Miscellany*: Long Run Behavior of Bayesian Methods; Bayesian Experimental Design

(Lecture 4)

Tom Loredo

Dept. of Astronomy, Cornell University

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

*Bayesian use of sample space integrals

The Frequentist Outlook

Probabilities for hypotheses are meaningless because hypotheses are not “random variables.”

Data *are* random, so only probabilities for data can appear in calculations.

These probabilities must be interpreted as long-run frequencies.

$\Rightarrow$ Seek to identify procedures that have good behavior in the long run.

What is good for the long run
is good for the case at hand.

The Bayesian Outlook

Quantify information about the case at hand as completely and consistently as possible.

No explicit regard given to long run performance.

But a result that claims to be optimal in each case should behave well in the long run.

Is what is good for the case at hand also good for the long run?

Long Run Behavior of Bayesian Methods

Agenda

- Bayesian calibration
- Consistency & convergence of Bayesian methods

Bayesian Calibration

Credible region $\Delta(D)$ with probability P:

$$P = \int_{\Delta(D)} d\theta \; p(\theta|I) \frac{p(D|\theta, I)}{p(D|I)}$$

What fraction of the time, Q, will the true θ be in $\Delta(D)$?

1. Draw θ from $p(\theta|I)$
2. Simulate data from $p(D|\theta, I)$
3. Calculate $\Delta(D)$ and see if $\theta \in \Delta(D)$

$$Q = \int d\theta \; p(\theta|I) \int dD \; p(D|\theta, I) \, [\theta \in \Delta(D)]$$

$$Q = \int d\theta\, p(\theta|I) \int dD\, p(D|\theta, I)\, [\theta \in \Delta(D)]$$

Note appearance of $p(\theta, D|I) = p(\theta|D, I) p(D|I)$:

$$\begin{aligned}
Q &= \int dD \int d\theta\, p(\theta|D, I)\, p(D|I)\, [\theta \in \Delta(D)] \\
&= \int dD\, p(D|I) \int_{\Delta(D)} d\theta\, p(\theta|D, I) \\
&= P \int dD\, p(D|I) \\
&= P
\end{aligned}$$

Bayesian inferences are "calibrated." *Always.*
Calibration is with respect to choice of prior & $\mathcal{L}$.

Real-Life Confidence Regions

Theoretical (frequentist) confidence regions:

A rule $\delta(D)$ gives a region with covering probability:

$$C_\delta(\theta) = \int dD\, p(D|\theta, I)\, [\theta \in \delta(D)]$$

It's a *confidence region* iff $C(\theta) = P$, a *constant*.

Such rules almost never exist in practice!

The CR requirement is often relaxed: require $C(\theta) \geq P$ (conservative).

The actual coverage of many standard regions thus fluctuates (even for coin flipping—Brown et al. 2000).

Average coverage:

Intuition suggests reporting some kind of average performance: $\int d\theta\ f(\theta) C_\delta(\theta)$

Recall the Bayesian calibration condition:

$$
\begin{aligned}
P &= \int d\theta\, p(\theta|I) \int dD\, p(D|\theta, I)\, [\theta \in \Delta(D)] \\
&= \int d\theta\, p(\theta|I)\, C_\delta(\theta)
\end{aligned}
$$

provided we take $\delta(D) = \Delta(D)$.

- If $C_\Delta(\theta) = P$, the credible region *is* a confidence region.
- Otherwise, the credible region's probability content accounts for a priori uncertainty in θ—we *need* priors for this.

Calibration for Bayesian Model Comparison

Assign prior probabilities to N_M different models.

Choose as the true model that with the highest posterior probability, but only if the probability exceeds $P_{\rm crit}$.

Iterate via Monte Carlo:

- 1.Choose a model by sampling from the model prior.
- 2.Choose parameters for that model by sampling from the parameter prior *pdf*.
- 3.Sample data from that model's sampling distribution conditioned on the chosen parameters.
- 4.Calculate the posteriors for all the models; choose the most probable if its $P > P_{\rm crit}$.

$\Rightarrow$ Will be correct $\geq 100 P_{\rm crit}$% of the time that we reach a conclusion in the Monte Carlo experiment.

Robustness to model prior:

What if model frequencies $\neq$ model priors?

Choose between two models based on the Bayes factor, B, but let them occur with nonequal frequencies. Let

$$\gamma = \max\left[\frac{p(M_1 \mid I)}{p(M_2 \mid I)}, \frac{p(M_2 \mid I)}{p(M_1 \mid I)}\right]$$

Fraction of time a correct conclusion is made if we require $B > B_{\rm crit}$ or $B < 1/B_{\rm crit}$ is

$$Q > \frac{1}{1 + \frac{\gamma}{B_{\rm crit}}}$$

E.g., if $B_{\rm crit} = 100$:

- Correct ≥ 99% if $\gamma = 1$
- Correct ≥ 91% if $\gamma = 9$

A Worry: Incorrect Models

What if none of the models is “true”?

Comfort from experience: Rarely are statistical models precisely true, yet standard models have proved themselves adequate in applications.

Comfort from probabilists: Studies of consistency in the framework of nonparametric Bayesian inference show “good priors are those that are approximately right for most densities; parametric priors [e.g., histograms] are often good enough” (Lavine 1994).

One should worry somewhat, but there is not yet any theory providing a consistent, quantitative “model failure alert.”

Bayesian Consistency & Convergence

Parameter Estimation:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$.
- Using standard nonuniform "reference" priors can improve their performance to $O(n^{-1})$.
- For handling nuisance parameters, regions based on marginal likelihoods have superior long-run performance to regions found with conventional frequentist methods like profile likelihood.

Model Comparison:

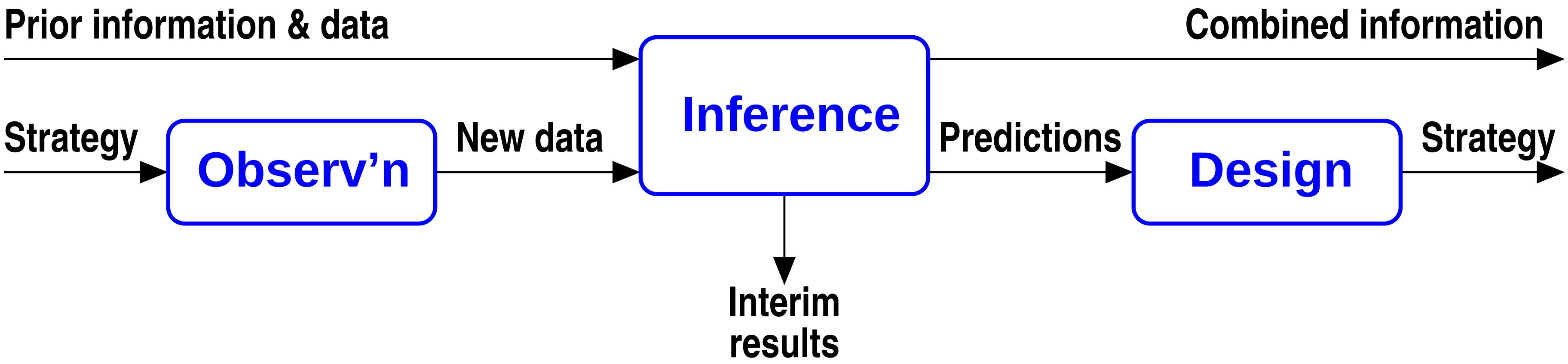
- Model comparison is asymptotically consistent. Popular frequentist procedures (e.g., χ^2 test, asymptotic likelihood ratio ($\Delta\chi^2$), AIC) are not.
- For separate (not nested) models, the posterior probability for the true model converges to 1 exponentially quickly.
- When selecting between more than 2 models, carrying out multiple frequentist significance tests can give misleading results. Bayes factors continue to function well.

Summary

Parametric Bayesian methods are typically excellent frequentist methods!

Not too surprising—methods that claim to be optimal for each individual case should be good in the long run, too.

Bayesian Adaptive Exploration



- Theory
  - Decision theory
  - Experimental design
- Proof of concept: Exoplanets
  - Motivation: SIM EPIc Survey
  - Demonstration: A few BAE cycles
- Challenges

Bayesian Decision Theory

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.

Utility $= U(c, o)$

- c: Choice of action (decide b/t these)
- o: Outcome (what we are uncertain of)

Deciding amidst uncertainty

We are uncertain of what the outcome will be
$\rightarrow$ average:

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

The best choice maximizes the expected utility:

$$\hat{c} = \arg\max_c EU(c)$$

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:

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

Information as Utility

Common goal: discern among the H_i.
$\rightarrow$ Utility $=$ information $\mathcal{I}(e,d)$ 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.

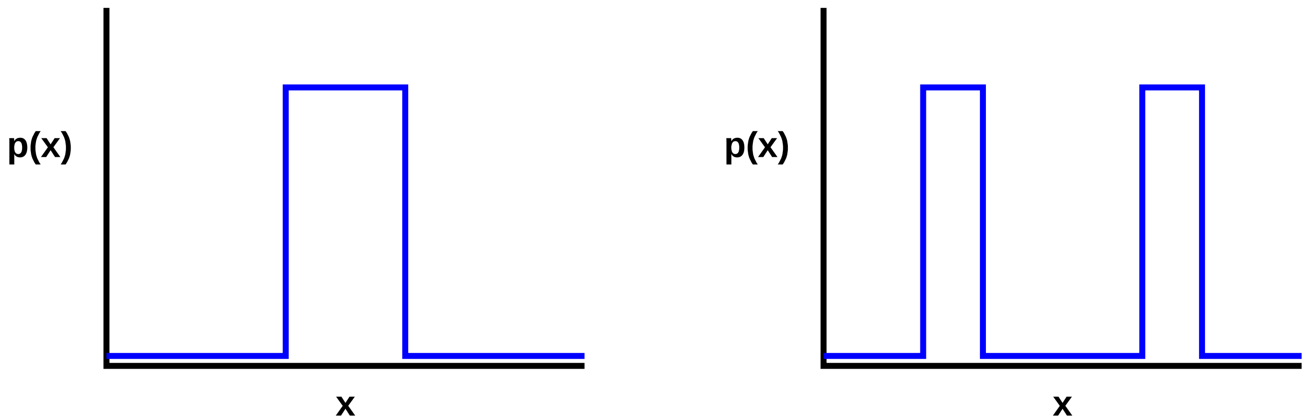
Measuring Information With Entropy

Entropy of a Gaussian

$$p(x) \propto e^{-(x-\mu)^2/2\sigma^2} \quad \longrightarrow \quad \mathcal{I} \propto -\log(\sigma)$$

$$p(\vec{x}) \propto \exp\left[-\tfrac{1}{2}\vec{x} \cdot \mathbf{V}^{-1} \cdot \vec{x}\right] \quad \longrightarrow \quad \mathcal{I} \propto -\log(\det \mathbf{V})$$

Entropy measures volume, not width

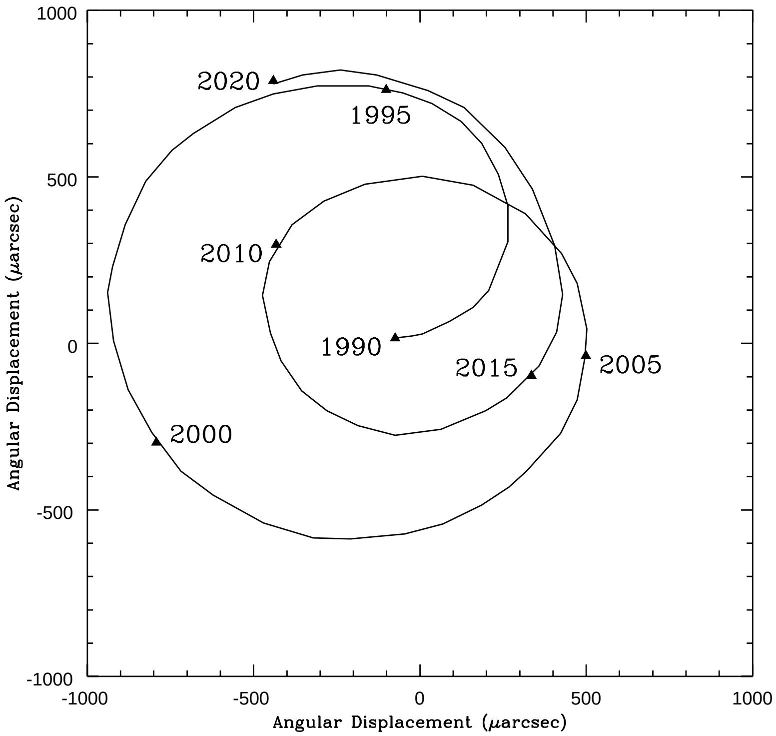


These distributions have the same entropy/amount of information.

Finding Exoplanets: The Space Interferometry Mission

SIM in 2009 (?)

The Sun’s Wobble From 10 pc



EPIcS: Extrasolar Planet Interferometric Survey

Tier 1

- Goal: Identify Earth-like planets in habitable regions around nearby Sun-like stars
- Requires 1 μas astrometry
  - Long integration times
  - Astrometrically stable reference stars
- ~ 75 MS stars within 10 pc, ~ 70 epochs per target

Tier 2

- Goal: Explore the nature and evolution of planetary systems in their full variety
- Requires 4 μas astrometry, short integration times
- ~ 1000 targets, “piggyback” on Tier 1

Preparatory observing

- High precision radial velocity and adaptive optics observing
- Identify science targets
- Identify reference stars (K giants? eccentric binaries?)

Huge resource expenditures
$\rightarrow$ must optimize use of resources

Example: Orbit Estimation With Radial Velocity Observations

Data are Kepler velocity 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}.

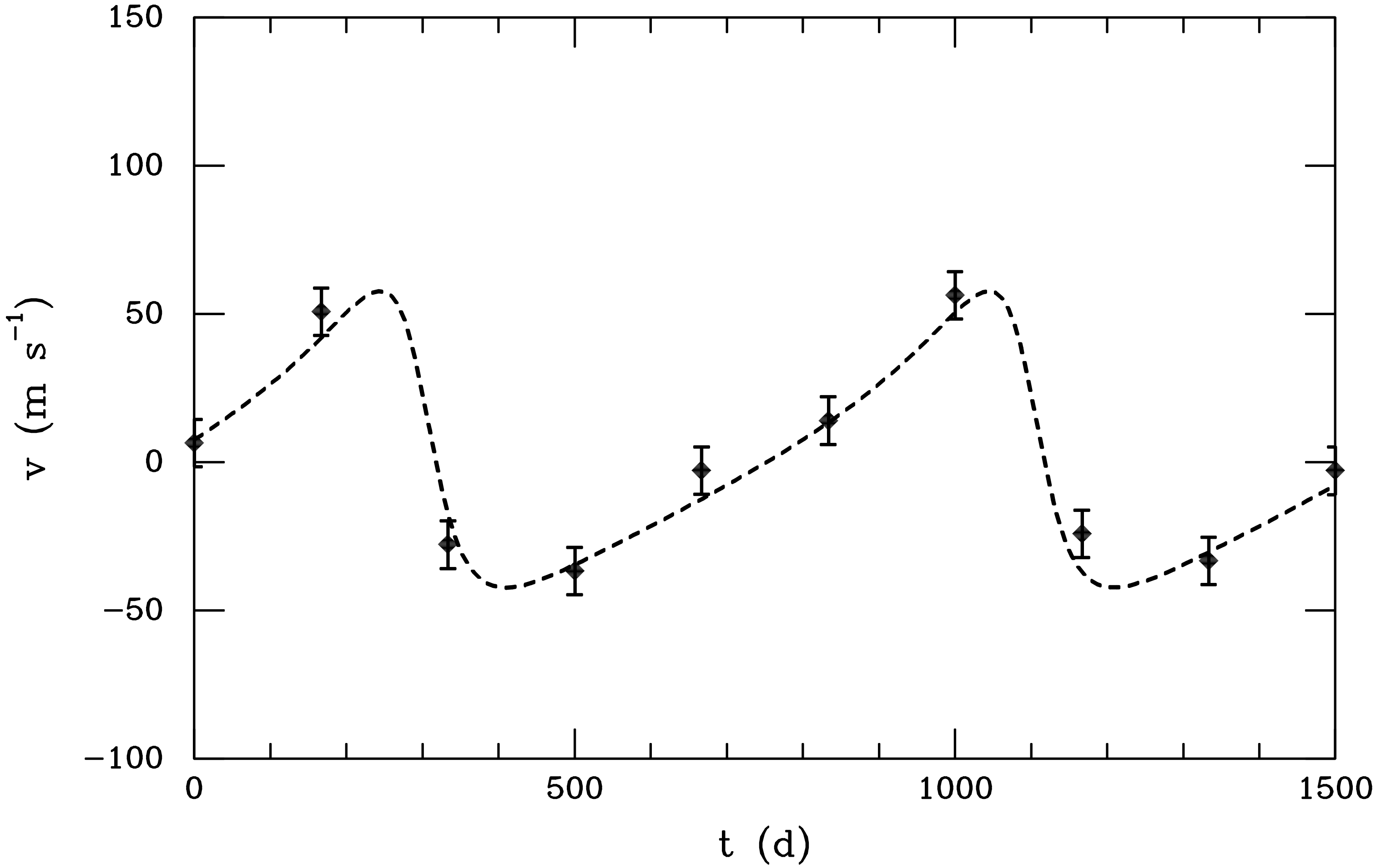
Simulate data with “typical” Jupiter-like exoplanet parameters:

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

Goal: Estimate parameters τ, e and K.

Cycle 1: Observation

Prior “setup” stage specifies 10 equispaced observations.



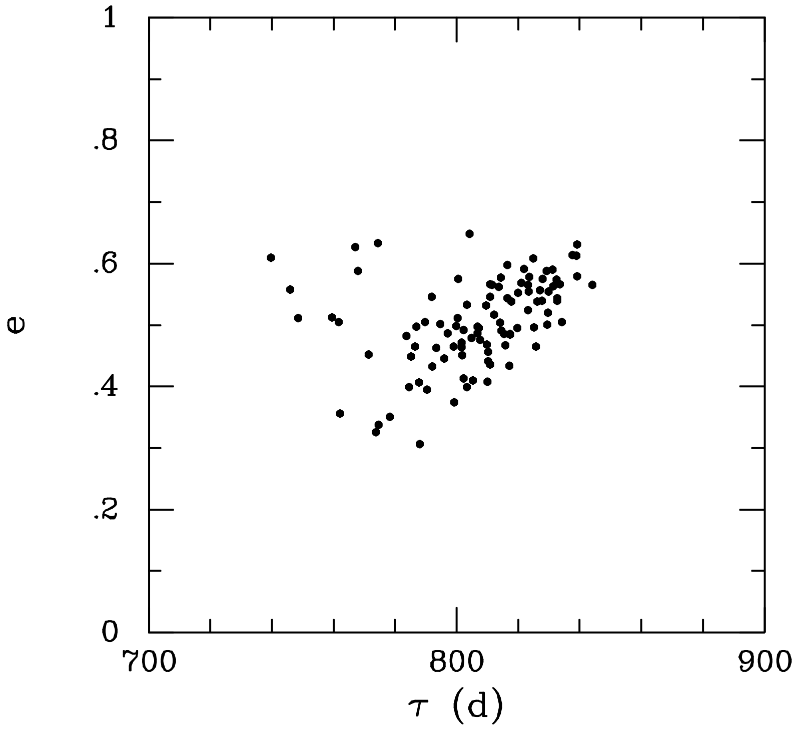
Cycle 1: Inference

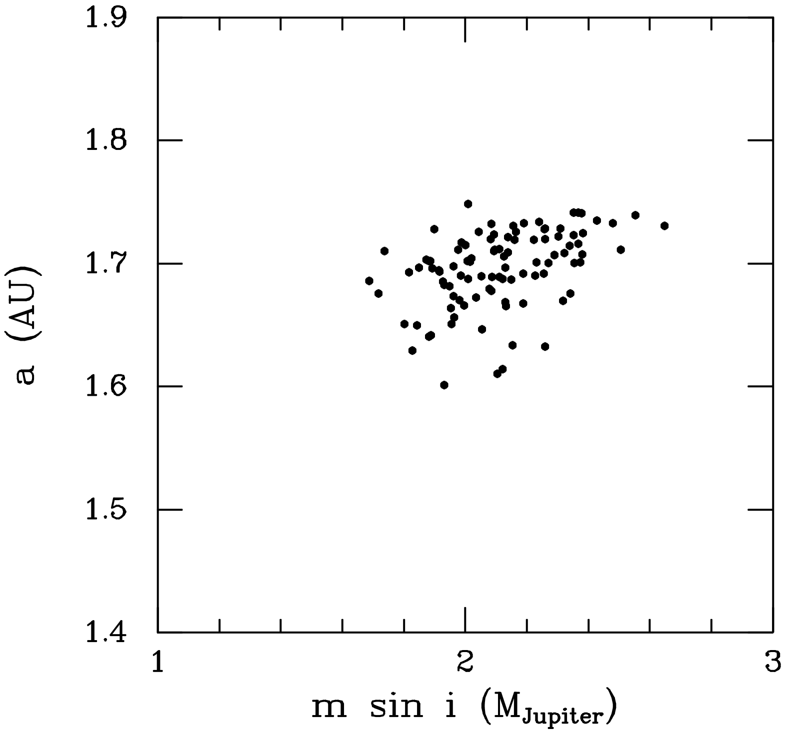
Use flat priors,

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

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

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





Aside: Kepler Periodograms

Keplerian radial velocity model:

$$V(t) = A_1 + A_2[e + \cos \upsilon(t)] + A_3 \sin \upsilon(t)$$

$$\upsilon(t) = f(t; \tau, e, T) \qquad \text{via Kepler's eqn}$$

Period τ and 2 other nonlinear parameters (e, T)
3 linear amplitudes (COM velocity, orbital velocity, λ)

Use Bretthorst algorithm. For $e = 0 \rightarrow$ L-S periodogram, the current standard tool, but the Bayesian generalization accounts for orbital eccentricity.

For astrometry, 2D data require $x(t)$, $y(t)$.
Extra parameters: inclination, parallax, proper motion.

Cycle 1: Design

Predict value of future datum at t

$$
\begin{aligned}
p(d|t, D, I) &= \int d\tau\; de\; dK\; p(\tau, e, K|D, I) \\
&\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[d - v(t;\tau, e, K)]^2}{2\sigma^2}\right) \\
&\approx \frac{1}{N} \sum_{\{\tau_j, e_j, K_j\}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[d - v(t;\tau_j, e_j, K_j)]^2}{2\sigma^2}\right)
\end{aligned}
$$

Effect of a datum on inferences

Information if we sample at t and get datum d:

$$\mathcal{I}(d, t) = \int d\tau\; de\; dK\; p(\tau, e, K|d, t, D, I) \log[p(\tau, e, K|d, t, D, I)$$

Average over unknown datum value

Expected information:

$$\mathcal{EI}(t) \;=\; \int dd\; p(d|t, D, I)\, \mathcal{I}(d, t)$$

Width of noise dist'n is independent of value of the signal$\longrightarrow$

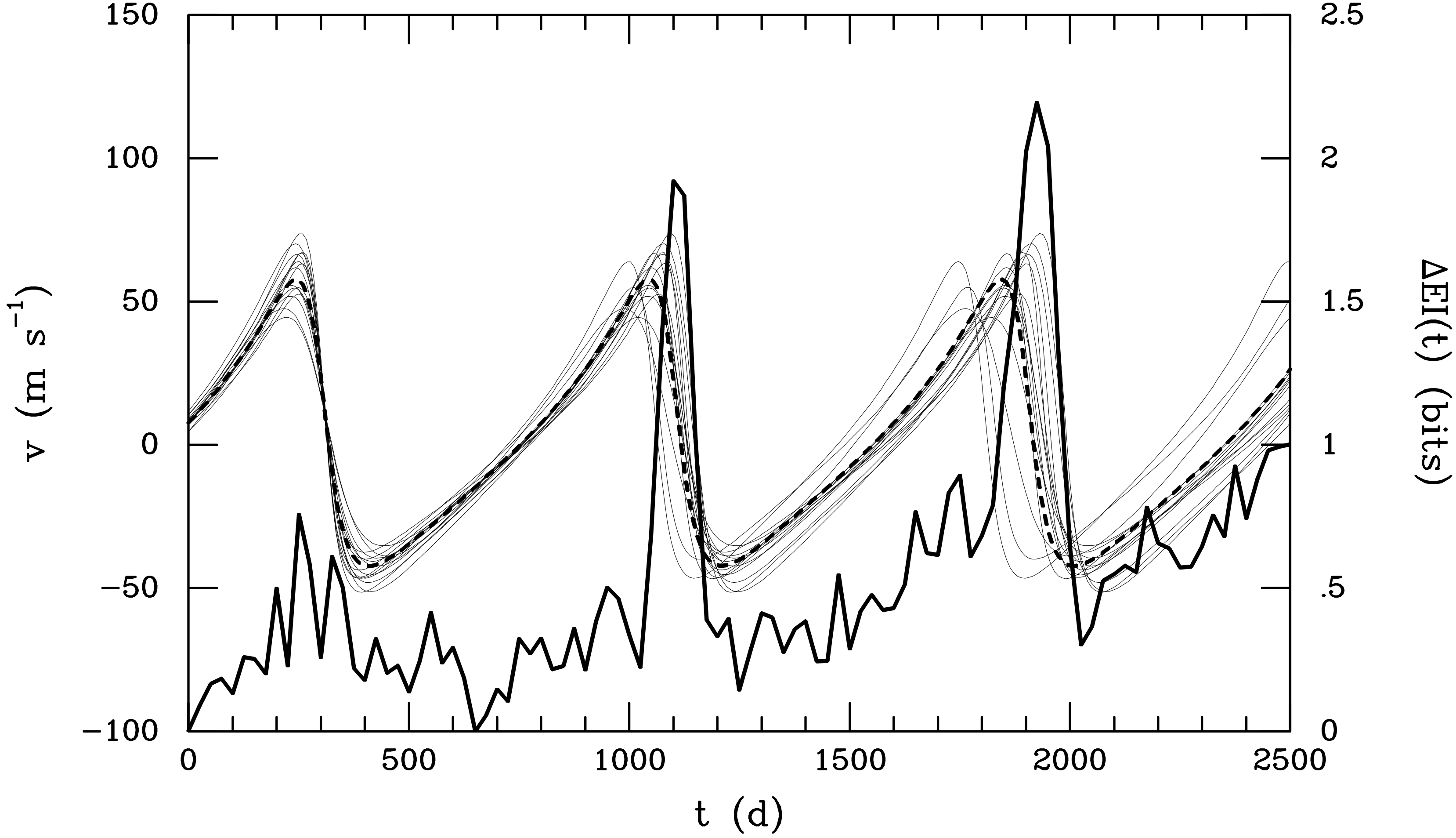
$$\mathcal{EI}(t) \;=\; -\int dd\; p(d|t, D, I) \log[p(d|t, D, I)]$$

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

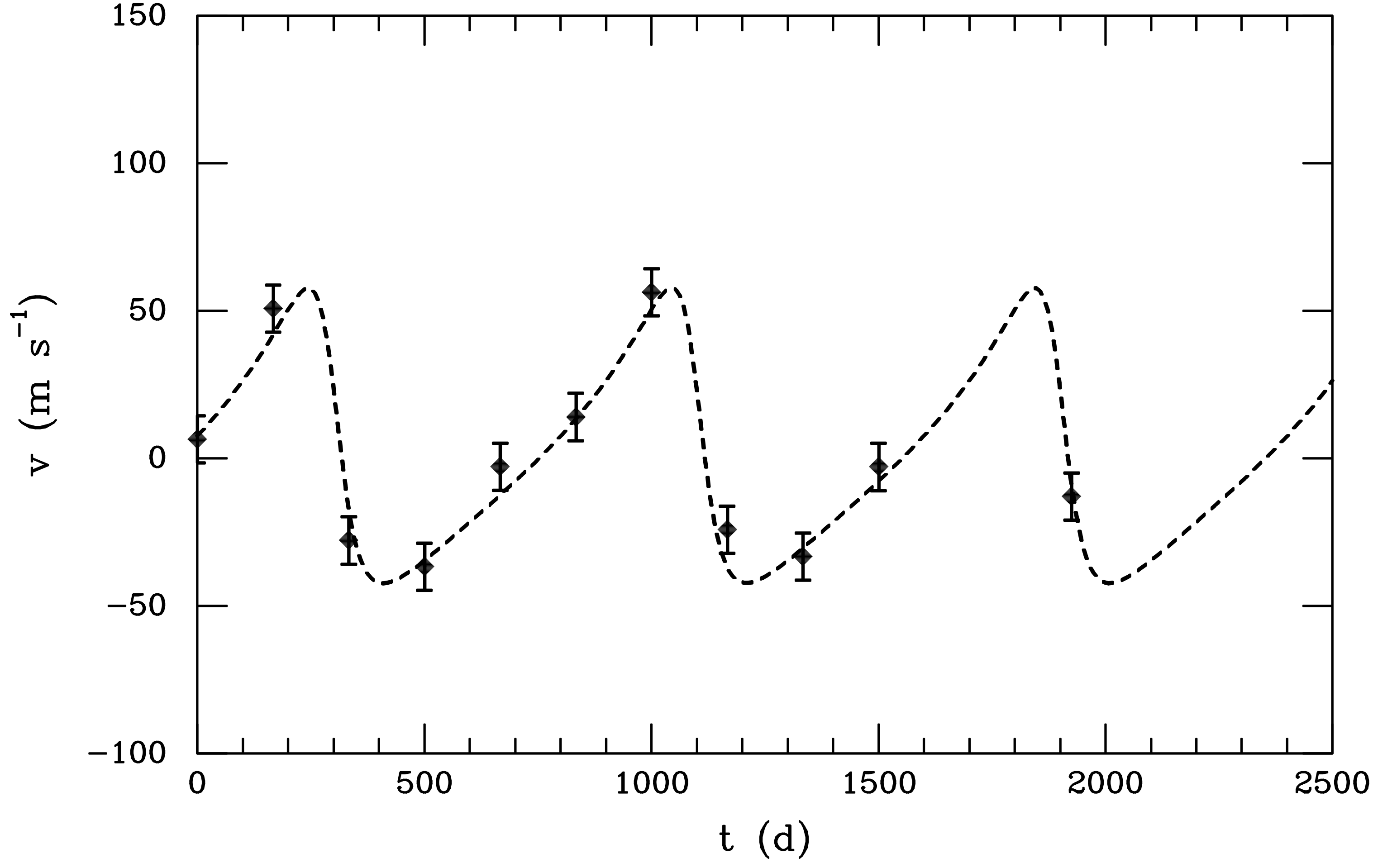
Evaluate by Monte Carlo using posterior samples & data samples.

Pick t to maximize $\mathcal{EI}(t)$.

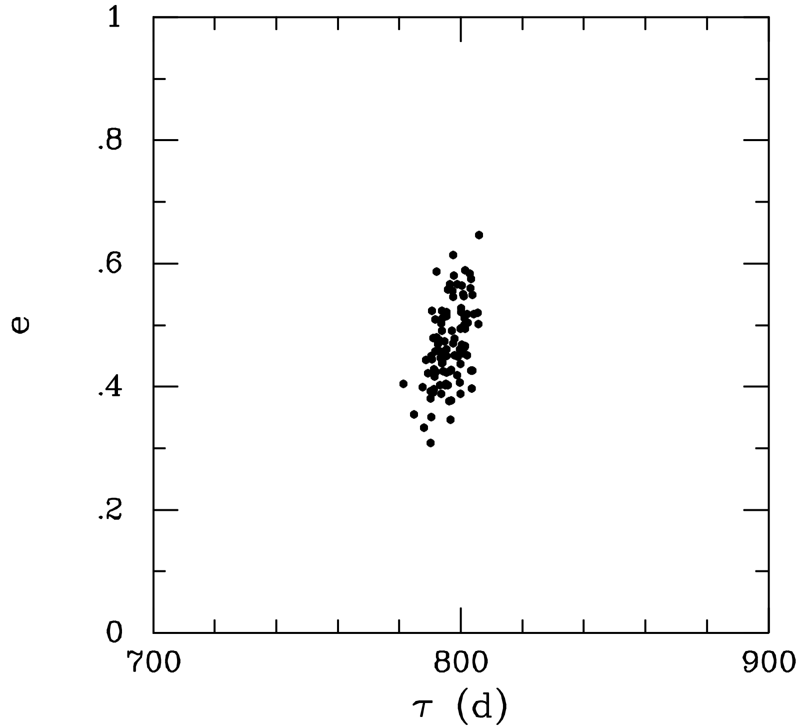
Design Results: Predictions, Entropy

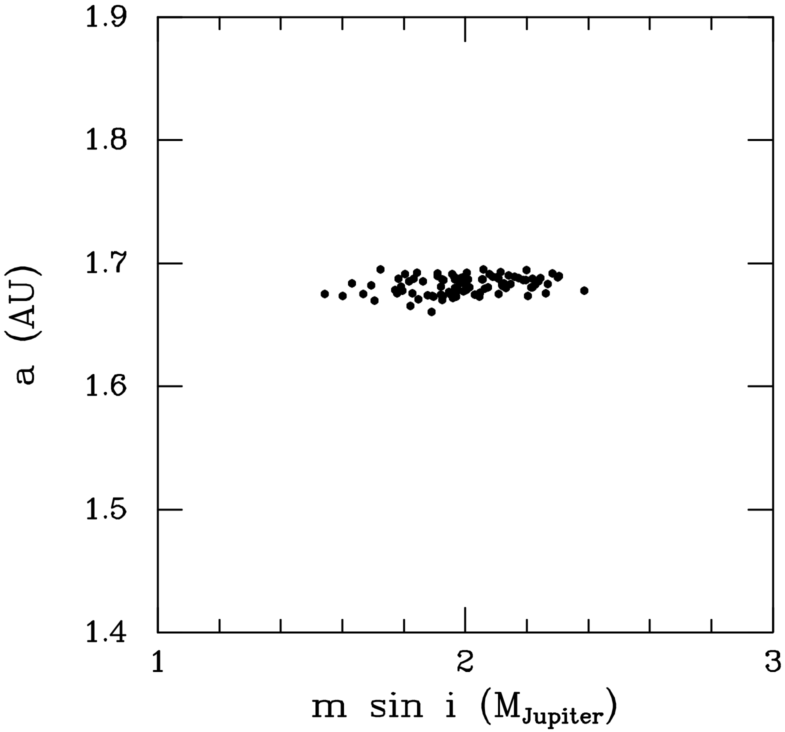


Cycle 2: Observation

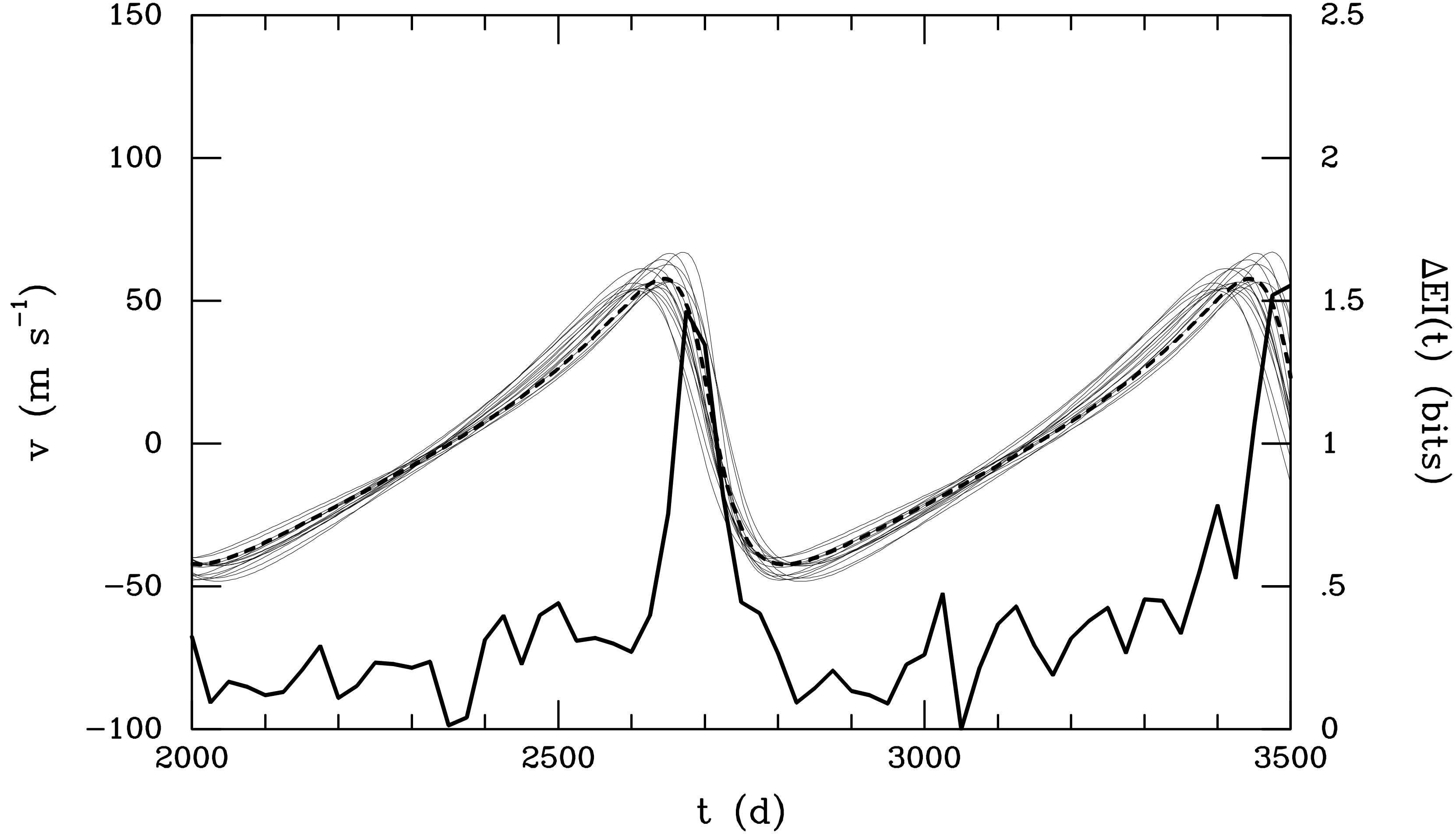


Cycle 2: Inference



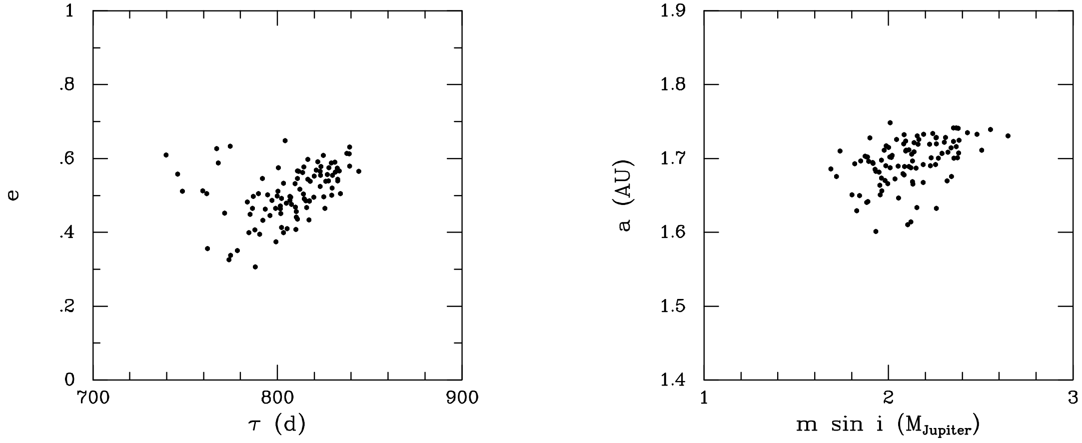


Cycle 2: Design

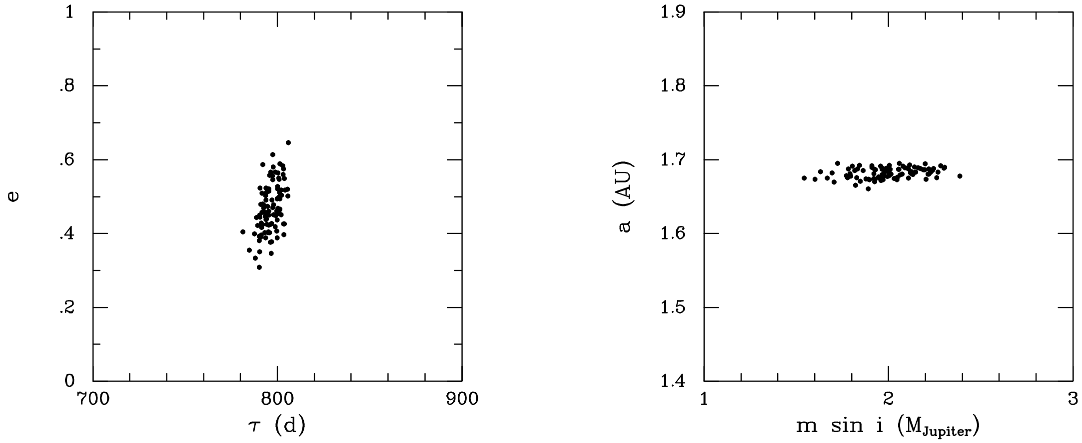


Evolution of Inferences

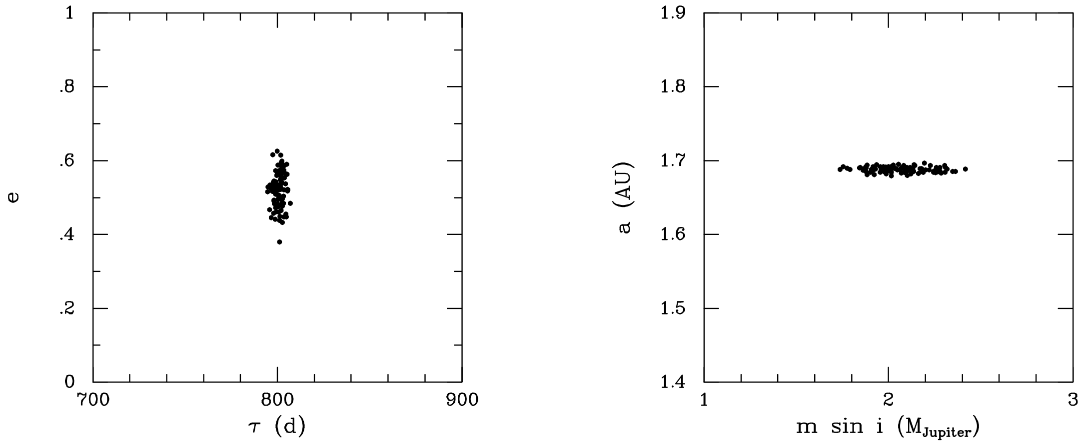
Cycle 1 (10 samples)



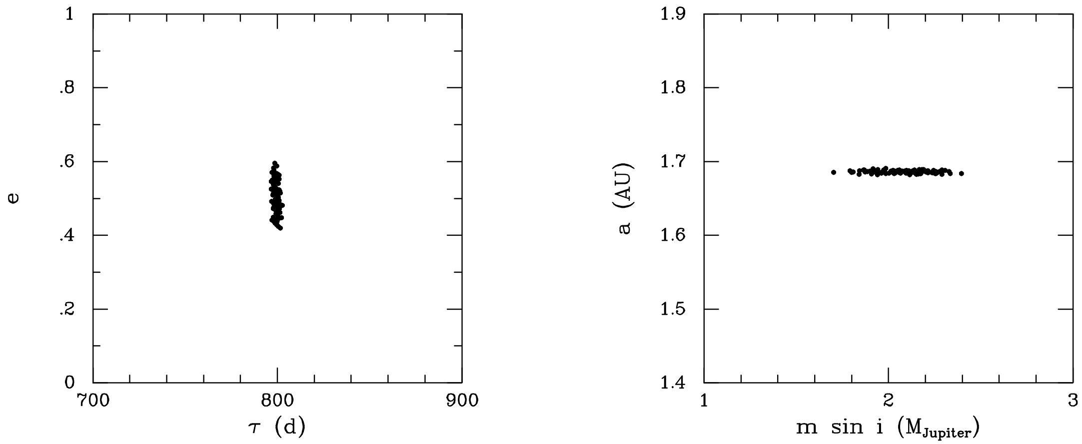
Cycle 2 (11 samples)



Cycle 3 (12 samples)



Cycle 4 (13 samples)



Challenges

Evolving goals for inference

Goal may originally be detection (model comparison), then estimation. How are these related? How/when to switch?

Generalizing the utility function

Cost of a sample vs. time or costs of samples of different size could enter utility. How many bits is an observation worth?

Computational algorithms

Are there MCMC algorithms uniquely suited to adaptive exploration? When is it smart to linearize?

Design for the “setup” cycle

What should the size of a setup sample be? Can the same algorithms be used for setup design?

When is it worth the effort?

Key Ideas

Sample space integrals are useful in a Bayesian setting.

- Long run behavior of Bayesian methods
  - Bayesian methods are calibrated
  - Parametric Bayesian methods have good frequentist behavior

  Bayes may be the best way to be frequentist!
- Bayesian adaptive exploration
  - Can provide dramatic benefits in nonlinear settings
  - Many challenges and open questions