Bayesian Miscellany

Tom Loredo

Dept. of Astronomy, Cornell University

http://www.astro.cornell.edu/staff/loredo/

Bayesian Miscellany

- Decision theory
- Experimental design
- A few more algorithms
  - Savage-Dickey density ratio
  - Data augmentation algorithms
- Toolkits

A Fundamental $\mathcal{B}$–$\mathcal{F}$ Relationship

($\mathcal{B} =$ Bayes, $\mathcal{F} =$ Frequentist)

$\mathcal{B}$ is both more general and more narrow than $\mathcal{F}$

- More general: Can (in princ.) contemplate p of *anything*, not just "random variables"
- More narrow: No freedom in how data appear in inferences—always through the *likelihood*, $\mathcal{L}_M(\theta) \equiv p(D_{\text{obs}}|\theta, M)$

$\mathcal{F}$ can always base a procedure on a $\mathcal{B}$ calculation

Analogy: $\mathcal{B}$ as the "Lorentz invariance" of statistics, restricting focus on a *particular class* of procedures

This viewpoint has a formal justification. . . .

Decision

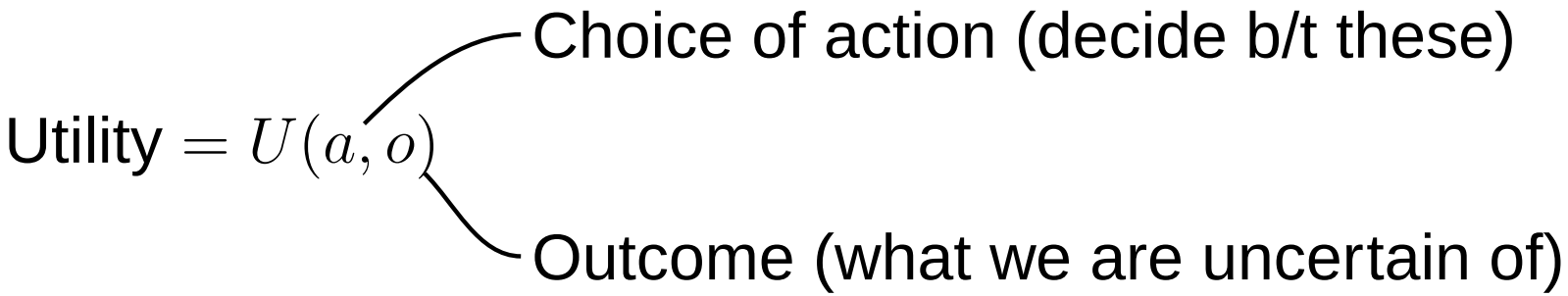
Acting Amidst Uncertainty

Decisions depend on consequences

Might bet on an improbable outcome provided the payoff is large if it occurs and the loss is small if it doesn't.

Utility and loss functions

Compare consequences via *utility* quantifying the benefits of a decision, or via *loss* quantifying costs.



Loss $L(a, o) = U_{\max} - U(a, o)$

Frequentist Decision Theory

Consider a *rule* $a(D)$ that chooses a particular action when D is observed.

Central quantity: *Risk*

$$R(o) = \sum_D p(D|o) L[a(D), o]$$

Seek rules with small risks for anticipated outcomes

Admissable rules: $a(D)$ is admissiable if there is no other rule that is at least as good for all o, *and* better for at least one o

Frequentist Inference and Decision

Inference: Frequentist calibration

In repeated practical use of a statistical procedure, the long-run average *actual* accuracy should not be less than (and ideally should equal) the long-run average *reported* accuracy (procedures should be calibrated).

Many procedures/rules can be created that are calibrated this way. How to choose among them?

Decision: Optimal Rules

- Devise a family of rules with desired performance
- Specify a loss function
- Find the rule with the “best” risk

Optimal $\mathcal{F}$ inference and decision are inseparable

Example—Parameter Estimation

Estimate a normal mean, μ, from N observations, $x = \{x_i\}$; σ known

Point and interval estimates

- $\bar{x}$ is best linear unbiased estimator (*BLUE*; squared-error loss)
- $I(x) = [\bar{x} \pm \sigma/\sqrt{N}]$ is shortest 68.3% *confidence interval*: $p(I(x)\,\text{covers}\,\mu|\mu) = 0.683$

Example—Testing

Is $\mu = 0$ (M_0, "null hypothesis) or $\mu \neq 0$?

Neyman-Pearson Significance test

- Procedure: Accept M_0 if $-2\sigma/\sqrt{N} < \bar{x} < 2\sigma/\sqrt{N}$
- Type I error probability $\alpha =$ "false alarm rate" $=$ 5%
- Uniformly most powerful (UMP) test (has smallest Type II error rate for any test with $\alpha = 0.05$ against $\mu \neq 0$ Normal alternatives)

Bayesian Decision Theory

We are uncertain of what the outcome will be

$\rightarrow$ average:

$$EU(a) = \sum_{\text{outcomes}} P(o|\ldots)\, U(a, o)$$

The best action maximizes the expected utility:

$$\hat{a} = \arg\max_a EU(a)$$

I.e., minimize expected loss

Inference and decision are separate stages in $\mathcal{B}$—can formulate and report inferences without considering decisions.

Well-Posed Inference Problems

Well-posed problems have unique solutions.

Both approaches require specification of models giving $p(D|\ldots)$. They differ in what *else* is needed.

Frequentist

- Primary measure of performance (bias, coverage, α)
- Family of procedures providing desired performance
- Loss function comparing different procedures with same primary measure (squared error, interval size, power)

Bayesian

- Information specifying priors

Wald's Complete Class Theorem

Abraham Wald, *Statistical Decision Functions* (1950):

Admissable decision rules are Bayes rules!

("Lorentz invariance" for statistics)

Little impact on $\mathcal{F}$ practice—an admissable rule can be "worse" than an inadmissable rule:

- Admissable: Estimate μ with $\hat{\mu} = 5$, *always*
- Inadmissable: Use $\hat{\mu} = (x_1 + x_N)/2$

Wald's theorem can eliminate many bad rules, but not all

Suggests $\mathcal{F}$ approach: Study $\mathcal{F}$ performance of classes of Bayes rules.

Suggests $\mathcal{B}$ approach: Identify "good" priors by studying frequentist performance of Bayes rules.

Historical perspective: Dennis Lindley

> [Fisherian methods] lacked the cohesion we expected of a mathematical discipline, where there was a set of axioms from which theorems could be proved. . . the theorems would include the mixed collection of results that we had acquired from the masters. . . and would provide a methodology whereby new theorems and useful implementations could be found. . .
>
> The real advance came with Savage (*The Foundations of Statistics*, 1954) who. . . accomplished everything we mathematicians had hoped for. . . But by 1971 the dream had shattered and Savage was to write in his second edition, more honestly than most scientists can manage, that his attempt. . . to justify the ideas of Fisher and others, which he termed frequentist, had failed. . .
>
> The conclusion was therefore exactly the opposite of what had been the object of the original exercise, to support frequentist ideas, and had ended up. . . destroying them. The Savage school had produced a constructive approach that began to be explored and was found. . . to work, but which largely disagreed with Fisher.

Bayesian Miscellany

- Decision theory
- Experimental design
- A few more algorithms
  - Savage-Dickey density ratio
  - Data augmentation algorithms
- Toolkits

Bayesian Experimental Design

Basic principles

Choices $= \{e\}$, possible experiments (sample times, sample sizes. . .).

Outcomes $= \{d\}$, values of future data.

Utility balances value of d for achieving experiment goals against the cost of the experiment.

Choose the experiment that maximizes

$$EU(e) = \sum_{d} p(d_e|I)\, U(e,d)$$

To predict d we must know which of several hypothetical "states of nature" H_i is true. $\rightarrow$ Average over H_i:

$$EU(e) = \sum_{H_i} p(H_i|I) \sum_{d} p(d_e|H_i, I)\, U(e,d)$$

Information as Utility

Common goal: discern among the H_i.
$\rightarrow$ Utility $=$ information $\mathcal{I}(H|d_e)$ in $p(H_i|d_e, I)$:

$$\begin{aligned} U(e,d) &= \sum_{H_i} p(H_i|d_e, I)\ \log\left[p(H_i|d_e, I)\right] \\ &= -\text{Entropy of posterior} \end{aligned}$$

Design to maximize expected information.
(Dennis Lindley, 1957)

Expected Information

$$\begin{aligned} E\mathcal{I}(e) &= \sum_{d}\sum_{H_i} p(H_i|I)p(d_e|H_i, I) \\ &\quad \times \sum_{H_i'} p(H_i'|d_e, I)\ \log\left[p(H_i'|d_e, I)\right] \end{aligned}$$

Expected Information

$$E\mathcal{I}(e) = \sum_d p(d_e|I)\,\mathcal{I}[H|d_e, I]$$

Shannon's theorem (BT for entropy):

$$\begin{aligned}
\mathcal{I}[D, H|I] &= \sum_{d,i} p(d, H_i|I) \log p(d, H_i|I) \\
&= \sum_{d,i} p(d, H_i|I) \log p(H_i|I) + \sum_{d,i} p(d|H_i, I) \\
&= \mathcal{I}[H|I] + \sum_i p(H_i|I)\,\mathcal{I}[D|H_i, I] \\
\text{exchange } D, H &= \mathcal{I}[D|I] + \sum_d p(d|I)\,\mathcal{I}[H_i|D, I]
\end{aligned}$$

$$E\mathcal{I}(e) \;=\; \mathcal{I}[H|I] + \sum_i p(H_i|I)\mathcal{I}[D_e|H_i, I] - \mathcal{I}[D_e|I]$$

Info in prior is independent of the experiment.

In many cases info in sampling distribution is independent of experiment (e.g., noise statistics do not depend on sample location).

$$E\mathcal{I}(e) \;=\; C - \int dd_e\; p(d_e|D, I) \log[p(d_e|D, I)]$$

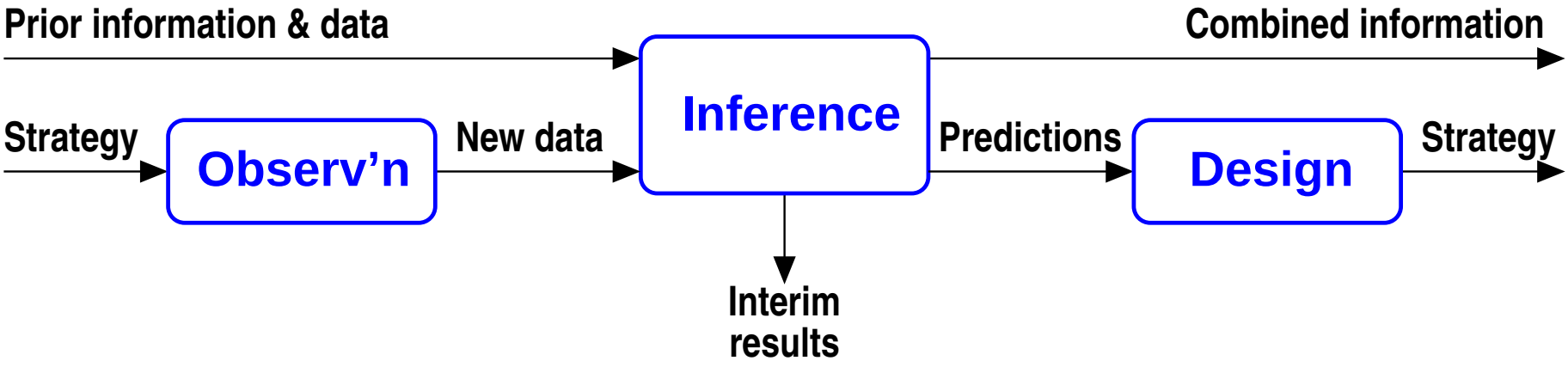
Maximum entropy sampling.
(Sebastiani & Wynn 1997, 2000)

To learn the most, sample where you know the least.

What's Old; What's New

- Linear models, normal errors, flat priors $\rightarrow$ frequentist and Bayesian designs identical
- Nearly all literature treats this case, or nonlinear models linearized about best-fit
- Bayes generalizes straightforwardly to strongly nonlinear models, strong priors
- Interim inference simple in Bayesian framework
- Posterior sampling/MCMC is making rigorous nonlinear design tractable

Bayesian Adaptive Exploration



One-step-at-a-time Bayesian experimental design ("myopic")

Example: Orbit Estimation With Radial Velocity Observations

Data are Kepler velocities plus noise:

$$d_i = V(t_i; \tau, e, K) + e_i$$

3 remaining geometrical params (t_0, λ, i) are fixed.

Noise probability is Gaussian with known $\sigma = 8$ m s^{-1}.

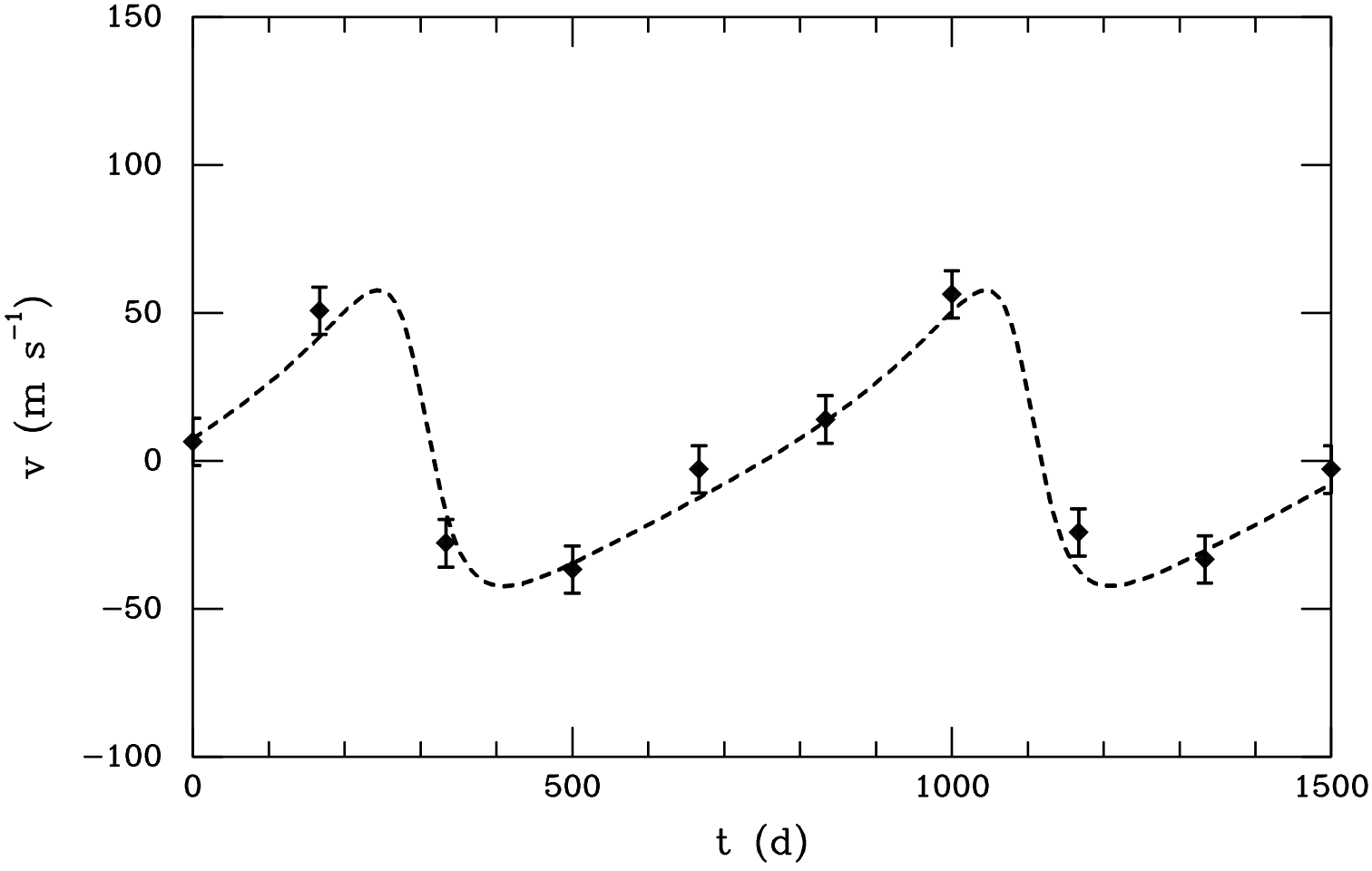
Simulate data with "typical" Jupiter-like exoplanet parameters:

$$\begin{aligned} \tau &= 800 \text{ d} \\ e &= 0.5 \\ K &= 50 \text{ ms}^{-1} \end{aligned}$$

Goal: Estimate parameters τ, e and K.

Cycle 1: Observation

Prior “setup” stage specifies 10 equispaced observations.



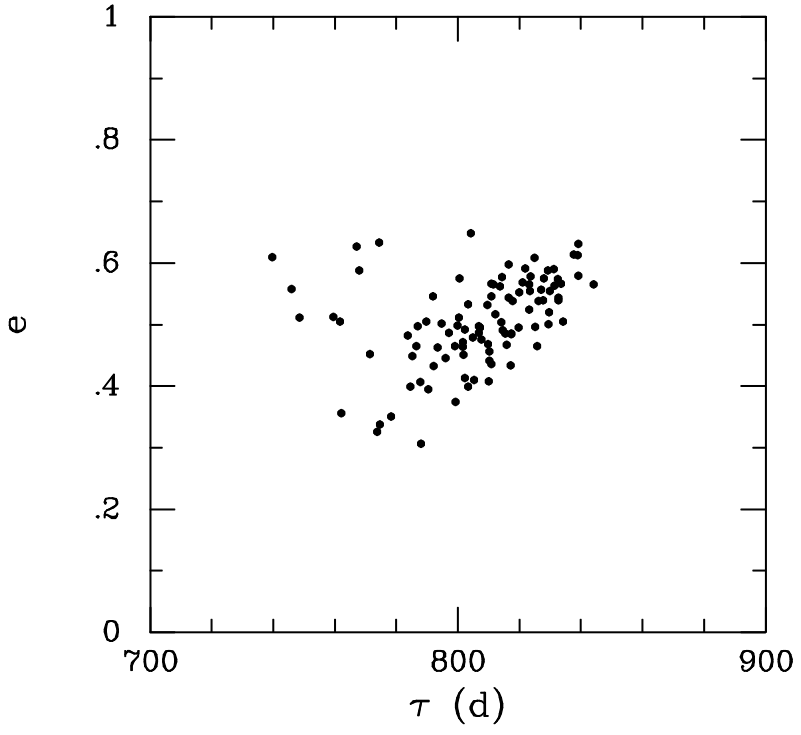
Cycle 1: Inference

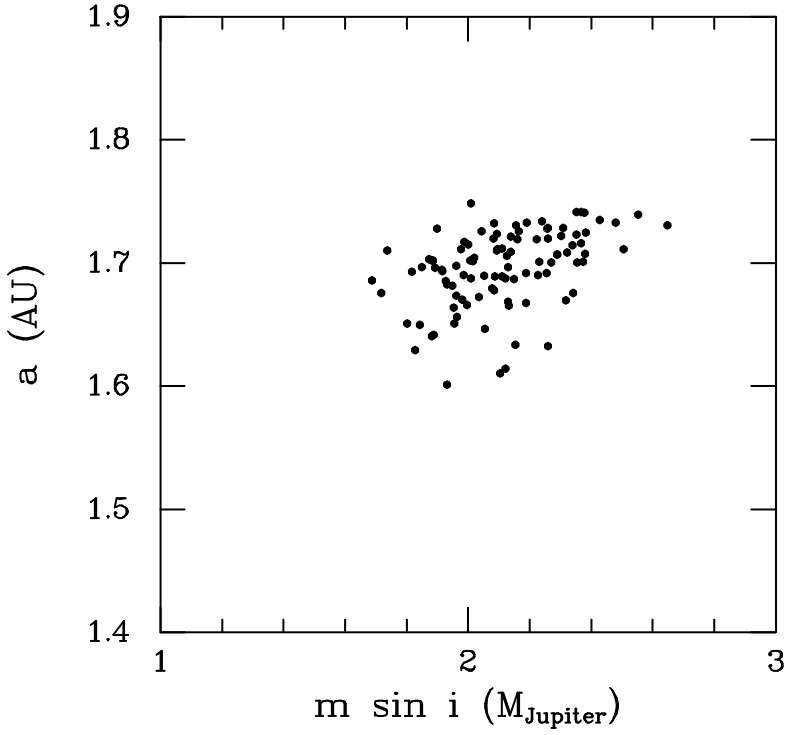
Use flat priors,

$$p(\tau, e, K|D, I) \propto \exp[-Q(\tau, e, K)/2\sigma^2]$$

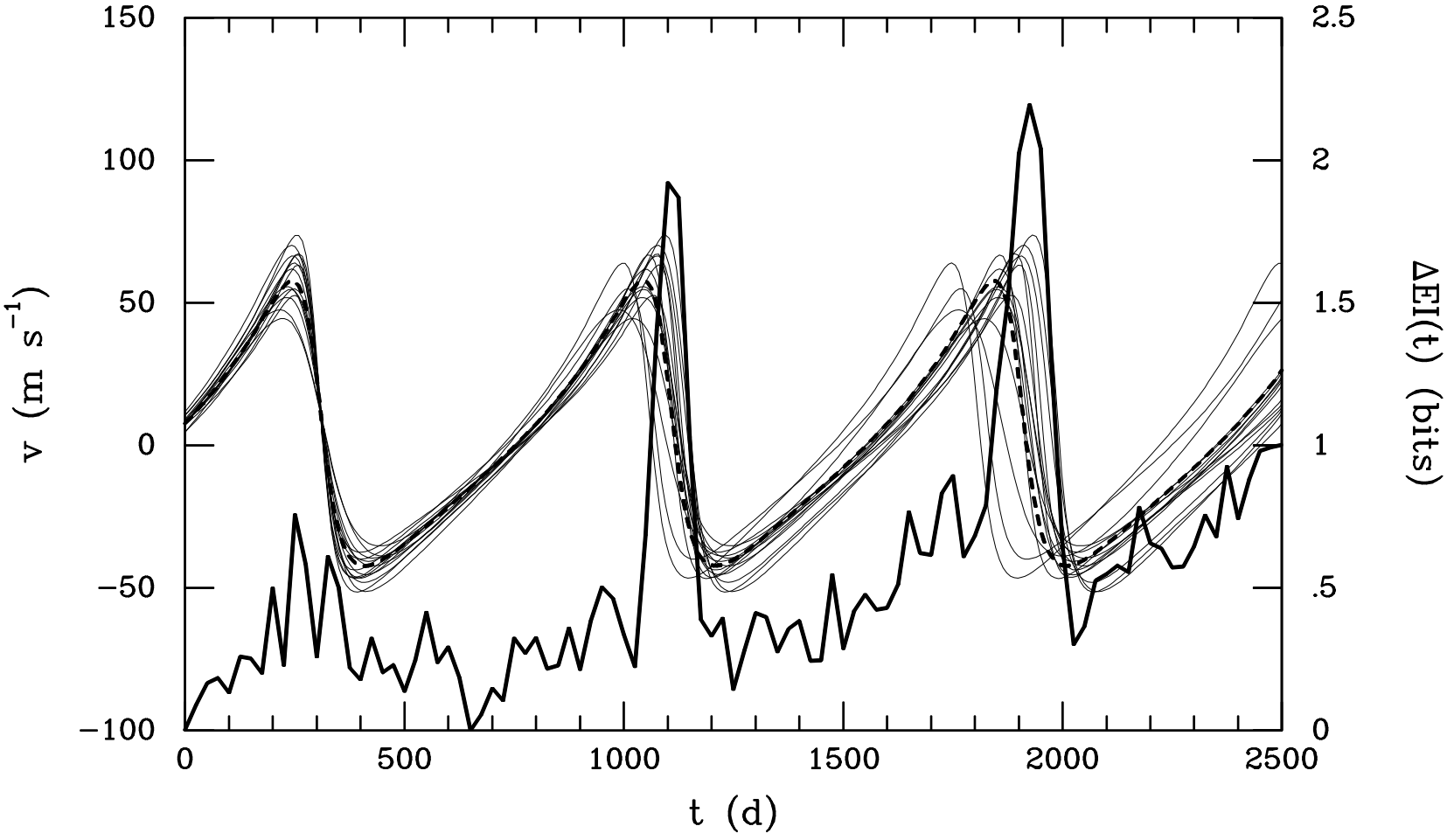
$Q =$ sum of squared residuals using best-fit amplitudes.

Generate $\{\tau_j, e_j, K_j\}$ via posterior sampling.

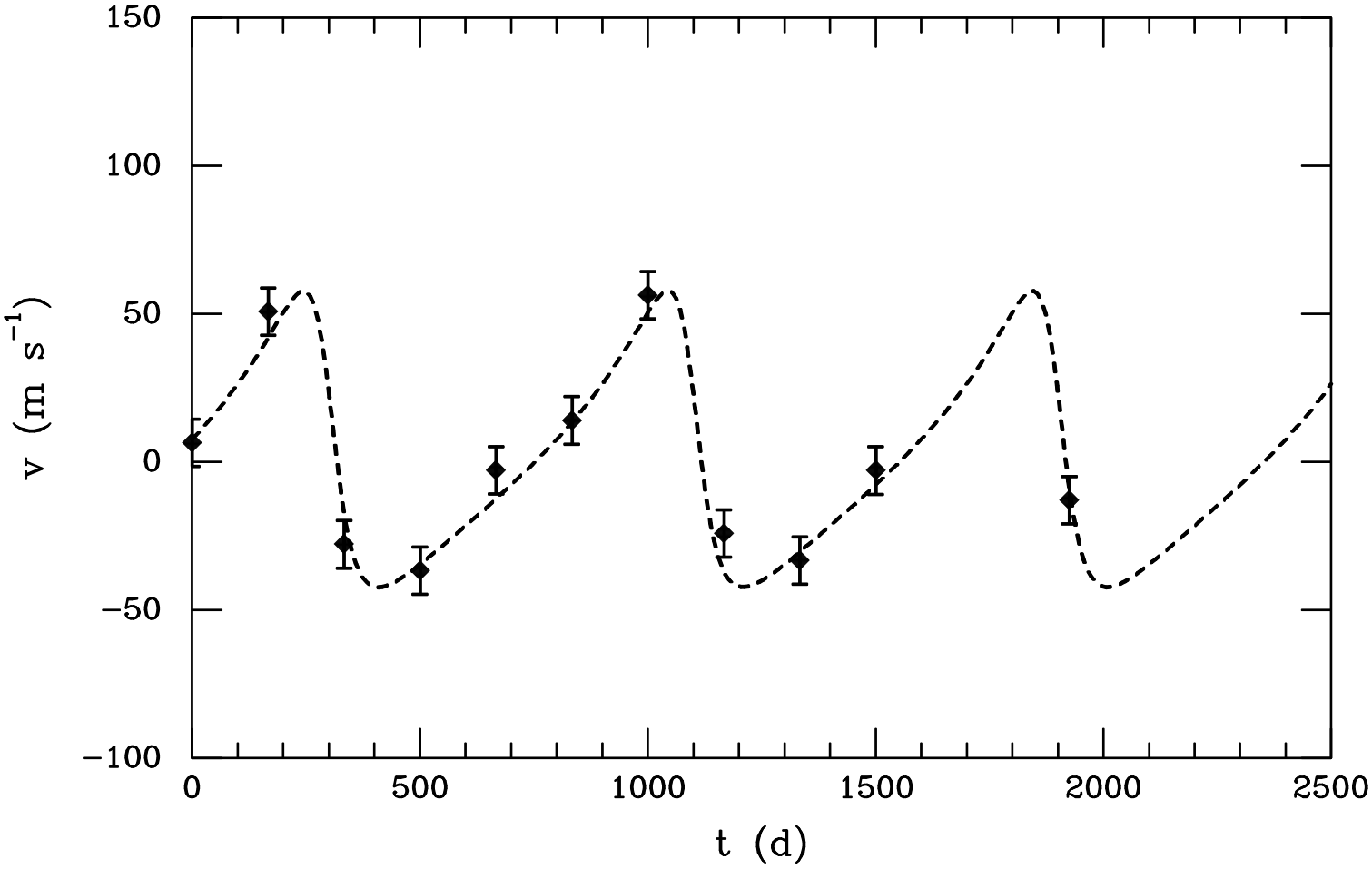




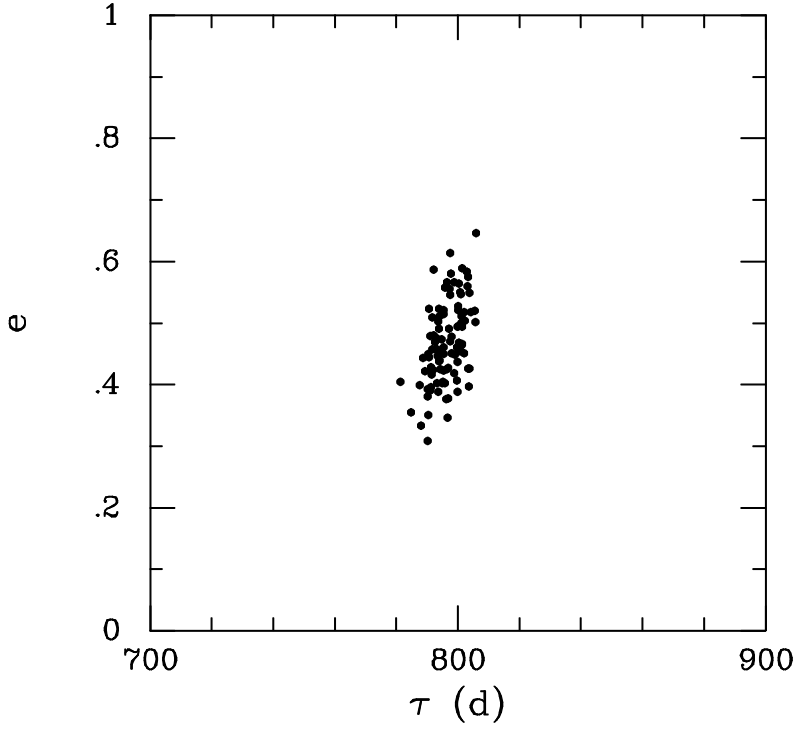
Cycle 1 Design: Predictions, Entropy

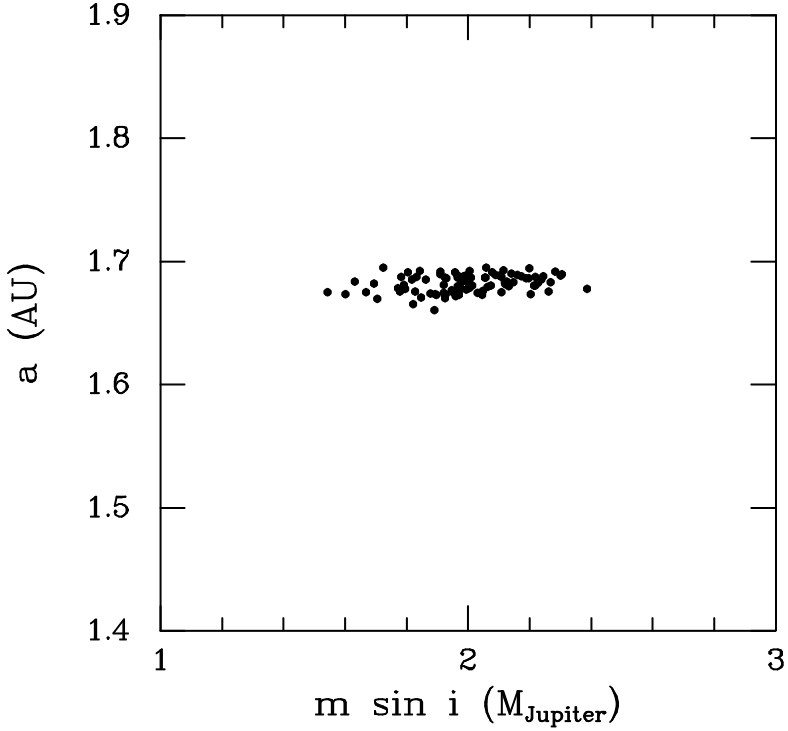


Cycle 2: Observation



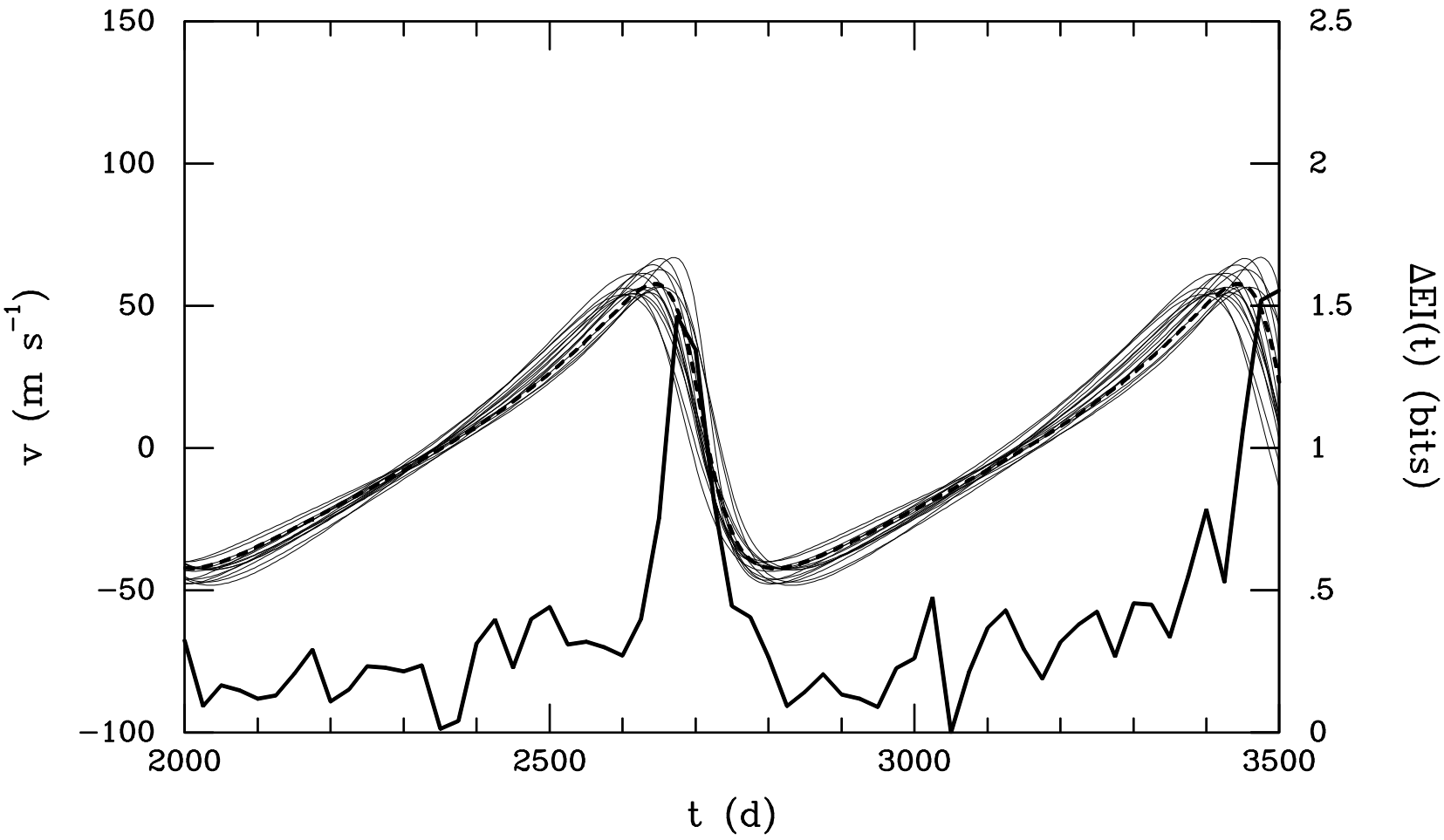
Cycle 2: Inference





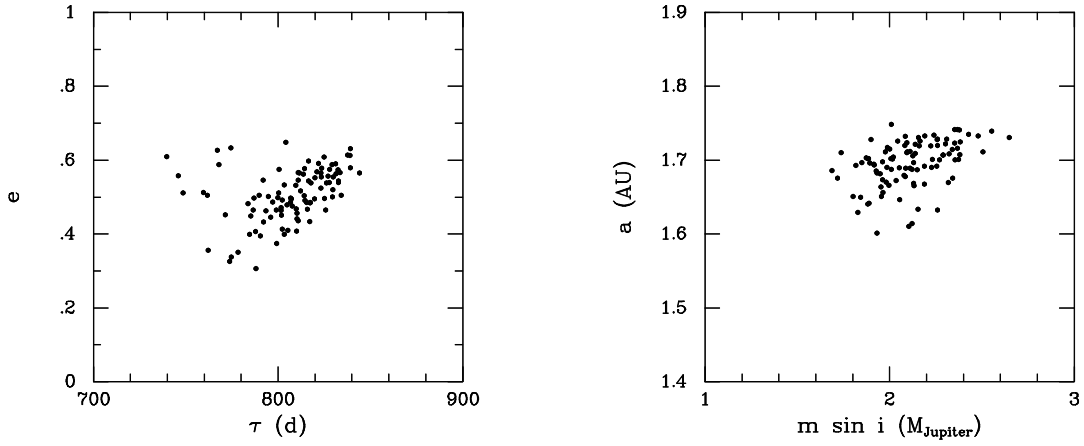
"Volume" $V_2 = V_1/5.8$

Cycle 2: Design

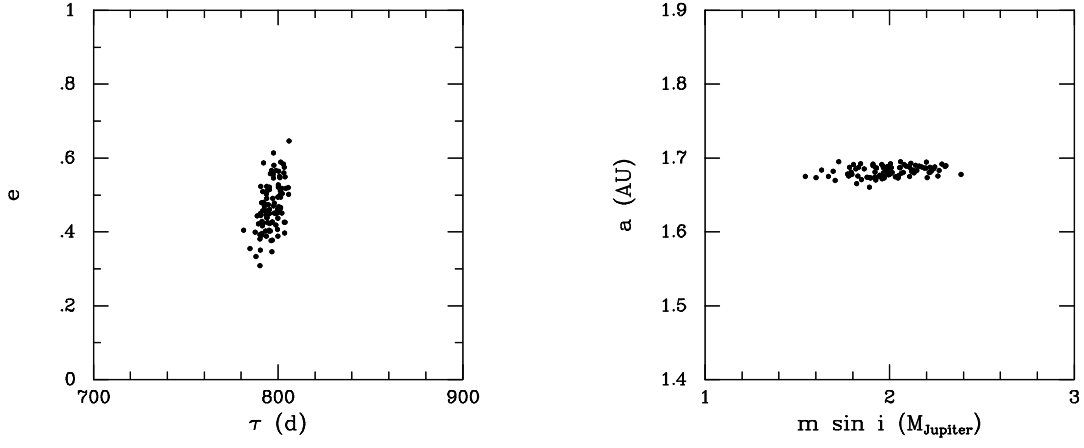


Evolution of Inferences

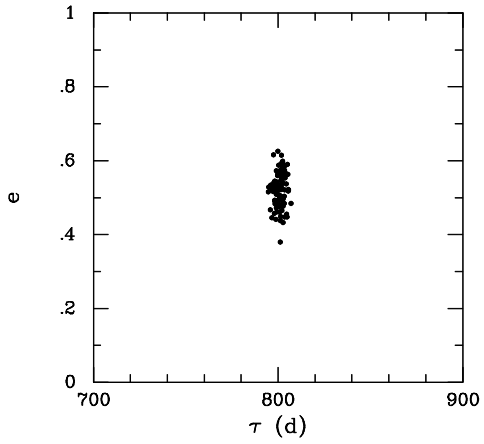
Cycle 1 (10 samples)

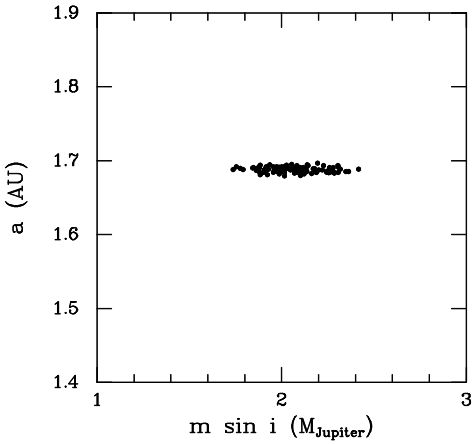


Cycle 2 (11 samples; $V_2 = V_1/5.8$)

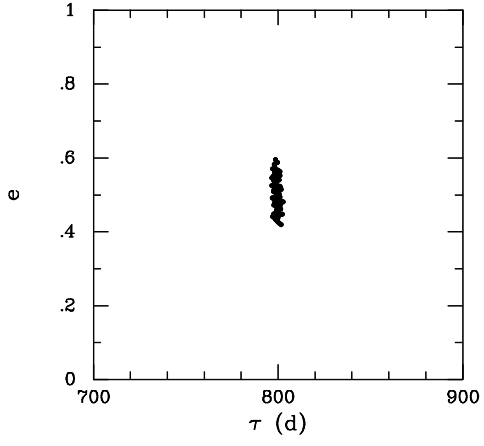


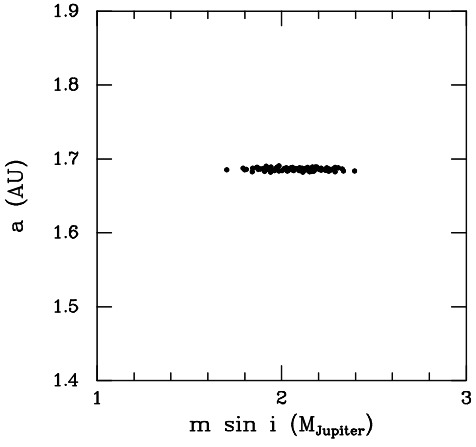
Cycle 3 (12 samples; $V_3 = V_2/3.9$)





Cycle 4 (13 samples; $V_4 = V_3/1.8$)





Bayesian Miscellany

- Decision theory
- Experimental design
- A few more algorithms
  - Savage-Dickey density ratio
  - Data augmentation algorithms
- Toolkits

Noteworthy Algorithms

Savage-Dickey Density Ratio

For model comparison with *nested models:*

M_1: Parameters θ, likelihood $\mathcal{L}_1(\theta)$

M_2: Parameters (θ, ϕ), likelihood $\mathcal{L}_2(\theta, \phi)$

Let $\phi_0 =$ value of ϕ assumed by M_1:

$$\mathcal{L}_1(\theta) = \mathcal{L}_2(\theta, \phi_0)$$

Priors:

$$\begin{aligned} p(\theta|M_1) &= f(\theta) \\ p(\theta, \phi|M_2) &= f(\theta)\, g(\phi) \end{aligned}$$

(may be relaxed).

Compare models via marginal likelihoods:

$$\mathcal{L}(M_1) = \int d\theta \; f(\theta)\, \mathcal{L}_2(\theta, \phi_0)$$

$$\mathcal{L}(M_2) = \int d\theta d\phi \; f(\theta)\, f(\phi)\, \mathcal{L}_2(\theta, \phi)$$

Due to nesting, they appear similar! Note:

$$p(\phi|D, M_2) = \frac{1}{\mathcal{L}(M_2)} \int d\theta \; f(\theta) g(\phi) \mathcal{L}_2(\theta, \phi)$$

Thus

$$\begin{aligned} \mathcal{L}(M_1) &= \int d\theta \; f(\theta) g(\phi_0) \mathcal{L}_2(\theta, \phi_0) \times \frac{\mathcal{L}(M_2)}{g(\phi_0)} \\ &= \frac{p(\phi_0|D, M_2)}{p(\phi_0|M_2)} \mathcal{L}(M_2) \\ \rightarrow B_{21} &= \frac{p(\phi_0|M_2)}{p(\phi_0|D, M_2)} = \frac{\text{penalize } M_2 \text{ for } \phi_0}{\text{penalize } M_1 \text{ for misfit}} \end{aligned}$$

Can approximate this via MCMC with *only* M_2, as long as ϕ_0 isn't too far in tail.

"Missing data" algorithms

Basic idea: "If only we knew x . . . "

Working with $p(\theta|D,M)$ is hard; working with $p(\theta|x,D,M)$ is much easier.

$$p(\theta|D,M) = \int dx\; p(\theta|x,D,M)\, p(x|D,M) \quad \text{(continuous } x\text{)}$$

$$= \sum_i p(\theta|x_i,D,M)\, p(x_i|D,M) \quad \text{(discrete } x\text{)}$$

- **EM algorithm**: Can find mode via 2-step iteration; converges monotonically
- **Data augmentation**: An MCMC approach like the Gibbs sampler that finds the whole marginal for θ

Astronomy applications: CHASC group (David van Dyk, Alanna Connors)

Bayesian Miscellany

- Decision theory
- Experimental design
- A few more algorithms
  - Savage-Dickey density ratio
  - Data augmentation algorithms
- Toolkits

Tools for Computational Bayes

Python

- **PyMC** `http://trichech.us/pymc`
  A framework for MCMC via Metropolis-Hastings; also implements Kalman filters. Targets biometrics, but is general.
- **SimPy** `http://simpy.sourceforge.net/`
  Intro to SimPy http://heather.cs.ucdavis.edu/ matloff/simpy.html
  SimPy (rhymes with "Blimpie") is a process-oriented public-domain package for discrete-event simulation.
- **RSPython** `http://www.omegahat.org/`
  Bi-directional communication between Python and R
- **MDP** `http://mdp-toolkit.sourceforge.net/`
  Modular toolkit for Data Processing: Current emphasis is on machine learning (PCA, ICA...). Modularity allows combination of algorithms and other data processing elements into “flows.”

R and S

- **CRAN Bayesian task view**
  `http://cran.r-project.org/src/contrib/Views/Bayesian.ht`
  Overview of many R packages implementing various Bayesian models and methods
- **Omega-hat** `http://www.omegahat.org/`
  RPython, RMatlab, R-Xlisp
- **BOA** `http://www.public-health.uiowa.edu/boa/`
  Bayesian Output Analysis: Convergence diagnostics and statistical and graphical analysis of MCMC output; can read BUGS output files.
- **CODA**
  `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/coda03/`
  Convergence Diagnosis and Output Analysis: Menu-driven R/S plugins for analyzing BUGS output

Java

- **Omega-hat** `http://www.omegahat.org/`
  Java environment for statistical computing, being developed by XLisp-stat and R developers
- **Hydra** `http://research.warnes.net/projects/mcmc/hydra/`
  HYDRA provides methods for implementing MCMC samplers using Metropolis, Metropolis-Hastings, Gibbs methods. In addition, it provides classes implementing several unique adaptive and multiple chain/parallel MCMC methods.

C/C++/Fortran

- **BayeSys 3** `http://www.inference.phy.cam.ac.uk/bayesys/`
  Sophisticated suite of MCMC samplers including transdimensional capability, by the author of MemSys
- **fbm**
  `http://www.cs.utoronto.ca/~radford/fbm.software.html`
  Flexible Bayesian Modeling: MCMC for simple Bayes, Bayesian regression and classification models based on neural networks and Gaussian processes, and Bayesian density estimation and clustering using mixture models and Dirichlet diffusion trees
- **BayesPack, DCUHRE**
  `http://www.sci.wsu.edu/math/faculty/genz/homepage`
  Adaptive quadrature, randomized quadrature, Monte Carlo integration
- **BIE**, **CDF Bayesian limits** (see below)

Statisticians' Tools

- **BUGS/WinBUGS** `http://www.mrc-bsu.cam.ac.uk/bugs/`
  Bayesian Inference Using Gibbs Sampling: Flexible software for the Bayesian analysis of complex statistical models using MCMC
- **OpenBUGS** `http://mathstat.helsinki.fi/openbugs/`
  BUGS on Windows and Linux, and from inside the R
- **XLisp-stat**
  `http://www.stat.uiowa.edu/~luke/xls/xlsinfo/xlsinfo.htm`
  Lisp-based data analysis environment, with an emphasis on providing a framework for exploring the use of dynamic graphical methods

Astronomer/Physicist Tools

- **BIE** `http://www.astro.umass.edu/~weinberg/proto_bie/`
  Bayesian Inference Engine: General framework for Bayesian inference, tailored to astronomical and earth-science survey data. Built-in database capability to support analysis of terabyte-scale data sets. Inference is by Bayes via MCMC.
- **XSpec**, **CIAO/Sherpa**
  Both environments have some basic Bayesian capability.
- **CDF Bayesian Limit Software**
  `http://www-cdf.fnal.gov/physics/statistics/statistics_s`
  Limits for Poisson counting processes, with background & efficiency uncertainties
- **root** `http://root.cern.ch/`
  Bayesian support? (`BayesDivide`)

Closing Reflections

Philip Dawid (2000)

> What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria. I have heard people use this 'dullness' as an argument against Bayesianism. One might as well complain that Newton's dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy's epicyclic system.
>
> All my experience teaches me that it is invariably more fruitful, and leads to deeper insights and better data analyses, to explore the consequences of being a 'thoroughly boring Bayesian'.

Dennis Lindley (2000)

> The philosophy places more emphasis on model construction than on formal inference. . . I do agree with Dawid that 'Bayesian statistics is fundamentally boring'. . . My only qualification would be that the theory may be boring but the applications are exciting.