# Overview of Bayesian methods for multiwavelength gamma-ray astronomy

Tom Loredo
Cornell Center for Astrophysics and Planetary Science,
Carl Sagan Institute,
& Dept. of Statistics and Data Science, Cornell U.
http://hosting.astro.cornell.edu/~loredo/

PhyStat-Gamma — 27–30 Sep 2022

Includes work with Tamás Budavári & Robert Wolpert
Grant support from NSF Astronomy & Statistics (AST-1814840, DMS-2015386)

Bayesian data analysis gets its name from ***Bayes's theorem***:

$$p(\theta|D) = \frac{p(\theta)\, p(D|\theta)}{p(D)} = \frac{p(\theta)\, \mathcal{L}(\theta)}{p(D)}$$

So it's basically about ***modulating maximum likelihood with priors***…

Bayesian data analysis gets its name from ***Bayes's theorem***:

$$p(\theta|D) = \frac{p(\theta)\, p(D|\theta)}{p(D)}$$
$$= \frac{p(\theta)\, \mathcal{L}(\theta)}{p(D)}$$

So it's basically about ***modulating maximum likelihood with priors***…

# Bayesian inference in a nutshell

*Probability as generalized logic*

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

*Bayes's theorem*

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis $\propto$ ability of hypothesis to *predict* the observed data

*Law of total probability*

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$$

The support for a *composite* hypothesis must account for all the ways it could be true, via *marginalization*

# On the key role of marginalization

Bayesian statistics uses all of probability theory, not just Bayes's theorem, and not even primarily Bayes's theorem. . . . Perhaps the most important theorem for doing Bayesian calculations is the *law of total probability* (LTP) that relates marginal probabilities to joint and conditional probabilities. . . . Arguably, if this approach to inference is to be named for a theorem, "total probability inference" would be a more appropriate appellation than "Bayesian statistics." It is probably too late to change the name. But it is not too late to change the emphasis.

— Loredo (2013)

The key distinguishing property of a Bayesian approach is marginalization instead of optimization, not the prior, or Bayes rule. . . . Broadly speaking, what makes Bayesian approaches distinctive is a posterior weighted marginalization over parameters. . . . Moreover, basic probability theory indicates that marginalization is desirable.

— Wilson (2020), Wilson & Izmailov (2020)

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**
2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment
3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. **Closing thoughts**

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**

2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment

3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem

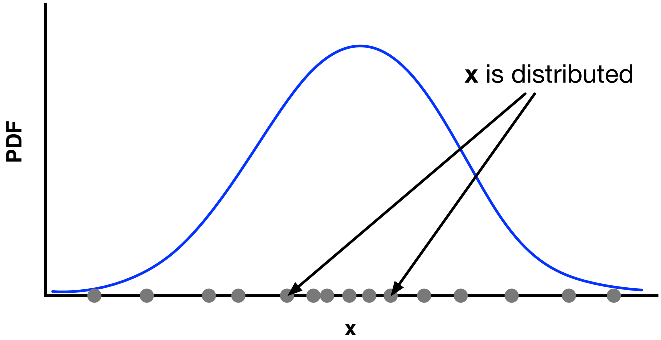
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning

5. **Closing thoughts**

# Interpreting PDFs

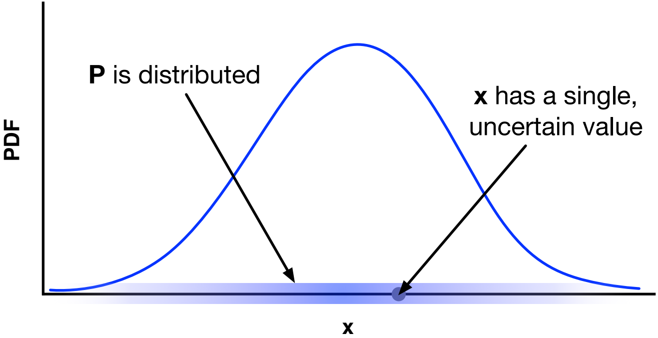
*Frequentist*

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. $p(x)$ describes how the *values of x* would be distributed among *infinitely many trials*:



## *Bayesian*

Probability *quantifies uncertainty* in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values $x$ might have taken in *the single case before us*:




This interpretation holds whether $x$ labels data or hypotheses.

# Probability & frequency in IID settings

Consider a setting where we assign the same probability to many independent outcomes (flips of a coin, rolls of a die, searches for an Earth around a G dwarf...):

- If the probability is high, we expect the outcomes to occur frequently
- If the probability is low, we expect the outcomes to occur rarely

In IID repeated trial settings, it seems there should be a relationship between single-trial probability and multiple-trial (relative) frequency

Early probabilists—Bernoulli, Bayes, Laplace, etc.—interpreted probability in a Bayesian way, but sought to derive connections to frequency in replication settings

## *Frequency from probability*

Bernoulli's (weak) law of large numbers: In repeated IID trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \to \alpha \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $P(\text{success}|\ldots)$ does not change from sample to sample, it may be interpreted as the expected relative frequency

## *Probability from frequency*

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\to$ First use of Bayes's theorem:

Probability for success in next trial of IID sequence:

$$\mathrm{E}(\alpha) \to \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $P(\text{success}|\ldots)$ does not change from sample to sample, it may be estimated using relative frequency data

There is nothing more Bayesian than to be interested in the role of frequency in inference. But probability is not identified with frequency—the former is an abstract measure of argument strength; the latter is (potentially) observable.

Probability as a measure of strength of a data-based argument is separate from *calibration*—quantifying long-run performance of a procedure used in a replication setting. When calibration properties are of interest, they need to be separately computed.

# Frequentist vs. Bayesian statements

"The data $D_{\text{obs}}$ support hypothesis $H$ . . . "

## *Frequentist assessment*

*"H was selected with a procedure that's right 95% of the time over a set $\{D_{hyp}\}$ that includes $D_{obs}$."*

Probabilities are properties of *procedures*, not of particular results. Guaranteed long-run performance is the *sine qua non*.

## *Bayesian assessment*

*"The strength of the chain of reasoning from the model and $D_{obs}$ to H is 0.95, on a scale where 1= certainty."*

Probabilities are associated with arguments based on *specific, observed data*.

Long-run performance must be separately evaluated (and is typically good by frequentist criteria in parametric settings).

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**

2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment

3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem

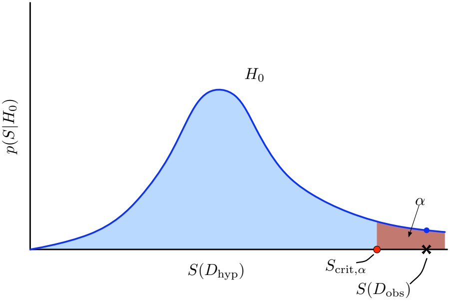
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning

5. **Closing thoughts**

# Null hypothesis significance testing (NHST)

## *Neyman-Pearson testing*

- Specify simple null hypothesis $H_0$ such that rejecting it implies an interesting effect is present
- Devise statistic $S(D)$ measuring departure from null predictions
- Divide sample space into probable and improbable parts (for $H_0$); $p(\text{improbable}|H_0) = \alpha$ (Type I error rate), with $\alpha$ specified a priori
- If $S(D_{\rm obs})$ lies in improbable region, reject $H_0$; otherwise accept it
- Report: "$H_0$ was rejected (or not) with a procedure with false-alarm frequency $\alpha$"



Neyman and Pearson devised this approach guided by Neyman's *frequentist principle*:

> *In repeated practical use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error. (Berger 2003)*

A *confidence region* is an example of a familiar procedure satisfying the frequentist principle

They insisted that one also specify an alternative, and find the error rate for falsely rejecting it (Type II error)

For *simple* null and alternative hypotheses, the optimal $S(D)$ is the (log) *likelihood ratio*. For composite hypotheses, the *maximum* likelihood ratio is popular (not necessarily optimal).

## Fisher's p-value testing

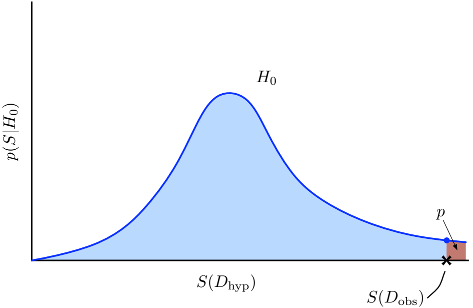
Fisher (and others) felt reporting a rejection frequency of $\alpha$ no matter where $S(D_{\rm obs})$ lies in the rejection region does not accurately communicate the strength of evidence against $H_0$

He advocated reporting the *p-value*:

$$p = P(S(D) > S(D_{\rm obs})|H_0)$$

Smaller $p$-values indicate stronger evidence against $H_0$

Astronomers call this the *significance level* or the *false-alarm probability* (FAP). Statisticians don't—for good reason!



## *ASA 2016 statement on statistical significance and p-values*

- ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**
- **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**
- **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**
- **. . .**

# $p$-values and the FAP fallacy

From the exoplanets literature:

> "...the false alarm probability for this signal is rather high at a few percent."

> "This signal has a false alarm probability of $< 4$ % and is consistent with a planet of minimum mass 2.2 $M_{\odot}$..."

> "This detection has a signal-to-noise ratio of 4.1 with an empirically estimated upper limit on false alarm probability of 1.0%."

> "We find a false-alarm probability $< 10^{-4}$ that the RV oscillations attributed to CoRoT-7b and CoRoT-7c are spurious effects of noise and activity."

*All of these statements incorrectly describe the weight of evidence for a planet, and almost certainly greatly exaggerate the weight of the evidence*

Similar misuses of $p$-values appear throughout astronomy, including in Nobel prize winning work discovering the accelerated expansion of the universe, and the first gravitational wave sources.

*We can* (**must**) *do better!*

# What's wrong?

"This *signal, with* $S(D_{obs}) = X$*, has a* FA*P of* $p \ldots$"

$$p = P(\{D_{\rm hyp} : S(D_{\rm hyp}) \geq S(D_{\rm obs})\} | H_0)$$

*Probability* ... given $\boldsymbol{H}_0$

$p$ is computed assuming that $H_0$ *always operates*

*Every* alarm is false (i.e., with FAP$= 1$) in this "world"

For any signal to have FAP$\neq 1$, alternatives to the null must sometimes act; the FAP will depend on how often they do, and what they are

*Probability*... *including* worse *departures from null predictions*

$p$ is not a property of *this* signal; it's the size of the *ensemble* of possible null-generated datasets with $S(D) > S(D_{\rm obs})$

$D_{\rm obs}$ bounds this set on the *weakest* side

# What a $p$-value really means

In the voice of Don LaFontaine or Lake Bell:

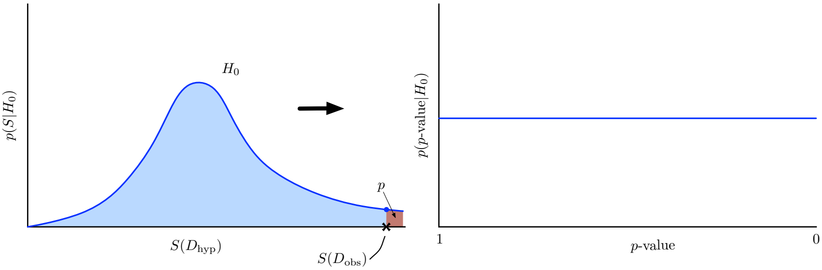
In a world. . . *with absolutely no sources,*
*with a threshold set so we wrongly claim to*
*detect sources* $100 \times p\%$ *of the time,*
*this data would wrongly be considered a*
*detection—and it would be the data*
*providing the* weakest *evidence for a source*
*in that world.*

Who wants to say *that*?! Whence "$p$-value," a measure of "surprisingness" under the null.

# $p$'s one intuitive property

Under the null, the fraction of time $p > X$ is. . . $X$

Think of $p$ as an alternative test statistic—a nonlinear mapping of $S(D)$ that has a *uniform distribution under the null*




$p$ is a surprise-ordered relabeling of the data, with a $U(0, 1)$ PDF, and a linearly rising CDF

# Surprise isn't enough

The rarity of data "like" $D_{\text{obs}}$ under $H_0$ is evidence against $H_0$ only if *plausible alternatives* make $D_{\text{obs}}$ *less* surprising

Expand the "world" of the $p$-value calculation:

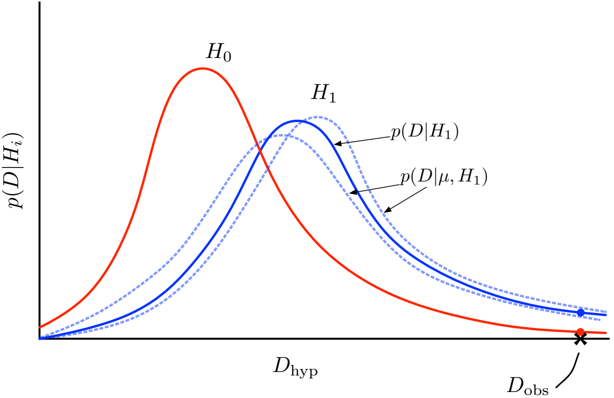
- Let an alternative, $H_1$, sometimes operate, with probability $\pi_1$ (with null prevalence $\pi_0 = 1 - \pi_1$)
- Compare the rates for getting the observed $p$-value under $H_0$ and $H_1$ (*not* "observed or smaller $p$-value")
- Equivalently: Compare the rates for getting $S(D_{\text{obs}})$ under $H_0$ and $H_1$

This *conditional frequentist* approach can produce genuine FAPs; it uses $P(S(D_{\text{obs}})|H_i)$, not tail areas

If the hypotheses are simple and $S(\cdot)$ is sufficient, this corresponds to using Bayes factors

For *composite* hypotheses ($H_1$ here), the *marginal likelihood* accounts for parameter uncertainty that is ignored by $p$-values (which typically set parameters equal to best-fit values):

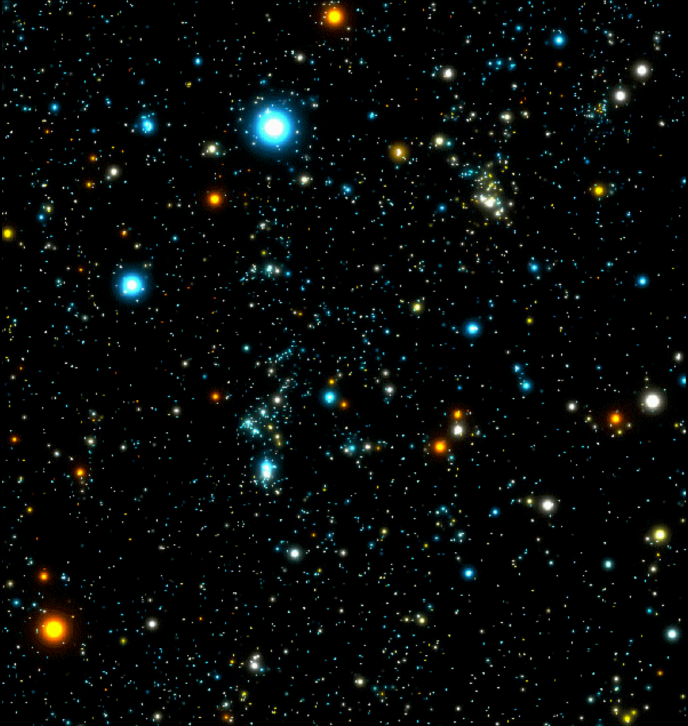
$$p(D|H_i) = \int d\theta_i \; p(\theta_i) \; p(D|\theta_i, H_i)$$

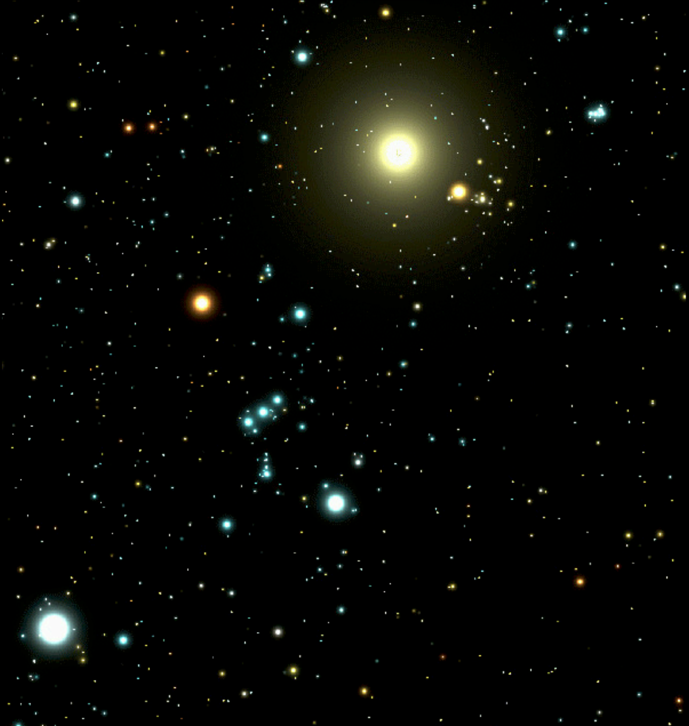


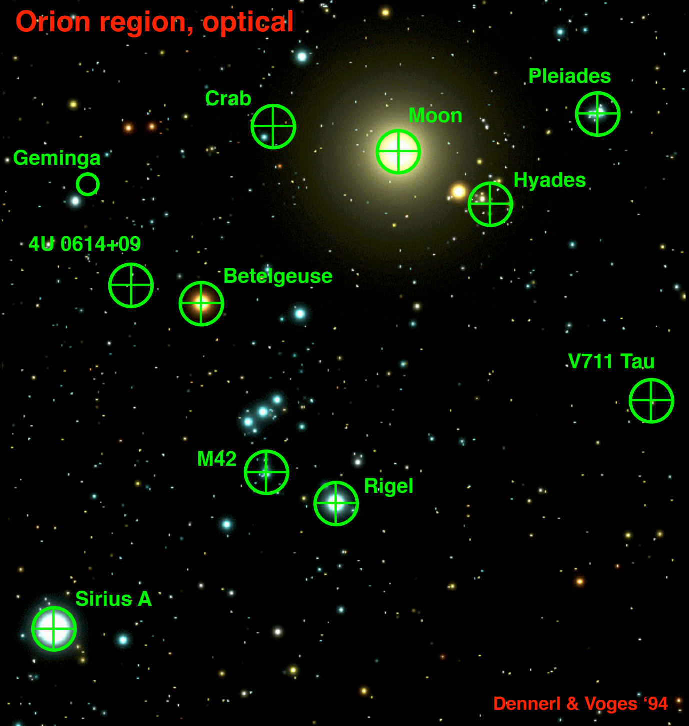

Also, the marginal likelihood uses *all* of the data, not just the value of a test statistic: in general $p(D|H_i) \neq p(S(D)|H_i)$

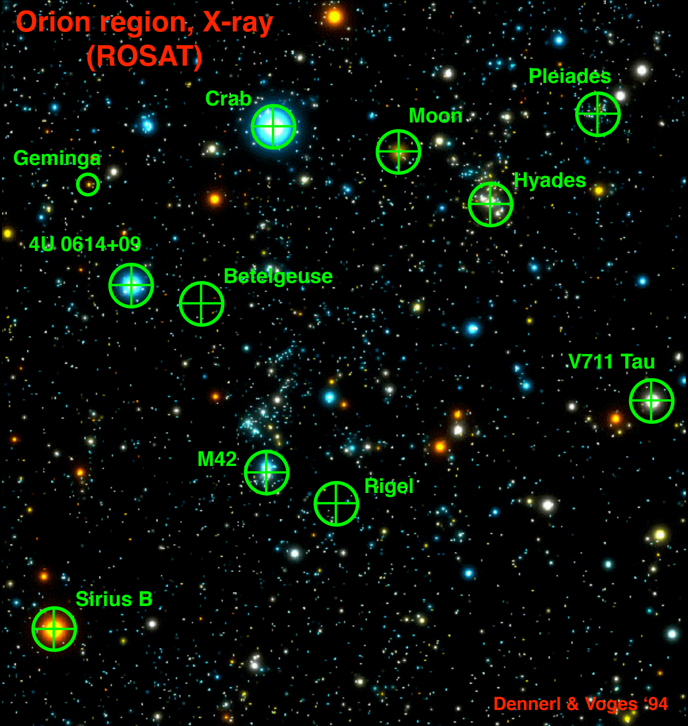
# Agenda

1. Understanding Bayesian vs. frequentist inferences
2. **Cross-identification: $p$-values and alternatives**
   - $p$-values are not FAPs
   - Spatio-temporal coincidence assessment
3. Nuisance parameters
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. Priors: More than penalties
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. Closing thoughts



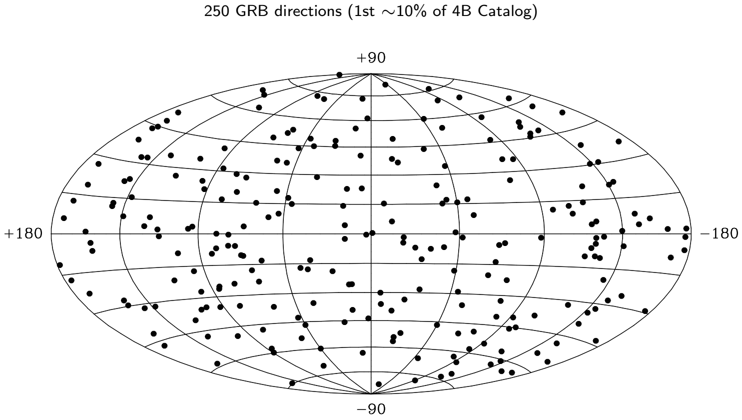



Orion region, optical
Crab
Moon
Pleiades
Geminga
Hyades
4U 0614+09
Betelgeuse
V711 Tau
M42
Rigel
Sirius A
Dennerl & Voges '94

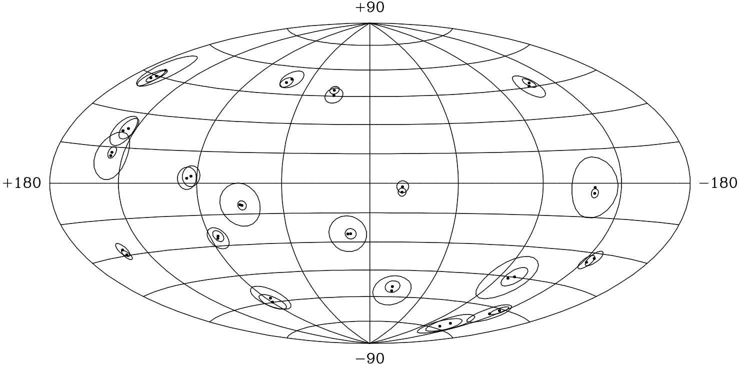


Orion region, X-ray
(ROSAT)
Crab
Moon
Pleiades
Geminga
Hyades
4U 0614+09
Betelgeuse
V711 Tau
M42
Rigel
Sirius B
Dennerl & Voges '94

# Do GRB sources repeat?



Subset with neighbor within 3° (39)




- 485 out of the 1st 1000 are this close
- 2280 out of the total 2702 are this close

Are there too many close pairs, presuming independence?

Various statistics (nearest neighbor, angular correlation) gave $p \sim 0.001$ to 0.01 assuming independence, isotropy—some also using antipodal correlations!

# Coincidence assessment in astronomy

We observe the same region of the sky through various "windows:"

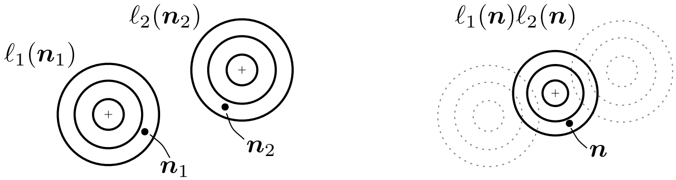
- **Multiwavelength astronomy** — Different regions of the electromagnetic spectrum
- **Multi-messenger astronomy** — Different types of radiation
  - Electromagnetic
  - Neutrinos
  - Cosmic rays
  - Gravitational radiation
- **Time-domain astronomy** — Different periods of time

Fundamental questions:

- Are objects/events associated ("counterparts")? $\rightarrow$ Pool information to better characterize underlying phenomenon
- Are objects/events distinct? $\rightarrow$ Discovery!

Fundamental difficulties: Uncertainties in directions and other observables, measures of closeness, number of candidate matches. . .

# Bayesian Coincidence Assessment


$$
\begin{aligned}
p(d_1, d_2|H_0) &= \int d\boldsymbol{n}_1 \; p(\boldsymbol{n}_1|H_0)\, \ell_1(\boldsymbol{n}_1) \\
&\times \int d\boldsymbol{n}_2 \cdots
\end{aligned}
$$

$$
p(d_1, d_2|H_1) = \int d\boldsymbol{n} \; p(\boldsymbol{n}|H_1)\, \ell_1(\boldsymbol{n})\, \ell_2(\boldsymbol{n})
$$

# Multiplet Bayes Factors

Analytical result using Fisher dist'n (isotropic prior):

$$B_{ij} = \frac{\kappa_i \kappa_j}{(4\pi)^2 \sinh(\kappa_i) \sinh(\kappa_j)} \frac{\sinh(R)}{R},$$

$$R^2 = \kappa_i^2 + \kappa_j^2 + 2\kappa_i \kappa_j \cos(\boldsymbol{n}_i \cdot \boldsymbol{n}_j)$$

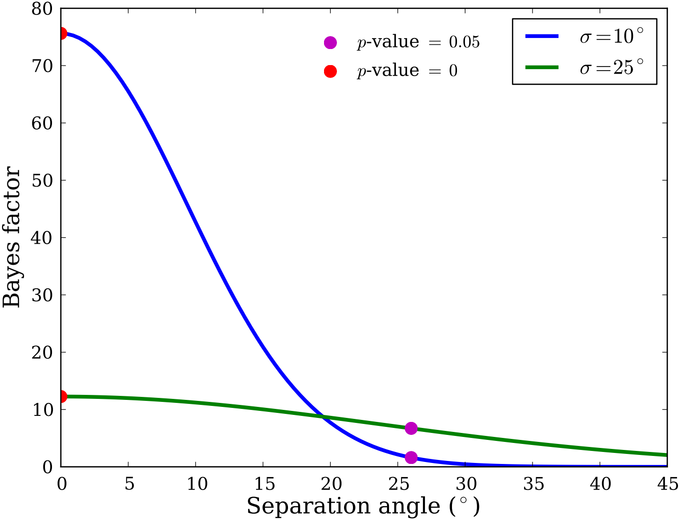
Generalization to multiplet of size $k$:

$$B_{ij\ldots l} = \frac{1}{(4\pi)^k} \frac{\sinh(R)}{R} \left( \frac{\kappa_i}{\sinh(\kappa_i)} \right) \left( \frac{\kappa_j}{\sinh(\kappa_j)} \right) \times \cdots \times \left( \frac{\kappa_l}{\sinh(\kappa_l)} \right)$$

$$R^2 = (\kappa_i \boldsymbol{n}_i + \kappa_j \boldsymbol{n}_j + \cdots + \kappa_l \boldsymbol{n}_l)^2$$

# Doublet Bayes factor behavior

vs. nearest-neighbor $p$-value



# Challenge: Large hypothesis spaces

For $N = 2$ events, there was a single coincidence hypothesis, $H_1$

For $N = 3$ events:

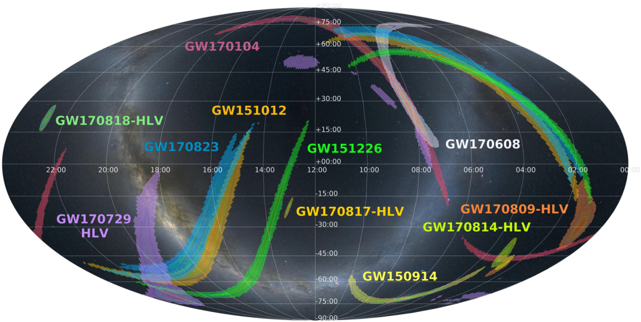
- Three doublets: $1 + 2$, $1 + 3$, or $2 + 3$
- One triplet

The number of alternatives (partitions, $\varpi$) grows combinatorially!

- *Model building:* Assign sensible priors to partitions
- *Computation:* Find & sum over important partitions

# Challenge: Large, complex localizations

LIGO+VIRGO GW source localizations




See Friday's talks by Budavári and Salvato for more...

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**
2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment
3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. **Closing thoughts**

# Nuisance parameters and marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

*Example*

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal $s$ and a background $b$.

We have additional data just about $b$.

What do the data tell us about $s$?

# Marginal posterior distribution

To summarize implications for $s$, accounting for $b$ uncertainty, *marginalize*:

$$\begin{aligned} p(s|D,M) &= \int db\, p(s,b|D,M) \\ &\propto p(s|M) \int db\, p(b|s,M)\, \mathcal{L}(s,b) \\ &= p(s|M)\mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function for* $s$:

$$\mathcal{L}_m(s) \equiv \int db\, p(b|s)\, \mathcal{L}(s,b)$$

Maximum likelihood suggests instead computing the *profile likelihood*:

$$\mathcal{L}_p(s) \equiv \mathcal{L}(s,\hat{b}_s), \qquad \hat{b}_s = \text{best } b \text{ given } s$$

# Marginalization vs. profiling

*For insight:* Suppose the prior is broad compared to the likelihood $\rightarrow$ for a fixed $s$, we can accurately estimate $b$ with max likelihood $\hat{b}_s$, with small uncertainty $\delta b_s$.

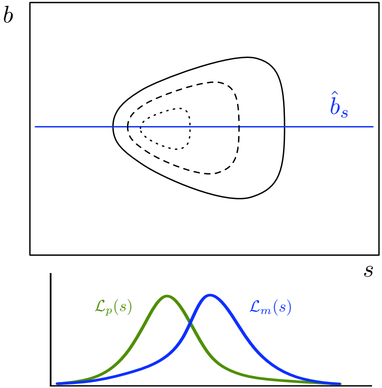
$$\begin{aligned}
\mathcal{L}_m(s) &\equiv \int db\; p(b|s)\,\mathcal{L}(s,b) \\
&\approx p(\hat{b}_s|s)\,\mathcal{L}(s, \underbrace{\hat{b}_s}_{\text{best } b \text{ given } s})\; \underbrace{\delta b_s}_{b \text{ uncertainty given } s}
\end{aligned}$$

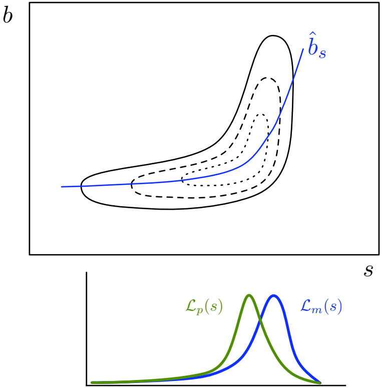
Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\quad \sigma_s^2 = \sigma_r^2 + \sigma_b^2$, and $\delta b_s$ is *const.*

Background *subtraction* is a special case of background *marginalization*.

Flared/skewed/bannana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ *differ*




General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$.
Otherwise, they will likely *differ*.

In *measurement error problems* the difference can have dramatic consequences (due to proliferation of latent parameters)

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**
2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment
3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. **Closing thoughts**

# The on/off problem for Poisson counting data

*Basic problem*

- Look off-source; unknown background rate $b$
  Count $N_{\rm off}$ photons in interval $T_{\rm off}$
- Look on-source; rate is $r = s + b$ with unknown signal $s$
  Count $N_{\rm on}$ photons in interval $T_{\rm on}$
- Infer $s$

*Conventional solution*

$$\begin{aligned}
\hat{b} &= N_{\rm off}/T_{\rm off}; & \sigma_b &= \sqrt{N_{\rm off}}/T_{\rm off} \\
\hat{r} &= N_{\rm on}/T_{\rm on}; & \sigma_r &= \sqrt{N_{\rm on}}/T_{\rm on} \\
\hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2}
\end{aligned}$$

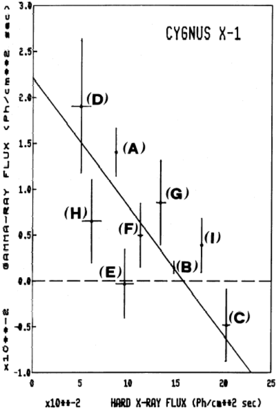
But $\hat{s}$ can be *negative*!

Multiple ad hoc fixes (ca. 1989) all failed in some regime

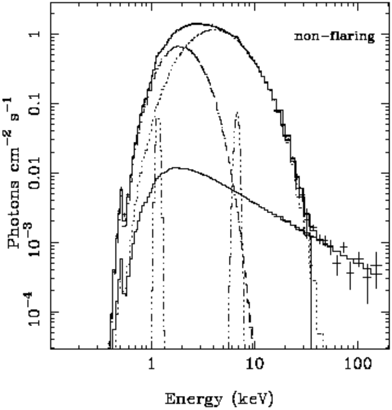
# Examples

Spectra of X-ray, $\gamma$-ray sources

Bassani et al. 1989




Di Salvo et al. 2001




Sample sizes are never large. . . once $N$ is "large enough," you can start subdividing the data to learn more. . . . $N$ is never enough because if it were "enough" you'd already be on to the next problem for which you need more data. — Andrew Gelman (blog entry, 31 July 2005)

# Bayesian solution to on/off problem

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

$$\mathcal{L}(s,b) = \frac{(bT_{\rm off})^{N_{\rm off}}}{N_{\rm off}!}e^{-bT_{\rm off}} \times \frac{[(s+b)T_{\rm on}]^{N_{\rm on}}}{N_{\rm on}!}e^{-(s+b)T_{\rm on}}$$

Adopting flat priors for $(s, b)$, the joint posterior is

$$p(s,b|N_{\rm on},N_{\rm off},\mathcal{C}) \quad \propto \quad (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}$$

Note if $b = 0$, the (normalized) posterior distribution is a gamma distribution,

$$p(s,b=0|N_{\rm on},N_{\rm off},\mathcal{C}) = \frac{T_{\rm on}(sT_{\rm on})^{N_{\rm on}}}{N_{\rm on}!}e^{-sT_{\rm on}}$$

Now marginalize over $b$;

$$
\begin{aligned}
p(s|N_{\rm on}, N_{\rm off}, \mathcal{C}) &= \int db\; p(s, b \mid N_{\rm on}, \mathcal{C}) \\
&\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}
\end{aligned}
$$

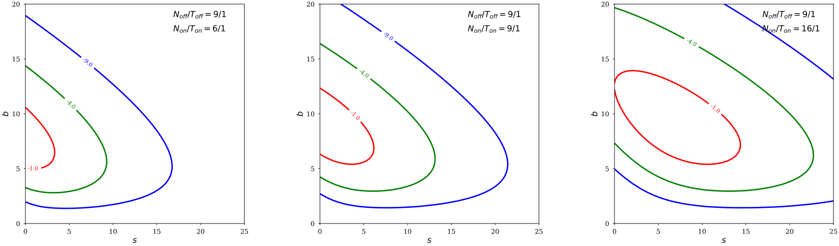
Expand $(s+b)^{N_{\rm on}}$ and do the resulting $\Gamma$ integrals:

$$
\begin{aligned}
p(s|N_{\rm on}, N_{\rm off}, \mathcal{C}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\
C_i &\propto \left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on} + N_{\rm off} - i)!}{(N_{\rm on} - i)!}
\end{aligned}
$$

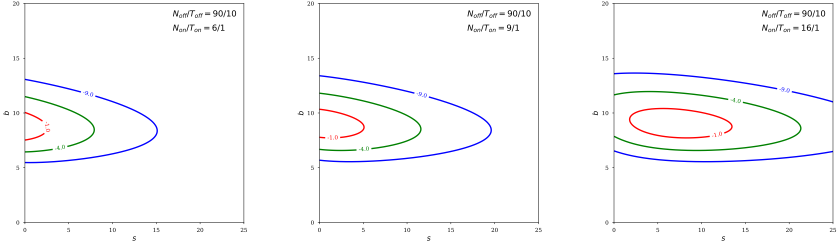
Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

# Example on/off joint PDFs

## $T_{\text{on}} = T_{\text{off}} = 1$

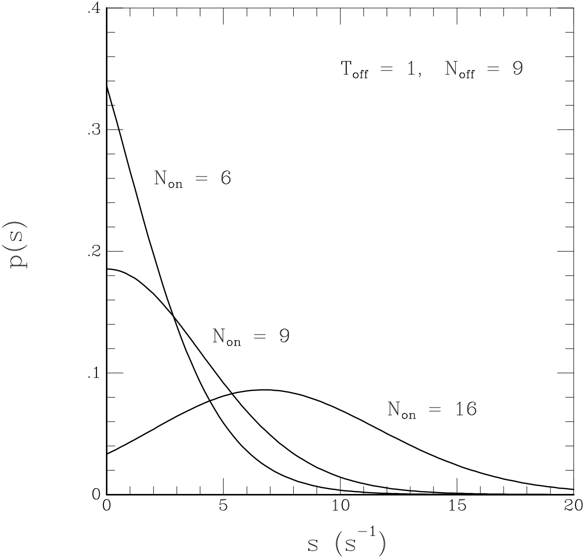



## $T_{\text{on}} = 1,\ T_{\text{off}} = 10$



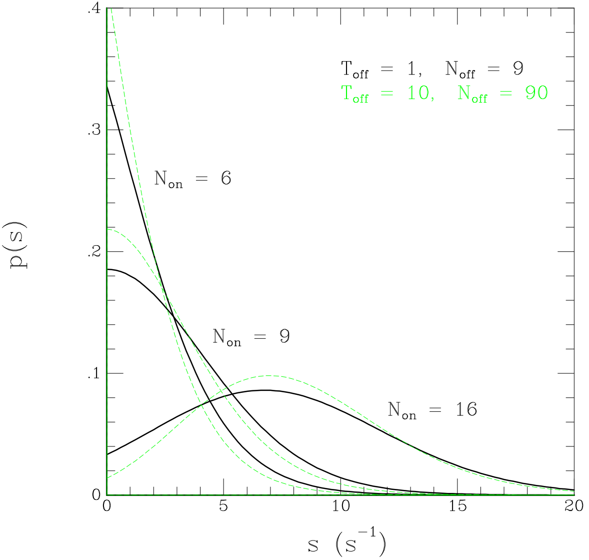
# Example on/off marginal PDFs—Short integrations

$T_{\mathrm{on}} = T_{\mathrm{off}} = 1$



# Example on/off marginal PDFs—Long background integrations

$T_{\rm on} = 1$, $T_{\rm off} = 10$



# Credible vs. confidence regions

Bayesian credible regions are *not* frequentist confidence regions:

- Credible regions guarantee exact *average* coverage, averaging over true rates wrt the prior
- Confidence regions guarantee *minimum* coverage—infimum over all possible true rates (conservative)
- Parametric model credible regions using flat priors are approximate confidence regions, with coverage error $O(1/\sqrt{N})$. Using a *reference prior* improves this. Sometimes there is a "probability matching prior" that makes it exact.

# Testing Coverage

### Frequentist MC

```
CL = 0.9    //confidence level
N = 1000  //experiments
~~mu1 = mu2 = 0~~
for( every possible true mu0 ){
   //test coverage for this mu0
   coverage = 0

   //simulate experiments
   for(i=0;i<N;i=i+1){
      x0 ~ p(x|mu0)
      FreqLimit(CL,x0,mu1,mu2)
      if( mu1<=mu0<=mu2 )
        coverage = coverage + 1
   }
  coverage = coverage/N
  //~~coverage should equal CL~~
}
```

conditional coverage

min. cond'l coverage = CL

### Bayesian MC

```
CL = 0.9    //~~confidence~~ credible level
N = 1000  //attempts
~~mu1 = mu2 = 0~~
for( ~~every possible x0~~ ){
   //test coverage for this x0
   coverage = 0
   BayesLimit(CL,x0,mu1,mu2)
  //sample posterior
  for(i=0;i<N;i=i+1){
     mu0 ~ p(mu|x0)p(mu)/p(x0)

     if( mu1<=mu0<=mu2 )
       coverage = coverage + 1
  }
  coverage = coverage/N
  //~~coverage should equal CL~~
}
```

set of x0 vals drawn from prior predictive

conditional coverage

avg. cond'l coverage = CL

more common in this order:

$$\mu \sim p(\mu)$$
$$x_0 \sim p(x_0|\mu)$$

# Agenda

1. **Understanding Bayesian vs. frequentist inferences**
2. **Cross-identification: *p*-values and alternatives**
   - *p*-values are not FAPs
   - Spatio-temporal coincidence assessment
3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. **Closing thoughts**

# Roles of the prior

*Prior has two roles*

- Modulate the likelihood to incorporate relevant prior information
- Convert likelihood from "intensity" to "measure" $\rightarrow$ enable accounting for *size of parameter space*

*Physical analogy*

$$\begin{aligned} \text{Heat} \quad Q &= \int d\vec{r}\, c_v(\vec{r}) T(\vec{r}) \\ \text{Probability} \quad P &\propto \int d\theta\; p(\theta)\mathcal{L}(\theta) \end{aligned}$$

Maximum likelihood focuses on the "hottest" parameters.
Bayes focuses on the parameters with the most "heat."

A high-$T$ region may contain little heat if its $c_v$ is low or if its volume is small.

A high-$\mathcal{L}$ region may contain little probability if its prior is low or if its volume is small.

Frequentist *penalized maximum likelihood* methods multiply the likelihood by a penalty function, $r(\theta)$ (e.g., a regularizer):

$$\arg\max r(\theta)\,\mathcal{L}(\theta)$$

The penalty function shifts the location of the maximum

This *looks* like a prior, but because Bayesian calculations integrate over $\theta$, the prior can do much more than shift the location of the mode.

Relevant ideas:

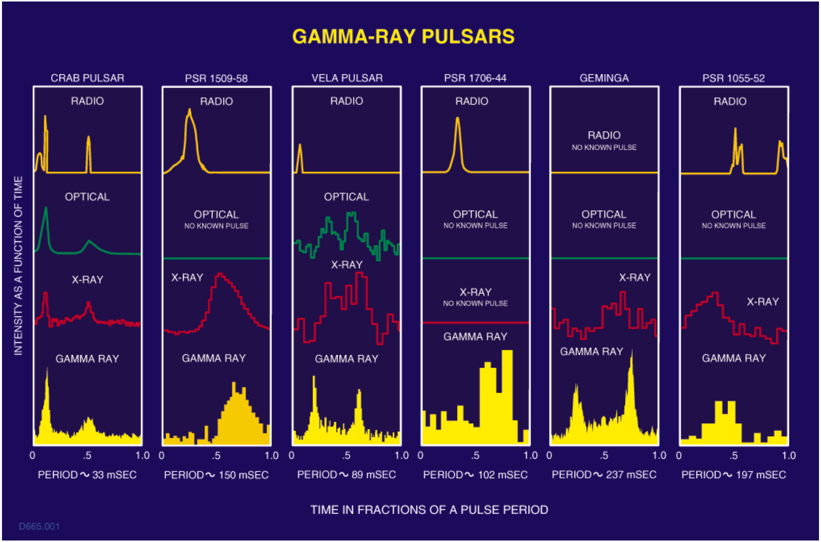
- Curse of dimensionality (hi-D geometry)
- Concentration of measure (measure theory)
- Typical sets (information theory)

These all indicate that, in hi-D spaces with a kind of symmetry (product spaces), volume (probability!) can accumulate in unanticipated ways

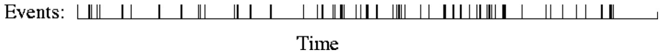
# Agenda


1. **Understanding Bayesian vs. frequentist inferences**
2. **Cross-identification: $p$-values and alternatives**
   - $p$-values are not FAPs
   - Spatio-temporal coincidence assessment
3. **Nuisance parameters**
   - Marginalizing vs. profiling
   - Poisson on/off problem
4. **Priors: More than penalties**
   - Impacts of priors
   - TTE data: Period searching with adaptive binning
5. **Closing thoughts**

# Pulsars from Radio to Gamma Rays



# Pulsar Searching: Entry-Level Nonparametrics




X-ray/$\gamma$-ray arrival time series, $N =$ dozens to millions

Goal: Detect periodicity

Rate = avg. rate $A \times$ periodic shape $\rho$ (params $\mathcal{S}$)

$$r(t) = A\rho(\omega t - \phi)$$

Inhomogeneous point process likelihood (for $T \gg$ period)

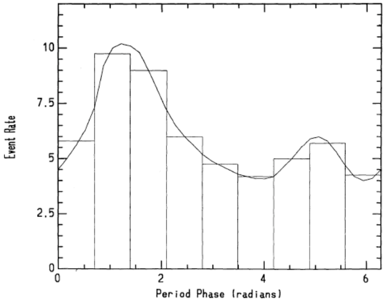
$$\mathcal{L}(A, \omega, \phi, \mathcal{S}) = \left[A^N e^{-AT}\right] \prod_i \rho(\omega t_i - \phi)$$

Marginal likelihood for $\omega$, $\phi$, $\mathcal{S}$

$$\mathcal{L}(\omega, \phi, \mathcal{S}) = \prod_i \rho(\omega t_i - \phi)$$

Various models implemented . . .

## *Piecewise-constant model (Gregory & Loredo 1992)*

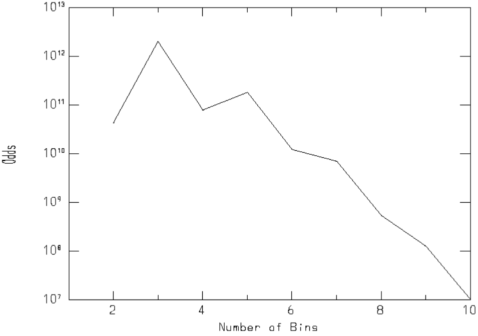


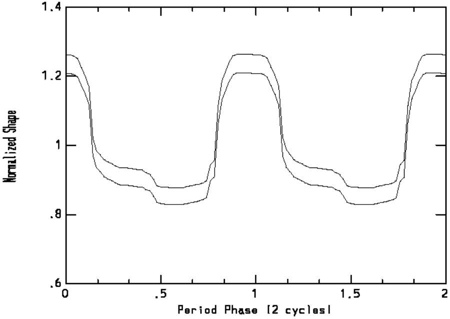

- Take $\rho(\theta) = f_k$ in $M$ phase bins
- Use *flat prior* on $f_k$ over simplex $\sum_k f_k = 1$
- *Analytically* marginalize over shape $\rightarrow$

$$p \propto \frac{(M-1)!}{(N+M-1)!} \left[ \frac{n_1!\; n_2!\; \ldots n_M!}{N!} \right] \qquad \text{entropy!}$$

- Numerically marginalize over phase, frequency
- Model-average over $M$ to predict light curve

X-Ray Pulsar PSR 0540-693 (Gregory & TL 1996)
3300 events over $10^5$ s, many gaps, Rayleigh test fails



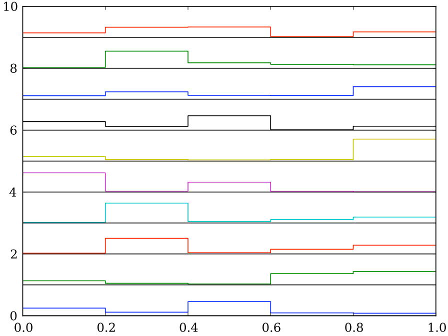




David MacKay & John Skilling observed that the odds falls surprisingly quickly with increasing # of bins. . .

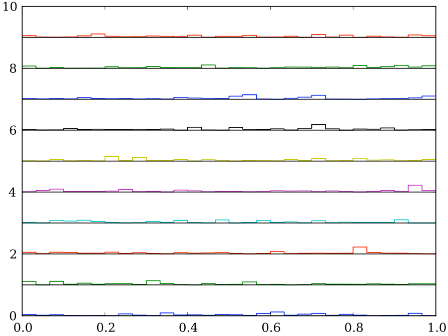
# Can We Do Better?

The flat stepwise shape prior is. . . *flat!*

Flat prior, $m$=5




Flat prior, $m$=30

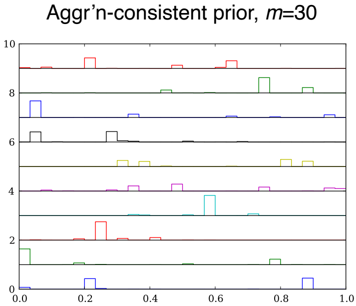


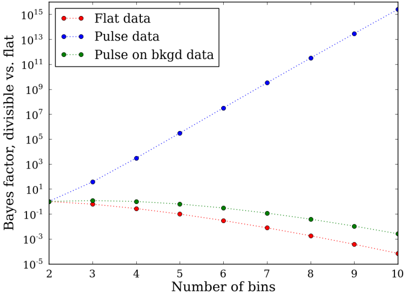
- Adopt symmetric Dirichlet prior:

$$p(\boldsymbol{f}) = \delta\left(1 - \sum_k f_k\right) \prod_k f_k^{\alpha-1}$$

- Cross-model consistency requirement:
  4-bin prior should become 2-bin prior when binned up, etc.

- Aggregation consistency $\rightarrow \alpha = C/M$




*Still work to do. . .*

# Theme: Parameter space volume

*Bayesian calculations sum/integrate over parameter/hypothesis space!*

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Marginal likelihoods have parameter space volume factors that can penalize models for unncecessary complexity
- Prediction, uncertainty propagation, model averaging...

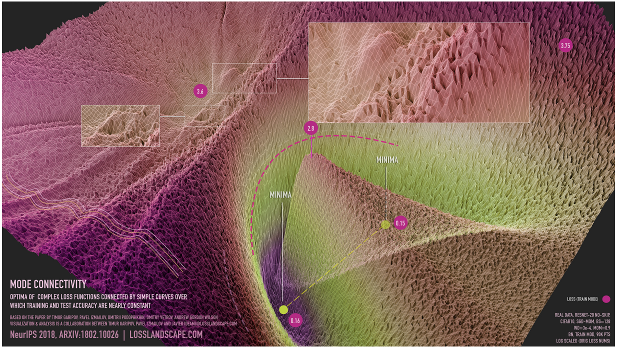
Many virtues of Bayesian methods can be attributed to this accounting for the "size" of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added "by hand"—ignoring Fisher!).

# A frontier: Bayesian neural nets

Neural nets are large, composite models with thousands to millions of weight parameters, $w$:

$$\begin{aligned}\log[p(w|D)] &= \log[\pi(w)] + \log[\mathcal{L}(w)] + C \\ &= \log[\pi(w)] - \mathrm{Loss}(w) + C\end{aligned}$$

Deep neural net loss landscape




See: Loss surfaces… and What Are Bayesian Neural Network Posteriors Really Like?