Parameter Estimation

William H. Jefferys
University of Texas at Austin
bill@bayesrules.net

Elements of Inference

- Inference problems contain two indispensable elements:
  - Data $x \in X$ (known)
  - Parameters $\theta \in \Theta$ (unknown)
- The challenge is to go from one to the other
  - Inference: $x \Rightarrow \theta$
- [The other direction, $\theta \Rightarrow x$, is also important: This is the *predictive problem*. Generally results in a probability distribution on x, given θ]

Models

- Any inference procedure requires two kinds of models:
  - Physical Model
  - Statistical Model

Physical Model

- Describes the physics and geometry of the data-taking process, from the origination of the signal in the object being studies, through its propagation to the detector, the optics of the telescope, and the detection process
- May involve unknown parameters that describe the physics
  - Unknown parameters are often *nuisance parameters:* We aren't interested in them for themselves, but they have to be estimated in order to measure the parameters we *do* care about
- Example: linear model relating error-free time t to error-free position x and unknown velocity a and offset b. We may care only about a; then b would be a nuisance parameter

$$x = at + b$$

Physical Model

- Describes the physics and geometry of the data-taking process, from the origination of the signal in the object being studies, through its propagation to the detector, the optics of the telesc
- May invo e the physics
  - Unkno *rameters:* We aren't t they have to be esti eters we *do* care about

Accurate modeling is important! For example, in Jay Anderson's problem: Centroiding is sensitive to the PSF model adopted

- Example: linear model relating error-free time *t* to error-free position *x* and unknown velocity *a* and offset *b*. We may care only about *a*; then *b* would be a nuisance parameter

$$x = at + b$$

Statistical Model

- Relates the values of the observations x_0 that we actually *observed* (recorded) to the physical model and the parameters that describe the data taking process (e.g., variances)
- Models the random nature of the data collection process
  - Random arrival times of photons (Poisson process?)
  - Atmospheric fluctuations (Gaussian process?)
  - Unsteadiness in telescope pointing, etc.
- Described by the *likelihood function*
- Example: Error of recorded data has normal distribution:

$$p(x_0 \mid x) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (x^{(i)} - x_0^{(i)})^2 \right)$$

Statistical Model

- Relates the values of the observations x_0 that we actually *observed* (recorded) to the physical model and the parameters that describe the data taking process (e.g., variances)
- Models th process
  - Rando rocess?)
  - Atmos ?)
  - Unstea
- Describec
- Example: Error of recorded data has normal distribution:

$$p(x_0 \mid x) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^{(i)} - x_0^{(i)})^2\right)$$

Accurate statistical models are also important!

Not all data are normally distributed!

Statistical Model

- The probability of observing a particular data set (vector) x_0 depends on both the physical and statistical models:

$$p(x_0 \mid \theta, B)$$

- Here, p is the *sampling distribution*, the probability of observing data x_0 given the *true* parameters describing both the physical and statistical models, which are included in the vector θ (state of nature). B is background information.
- In the previous example, with $x_0 = \{x_0^{(i)}\},\ t = \{t^{(i)}\}$

$$p(x_0 \mid \theta, t) = p(x_0 \mid a, b, \sigma, t)$$

$$= \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (a t^{(i)} + b - x_0^{(i)})^2 \right)$$

$$\theta = \{a, b, \sigma\}$$

Likelihood Function

- Once we have observed a *particular* data set x_0, the fact that we *might* have observed a different set x', but didn't, should be irrelevant to our inference from x_0 to θ.
- x_0 is *not* regarded as a random variable. It is known exactly, even though it arose from a random process.
- The important thing is how $p(x_0|\theta)$ varies with θ, given the *actual* x_0 that we have observed.

$$p(x_0 \mid \theta) = L(\theta; x_0)$$

- The function $L(\theta; x_0)$ is the *likelihood function*. It measures how well each possible value of θ is supported by the actual data x_0 we have observed.

Likelihood Principle

- The *Likelihood Principle* says that the likelihood function contains all of the information about θ that is contained in the data x_0.
- This means that inference should be based on $p(x_0|\theta)$, considered as a function of θ.
  - Values of θ that make the likelihood big are better supported by the data than values of θ that make it small
  - Therefore, to first order we would like to concentrate on values of θ that make $p(x_0|\theta)$ *big*.

Maximum Likelihood

- This leads to the first method of inference: *Maximum Likelihood*
  - Simply choose the value $\hat{\theta}$ of θ that maximizes $p(x_0|\theta)$
  - $\hat{\theta}$ might not be unique
  - (In normal case, justifies minimizing chi-square)
- Note that considered as a function of θ, $p(x_0|\theta)$ is *not* a probability density.
  - It is not normalized
  - It cannot be used as if it were a probability in order to estimate the errors in θ

Maximum Likelihood

- The method of maximum likelihood has many satisfying features:
  - It is easy to apply (you just need a program to maximize some arbitrary multivariate function, and these are readily available)
  - In many, even most situations involving *parameter estimation*, it gives satisfactory values of $\hat{\theta}$
- However
  - In some circumstances it gives bad answers
  - It does not estimate the error in $\hat{\theta}$
  - There are both logical and practical objections to estimating the error by standard methods

Estimating the Error

- A practical method of estimating the error is to note that the procedure that takes us from x_0 to $\hat{\theta}$ generally displays $\hat{\theta}$ as a function of x_0:

$$\hat{\theta} = f(x_0)$$

- But then we can consider generating a large *bootstrap sample* of x from the sampling distribution, assuming that $\hat{\theta}$ is the *true* value of θ:

$$\{x\} = (x_1, x_2, \ldots, x_N) \sim p(x \mid \hat{\theta})$$

- To each x_i there corresponds a θ_i:

$$\theta_i = f(x_i)$$

- It is *plausible* that the distribution of these θ_i displays the uncertainty in our estimated $\hat{\theta}$

Estimating the Error

- This rather roundabout method of looking at the error in $\hat{\theta}$ has a drawback: It violates the Likelihood Principle!
  - It is computed using data that were not observed, but might have been observed, instead of considering only the data that were actually observed
- Nonetheless, the idea is commonly used and (usually) gives reasonable answers.
- In special cases (e.g., linear models with normally distributed errors) this error analysis can be done exactly, rather than by bootstrap simulation as outlined. But simulation can always be used.
  - Exact or simulated, it still violates the Likelihood Principle

Direct Use of Probability Theory

- My choice: Use probability theory directly.
- This goes by the name *Bayesian inference*
- Requires additional input (background information)
- If we have a prior probability density $\pi(\theta)$ on θ, indicating what we know about θ before looking at the data, we can turn $p(x_0|\theta)$ into a probability density on θ:

$$p(\theta \mid x_0) = \frac{p(x_0 \mid \theta)\pi(\theta)}{\int p(x_0 \mid \theta)\pi(\theta)d\theta} = \frac{p(x_0 \mid \theta)\pi(\theta)}{p(x_0)}$$

- This is *Bayes' Theorem*. $\pi(\theta)$ is the *prior distribution* of θ, or *prior*. $p(\theta|x_0)$ is the *posterior distribution* of θ, or *posterior*. The denominator is just a normalizing constant.
- θ is regarded as a *random variable*; x_0 is not

Direct Use of Probability Theory

- My choice: Use probability theory directly.
- This goes by the name *Bayesian inference*
- Requires a)
- If we have dicating what we kn ve can turn $p(x_0|\theta)$ into

$$p(\theta \quad)$$

- This is *Bay on* of θ, or *prior*. $p(\theta|x_0)$ is the *posterior distribution* of θ, or *posterior*. The denominator is just a normalizing constant.
- θ is regarded as a *random variable*; x_0 is not

Conceptually simple
Logically consistent
Powerful

Maximum a Posteriori Inference

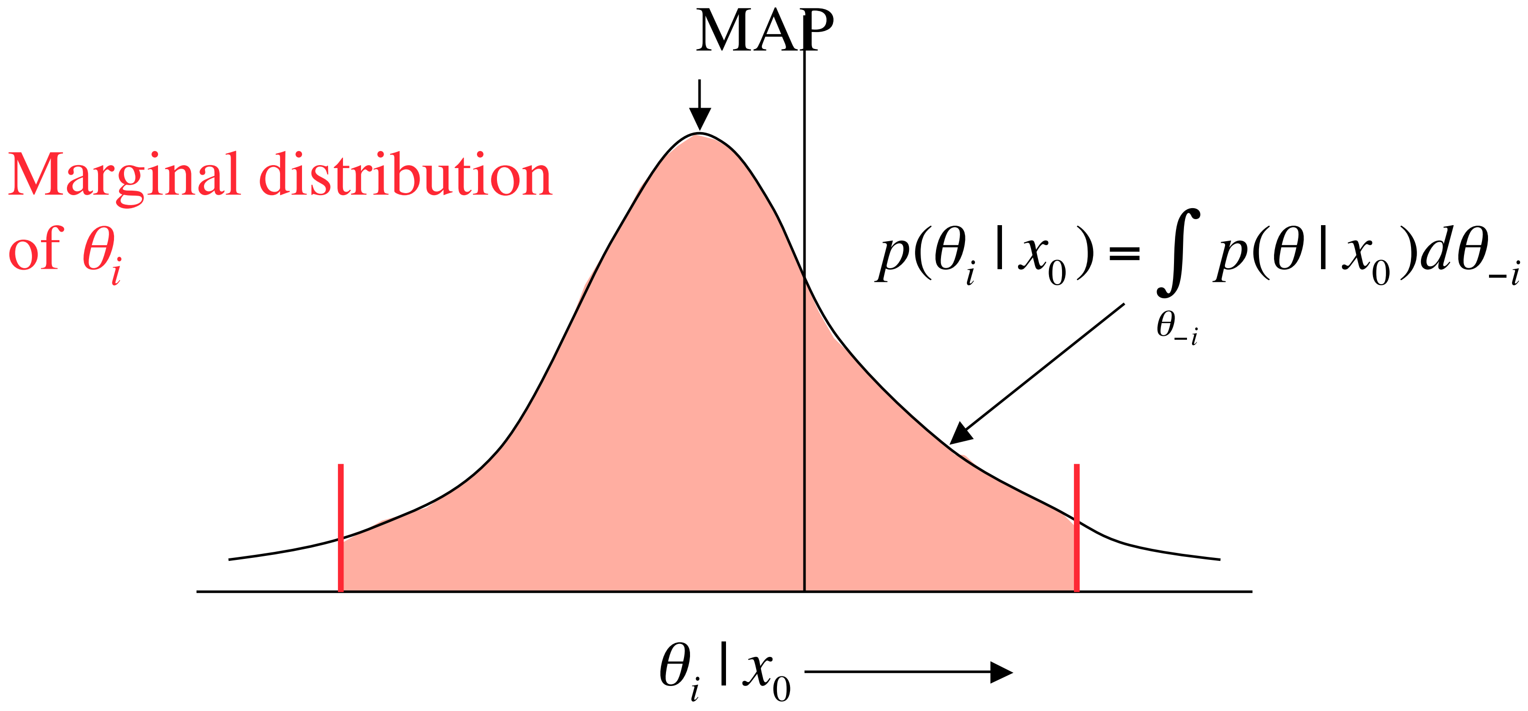
- Suppose we choose a prior, for example, $\pi(\theta)\propto \mathit{constant}$.
  - This particular prior cannot be normalized; but as long as the posterior is normalized, this is not a problem
- Then the posterior distribution will be proportional to the likelihood (considered as a function of θ):

$$p(\theta \mid x_0) \propto p(x_0 \mid \theta)\pi(\theta) \propto p(x_0 \mid \theta)$$

- By maximizing this posterior distribution we obtain the *Maximum a Posteriori Estimate* of θ, also known as MAP
- Since the posterior density is a *genuine* probability density, it can be used *directly* to analyze the error distribution of θ
- Would be like using the likelihood to get the error distribution. It legitimizes this notion.

Using the Posterior Distribution

- 95% Credible Interval: 95% of the total probability lies between the two red lines



The Importance of Priors

- However, $\pi(\theta)\propto constant$ is not reasonable under most circumstances
  - You may have significant prior information about θ. If so, it would be foolish not to use it. For example, you may already have a good idea where θ lies. You can put a prior distribution $\pi(\theta)$ on θ that reflects that information
    - For example, we already have a good idea of the value of the Hubble constant, so to ignore this information might be problematic
  - There may be other constraints that provide important prior information, e.g.,
    - Fluxes are non-negative
  - Without prior information, use objective priors

The Importance of Priors

- An example is when measuring the distance to an object. If we believe that the particular kind of object is distributed uniformly in space (out to some limit, say), then before you even look at the data, you ought to regard the distribution in the distance *s* to be distributed proportionally to

$$\pi(s)ds \propto s^2 ds$$

- Failure to recognize this leads to the so-called *Lutz-Kelker bias*, which was recognized by Trumpler and Weaver in the 1950s but only became widely understood when Lutz and Kelker rediscussed it in the 1970s
  - Lutz and Kelker's discussion was not based on Bayesian priors; but the Bayesian method forces one to consider this when setting up the problem
  - The Lutz-Kelker prior is an objective prior

Hierarchical Bayes Models

- A prior distribution may depend on other unknown parameters not mentioned in the likelihood. If so, these new unknown parameters will themselves need priors
  - In general any unknown parameter needs a prior
- Thus for example with the likelihood $p(x_0|\theta)$, the prior may be $\pi(\theta|\sigma)$ so that we also need a prior $\pi(\sigma)$, obtaining a posterior distribution (excluding the normalizing factor)

$$p(\theta,\sigma \mid x_0) \propto p(x_0 \mid \theta)\pi(\theta \mid \sigma)\pi(\sigma)$$

- The hierarchical Bayes idea provides a very rich class of statistical models that enables the investigator to model the problem of interest more closely

Model Selection

- One may be interested in a class of models with differing numbers of parameters
  - For examples, polynomials of unknown rank, e.g.,

$$x = at + b + \gamma t^2 : \quad \text{Do we need the term in } \gamma ?$$

- Maximum likelihood has difficulty with this
  - Methods are basically *ad hoc*
- The Bayesian approach is straightforward. If *m* denotes the model, then the posterior probability of θ and *m* is given by

$$p(\theta, m \mid x_0) \propto p(x_0 \mid \theta)\pi(\theta \mid m)\pi(m)$$

- The posterior probability of a model *m* is given by integrating out θ:

$$p(m \mid x_0) = \int p(\theta, m \mid x_0) d\theta$$

Model Selection

- One may be interested in a class of models with differing numbers of parameters
  - For exa , e.g.,

    $x =$ $\gamma?$
- Maximum
  - Metho
- The Bayes denotes the model, the is given by

Cluster membership! For each star

-member, not member (model)

-probability it is a member

-decision rule for membership

- The posterior probability of a model *m* is given by integrating out θ:

$$p(m \mid x_0) = \int p(\theta, m \mid x_0) d\theta$$

Model Averaging

- Conversely, none of the models being compared may be believable, as for example when polynomials are being used to approximate some unknown function. In this case one can make estimates of the unknown parameter θ by *model averaging*:

$$p(\theta \mid x_0) = \sum_m p(\theta, m \mid x_0)$$

- This allows us to avoid committing to a particular *m*, but instead to allow the analysis to weight the contributions of each individual model by the posterior probability of the model

$$m_0 : x = at + b$$

versus

$$m_1 : x = at + b + \gamma t^2$$

Simulation

- In practice, it has been very difficult to evaluate the normalizing constant required to produce a normalized posterior distribution
- In the past 15 years, it has become possible to avoid this problem by using computationally intensive simulation to draw a sample from the posterior distribution
- Thus we replace the posterior distribution $p(\theta|x_0)$ by a sample

$$\{\theta\} = (\theta_1, \theta_2, \ldots, \theta_n) \sim p(\theta \mid x_0)$$

- Once we have the sample we can compute the inferences directly from the sample, e.g.,

$$\hat{\theta} = E(\theta) \approx \frac{1}{n} \sum_i \theta_i$$

Simulation

- The methods for simulation are beyond the scope of this talk; The simplest are
  - Importance sampling
  - Acceptance-rejection sampling
- More generally we have *Markov chain Monte Carlo* (MCMC)
  - Gibbs Sampling
  - Metropolis-Hastings sampling
- Other important methods are
  - Reversible-jump MCMC
  - Metropolis-coupled MCMC (MCMCMC=$[MC]^3$)
- The key idea is that one does not need to know the normalizing constant in order to draw the sample

References

- My course notes:
  http://bayesrules.net/ast383.html
- Michael Lavine's Book:
  http://www.stat.duke.edu/~michael/book.html
- Data Analysis: A Bayesian Tutorial (D. S. Sivia. Oxford: Clarendon Press)
- Bayesian Data Analysis, Second Edition (Andrew Gelman, John B. Carlin, Hal S. Stein and Donald B. Rubin. London: Chapman and Hall)