Introduction to Bayesian inference in astronomy

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://www.astro.cornell.edu/staff/loredo/bayes/

CASt Summer School — 2–3 June 2016
Agenda

1 Motivating example: \( \bar{x} \pm \sigma / \sqrt{N} \)
   Confidence intervals vs. credible intervals
   A tale of two spaces

2 Probability theory as generalized logic

3 Probability theory for data analysis: Three theorems

4 Inference with parametric models
   Parameter Estimation
   Model Uncertainty

5 Multilevel models for measurement error

6 Bayesian computation
Agenda

1 Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$
   Confidence intervals vs. credible intervals
   A tale of two spaces

2 Probability theory as generalized logic

3 Probability theory for data analysis: Three theorems

4 Inference with parametric models
   Parameter Estimation
   Model Uncertainty

5 Multilevel models for measurement error

6 Bayesian computation
Inference With Parametric Models

Models $M_i$ ($i = 1$ to $N$), each with parameters $\theta_i$, each imply a sampling dist’n (conditional predictive dist’n for possible data):

$$p(D|\theta_i, M_i)$$

The $\theta_i$ dependence when we fix attention on the observed data is the likelihood function:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about $i$ (model uncertainty) or $\theta_i$ (parameter uncertainty).

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{\text{obs}}$. 
Classes of Problems

Single-model inference

Premise = choice of single model (specific $i$)

*Parameter estimation*: What can we say about $\theta_i$ or $f(\theta_i)$?

*Prediction*: What can we say about future data $D'$?

Multi-model inference

Premise = $\{M_i\}$

*Model comparison/choice*: What can we say about $i$?

*Model averaging*:
  - *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; $\phi$ is common to all
    What can we say about $\phi$ w/o committing to one model?
  - *Prediction*: What can we say about future $D'$, accounting for model uncertainty?

Model checking

Premise = $M_1 \lor$ “all” alternatives

Is $M_1$ adequate? (predictive tests, calibration, robustness)
**Parameter Estimation**

**Problem statement**

\[ I = \text{Model } M \text{ with parameters } \theta \text{ (+ any add’l info)} \]

\[ H_i = \text{statements about } \theta; \text{ e.g. } \theta \in [2.5, 3.5], \text{ or } \theta > 0 \]

Probability for any such statement can be found using a *probability density function* (PDF) for \( \theta \):

\[
P(\theta \in [\theta, \theta + d\theta] | \cdots) = f(\theta) d\theta
\]

\[= p(\theta | \cdots) d\theta\]

**Posterior probability density**

\[
p(\theta | D, M) = \frac{p(\theta | M) \mathcal{L}(\theta)}{\int d\theta \ p(\theta | M) \mathcal{L}(\theta)}
\]
Summaries of posterior

- “Best fit” values:
  - Mode, \( \hat{\theta} \), maximizes \( p(\theta|D, M) \)
  - Posterior mean, \( \langle \theta \rangle = \int d\theta \theta p(\theta|D, M) \)

- Uncertainties:
  - Credible region \( \Delta \) of probability \( C \):
    \[
    C = P(\theta \in \Delta|D, M) = \int_{\Delta} d\theta \, p(\theta|D, M)
    \]
    Highest Posterior Density (HPD) region has \( p(\theta|D, M) \) higher inside than outside
  - Posterior standard deviation, variance, covariances

- Marginal distributions
  - Interesting parameters \( \phi \), nuisance parameters \( \eta \)
  - Marginal dist’n for \( \phi \):
    \[
    p(\phi|D, M) = \int d\eta \, p(\phi, \eta|D, M)
    \]
Estimating a Normal Mean

Problem specification

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

\[
L(\mu) \equiv p(D|\mu, \sigma, M) = \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-(d_i-\mu)^2/2\sigma^2}; \quad \sigma = 1
\]

\[
\propto \exp \left( - \frac{N(\mu - \bar{d})^2}{2\sigma^2} \right)
\]

Likelihood function is a Gaussian function at \( \bar{d} \), width \( w = \sigma/\sqrt{N} \)
**Informative Conjugate Prior**

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

*Conjugate* because the posterior turns out also to be normal.

$w_0 \to \infty$ is the “uninformative” flat prior limit; posterior remains normal and proper (normalizable)

**Posterior**

Normal $N(\tilde{\mu}, \tilde{w}^2)$, but mean, std. deviation “shrink” towards prior.

Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when $w_0$ is large.

Then

\[
\tilde{\mu} = \bar{d} + B \cdot (\mu_0 - \bar{d}) \\
\tilde{w} = w \cdot \sqrt{1 - B}
\]

“*Principle of stable estimation*” — The prior affects estimates only when data are not informative relative to prior
Conjugate normal examples:

- Data have $\bar{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$
Supplement:

- Binomial example
  - Bernoulli trials, binomial & negative binomial dist’ns
  - Beta-binomial conjugate model
  - Likelihood principles

- Normal example
  - Analytical details for normal example
  - Sufficiency; sample mean and variance as sufficient statistics
  - Connection to least-squares curve fitting
  - Handling $\sigma$ uncertainty by marginalizing over $\sigma$; Student’s $t$ distribution
To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

**Example**

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal $s$ and a background $b$.

We have additional data just about $b$.

What do the data tell us about $s$?
Marginal posterior distribution

To summarize implications for $s$, accounting for $b$ uncertainty, marginalize:

$$p(s|D, M) = \int db \ p(s, b|D, M)$$

$$\propto p(s|M) \int db \ p(b|s, M) \mathcal{L}(s, b)$$

$$= p(s|M) \mathcal{L}_m(s)$$

with $\mathcal{L}_m(s)$ the marginal likelihood function for $s$:

$$\mathcal{L}_m(s) \equiv \int db \ p(b|s) \mathcal{L}(s, b)$$
Marginalization vs. Profiling

For insight: Suppose the prior is broad compared to the likelihood for a fixed $s$, we can accurately estimate $b$ with max likelihood $\hat{b}_s$, with small uncertainty $\delta b_s$.

\[
\mathcal{L}_m(s) \equiv \int db \ p(b|s) \mathcal{L}(s, b)
\]

\[
\approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s
\]

Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a parameter space volume factor

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}, \quad \sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background subtraction is a special case of background marginalization.
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$

$\delta b_s$ is const. vs. $s$

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$
Flared/skewed/banana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ differ

General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$. Otherwise, they will likely differ.

In “measurement error problems” the difference can be dramatic
The On/Off Problem for Poisson counting data

**Basic problem**

- Look off-source; unknown background rate \( b \)
  Count \( N_{\text{off}} \) photons in interval \( T_{\text{off}} \)

- Look on-source; rate is \( r = s + b \) with unknown signal \( s \)
  Count \( N_{\text{on}} \) photons in interval \( T_{\text{on}} \)

- Infer \( s \)

**Conventional solution**

\[
\hat{b} = \frac{N_{\text{off}}}{T_{\text{off}}}; \quad \sigma_b = \sqrt{\frac{N_{\text{off}}}{T_{\text{off}}}}
\]

\[
\hat{r} = \frac{N_{\text{on}}}{T_{\text{on}}}; \quad \sigma_r = \sqrt{\frac{N_{\text{on}}}{T_{\text{on}}}}
\]

\[
\hat{s} = \hat{r} - \hat{b}; \quad \sigma_s = \sqrt{\sigma_r^2 + \sigma_b^2}
\]

But \( \hat{s} \) can be *negative!*
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

Di Salvo et al. 2001
Sample sizes are never large. If \( N \) is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once \( N \) is ‘large enough,’ you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). \( N \) is never enough because if it were ‘enough’ you’d already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)
Sample sizes are never large. If $N$ is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once $N$ is ‘large enough,’ you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). $N$ is never enough because if it were ‘enough’ you’d already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)
Bayesian Solution to On/Off Problem

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

\[
\mathcal{L}(s, b) = \frac{(b T_{\text{off}})^{N_{\text{off}}}}{N_{\text{off}}!} e^{-b T_{\text{off}}} \times \frac{[(s + b) T_{\text{on}}]^{N_{\text{on}}}}{N_{\text{on}}!} e^{-(s+b) T_{\text{on}}}
\]

 Adopting flat priors for \((s, b)\), the joint posterior is

\[
p(s, b|N_{\text{on}}, N_{\text{off}}, I) \propto (s + b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-s T_{\text{on}}} e^{-(s+b) T_{\text{on}}}
\]

If \(b = 0\), the (normalized) posterior distribution is a gamma distribution,

\[
p(s, b = 0|N_{\text{on}}, N_{\text{off}}, I) = \frac{T_{\text{on}}(s T_{\text{on}})^{N_{\text{on}}}}{N_{\text{on}}!} e^{-s T_{\text{on}}}
\]
Now marginalize over $b$;

$$p(s|N_{on}, N_{off}, I) = \int db \ p(s, b | N_{on}, I_{all})$$

$$\propto \int db \ (s + b)^{N_{on}} b^{N_{off}} e^{-sT_{on}} e^{-b(T_{on} + T_{off})}$$

Expand $(s + b)^{N_{on}}$ and do the resulting $\Gamma$ integrals:

$$p(s|N_{on}, I_{all}) = \sum_{i=0}^{N_{on}} C_i \frac{T_{on}(sT_{on})^i e^{-sT_{on}}}{i!}$$

$$C_i \propto \left(1 + \frac{T_{off}}{T_{on}}\right)^i \frac{(N_{on} + N_{off} - i)!}{(N_{on} - i)!}$$

Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)
Example On/Off Posteriors—Short Integrations

$T_{on} = 1$

$T_{off} = 1, \ N_{off} = 9$

$N_{on} = 6$

$N_{on} = 9$

$N_{on} = 16$

$p(s)$

$s \ (s^{-1})$
Example On/Off Posteriors—Long Background Integrations

\[ T_{\text{on}} = 1 \]

\[ T_{\text{off}} = 1, \quad N_{\text{off}} = 9 \]

\[ T_{\text{off}} = 10, \quad N_{\text{off}} = 90 \]

\[ N_{\text{on}} = 6 \]

\[ N_{\text{on}} = 9 \]

\[ N_{\text{on}} = 16 \]
Supplement:

- Analytical details for Poisson dist’n inference
- Gamma-Poisson conjugate model
- Alternative (equivalent) solutions to the on/off problem
- Multibin case
Bayesian Curve Fitting & Least Squares

Setup

Data $D = \{d_i\}$ are measurements of an underlying function $f(x; \theta)$ at $N$ sample points $\{x_i\}$. Let $f_i(\theta) \equiv f(x_i; \theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2)$$

We seek learn $\theta$, or to compare different functional forms (model choice, $M$)

Likelihood

$$p(D|\theta, M) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right]$$

$$\propto \exp \left[ -\frac{1}{2} \sum_i \left( \frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right]$$

$$= \exp \left[ -\frac{\chi^2(\theta)}{2} \right]$$
Posterior

For prior density $\pi(\theta)$,

$$p(\theta|D, M) \propto \pi(\theta) \exp \left[ -\frac{\chi^2(\theta)}{2} \right]$$

If you have a least-squares or $\chi^2$ code:

- Treat $\chi^2(\theta)$ as $-2 \log L(\theta)$

- Bayesian inference amounts to exploration and numerical integration of $\pi(\theta)e^{-\chi^2(\theta)/2}$
Important Case: Separable Nonlinear Models

A (linearly) separable model has parameters $\theta = (A, \psi)$:

- Linear amplitudes $A = \{A_\alpha\}$
- Nonlinear parameters $\psi$

$f(x; \theta)$ is a linear superposition of $M$ nonlinear components $g_\alpha(x; \psi)$:

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

or

$$\vec{d} = \sum_{\alpha} A_\alpha \vec{g}_\alpha(\psi) + \vec{\epsilon}.$$

Why this is important: You can marginalize over $A$ analytically

→ Bretthorst algorithm ("Bayesian Spectrum Analysis & Param. Est’n" 1988)

Algorithm is closely related to linear least squares, diagonalization, SVD; for sinusoidal $g_\alpha$, generalizes periodograms
Many Roles for Marginalization

Eliminate nuisance parameters

\[ p(\phi|D, M) = \int d\eta \ p(\phi, \eta|D, M) \]

Propagate uncertainty

Model has parameters \( \theta \); what can we infer about \( F = f(\theta) \)?

\[ p(F|D, M) = \int d\theta \ p(F, \theta|D, M) = \int d\theta \ p(\theta|D, M) \ p(F|\theta, M) \]

\[ = \int d\theta \ p(\theta|D, M) \delta[F - f(\theta)] \quad \text{[single-valued case]} \]

Prediction

Given a model with parameters \( \theta \) and present data \( D \), predict future data \( D' \) (e.g., for experimental design):

\[ p(D'|D, M) = \int d\theta \ p(D', \theta|D, M) = \int d\theta \ p(\theta|D, M) \ p(D'|\theta, M) \]

Model comparison...
Model Comparison

**Problem statement**

\[ I = (M_1 \lor M_2 \lor \ldots) \] — Specify a set of models.
\[ H_i = M_i \] — Hypothesis chooses a model.

**Posterior probability for a model**

\[
p(M_i|D, I) = p(M_i|I) \frac{p(D|M_i, I)}{p(D|I)}
\]

\[
\propto p(M_i|I) L(M_i)
\]

\[
\mathcal{L}(M_i) = p(D|M_i) = \int d\theta_i \ p(\theta_i|M_i)p(D|\theta_i, M_i).
\]

Likelihood for model = Average likelihood for its parameters

\[
\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle
\]

Varied terminology: Prior predictive = Average likelihood = Global likelihood = Marginal likelihood = (Weight of) Evidence for model
Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the **odds** favoring one over the other:

\[
O_{ij} \equiv \frac{p(M_i|D, I)}{p(M_j|D, I)} = \frac{p(M_i|I) \times p(D|M_i, I)}{p(M_j|I) \times p(D|M_j, I)}
\]

The data-dependent part is called the **Bayes factor**:

\[
B_{ij} \equiv \frac{p(D|M_i, I)}{p(D|M_j, I)}
\]

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods.
An Automatic Ockham’s Razor

Consider *nested models*:

- Simpler model $M_1$ with parameters $\theta_1$
- “Larger” rival $M_2$ with parameters $\theta_2 = (\theta_1, \eta)$

⇒ $\mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$

But what about $p(D|M_i) = \int d\theta_i \ p(\theta_i|M) \ \mathcal{L}(\theta_i)$?

*Prior predictive distributions*

Normalization implies *there must be data that favor $M_1$*:
The Ockham Factor

Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for “wasted” *volume of parameter space*

Quantifies intuition that models shouldn’t require fine-tuning
**Bayes factors vs. hypothesis testing**

\[ p(D | M_i) = \int d\theta_i \, p(\theta_i | M) \, L(\theta_i) \approx p(\hat{\theta}_i | M) L(\hat{\theta}_i) \delta \theta_i \]

\[ \approx \frac{\delta \theta_i}{\Delta \theta_i} \]

\[ = \text{Maximum Likelihood} \times \text{Ockham Factor} \]

Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for “wasted” volume of parameter space

Quantifies intuition that models shouldn’t require fine-tuning
Neyman-Pearson test with Type I error rate $\alpha$

- Optimize over $\theta$
- Also compute Type II rate for alternative, $H_1$
- $p(S|H_0)$
- $S(D_{hyp})$
- $S_{crit,\alpha}$
- $S(D_{obs})$

Fisherian $p$-value

- Optimize over $\theta$
- $p(S|H_0)$
- $S(D_{hyp})$
- $S(D_{obs})$
- $\alpha$
- $p$

Bayes factors

- $p(D|H_0)$
- $p(D|\theta, H_0)$
- $D_{hyp}$
- $D_{obs}$

- NP & Fisher give $H_0$ a special role
- NP & Fisher optimize over $\theta$, integrate over $D_{hyp}$
- Bayes considers rival $H_i$ symmetrically
- Bayes integrates over $\theta$, uses only $D_{obs}$

Bayes factors can only compare rival models; they don’t measure “goodness-of-fit”

Posterior predictive $p$-values are a BDA alternative for measuring “surprisingness” of data for model checking; they integrate over both data and parameter spaces
Neyman-Pearson test with Type I error rate $\alpha$

- Optimize over $\theta$
- Also compute Type II rate for alternative, $H_1$

$$p(S|H_0)$$

$$S(D_{hyp})$$

$$S_{crit, \alpha}$$

$$S(D_{obs})$$

Fisherman $p$-value

- Optimize over $\theta$

$$p(S|H_0)$$

$$S(D_{hyp})$$

$$S(D_{obs})$$

Bayes factors

- NP & Fisher give $H_0$ a special role
- NP & Fisher optimize over $\theta$, integrate over $D_{hyp}$
- Bayes considers rival $H_1$ symmetrically
- Bayes integrates over $\theta$, uses only $D_{obs}$

Bayes factors can only compare rival models; they don’t measure “goodness-of-fit”

Posterior predictive $p$-values are a BDA alternative for measuring “suprisingness” of data for model checking; they integrate over both data and parameter spaces

See “$p$-value note” online
Sirius A
Rigel
Betelgeuse
Moon
Pleiades
Hyades
M42
Crab
4U 0614+09
V711 Tau
Geminga
Dennerl & Voges '94
Orion region, optical

Sirius B
Rigel
Betelgeuse
Moon
Pleiades
Hyades
M42
Crab
4U 0614+09
V711 Tau
Orion region, X-ray
(ROSAT)

Dennerl & Voges '94
Bayesian Coincidence Assessment

\[ p(d_1, d_2 | H_0) = \int d n_1 p(n_1 | H_0) \ell_1(n_1) \times \int d n_2 \ldots \]

\[ p(d_1, d_2 | H_1) = \int d n p(n | H_1) \ell_1(n) \ell_2(n) \]
Doublet Bayes factor behavior vs. nearest-neighbor $p$-value

![Graph showing the relationship between separation angle and Bayes factor/nearest-neighbor $p$-value for different values of $\sigma$.]
Challenge: Large hypothesis spaces

For $N = 2$ events, there was a single coincidence hypothesis, $H_1$

For $N = 3$ events:

- Three doublets: $1 + 2$, $1 + 3$, or $2 + 3$
- One triplet

The number of alternatives (partitions, $\mathcal{P}$) grows combinatorially!

- **Model building**: Assign sensible priors to partitions
- **Computation**: Find & sum over important partitions

**IVOA Open SkyQuery** implements catalog cross-matching using Bayesian model comparison

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space.
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters.
- Model likelihoods have Ockham factors resulting from parameter space volume factors.

Many virtues of Bayesian methods can be attributed to this accounting for the “size” of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added “by hand”).
Roles of the prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from “intensity” to “measure”  
  → account for size of parameter space

Physical analogy

\[
\text{Heat } Q = \int \, dr \, c_v(r) \, T(r)
\]

\[
\text{Probability } P \propto \int \, d\theta \, p(\theta) \, \mathcal{L}(\theta)
\]

Maximum likelihood focuses on the “hottest” parameters.
Bayes focuses on the parameters with the most “heat.”

A high-\(T\) region may contain little heat if its \(c_v\) is low or if its volume is small.

A high-\(\mathcal{L}\) region may contain little probability if its prior is low or if its volume is small.
Supplement:

- Assigning priors
- Rule-based non-informative priors: Jeffreys, reference
Recap of Key Ideas

*Probability as generalized logic*

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

*Bayes’s theorem*

\[ p(\text{Hypothesis} | \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} | \text{Hypothesis}) \]

Data *change* the support for a hypothesis \( \propto \) ability of hypothesis to *predict* the data

*Law of total probability*

\[ p(\text{Hypotheses} \mid \text{Data}) = \sum p(\text{Hypothesis} \mid \text{Data}) \]

The support for a *compound/composite* hypothesis must account for all the ways it could be true
Agenda

1 Motivating example: \( \bar{x} \pm \frac{\sigma}{\sqrt{N}} \)
   Confidence intervals vs. credible intervals
   A tale of two spaces

2 Probability theory as generalized logic

3 Probability theory for data analysis: Three theorems

4 Inference with parametric models
   Parameter Estimation
   Model Uncertainty

5 Multilevel models for measurement error

6 Bayesian computation
Supplement:

- Surveying cosmic populations: selection effects & measurement error
- Graphical/hierarchical/multilevel models and latent parameters

*See Angie Wolfgang’s lectures!

More information:

CAS 2014 Supplement Session
Overview of MLMs in astronomy: arXiv:1208.3036
In progress: GPU software (Szalai-Gindl, Budavari, Kelly, TL)
Agenda

1 Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$
   Confidence intervals vs. credible intervals
   A tale of two spaces

2 Probability theory as generalized logic

3 Probability theory for data analysis: Three theorems

4 Inference with parametric models
   Parameter Estimation
   Model Uncertainty

5 Multilevel models for measurement error

6 Bayesian computation
Notation focusing on computational tasks

\[ p(\theta|D, M) = \frac{p(\theta|M)p(D|\theta, M)}{p(D|M)} = \frac{\pi(\theta)L(\theta)}{Z} = \frac{q(\theta)}{Z} \]

- \( M \) = model specification
- \( D \) specifies observed data
- \( \theta \) = model parameters
- \( \pi(\theta) \) = prior pdf for \( \theta \)
- \( L(\theta) \) = likelihood for \( \theta \) (likelihood function)
- \( q(\theta) = \pi(\theta)L(\theta) = \) “quasiposterior”
- \( Z = p(D|M) = \) (marginal) likelihood for the model
Parameter space integrals

For model with \( m \) parameters, we need to evaluate integrals like:

\[
\int d^m \theta \ g(\theta) \pi(\theta) \mathcal{L}(\theta) = \int d^m \theta \ g(\theta) \ q(\theta)
\]

- \( g(\theta) = 1 \rightarrow Z = p(D|M) \) (norm. const., model likelihood)
- \( g(\theta) = \theta \rightarrow \) posterior mean for \( \theta \)
- \( g(\theta) = \text{‘box’} \rightarrow \) probability \( \theta \in \) credible region
- \( g(\theta) = 1, \) integrate over subspace \( \rightarrow \) marginal posterior
- \( g(\theta) = \delta[\psi - \psi(\theta)] \rightarrow \) propagate uncertainty to \( \psi(\theta) \)

Except for optimization, Bayesian computation amounts to computing the expectation of some function \( g(\theta) \) with respect to the posterior dist’n for \( \theta \)

Contrast with frequentist computation, which integrates over sample space, e.g., via Monte Carlo simulation of data
Bayesian Computation Menu

**Large sample size, $N$: Laplace approximation**
- Approximate posterior as multivariate normal $\rightarrow \det(\text{covar})$ factors
- Uses ingredients available in $\chi^2/\text{ML}$ fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

**Modest-dimensional models ($m \lesssim 10$ to 20)**
- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

**High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo**
- Posterior sampling — create RNG that samples posterior
  - Markov Chain Monte Carlo (MCMC) is the most general framework
- Sequential Monte Carlo (SMC)
- Approximate(ly) Bayesian computation (ABC)
- ...
The Laplace approximation (1-D)

**Motivation**

- Approximate integrand in neighborhood of the peak, \( \hat{\theta} \)
- Approximate the \( \log \) integrand, since we want PDFs to be non-negative: For \( e^{\Lambda(\theta)} = g(\theta)q(\theta) \), Taylor series to 2nd order gives

\[
\Lambda(\theta) \approx \Lambda(\hat{\theta}) + \Lambda'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\Lambda''(\hat{\theta})(\theta - \hat{\theta})^2
\]

vanishes

Leading order dependence on \( \theta \) is **Gaussian** with mean \( \hat{\theta} \) and 
\( \sigma^2 = -\frac{1}{\Lambda''(\hat{\theta})} \)

- In many settings asymptotics \( \rightarrow \) expect \( q(\theta) \) to be 
  \( \approx \) Gaussian so \( gq \approx \) Gaussian if \( g(\theta) \) varies slowly

Fits a Gaussian function to the peak of the integrand, and estimates the original integral using the fitted Gaussian
Examples—Gamma, Beta

\[ p(r|n, T) \]
\[ n = 8, \quad T = 1 \]
\[ p(\alpha|n, N) \]
\[ n = 8, \quad N = 12 \]

\[ (N - n) = 32 \quad N = 48 \]

Laplace approx’n
Laplace Approximations (multi-D)

Suppose posterior has a single dominant (interior) mode at $\hat{\theta}$. For large $N$,

$$
\pi(\theta)\mathcal{L}(\theta) \approx \pi(\hat{\theta})\mathcal{L}(\hat{\theta}) \exp \left[ -\frac{1}{2}(\theta - \hat{\theta})\hat{I}(\theta - \hat{\theta}) \right]
$$

where

$$
\hat{I} = -\left. \frac{\partial^2 \ln[\pi(\theta)\mathcal{L}(\theta)]}{\partial^2 \theta} \right|_{\hat{\theta}}
$$

$\hat{I}$ = Negative Hessian of $\ln[\pi(\theta)\mathcal{L}(\theta)]$

$\hat{I}$ = “Observed Fisher info. matrix” (for flat prior)

$\approx$ Inverse of covariance matrix

E.g., for 1-d Gaussian posterior, $\hat{I} = 1/\sigma^2_\theta$
Marginal likelihoods

$$\int d\theta \pi(\theta)\mathcal{L}(\theta) \approx \pi(\hat{\theta})\mathcal{L}(\hat{\theta}) (2\pi)^{m/2} |\hat{\mathcal{I}}|^{-1/2}$$

Marginal posterior densities

Profile likelihood $\mathcal{L}_p(\phi) \equiv \max_\eta \mathcal{L}(\phi, \eta) = \mathcal{L}(\phi, \hat{\eta}(\phi))$

$$\rightarrow p(\phi|D, M) \propto \pi(\phi, \hat{\eta}(\phi))\mathcal{L}_p(\phi) |\mathcal{I}_\eta(\phi)|^{-1/2}$$

with $\mathcal{I}_\eta(\phi) = \partial_\eta \partial_\eta \ln(\pi\mathcal{L})|\hat{\eta}$

Posterior expectations

$$\int d\theta f(\theta)\pi(\theta)\mathcal{L}(\theta) \propto f(\tilde{\theta})\pi(\tilde{\theta})\mathcal{L}(\tilde{\theta}) (2\pi)^{m/2} |\tilde{\mathcal{I}}|^{-1/2}$$

where $\tilde{\theta}$ maximizes $f\pi\mathcal{L}$

Features

Uses output of common algorithms for frequentist methods (optimization, Hessian*)

Uses ratios → approximation is often $O(1/N)$ or better

Includes volume factors that are missing from common frequentist methods (better inferences!)

Related to BIC—*but better!*

*Some optimizers provide approximate Hessians, e.g., Levenberg-Marquardt for modeling data with additive Gaussian noise. For more general cases, see Kass (1987) “Computing observed information by finite differences” (beware typos): central 2nd differencing + Richardson extrapolation.
**Drawbacks**

Posterior must be smooth and unimodal (or well-separated modes)

Mode must be away from boundaries (can be relaxed)

Result is parameterization-dependent—try to reparameterize to make things look as Gaussian as possible (e.g., $\theta \rightarrow \log \theta$ to straighten banana-shaped contours)

Asymptotic approximation with no simple diagnostics (like many frequentist methods)

Empirically, it often does not work well for $m \gtrsim 10$
Trapezoid and Simpson rules

Trapezoid rule
Piecewise-linear approximation

Simpson's rule
Piecewise-parabolic approximation

Trapezoid rule:
\[
\int f(x) \approx \Delta x \left[ \frac{1}{2} f(x_0) + f(x_1) + f(x_2) + \cdots + \frac{1}{2} f(x_n) \right]
\]

Simpson’s rule:
\[
\int f(x) \approx \frac{\Delta x}{3} \left[ f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + f(x_n) \right]
\]
Adaptive quadrature

1. Estimate integral over \([a, b]\)
2. Estimate error (e.g., using higher-\(n\) rule that reuses nodes)
3. If error is large, subdivide interval, and repeat in subintervals
4. When error criterion met, sum subinterval quadratures

scipy.integrate.quad() uses adaptive Clenshaw-Curtis or Fourier rules
Monomial Cubature Rules

Seek rules exact for multinomials \((\times \text{ weight})\) up to fixed monomial degree with desired lattice symmetry; e.g.:

\[ f(x, y, z) = \text{MVN}(x, y, z) \sum_{ijk} a_{ijk} x^i y^j z^k \quad \text{for } i + j + k \leq 7 \]

Number of points required grows much more slowly with \(m\) than for Cartesian rules (but still quickly)

A 7th order rule in 2-d
Adaptive Cubature

- Subregion adaptive cubature: Use a pair of monomial rules (for error estimation); recursively subdivide regions with large error (ADAPT, CUHRE, BAYESPACK, CUBA). Concentrates points where most of the probability lies.

- Adaptive grid adjustment: Naylor-Smith method
  Iteratively update abscissas and weights to make the (unimodal) posterior approach the weight function.

These provide diagnostics (error estimates or measures of reparameterization quality).

\[
\text{\# nodes used by ADAPT's 7th order rule} = 2^d + 2d^2 + 2d + 1
\]

<table>
<thead>
<tr>
<th>Dimen</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>17</td>
<td>33</td>
<td>57</td>
<td>93</td>
<td>149</td>
<td>241</td>
<td>401</td>
<td>693</td>
<td>1245</td>
</tr>
</tbody>
</table>
Analysis of Galaxy Polarizations

TJL, Flanagan, Wasserman (1997)
Modest-D: IID Monte Carlo Integration

\[ \int g \times p \] is just the expectation of \( g \); suggests approximating with a sample average based on IID draws from \( p \):

\[
\int d\theta \, g(\theta)p(\theta) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta)} g(\theta_i) + O(n^{-1/2})
\]

This is like a cubature rule, with equal weights and random nodes

Ignores smoothness \( \rightarrow \) poor performance in 1-D, 2-D vs. quadrature rules

Avoids curse of dimensionality: \( O(n^{-1/2}) \) regardless of dimension
Why/when it works

- Independent sampling & law of large numbers → asymptotic convergence in probability
- Error term is from CLT; requires finite variance

Practical problems

- $p(\theta)$ must be a density we can draw IID samples from—perhaps the prior or a simple posterior, but...
- $O(n^{-1/2})$ multiplier (std. dev’n of $g$) may be large

→ IID* Monte Carlo can be hard if dimension $\gtrsim 5–10$

*IID = independently, identically distributed
Posterior sampling

\[ \int d\theta \ g(\theta)p(\theta|D) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta|D)} g(\theta_i) + O(n^{-1/2}) \]

When \( p(\theta) \) is a posterior distribution, drawing samples from it is called posterior sampling:

- One set of samples can be used for many different calculations (so long as they don’t depend on low-probability events)
- This is the most promising and general approach for Bayesian computation in high dimensions—though with a twist (MCMC!)

**Challenge:** How to build a RNG that samples from a posterior?
Abandon independence → *Markov chain Monte Carlo (MCMC)*

*More in presentations by Murali Haran, Eric Ford!*

See [SCMA 5 Bayesian Computation tutorial notes](#), and notes from [CASSt 2014 Supplement Sessions](#), for more on computation!

See [online resource list](#) for an annotated list of Bayesian books and software