theta_1}$$

- The term $p(\hat{\theta}_0)/p(\hat{\theta}_1)$ is sensitive to dimension and scale.
  - At mode, multivariate normal prior $\propto 1/|\Sigma|^{d/2}$.
- Bayes Factor penalizes larger models. *...and depends strongly on choice of prior.*
- Don't hide your priors!

A Taxonomy of Tests

Different types of tests involve different methods

- Simple vs Simple
- Nested I: One-sided tests
- Nested II: Precise null (with H_0 on boundary)
- Nested III: Parameters undefined under H_0
- Non-Nested

In each case we can consider relative advantages of p-values an Bayesian methods.

Imperial College London

Simple vs Simple

Simple vs Simple

- H_0 and H_1 are fully specified: no unknown parameters.
- E.g., normal hierarchy vs inverted hierarchy.
- Bayes Factor $= \frac{p_0(y)}{p_1(y)} =$ Likelihood Ratio

P-values: log(Likelihood Ratio) $\sim$ NORMAL (large n) (ANDREY)

(Must use Monte Carlo to specify two null distributions.)

Bayesian: No problem with priors!

Methods give consistent results.

Andrey gives example when used for trigger

... but with small n not Gaussian

Everything is simple, but models rarely fully specified

Imperial College London

Nested I: One-sided Tests

Nested I: One-sided Tests (JIM B)

- $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_0$.
- E.g., $H_0 : \Delta m^2_{32} \leq 0$ versus $H_1 : \Delta m^2_{32} > 0$.

P-values: p-value $= \sup_{\theta \leq \theta_0} \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid \theta\right)$ *(Use Wilks Thm.)*

Bayesian: Avoid $p_0(y)$ and $p_2(y)$: $\Pr(H_0 \mid y) = \Pr(\theta \leq \theta_0 \mid y)$.

- Requires only one model and one prior specification.
- *Can incorporate external knowledge into Bayesian analysis via prior, e.g.,* $|\Delta m^2_{32}| = 2.43 \pm 0.13$.
- *Mass hierarchy can be handled this way* *(frequency or Bayesian)*

...much easier than non-nested model comparison.

Again methods give consistent results.

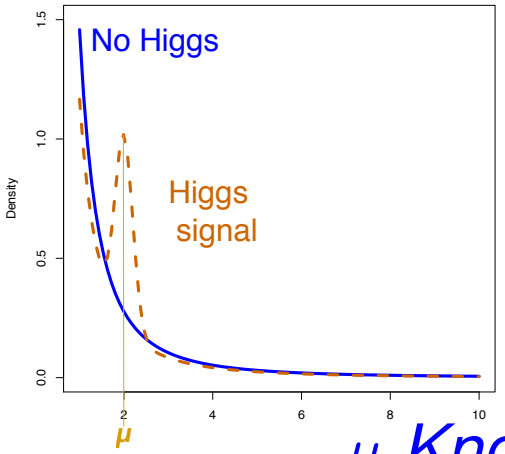
Imperial College London

Nested II: Precise Null (with H_0 on boundary)

Nested II: Precise Null

$$f(y_i|\theta) = (1-\lambda)f_0(y_i|\alpha) + \lambda f_1(y_i|\mu)$$
$$= \text{background} + \text{Higgs}$$

- f_1 is fully specified: i.e., μ is known.
- $H_0 : \lambda = 0$ (no discovery)
  $H_1 : \lambda > 0$ (discovery)



μ Known!

P-values: LRT: Wilks does not apply, use Chernoff.

Bayesian: Choice of priors on λ matters! (JIM B)

P-values $\ll \Pr(H_0 \mid y)$.

Example: Neutrino-less Double beta decay (MATTEO A)

Bayes and P-value differ – why not report both?
"Important to understand" what each means.

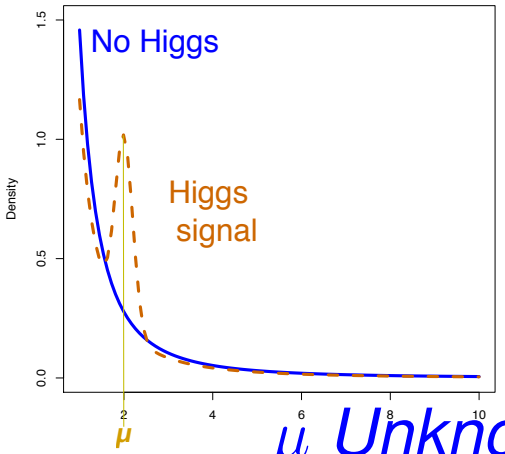
Imperial College London

Nested III: Precise Null (with Parameters Undefined Under H_0)

Nested III: μ undefined under H_0

$$f(y_i|\theta) = (1-\lambda)f_0(y_i|\alpha) + \lambda f_1(y_i|\mu)$$

- μ unknown, no value under H_0.
- $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$



μ *Unknown!*

P-values: Bound global p-value

- Look elsewhere effect, method of GV. (PHILLIP L)

Bayesian: Choice of priors on λ and μ matter!

- Use prior on μ to correct for LEE.

Examples: How to correct for harmonic bumps? (PHILLIP L)

check out Sara Algeri's arXiv:1701.06820 and 1803.03858

other examples (CHAO Z, NICHOLAS W, BIRGIT N)

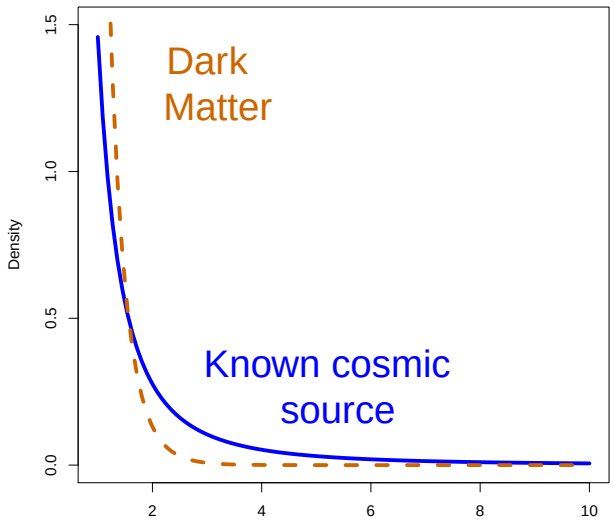
Why are local p-values still reported?

Imperial College London

Non-nested models

Non-nested models

- two *parameterized* non-nested models
- H_0: γ-ray energy of known cosmic sources
  H_1: γ-ray energy of dark matter.
- H_0: normal hierarchy
  H_1: inverted hierarchy .
- Is there a *null* model?



P-values: Embed in mixture model and bound global p-value or Monte Carlo (toys).

Bayesian: No problems in principle.

... but choice of prior may cause difficulties.

Best to avoid if you can! E.g., mass hierarchy.

Imperial College London

Outline

Imperial College London

Strategies

What is a physicists to do?

- Controlling false discovery is critical in physical sciences.
  and itervals that contain the truth... some of the time!
- Comparing p-values with a predetermined significant level can control false discovery.... *if used with care, e.g., no cherry picking!*
- When confronted with small p-values researchers *...even statisticians!!...* may believe H_0 is unlikely.
- Bayesian solutions can better quantify likelihood of H_0 / H_1.
- **Solution:** Compute both *global* p-value *and* Bayes Factor.

Careful

1. *global corrections for p-values*
2. *choice and validation of prior distributions*

remain challenging!

Imperial College London