Statistical Issues in Astronomical Searches:
A Statistician's Perspective

David A. van Dyk*

Statistics Section, Imperial College London

Progress on Statistical Issues in Searches, June 2012

*I thought it would be safest to use an assumed identity

The Model Selection & Checking Problems

- Typically begin with baseline, default, or presumed model:
  Null Hypothesis: There is no source.
  - Model Checking: Is the model consistent with the data?
  - If not, characterize inconsistency, improve model, recheck.
- May have another model that we suspect or hope is better:
  Alternative Hypothesis: There is a source.
  - Model Selection / Comparison: Decide between or weigh the evidence for the two (or more?) models.
- These are surprisingly subtle problems:
  - No consensus exists on how to proceed.
  - Disagreement between Bayesian and Frequentist methods.

Neyman-Pearson

Model Selection:

H_0 There is no source.

H_A There is a source.

- Need test statistic, T, with known distribution under H_0.
- Threshold T* is the smallest value such that

$$\Pr(T > T^* \mid \text{no source}) \leq \alpha,$$

If $T > T^*$ sufficient evidence to declare a detection.

Assessment?

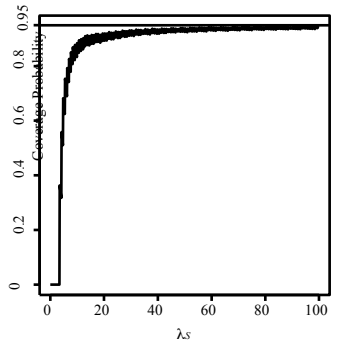
Pro: Frequency properties: Bounded Pr(false positive).

Con: No characterization of the strength of evidence.

How to find T??

What should be Reported?

- Confidence Interval is often only reported if source is detected.
- But deciding whether to report an interval based on the data alters its frequency properties.



Unfortunately, frequency properties depend on what you would have done, had you had a different data set.

Bayes Factors and Posterior Probabilites

Bayesian methods have no trouble with unknowns parameters

- The prior predictive distribution:

$$p_i(x) = \int p_i(x|\theta)p_i(\theta)d\theta$$

- How likely is X under model i (likelihood + prior dist'n).
- Compare two models with the Bayes Factor:

$$\text{Bayes Factor} = \frac{p_0(x)}{p_A(x)}.$$

or the posterior probability of H_0:

$$\Pr(H_0|x) = \frac{p_0(x)\pi_0}{p_0(x)\pi_0 + p_A(x)(1-\pi_0)}.$$

add: what if neither model is acceptable?

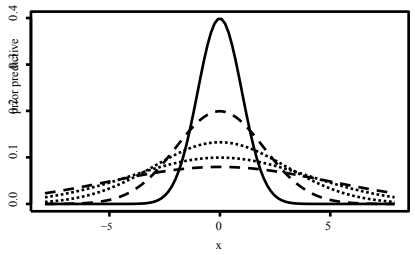
The Choice of Prior Dist'n Matters!

Example:

Likelihood: $X \sim N(\mu,1)$.

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

Prior Pred.: $X \sim N(0,1+\tau^2)$.



Value of $p_A(x)$ depends on τ^2!

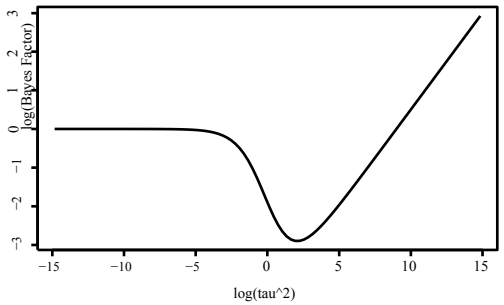
Must think hard about choice of prior and report!

The Choice of Prior Dist'n Matters!

Bayes Factor:

H_0 : $X \sim N(0,1)$.

H_A : $X \sim N(0,1+\tau^2)$.



Assessment of Bayes Factors.

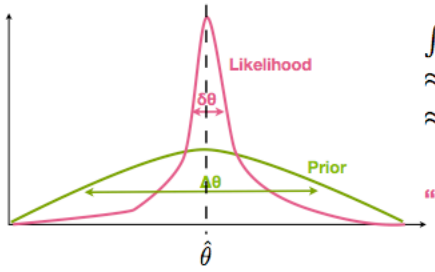
Cons: Bayes Factor depends heavily on the prior scale.

Pros: Probability based principled method, answers right question, no problem with nuisance parameters.

An in-built Occam's razor

Imperial College London

- The Bayesian evidence balances *quality of fit* vs *extra model complexity*.
- It rewards highly predictive models, penalizing "wasted" parameter space.
- **The prior here is important:** it quantifies the *predictive power* of the model.



$$P(d|M) = \int d\theta L(\theta) P(\theta|M)$$
$$\approx L(\hat{\theta}) \delta\theta P(\hat{\theta})$$
$$\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})$$

"Occam's factor"

Quality of fit

How to Choose the Prior Dist'n.

- Unlike with parameter inference, prior must be proper.
  - Prior Predictive Distribution is improper with improper prior!
- There is no default prior distribution.
- Possible Solutions
  1. Minimize Bayes Factor over a class of priors (see below).
  2. Use a subjective prior distribution.
- Subjective prior distributions are especially illusive:
  What are likely values a parameters in a possible model?
- Problem is even more complicated when:
  - Parameter space is large.
  - H_0 and H_A have different (non-nested) parameters.

A Dangerous Method....

Although the use of p-values is endemic in data analysis, they are not easily interpreted (for a precise H_0[1]):

1. When compared to Bayes Factors or $Pr(H_0|data)$, p-values vastly overstate the evidence for H_1.
   - Even using the prior most favorable to H_1 (in a large class).
2. Computed given data as extreme or more extreme than X.
   - This is much stronger evidence for H_1 than X.
   - Agree with Bayes measures given "as/more extreme".
3. P-values cannot be easily calibrated with Bayes Measures
   - Depends on sample size, model, and precision of H_0.

P-values bias inference in the direction of false discovery.

[1] Berger & Delampady, Testing Precise Hypotheses, Stat. Sci., 1987

Not a Frequentist Method...

> "... a rough rule known to astronomers, i.e., that differences up to twice standard error usually disappear when more or better observations become available, and that thoes of three or more times usually persist." [2]

- Suppose over time, H_0 is true about half the time.
- Looking back over results with $1.96 < \text{p-value} < 2.00$, the astronomer might find H_0 to be true 30% of the time.
- The absolute minimum limiting proportion is 22%.
- Compare with "5% significance" associated with p-value.

Why are p-values so popular?

[2] Jeffrey (1980) in Berger & Delampady (1987)

Why are p-values so popular?

Maybe it is just a bad habit....

Assessment of P-values

Cons: Biased toward discovery and uninterpretable.

Pros: Everyone is doing it...

The Likelihood Ratio Test Statistic

Neyman-Pearson Testing and P-values require a Test Statistic.

- Often derived on a case-by-case basis.
- An important general Test Statistic: The Likelihood Ratio:

$$\text{Likelihood Ratio} = \frac{\sup_{\theta \in \Theta_0} p(x|\theta)}{\sup_{\theta \in \Theta_A} p(x|\theta)}.$$

- Compare with Bayes Factor: how small is small enough?
- *Asymptotic* chi-square distribution
  1. Θ_0 must be in the interior of Θ_A.
  2. If Θ_0 is on the boundary, but all parameters are identified under H_0, $-2\log(\text{likelihood ratio}) \overset{asy}{\sim} \text{mixture of} \chi^2$.

An alternative – the Rao score test

Based on $\partial \ell / \partial(\theta_0)$

Only requires parameter estimates under the null, which can be a big computational saving

Asymptotically chi-square (even in some conditions for which this is not true for LRT)

Locally most powerful

See also the closely related Neyman C-alpha tests

Ref: Lehman and Romano, Testing Statistical Hypotheses

Posterior Predictive P-values

Hybrid Methods: Recall the definition of the p-value:

$$\text{p-value} = \Pr(T > T_{obs}|H_0).$$

How do we compute p-value with unknown param's under H_0?

1. Careful choice of T, dist'n may not depend on unknowns.
2. Use estimates of unknowns under H_0.
3. Average over the posterior dist'n of unknowns under H_0:

$$\text{ppp-value} = \int \Pr(T > T_{obs}|H_0)p(\theta|x)d\theta.$$

ppp-values may be very weak with poor choice of T. Use LRT!

Other Methods

There are Many other methods....

1. Bayesian Model Averaging
   - Pros: Bayesian, but less dependent on the choice of prior.
   - Cons: More appropriate for prediction than model selection.
2. Decision Theory
   - Pros: Derives rules tailored to specific scientific goals.
   - Cons: Sensitive to choice of Loss Function and Prior.
3. Information Criteria (e.g., AIC, BIC, etc.)
   - Pros: Simple to compute with an intuitive form!
   - Cons: Ad hoc—with questionable statistical properties.
4. Conditional Error Probabilities
   - Pros: Bayesian methods with frequency interpretation!
   - Cons: Frequency conditional prob's make eyes glaze over.

Other Methods

There are Many other methods....

- "Default Bayes Factors"
  - Pros: Derive a proper prior dist'n based on training sample.
  - Cons: Result depends on the choice of training sample.

Decision Theory

A decision theoretic approach begins with a "Loss" Function, perhaps with c << C.

	Decision	
Truth	H_0	H_A
H_0	0	C
H_A	c	0

Derive decision rule, for example minimizing the Bayes Risk:

Bayes Risk = π_0 E(Loss|decision,H_0)+(1−π_0) E(Loss|decision,H_1)

Assessment of Decision Theory

Pros: Derives rules tailored to specific scientific goals.

Cons: Sensitive to choice of Loss Function and Prior.

(JR - Anyone up for least favorable priors and minimax rules?)

Can we abandon formal model selection all together?

- Nested Models:

  H_0: $\theta = \theta_0$ (a special case of H_A)

  H_A: $\theta = \theta_0$

  1. E.g., In cosmology, $\Omega_\kappa = 0$ vs. $\Omega_\kappa > 0$ or $\Omega_\kappa < 0$.
  2. Fit the larger model and give an interval for θ: **No Testing!**

- Does this answer the larger question?
  1. Is θ_0 a special value?
  2. Should extra weight be put on default / presumed model?
     - If not an interval may suffice.
     - If yes some sort of formal model selection may be needed.
- "Nested models are fairly common in cosmology"
  1. "flat or near flat universe is predicted by inflation"
  2. testing for infinite universe, $\Omega_\kappa \leq 0$.

Examples of model comparison questions

Imperial College London

ASTROPARTICLE

Gravitational waves detection
Do cosmic rays correlate with AGNs?
Which SUSY model is 'best'?
Is there evidence for DM modulation?
Is there a DM signal in gamma ray/ neutrino data?

COSMOLOGY

Is the Universe flat?
Does dark energy evolve?
Are there anomalies in the CMB?
Which inflationary model is 'best'?
Is there evidence for modified gravity?
Are the initial conditions adiabatic?

Many scientific questions are of the model comparison type

ASTROPHYSICS

Exoplanets detection
Is there a line in this spectrum?
Is there a source in this image?

Model Selection & Model Checking are not for the faint of heart...

- Approach Model Selection with humility.
- If possible it should simply be avoided...
- This seems possible—at least in some cases—in cosmology.

If model comparison is necessary.....

1. It is hard to justify p-values—they are simply not calibrated

   We feel that the correct interpretation of a P-value, although perhaps objective, is nearly meaningless, and that the actual meaning usually ascribed to a P-value by practitioners contains hidden and extreme bias.
   — J. Berger and M. Delampady (Stat Sci., 1987).

2. Bayes Factors are highly dependent on choice of prior.

   Bayesians address the question everyone is interested in by using assumptions no one believes, while frequntists use impeccable logic to deal with an issue on interest to anyone. — L. Lyons (via R. Trotta).

D.A. Freedman. “Some issues in the foundation of statistics.” Foundations of Science, vol. 1 (1995) pp.19–83.

My own experience suggests that neither decision-makers nor their statisticians do in fact have prior probabilities. A large part of Bayesian statistics is about what you would do if you had a prior.For the rest, statisticians make up priors that are mathematically convenient or attractive. Once used, priors become familiar; therefore, they come to be accepted as "natural" and are liable to be used again; such priors may eventually generate their own technical literature.

Similarly, a large part of objectivist statistics is about what you do if you had a model; and all of us spend enormous amounts of energy finding out what would happen if the data kept pouring in. I wish we could learn to look at the data more directly, without the fictional models and priors.

If model comparison is necessary.....

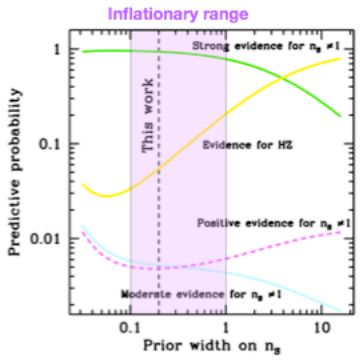
1. At least the Bayesian can clearly identify the assumptions.
2. So... I prefer Bayes Factors—but with:
   1. Careful choice of prior distribution.
   2. Clearly identified prior distribution.
   3. Comprehensive analysis of sensitivity to prior.
3. If no informative prior is available, identify classes of prior distribution that lead to one choice or the other.

As Always: Try several methods and compare results!!!

Example of reasonable sensitivity analysis

Imperial College London

- The favoured model (non-scale invariant CMB spectrum) is robust for physically reasonable changes (motivated by inflation) in the prior width



Trotta (2007)

My view of statistical models

Statistics is largely a “what if” game. For example:

•What if the observations were realizations of random variables from some probability distribution F?

•What if those random variables were independent?

•What if the probability distribution depended on parameters of interest in some specific way?

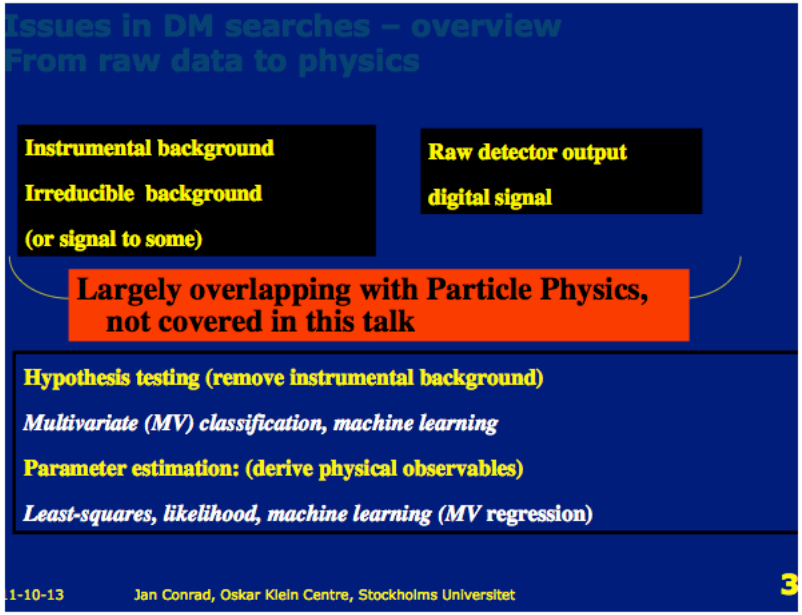
•What would I do if I had to make a decision?

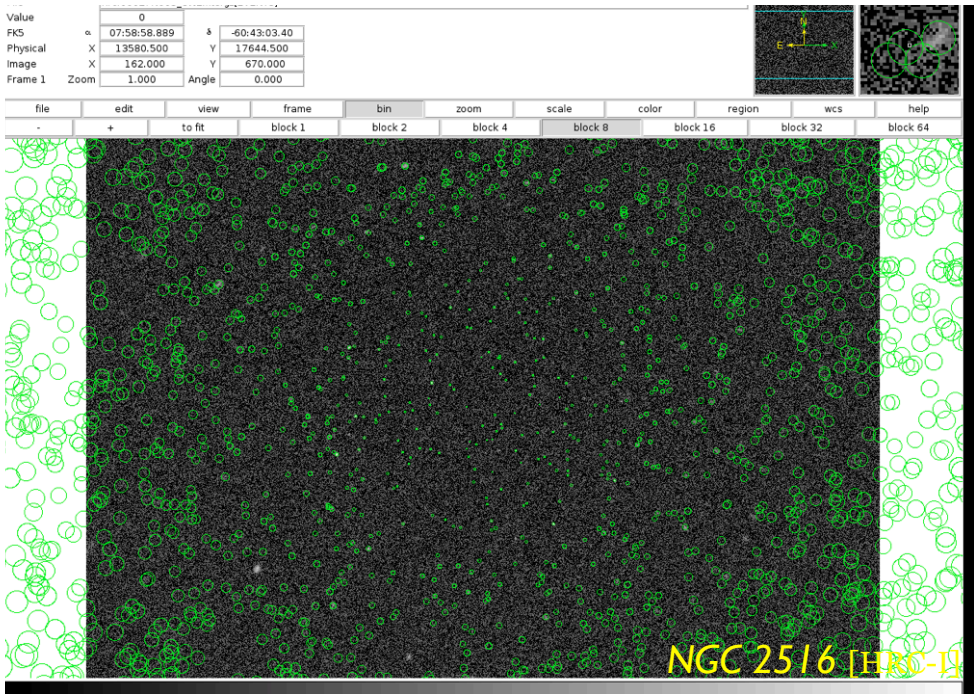
•What would be the behavior of parameter estimates if another draw were made from that probability distribution?

•What if I represented my state of knowledge as a probability distribution? For example, this one - G. How then would my state of knowledge change if I observed a realization from F?

Utility of these constructs depends on scientific context

From whence come these hypothetical iid random variables?




Value 0
FK5 α 07:58:58.889 δ -60:43:03.40
Physical X 13580.500 Y 17644.500
Image X 162.000 Y 670.000
Frame 1 Zoom 1.000 Angle 0.000
file edit view frame bin zoom scale color region wcs help
- + to fit block 1 block 2 block 4 block 8 block 16 block 32 block 64
NGC 2516 [HRC-I]

In defense of p-values

It's not so hard to understand what a p-value is: P(T > t |H), or for Neyman and Pearson the smallest significance level at which the test based on T would reject (Fisher had no use for the NP paradigm)

It's not so hard to understand what a p-value isn't: P(H|T=t). For those who are upset by this: *why are you so hung up -- get over it! Would it be such a disaster if it were impossible to find P(H|T=t)*?*

Some virtues of the humble p-value:

•It is calculated without reference to the alternative, which may be impossible to formulate in a fully specified manner. Hence, for example, the utility of permutation tests (the lady tasting tea, rank tests).

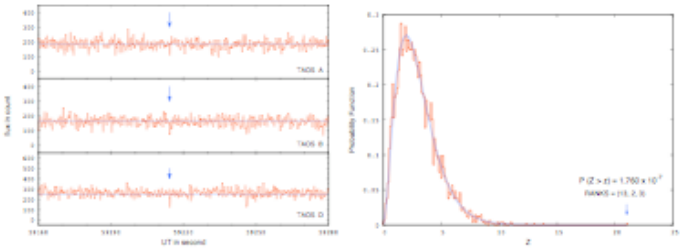
•If H holds, it is uniformly distributed

*Is this aim really so important? (Juries do reach verdicts).

Rank tests: the Taiwanese American Occultation Survey

Monitors light from hundreds of stars at 5 Hz on 3-4 telescopes, searching for occultations by KBO's.

(Lehner et al 2010, PSAP)

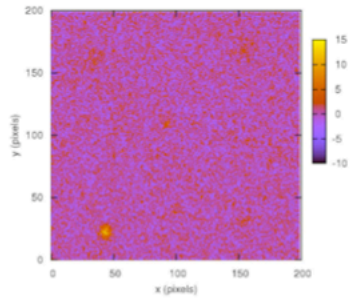


Astro example: how many sources?

Imperial College London

Feroz and Hobson (2007)

Signal + Noise

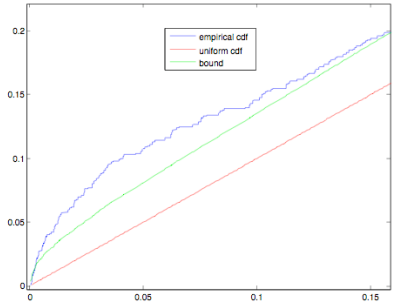


P-values are uniform under the null

P-values of test statistics from m potential sources

Donoho and Jin (2004 Annals of Statistics): is there evidence that there are any sources?
Meinshausen and Rice (2006 Annals of Statistics): lower confidence bound on the number of sources

There are situations in which it can be determined that there are sources, and lower bounds for the number can be found, but in which it can not be determined which ones they are



Another paradigm

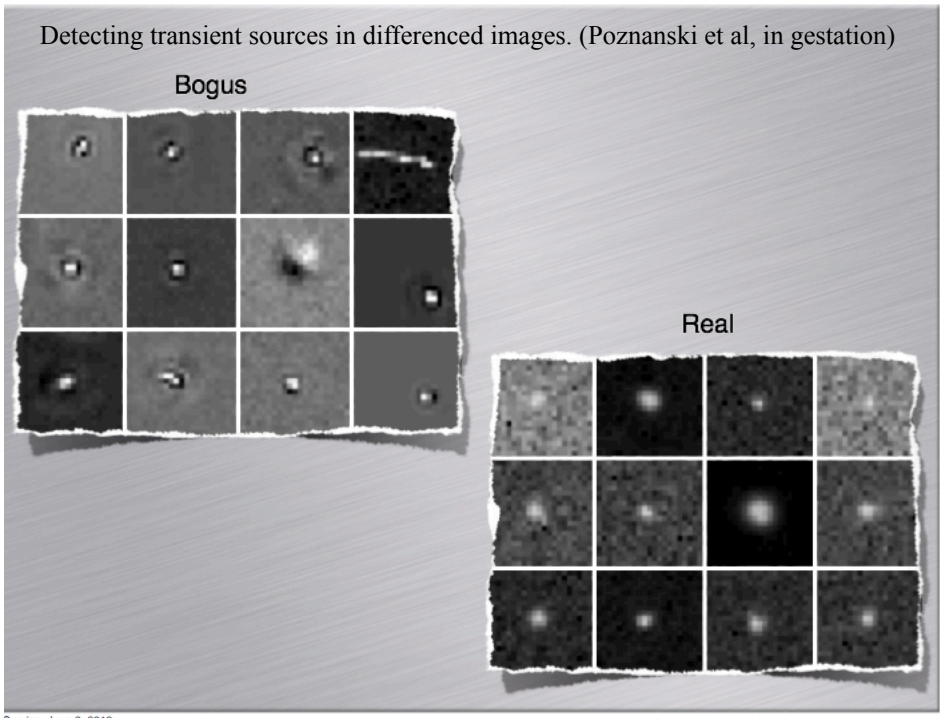
Models are not represented by analytic forms, but by complex algorithms (eg random forests, support vector machines), and are evaluated not by likelihood-based calculations, but by predictive accuracy, as measured by cross-validation, (Breiman, Statistical Science, 2001)

Djorgovski (clustering classification), Morgan

Trotta: In fundamental physics, models and parameters (and their priors) are supposed to represent (albeit in an idealized way) the real world, i.e., they are not simply useful representation of the data (as they are in other statistical problems).

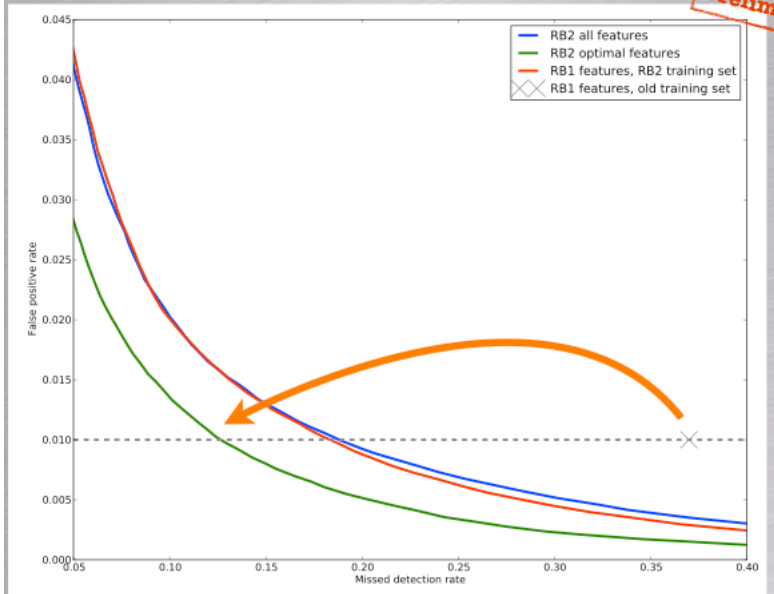
But suppose the models have hundreds of parameters (Conrad). Is there some utility for this paradigm? Can the gap between this and the traditional paradigms be bridged?

Detecting transient sources in differenced images. (Poznanski et al, in gestation)



Some Results

Preliminary



In Conclusion

The problems are fascinating. No elegant, coherent, convincing unifying statistical methodology has emerged that is universally applicable.

The scientific achievements are substantial, as are outstanding challenges

Open questions

Imperial College London

- Is Bayesian model selection the correct framework?

  "Bayesians address the question everyone is interested in by using assumptions no–one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone" Louis Lyons

  - Popperian view according to which models start off being infinitely improbable (and stay like that no matter the confirmative evidence) is untenable. Falsification only half of the story!
  - Criticisms from e.g. G. Efstathiou (arXiv:0802.3185) and Bob Cousins (Phys.Rev.Lett.101,029101, 2008)
  - Bayesian model selection matches goodness-of-fit tests for specific choices of priors

- How do we deal with Lindley's paradox? (Lindley, 1954)

  It is easy to construct cases where frequentist hypothesis testing and Bayesian model comparison disagree. How should we interpret the result?

- What do we do when Frequentist (profile likelihood) and Bayesian (marginal posterior) inferences disagree even at the level of parameter inference? What is the scientific outcome/conclusion of the measurement?
- How do we assess the completness of the set of known models in a Bayesian context?

  Bayesians maintain that it is useless to reject a model unless a better alternative is available. However, an absolute scale of model adequacy seems useful. Can Bayesian model comparison be extended to the space of unknown models? (e.g., March, RT et al, 2010)

More questions...

Imperial College London

- Is Bayesian model averaging useful?

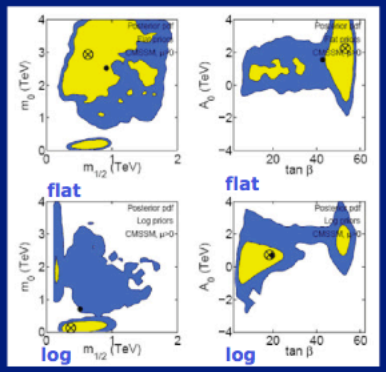
  Are model-averaged constraints useful, and if so in which context? Is the propagation of the Occam's razor effect onto parameter inference problematic?

- Is there such a thing as a "correct" prior?

  In fundamental physics, models and parameters (and their priors) are supposed to represent (albeit in an idealized way) the real world, i.e., they are not simply useful representation of the data (as they are in other statistical problems).
  One could imagine that there exist a "correct" (i.e., tied to physics) prior for parameters of our cosmological model, which could in principle be derived from fundamental theories such as string theory.

Challenges even in simplest Supersymmetric (4 parameters. CMSSM) theory

- **Prior dependence**
  - Flat vs. Log priors give signifcantly different results.
  - Remedied when including more data (LHC for CMSSM, but what happens if we have to go to 100 parameters?)

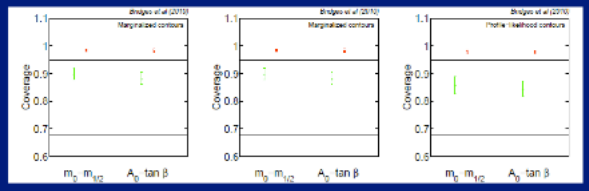


Challenges even in simplest Supersymmetric(4 parameters. CMSSM) theory

- **Frequentist properties**
  - Both over and undercoverage

Bridges+, JHEP 1103(2011) 012, LHC
Akrami+, JCAP 1107 (2011) 002

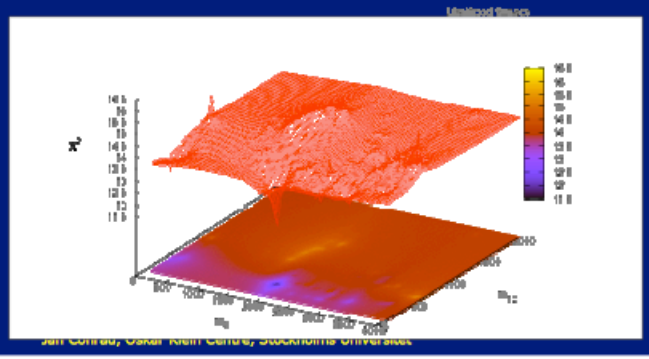
  - Bad sampling of the likelihood, boundaries on the parameters, flat prior in many dimensions (my guess) ….



Challenges even in simplest Supersymmetric(4 parameters. CMSSM) theory

- **Sensitivity to fine-tuning (especially for profile likelihood)**
  - PL picks "false" or "true" likelihood peaks
  - PL much more sensitive to adequate sampling of the likelihood
  - Can learning machines help ??

e.g. Feroz+, JHEP 1106:042,2011



In Conclusion

Much remains to be done. What will be the content and themes of such a conference 20 years from now? (I would love to attend)

The glass is t!

Special thanks to Louis and Jeff!