# From Gaussians to Periodograms
## *(Lecture 2)*

Tom Loredo

Dept. of Astronomy, Cornell University

# Summary of Lecture 1

*Overview of Bayesian inference*

- What to do: Calculate $p(H_i|D,\ldots)$
- What's different about it:
  - Problem solving vs. solution characterization
  - Parameter space vs. sample space integrals
- How to do it:
  - Large $N$: Laplace Approximation
  - Exact: Analytic, Adaptive/randomized quadrature, posterior sampling

# Today's Lecture

- Simple example: Coin flipping
- Inference with Gaussians
  - Estimating the mean ($\sigma$ known & unknown)

$$d_i = \mu + \epsilon_i$$

  - Superposed nonlinear models (Bretthorst algorithm)

$$d_i = \sum_{\alpha=1}^{M} A_\alpha g_\alpha(x_i;\theta) + \epsilon_i$$

# Example: Coin Flipping!

Parameter Estimation

$M =$ Assumed independence of flips

$H_i =$ Statements about $a$, the probability for heads on the next flip $\rightarrow$ seek $p(a|D, M)$

$D =$ Sequence of results from $N$ previous flips:

THTHHHTHHHHT ($n = 8$ heads in $N = 12$ flips)

*Likelihood:*

$$\begin{aligned} p(D|a, M) &= p(\text{tails}|a, M) \times p(\text{heads}|a, M) \times \cdots \\ &= a^n (1-a)^{N-n} \\ &= \mathcal{L}(a) \end{aligned}$$

*Prior:*

Starting with no information about $a$ beyond its definition, use as an "uninformative" prior $p(a|M) = 1$. Justifications:

- Intuition: Don't prefer any $a$ interval to any other of same size
- Bayes's justification: "Ignorance" means that before doing the $N$ flips, we have no preference for how many will be heads:

$$P(n\,\text{heads}|M) = 1/N \rightarrow p(a|M) = 1$$

Consider this a *convention*—an assumption added to $M$ to make the problem well posed.

*Prior Predictive:*

$$\begin{aligned} p(D|M) &= \int da\; a^n (1-a)^{N-n} \\ &= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

A *Beta integral*, $B(a,b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

*Posterior:*

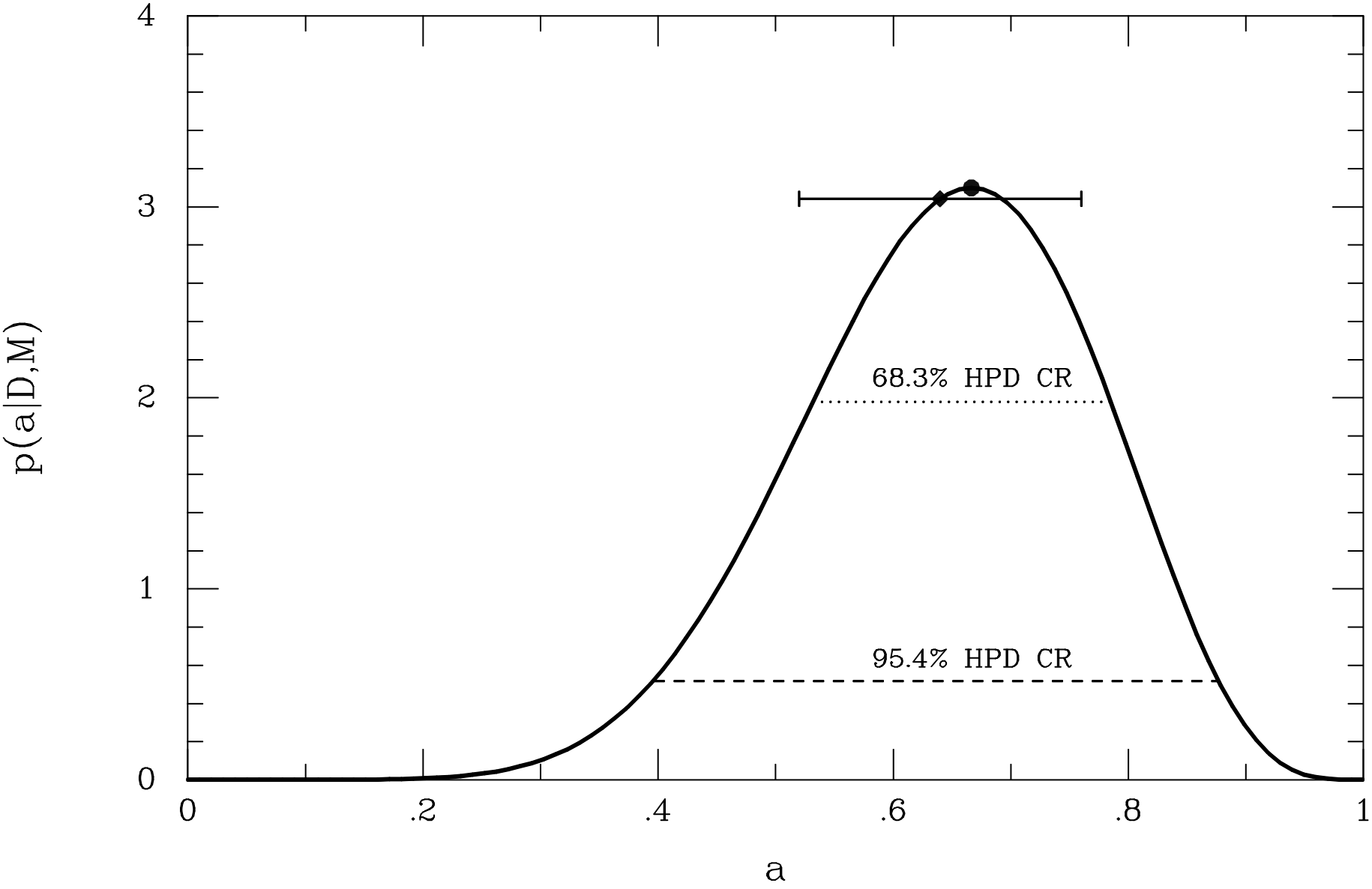
$$p(a|D,M) \;=\; \frac{(N+1)!}{n!(N-n)!}a^n(1-a)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: $\hat{a} = \frac{n}{N} = 2/3$; $\langle a \rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_a = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$

  Find credible regions numerically, or with incomplete beta function

Note that the posterior depends on the data only through $n$, not the $N$ binary numbers describing the sequence. $n$ is a (minimal) *Sufficient Statistic*.


4
3
2
1
0
p(a|D,M)
68.3% HPD CR
95.4% HPD CR
0
.2
.4
.6
.8
1
a

# Model Comparison: Fair Flips?

$M_1$: $a = 1/2$

$M_2$: $a \in [0, 1]$ with flat prior.

*Maximum Likelihoods:*

$$M_1 : \qquad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2 : \quad \mathcal{L}(2/3) \;=\; \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{\mathcal{L}(2/3)} = 0.51$$

Maximum likelihoods favor $M_2$ (biased flips).

*Bayes Factor (ratio of model likelihoods):*

$$p(D|M_1) = \frac{1}{2^N}; \qquad \text{and} \qquad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\begin{aligned} \rightarrow B_{12} \equiv \frac{p(D|M_1)}{p(D|M_2)} &= \frac{(N+1)!}{n!(N-n)!2^N} \\ &= 1.57 \end{aligned}$$

Bayes factor (odds) favors $M_1$ (fair flips).

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure the coin is fair.

If $n = 0$, $B_{12} \approx 1/315$; if $n = 3$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect the coin flipping is not fair.

# Coin Flipping: Binomial Distribution

Suppose $D = n$ (number of heads in $N$ flips), rather than the actual sequence. What is $p(a|n, M)$?

*Likelihood:*

Let $S =$ a sequence of flips with $n$ heads.

$$\begin{aligned} p(n|a, M) &= \sum_S p(S|a, M)p(n|S, a, M) \\ &= a^n(a-a)^{N-n}C_{n,N} \end{aligned}$$

$C_{n,N} =$ # of sequences of length $N$ with $n$ heads.

$$\rightarrow p(n|a, M) = \frac{N!}{n!(N-n)!}a^n(1-a)^{N-n}$$

The *Binomial Distribution* for $n$ given $a$, $N$.

*Posterior:*

$$p(a|n,M) = \frac{\frac{N!}{n!(N-n)!}a^n(1-a)^{N-n}}{p(n|M)}$$

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int da\; a^n(1-a)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

$$\rightarrow p(a|n,M) = \frac{(N+1)!}{n!(N-n)!}a^n(1-a)^{N-n}$$

Same result as when data specified the actual sequence.

# Lessons from Coin Flip Analyses

- Sufficiency: Calculation of the likelihood identified a sufficient statistic
- Only dependence of likelihood on *parameters* matters
- Demonstration of “Occam factors”:
  - Simpler model can be favored
  - Experiments can have limited strength

# Inference With Gaussians

*Gaussian PDF:*

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ over } [-\infty,\infty]$$

Common abbreviated notation: $x \sim N(\mu,\sigma^2)$

*Parameters:*

$$\mu = \langle x \rangle \equiv \int dx\; x\, p(x|\mu,\sigma)$$

$$\sigma^2 = \langle (x-\mu)^2 \rangle \equiv \int dx\; (x-\mu)^2\, p(x|\mu,\sigma)$$

# Gauss's Observation: Sufficiency

Suppose our data consist of $N$ measurements, $d_i = \mu + \epsilon_i$. Suppose the noise contributions are independent, and $\epsilon_i \sim N(0, \sigma^2)$.

$$
\begin{aligned}
p(D|\mu, \sigma, M) &= \prod_i p(d_i|\mu, \sigma, M) \\
&= \prod_i p(\epsilon_i = d_i - \mu|\mu, \sigma, M) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q/2\sigma^2}
\end{aligned}
$$

Find dependence of $Q$ on $\mu$ by completing the square:

$$\begin{aligned} Q &= \sum_i (d_i - \mu)^2 \\ &= \sum_i d_i^2 + N\mu^2 - 2N\mu\bar{d} \qquad \text{where } \bar{d} \equiv \frac{1}{N}\sum_i d_i \\ &= N(\mu - \bar{d})^2 + Nr^2 \quad \text{where } r^2 \equiv \frac{1}{N}\sum_i (d_i - \bar{d})^2 \end{aligned}$$

Likelihood depends on $\{d_i\}$ only through $\bar{d}$ and $r$:

$$\mathcal{L}(\mu, \sigma) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right)$$

The sample mean and variance are *sufficient statistics*.

# Estimating a Normal Mean

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $\sigma$ is known $\rightarrow I = (\sigma, M)$.

Parameter space: $\mu$; seek $p(\mu|D, \sigma, M)$

*Likelihood:*

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \end{aligned}$$

*"Uninformative" prior:*

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.
This prior is *improper* unless bounded.

*Prior predictive:*

$$\begin{aligned} p(D|\sigma, M) &= \int d\mu \, C \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \\ &= C(\sigma/\sqrt{N})\sqrt{2\pi} \end{aligned}$$

. . . minus a tiny bit from tails, using a proper prior.

*Posterior:*

$$p(\mu|D,\sigma,M) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} \exp\left(-\frac{N(\mu-\bar{d})^2}{2\sigma^2}\right)$$

Posterior is Gaussian at $\bar{d}$ with standard deviation $w = \sigma/\sqrt{N}$.

68.3% HPD credible region for $\mu$ is $\bar{d} \pm \sigma/\sqrt{N}$.

Note that $C$ drops out $\rightarrow$ limit of infinite prior range is well behaved.

*Informative Conjugate Prior:*

Use a Gaussian prior, $\mu \sim N(\mu_0, w_0^2)$

*Posterior:*

Remains Gaussian with

$$\begin{aligned} \hat{\mu} &= \frac{\bar{d}}{1 + \frac{w^2}{w_0^2}} + \frac{\mu_0}{1 + \frac{w_0^2}{w^2}} \\ w' &= w \frac{1}{\sqrt{1 + w^2/w_0^2}} \end{aligned}$$

"Principle of stable estimation:" The prior affects inferences only when data are not informative.

# Estimating a Normal Mean: Unknown $\sigma$

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $\sigma$ is *unknown*

Parameter space: $(\mu, \sigma)$; seek $p(\mu|D, \sigma, M)$

*Likelihood:*

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \\ &\propto \frac{1}{\sigma^N} e^{-Q/2\sigma^2} \end{aligned}$$

where $Q = N\left[r^2 + (\mu - \bar{d})^2\right]$

*Uninformative Priors:*

Assume priors for $\mu$ and $\sigma$ are independent.

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.

Scale invariance $\Rightarrow p(\sigma) \propto 1/\sigma$.

*Joint Posterior for $\mu$, $\sigma$:*

$$p(\mu, \sigma | D, M) \propto \frac{1}{\sigma^{N+1}} e^{-Q/2\sigma^2}$$

## *Marginal Posterior:*

$$p(\mu|D,M) \propto \int d\sigma \; \frac{1}{\sigma^{N+1}} e^{-Q/2\sigma^2}$$

Let $\tau = \frac{Q}{2\sigma^2}$ so $\sigma = \sqrt{\frac{Q}{2\tau}}$ and $|d\sigma| = \tau^{-3/2}\sqrt{\frac{Q}{2}}$

$$\begin{aligned} \Rightarrow p(\mu|D,M) &\propto 2^{N/2} Q^{-N/2} \int d\tau \; \tau^{\frac{N}{2}-1} e^{-\tau} \\ &\propto Q^{-N/2} \end{aligned}$$

Write $Q = Nr^2\left[1 + \left(\frac{\mu - \bar{d}}{r}\right)^2\right]$ and normalize:

$$p(\mu|D,M) = \frac{\left(\frac{N}{2}-1\right)!}{\left(\frac{N}{2}-\frac{3}{2}\right)!\sqrt{\pi}}\frac{1}{r}\left[1 + \frac{1}{N}\left(\frac{\mu - \bar{d}}{r/\sqrt{N}}\right)^2\right]^{-N/2}$$

"Student's $t$ distribution," with $t = \frac{(\mu - \bar{d})}{r/\sqrt{N}}$

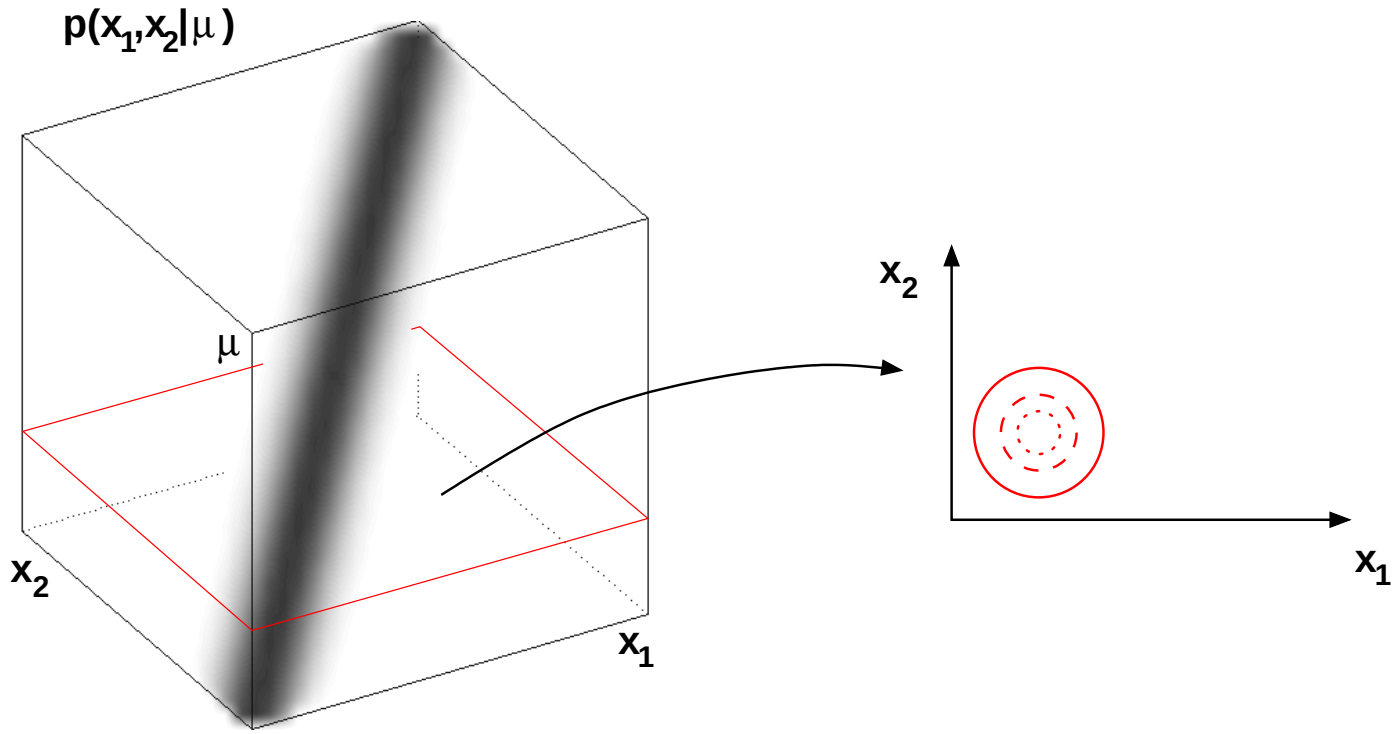
A "bell curve," but with power-law tails

Large $N$:

$$p(\mu|D,M) \sim e^{-N(\mu-\bar{d})^2/2r^2}$$

# A Frequentist Confidence Region

Infer $\mu$ : $\quad x_i = \mu + \epsilon_i; \qquad p(x_i|\mu, M) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(x_i - \mu)^2}{2\sigma^2}\right]$




68% confidence region: $\bar{x} \pm \sigma/\sqrt{N}$

## Monte Carlo Algorithm:

1. Pick a null hypothesis, $\mu = \mu_0$
2. Draw $x_i \sim N(\mu_0, \sigma^2)$ for $i = 1$ to $N$
3. Find $\bar{x}$; check if $\mu_0 \in \bar{x} \pm \sigma/\sqrt{N}$
4. Repeat $M >> 1$ times; report fraction ($\approx 0.683$)
5. *Hope result is independent of $\mu_0$!*

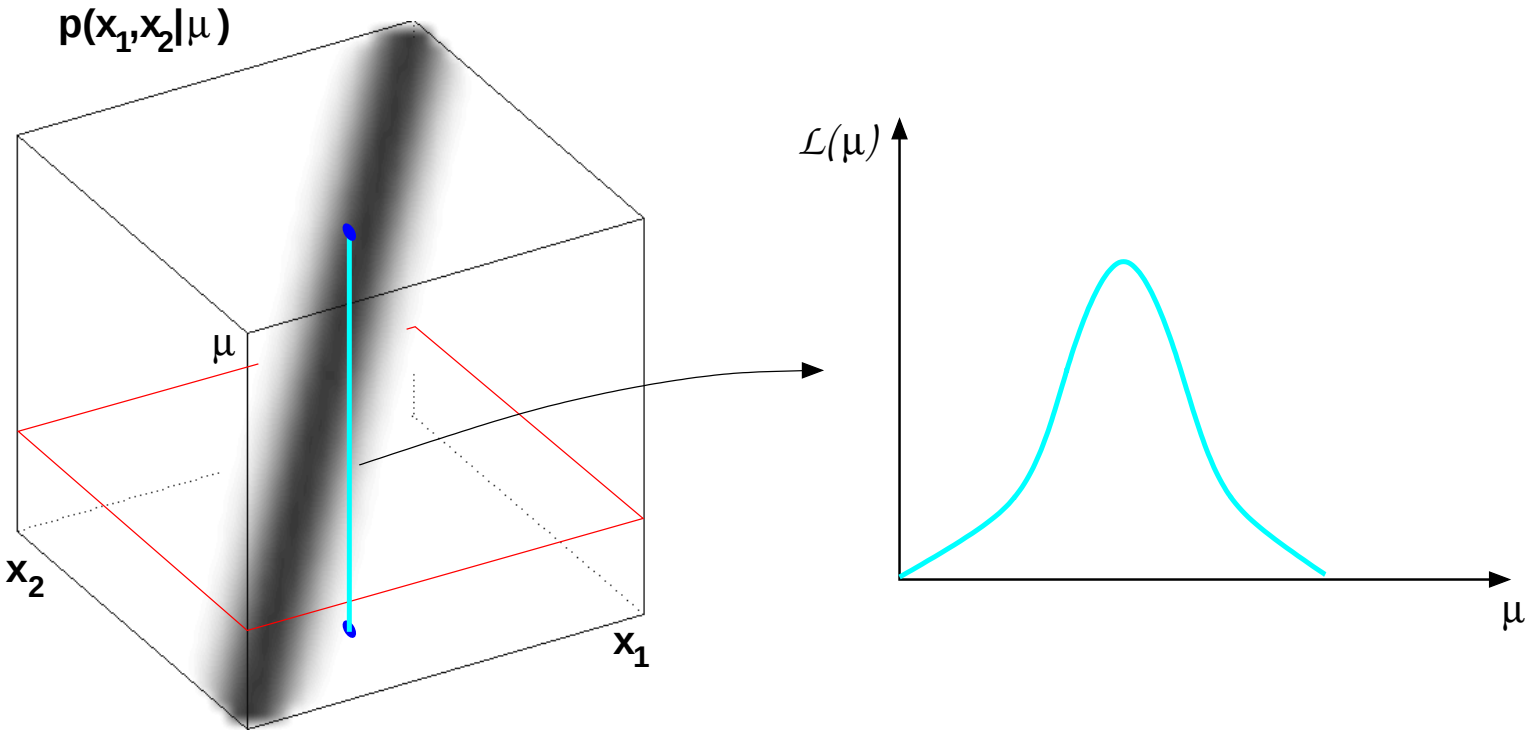
A Monte Carlo calculation of the $N$-dimensional integral:

$$\int dx_1 \frac{e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \cdots \int dx_N \frac{e^{-\frac{(x_N-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \times [\mu_0 \in \bar{x} \pm \sigma/\sqrt{N}]$$

$$= \int d(\text{angles}) \int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\bar{x} \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right] \cdots \approx 0.683$$

# A Bayesian Credible Region

Infer $\mu$ : Flat prior; $\mathcal{L}(\mu) \propto \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]$




68% credible region: $\bar{x} \pm \sigma/\sqrt{N}$

68% credible region: $\bar{x} \pm \sigma/\sqrt{N}$

$$\frac{\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\mu \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]}{\int_{-\infty}^{\infty} d\mu \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]} \approx 0.683$$

Equivalent to a Monte Carlo calculation of a 1-d integral:

1. Draw $\mu$ from $N(\bar{x}, \sigma^2/N)$ (i.e., prior $\times \mathcal{L}$)
2. Repeat $M >> 1$ times; histogram
3. Report most probable 68.3% region

This simulation uses hypothetical *hypotheses* rather than hypothetical *data*.

# When Will Results Differ?

When models are linear in the parameters and have additive Gaussian noise, frequentist results are identical to Bayesian results with flat priors.

This mathematical coincidence will *not* occur if:

- The choice of statistic is not obvious (no sufficient statistics)
- There is no identity between parameter space and sample space integrals (due to nonlinearity or the form of the sampling distribution)
- There is important prior information

Also, some problems can be quantitatively addressed only from the Bayesian viewpoint; e.g., systematic error.

# The Bretthorst Algorithm

## *Superposed Nonlinear Models*

$N$ samples of a superpos'n of nonlinear functions plus Gaussian errors,

$$\begin{aligned} d_i &= \sum_{\alpha=1}^{M} A_\alpha g_\alpha(x_i;\theta) + \epsilon_i \\ \text{or} \qquad \vec{d} &= \sum_{\alpha} A_\alpha \vec{g}_\alpha(\theta) + \vec{\epsilon}. \end{aligned}$$

E.g., sinusoidal signal:

$$\begin{aligned} f(x) &= A\cos(\omega x + \phi) \\ &= A_1 \cos\omega x + A_2 \sin\omega x \\ &= A_1 g_1(x,\omega) + A_2 g_2(x,\omega) \end{aligned}$$

The log-likelihood is a quadratic form in $A_\alpha$,

$$\mathcal{L}(A,\theta) \propto \frac{1}{\sigma^N}\exp\left[-\frac{Q(A,\theta)}{2\sigma^2}\right]$$

$$\text{with} \quad Q = \left[\vec{d} - \sum_\alpha A_\alpha \vec{g}_\alpha\right]^2$$

$$= d^2 - 2\sum_\alpha A_\alpha \vec{d}\cdot\vec{g}_\alpha + \sum_{\alpha,\beta} A_\alpha A_\beta \eta_{\alpha\beta}$$

$$\text{where} \quad \eta_{\alpha\beta} = \vec{g}_\alpha \cdot \vec{g}_\beta$$

Goals:

- Estimate $\theta$ given a prior, $\pi(\theta)$.
- Estimate amplitudes.
- Compare rival models.

*The algorithm:*

- Switch to orthonormal set of models, $\vec{h}_\mu(\theta)$ by diagonalizing $\eta_{\alpha\beta}$; new amplitudes $B = \{B_\mu\}$.

$$Q = \sum_\mu \left[ B_\mu - \vec{d} \cdot \vec{h}_\mu(\theta) \right]^2 + r^2(\theta, B)$$

residual $\qquad \vec{r}(\theta, B) = \vec{d} - \sum_\mu B_\mu \vec{h}_\mu$

$$p(B, \theta | D, I) \quad \propto \quad \frac{\pi(\theta) J(\theta)}{\sigma^N} \exp\left[ -\frac{r^2}{2\sigma^2} \right] \exp\left[ \frac{-1}{2\sigma^2} \sum_\mu (B_\mu - \hat{B}_\mu)^2 \right]$$

where $\qquad J(\theta) = \prod_\mu \lambda_\mu(\theta)^{-1/2}$

- Marginalize $B$'s analytically.

$$p(\theta|D,I) \propto \frac{\pi(\theta)J(\theta)}{\sigma^{N-M}} \exp\left[-\frac{r^2(\theta)}{2\sigma^2}\right]$$

$$r^2(\theta) = \text{residual sum of squares from least squares}$$

- If $\sigma$ unknown, marginalize using $p(\sigma|I) \propto \frac{1}{\sigma}$.

$$p(\theta|D,I) \propto \pi(\theta)J(\theta)\left[r^2(\theta)\right]^{\frac{M-N}{2}}$$

# Detecting Periodic Signals

*Conventional approach:*

FT of signal $\approx$ DFT of data

$$\rightarrow \text{use } S(\omega) = \frac{2}{N}\left[\left(\sum_i d_i \cos\omega t_i\right)^2 + \left(\sum_i d_i \sin\omega t_i\right)^2\right]$$

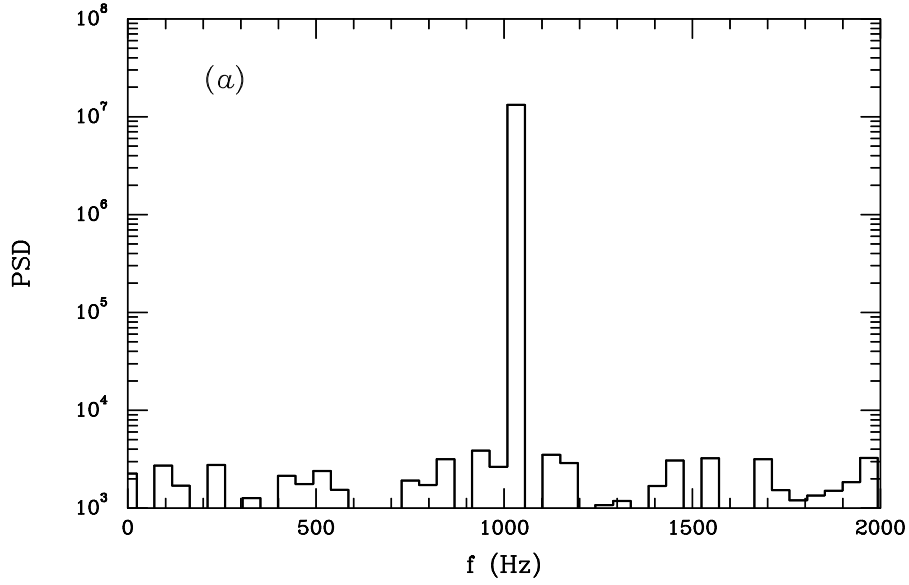
Schuster periodogram (evenly spaced data)
Lomb-Scargle periodogram (unevenly spaced data)

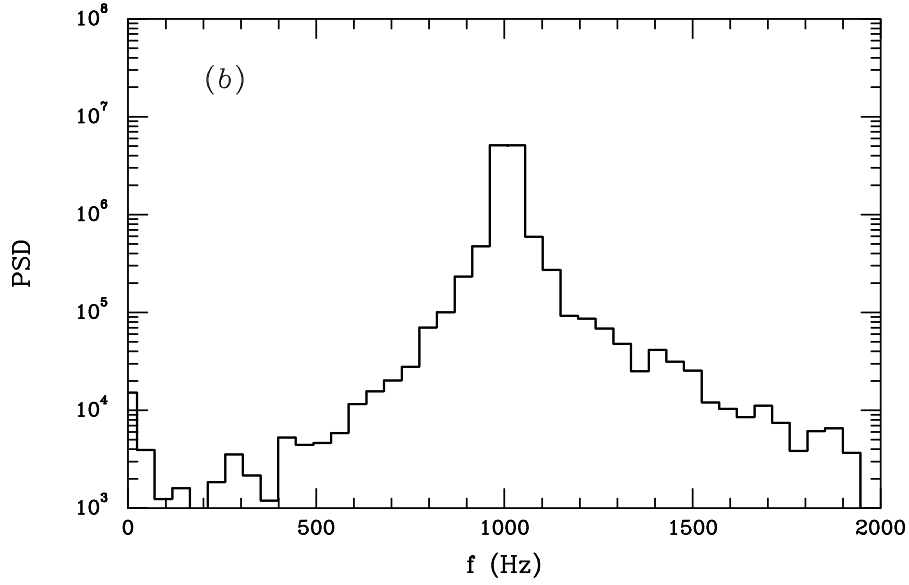
Interpret data power spectrum as signal power spectrum corrupted by noise and sampling “window”

Under simple null hypotheses, $N/2$ values given by DFT are *statistically independent* $\Rightarrow$ focus on those

# Problem: Leakage in the Power Spectrum

1024 samples at 48 kHz sampling rate
$S/N = 5$, white noise
$f = 1031.25$ Hz (Fourier frequency) and $f = 1008$ Hz




$\Rightarrow$ Attempt to reduce leakage by tapering/smoothing data

*Bayesian approach (Jaynes & Bretthorst):*

Explicit periodic signal model:

$$\begin{aligned} f(t) &= A\cos(\omega t - \phi) && \text{parameters } \omega, A, \phi \\ &= A_1 \cos\omega t + A_2 \sin\omega t && \text{parameters } \omega, A_1, A_2 \end{aligned}$$

$$d_i = f(t_i) + e_i \qquad \text{Gaussian error } \textit{pdf}\text{s; rms} = \sigma$$

Estimate $\omega$ *assuming signal present*:

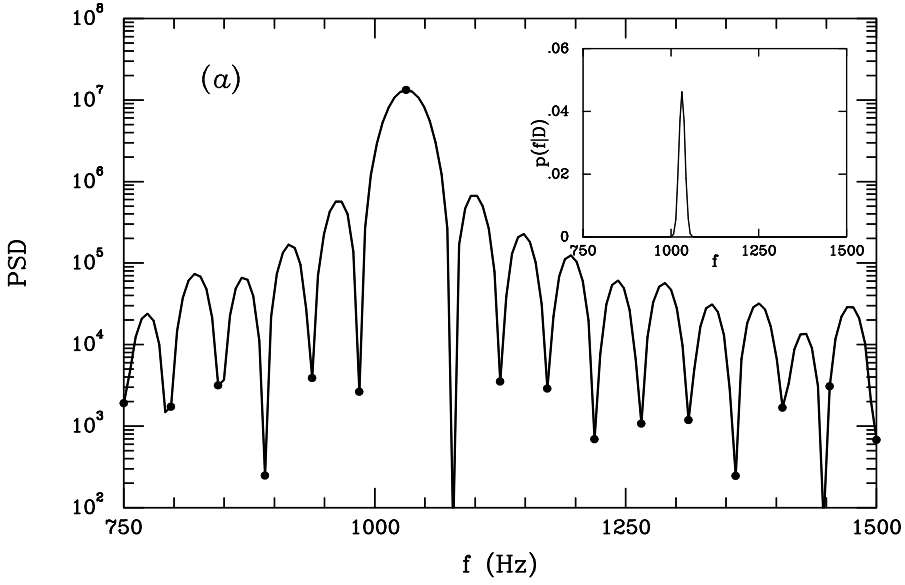
$$\begin{aligned} p(\omega|D) &\propto \int dA_1 \int dA_2 \; p(\omega, A_1, A_2)\mathcal{L}(\omega, A_1, A_2) \\ &\propto p(\omega) J(\omega) \exp\left[\frac{S(\omega)}{\sigma^2}\right] \end{aligned}$$

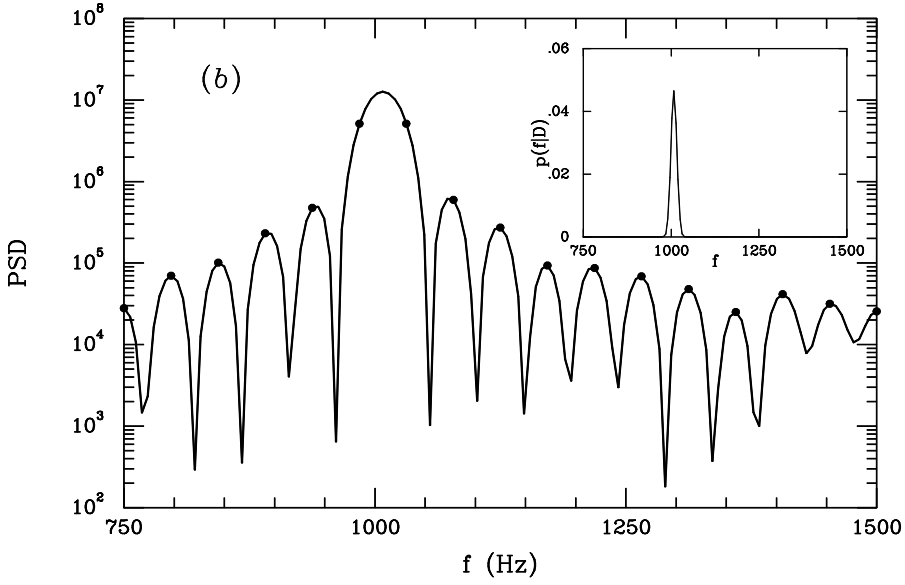
# Continuous Periodogram and Posterior Distribution for $\omega$

1024 samples at 48 kHz sampling rate

$S/N = 5$

$f = 1031.25$ Hz (Fourier frequency) and $f = 1008$ Hz



*New aspects:*

- Periodogram is not a power spectrum, but the logarithm of the marginal *pdf* for $\omega$
- No special role for Fourier frequencies
- No "leakage;" log *pdf* has similar structure for *all* signal frequencies, but sidelobes get *exponentiated* away.
- Detect signal using signal model likelihood:

$$\mathcal{L}(signal) \approx \exp\left[\frac{S_{\max}}{\sigma^2}\right] \times \frac{\text{peak width}}{\text{prior search range}}$$

- Conventional periodograms optimal only for *single sinusoids*
- Noise estimated using *peak amplitude* (rms residual)

# Astrophysical Applications

*Bayes* + *Gaussians* + *linear & nonlinear parameters*

- Periodic signals:
  - Cepheid/RR Lyrae variables — Bretthorst, Jefferys & Berger
  - X-ray flare stars — Gregory
  - Planet detection — TJL, Scargle, Bretthorst
- Gravitational radiation from binary inspiral (LIGO) — Finn, Flanagan, Cutler...
- Cosmic background radiation (signal=variance) — Readhead, Lasenby, Bond, Jaffe...

# Key Ideas

- Sufficiency—For some problems, inferences depend on only a few features of the data. The Bayesian algorithm identifies these automatically.
- For simple Gaussian problems, Bayesian and frequentist results are similar; but this is a mathematical coincidence.
- A powerful modeling framework: Superposed nonlinear models $+$ Gaussian noise
- A new interpretation of the Fourier power spectrum with significant practical consequences