Putting the newest of Bayes into context for astronomers

Alanna Connors

ABSTRACT In his paper, J. Berger has issued a friendly invitation to Bayesian methods, both classical and new. In this paper I try to put some of those concepts into context for astronomers. Particularly for those for whom Bayesian inference is new, I hope to help translate why it might be of interest to invest the significant amount of intellectual and software effort involved in retooling. I highlight some of the standard benefits and objections to classical Bayesian inference, then sketch out two simple examples. For the first, because we are astrophysicists, everything works. For the second, more complicated example: maybe physicists could use thoughtful expert help after all. I conclude with a few personal thoughts on moving towards likelihood ratios, either frequentist or Bayesian; and towards leaving the appealing but "ad hoc" statistics for data–exploration.

1 Goals/Context

1.1 Introduction

Why might recent developments in Bayesian analysis, or even standard Bayesian procedures, be of interest to astronomers and physicists? J. Berger, in [BE96] presents some examples, from the point of view of a statistician. In this paper, I try to translate these concepts to a point of view more familiar to astronomers and physicists. [BE96] focuses on hypothesis testing and model selection. I try to start more slowly. I first highlight terms that may be unfamiliar; and then very briefly sketch out standard Bayes parameter estimation and likelihood ratios for two examples from γ–ray astrophysics. With these in mind, one can see where [BE96] presents classic examples of Bayesian hypothesis testing; plus both some intriguing new ideas on the difficult area of priors; and new developments in computer techniques. I hope this might also briefly give statisticians some of the flavor of of trying to eke out inferences about physical conditions of objects in the distant sky; and where Bayesian methods might be more practical. I close with a few personal thoughts on moving towards the use of likelihood ratios.

1.2 What is it?

Bayesian inference is a clear procedure for building measurement tools (probabilities and their ratios) for: 1) parameter estimation; 2) model selection and hypothesis testing; 3) robustness and sensitivity of results to model choice, and prior information; and 4) prediction. Many astrophysicists are more familiar with *sampling statistics*: the probability of the data $\mathcal{X}$, given a model or hypothesis $\mathcal{M}$ and parameters Θ, $p(\mathcal{X}|\Theta\mathcal{M}I)$ (or $p(\mathcal{X}|\mathcal{M}I)$). With Bayesian inference one works with the inverse: the probability of a model or hypothesis $\mathcal{M}$ and parameters Θ given the data, $p(\Theta|\mathcal{M}\mathcal{X})$ (or $p(\mathcal{M}|\mathcal{X})$). One gets from one (*data–space, on the right*) to the other (*parameter– or hypothesis–space, on the left*) via *Bayes's Theorem:*

$$p(\Theta|\mathcal{X}I) = \frac{p(\Theta|I)}{p(\mathcal{X}|I)} p(\mathcal{X}|\Theta I), \text{ or } p(\mathcal{M}|\mathcal{X}I) = \frac{p(\mathcal{M}|I)}{p(\mathcal{X}|I)} p(\mathcal{X}|\mathcal{M}I). \tag{1.1}$$

Here "I" represents prior measurements and information; $p(\Theta|I)$ (or $p(\mathcal{M}|I)$) is called the *prior probability*; $p(\mathcal{X}|\Theta I)$ the *direct probability* or *sampling statistic*; $p(\Theta|\mathcal{X}I)$ (or $p(\mathcal{M}|\mathcal{X}I)$) is the *posterior probability*; and $p(\mathcal{X}|I)$ serves as a normalization term.

The references cited by [BE96] give fine overviews and bibliographies. I would like to highlight two: [JA78] contains a classic historical account from the perspective of a physicist. Perhaps the earliest modern use of Bayesian inference in astronomy is [BI71].

1.3 How is it different from what I'm used to doing?

Sampling statistics is based on the long–term (asymptotic) frequency of occurrence of a particular pattern of data, assuming the model is true. Many astronomers use the recipes for likelihood ratios in [LM79],[CA78] to generate confidence intervals, which are based on the Central Limit Theorem asymptotically holding. (Also, some astrophysicists might be more comfortable with the applied math term "inverse problems" [CB86]. Or, they may not have realized that "forward–fitting", using χ^2, is a maximum–likelihood method that assumes a Gauss–Normal form for the sampling statistic.) By contrast, Bayesian inference calculates the probability of the parameters (or model) given any prior information, plus just the data one has.

The concept of *priors*, of assigning probability distributions to parameters before making inferences from the data, may also be new to astrophysicists. [BE96] lists many standard options then spends some time discussing new "one-size-fits-all" priors; usually they are "custom–built". I want to highlight two distinctions: *informative* versus *uninformative* priors; and *proper* versus *improper* priors. When one has significant prior information (such as a previous background measurement, or knowledge of atomic line strengths), one can use an *informative prior*. Without such knowledge, one uses an *uninformative prior*. In the latter case, a physicist or astronomer can often constrain the form of the prior from knowledge of the geometry of the physical system, or physics theory, or invariance arguments (see also [JA78]). A *proper prior* is one that is normalized to one; while an *improper prior* is a handy analytic form (such as a constant or log distribution) that, when integrated over all parameter space, tends to ∞ and so is not normalizable. [BE96] notes that the latter can work well for parameter estimation, but has drawbacks for model comparison and hypothesis testing. This drives his "intrinsic Bayes factor" approach.

When working in parameter–space one can integrate over uninteresting (or "nuisance") parameters; or indeed over all parameters. This is called *marginalization*; another potentially unfamiliar term. Note that (by marginalizing over all parameter space) one can directly calculate and compare the global probabilities of two hypotheses with differing numbers of parameters. There is no need to add an extra factor for each degree of freedom (e.g. in sampling statistics one might require the difference in χ^2, equivalent to $-2\log\big[p(\mathcal{X}|\Theta I)\big]$, to be more than 1). As [BE96] illustrates, integrating over each extra dimension intrinsically takes this into account.

2 Benefits / Objections

2.1 Benefits

It gives a clear mechanism to build a tool to get the best measure of distance between two clearly stated hypotheses. It is always a *sufficient statistic*; that is, it incorporates all the information about the hypotheses that is available in the data; and it includes a mechanism to optimally incorporate prior information. For example, [BH93] suggests an appealing but "ad hoc" statistic for incorporating imaging information when searching for periodic γ-ray emission from a known radio pulsar. Each γ-ray photon is weighted by its angular distance from the source according to a telescope point–spread function, before the data are binned at the pulsar period into a phase histogram, and a χ^2 test for a flat light–curve is performed. This seems intuitive, but how does one know whether it incorporates all of the information available in the data, and in one's prior information?

One can tackle any problem where the hypotheses are clearly stated. For example, many image processing applications have very large numbers of parameters, comparable to the number of data points. This can be a numerically intractable "inverse problem", until one notices that with Bayesian methods one has a prior that can act as a regularizer.

It is valid for moderate and small data sets (no asymptotics required). The familiar recipes used by astronomers to generate confidence intervals are based on the Central Limit Theorem [CA78], [LM79]. Often this does not strictly hold. For example there may be multiple peaks in the probability–

space. Or, the sample size may be very small and the measurement not repeatable: [LO92] points out there was only one chance to measure neutrinos from SN 1987A; there were roughly two dozen neutrinos, and approximately 8 parameters.

One can reduce dimension of problems by integrating over uninteresting parameters. A common example: an interesting source energy spectrum might have $\sim 10^2$ energy bins, low Poisson counts per channel, plus measurements of the $\sim 10^2$ background rates in each bin. One does not subtract the background rate from each energy channel in the source spectrum, but instead *marginalizes* over the imperfectly known background rates [LO92]. It also clears up what to do with the "number of trials" question: one *integrates* over a range of trial parameters.

One can compare the likelihoods of non–nested models with different numbers of parameters. [BE96]. Also, by definition, *one can handle uncertainties in the model or in prior information.* Examples include uncertainties in stellar coronal models; or in energy response matrices.

2.2 Objections

Learning the language, retooling. "It's not in Bevington." No, it's not; but neither are most of the techniques discussed in these proceedings. Becoming familiar with the language of priors, posteriors, marginalizations, and credible regions requires a significant effort.

Getting practical, reliable priors. This is an active area of research, as [BE96] makes clear. One approach is to report one's results in a form where the effect of using different priors is easy to calculate.

Computation time. "Rev. Thomas Bayes started his calculation in 1783, and they're just now finishing." – D. J. Forrest on the recent rise in interest in Bayesian methods. Although marginalization is a Bayesian technique of great power, it requires integrating over parameter space. Numerical integration in high dimensions is one of the classic high–CPU problems. [BE96] touches on some new techniques. However, when the integration can be done analytically, marginalizing can actually speed up a calculation [LO92].

No general "goodness of fit" like χ^2. "That's an objection?" – standard Bayesian response. Standard significance tests use the tail of the distribution. [BE96] works through an example showing this is often not a very good discriminator between two hypotheses. Instead a Bayesian analysis specifically calculates the probability or likelihood of two (or more) hypotheses.

3 Simple example: Astrophysicists have it easier than statisticians

3.1 Specifying the problem

Periodic Time Series Analysis. Suppose one is searching for γ-ray emission from a known pulsar, with position, period, and all period derivatives known from radio data. Given a set of γ-ray data, what is the likelihood that a periodic signal has been detected? This is a quick sketch. For more details, [GL92] carefully treat a problem that is similar but has a different shape function.

Data. The data are in the form of time–tagged events (point Poisson process): a list of photon arrival times with a 3° window around the source position, and standard data quality cuts on the other parameters [MU95]. The two sets I show here are 1–3 MeV and 10–30 COMPTEL data on the well–known 33 ms Crab pulsar. It is a 14 day observation. There are 54626 photons in this 1–3 MeV dataset (about 1 every 20 seconds); and 1981 in the 10–30 MeV data (about 1 every 10 minutes). There is known to be a significant ($> 80\%$ of the events) background component. For this example we look for the total pulsed fraction of the source + background rate.

Null hypothesis, $\mathcal{M}_0$. The photon arrival times are completely random, and can be described by a Poisson process with a constant rate $\mu_0(t) \;=\; B$.

Interesting hypothesis, $\mathcal{M}_1$. The photon arrival times are periodic, with a shape described by $\rho(t)$, with $< \rho(t) > \equiv 1$ when averaged over one cycle; and total normalization described by B: $\mu_1(t) \;=\; B\rho(t)$.

Shape function for interesting hypothesis. Since this is a Poisson process, it is convenient to describe the periodic shape by an exponentiated Fourier series, or generalized von Mises distribution. For one component, $\rho(t) \propto \exp\left[-\kappa \cos(\Theta(t) + \phi)\right]$, with $\Theta(t)$ the pulsar phase from radio data; and ϕ the unknown phase difference between the radio and gamma–ray energies. The parameter κ is known as the shape or concentration parameter, with pulsed fraction $f \;=\; \tanh(\kappa)$. The normalization condition $< \rho(t) > = 1$ requires $\rho(t) \;=\; \exp\left[-\kappa \cos(\Theta(t) + \phi)\right]/I_0(\kappa)$, where I_0 is the modified Bessel function of order zero.

3.2 Assigning probabilities

Priors. Knowing the physical meaning of the pulsed fraction $f \in [0, 1]$ and relative phase $\phi \in [0, 2\pi]$ allows one to assign unambiguous properly normalized prior probabilities, even when one has no previous measurements. From symmetry, one argues that the prior for the phase ϕ should be $p(\phi|I)\mathrm{d}\phi \;=\; \mathrm{d}\phi/(2\pi)$. Likewise, the prior on the pulsed fraction can be given by $p(f|I)\mathrm{d}f \ = \ \mathrm{d}f$. The prior on $B \in [0, B_0]$ is the only ambiguous assignment. Should it be a uniform prior, $p(B|I)\mathrm{d}B \ = \ \mathrm{d}B/B_0$? A log–uniform prior, $p(B|I)\mathrm{d}B = \mathrm{d}B/(B\log(B_0))$? However, whatever the choice, all dependence on B will be exactly the same for the null and interesting hypotheses, and so will cancel when a likelihood ratio is taken. For this example, I chose the former, and let $B_0 \to \infty$ at the end of the calculation.

Direct probability. For both null and interesting hypotheses, one uses the Poisson probability, given a model rate $\mu(t)$, and detection of N photons at times $\{t_k\}$, in a total live–time T_L, in (very small) time bins δt [GL92]:

$$p(\{t_k\}|\mu(t), I) \ = \ \exp\left[-\int_{T_L}\mu(t)\mathrm{d}t\right]\prod_{k=1}^{N}\mu(t_k)\delta t. \qquad (1.2)$$

Turning the crank. For each hypothesis, one applies Bayes's Theorem, integrates analytically over the amplitude and phase parameters B and ϕ, and then takes the ratio. (The normalization term $p(\mathcal{X}|I)$ cancels, and so is not calculated.) This gives $\lambda(f)$, the log likelihood for parameter estimation:

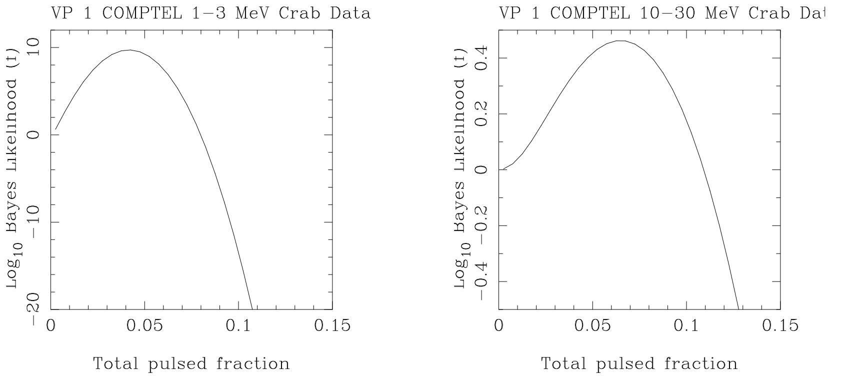
$$\lambda(f) \ = \ \log_{10}\Big[I_0(\kappa\mathcal{S}_N)/I_0(\kappa)^N\Big], \qquad (1.3)$$

where $\mathcal{S}_N$ is defined as $\mathcal{S}_N \ \equiv \ \frac{1}{N}\sum_{k=1}^{N}\cos^2\Theta(t_k) + \sin^2\Theta(t_k)$. For $\kappa = 1$, this is analogous to a frequentist Rayleigh statistic.

For hypothesis testing, one obtains the Bayes factor, or ratio of the total probabilities of the interesting to null hypotheses:

$$\mathcal{L} \ = \ \int_0^1 \mathrm{d}f\frac{I_0(\kappa\mathcal{S}_N)}{I_0(\kappa)^N}; \quad f \ = \ \tanh(\kappa). \qquad (1.4)$$

3.3 Application to data



Here we plot $\lambda(f)$ for two different datasets. Both are from a two week CGRO–COMPTEL observation of the Crab pulsar. The first shows the 1–3 MeV band, where it was detected very significantly (total pulsed fraction $f = 0.042 \pm 0.006$; Bayes factor $\mathcal{L} = 10^{7.8}$). The second shows the 10–30 MeV Crab data. The total pulsed fraction $f = 0.063 \pm 0.03$ is suggestive, but not a formally significant detection (Bayes factor $\mathcal{L} = 10^{-0.6} < 1$).

4 Adding a complication: astrophysicists need help from statisticians

Joint imaging and timing analysis. With Bayesian inference, it is straightforward to add more information. Since these data were from an imaging telescope, why not use the imaging response on the full dataset, rather than just an angular window around the source? One should be able to derive a likelihood ratio for joint imaging and timing analysis, and at once obtain credible regions for both the source flux and pulsed fraction. The data are the same, save that a much wider angular window was used. There are 157175 photons in this 1–3 MeV dataset (about 1 every 8 seconds); and 7096 in the 10–30 MeV data (about 1 every 3 minutes). The models are a little more complicated. Let j be the index for the spatial imaging bins; β_j the shape of the background as a function of bin position, with $\sum_j \beta_j \equiv 1$; $\mathcal{R}_j$ the instrument response (or point–spread function) in bin j, given the known pulsar position; and A the source flux (photons-cm^{-2}-s^{-1}). Note that the shape of the instrument background β_j and the response $\mathcal{R}_j$ are both known a priori. The rate for the null hypothesis, $\mathcal{M}_0$, is still one component: $\mu_{0j}(t) = B\beta_j(t)$. However, the rate for the interesting hypothesis, $\mathcal{M}_1$, is now two (background + source): $\mu_{1j}(t) = B\beta_j(t) + A\mathcal{R}_j\rho(t)$.

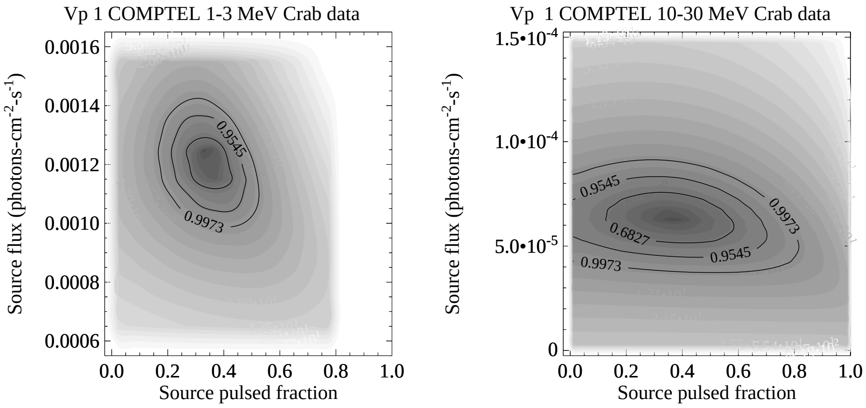
Assigning probabilities. One assigns the same priors for ϕ, f, and B one did previously, but how does one assign a prior for A? There is no one unambiguous choice, and dependence on A will not cancel when the likelihood ratio is taken. For this calculation, I used a uniform prior on $A \in [0, A_0]$, with $A = 10^4$ photons-cm^{-2}-s^{-1}. Once the μ are given, the direct probabilities have the same form as before.

Turning the crank. This gives $\lambda(f)$, the log likelihood for parameter estimation:

$$\begin{aligned}\lambda(f, A) \;&=\; \log_{10}\Big[p(A|I)\int_0^{B_0} \mathrm{d}B \frac{{T_L}^{N+1}}{N!} \;\times \\ &\exp\big[-T_L \sum_j \big(B\beta_j + AR_j\big)\Delta V_j\big] \prod_{k=1}^{N}\Big(B\beta_{j_k} + AR_{j_k}\rho(t_k)\Big)\Big], \qquad (1.5)\end{aligned}$$

and global Bayes factor $\mathcal{L} = \int_0^1 \mathrm{d}f \int_0^{A_0} \mathrm{d}A\; 10^{\lambda(f,A)}$, where the integrations over B, f and A are performed numerically.

Application to data. The results (68.27, 95.45, and 99.73% posterior probability credible regions) are displayed for the same CGRO–COMPTEL Crab observations as before.



The detections appear more significant. For the 1–3 MeV data, one finds a source flux $A = 1.2 \times 10^{-3} \pm 6 \times 10^{-5}$ photons-cm^{-2}-s^{-1}; a source pulsed fraction $f = 0.35 \pm 0.05$; and a global Bayes factor $\mathcal{L} = 10^{75.8}$. For the 10–30 MeV data, one finds a source flux $A = 6.1 \times 10^{-5} \pm 7.6 \times 10^{-6}$ photons-cm^{-2}-s^{-1}; a source pulsed fraction $f = 0.36 \pm 0.15$; and a global Bayes factor $\mathcal{L} = 10^{5.99}$. However, without a prior for A with an unambiguous normalization, it is hard to interpret the total likelihood of the hypothesis that there is a pulsed γ-ray source. A different choice of prior and A_0 would have given about the same parameter constraints, but different global Bayes factors. This was the problem addressed by J. Berger's "intrinsic Bayes factor" method.

5 Future thoughts

For the future. Clearly thoughtful priors are an active area of concern for the future. For many problems, an astrophysicist may be able to use physical knowledge of a system to assign reasonable, proper priors; for others, the choice may be ambiguous, so much remains to be worked out. We are aided by both increases in computation speed, and by new numerical integration techniques such as MCMC. This allows a greater flexibility in the kinds of problems one can tackle in a reasonable amount of time.

Personal thoughts. I often find that, once having derived a Bayesian likelihood ratio, I later see a relation to a standard maximum likelihood statistic. I find the Bayes prescription clearer, especially when exploring the problem. [TA93] coined term "likelihoodist" to describe those basing their inference on the shape of a likelihood, Bayesian or otherwise. Astronomers are clever people, and come up with many ingenious, intuitive, and speedy ad–hoc statistics. I am coming to consider these as methods of data exploration and visualization; but for the final calculations of probabilities and uncertainties, I encourage astrophysicists to make more use of a "likelihoodist" perspective.

Acknowledgments: I thank T Loredo, E Linder and D Sinha for pivotal discussions. T Loredo provided software for calculating the Bayesian credible regions shown in the figures. AC is supported through the CGRO–COMPTEL project, which is supported in part through NASA grant NAS 5-26646, DARA grant 50 QV 90968, and the Netherlands Organization for Scientific Research (NWO).

6 References

[BE96] J. Berger, these proceedings

[BE69] P. R. Bevington. *Data Reduction and Error Analysis for the Physical Sciences.* McGraw-Hill, 1969.

[BH93] L. E. Brown and D. H. Hartman. *Astrophys. & Space Science*, **209**, 285, 1993.

[BI71] A. B. Bijaoui. *Astron. & Astrophys.*, **13**, 226, 1971.

[CB86] I. J. D. Craig and J. C. Brown. *Inverse Problems in Astronomy.* Hilger, 1986.

[CA78] W. Cash. *Astrophys. J.*, **228**, 939, 1978.

[GL92] P. Gregory and T. Loredo. *Astrophys. J.*, **398**, 146, 1992.

[JA78] E. T. Jaynes. Where do we stand on Maximum Entropy?. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics.* Kluwer, 1978.

[LM79] M. Lampton, B. Margon and S. Bowyer. *Astrophys. J.*, **208**, 177, 1976.

[LO92] T. Loredo. In *Statistical Challenges in Modern Astronomy*, Springer-Verlag, 1992.

[MU95] R. Much *et al.* In *Proceedings of the Compton Symposium, Munich* 1995.

[TA93] M. A. Tanner. *Tools for Statistical Inference.* Kluwer, 1993.