Statistics Perspective

James O. Berger
Duke University
Statistical and Applied Mathematical Sciences Institute

SAMSI Astrostatistics Program
(with CASt; indeed Jogesh was the Program Leader)

- Tutorials
- Workshops
- Working Groups

Tutorials (January 18-22, 2006): Soon to be on the web!

Bayesian Astrostatistics: This three-day session - taught by Tom Loredo, Bill Jefferys and Philip Gregory - covered the basic theory and practice of Bayesian statistics, using examples from astronomy.

Nonparametric statistics and machine learning: This two-day tutorial, taught by Chad Schafer and Larry Wasserman, introduced astronomers to modern methods in nonparametric statistics.

Astronomy for statisticians: This two-day tutorial, taught by Bill Jefferys and Eric Feigelson, reviewed modern understanding of our universe and raised key statistical issues underlying astronomical studies.

Working Groups and Intensive Study Sessions

Exoplanets: Discussed in the presentations of Eric Ford, Merlise Clyde and Bill Jefferys.

Surveys and Population Studies: Discussed in the presentation of Tom Loredo and Woncheol Jang.

Source Detection and Feature Detection: Discussed in the presentations of Alanna Connors, David van Dyk, Rebecca Willet.

Gravitational Lensing: led by Arlie Petters, studied

- magnification of probability distributions with applications to dark matter substructures on galactic scales;
- statistical methods in cluster mass reconstruction;
- statistics of image counting in stochastic microlensing;
- applications of spatial statistics to the distribution of dark matter structures.

Intensive Session on Statistical Issues in Particle Physics: Discussed in the presentations of Louis Lyons? and Michael Woodroofe.

Intensive Session on Stellar Evolution: led by Bill Jefferys and studied improving MCMC sampling, handling field stars, and (very successfully) handling binary stars.

Workshops

Planning meeting: NASA-Ames, July 14-15, 2005, organized by Jeff Scargle.

Some important themes that came up that we didn't pursue:

- How do we summarize data so that we can best use the conclusions for future scientific investigations?
- Complex computer models/data interface; identifying and adjusting for systematic errors

Opening Workshop: SAMSI, January 23-25, 2006, with primary purpose of forming and advising the working groups.

Closing Workshop: Here!

Reprise/Reunion Events: Possible over the next 2 years.

Future SAMSI Programs

Multiplicity and Reproducibility in Scientific Studies: July 10-28, 2006

High Dimensional Inference and Random Matrices: Fall, 2006, with a Machine Learning component in Spring, 2007.

Development, Assessment and Validation of Computer Models: Fall, 2006 and Spring, 2007

Who is in a SAMSI program?

- Postdoctorals
- Many visiting and local faculty and scientists
- Graduate students
- Short term visitors and workshop participants

www.samsi.info

The Current Status of Astrostat

- Many astronomers and astrophysicists are doing (and developing) very good statistics.
- Some statisticians (not enough) are heavily involved with astronomers and astrophysicists.
- Technology transfer is happening (and meetings like this and the various tutorials and schools are a great help).
- Lacking, however, is a regular structure for forming a 'team environment,' where statisticians are involved in major astronomical projects from the beginning. Barriers include
  - Astronomical funding mechanisms (little money for statistics at all, much less for statisticians)
  - Shortage of statisticians
  - The 'astronomers can figure it out for themselves' syndrome.

A Few Commonly Occurring Science Themes

- 'Doing it right' versus 'facing up to reality.'
- Miscellaneous Bayesian Comments
- Basics of Bayesian hypothesis testing
- Comments on model selection
- Nonparametrics

‘Doing it right’ versus ‘facing up to reality.’

Alternative descriptions of ‘Doing it Right:’

- Alanna Connors and David van Dyk: Put in physical/statistical models of structure at the beginning, and try to carry them through all the data accumulation and processing stages.
- Tom Loredo: forward analysis, leading to true likelihood analysis.

Alternative descriptions of ‘Facing up to Reality.’

- Tim Axelrod: “swimming upstream”
- Robert Lupton: “sleazy analysis”

My view:

- The forward process gets uncertainties right and reduces possibilities of bias, and so is a great way to begin with a problem.
- The forward process is often tough; thank heavens for ‘sleaze.’

Miscellaneous Bayesian Comments

- Bayes/MCMC is principled, and realistically accounts for uncertainty; uncertainties from other analyses are rarely large enough. (Gary Hinshaw: "half of all 3 sigma results are wrong.")
- Worries about priors
  - Proper subjective priors are often tough to come by: true!
  - And statisticians say "flat priors are evil" (Harrison Prosper). Not so! Objective Bayes has a wonderful 240 year history, and is the dominant practical Bayesian philosophy today. Two prominent objective Bayesian approaches are
    - The maximum entropy approach (especially useful when you know moments of the system)
    - The reference prior approach: Example: For a Poisson(λ) distribution, $\pi(\lambda) = \lambda^{-1/2}$ is the reference prior.

- Bayes is often viewed as not being doable on really complex problems (e.g. Gary Bernstein's weak gravitational lensing).
  - MCMC's can often be made amazingly efficient (with work).
  - There are a variety of approximations (e.g. the Laplace approximation, as mentioned by Jiayang Sun).
  - 'Fiddles' are fair if absolutely needed (e.g., Christopher Kochanek did forward Bayes, but needed to use maximum likelihood estimates of nuisance parameters in MCMC loops).
  - New very fast approximations such as *variational methods* from the machine learning community can work with very large problems.
- 'Max-Ent Bayesian' versus 'Statistical Bayesian' terminology:
  - posterior likelihood $\rightarrow$ posterior density
  - evidence $\rightarrow$ marginal likelihood
  - ...

Basics of Bayesian Hypothesis Testing

Key issue 1: Is the hypothesis being tested believable?

Example 1. Gary Hinshaw, in discussing WMAP, posed the question of whether the inflation parameter n_s is < 1. Is $n_s = 1$ a believable hypothesis so that one is testing

$$H_0 : n_s = 1 \text{ versus } H_1 : n_s \neq 1,$$

or is it not particularly plausible in which case one is, say, testing

$$H_0 : n_s < 1 \text{ versus } H_1 : n_s \geq 1?$$

Example 2. Glen Cowan's physics test of $H_0 : s = 0$, where s is the mean signal. Here $s = 0$ is clearly plausible (e.g., no Higgs).

Key issue 2: There is no need to assign prior probabilities to hypotheses; one can give Bayes factors, and one can even give useful bounds on Bayes factors that are completely independent of priors.

Example: Observe $n =$ Poisson$(s + b)$, $b =$ background mean.

Test: $H_0 : s = 0$ versus $H_1 : s > 0$.

P-value: $P(N \geq n \mid b, s = 0)$ (=0.0014 if $n = 5$ and $b = 0.8$).

Bayes factor of H_0 to H_1: $B_{01} = \frac{b^n \ e^{-b}}{\int_0^\infty (s+b)^n \ e^{-(s+b)} \pi(s) \ ds}$.

Subjective approach: Choose $\pi(s)$ subjectively (e.g., using the standard model predictions of the mass of the Higgs).

Absolute lower bound on the Bayes factor: choose $\pi(s)$ to be a point mass at $\hat{s}$, yielding (the observed likelihood ratio)

$$B_{01} \geq \min\{1, \left(\frac{b}{n}\right)^n e^{n-b}\} \quad (=0.007 \text{ if } n = 5 \text{ and } b = 0.8).$$

An objective lower bound on the Bayes factor: more reasonable (at least for near-Gaussian problems) is

$$B_{01} \geq -e \ p \ \log p \quad (=0.025 \text{ if } n = 5 \text{ and } b = 0.8).$$

Comments on Model Selection

- Bayesian model selection (talks by Eric, Bill, Merlise Clyde, Graham Woan) applies to models of any types (e.g. Chris Koen's)
- Chris Genovese discussed the difference between AIC (most useful for prediction) and BIC (useful for ascertaining the 'true model.')
- But is BIC $\equiv 2l(\hat{\boldsymbol{\theta}}) - p \log n$ a good approximation to a Bayesian answer? The full Bayesian marginal likelihood of a model is $m(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = c_\pi e^{BIC/2}(1 + o(1))$, as $n \to \infty$.
  - n is often ill-defined (e.g. with binning, as Gary noted).
  - p is often ill-defined, and can grow with n.
  - The constant c_π is often very far from 1.
- Harrison Prosper: is Bernardo's discrepency measure useful? Yes, if one is trying to decide if a smaller model is adequate as an approximation to a larger model; it is not designed for 'discovery.'

- What about False Discovery Rate as a tool for dealing with a huge number of possible 'discoveries'? (Istvan had 1.6 million.)
  - As mentioned by Chris Genovese, FDR is only useful for controlling FDR, not for establishing a 'discovery.'
  - Indeed, controlling Family Wide Error Rate by usual methods gives greater power for 'proving' that a discovery exists (e.g. Chris Koen's testing for a significant peak in periodograms).
  - FDR is reasonable for obtaining a set of candidates for further study, although it lacks a decision-theoretic justification.
  - FDR is quite conservative, unless one incorporates estimates of the proportion of discoveries that exist, which is difficult - although newer methods (mentioned by Chris Genovese, Laura Cayon, and John Rice) attack this issue.
  - Bayesian approaches exist and can have great power (e.g., Graham Woan's talk); multiplicity is automatically handled.

Nonparametrics

- When astrophysics gives a parametric model, great! (Eric)
- When not, consider nonparametrics; several different cases arose:
  - Nonparametric function estimation: often this is estimating a nice function, about which much may be known, but not an exact functional form.
  - Dealing with a mess, like cosmic structures (Vicent Martinez and Ofer Lahav).
  - Nonparametric testing (William Romanishin and John Rice)

Nonparametrics for nice functions

- Usual nonparametrics tries to accommodate nasty functions, and can give conservative answers when applied to nice functions.
- Imposing constraints on the functions, such as monotonicity or convexity (arising from astrophysics), can be much better (as in

Michael's talk, Martin's disc?). There are also Bayesian methods.

Nonparametrics for messes: There are many tools such as spatial tools (Adrian Baddeley's talk), clustering tools, ...

One I haven't heard discussed here: *Dirichlet process mixtures* for clustering (Mike West originally, and lately a big segment of the machine learning community, e.g. Michael Jordan).

A simple case: start with standard *finite normal mixtures:* vectors $\boldsymbol{y}_i, i = 1, \ldots, n$, are modeled by a mixture of k multivariate normal distributions, $\boldsymbol{y}_i \sim \sum_{j=1}^{k} w_j N(\boldsymbol{\mu}_j, \Sigma_j)$, where the mixture weights w_j are assigned a Dirichlet$(\alpha, \ldots, \alpha)$ distribution. Letting $k \to \infty$ and $\alpha \to 0$ at the right rates, one ends up with the nonparametric Dirichlet process mixture, a flexible clustering method with fast computation.

Nonparametric testing: Be at least a closet Bayesian to guide the test, and maybe it works to be a real Bayesian, but we need comparisons.