Astronomy 8824: Statistics Notes 2
Bayesian Parameter Estimation

In Bayesian parameter estimation, one has a model that is assumed to describe the data, and the task is to determine its parameters.

Hypothesis is "true value of parameter is $\theta_{\rm true} = \theta$" (discrete) or "true value of parameter is $\theta \le \theta_{\rm true} \le \theta + d\theta$" (continuous).

$$p(\theta|DI) = p(\theta|I)\frac{p(D|\theta I)}{p(D|I)}.$$

If θ is continuous, then, technically, $p(\theta|DI)$ and $p(\theta|I)$ both have a $d\theta$ attached.

A Bayesian searches for the parameter value with *maximum posterior probability* $p(\theta|DI)$.

If $p(\theta|I)$ is flat, then this is also the value with *maximum likelihood* $p(D|\theta I)$.

Maximum likelihood estimators play a major role in both Bayesian and classical approaches.

A simple example: mean of data

Estimate mean from N measurements x_i, when dispersion σ is known, and x_i are Gaussian distributed and independent. (Following Loredo, §5.2.2; see also Ivezic et al. §5.6.1)

Flat prior: $p(\mu|I) = (\mu_{\rm max} - \mu_{\rm min})^{-1}$.

Likelihood:

$$\begin{aligned} p(\{x_i\}|\mu I) &= \prod_i (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{Ns^2}{2\sigma^2}\right] \exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right], \end{aligned}$$

where $\overline{x} = \frac{1}{N}\sum x_i$ is sample mean and $s^2 = \frac{1}{N}\sum (x_i - \overline{x})^2$ is sample variance.

Global likelihood: $p(\{x_i\}|I) = \int_{\mu_{\rm min}}^{\mu_{\rm max}} p(\{x_i\}|\mu I) d\mu$.

Final result is

$$p(\mu|\{x_i\}I)d\mu = K\left(\frac{N}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right], \quad \mu_{\rm min} \le \mu \le \mu_{\rm max},$$

a Gaussian with mean $\overline{x}$ and dispersion $\sigma/\sqrt{N}$, truncated at $\mu_{\rm min}$ and $\mu_{\rm max}$, with K a normalization constant such that the probability integrates to one.

Some comments on priors

In the above example, as long as prior range is big compared to $\sigma/\sqrt{N}$, the prior doesn't matter.

Otherwise it does, by truncation and normalization $K > 1$.

If new measurements come in, they can be incorporated by taking output of this result as *prior* for new analysis.

At least at informal level, this is often done, e.g., H_0 priors on CMB analyses.

To have $p(\theta|DI) \propto p(D|\theta I)$, we need the prior $p(\theta|I)$ to be flat in the range allowed by the data, not universally.

For example, we may know that $\mu > 0$ on physical grounds. If $\bar{x} \gg \sigma/\sqrt{N}$, then $p(\mu|I)$ is approximately flat in the allowable range if it is "broad" compared to $\sigma/\sqrt{N}$. But if $\bar{x} \sim \sigma/\sqrt{N}$, then a flat prior cannot be a good approximation.

For a positive-definite parameter where we have essentially no prior knowledge about its value, a common choice of prior is $p(\theta|I) \propto 1/\theta$, i.e., flat in $\ln\theta$ instead of θ itself.

Maximum Posterior vs. Maximum Likelihood

From a Bayesian point of view, the end result of a parameter estimation calculation *is* the posterior probability distribution $p(\theta|DI)$. Recall that

$$p(\theta|DI) = p(\theta|I)\frac{p(D|\theta I)}{p(D|I)}.$$

For a flat prior $p(\theta|I)$, the posterior is proportional to the likelihood $P(D|\theta I)$.

Frequentist parameter estimation methods often focus on maximum likelihood estimators, so there is much in common between frequentist and Bayesian approaches. Parameter estimates based on $p(\theta|DI)$ or $p(D|\theta I)$ will be similar if $p(\theta|I)$ is flat in the region of parameter space allowed by the data.

If you give an expression for, table of, or plot of the likelihood function, then you have presented all of the evidence of the data, and others can apply prior probabilities or frequentist assessments as they wish. Thus, if statistics are important to your answer, there is much to be said for presenting things this way if you can.

A maximum likelihood example: weighted mean

Suppose that we are estimating the mean from N data points that have different errors ("heteroscedastic" data):

Likelihood:

$$p(\{x_i\}|\mu I) = \prod_i (2\pi\sigma_i^2)^{-1/2} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right] .$$

Take log and set $d\ln L(\mu)/d\mu = 0$.

A few-line derivation shows that the maximum likelihood estimator is

$$\hat{\mu}_w = \frac{1}{\sum_i 1/\sigma_i^2} \sum_i \frac{x_i}{\sigma_i^2} .$$

The contribution of each data point is weighted by its inverse variance.

The variance of this estimator is

$$\mathrm{Var}(\hat{\mu}_w) = \frac{1}{\sum_i 1/\sigma_i^2} \ .$$

ML vs. MAP example: mean of Poisson data

(From Bailer-Jones, §4.4.5)

Suppose we have data $\{y_i\}$ drawn from a Poisson distribution with unknown λ.

For example, we might have a few X-ray photons detected from an astronomical source, and we want to estimate its flux.

We have (for some reason) a prior $P(\lambda) = \exp(-\lambda/a)$ where a is known.

Likelihood:

$$L(\lambda) = \prod_i \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \ .$$

Thus

$$\ln L(\lambda) = \sum_i [y_i \ln\lambda - \lambda - \ln(y_i!)] \ .$$

Setting $d\ln L(\mu)/d\lambda = 0$ gives

$$\sum_i \left(\frac{y_i}{\lambda} - 1\right) = 0 \quad \Longrightarrow \quad \lambda = \frac{1}{N}\sum_i y_i.$$

The maximum likelihood estimate is simply the mean of the data.

For the maximum posterior estimate we want to maximize

$$\ln P(\lambda|\{y_i\}) = \ln L(\lambda) + \ln P(\lambda) + \mathrm{const}$$

with the normalized prior

$$P(\lambda) = \frac{1}{a}e^{-\lambda/a} \quad \Longrightarrow \quad \frac{d\ln P(\lambda)}{d\lambda} = -\frac{1}{a} \ .$$

Differentiating and setting to zero gives

$$\sum_i \left(\frac{y_i}{\lambda} - 1\right) - \frac{1}{a} = 0 \quad \Longrightarrow \quad \lambda = \frac{1}{N + 1/a}\sum_i y_i.$$

The prior favors smaller λ and therefore reduces the estimate.

As the number of data points increases, the influence of the prior decreases because $1/a$ must become smaller relative to N.

In the limit $a \to \infty$ the prior is flat and the solution goes to the maximum likelihood solution.

Confidence intervals

We often *summarize* the results of a calculation with an estimate and a confidence interval. Bayesians seem to prefer the term "credible region" to "confidence interval," but they seem to me nearly interchangeable, even though they are based on different concepts of probability.

Typically, one would quote the maximum likelihood (or maximum posterior probability) value as the estimate, though if the likelihood function is far from Gaussian people sometimes quote the likelihood weighted mean.

The confidence interval is a region of highest likelihood (or posterior probability) and is characterized by the fraction of the probability that it contains.

For a 1-dimensional problem (1 parameter), this is usually straightforward, though even here a complicated likelihood function may have multiple maxima.

For a Gaussian likelihood function,

$$\ln\mathcal{L} = \ln\mathcal{L}_{\rm max} - \frac{1}{2}\Delta\chi^2, \qquad \mathcal{L} = \mathcal{L}_{\rm max} e^{-\Delta\chi^2/2}.$$

(We'll have more to say about χ^2 and $\Delta\chi^2$ shortly, but for now you can regard $\Delta\chi^2$ as a measure of the deviation of $p(D|\theta I)$ from its maximum value.)

The regions $\Delta\chi^2 \le 1$, $\Delta\chi^2 \le 4$, and $\Delta\chi^2 \le 9$ contain 68.3%, 95.4%, and 99.73% of the probability. Since a Gaussian is $(2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2}$, these values of $\Delta\chi^2$ correspond to 1σ, 2σ, and 3σ deviations.

For a non-Gaussian likelihood function, it can be useful instead to quote the values where $\mathcal{L}$ falls to some fraction of its maximum value, say 0.1, in which case the parameter value is 10 times less probable than its most probable value. This particular fraction corresponds in the Gaussian case to 2.14σ, since $e^{-2.14^2/2} = 0.1$.

If there are multiple parameters, errors on different parameters may be correlated.

Confidence intervals are defined by contours in a multi-dimensional parameter space.

If the likelihood function is a multi-variate Gaussian, then the confidence contours are ellipses, with the direction of the ellipse axes depending on the covariance of the errors in the parameters.

For the 2-d case, the contours $\Delta\chi^2 = 2.30$, 6.17, and 11.80 enclose 68.3%, 95.4%, and 99.73% of the probability. ($\Delta\chi^2 = 0.21$ contains 99%.) See the *Numerical Recipes* chapter on "Modeling of Data" for higher dimensions and more discussion.

These are sometimes referred to as "1σ", "2σ", and "3σ" regions, when there are multiple dimensions this usage is loose at best and can be misleading.

In some cases, a sensible choice of parameters will eliminate or minimize covariance, making results easier to interpret. An obvious case is the slope and intercept of a linear fit. These

are usually highly correlated, but the covariance can be eliminated by defining the intercept at an appropriate "pivot point," fitting $y = a(x - x_p) + b$ instead of $y = ax + b$.

Marginalization

Suppose that we are simultaneously fitting multiple parameters θ_i but that we would like to know the confidence interval for one of them in particular, e.g., θ_1.

One of the strengths of Bayesian statistics is that it offers a clear way of doing this:

$$p(\theta_1|DI) = \int p(\{\theta_i\}|DI) d\theta_2 d\theta_3 ... d\theta_n.$$

This procedure of integrating over "nuisance parameters" is called "marginalization." (The above expression is the marginal pdf of θ_1.)

The approach doesn't make sense in the frequentist framework because one cannot talk about the probability of a parameter value.

Example: Suppose we have data that we are using to estimate the slope a, intercept b, and intrinsic scatter σ of a linear relation between x and y.

If we just want to know the posterior distribution for the slope, we can find it from

$$p(a|DI) = \int_{-\infty}^{\infty} db \int_0^{\infty} d\sigma p(ab\sigma|DI).$$

We don't have to go down to a single dimension, e.g., if we don't care about the dispersion σ but would like to know the joint distribution of a and b:

$$p(ab|DI) = \int_0^{\infty} d\sigma p(ab\sigma|DI) d\sigma.$$

Marginalization plays a crucial role in, for example, cosmological analyses of CMB and large scale structure data, where the cosmological model being fit typically has 6-10 free parameters but we are often interested in learning about constraints on specific ones, such as H_0 or the effective number of neutrino species.

Systematic uncertainties in the measurements can often be treated by introducing a nuisance parameter that describes them, such as a calibration offset, imposing some prior, and then marginalizing over these nuisance parameters when fitting for other parameters of physical interest.

Of course, sometimes one astronomer's nuisance is another astronomer's science, and vice versa.

Straight line fitting: the "standard" case

Determine maximum likelihood values of a and b in a linear fit $y = ax + b$, given data points with known errors on y, assuming Gaussian error distribution:

$$p(\hat{y}_i|y_i) = (2\pi\sigma_{y,i}^2)^{-1/2} \exp\left[\frac{-(\hat{y}_i - y_i)^2}{2\sigma_{y,i}^2}\right],$$

where y_i is the true value and $\hat{y}_i$ is the observed value.

Likelihood

$$\mathcal{L} = p(\{\hat{y}_i\}|a,b) = \prod_i p(\hat{y}_i|ax_i+b)$$

$$= \prod_i (2\pi\sigma_{y,i}^2)^{-1/2} \exp\left[-\frac{(\hat{y}_i - ax_i - b)^2}{2\sigma_{y,i}^2}\right].$$

It is often convenient to work with the logarithm of the likelihood

$$\ln\mathcal{L} = -\frac{1}{2}\sum \frac{(\hat{y}_i - ax_i - b)^2}{2\sigma_{y,i}^2} + C,$$

where C depends on the (known) errors $\sigma_{y,i}$ but is independent of a and b.

The maximum likelihood solution is thus the solution with minimum

$$\chi^2 = \sum \frac{(\hat{y}_i - ax_i - b)^2}{\sigma_{y,i}^2},$$

and $\ln\mathcal{L} = -\chi^2/2 + C$.

For this problem, one can find standard analytic expressions (e.g., Numerical Recipes §15.2) for a and b in terms of the data and error bars by solving the equations that define the maximum of the likelihood function,

$$\frac{\partial\ln\mathcal{L}}{\partial a} = 0, \qquad \frac{\partial\ln\mathcal{L}}{\partial b} = 0.$$

Straight line fitting: a non-standard but very useful case

Now consider a more complicated variation of this problem: fit $\overline{y} = ax + b$, with measurement errors in x and y *and* intrinsic scatter in the relation between y and x.

A model with intrinsic scatter (here assumed constant from point to point and denoted σ) is usually more realistic than the commonly adopted, perfect correlation model.

If all of the scatters are Gaussian distributed, we have

$$p(y_i|x_i) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(y_i - ax_i - b)^2}{2\sigma^2}\right]$$

$$p(\hat{y}_i|y_i) = (2\pi\sigma_{y,i}^2)^{-1/2} \exp\left[\frac{-(\hat{y}_i - y_i)^2}{2\sigma_{y,i}^2}\right]$$

$$p(\hat{x}_i|x_i) = (2\pi\sigma_{x,i}^2)^{-1/2} \exp\left[\frac{-(\hat{x}_i - x_i)^2}{2\sigma_{x,i}^2}\right].$$

In this case we want to maximize

$$\mathcal{L} = \prod_i p(\hat{y}_i|\hat{x}_i) \qquad \Longrightarrow \qquad \ln\mathcal{L} = \sum_i \ln p(\hat{y}_i|\hat{x}_i).$$

So we need the expression for $p(\hat{y}_i|\hat{x}_i)$.

$$p(\hat{y}_i|\hat{x}_i) = \int_{-\infty}^{\infty} dy_i\, p(\hat{y}_i|y_i)\, p(y_i|\hat{x}_i)$$
$$= \int_{-\infty}^{\infty} dy_i\, p(\hat{y}_i|y_i) \int_{-\infty}^{\infty} dx_i\, p(y_i|x_i)\, p(x_i|\hat{x}_i).$$

Now assume a flat prior on x_i, $p(x_i)$ =const., so that $p(x_i|\hat{x}_i) = p(\hat{x}_i|x_i)$ (by Bayes' theorem and the requirement that probabilities integrate to one). This assumption is non-trivial, but usually OK because we only require flatness over the range allowed by $\hat{x}_i$.

We can now substitute our expressions for the probabilities, and several pages of algebra and integrals lead eventually to the expression

$$p(\hat{y}_i|\hat{x}_i) = (2\pi)^{-1/2}(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)^{-1/2} \exp\left[-\frac{(\hat{y}_i - a\hat{x}_i - b)^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)}\right].$$

This expression looks eminently sensible. For $\sigma_{x,i} = 0$, we get a Gaussian whose width is the quadrature sum of the intrinsic and observational scatter in y. Non-zero $\sigma_{x,i}$ increases the probability of larger deviation between observed and predicted y_i by allowing the true value of $ax_i + b$ to be closer to $\hat{y}_i$ than $a\hat{x}_i + b$.

A deviation $\Delta y_i/\sigma_{y,i}$ has similar weight to a deviation $a\Delta x_i/\sigma_{x,i}$. If you think of x and y as having different units, then it is obvious that a factor of a is needed to give $\sigma_{y,i}$ and $a\sigma_{x,i}$ the same dimensions.

The maximum likelihood solution requires maximizing

$$\sum_i \ln p(\hat{y}_i|\hat{x}_i) = -\frac{1}{2}\sum_i \ln(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)$$
$$- \sum_i \frac{(\hat{y}_i - a\hat{x}_i - b)^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)} + \text{constant},$$

and thus solving the equations

$$\frac{\partial \ln\mathcal{L}}{\partial a} = 0, \qquad \frac{\partial \ln\mathcal{L}}{\partial b} = 0, \qquad \frac{\partial \ln\mathcal{L}}{\partial \sigma} = 0.$$

There is a straightforward algebraic solution for b,

$$b = \frac{-\sum_i (a\hat{x}_i - \hat{y}_i) W_i}{\sum_i W_i},$$

where the weights are

$$W_i = \frac{1}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)}\ .$$

This is just an inverse-variance weighted average of the individual estimates of b.

I couldn't find algebraic solutions for a and σ, but it is straightforward to search a grid of (a, σ), finding the best b for each (a, σ) from the above equation and evaluating the overall likelihood.

There are a couple of points worth noting about the likelihood expression.

First, you might naively have thought that with intrinsic scatter as a free parameter, the maximum likelihood solution would be to have a very large intrinsic scatter, since then each deviation would contribute very little to χ^2.

However, while the second term in the likelihood always rewards large σ^2, the first term penalizes it, basically because the prediction $ax + b$ is diluted by being spread over a large range, so it doesn't get much "credit" when it is close.

If a significant fraction of points have deviations that put them on the exponential tail of the Gaussian, then raising σ will increase the likelihood, but once the typical deviation falls to $\sim 1\sigma$, raising σ will decrease the likelihood.

This is, of course, what ought to happen. If the prediction is a scatterplot (as happens in the limit of large intrinsic scatter), then it is unlikely to actually have the points lie close to a line.

Second, if we reverse the roles of y and x, letting the intrinsic scatter be on x rather than y, then the solution for a and b (especially a) will be different.

Intrinsic scatter on y is a *different* hypothesis from intrinsic scatter on x, and the corresponding best fit slopes and intercepts are different.

The difference goes away if σ is small compared to the observational errors.