Statistics for Twenty-first Century Astrometry

William H. Jefferys
University of Texas at Austin, USA

Abstract

H.K. Eichhorn had a lively interest in statistics during his entire scientific career, and made a number of significant contributions to the statistical treatment of astrometric problems. In the past decade, a strong movement has taken place for the reintroduction of Bayesian methods of statistics into astronomy, driven by new understandings of the power of these methods as well as by the adoption of computationally-intensive simulation methods to the practical solution of Bayesian problems. In this paper I will discuss how Bayesian methods may be applied to the statistical discussion of astrometric data, with special reference to several problems that were of interest to Eichhorn.

Bayesian Analysis and Astronomy

- Bayesian methods offer many advantages for astronomical research and have attracted much recent interest.
  - Astronomy and Astrophysics Abstracts lists 117 articles with the keywords ‘Bayes’ or ‘Bayesian’ in the past 5 years, and the number is increasing rapidly (there were 33 in 1999 alone).
  - At the June 1999 AAS meeting in Chicago, there was a special session on Bayesian and Related Likelihood Techniques. Another session at the June 2000 meeting will also feature Bayesian methods.

Advantages of Bayesian Methods

- Bayesian methods allow us to do things that would be difficult or impossible with standard (frequentist) statistical analysis.
  - It is easy to incorporate prior physical or statistical information
  - Results depends only on what is actually observed, not on observations that might have been made but were not.
  - We can compare models and average over both nested and unnested models.
  - Correct interpretation of results is very natural, especially for physical scientists.

Advantages of Bayesian Methods

- It is a *systematic* way of approaching statistical problems, rather than a collection of ad hoc techniques. Very complex problems (difficult or impossible to handle classically) are straightforwardly analyzed within a Bayesian framework.
- Bayesian analysis 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.
- With proposed astrometric missions (e.g., FAME) where the signal can be very weak, analyses based on normal approximations may not be adequate. Bayesian analysis of Poisson-distributed data, for example, may be a better choice.

Basic Method

- In a nutshell, Bayesian analysis entails the following systematic 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 integrating over the posterior distribution to produce marginal distributions or integrals of interest (e.g., means, variances).

Priors

- Choose prior distributions ("priors") that reflect your knowledge about each parameter and model prior to looking at the data
  - The investigator is *required* to provide *all* relevant prior information that he has before proceeding
  - There is *always* prior information. 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.
  - Prior information can be statistical in nature, e.g., we may have statistical knowledge about the spatial or velocity distribution of stars, or the variation in a telescope's plate scale.

Prior Information

- In Bayesian analysis, our knowledge about an unknown quantity is encoded in a *prior distribution* on the quantity in question, e.g., $p(\theta|B)$, where B is background information.
- Where prior information is vague or uninformative, a vague prior generally recovers results similar to a classical analysis.
  - » EXCEPTION: Model selection/model averaging
- Sensitive dependence of the result on reasonable variations in prior information indicates that no analysis, Bayesian or other, can give reliable results.

Prior Information

- The problem of prior information of a statistical or probabilistic nature was addressed in a classical framework by Eichhorn (1978: "Least-squares adjustment with probabilistic constraints," MNRAS **182**,366-360) and by Eichhorn and Standish (1981: "Remarks on nonstandard least-squares problems," AJ **86**, 156-159).
- They considered adjusting astrometric data given prior knowledge about some of the parameters in the problem, e.g., that the plate scale values only varied within a certain dispersion.
- For the cases studied in these papers (multivariate normal distributions), their result is similar to the Bayesian one (but the interpretation is different).

Prior Information

- In another example, Eichhorn and Smith studied the Lutz-Kelker bias (1996: MNRAS **281**, 211-218).
- The classical way to understand the Lutz-Kelker bias is that we are more likely to have observed a star a slightly farther away with a negative error that brings it in to the observed distance, than we are to have observed a slightly nearer star with a positive error that pushes it out to the observed distance, because the number of stars increases with increasing distance.
- The Bayesian notes that is more likely *a priori* that a star of unknown distance is farther away than that it is nearer. The mathematical analysis gives a similar result, but the Bayesian approach, by demanding at the outset that we think about prior information, inevitably leads us to consider this phenomenon, which classical analysis missed for a century.

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 proportional to the sampling distribution for observing the data *Y*, 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.

Likelihood Function

- Like Bayesian estimation, maximum likelihood estimation (upon which Eichhorn based many of his papers) is founded upon using the likelihood function.
- This is good, because the likelihood function is always a *sufficient statistic* for the parameters of the problem.
- However, it is not the whole story
  - Maximum likelihood does not take prior information into account
  - It fails in some notorious situations, like errors-in-variables problems where the variance of the observations is estimated (Bayesian analysis gets the right answer; classical analysis relies on an *ad hoc* factor of 2 correction)
  - There are other difficulties as well

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

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

  - 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. Thus in Bayesian analysis
  - Results take into account all known information
  - Results do not depend on the order in which the data (e.g, Y and Z) are obtained
  - Results are consistent with common sense *inductive* reasoning as well as with standard *deductive* logic, e.g., if A entails B, then observing B should support A (inductively), and observing ~B should refute A (logically)

Integration and Marginalization

- Derive summaries of quantities of interest from the full posterior distribution by integrating over the posterior distribution to produce marginal distributions or integrals of interest (e.g., means, variances).
  - Bayesian methodology provides a *simple* and *systematic* way of handling nuisance parameters required by the analysis but which are of no interest to us. We simply integrate them out (marginalize them) to obtain the marginal distribution of the parameter(s) of interest:

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

  - Likewise, computing summary statistics is simple : e.g., posterior means and variances

$$\bar{\theta}_1 \mid Y, B = \int \theta_1 p(\theta_1, \theta_2 \mid Y, B)d\theta_1 d\theta_2$$

The Problem of Model Selection

- Eichhorn and Williams (1963: "On the systematic accuracy of photographic astrometric data," AJ **68**, 221-231) studied the problem of choosing between competing astrometric models. Often the models are empirical, e.g., polynomial expansions in the coordinates.
- The problem is to avoid the Scylla of underfitting the data, resulting in a model that is inadequate, and the Charybdis of overfitting the data.
- Navigating between these dangerous shoals is by no means trivial, and standard statistical methods such as the F-test and stepwise regression are not to be trusted (they too easily reject adequate models in favor of overly complex ones).

The Problem of Model Selection

- Eichhorn and Williams proposed a criterion based on trading off the decrease in average residual against the increase in average error introduced by the plate constants.
- The Bayesian approach reveals how these two effects should be traded off against each other, producing a sort of Bayesian Ockham's razor that favors the simplest adequate model.
- Eichhorn and Williams' basic notion is sound; but in my opinion the Bayesian approach to this problem is simpler and more compelling, *and*
  - It is not limited to nested models
  - It allows for *model averaging*, unavailable with any classical approach.

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

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

- 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

- 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 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, it could 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 summary statistics on ϑ_1 we can (for example) compute its posterior mean and variance:

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

Practical Application

- Until recently, a major practical difficulty has been computing the required integrals, limiting the method to situations where exact results can be obtained with analytic approximations
- In the past decade, considerable progress has been made 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 be calculated conveniently.
- 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 sample drawn from the underlying posterior 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 sample from the conditional distributions

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

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 X_i, $i=1,\ldots,N$, of a parameter x, with unknown variance σ. The likelihood is
  $p(X|x,\sigma) \sim \sigma^{-N} \exp(-\Sigma(X_i-x)^2/2\sigma^2)$
- Assume a flat (uniform) prior for x and a "Jeffreys" prior $1/\sigma$ for σ. The posterior is proportional to prior times likelihood:
  $p(x,\sigma|X) \sim \sigma^{-(N+1)} \exp(-\Sigma(X_i-x)^2/2\sigma^2)$
- The conditional distributions are: for x, a normal distribution with mean equal to the average of the X's and variance equal to σ^2 (which is known at each Gibbs step); and for σ^2, an inverse chi-square distribution with N-1 degrees of freedom.

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 $q(\theta^*|\theta)$.

Example: Cepheid Distances

- With T.G. Barnes of McDonald Observatory and J.O. Berger and P. Müller of Duke University's Institute for Statistics and Decision Sciences, I have been working on a Bayesian approach to the problem of estimating distances to Cepheid variables using the surface-brightness method.
- We use photometric data in several colors as well as Doppler velocity data on the surface of the star to determine the distance and absolute magnitude of the star.
- The problem, although not an astrometric problem *per se*, is nonetheless a good example of the application of Bayesian ideas to problems of this sort and illustrates several of the points made earlier (prior information, model selection, model averaging).

Cepheid Distances

- We model the radial velocity and V-magnitude of the star as Fourier polynomials of unknown order, 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

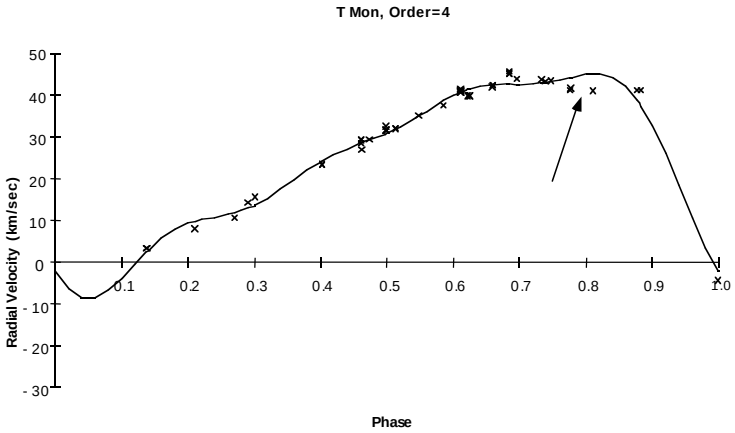
$\overline{v_r}$ is the mean radial velocity and

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

- This becomes a model selection/averaging problem because we want to use the optimal number of coefficients and/or we want to average over models in an optimal way.

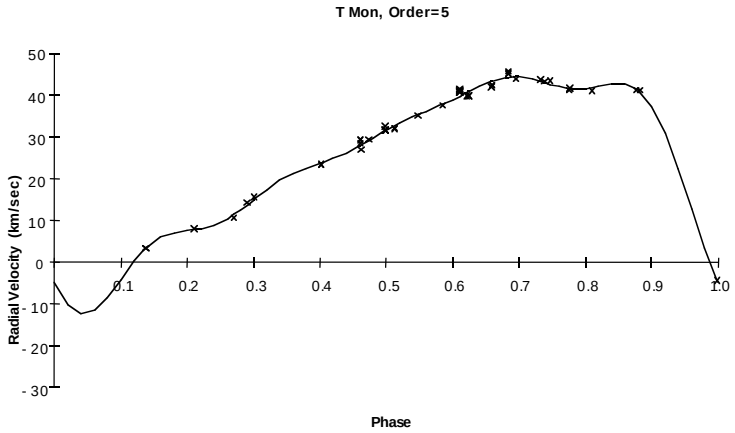
Velocity Fit, Fourth Order

- Arrow shows physically real "glitch"



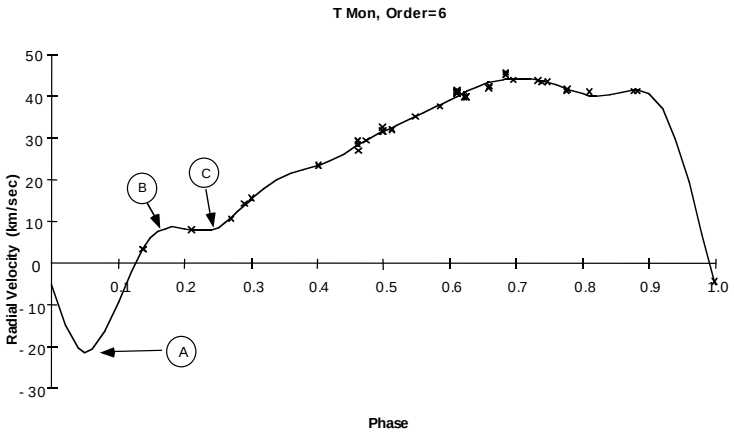
Velocity Fit, Fifth Order

- Seems to be a good fit.



Velocity Fit, Sixth Order

- Arrows (particularly A) show evidence of overfitting



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.

Cepheid Distances

- The resulting model is fairly complex, simultaneously estimating a number of Fourier coefficients and nuisance variables (up to 40 variables) for a large number of distinct models, along with the variables of interest (distance and absolute magnitudes).
- The Markov chain provides a sample drawn from the posterior distribution for our problem as a function of all of these variables, including model specifier.
- From it we obtain very simply the marginal distributions of parameters of interest as the marginal distributions of the sample, and means and variances of parameters (or any other desired quantities) as sample means and sample variances based on the sample.

Sample Run

- Note: Here in the talk I showed a selection of charts based on the cepheid variable simulations. A copy of the charts can be found as a .pdf file associated with this lecture on my web page. It is a large download, about 2.5 MB, so be warned!

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 Bayesian solution is to recognize that our prior distribution on the distance of stars depends on the distance
  - For a uniform distribution it would be proportional to $s^2 ds$, which although an improper distribution, gives a reasonable answer if the posterior distribution is normalizable.

Significant Issues on Priors

- In our problem we chose 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 is

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

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

Significant Issues on Priors

- The priors on the Fourier coefficients must be chosen carefully. If they are too vague, significant terms may be rejected. If too sharp, overfitting may result.
- For our models we have used 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 a parameter to estimated.

- This maximum entropy prior expresses the proper degree of ignorance about the Fourier coefficients.

Conclusions

- Bayesian analysis is a promising statistical tool for discussing astrometric data.
- It requires us to think clearly about prior information, e.g., it naturally requires us to consider the Lutz-Kelker phenomenon from the outset, and tells us how to build it into the model using our knowledge of the spatial distribution of stars
- It effectively solves the problem of accounting for competing astrometric models by Bayesian model averaging
- We can expect Bayesian and quasi-Bayesian methods to play important roles in missions such as FAME, which challenge the state of the art of statistical technology