Bayesian inference

Eric Feigelson

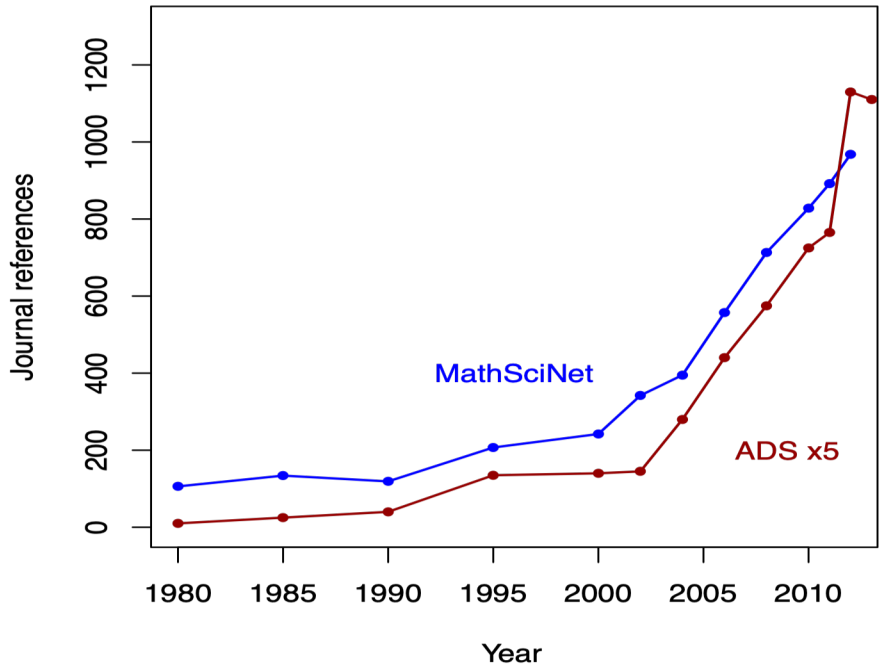
NARIT-EACOA Summer Workshop on Astrostatistics & Astroinformatics
August 2019

Bayesian inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference derives the posterior probability as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data.

-- Wikipedia 2019

Rapid rise of Bayesian analyses in mathematics and astronomy



Outline

- Derivation of Bayes' Theorem
- A simple astronomical example
- Priors
- Posteriors
- Parameter estimation & credible intervals
- Model selection
- Marginalization
- Hierarchical models
- Bayesian computation (stochastic processes, MCMC, SNIa cosmology)
- Philosophical considerations: frequentist or Bayesian?
- Advantages & disadvantages of Bayes
- Final advice

Axioms of probability

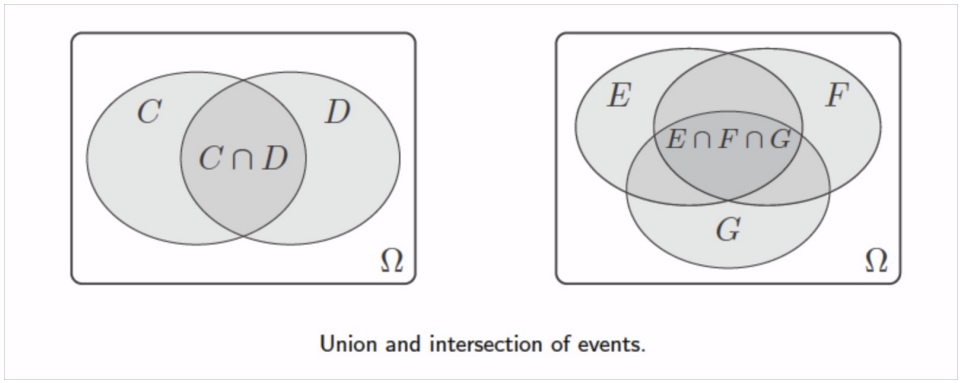
Axiom 1: For an event A, $0 \leq P(A) \leq 1$,

Axiom 2: For a sample space Ω, $P(\Omega) = 1$

Axiom 3: For mutually exclusive events

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

$\cup$ = "or", union $\cap$ = "and", intersection $|$ = "given", conditioned on



Union and intersection of events.

Conditional probability

The probability of event A given event B is equal to the intersection of A and B normalized by the probability of B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

$$P(A \cap B) = P(A \mid B)P(B)$$

Example: The conditional probability that a star has solar mass and a Jovian planet with ellipticity 0.5-0.6 is equal to the product of the probability that a star is a G star (say P~1%) and the probability that any star has a ε=0.5-0.6 Jovian planet (say P~20%). The conditional probability is thus P~0.2%.

,

The 'multiplication rule' easily extends to n events:

$$P(A_1 \cap A_2 \cap \ldots A_n) = P(A_1)\, P(A_2 \mid A_1) \ldots P(A_{n-1} \mid A_1, \ldots A_{n-2}) \times P(A_n \mid A_1, \ldots A_{n-1}).$$

Another astronomical example:

Except for the rare circumstance when an entirely new phenomenon is discovered, astronomers are measuring properties of celestial bodies or populations for which some distinctive properties are already available. Consider, for example, a subpopulation of galaxies found to exhibit Seyfert-like spectra in the optical band (property A) that have already been examined for nonthermal lobes in the radio band (property B). Then the conditional probability that a galaxy has a Seyfert nucleus given that it also has radio lobes is given by equation (2.11), and this probability can be estimated from careful study of galaxy samples. Similar inferences can be made with

Bayes' Theorem

Let $B_1, \ldots, B_k$ be a partition of the sample space Ω.

If A is any event in Ω, then to compute

$P(A)$, one can use probabilities of pieces of A on each of the sets B_i and add them together to obtain the Law of Total Probability

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k). \qquad (2.13)$$

For B_k possible outcomes, using the definition of conditional probability and the Law of Total Probability,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_k)P(B_k)}.$$

Bayes' Theorem is thus just a necessary result of probability theory, or logic, based on the three Axioms. In Bayesian inference the terms are given a specific meaning (MSMA, p.63-64)

$$P(M_i(\theta) \mid X) = \frac{P(X \mid M_i(\theta))P(M_i(\theta))}{P(X \mid M_1(\theta))P(M_1(\theta)) + \cdots + P(X \mid M_k(\theta))P(M_k(\theta))}.$$

Let X represent the data, and M represent the space of models or hypotheses that depend on parameters theta

$P(M_i|X)$ = conditional probability of M_i given X = posterior probability density
$P(X|M_i)$= conditional probability of X given M_i = likelihood function
$P(M_i)$ = prior or marginal probability of M_i = prior information
P(X) [denominator] = marginal probability of X = normalizing constant

Bayes' Theorem in English:
The posterior distribution of a chosen model given the data is equal to the normalized product of (the likelihood of the data for that model) and (the prior probability that the model is true without reference to the data)

Posterior distribution

Once the prior distribution and alternative hypotheses (range of θ parameters) are specified, and the data are obtained, the posterior distribution can be calculated. This distribution can be plotted in the *p*-space of its parameters. These plots give information on any non-Gaussianity and multimodality of the posterior. Typically, the scientist is interested in the `best' Bayesian estimator for the parameters; i.e. the maximum (mode) of the posterior. The *credible region* around this value is then estimated.

Prior distributions

This is often the controversial aspect of Bayesian inference, because subjective judgment or simplistic uninformative priors are often used. For uniform priors, maximizing the posterior often gives the same result as maximum likelihood estimation, although interpretation of results differ. Bayesian inference is most effective when the scientist ***wants*** to bias the likelihood based on the data using scientifically meaningful prior constraints on the the parameters.

An astronomical example of Bayesian inference

Is this new active galactic nucleus radio-loud?

Let X be a random variable taking two values: X=1 indicates *Yes* and X=0 is *No.*

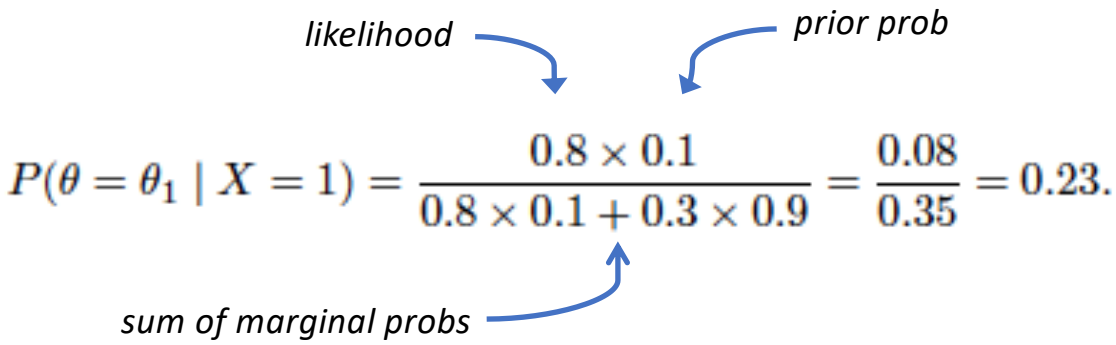
Let θ be a parameter denoting AGN radio-loudness: θ_1 indicates *Yes*, θ_2 is *No*

From previous AGN surveys in the visible & X-ray bands, the astronomer expects a probability of radio-loudness: $P(\theta = \theta_1) = 0.1$.

The new AGN under study was observed with a radio telescope sensitive enough to measure radio-loudness 80% of the time in radio-loud AGN:

$P(X=1|\theta_1) = 0.8$. However, 30% of the time the telescope detects irrelevant radio emission from star formation in the host galaxy: $P(X=1|\theta_2) = 0.3$.

Use Bayes' Theorem to calculate the chances than an AGN with detected radio emission is truly a radio-loud AGN:



Wow! Only 23% of true radio-loud AGN are clearly identified in this survey: 77% are either false negatives or false positives. Trying different assumptions shows that the result is moderately sensitive to the value of the prior (0.10) but is very sensitive to the false positive fraction (0.30). If this is reduced to 0.05 (e.g. through radio polarization & spectral study), then the discovery fraction of true radio-loud AGN rises to 95%. Bayesian calculations can help the astronomer evaluate how the science goals can be better achieved in a future experiment.

Prior distributions: Uninformative priors

Many Bayesian studies (in astronomy and elsewhere) do not have an empirical or subjective basis for specifying the distribution of a model parameter in advance of the experiment/observation at hand. In such cases, an uninformative prior is used to weight the likelihood in Bayes' Theorem.

These priors make few or no assumptions about the distribution of model parameters. Two common choices:

- The uniform distribution over the full space of possible values. This often reproduces results from maximum likelihood estimation.

- Jeffreys prior $\pi(\theta) = | \mathbf{I}(\theta) |^{1/2}$, I is the Fisher Information Matrix

However, use of uninformative priors is controversial and many statisticians do not support their use:

1. Many are **improper priors** that do not integrate to unity (often the integral is infinite). Thus they are not p.d.f.'s and should not be used.
2. The results depend on arbitrary choices. In an astrophysical model, is the prior of *X* or *log(X)* assumed to be uniform? For the normal model, is the prior of the variance or the standard deviation assumed to be uniform? A uniform s.d. allows Bayesian calculations to reproduce many classical results.

A statistician's worried viewpoint about uninformative priors

"Because the prior is inescapably part of the model in the Bayesian approach, marginal likelihoods, Bayes factors and posterior model probabilities are inescapably sensitive to the choice of prior. In consequence, it is only when those priors that differ between alternative models are really precise and meaningful representations of prior knowledge that we can justify using Bayes factors and posterior model probabilities for model selection. Even then the computation of the marginal likelihood is often difficult."

Simon Wood, *Core Statistics* (2015)

Proper use of priors

A reasonable alternative is to try different reasonable proper priors and, if the results are compatible, report them as scientifically reliable results.

When flat or uninformative priors are used together with estimation using the mode of the posterior (MAP or HPD best fit), then we recommend that the Bayesian approach be dropped and the Maximum Likelihood Estimation formulation be used instead.

When the prior can be reliably established from detailed scientific information available from earlier observations or from astrophysical theory, then we encourage use of these informative priors with a Bayesian approach. However, it is wise to examine the relative influence of different reasonable priors, together with the data, on the scientific result for the particular situation at hand.

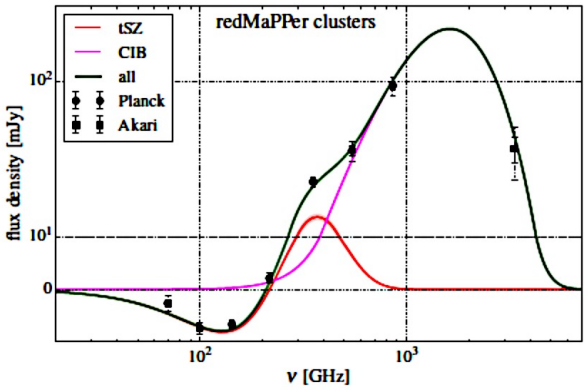
A published Bayesian analysis should communicate the statistical model (likelihood) and the prior distribution of all parameters in sufficient detail that the inferential calculation is reproducible by other scientists.

Bayesian posterior

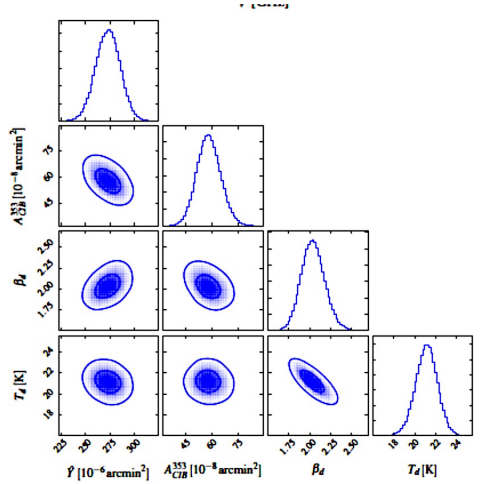
The result of a calculation of Bayes' Theorem for a dataset and a model space is the distribution of the posterior. Astronomers often plot univariate and bivariate projections of a multivariate posterior estimated by MCMC sampling.

Example: Model of Sunyaev-Zel'dovich distortion to the cosmic microwave background spectrum (taking the dust-induced cosmic infrared background (CIB) variations into account) applied to 26,111 galaxy clusters from the Sloan Digital Sky Survey (Soergel et al. 2017)

Best fit model

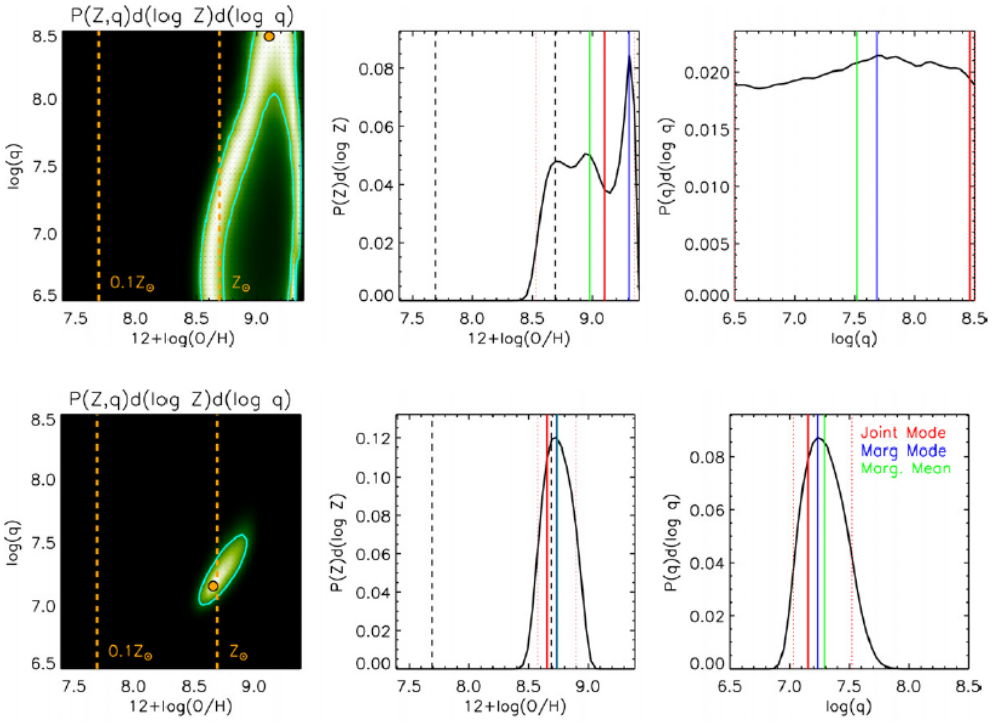


Marginal posterior distributions



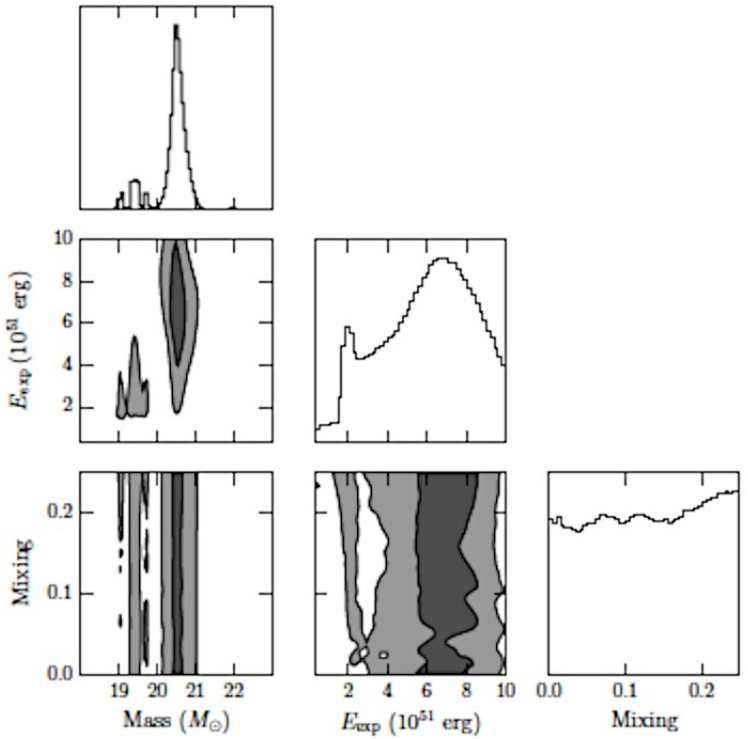
Some messier posteriors

Bayesian analysis of ionization and metallicity in HII regions. The top panels show joint and marginal posteriors of ionization and O/H abundance using 2 emission lines. The bottom panels show the posterior using 8 emission lines.



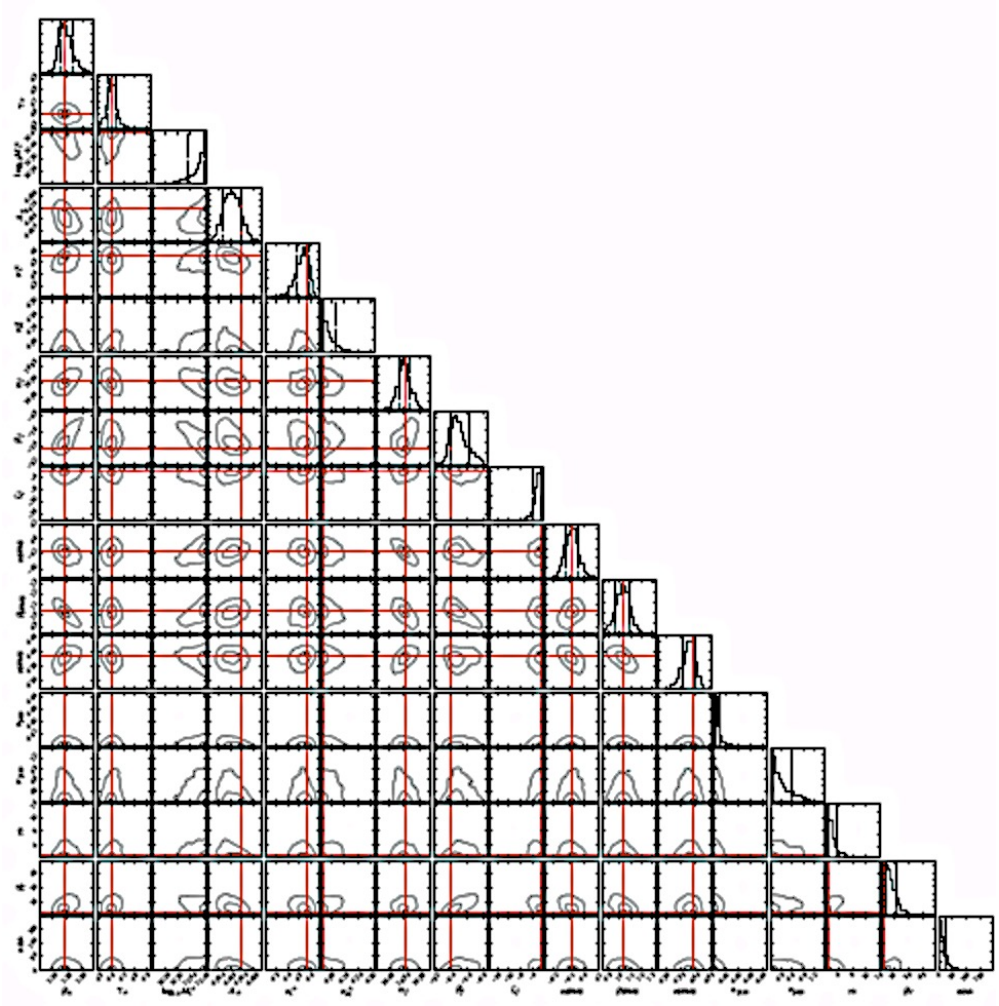
(Blanc et al. 2015)

Modeling a high-redshift metal-poor damped Lyα absorption (DLA) system



Cooke et al. 2017

A high-dimensional posterior



Bull 2017

Bayesian parameter estimation

(or Summarizing the posterior distribution)

Scientists often seek a single `best' model giving 'best-fit' parameters for the dataset and the model space, rather than a multivariate distribution of model probabilities.

Three approaches are commonly used in Bayesian inference. The choice should be based on the a previously specified 'loss function' (or 'risk function') that quantifies the scientific value of alternative models. The principles arise from Bayesian decision theory, a branch of information theory.

- The **mode** of the posterior distribution. This is sometimes called the **maximum a posteriori (MAP)** or the **highest posterior density (HPD)** estimate. For uniform priors or very large datasets, the posterior mode gives model parameter values approaching the classical maximum likelihood estimators. For an informative prior, the MAP solution is a weighted average of the MLE of the prior and the MLE of the data. In decision theory, the mode is preferred when the cost of a wrong answer is high (posterior loss is binary).

- The **median** of the posterior is preferred when the cost of a wrong answer is low (posterior loss scales as the linear 'distance' between models)

- The **mean** of the posterior distribution. This a weighted mean of the likelihood and the prior:

$$E(\theta|\mathbf{X}) = \frac{\int_\Omega \theta L_{\mathbf{X}}(\theta)\pi(\theta)d\theta}{\int_\Omega L_{\mathbf{X}}(\theta)\pi(\theta)d\theta}$$

This is simply the expectation of the posterior distribution. The mean is perferred for intermediate cost functions (posterior loss scales as the squared distance between models).

J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed 1985

C. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation,* 2nd ed 2007

Unfortunately, astronomical research does not usually have a clear loss function, so astronomers are subjectively choosing to report medians, means and modes.

Essentially, the research community must choose a consistent summary statistic, much as it chooses a consistent significance level (3-sigma) for reporting results of hypothesis tests.

Some experts have suggested the ***posterior median*** *for general use. Other experts suggest avoiding summarizing the posterior and use/discuss the entire distribution.*

Bayesian credible intervals

The Bayesian **credible interval** of parameter values (or **credible region** for *p*-dimensional models) around the MAP value can be estimated from the analytical or computed posterior distribution. This plays the role of the *confidence interval* in classical statistical inference. The credible interval can be found by solving for lower and upper functions such that

$$P(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X}) \mid \mathbf{X}) = 1-\alpha$$

where α=0.05, 0.01, etc. In realistic cases, it is computed numerically by computing values of the posterior distribution around the best fit value.

Bayesian model selection

An important class of hypothesis tests is **model selection**, the comparison of two alternative models for a given dataset. When applied to nested models, this problem is important for deciding how many parameters is needed to adequately fit the data in a parsimonious fashion.

Bayesian model selection is based on the ***Bayes factor,*** or ratio of posteriors, given by

$$B_{12} = \frac{P(\mathbf{X}|M_1, \pi)}{P(\mathbf{X}|M_2, \pi)}$$

The ratio of the probabilities of the two models, or **odds ratio,** is the product of the ratio of likelihoods and the Bayes factor:

$$O_{12} = \frac{P(M_1|\pi)}{P(M_2|\pi)} B_{12}$$

The Bayesian odds ratio is equal to the classical likelihood ratio test (LRT) when the priors for the two models are equal. This is often the case when the priors are uninformative.

History: The LRT was established by theorem by Neyman & Pearson (1933), and Wilks (1938) showed it asymptotically follows a chi-squared distribution.

Model selection is an example of Bayesian hypothesis testing and has a number of advantages over classical hypothesis testing:

- The Bayes factor automatically accounts for the number of parameters, favoring the more complicated model only if the ratio of the likelihoods is sufficiently high. In classical MLE, the penalty for model complexity is debated, and does not arise naturally from the mathematics.

- Both classical and Bayesian analysis often use the Bayesian Information Criterion (BIC), which is an approximation to the Bayes factor.

- Bayes factors allow comparison of nonnested models, and Bayesian model averaging can be used to account for model uncertainty.

To compute the Bayes factor, we need to calculate the *marginal likelihood* of each model for the available data

$$f(\mathbf{y}|M_1) = \int f(\mathbf{y}|\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1$$

Use of these marginal likelihoods in model selection accounts for differences in model complexity – models with larger 'volumes' of parameter space are not automatically favored, as they are when the LRT is used.

However, two difficulties arise. First, it may be hard to compute the marginal likelihoods for all parameters and (possibly) for a wide range of models. Second, the marginal likelihoods are sensitive to the prior and will change values for different uninformative priors unless the same improper priors are used in all models.

In practice, it is often easier to compute the approximate than the full odds ratio.

Bayesian marginalization

Many problems have variables or parameters of little scientific interest (e.g. detector background vs. astronomical signal). Bayesian formulations allow direct ***marginalization*** (integration) over nuisance variables.

Consider a case of a vector of *k* parameters where we are interested in the i^{th} parameter and not the others:

$$P(\theta_i) = \int d^{k-1}\theta \; P(\theta_1, \theta_2, ..., \theta_k | x_1, x_2, ..., x_n)$$

The distribution of the interesting parameter θ_i takes into account information about all of the other parameters. This `averages out' the influence of other parameters.

Hierarchical Bayes' modeling

When a Bayesian model is based on a prior distribution that itself has unknown parameters, the calculation must simultaneously solve for the model parameters and the prior **hyperparameters**. This is a type of **hierarchical Bayes' model**. The acquisition of additional data simultaneously updates the prior distribution and constrains the model parameters of interest. Examples:

- the prior is a mixture of two distributions with the mixing fraction serving as a hyperparameter
- the model parameters θ are generated from a process governed by a hyperparameter ψ. Then (ignoring normalizations)

$$P(\theta, \phi|\mathbf{X}) = P(\mathbf{X}|\theta)P(\theta, \phi)P(\phi)$$

Hierarchical Bayesian models are increasingly common in the astronomical literature.

Introduction to Bayesian computing

While the concepts of inverse probability and Bayesian inference were introduced Simon Pierre Laplace two centuries ago, prior to ~1990, Bayesian inferential applications were largely restricted to simple problems.

Bayesian estimation requires considerable more computation than least squares estimation (system of linear equations) or maximum likelihood estimation (optimization of a single likelihood function) because it often requires examination of, and sometimes integration over, the full space of possible models. Modern astrophysical models can have dozens (or more) parameters, requiring mapping of the prior-weighted likelihood function in high dimensions.

Markov chain Monte Carlo (MCMC) methods can, with varying degrees of efficiency, map the posterior draw sequential samples from the parameter space where the likelihood and prior are evaluated. For simpler problems, the Laplacian approximation can be much more efficient. Integrated Nested Laplacian Approximation (INLA) can be effective for many situations in astronomy (arxiv:1802.06280).

Strong background on Bayesian computation

Vignettes for R/CRAN package LaplacesDemon including

https://web.archive.org/web/20150531112558/http://www.bayesian-inference.com:80/mcmc

Interactive visualization of several MCMC algorithms

https://chi-feng.github.io/mcmc-demo/app.html

Stochastic processes I

Consider measurements of the Doppler motion of a star orbited by a companion star or exoplanet. A collection of Doppler velocities is a function of *t* is an example of a *stochastic process*: for each observed time *t, X(t)* is a random variable. t can represent any fixed variable: e.g. time, space (1D, 3D, …), space-time, or a lattice parameter space. Astronomers encounter them as functions of fixed time-like variables such as:

Brightness B(RA,Dec) defines an image

Brightness B(wavelength) defines a spectrum

Brightness B(time) defines a light curve

Spectral index or radial velocity (RA,Dec) defines other images

Density $\rho(x,y,z,v_x,v_y,v_z)$ defines a fluid flow

Stochastic processes II

Random variables can be functions of discrete or continuous time-like variables (e.g. pixelated images or lightcurve with accurate timestamps)

The r.v.'s themselves can assume discrete or continuous values (e.g. photon arrivals or real-valued brightness).

The observations can be sequences of i.i.d. r.v.'s, or they can exhibit dependencies. These dependencies can arise either from the instrument (e.g. point spread function in an image) or from the underlying physical process (e.g. timescale for brightness variations in an accretion disk).

A stochastic process is ***stationary*** if $(X(t_1), X(t_2), \ldots X(t_k))$ and $(X(t_1+\delta), X(t_2+\delta), \ldots X(t_k+\delta))$ have the same joint distribution for all δ, $t_1, \ldots, t_k$ and $k \geq 1$.

Note that trends in the mean cause nonstationarity. The spatial structure of an image with stars and galaxies is nonstationary. The brightness variations of a variable star may or may not be stationary.

Markov chains

A stochastic process $\{X_n\}$ is called a *discrete time Markov chain* if the distribution of X_{n+1} given the past $X_n, \ldots, X_0$ depends only on the immediate past. Further suppose that the probability for transitions from states *i* to *j* are fixed, P_{ij}. Formally, this process is written:

$$P\{X_{n+1} = j \mid X_n=i, X_{n-1}=i_{n-1}, \ldots X_0=i_0\} = P(X_{n+1}=j|X_n=i) = P_{ij}$$

For all states $i=i_0$ to i_{n-1} and all times $n \geq 0$

This process is called a *Markov chain.* The values of X_n are called *states*. They need not be integers.

A simple Markov chain is the *random walk*, $P_{i,i+1} = p = 1-P_{i,i-1}$.

Markov Chains II

$\mathbf{P}_{ij}$ can be written as a matrix of one-step transition conditional probabilities from *i* to *j*. An initial state at time 0 needs to be specified to give unconditional probability distributions at time *n*. Note that state *i* may communicate with some states *j* but not with some other states *k*. States that communicate are in the same *class*

Constructing Markov Chains with random numbers to generate i.i.d. sequences of values with a specified p.d.f. provide a suite of algorithms for simulating complicated probability distributions. These are called ***Markov Chain Monte Carlo*** techniques.

Sometimes a Markov process is not directly visible, but some outcome from the chain (e.g. a signal when it visits state *i*) is available. These ***hidden Markov models*** are valuable for a variety of inference problems.

Markov chain Monte Carlo techniques

- **Gibbs sampler** Here the multivariate problem is simplified to a sequence of univariate function evaluations. Consider a 3-dimensional parameter space (θ_1, θ_2,θ_3). Starting at an initial location θ^0, make a random step, simulate $\theta^1_1 \mid (\theta^0_2, \theta^0_3,$**X**), $\theta^1_2 \mid (\theta^1_1, \theta^0_3,$**X)** and $\theta^1_3 \mid (\theta^1_1, \theta^1_2,$**X**), and calculate the posterior (prior-weighted likelihood) at the new location. Continue similar iterations to form a chain, and create multiple chains with different starting points and random steps. For high-dimensional problems, the sequence of variable updates can be varied, and the space can be blocked into subspaces that are updated sequentially.

- **Metropolis-Hastings algorithm** This procedure increases the efficiency of the chain's mapping of the posterior distribution by accepting the next step forward if it increases the posterior or satisfies some probability rule. Strategies for jumping around the parameter space avoid being trapped in small regions of the distribution. An early and common procedure is to combine the Gibbs and Metropolis strategies

Metropolis, Rosenbluth, Teller 1953, "Equation of State Calculations by Fast Computing Machines." *J. Chem. Phys.* MANIAC I computer

Convergence measures and stopping rules for MCMC simulations are very important. Unlike the EM Algorithm for MLE, *there are no theorems guaranteeing convergence* on a maximum in the posterior distribution. Millions of iterations may be needed for a single MCMC chain, and many chains may to needed to obtain reliable results.

Common stopping criteria include:

- chain standard deviation becomes small
- autocorrelation within the chains becomes small
- within-chain and between-chain variances approach equality (Gelman-Rubin diagnostic)

Dozens of MCMC-type methods have been developed in the past ~20 years, and many are implemented in ~100 R/CRAN packages.

Algorithms

MCMC algorithms in the LaplacesDemon CRAN package:

- Adaptive Directional Metropolis-within-Gibbs (ADMG)
- Adaptive Griddy-Gibbs (AGG)
- Adaptive Hamiltonian Monte Carlo (AHMC)
- Adaptive Metropolis (AM)
- Adaptive Metropolis-within-Gibbs (AMWG)
- Adaptive-Mixture Metropolis (AMM)
- Affine-Invariant Ensemble Sampler (AIES)
- Automated Factor Slice Sampler (AFSS)
- Componentwise Hit-And-Run Metropolis (CHARM)
- Delayed Rejection Adaptive Metropolis (DRAM)
- Delayed Rejection Metropolis (DRM)
- Differential Evolution Markov Chain (DEMC)
- Elliptical Slice Sampler (ESS)
- Gibbs Sampler (Gibbs)
- Griddy-Gibbs (GG)
- Hamiltonian Monte Carlo (HMC)
- Hamiltonian Monte Carlo with Dual-Averaging (HMCDA)
- Hit-And-Run Metropolis (HARM)
- Independence Metropolis (IM)
- Interchain Adaptation (INCA)
- … … …

Two variants with code developed by astrostatisticians have acquired sudden popularity in astronomy:

- **MultiNest** This method was designed for complex posterior distributions with many modes (peaks) or degeneracies in high dimensions by Feroz & Hobson (Mon Not R Astro Soc 2008-09). A clustered nested sampling procedure reduces the computations for calculating the Bayesian evidence and posteriors. Written in Fortran 90, it has wrappers for C, C++, R, Python and Matlab.

- **Affine-Invariant Ensemble Sampler** This method was designed for badly scaled parameter spaces and skewed posterior distributions by Goodman & Weare (Comm Appl Math Comp Sci 2010). Here $>2K$ chains are simultaneous run, each with k starting points. For each iteration, the walkers are assigned new positions based on a scaled distance to another randomly-selected chain. Foreman-Mackey, Hogg, Lang, Goodman (Pub Astro Soc Pacific 2013) introduced a public-domain parallelized Python implementation called *emcee* with enthusiastic usage by astronomers.

Characteristics of MCMC algorithms

Chain properties: Non-Markovian (e.g. adaptive, new values not just based on last value), recurrent (retuens to a chosen state), periodic (cyclical);y recurrent, irreducible (all states accessible). A Markov chain with a stationary, aperiodic, irreducible distribution is called *ergodic* with advantageous properties (central limit theorem, convergence).

Proposal generation: multivariate proposal with all parameters, or proposal for individual parameters (slower)

Acceptance rate: ratio of accepted proposals to total iterations

Blockwise sampling: Model parameters are divided into groups of correlated variables that are sampled separately. Allows higher acceptance rates, and tuning algorithms for different blocks. However convergence may be inhibited by inter-block correlation.

Highest posterior density intervals

Metropolis-Hastings algorithm

Consider a function *y* of a time-like series $x^{(t)}$. We want to construct a sequence of y values that sample a target distribution. We start with a 'proposal' distribution *q* that is simpler, and wider, than the (often unknown) target distribution. At each iteration *t* of the chain, perform two operations:

Sample $y \sim q\left(y|x^{(t)}\right)$ with probability $\alpha(x^{(t)}, y) = \min\left\{1, \frac{\pi(y)q\left(x^{(t)}|y\right)}{\pi\left(x^{(t)}\right)q\left(y|x^{(t)}\right)}\right\}$

If accepted, assign *y* to be $x^{(t+1)}$. If reject, do nothing and try again.

If q(y|x)=π(y), then the samples are independent
If q(y|x_ = q(y), we have the independence sampler
If q(y|x) = q(|y-x|), the we have a Metropolis random-walk

Convergence measures and stopping rules for MCMC simulations are very important. Unlike the EM Algorithm, there are no theorems guaranteeing convergence on a maximum in the posterior distribution. Millions of iterations may be needed for a single MCMC chain, and many chains may to needed to obtain reliable results.

Common stopping criteria include:

- chain standard deviation becomes small
- autocorrelation within the chains becomes small
- within-chain and between-chain variances approach equality (Gelman-Rubin diagnostic)

Dozens of MCMC-type methods have been developed in the past ~20 years, and many are implemented in ~100 R/CRAN packages. Two variants developed by astrostatisticians have acquired sudden popularity in astronomy:

Stopping rules and convergence diagnostics

There is no theorem to establish

when a Markov chain has converged.

All convergence diagnostics and stopping rules

are suggestive only.

A simple measure of convergence is when the MCMC reaches a user-specified scatter level. However, due to the autocorrelation, the number of iterations N overestimates the effective sample size. There are various suggested corrections to the standard deviation involving the correlation coefficient or ACF:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \mathrm{E}\left[s^2\right] = \sigma^2\left[1 - \frac{2}{n-1}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right] \quad f = \sqrt{\frac{1+\rho}{1-\rho}},$$

$$\mathrm{ESS} = \frac{n}{1 + 2\sum_{k=1}^{\infty}\rho(k)} \qquad \mathrm{nse}\left(\bar{h}_N\right) \approx \sqrt{\frac{\sigma_h^2}{N}\left\{1 + 2\sum_{l=1}^{N-1}\rho_l(h)\right\}}$$

Geyer 1992

Gelman-Rubin diagnostic

Names: Potential scale reduction factor, G-R diagnostic, G-R shrink factor, R-hat

Concept: Convergence is reasonably achieved when chains have 'forgotten' starting values and recent outputs of different chains are indistinguishable. The variance of the chain ensemble is the sum of within-chain and between-chain variances for n iterations/chain. If the chains have not converged, **W** (mean of the variances within each chain) will be too small and **B** too large. The initial values for the chains must be overdispersed compared to the final posterior distribution (including possible multiple modes).

R diagnostic: Convergence $\hat{\sigma}^2 = \frac{(n-1)W}{n} + \frac{B}{n}$ when $1.0 < R \leq 1.1$ where chain ensemble variance:

$$R = \sqrt{\frac{(d+3)\widehat{V}}{(d+1)W}} \qquad d = \frac{2 * \widehat{V}^2}{\mathbf{Var}(\widehat{V})} \qquad \widehat{V} = \hat{\sigma}^2 + \frac{B}{mn}$$

Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, p. 457–511

Brooks, S.P. & Gelman, A. (1998) General methods for monitoring convergence of interactive simulations*, J. Computational & Graphical Statistics*, 7 434-455

Convergence diagnostics graphics

PDF estimate plot: Univariate or bivariate kernel density estimator plots of posterior from MCMC chains. Note assumptions underlying bandwidth selection. Histograms for discrete valued posteriors.

Trace plots: Time series-like plot of values for each variable in a chain

Autocorrelation function plot: Plot of ACF for each variable in each chain; high autocorrelation shows slow mixing and slow convergence.

Cross-correlatkon plot: Tile plot of correlations between parameters

ROC curve and separation plot: For problems with binary response variable

Caterpillar plot: For high-dimensional problems, stacked boxplot showing HPD & quantiles for each variable

Cumulative quantile plot: Shows evolution of 50%, 99%, … quantiles for n iterations

Gelman-Rubin-Brooks plot: G-R diagnostic vs. n iterations. Important to see recent fluctuations, rather than just test for R~1.0.

Geweke-Brooks plot: Shows Z-score (measuring similarity of beginning & end of a Markov chain) as increasing fractions of the early chain are omitted.

MCMC convergence in R

Diagnostics:

- Gelman and Rubin
- Geweke
- Heidelberger and Welch
- Raftery and Lewis
- Brooks and Gelman multivariate shrink factors

CRAN packages:

- coda
- LaplacesDemon
- ggmcmc
- boa

Many astronomers are not conducting sufficient tests to insure MCMC convergence

Astronomical example of Bayesian computation: Supernova Type Ia cosmology

See R script running Stan code with Hamiltonian MCMC in files:

Hilbe_SNIa.pdf

Hilbe_SNIa.R

excerpted from

Bayesian Models for Astrophysical Data using R, Python, JAGS and Stan

by Joseph Hilbe, Rafael de Souza & Emille Ishida (2017)

Philosophical considerations for model fitting: Comparing classical and Bayesian approaches

A Bayesian views probability as the plausibility of a situation or interpretation based on a combination of current and past information.

A frequentist views probability as the chance of a situation or interpretation assuming many hypothetical experiments were made, without consideration of past information.

Bayesian calculations average over model space, while frequentist calculations averages over sample space:

Bayesian: Data are fixed, hypotheses vary
Frequentist: Hypothesis is fixed, data vary

`Why isn't every physicist a Bayesian?' (particle physicist R. Cousins)

- For many simple situations, frequentist and Bayesian solutions are (nearly) identical

- Frequentist estimation is typically simpler mathematically and computationally. Except for trivial problems, Bayesian estimation often require arduous computation for the calculation of posteriors in the full space of possible parameter values

- Bayesian estimation will be biased if the prior distributions are misspecified. If priors are not known, MLEs may be preferred. Informative Bayesian priors can arise from astrophysical theory and/or previous empirical study.

- Bayesian model selection has advantages over frequentist model selection. The Bayes Factor and BIC can evaluate evidence in *favor of* (not just against) a model; be applied to non-nested model alternatives; incorporates external (prior) information; and has a natural compensation for model compexity (Occam's Razor). However, Bayesian model selection does not give formal probabilities.

- Markov chain Monte Carlo (MCMC) calculations is not intrinsically related to Bayesian inference … they are just convenient numerical tools for some applications. Other, simpler calculations (such as the Laplace approximation, INLA) or more complex calculations (such as Approximate Bayesian Computation) may be appropriate for a given problem.

Some disadvantages of Bayesian inference

Bayesian inference depends on the specification of a large, and often ill-defined, space of possible models.

For each possible model, Bayesian inference requires quantitative statement of the distribution of each models of interest prior to the acquisition of data. This often gives a subjective element to the procedure.

Bayesian model fitting requires specification of, and integration over, a universe of alternative theories. This is often both conceptually and computationally difficult. Simulations may take millions of iterations and may not converge. MLE is much less computationally demanding.

Some advantages of Bayesian inference

When the scientist indeed have prior knowledge (from previous observations or from astrophysical theory) of the parameter distributions, this can readily be incorporated into the Bayesian prior. Bayesian inference takes full account of this information. Such ancillary information and can only be included into frequentist calculations with difficulty.

Bayesian `marginalization' can treat the effects of nuisance variables (e.g systematic error, unobservable or uninteresting variables) with greater transparency than is often achieved with frequentist calculations.

Bayesian hypothesis tests treat hypotheses symmetrically, and Bayesian model selection can give probabilities that a model is correct. Classical hypothesis tests only give probabilities that a model is incorrect.

An astronomer's viewpoint about Bayesian applications (Feigelson)

Bayesian inference is often used in decision theory where a decision *must* be made, even a decision to do nothing. In astronomical research, a conclusion does not have to be reported unless new scientific insights emerge.

- If there is no prior evidence, then use maximum likelihood estimation, and report results only if the parameter estimates are scientifically interesting.
- If there is prior evidence, then use Bayesian inference, and report results only if the parameter estimates are improved by the new evidence to a scientifically interesting degree.

Broad advice on choosing inferential approach

When little is known about a scientific problem and the questions addressed are straightforward, then nonparametric statistics and hypothesis tests may be most appropriate. (MSMA Chpts 3.5 & 5)

When a parametric model, either heuristic or astrophysical, can be reasonably applied, and the experimental situation is relatively simple, then frequentist point estimation may be valuable (least squares & MLE). (MSMA Chpts 3.4 & 4)

Bayesian inference is best pursued when the situation is known to have external constraints (informative priors based on real knowledge from previous astronomical observations or from astrophysical theory), nuisance variables are present, and/or hierarchical relationships exist between variables. (MSMA Chpt 3.8)

Don't hesitate to pursue multiple avenues of analysis

1. Nonparametrics – 'Let the data speak for themselves' (Fisher?)

1. Maximum likelihood estimation – Can the data, considered in isolation, be well-fit by a chosen mathematical model? What are the best-fit parameter values? Is the best fit a good fit?

1. Bayesian with simple priors – What influence does prior knowledge about the parameters have on the best-fit solution?

1. Hierarchical Bayes – What can we learn about more complicated models with latent variables, prior hyperparameters, several modeling stages, etc.