A Bayesian perspective on the unmodelled burst problem

Antony Searle

LIGO Caltech

Overview

- [Not a detailed derivation of the bursts Bayesian search]
  - arXiv:0712.0196
- Maximum, constrained or regularized likelihood methods
  - Are implicitly subjective
- Every Monte-Carlo simulation has a corresponding optimal search
  - The likelihood/Bayes factor search using the same distributions
- Priors intimately related to physics knowledge
  - Bayesian analysis defined by physics model
  - No "statistics" decisions involved
- Implications? Where am I hopelessly naive?

The essential problem

- A signal model is necessary for any analysis
  - Even supposedly model-independent methods have implicit signal models
- We have limited (not zero) knowledge of it
- What do we do?

Notation

- Define a compact notation for multivariate normal probability distribution functions

$$N(\mu, \Sigma, x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

  - μ is an *n*-vector describing the centre of the distribution
  - Σ is an $n \times n$ covariance matrix describing the shape of the distribution
  - x is an *n*-vector at which we evaluate the probability distribution function

Physical system

- Use a linear model for the global network

$$x = F \cdot h + \varepsilon$$

$$p(\varepsilon) = N(0, \Sigma, \varepsilon)$$

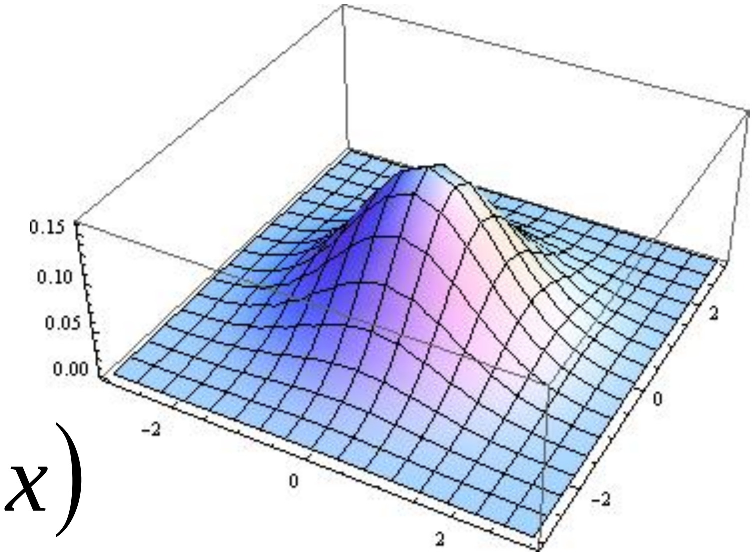
  - x is the set of observations
  - F is the response of the network to strain
  - h the incident strain
  - ε is a random variable representing noise
  - Σ is the covariance of the noise

Distributions

- Probability density functions:

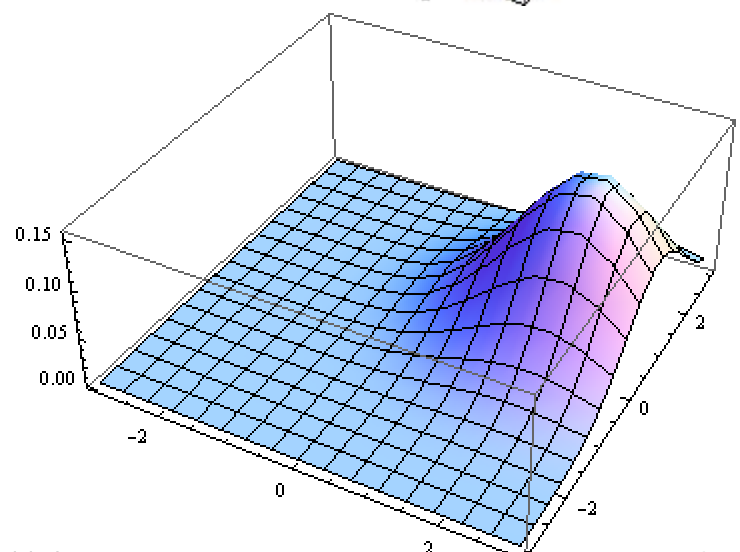
$$p(x \mid H_0) = N(0, \Sigma, x)$$

$$p(x \mid h, H_1) = N(F \cdot h, \Sigma, x)$$



- Likelihood test:

$$\frac{p(x \mid H_1)}{p(x \mid H_0)} > \lambda$$



What next?

- Standard, constraints, regularized and Bayesian all agree up to this point
- The problem is that we have

$$p\left(x \mid H_1\right) = ?$$

$$p\left(x \mid h, H_1\right) = N\left(F \cdot h, \Sigma, x\right)$$

  – H_1 is not a *simple* hypothesis; it is parameterised by the unknown strain h

Maximum likelihood

- We “need” to choose a value of h to evaluate p
  - Use the “most probable” value

$$\hat{h}(x) = \arg\max\ p(x \mid h, H_1)$$

$$p(x \mid H_1) = N\left(F \cdot \hat{h}(x), \Sigma, x\right) = \max_h\ p(x \mid h, H_1)$$

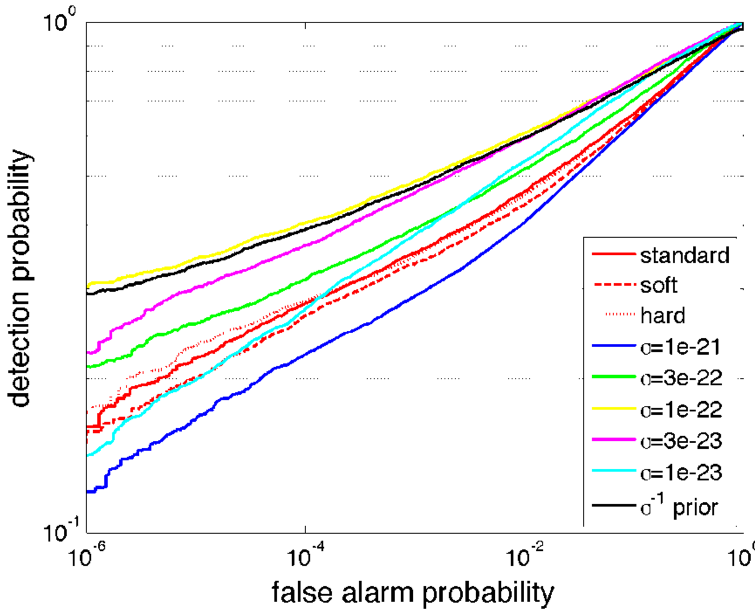
  - Often, it works fine
- Problems
  - We centre* the distribution on the data, then use that distribution to compute the probability of the data we centred it on?
    - “Drawing the bullseye around the bullethole”
  - $p(x \mid H_1)$ isn’t a distribution anymore (unnormalizable)

Maximum other things

- Instead, use the value of h that maximises something else
  - Hard constraint: constrain $h_2 = 0$
    - "Only move the bullseye vertically"
  - Tikhonov: penalise by h^2
    - "Don't move the bullseye too far"
  - Soft constraint: penalise by $(h_2)^2$
    - "Move the bullseye vertically freely but not too far horizontally"
  - (This doesn't mean that they are *ineffective*)
- The process of maximising a function of x inherently "peeks" at the data before performing the hypothesis test

Comparing methods

- Since the methods aren't "objective" or "unique," how do we choose?
- The community uses ROC curves produced by Monte-Carlo simulations to compare classification statistics ("noise" or "signal")
- The curve alone proves nothing; it is contingent on the signal (and) noise models chosen



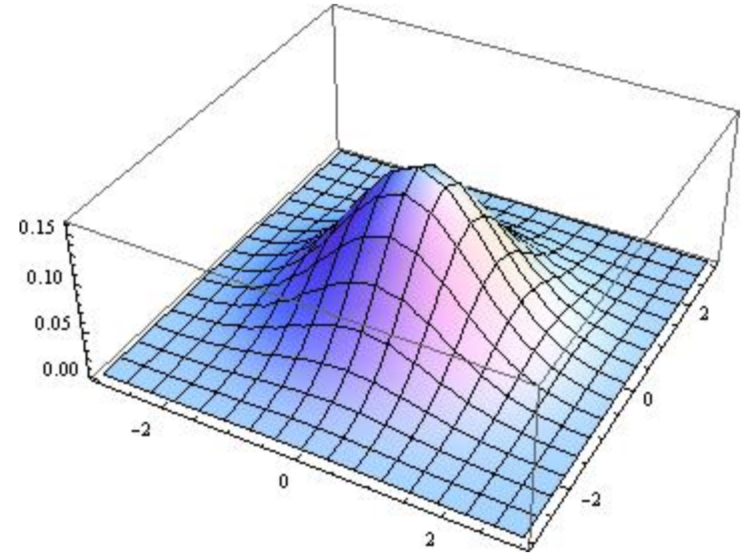
Formal definition

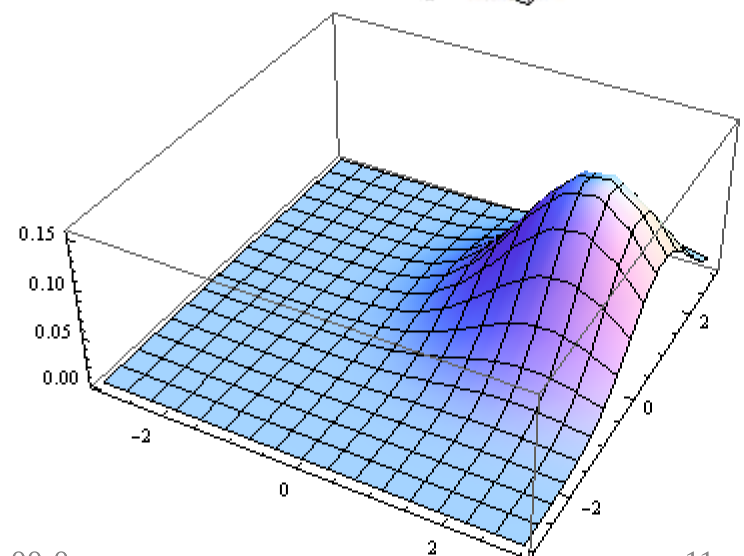
- A Monte Carlo analysis samples (at least) two distributions
  - Noise distribution

$$p(x \mid H_0)$$

  - Signal distribution

$$p(x \mid H_1)$$



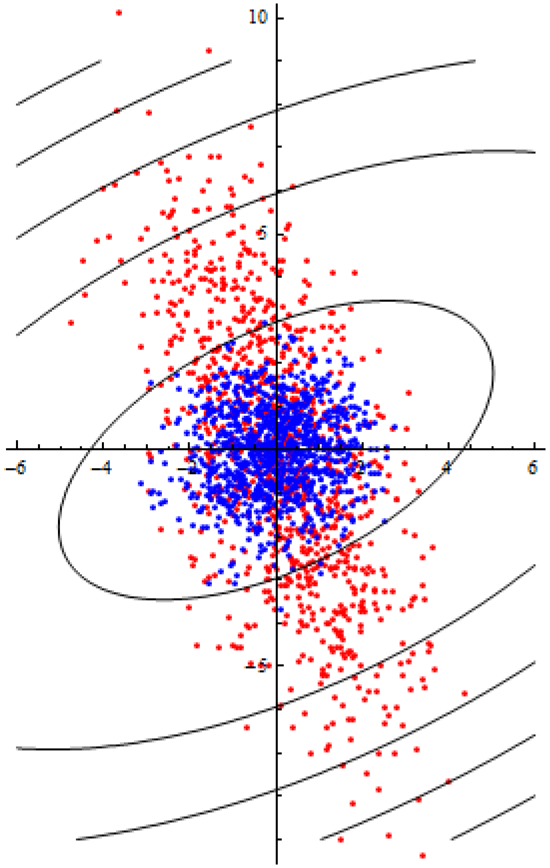


Optimal statistic

- For a given Monte-Carlo simulation, the optimal statistic is the likelihood ratio

  $$p(x \mid H_1) \,/\, p(x \mid H_0)$$

  – Any perturbation worsens performance for a given false alarm rate

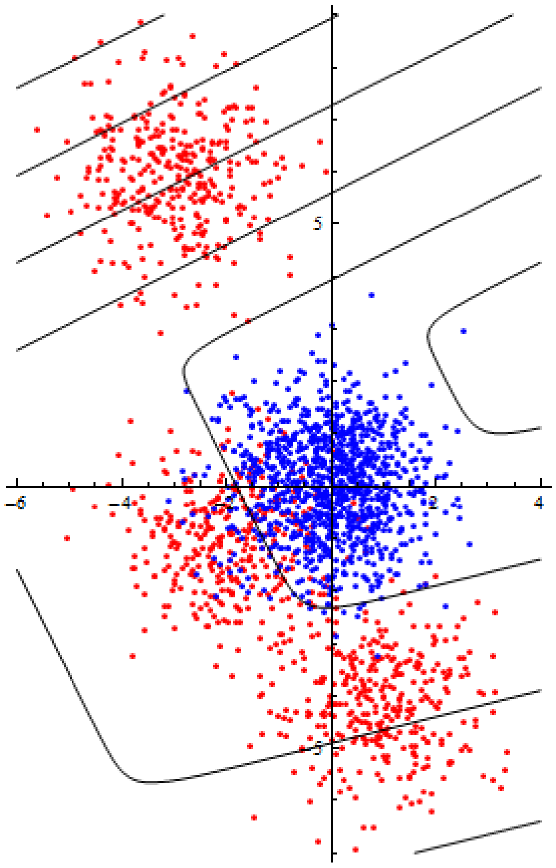


- Currently we
  - Design various methods
  - Design a Monte-Carlo simulation
    - Defines an optimal method
  - Test methods with the simulation
    - All guaranteed to be less effective than the optimal method

- Should we instead
  - Design a Monte Carlo simulation
  - Implement the method that is optimal
  - Not need to run it (except for validation)

Not best practice?

- We often use Monte-Carlo simulations that inject only a handful of waveforms and amplitudes, each many times
- Optimal statistic clearly “cheats” in this case
  - Goes to something like a template bank!
- The problem is not the optimal statistic, but the unreality of the simulation



- Why do we attach any credibility to the simulations?
  - Maybe we believe that the distribution is in some sense “smooth” over the space of waveforms, so our small zoo sufficiently samples it
    - Given the implicitness of this assumption, how do we know current methods aren’t cheating? (Unlikely, because the methods don’t seem to contain enough information; they too are “smooth”)
- If we believe this, we should use this smooth distribution in the Monte-Carlo simulation
  - The optimal statistic then won’t cheat
  - We’ll measure what we hope we’re measuring

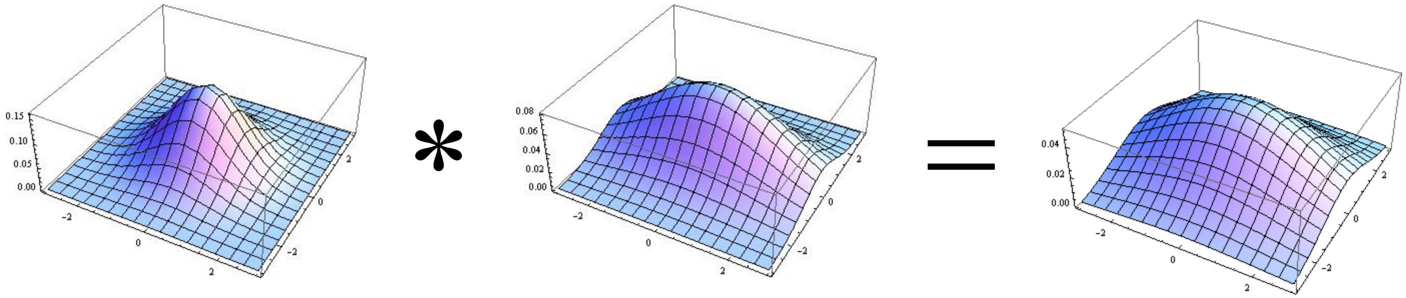
Better practice

- We should simulate a signal population that represents everything we know about the real world $p(h \mid H_1)$
- There's no unique right model
  - However, every model makes many definite physical predictions that we can use to rule out many
    - Rate doesn't fall as energy increases
    - Rate varies with direction
    - Amplitude varies with direction
  - People will disagree on the details
  - Robustness over different models doesn't require multiple simulations, just that we average the models together
    - If we're ignorant, this will result in an uninformative general signal model
- This is the best we can do in absence of knowledge of the real distribution

Implementing

- To sample from $p(x \mid H_1)$
  - Draw a sample strain h from $p(h \mid H_1)$
  - Next, draw a sample x from $p(x \mid h, H_1) = N(F \cdot h, \Sigma, x)$
- This convolves the distributions

$$p\left(x \mid H_1\right) = \int_{-\infty}^{\infty} p\left(x \mid h, H_1\right) p\left(h \mid H_1\right) \mathrm{d}h$$



Prior

- We've just made a Bayesian prior plausibility distribution $p(h \mid H_1)$
  - We don't assert that $p(h \mid H_1)$ represents the limiting behaviour of an infinite number of experiments (i.e. the real distribution of gravitational wave bursts)
  - We only assert that it represents our prior expectation for the first observation
    - That observation changes our state of knowledge, and we postulate a new prior for the second observation and eventually converge on the truth

Marginalization

- Drawing *h* then *x* is the same as drawing from the Bayesian marginalized distribution
  - Marginalization isn't an arbitrary process like (constrained) maximisation
  - The subjectivity is wholly contained within the prior, where it has an immediate physical interpretation
- The Bayesian analysis is the optimal statistic of the Monte-Carlo simulation with the same prior
  - A big target, but drawn before we shoot!
- Why not go directly to the Bayesian search?

Practicality

- Typically, the superiority of Bayesian analysis is moot because the marginalization integral is too expensive to compute
  - Hence the approximation schemes of the field of *multivariate analysis*, and the association of Bayesian methods with Metropolis-Hastings and other numerical methods
- However, the integral is soluble if we choose

$$p(h \mid H_1) = N(0, Z, h), \text{so that}$$

$$p(x \mid H_1) = N\left(0, \left(\Sigma^{-1} - \Sigma^{-1}F\left(F^T\Sigma^{-1}F + Z^{-1}\right)^{-1}\left(\Sigma^{-1}F\right)^T\right)^{-1}, x\right)$$

Multivariate normal distribution prior

- The prior is quite flexible
  - Sum of arbitrary number of template waveforms
- Examples
  - Matched filter
    - One waveform
  - White noise
    - $\{e_i\}$, Fourier basis, ...
    - Expresses as much ignorance as possible
  - Spectra / bands / tiles / clusters
    - e.g. weighted subset of Fourier basis vectors
  - Interpolation between simulated waveforms
    - Set of a small number of waveforms
  - Conservative enclosure of a zoo of waveforms
    - Distribution inferred by treating zoo as samples
- Well suited to searches where we are quite ignorant
  - Not well-understood non-linear subspaces (inspiral search)

Independence?

- Gursel-Tinto claims it

  “does not rely on any assumptions about the waveforms and in fact works for gravitational-wave bursts of any kind”

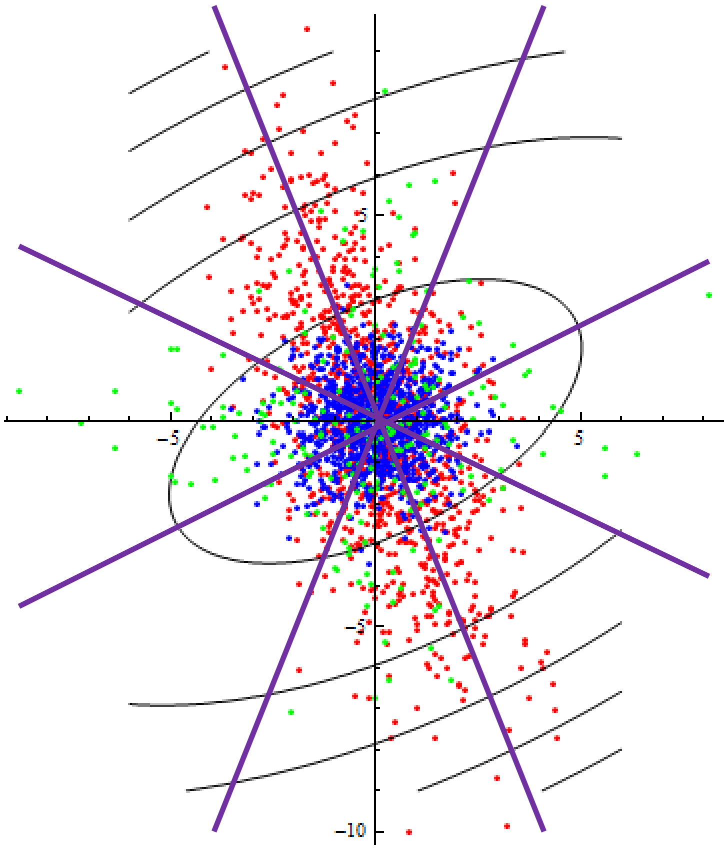
- Unfortunately, not true
  - Impossible in principle under Bayesian paradigm
  - We can even explicitly determine the priors

Implicit priors

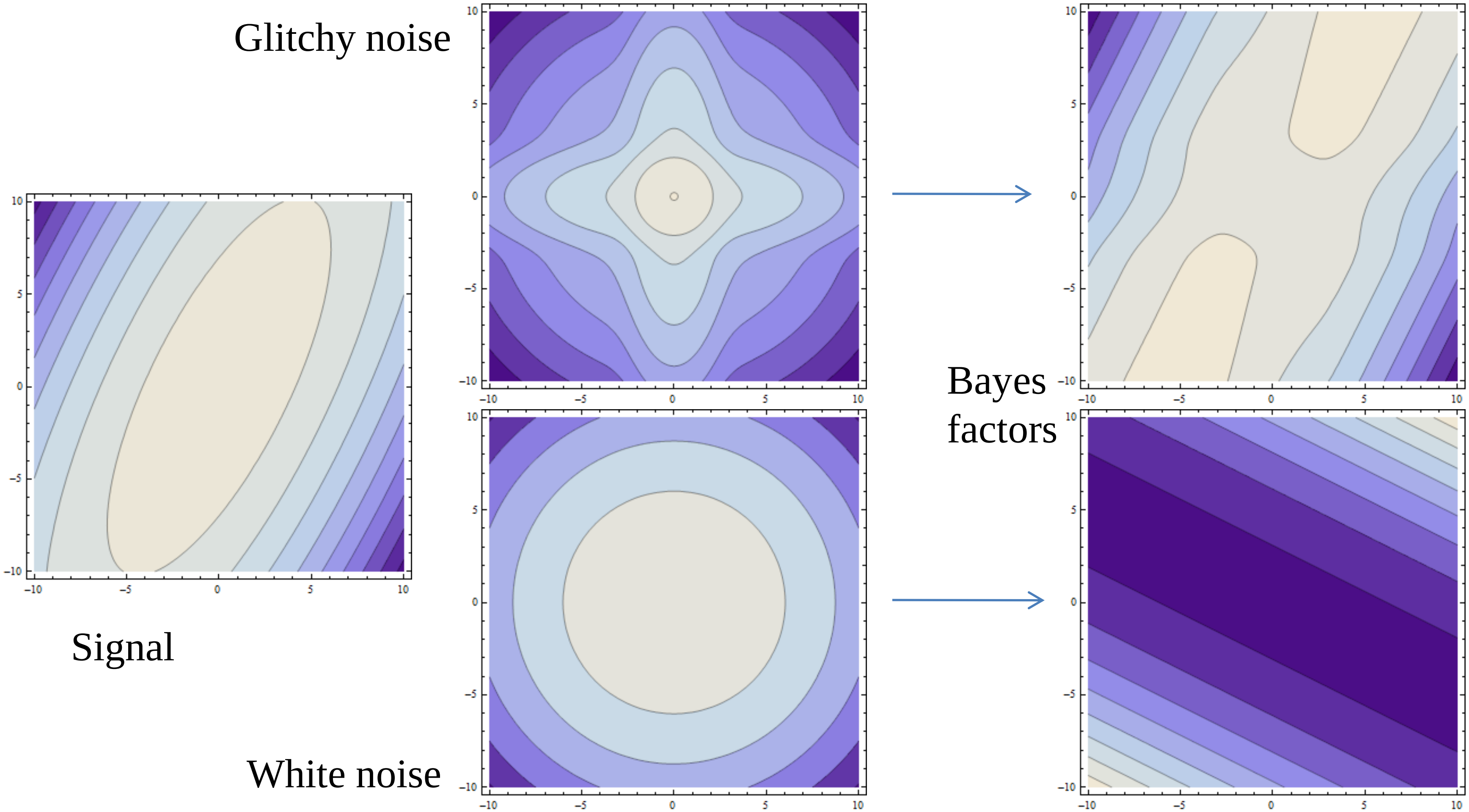
- We can reverse the derivation to extract the signal prior for a given statistic:
  - Gursel Tinto: (improper) $p(h \mid H_1) = 1$
    - Implies infinite gravitational waves
  - Tikhonov with regularizer α: $p(h \mid H_1) = N(0, \alpha^{-1} I, h)$
    - Implies a particular signal energy
  - Soft constraint: $\lim_{\sigma \to 0}$ for $p(h \mid H_1) = N(0, \sigma I, h)$
    - Implies infinitesimal signals
  - Hard constraint:
    - Implies infinitesimal, optimally oriented, linearly polarised signals
- These aren't "wrong", but they do contradict the state of knowledge of their creators

Glitches and cuts

- Everything also applies to the noise model
  - Except we have much more information (the data)
- We should include our best knowledge of glitches in the simulation
- "Cuts" to suppress glitches exclude parts of observation space in attempt to repair our use of a noise model we know is wrong
  - For example, incoherent energies
  - They are unnecessary / automatic if we can include them in the noise model
  - They are a kind of multivariate analysis by hand

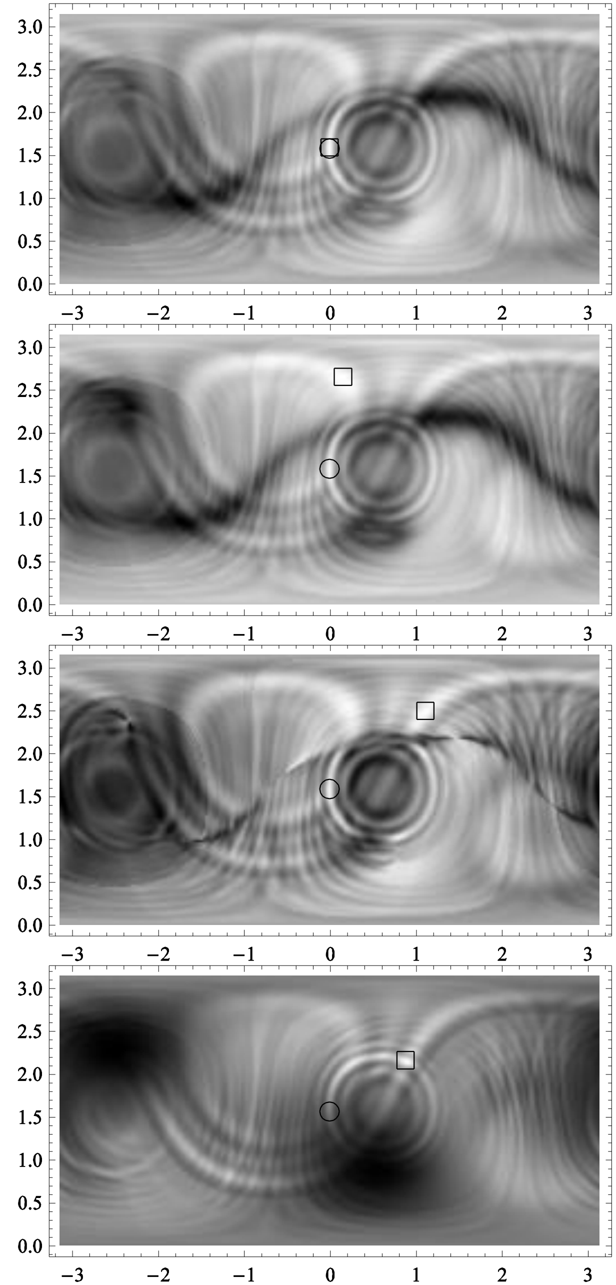


Glitchy noise

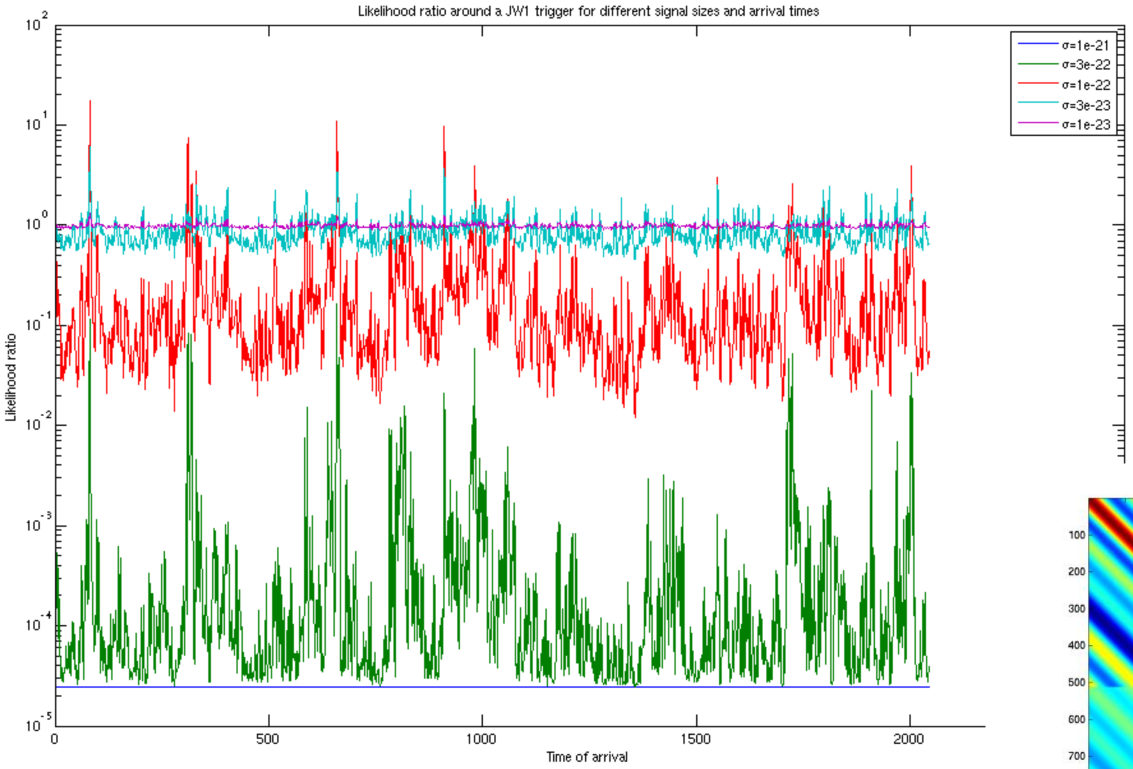


Directions

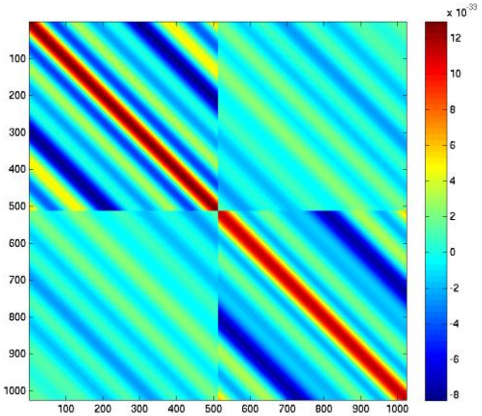
- Same problems as *h*
  - Actually $p(x \mid \theta, \phi, h, H_1)$
  - Maximum (constrained) likelihoods similarly find best-fit θ, ϕ
    - Implicit unphysical bias (rate varies with direction)
  - Explicit $p(\theta, \phi, \mid H_1)$ for Monte-Carlo and Bayesian
    - Physical prior performs better



Nascent Bayesian search



Direction marginalized Bayes factor as a function of arrival time and signal size, in restricted X-Pipeline search on a timeshifted JW1 peak correlator triple trigger



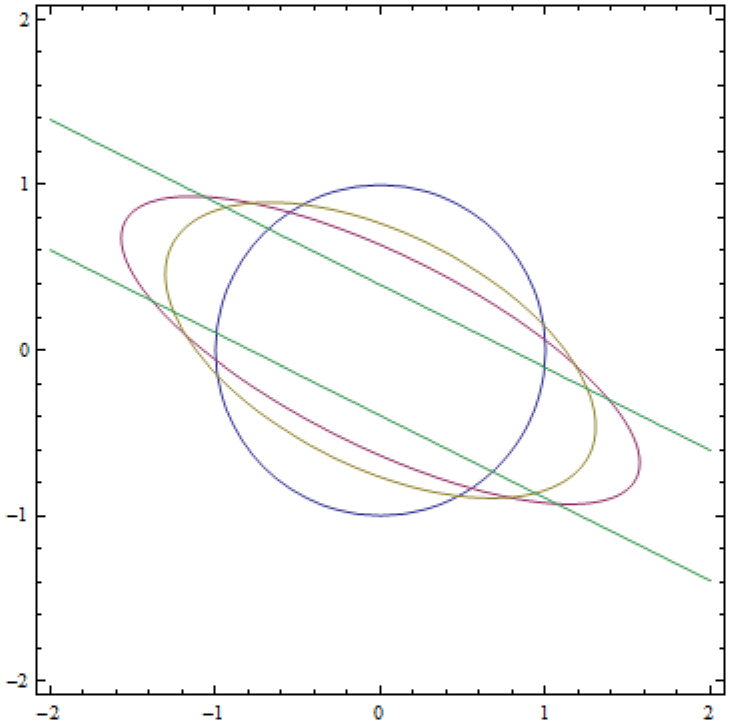
The covariance matrix Σ for simulated H1L1 in prototype full search

Implications?

- Current methods are biased (workarounds in X-Pipeline)
  - Can be fixed with very minor code changes and no performance penalty
  - Use the Tikhonov weights (a Bayesian special case)
    - Makes a Bayesian search with a proper prior for signal energy that can be peaked at threshold of detection
    - Minimal changes to current pipelines
  - Add a normalizing function of direction
    - Very cheap and simple change removes directional bias
  - Marginalize rather than maximise over direction
    - Very cheap simple change
  - Marginalize over several Tikhonov regularizer values
    - Expensive but removes expected energy bias
- Do we need a consensus physically realistic signal model?
- Instead of ad-hoc cuts and statistics, it is practical to use physical models of signals and glitchy noise

Quadratic form

- For stationary signals, the Bayesian statistic is similar in form
  - Easily retrofitted into other pipelines
- Statistic isosurfaces for
  - Standard
  - Bayesian
  - Soft
  - Hard



Bad practice

- Forming ROC curves for
  - Particular waveforms
  - Particular SNRs
- This is the performance of the method on a tiny subspace of any credible signal population

Bayesian

- Bayesian inference lets us turn $p(x \mid h, H_1)$ into $p(x \mid H_1)$ without peeking at x.
  - Instead we specify *a priori* how *plausible* (not *probable*) a particular h is with a *prior plausibility distribution* $p(h)$
- This is a subjective choice
  - Choosing what to maximise is no less subjective
    - In fact, all the maximisation schemes above are precisely equivalent to certain (weird) choices of prior
  - $p(h)$ has an immediate physical interpretation

Physicality of Priors

- Priors are not probability distributions
  - $p(h)$ does not purport to be the limiting frequency of many gravitational wave experiments
  - $p(h)$ it represents how plausible we think the occurrence of a waveform is, and hence
    - Physical statements about bursts we agree with
    - How much we would bet on a particular waveform occurring
    - A waveform generator that we would find credible in a Monte Carlo simulation
  - Priors are capable of representing ignorance

Monte-Carlo

- Given two statistics, how do we evaluate their relative performance?
- We perform a Monte-Carlo simulation and construct the ROC curves
- To do so, we must use some kind of signal model
- The usefulness of the simulation is contingent on the credibility of the signal model

Bayesian analysis

- Form the posterior odds ratio for two hypotheses H_i given data **x**

$$\frac{p(H_1 \mid \mathbf{x})}{p(H_0 \mid \mathbf{x})} = \frac{p(H_1)}{p(H_0)} \frac{p(\mathbf{x} \mid H_1)}{p(\mathbf{x} \mid H_0)} > 1$$

- Requires an explicit distribution for both hypotheses to **form** the likelihood ratio

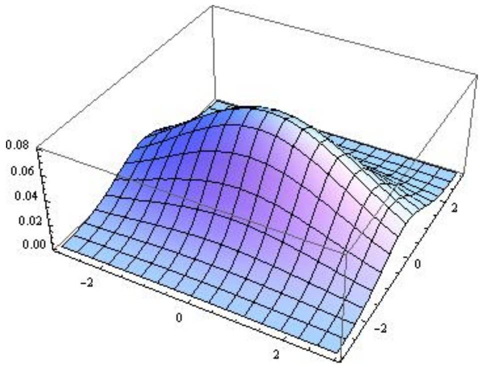
$$\frac{p(\mathbf{x} \mid H_1)}{p(\mathbf{x} \mid H_0)} > \frac{p(H_0)}{p(H_1)} = \lambda$$

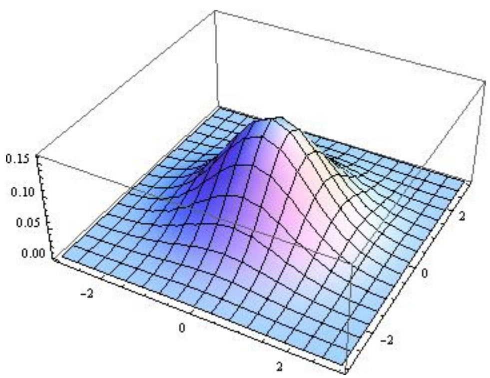
- Not particularly “Bayesian” yet

Plausibility vs. probability

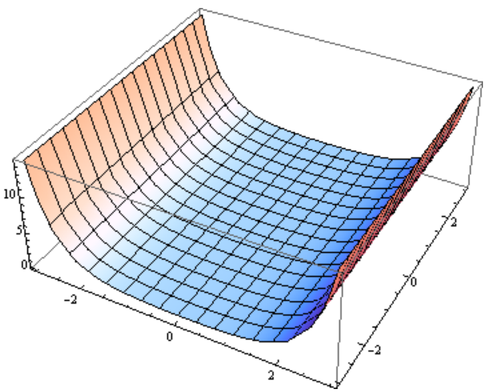
- The *probability* distribution for bursts is unknown*
- Use a *plausibility* distribution
  - Does not assert that it is the limiting behaviour over many experiments
    - The results of the experiment inform the prior plausibility distribution for the next experiment, and we asymptote to the true probability distribution*
  - Does reflect our state of knowledge, and encodes many concrete physical predictions
    - For example, the prior should imply that signals are rare and large signals are rarer than small signals

Bayesian





=



$$> \frac{p(H_0)}{p(H_1)}$$

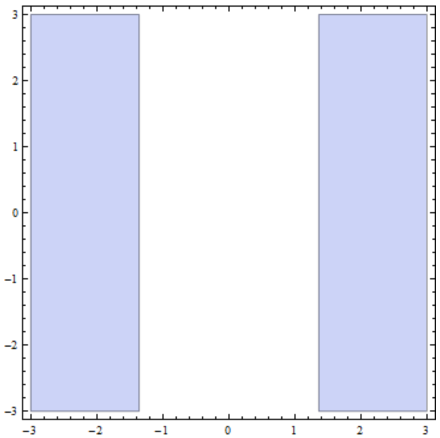
The statistic is the quotient of the signal and noise distributions

Monte-Carlo analogy

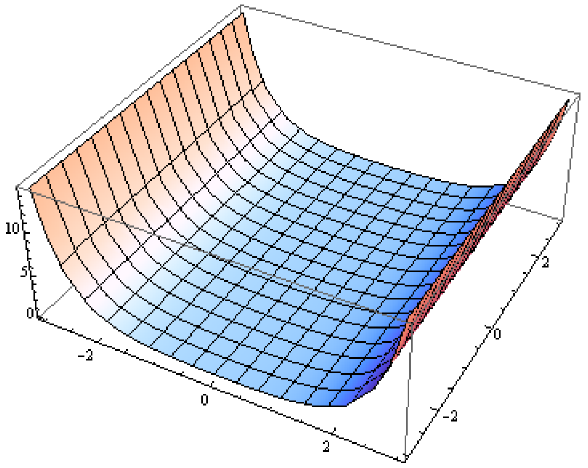
  - Monte-Carlo simulations perform random draws from signal and noise distributions
- For any Monte-Carlo simulation there is a unique optimal Bayesian search and vice versa
  - Both defined by $p(\mathbf{x} \mid H_i)$
- Insofar as we believe that a Monte-Carlo simulation is useful, we should use its corresponding Bayesian search

Classification

- All detection algorithms, Bayesian or otherwise, partition space of possible observations **x** into detection and non-detection

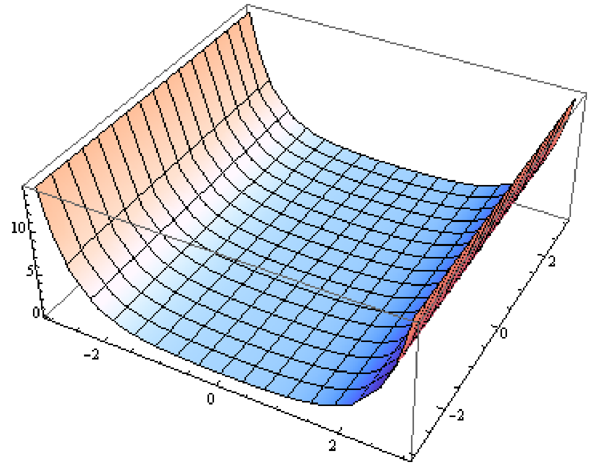


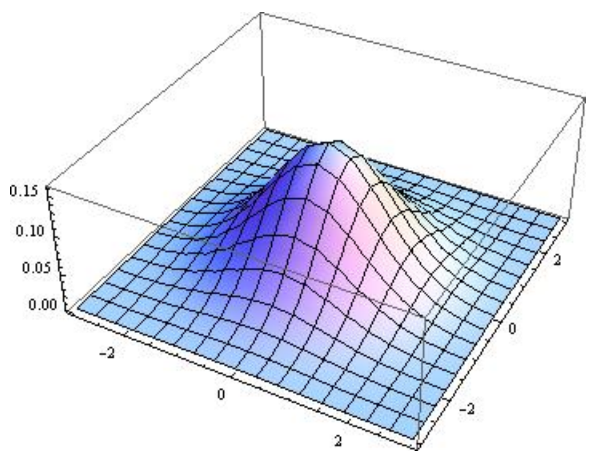
- Most non-Bayesian methods do this by generating a statistic whose isosurfaces form partitions of different confidence levels

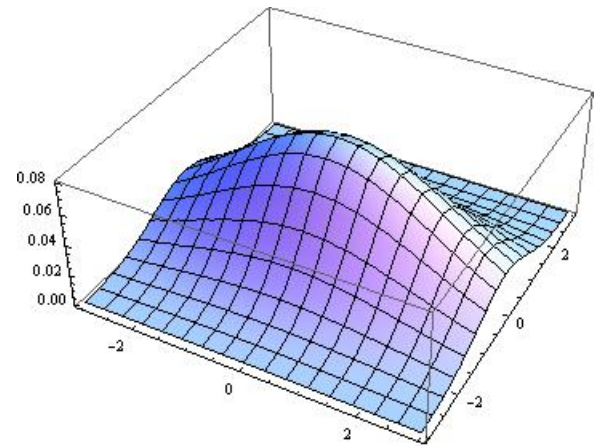


Working backwards

The signal distribution is the quotient of the statistic and the noise distribution (up to monotonic transformation)







Implicit signal priors

- We can deduce the implicit signal prior plausibility distribution for non-Bayesian methods!
    - GT asserts infinite signal energy
    - Constraint asserts infinitesimal energy
    - Tikhonov asserts a specific energy (= regularizer^{-1})
    - All optimal for (rate and/or energy) varying across the sky
- These are not credible priors
    - We would not find the results of corresponding MC tests particularly compelling
        - We can immediately think of better MC simulations to run
- This is not to say that existing searches perform poorly

Equivalence

- Families of partitions of space of observations
- Equivalence between a partition and a signal model
  - All methods assume a signal model from Bayesian perspective
  - Prior plausibility is not probability – does not assert that it represents distribution of many experiments

Credibility

- The credibility of a Monte-Carlo simulation depends on the injections and noise used
  - The simulation on the right tests only one large waveform and has glitchless noise
  - It tells us almost nothing about real-world performance