Bayesian Model Selection and Model Averaging, With Applications to the Cepheid Distance Scale

William H. Jefferys
University of Texas at Austin, USA

Collaborators

Thomas G. Barnes, III
University of Texas at Austin

James O. Berger
Peter Müller
Duke University

Bayesian Analysis and Astronomy

- Bayesian analysis offers many advantages for astronomical research and has attracted much recent interest.
- Astronomy and Astrophysics Abstracts lists 88 articles with the keywords 'Bayes' or 'Bayesian' in the past 5 years, and the number is increasing rapidly.
- This June, at the AAS meeting in Chicago, there was a special session on Bayesian and Related Likelihood Techniques. See

  http://www.aas.org/meetings/aas194/prelim/statistics.html

  My Duke colleague, Jim Berger, was the invited statistician for that session.

Advantages of Bayesian Methods

- Bayesian methods allow us to do things that would be difficult or impossible with standard (frequentist) analysis.
  - It is easy to incorporate prior physical or statistical information
  - It is coherent: we will not find ourselves in a situation where the analysis tells us that two contradictory things are simultaneously likely to be true.
  - Analysis depends only on what is actually observed, not on observations that might have been made but were not.
  - It can compare models and average over models, whether nested or not.
  - Correct interpretation of results is much more natural, especially for physical scientists.

Advantages of Bayesian Methods

- Bayesian analysis naturally incorporates prior information into the analysis. Indeed, the investigator is *required* to provide relevant prior information.
  - Prior information can include physical constraints, e.g., that both background and signal are greater than zero, or that a photon does not arrive after we detect it.
  - Prior information can also include statistical information, e.g., we already have some prior knowledge of the spatial distribution of stars, or of the value of the Hubble constant, or other information.
  - Sensitive dependence on reasonable prior information indicates that no analysis, Bayesian or other, can give reliable results.

Basic Method

- In a nutshell, a Bayesian analysis entails the following steps:
  - Choose prior distributions ("priors") that reflect your knowledge about each parameter and model prior to looking at the data
  - Determine the *likelihood function* of the data under each model and parameter value
  - Compute and normalize the full posterior distribution, conditioned on the data, using Bayes' theorem
  - Derive summaries of quantities of interest from the full posterior distribution by marginalization and/or computation of means.

Priors

- Choose prior distributions ("priors") that reflect your knowledge about each parameter and model prior to looking at the data
  - There is always some prior information about the problem. For example, we cannot count a negative number of photons. Parallaxes are greater than zero. We now know that the most likely value of the Hubble constant is in the ballpark of 60-80 km/sec/mpc (say) with smaller probabilities of its being higher or lower.
  - In Bayesian analysis, ones uncertainty about an unknown quantity is expressed by setting a *prior distribution* on the quantity in question, e.g., $p(\theta \mid B)$, where B is further background information.

Likelihood Function

- Determine the *likelihood function* of the data under each model and parameter value.
  - The likelihood function describes the statistical properties of the mathematical model of our problem. It tells us how the statistics of the observations (e.g., normal or Poisson data) are related to the parameters and any background information.
  - It is nothing but the sampling distribution for observing the data, given the parameters, but we are interested in its functional dependence on the parameters:

$$L(\theta; y \mid B) \propto p(y \mid \theta, B)$$

  - The likelihood is known up to a constant but arbitrary factor which cancels out in the analysis.

Posterior Distribution

- Compute and normalize the full posterior distribution, conditioned on the data, using Bayes' theorem.
  - The posterior distribution encodes what we know about the parameters and model after we observe the data. Thus, Bayesian analysis models learning .
  - Bayes' theorem says that

$$p(\theta \mid y, B) = \frac{p(y \mid \theta, B) p(\theta \mid B)}{p(y \mid B)}$$

  - Bayes' theorem is a trivial result of probability theory. The denominator is just a normalization factor and can often be dispensed with

$$p(y \mid B) = \int p(y \mid \theta, B) p(\theta \mid B) d\theta$$

Bayes' Theorem (Proof)

- By standard probability theory,

$$p(\theta \mid y, B) p(y \mid B) = p(\theta, y \mid B) = p(y \mid \theta, B) p(\theta \mid B)$$

from which Bayes' theorem follows immediately.

Posterior Distribution

- The posterior distribution after observing data y can be used as the prior distribution for new data z , which makes it easy to incorporate new data into an analysis based on earlier data.
- It can be shown that any *coherent* model of learning is equivalent to Bayesian learning.

Marginalization

- Derive summaries of quantities of interest from the full posterior distribution by marginalization and/or computation of means.
  - Bayesian methodology provides a simple and uniform way of handling nuisance parameters that are required by the analysis but are of no interest to us. We simply integrate them out (marginalize them) to obtain the marginal distribution of any parameter(s) we are interested in:

$$p(\theta_1 \mid y, B) = \int p(\theta_1, \theta_2 \mid y, B) d\theta_2$$

Bayesian Model Selection/Averaging

- Given models M_i, which depend on a *vector* of parameters ϑ, and given data Y, Bayes' theorem tells us that

$$p(\vartheta, M_i \mid Y) \propto p(Y \mid \vartheta, M_i)\, p(\vartheta \mid M_i)\, p(M_i),$$

  where the proportionality constant is chosen so that the left hand side is a normalized probability. The probabilities $p(\vartheta \mid M)$ and $p(M)$ are the prior probabilities of the parameters given the model and of the model, respectively; $p(Y \mid \vartheta, M)$ is the likelihood function, and $p(\vartheta, M \mid Y)$ is the joint posterior probability distribution of the parameters and models, given the data.
- Note that some parameters may not appear in some models, and there is no requirement that the models be nested.

Bayesian Model Selection

- We assume for the moment that we have supplied priors and performed the necessary integrations to produce a normalized posterior distribution. In practice this is often done by simulation using Markov Chain Monte Carlo (MCMC).
- Once this has been done, it is simple in principle, if more difficult in practice, to compute posterior probabilities of the models:

$$p(M_i \mid Y) = \int p(\vartheta, M_{i,} \mid Y)\, d\vartheta$$

- This set of numbers summarizes our degree of belief in each of the models, after looking at the data. If doing model selection, we choose the model with the highest posterior probability

Bayesian Model Averaging

- Suppose that one of the parameters, say ϑ_1, is common to all models and is of particular interest. For example, in the present application it might be the distance to a star. Then instead of choosing the distance as inferred from the most probable model, it may be better (especially if the models are empirical) to compute its marginal probability density over all models and other parameters:

$$p(\vartheta_1 \mid Y) = \sum_i \int p(\vartheta_1, \ldots \vartheta_n, M_{i,} \mid Y)\, d\vartheta_2 \ldots d\vartheta_n$$

- Then, if we are interested in an estimate of ϑ_1 we can (for example) compute its posterior mean and variance:

$$\hat{\vartheta}_1 = \int \vartheta_1 p(\vartheta_1 \mid Y)\, d\vartheta_1, \ \mathrm{Var}(\vartheta_1) = \int (\vartheta_1 - \hat{\vartheta}_1)^2\, p(\vartheta_1 \mid Y)\, d\vartheta_1$$

Practical Application

- A major difficulty has been to carry out the integrals required to do the computations in practice, limiting the method to situations where exact results can be obtained analytically or approximately

Practical Application

- The first of these is no longer considered a serious problem by most statisticians. Not only is classical ("frequentist") statistical methodology also shot through with subjective decisions (though they are better disguised and more arbitrary than in Bayesian methodology, where the subjectivity is summarized publicly in the prior), but also most classical results (excepting p-values) have straightforward Bayesian interpretations using standardized "reference" or "automatic" priors.
- It is quite common, even amongst statisticians who consider themselves frequentists, to use Bayesian methods when they are more convenient (often the case) or provide capabilities unavailable to frequentist methods (also often the case, e.g., when incorporating prior information).

Practical Application

- Considerable progress has been made in the past decade in solving the computational difficulties, particularly with the development of Markov Chain Monte Carlo (MCMC) methods for simulating a random sample (draw) from the full posterior distribution, from which marginal distributions and summary means and variances (as well as other averages) can conveniently be calculated.
- These methods have their origin in physics. The Metropolis-Hastings and Gibbs sampler methods are two popular schemes that originated in early attempts to solve large physics problems by Monte Carlo methods.

Practical Application: Markov Chains

- Start from an arbitrary point in the space of models and parameters. Following a specific set of rules, which depend only on the *unnormalized* posterior distribution, generate a random walk in this space, such that the distribution of the generated points converges to a distribution drawn from the underlying probability distribution.
- The random walk is a *Markov Chain*: That is, each step depends only upon the immediately previous step, and not on any of the earlier steps.
- Many rules for generating the transition from one state to the next are possible. All converge to the same distribution. One attempts to choose a rule that will give efficient sampling with a reasonable expenditure of effort and time.

Gibbs Sampler

- The Gibbs Sampler is a scheme for generating a sample from the full posterior distribution by sampling in succession from the conditional distributions. Thus, let the parameter vector θ be decomposed into a set of subvectors $\theta_1, \theta_2, \ldots, \theta_n$. Suppose it is possible to write the conditional distributions

$$p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_n),$$
$$p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_n),$$
$$\ldots$$
$$p(\theta_n | \theta_1, \theta_2, \ldots, \theta_{n-1}).$$

Gibbs Sampler (2)

- Starting from an arbitrary initial vector

  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)})$,

  generate in succession vectors $\theta^{(1)}, \theta^{(2)}, \ldots$ by sampling in succession from the conditional distributions:

  $p(\theta_1^{(k)} | \theta_2^{(k-1)}, \theta_3^{(k-1)}, \ldots, \theta_n^{(k-1)})$,

  $p(\theta_2^{(k)} | \theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_n^{(k-1)})$,

  …

  $p(\theta_n^{(k)} | \theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{n-1}^{(k)})$, with

  $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_n^{(k)})$.
- In the limit, the sample thus generated will converge to a sample drawn from the full posterior distribution.

Gibbs Sampler (Example)

- Suppose we have normally distributed estimates y_i, $i=1,\ldots,N$, of a parameter x, with unknown variance σ. The likelihood is

  $p(y|x,\sigma) \sim \sigma^{-N}\exp(-\Sigma(y_i-x)^2/2\sigma^2)$
- Assume a flat (uniform) prior for x and a "Jeffreys" prior $1/\sigma$ for σ. The posterior is proportional to prior times likelihood:

  $p(x,\sigma|y) \sim \sigma^{-(N+1)}\exp(-\Sigma(y_i-x)^2/2\sigma^2)$
- The conditional distributions are, for x, a normal distribution with mean equal to the average of the y's and variance equal to σ^2 (which is known at each Gibbs step), and σ^2 proportional to a sample from an inverse chi-square distribution with N-1 degrees of freedom (divide $\Sigma(y_i-x)^2$ by a sample from a standard chi-square distribution). Again note that when sampling for σ^2, x will be known at each Gibbs step.

Metropolis-Hastings Step

- The example is simple because the conditional distributions are all standard distributions from which samples can easily be drawn. This is not usually the case, and we would have to replace Gibbs steps with another scheme.
- A Metropolis-Hastings step involves producing a sample from a suitable *proposal distribution* $q(\theta^*|\theta)$, where θ is the value at the previous step. Then a calculation is done to see whether to accept the new θ^* as the new step, or to keep the old θ as the new step. If we retain the old value, the sampler does not "move" the parameter θ at this step. If we accept the new value, it will move.
- We choose q so that it is easy to generate random samples from it, and with other characteristics.

Metropolis-Hastings Step (2)

- Specifically, if $p(\theta)$ is the target distribution from which we wish to sample, first generate θ^* from $q(\theta^*|\theta)$.
- Then calculate

  $\alpha = \min\{1, (p(\theta^*)\, q(\theta|\theta^*))/(p(\theta)\, q(\theta^*|\theta))\}$
- Then generate a random number r uniform on [0,1]
- Accept the proposed θ^* if $r \le \alpha$, otherwise keep θ.
- Note that if $p(\theta^*) = q(\theta^*|\theta)$ for all θ, θ^*, then we will always accept the new value. In this case the Metropolis-Hastings step becomes an ordinary Gibbs step.
- Generally, although the Metropolis-Hastings steps are guaranteed to produce a Markov chain with the right limiting distribution, one gets better performance the closer we can approximate $p(\theta^*)$ by $q(\theta^*|\theta)$.

Mathematical Model of Cepheid Problem

- We model the radial velocity and V-magnitudes as Fourier polynomials of unknown order K, where ϑ is the phase. Thus, for the velocities:

$$v_r = \overline{v_r} + \Delta v_r \text{ where}$$

v_r is the observed radial velocity

$\overline{v_r}$ is the mean radial velocity and

$$\Delta v_r = \sum_{j=1}^{K} (a_j \cos j\vartheta + b_j \sin j\vartheta)$$

- A major problem is to choose the optimal number of coefficients in the Fourier polynomial.

Mathematical Model

- The Δ –radius of the star is the integral of the Δ– radial velocity:

$$\Delta r = -f \sum_{j=1}^{K} (a_j \sin j\vartheta_j - b_j \cos \vartheta_j) / j$$

where f is a positive numerical factor.

- The relationship between the radius and the photometry is given by

$$V = 10(C - (A + B(V - R) - 0.5 \log_{10}(\phi_0 + \Delta r / s)))$$

where the V and R magnitudes are corrected for reddening, A, B and C are known constants, ϕ_0 is the angular diameter of the star and s is the distance to the star

The Traditional Approach

- Observe velocity curve, smooth by eye or Fourier series, integrate to obtain radial displacement, do a least squares fit by predicting $\phi_0 + \Delta r / s$ from observed flux and color.
- This approach is statistically inadequate.
  - Fitting of velocities is *ad hoc.* If eyeball, how well fit? If Fourier series, how many terms to take? We have a *model selection/averaging* problem.
  - There is error in the independent variable $\Delta r = -f \int \Delta v_r \, d\vartheta$ and color index: This is an *errors-in-variables* problem.
  - Therefore, the solution for s may be biased, and its error will be underestimated. A simultaneous errors-in-variables solution of velocity and photometry data would be preferable.

Application to Cepheid Variables

- We have applied a Bayesian analysis to the calculation of Cepheid distances and radii using the surface brightness (Baade-Wesselink) method. Our model correctly treats the errors-in-variables aspect of our data (both x and y data appear with errors), and includes full model averaging with respect to the empirical Fourier series used to represent the velocity and photometry data.
- Based on a representative sample of eight stars, our new analyses support the distance scale of Gieren, Barnes, & Moffett (1993, *ApJ*, **418**, 135) and do not show bias in the calculation of those distances suggested by Laney & Stobie (1995, *MNRAS*, **274**, 337). Tom Barnes will discuss this aspect of our work.

Sample Run

Significant Issues on Priors

- The priors on the Fourier coefficients must be chosen carefully. If too vague, significant terms may be rejected. If too sharp, overfitting may result. For our models we have used a Zellner G-prior, which is equivalent to a Maximum Entropy prior, of the form

$$p(a) \propto \exp(-a'X'Xa/2\sigma^2),$$

where a is the vector of Fourier coefficients, X is the design matrix of sines and cosines for the problem, and σ is an arbitrary parameter which gets its own prior distribution (technical details beyond the scope of this discussion).

Significant Issues on Priors

- Cepheids are part of the disk population of the galaxy, and for low galactic latitudes are more numerous at larger distances s. So distances calculated by MLE or with a flat prior will be affected by Lutz-Kelker bias, which can amount to several percent.
- The classical way to understand the Lutz-Kelker bias is that since the density of stars increases with distance, it is more likely that we have a star a bit farther away with a negative error that brings it to the observed distance than that we have a closer star with a positive error that pushes it further out to the observed distance.

Significant Issues on Priors

- The Bayesian approach is simply to recognize that our prior distribution on the distance of stars depends on the distance (for a uniform distribution it would be proportional to $s^2 ds$).
- In our problem we choose a spatial distribution of stars that is exponentially stratified as we go away from the galactic plane. We adopted a scale height of 97±7 parsecs, and sampled the scale height as well. Our prior on the distance looks like

$$p(s) \sim \rho(s) s^2 ds,$$

where $\rho(s)$ is the spatial density of stars.

For an exponential distribution we have

$$\rho(s) \sim \exp(-|z|/z_0),$$

where z_0 is the scale height, $z = s \sin\beta$, and β is the latitude of the star.

Other Fully Bayesian Approaches

- Since going to Duke on sabbatical, I and my Duke collaborators have been studying other approaches to this problem. These are all still “in work”.
  - We are looking at other functional forms to represent the velocity and photometry data. In particular, the velocity curve suffers a steep drop-off near $\vartheta=1$; a more local representation using wavelets looks very promising.
  - We are investigating other priors on the coefficients that have been used successfully in other contexts. One promising approach is the “Expected Posterior Prior” developed by Jim Berger with one of his students.

Discussion of Results

- Use of our rigorous mathematical method does not change the distances, radii, mean velocities, or optimal phase shifts compared to the simple method used in Gieren *et al.* 1993, *ApJ*, **418**, 135.
- Our method effectively and objectively selects and averages over models with the optimal number of terms in the Fourier series for the velocities and photometry. Each model contributes appropriatedly to the final result.
- The Bayesian method suggests natural ways to account for statistical effects such as the Lutz-Kelker effect by regarding them as resulting from incorrect choices of the prior.

Discussion of Results

- Bayesian methods show extraordinary promise for obtaining solutions of complex statistical problems in astronomy. They are likely to become an important tool in the astronomical toolbox.

Beowulf

- Note that Beowulf would be an ideal platform for doing MCMC calculations, as it could run a number of independent Markov chains. This would greatly speed up the calculations as well as improving the statistics of the sampling.