Frequentist and Bayesian parametric modeling

Eric Feigelson

3rd INPE Advanced School in Astrophysics:
Astrostatistics 2009

Outline

Parametric modeling and astronomy

- Simple characterizations of a dataset: mean, standard deviation, correlation
- Heuristic relationships: linear (= powerlaw or Pareto law for logged variables) associations between variables
- Astrophysical models: spectra (continua and lines); stellar evolution; cloud turbulence; gravitational N-body; ΛCDM cosmology; ...; ...

Random experiments

A *random experiment* is one what can be repeated indefinitely and whose individual outcomes cannot be predicted with certainty (e.g. coin tosses, radioactive decay). This contrasts with a *deterministic experiment* where the outcome is determined by mathematical laws and initial conditions (e.g. Keplerian orbits). Many astronomical studies are effectively random experiments as the examined objects (stars, galaxies, quasars, Kuiper Belt objects) are a tiny fraction of a huge population, and the physical processes are so complex that the behaviors are not exactly predictable.

Probability relating data to models

The scientific process of evaluating models of measured phenomena can be treated in this framework of probability theory. Here we assign S to be the space of possible explanations, A is the observed dataset D_i for $i = 1, \ldots, n$, B_j are scientific models M_j. Sometimes there are just two models or 'hypotheses': H_0 and $H_1 \neq H_0$.

Often the models are not separately enumerated $j = 1, \ldots, k$, but rather are continuous functions of a vector of parameters θ. In this case, we are engaged in $parametric\ modeling$. We seek the values of θ that are most compatible with the data by evaluating the probabilities $P(M(\theta)|D_i)$. Note that we must assume, without any verification, that the data in fact follow chosen parametric functional family.

Point estimators I

Statistical inference is a quantitative form of induction, where a specific dataset is used to infer the properties of the underlying population.

Point estimation is a major class of inference problems where one seeks a 'best' estimate of a quantity relating to the population; e.g., its probability density function, mean value, dependency on some other property. A point estimate is designated with a $\hat{}$ symbol: $\hat{\theta} = f(x_1, \ldots, x_n)$.

Point estimators II

We look for estimates with favorable properties:

Unbiasedness The estimator is unbiased if its expectation based on the data E is equal to the true value for the population: $E(\hat{\theta}) = \theta$.

Consistency The estimator is consistent if it approaches the true value as n increases: $\hat{\theta} \to \theta$ as $N \to \infty$.

Sufficiency The estimator is sufficient when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter.

Ancillarity The estimator exhibits ancillarity if its distribution does not depend on the true value θ.

Efficiency The estimator is optimally efficient if it has the smallest variance compared to other estimators.

Asymptotic normality The estimator is asymptotically normal if its

Example #1: Powerlaw slope

Many plausible estimators do not have these properties. Consider, for example, the common astronomical problem of seeking the powerlaw slope from a dataset x_i (e.g. stellar initial mass function, spectral index of a synchrotron source, stellar mass-luminosity relationship). We typically take logarithm of the variable(s), bin them into a histogram, and perform a weighted least-squares linear regression to obtain slope $\hat{\beta}_{LS}$, $logN = \hat{\alpha}_{LS} + \hat{\beta}_{LS} logx$.

Due to the very asymmetrical shape of the powerlaw (= Pareto) distribution, the nonlinearity of the logarithm transformation, and loss of information during binning, this common estimator is biased, inconsistent and inefficient. The recommended statistic is the minimum-variance unbiased estimator

$$\hat{\beta}_{MVUE} = \frac{n-2}{\sum_{i=1}^{n} ln(x_i/x_0)} \quad \hat{\sigma}_{\beta_{MVUE}} = \frac{(1-2/n)n\ \beta_{MVUE}}{(n-2)\sqrt{n-3}}.$$

This is close to, but not exactly, the MLE.

How estimators are constructed

Judicious guessing Curve fitting, prior knowledge (e.g. physics)

Method of moments $\hat{\theta}_k = (1/n) \sum_{i=1}^{n} x_i^k$ (mean, variance, kurtosis, ...). Easy to compute but often not optimal.

Least-squares estimators Satisfactory under some conditions, easily computed.

Unbiased, minimum-variance estimators Often excellent, but only available in simple situations.

Maximum-likelihood estimators Often excellent. EM Algorithm used for computation with guaranteed convergence.

Bayesian estimators Often excellent. MCMC simulations used for computation, but convergence guaranteed.

Decision- theoretic and information-theoretic estimators Often excellent, but rare (e.g. Kullback-Liebler distance).

Maximum likelihood estimators

In 1912 as a student, R.A. Fisher proposed the likelihood function:

$$L(\theta|x_i) = L(\theta) = \Pi_{i=1}^{n} P(x_i|\theta)$$

where $P(x)$ is the model probability distribution function of the population with parameters θ. Note the reversal of roles of the data and parameters: we estimate parameters from the joint probability distribution of the sample.

The MLE of θ is $\hat{\theta}_{MLE}$ where L is maximized.

General MLE properties

- The MLE is usually consistent and unbiased, but not always.
- For many nice functions g, if $\hat{\theta}$ is the MLE of ? then $g(\hat{\theta})$ is the MLE of $g(\theta)$. This property is called Invariance.
- For large n, $\hat{\theta}$ is asymptotically normal. This permits construction of confidence intervals for θ using the Fisher Information Matrix, $I(\theta) = -[E(d^2 \ln L(x_i|\theta)/d^2\theta)]$. Bootstrap resampling is often better.
- When the likelihood cannot be maximized analytically, the iterative *EM Algorithm* (EM = Expectation-Maximization) is proven to maximize the likelihood. Example: Lucy-Richardson algorithm used for image deconvolution.
- MLEs can be applied to many problems where a model is specified: regressions, multivariate classification, time series modeling, hypothesis testing, etc.

A simple MLE: Mean of a normal I

We seek to estimate the parameters $\hat{\theta}$ which maximize the likelihood that the data are drawn from the model: $L(\theta) = P(x|M(\theta))$ where $|$ means 'given'. Let us consider a simple case where we seek the mean and dispersion of n measurements assuming Gaussian statistics and using calculus:

$$x_i = \mu + \epsilon_i \text{ where } \epsilon \sim N(0, \sigma^2).$$

$$L(\mu, \sigma) = \Pi_{i=1}^{n} P(x_i|\mu, \sigma) = \Pi_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x_i - \mu)^2}{2\sigma^2}]$$

$$-2 \ln L(\mu, \sigma) = n \ln(\sqrt{2}\pi\sigma^2) + \frac{n\hat{\sigma}^2}{\sigma^2} + n\frac{(\mu - \bar{x})^2}{2\sigma^2}$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

A simple MLE: Mean of a normal II

$$d(\ln L)/d\mu = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE}) = 0$$

$$d(\ln L)/d(\sigma^2) = -\frac{n}{2\sigma^2_{MLE}} + \frac{1}{2\sigma^4_{MLE}}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2 = 0$$

Result:

$$\hat{\mu}_{MLE} = \bar{x}, \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We find that $E(\hat{\mu}_{MLE}) = \mu$ is unbiased, but $E(\hat{\sigma}^2_{MLE}) = \frac{n-1}{n}\sigma^2$ is biased. Instead of the MLE, we thus prefer the unbiased variance estimator

$$\hat{\sigma}^2_{MVUE} = \frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

A Bayesian calculation: Mean value

If σ is known and we choose an uninformative prior for the mean, $P(\mu) = 1$, then the posterior probability for μ is $N(\bar{D}, Var)$ and the 68% credible region is $\bar{D} \pm \sigma/\sqrt{n}$. This is a common situation: the Bayesian best-fit parameters under some simple prior are often the same as the MLE parameters.

If σ is known and we choose a normal prior $\mu \sim N(\mu_p, Var_p)$, then the posterior for μ is $\tilde{\mu} = (1 - \frac{Var}{(Var+Var_p)})\bar{D} + \frac{Var}{(Var+Var_p)}\mu_p$. When the width of the prior is narrow, the estimator reflects the prior's mean rather than the data's mean.

If σ is unknown and we choose uninformative priors, the marginal posterior for the mean is a Student's t distribution (normal with powerlaw tails) that is asymptotically normal for large-n.

Poisson proportions I: The MLE

Situation: We have an astronomical population divided into two values (Type I $vs.$ Type II). Given observations n_I and n_{II}, we seek the proportion p (and its confidence interval) of the population with each value. In statistics, this may be called quantile test, binomial proportion or Poisson proportion problem.

If $N = n_I + n_{II}$ is sufficiently large for the central limit theorem to apply, p and its $(1 - alpha)$ quantile confidence intervals:

$$\tilde{p} = \frac{n_I}{N} \pm z_{1-\alpha/2)} \sqrt{n_I(N - n_I)/N^3}, \tag{1}$$

where $z_{1-\alpha/2} = \Phi^{-1}(\alpha/2)$ is the quantile of the normal distribution. For small N, the binomial distribution is used. This solution is the maximum-likelihood (or Wald) estimator.

Poisson proportions II: MLE problems

Example 1: If 10 of 100 quasars quasars are radio-loud and we seek the 95% confidence interval,
$\tilde{p} = 0.100 \pm (1.960)\sqrt{10(90)/100^3} = 0.100 \pm 0.059$.

Example 2: In a sparse X-ray or gamma-ray image, a source has $N = 12$ counts of which $n_I = 7$ occur in the hard band. The 'hardness ratio' n_I/N ratio is then 0.58 ± 0.28 (95% confidence). The HR is analogous to the color index (e.g., $B - V$).

However, this MLE is both biased and gives incorrect confidence intervals where p is near the limits of 0 or 1, and when N as high as 100 even for p away from these extreme values. Due to the discreteness and skewness of the binomial & Poisson distributions, the statistic performance is "persistently chaotic and unacceptably poor", oscillating discontinuously as p is smoothly varied. The problem is elucidated by leading statisticians in Brown et al. (*Statistical Science* 16, 101, 2001 and discussants).

Poisson proportions III: Several alternatives

$$\mathbf{E_{AC}}[p] = \tilde{p} \pm z_{0.025}\sqrt{n_I(\tilde{N} - n_I)/\tilde{N}^3} \tag{2}$$

where $\tilde{N} = N + 4$, $\tilde{p} = (n_I + 2)/(N + 4)$ for the 95% confidence interval. Derived by Agresti-Coull (*Amer Stat* 52, 119, 1998) based on Wilson (*JASA* 22, 209,1927). For the hardness ratio with $n_H = 7$ and $N = 12$, this gives 0.56 ± 0.24.

Bayesian credible interval using the beta distribution as the standard conjugate prior for binomial distributions. If $n_I \sim Bin(N, p)$ and p has prior distribution $Beta(a_1, a_2)$, then the posterior distribution of p is $Beta(n_I + a_1, N - n_I + a_2)$. The noninformative Jeffreys prior is $Beta(1/2, 1/2)$ and $\hat{p} = (n_I + 1/2)/(N + 1)$. For the hardness ratio, this 0.57 ± 0.26.

The inversion of the likelihood ratio test also has nice properties.

Poisson proportions IV: Astronomical discussions

Debate over the best representation: (a) $0 < n_I/n_{II} < \infty$; (b) $-\infty < log(n_I/n_{II}) < \infty$; (c) $0 < n_I/N < 1$; and (d) $-1 < (n_I - n_{II})/N < 1$. (a) is often avoided due to its extreme range and (d) is often chosen for its symmetry around zero. Note that the scientific notion of 'hardness' is too vague to define a specific quantity.

The CHARC astrostatistics collaboration discuss these choices and evaluate several statistical estimators, with the complication that a background rate is subtracted from all counts (Park et al. ApJ 652, 610, 2006). They study Bayesian estimators with gamma priors calculated using a Gibbs sampler, and show they greatly outperform the asymptotically-normal MLE for small-N. They discuss the roles of priors and mean $vs.$ mode of the posterior. C code (http://hea-www.harvard.edu/AstroStat/BEHR/).

Frequentist *vs.* Bayesian approaches I

1. Both approaches seek to quantify uncertainty in inductive inference from data using principles of probability.
2. ML inference considers models to be fixes and calculates probabilities of populations of hypothetical datasets. Bayesian inference considers a dataset to be fixed and calculates probabilities of populations of hypothetical models.
3. MLEs require and permit fewer explicit assumptions than Bayesian estimators. Bayesian approaches are better when prior knowledge strong affects the statistical results.
4. MLEs finds the single most-likely fit but, if the likelihood surface is complex (e.g., not unimodal) the solution may not be meaningful and the confidence intervals inaccurate. Bayesian estimators averages over likelihoods and accounts for the expanse of parameter space, and posterior distributions can reveal complex behaviors.

Frequentist $vs.$ Bayesian approaches II

5. Both methods require writing likelihoods, which may not be straightforward in real, complex situations. Calculations based on likelihoods are relatively easy for MLEs (sometimes analytical, otherwise EM Algorithm and bootstrap), but are often very hard (rarely analytical, sometimes MCMC) for Bayesian inference. Many Bayesian calculations can not (yet) be formulated or solved.

6. When samples are small or signals are weak, both approaches suffer. Assumptions of asymptotic normality fail in frequentist methods, and Bayesian calculations become hyper-sensitive to prior assumptions.

7. Classical least-squares methods (e.g. parametric fitting using χ^2) do not appear in either method, except as approximations in some simple situations.

Frequentist *vs.* Bayesian approaches III

8. Bayesian methods have lots of criticisms: there are many incompatible variants of 'Bayesian statistics' (some like 'empirical Bayes' mixed with frequentist methods); the priors are 'subjective' or difficult to choose (what does 'uninformative' mean?); integrations over parameter spaces can be extremely difficult, coverage and convergence is unsure (compared to ease and guarantee of the EM Algorithm for MLE); Bayesian procedures miss the value of nonparametric estimators/procedures like the e.d.f. and the bootstrap; traditional measures of statistical procedures are lost (e.g., unbiasedness, efficiency, maximum likelihood, robustness); procedures require too much thought and expertise for average scientists faced with real problems; Bayesian decision theory is confused with Bayesian estimation theory; Bayesian model selection is not very settled; Bayesian solutions are simply not available for the wealth of situations treated daily with traditional methods.

Frequentist *vs.* Bayesian approaches IV

9. Frequentist methods have lots of criticisms: the philosophy of infinite datasets rather infinite hypotheses violates common sense; there are no 'objective' procedures and choice of prior is implicit within traditional methods (whereas they are explicit in Bayesian approaches); standard methods have difficulty incorporating even simple prior constraints (e.g. zero avoidance) and cannot at all deal with complex situations (e.g. hierarchical structures, meta-analysis); standard p-values are wrong in multiple tests and require clumsy FDR-like corrections; methods that permit negative physical values or variances are clearly wrong; removal of nuisance ('latent') variables is awkward; standard confidence intervals are much less useful than MCMC-derived posterior distributions; frequentist model selection is not very settled; why rely on the 'maximum' of the likelihood function when no one relies on the 'mode' of other distributions?

Frequentist *vs.* Bayesian approaches V

"Bayesian inference (or more generally, Bayesian data analysis) is a method for summarizing uncertainty and making estimates and predictions using probability statements conditional on observed data and an assumed model. ... Bayesian statistics is about making probability statements; frequentist statistics is about evaluating probability statements ... so, not only can a statistician be Bayesian and frequentist at different times, a single analysis can also be both at the same time. " (A. Gelman, 2008)

"Statistics has struggled for nearly a century over the issue of whether the Bayesian or frequentist paradigm is superior. The debate is far from over and, indeed, should continue, since there are fundamental philosophical and pedagogical issues at stake. At the methodological level, however, the debate has become considerably muted, with the recognition that each approach has a great deal to contribute to statistical practice" (Bayarri & Berger, 2004)

Comparisons of frequentist and Bayesian inference

Comparative Statistical Inference, Vic Barnett (3rd edition, 1999)

'The interplay of Bayesian and frequentist analysis', M. J. Bayarri & J. O. Berger, *Statist. Sciences*, 19, 58 (2004)

Modes of Parametric Statistical Inference, Seymour Geisser (2006)

Principles of Statistical Inference, D. R. Cox (2006)

'Objections to Bayesian statistics', A. Gelman & commentators, *Bayes. Anal.* 3, 445 (2008)