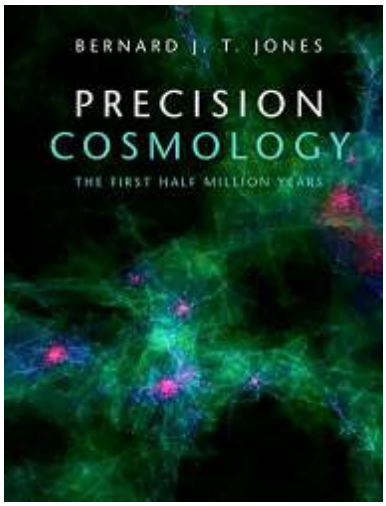
Likelihood Examples

A Supplement to “Precision Cosmology”

Bernard Jones



Further examples of the use of Likelihood

This is one of a set of Supplementary Notes and Chapters to “Precision Cosmology”.
Some of these Supplements might have been a chapter in the book itself, but were regarded either as being somewhat more specialised than the material elsewhere in the book, or somewhat tangential to the main subject matter.
The are mostly early drafts and have not been fully proof-read.
Please send comments on errors or ambiguities to “PrecisionCosmology(at)gmail.com”.

Contents

1 Likelihood in action

1.1 Likelihood Functions

1.1.1 Examples: Exponential, Gaussian and Bernoulli

If the measurements $\{x_i\}$ of an experiment are taken from a Gaussian distribution about a to-be-determined mean μ and variance σ we can write the probability of making those particular n measurements as

$$\mathcal{L}(\mu, \sigma \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp - \left(\frac{(x_i - \mu)^2}{2\sigma^2} \right) \tag{1.1}$$

Taking the log of this we get

$$\ln \mathcal{L}(\mu, \sigma \,|\, x_1, \ldots, x_n) = -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} - \frac{n}{2} \ln(2\pi\sigma^2) \tag{1.2}$$

We shall use this in the next section to determine μ and σ from the measurements $\{x_i\}$.

1.1.2 Examples: Bernoulli trials

Bernoulli trials

Ex 1.1.1 An experiment consists of n Bernoulli trials each with the same probability of success, θ. The result of a single experiment is either '1' or '0' with probability θ or $(1 - \theta)$. The result of the sequence of n such trials can be thought of as a sequence of 1's ("success") and 0's ("failure"), and so represented by a vector $\mathbf{x} = \{x_1, \ldots, x_n\}$ of zeros and ones. Show that the likelihood of θ is

$$\mathcal{L}(\theta \,|\, \mathbf{x}) = \theta^r (1 - \theta)^{n-r}, \quad r = (\Sigma_i x_i) \text{ is the number of 1's in } \mathbf{x} \tag{1.3}$$

Show that the log likelihood for this experiment is

$$\log \mathcal{L}(\theta \,|\, \mathbf{x}) = n\bar{x} \ln \theta + (n - r)(1 - \bar{x}) \ln(1 - \theta), \quad \bar{x} = r/n \tag{1.4}$$

1.2 Maximum Likelihood estimators

Check this carefully!

Suppose that we have a theory involving some to-be-determined parameters a_j ($j = 1, \ldots, n$). The theory predicts values of measurable quantities ξ_i ($i = 1, \ldots, N$) in terms of the a_j: $\xi_i = \xi_i(a_1, \ldots, a_n)$. Now perform an experiment the outcome of which is a measurement of the N quantities ξ_i, yielding values x_i with Gaussian distributed errors having standard deviation σ_i ($i = 1, \ldots, N$). How do we estimate the parameters a_j of the theory?

The Likelihood function is proportional to the probability of getting this data set, and given the Gaussian nature of the error distribution, this is

$$\mathcal{L}(a_1, \ldots, a_n) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2\sigma_i^2}\right] \tag{1.5}$$

Remember that the ξ_i are the quantities that our experiments would measure on the basis of a model parametrised by the parameters a_j: that is where the a_j come in on the right hand side.

We wish to select the parameters a_j so as to maximize this expression. This is obviously equivalent to minimizing the exponent, which we can write as

$$\frac{1}{2}\chi^2(a_j) = \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2\sigma_i^2} \tag{1.6}$$

Because we are minimizing a sum of squares, this is referred to as a *Least Squares Fit*. Differentiating with respect to each of the a_j yields a set of equations

$$\sum_{i=1}^{N} \frac{(x_i - \xi_i)}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_j} = 0, \quad j = 1. \ldots, n. \tag{1.7}$$

For a general function $\xi(a_1, \ldots, a_n)$ this is a nonlinear set of equations, and the fitting procedure is referred to more precisely as a *nonlinear least squares fit*.

We can make easy progress if the function $\xi(a_1, \ldots, a_n)$ is linear in the a_j. Write this linear relationship in matrix form for convenience:

$$\xi = \mathbf{C}\mathbf{a} \tag{1.8}$$

where the matrix $\mathbf{C}$ is not necessarily square. Define a new matrix $\mathbf{D}$ whose elements are

$$D_{ij} = \frac{C_{ji}}{\sigma_j^2} \tag{1.9}$$

and with this calculate the *data vector* $\mathbf{X}$ corresponding to the measured data $\mathbf{x}$:

$$\mathbf{X} = \mathbf{D}\mathbf{x} \tag{1.10}$$

This involves known quantities on the right hand side. It is then easy to show that

$$\mathbf{X} = \mathbf{M}\mathbf{a}, \qquad \mathbf{M} = \mathbf{D}\mathbf{C} \tag{1.11}$$

Even though **C** may not be not square, the *measurement matrix* **M** is square and symmetric and depends only on the errors.

The maximum likelihood estimator of the parameters a_j is then

$$\mathbf{a} = \mathbf{M}^{-1}\mathbf{X} \tag{1.12}$$

It can be shown that the standard error for the parameter a_j is given by the jth diagonal element of the inverse of **M**:

$$\Delta a_j = \sqrt{(\mathbf{M}^{-1})_{jj}} \tag{1.13}$$

(with no summation on repeated indices). Hence $\mathbf{M}^{-1}$ is called the *error matrix*.

Note that $\mathbf{M}^{-1}$ may have off-diagonal terms: this would happen if the parameters a_j were not statistically independent so that their errors were not independent.

1.3 Maximum Likelihood parameter estimation

Let us look at the examples of likelihood functions derived in section 1.1.

1.3.1 Exponential ML, again

Equation (**??**) gives the log-likelihood for the fitting the exponential distribution model to the decay lifetimes of particles. The maximum of the log-likelihood is given by

$$\frac{d}{dt}\ln\mathcal{L}(\tau) = \frac{1}{\tau^2}\sum_{i=1}^{n} t_i - \frac{n}{\tau} = 0 \tag{1.14}$$

so that the maximum likelihood estimator for the lifetime is

$$\tau_{ML} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{1.15}$$

This is just the mean of the observed lifetimes. Formally we should write this as $\tau_{ML}(t_1 \ldots, t_n)$ since its value depends directly on the data.

Ex 1.3.1 Show that

$$\mathbb{E}\left[\tau_{ML}(t_1, \ldots, t_n)\right] = \tau \tag{1.16}$$

This shows that this estimate is an *unbiased estimator* of τ. Not all maximum likelihood estimators are unbiased, except in the limit of infinitely large samples. Hint: Since the individual observations t_i are independent, you can express $\mathbb{E}\left[\tau_{ML}(t_1, \ldots, t_n)\right]$ as a product of integrals over the p.d.f's $p(t_i, \tau) = \exp(-t_i/\tau)$ of the i^{th} observation (see equation **??**).

Ex 1.3.2 The decay rate is the quantity $\lambda = 1/\tau$. Use the invariance property of the likelihood function to show that

$$\lambda_{ML}(t_1, \ldots, t_n) = \frac{n}{\Sigma_i t_i} \tag{1.17}$$

Note that this estimator of the decay rate is not unbiased[1] except in the limit of large samples:

$$\mathbb{E}[\lambda_{ML}] = \frac{n}{n-1}\lambda \tag{1.18}$$

To demonstrate this is more difficult than deriving (1.16) since you will need to know how to do a particular multiple integral[2].

1.3.2 Gaussian ML

There are two parameters to determine when the model for the data $\{x_1, \ldots, x_n\}$ is a Gaussian distribution for a single random variable: the mean μ and the variance σ. The maximum of the likelihood function occurs where both derivatives of $\ln \mathcal{L}(\mu, \sigma)$ are zero:

$$\frac{\partial}{\partial \mu}\mathcal{L}(\mu, \sigma) = 0 : \quad \Rightarrow \quad \mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.19}$$

$$\frac{\partial}{\partial \sigma}\mathcal{L}(\mu, \sigma) = 0 : \quad \Rightarrow \quad \sigma_{ML} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{ML})^2 \tag{1.20}$$

There is a simpler alternative to getting the value of μ_{ML} than to maximise $\ln \mathcal{L}(\mu, \sigma)$, which is to note that $\mathcal{L}$ as given in equation (1.1) can be written

$$\mathcal{L}(\mu, \sigma \,|\, x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \tag{1.21}$$

which reaches a maximum when the exponent reaches a minimum. We can write the exponent as

$$\frac{1}{2}\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \tag{1.22}$$

Minimising this with respect to μ yields the same estimate as before. It should be noted that this estimator of the variance of the Gaussian is a biased estimator. The interest of this is that it shows the close relationship between the maximum likelihood estimator and *least squares fitting* in the case of Gaussian distributions.

[1] Reminder: an estimator $X_{estimate}$ of a random variable X is unbiased if $X_{estimate} = \mathbb{E}[X]$. The estimate $X_{estimate}$ is said the be *consistent* if in the limit of large samples $\lim_{n\to\infty} X_{estimate} = \mathbb{E}[X]$.

[2] To do this you need to know that

$$\int_0^\infty \cdots \int_0^\infty \frac{e^{-(x_1 + \cdots + x_n)}}{x_1 + \cdots + x_n} dx_1 \ldots dx_n = \frac{1}{n-1}$$

This is derived from Gradshteyn and Ryzhik (2007, entry #4.638.1).

1.3.3 Bernoulli ML

Going back to Exercise (1.1.1), we can find the maximum of the log-likelihood by

$$\frac{d\ln\mathcal{L}(\theta)}{d\theta} = \frac{r}{\theta} - (n-r)\frac{1}{1-\theta} \Rightarrow \quad \theta_{ML} = \frac{r}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.23}$$

which is the fraction of 1's in the sequence $\{x_i\}$.

1.3.4 Lognormal ML

The p.d.f. for the lognormal distribution with mean μ and variance σ^2 is

$$f_X(x) = \frac{1}{x\sqrt{2\pi}\,\sigma}\exp -\frac{(\ln x-\mu)^2)}{2\sigma^2} \tag{1.24}$$

If we assume, for simplicity, that the variance is known, the only part of the expression for the log-likelihood of a dataset $\{x_i\}, i = 1,\ldots,n$ is the term containing μ:

$$\log\mathcal{L}(\mu) = \mu\text{-independent stuff } -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(\ln x_i-\mu)^2\right) \tag{1.25}$$

Differentiating with respect to μ gives

$$\frac{d\log\mathcal{L}(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum(\ln x_i-\mu) \tag{1.26}$$

and so the maximum likelihood estimator of μ is

$$\mu_{ML} = \frac{1}{n}\sum \ln x_i \tag{1.27}$$

This is just the mean of the log-data.

Ex 1.3.3 Given a lognormal model of known variance σ^2 and a dataset $\{x_i\}, i = 1,\ldots,n$, what is the maximum likelihood estimate of μ[3].

Ex 1.3.4 If data $\{x_i\}, i = 1,\ldots,n$ is to be modelled by a lognormal distribution of which neither the mean μ nor variance σ^2 is known, show that the maximum likelihood estimators of μ,σ are given by

$$\mu_{ML} = \frac{1}{n}\sum \ln x_i, \quad \sigma^2_{ML} = \frac{1}{n}\sum(\ln x_i-\mu_{ML})^2 \tag{1.28}$$

Hint: use the invariance property of the maximum likelihood estimators (1.19) and (1.20) of the parameters defining the Gaussian distribution from which the lognormal is derived.

1.4 Example of error estimates

1.4.1 Exponential distribution

An example where there is lack of symmetry in the likelihood function is given by our particle decay model. We have

$$\frac{\partial^2 \ln \mathcal{L}(\tau)}{\partial \tau^2} = -\frac{2}{\tau^3}\sum_{i=1}^{n} t_i + \frac{n}{\tau^2} = -\frac{n}{\tau^2} \tag{1.29}$$

This gives an error estimate based on the curvature of the likelihood function of

$$\sigma_\tau = \left(-\frac{\partial^2 \ln \mathcal{L}}{\partial \tau^2}\right)^{-1/2}\Bigg|_{\tau=\tau_{ML}} = \frac{\tau_{ML}}{n^{\frac{1}{2}}} \tag{1.30}$$

The problem with this in this case is that the log-likelihood function $\ln \mathcal{L}(\tau)$ is not symmetric about its maximum[3] and so the temptation to quote error bars on the determination of τ, on the basis of the data $\{t_i\}$, in the form $\tau_{ML} \pm \sigma$ would be misleading.

One way out of this in this case is not to use the curvature of the log-likelihood, but to use an alternative error estimate. An alternative to curvature is the width of $\log \mathcal{L}$ at half the maximum height $\log \mathcal{L}(\tau_{ML})/2)$. If we define the left and right errors σ_-, σ_+ by

$$\sigma = \sigma_-, \sigma_+ : \quad \log \mathcal{L}(\tau + \sigma) = \frac{1}{2}\log \mathcal{L}(\tau_{ML}) \tag{1.31}$$

we can then quote the result of the experiment as $\tau^{+\sigma_+}_{-\sigma_-}$, remembering to say that this is the half-height error estimate.

1.4.2 Fitting power laws: Pareto

A simple example of likelihood parameter estimation is given by the common astronomical problem of fitting a power law to a set of data taking positive values. This was first tackled in the astronomical literature by Crawford et al. (1970) who discussed it in the context of fitting radio source counts and has been taken up more recently by Maschberger and Kroupa (2009) who discussed it in the context of stellar mass functions. Other, more general treatments are found in Beg (1983) and Aban et al. (2006).

The statistical distribution to be fitted to a set of n data points is the power law

$$p(x; \alpha, x_{min}, x_{max}) = \frac{\alpha - 1}{x_{min}^{1-\alpha} - x_{max}^{1-\alpha}}\, x^{-\alpha}, \quad \alpha > 1,\ 0 < x_{min} < x < x_{max} \tag{1.32}$$

For $\alpha > 1$ the lower bound x_{min} is necessary for convergence. For the purpose of this example we shall consider x_{min} to be known and derive an equation for the estimator α_{ML} of the slope parameter α.

[3] It is easily verified that $\left.\frac{\partial^3 \ln \mathcal{L}}{\partial \tau^3}\right|_{\tau_{ML}} \neq 0$

The likelihood function for observations $\{x_i\}$ at n points is

$$\mathcal{L}(\alpha, x_{max}) = \prod_{i=1}^{n} p(x_i; \alpha, x_{max}) \tag{1.33}$$

from which we can calculate the log likelihood function as

$$\ln \mathcal{L} = n \ln(\alpha - 1) - n \ln(x_{min}^{1-\alpha} - x_{max}^{1-\alpha}) - \alpha \sum_{i=1}^{n} \ln x_i \tag{1.34}$$

This can be differentiated with respect to α to find the maximum likelihood estimator, α_{ML}.

We can simplify things further by making the simplification that $x_{max} \to \infty$. This then becomes the *Pareto Distribution* which is commonly seen in the world of economics. Our likelihood is then

$$\ln \mathcal{L}(\alpha) = n \ln(\alpha - 1) + n(\alpha - 1) \ln x_{min} - \alpha \sum_{i=1}^{n} \ln x_i \tag{1.35}$$

Here $\mathcal{L}(\alpha)$ does not contain an explicit dependence on x_{min} since we have temporarily assumed that the value of x_{min} is known. Notice that this is a nonlinear function of the model parameter α that we wish to determine, and so it does not fall into the form of equation (1.8) describing linear least squares fitting.

Maximizing this likelihood relative to the parameter α gives:

$$\frac{\partial L}{\partial \alpha} = \frac{N}{\alpha - 1} + N \ln x_{min} - \sum_{i=1}^{N} \ln x_i = 0 \tag{1.36}$$

which provides the maximum likelihood estimator $\hat{\alpha}$ of α:

$$\hat{\alpha} - 1 = \left[\frac{1}{N} \sum_{i=1}^{N} \ln \frac{x_i}{x_{min}} \right]^{-1} \tag{1.37}$$

It should be noted that $(\hat{\alpha} - 1)^{-1}$ is simply the mean of the logarithms of the normalized observations.

The error analysis follows simply by noting that

$$\sigma_\alpha^{-2} = -\left\langle \frac{\partial^2 L}{\partial \alpha^2} \right\rangle \simeq -\left(\frac{\partial^2 L}{\partial \alpha^2} \right)_{\alpha = \hat{\alpha}} \tag{1.38}$$

whence the expected statistical error is

$$\sigma_\alpha \simeq \frac{1}{\sqrt{N}} (\alpha - 1) \tag{1.39}$$

The situation with finite x_{max} is considerably more complex.

Note that this estimator is not an unbiased estimator, though the bias disappears as the sample size increases.

1.4.3 Estimating x_{min}

What happens if we do not know x_{min}, how do we determine it? That is part of the beauty of the likelihood approach: if we want an estimator for x_{min} we simply regard it as a parameter on the same footing as α and regard the log likelihood as a function of the variables to be determined: $L = L(x_{min}, \alpha)$. It is evident from equation (1.35) that $L(x_{min}, \alpha)$ is a monotonic increasing function of x_{min} for $\alpha > 1$. So the maximum of the log likelihood, viewed as a function of x_{min} is achieved for the smallest of the $\{x_i\}$. If we regard the data values $\{x_i\}$ as an ordered sequence then

$$\hat{x}_{min} = x_1, \tag{1.40}$$

the smallest of the data values. This should not occasion any surprise.

Ex 1.4.1 Show that for n observations $\{x_i, i = 1, \ldots, n\}$ of the model $f(x, \theta) = \theta^{-1},\ 0 < x < \theta$:

$$\mathcal{L}(\theta) = \theta^{-n}, \quad 0 < \max\{x_i\} < \theta \tag{1.41}$$

This is an example of a likelihood function that is not bell-shaped.

1.5 ML linear regression

As an important example of the maximum likelihood method we can look at the task of fitting a straight line through a set of points (x_i, y_i) in which the y_i values are subject to an uncertainty, or error, that is modelled as a Gaussian distribution of zero mean and variance σ^2 [4]. Note that the variance does not depend on the value of x_i. Our model for the distribution of points is then

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1.42}$$

where β is the to-be-determined slope of the line and α is its y-axis intercept. The errors ϵ_i are $N(0, \sigma)$ distributed random variables. The parameters to be determined for the data $\mathcal{D} : (\mathbf{x}, \mathbf{y}) = \{(x_i, y_i),\ i = 1, \ldots, n\}$ are α, β, σ^2. The likelihood $\mathcal{L}(\alpha, \beta, \sigma^2)$ is the product of the individual likelihoods:

$$\mathcal{L}(\alpha, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 \right) \tag{1.43}$$

[4] We could write this as $\sigma_{y|x}$ to emphasise that this is the variance in y given x

since, according to our model, $\epsilon_i = y_i - (\alpha + \beta x_i)$. The log-likelihood is

$$\ln \mathcal{L}(\alpha, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 \tag{1.44}$$

Finding the maximum of $\ln \mathcal{L}(\alpha, \beta, \sigma^2)$ by differentiating with respect to each of the parameters α, β, σ^2 in turn yields the system of equations[5]

Ex 1.5.1

$$\frac{\partial}{\partial \alpha}\ln \mathcal{L}(\alpha, \beta, \sigma^2): \quad \alpha n + \beta \Sigma x_i = \Sigma y_i, \tag{1.45}$$

$$\frac{\partial}{\partial \beta}\ln \mathcal{L}(\alpha, \beta, \sigma^2): \quad \alpha \Sigma x_i + \beta \Sigma x_i^2 = \Sigma x_i y_i \tag{1.46}$$

$$\frac{\partial}{\partial \sigma^2}\ln \mathcal{L}(\alpha, \beta, \sigma^2): \quad \sigma^2 = \frac{1}{n}\sum (y_i - \hat{y}_i)^2 \tag{1.47}$$

where α_{ML} and β_{ML} are the solutions of (1.46) and (1.46) and $\hat{y}_i = (\alpha_{ML} + \beta_{ML} x_i)$ is the maximum likelihood fit to the value of y at x_i.

There are two remarks to be made about this. Firstly, the maximum likelihood estimate (1.47) of the variance σ^2 is biased and underestimates the the variance. The unbiased sample variance is

$$s^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 \tag{1.48}$$

The divisor is $n - 2$ because there are two other disposable constants, α, β, that we can choose so as to make s^2 as small as possible. The second remark is that the matrix

$$\mathbf{I} = \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix} \tag{1.49}$$

is called the *information matrix*.

Ex 1.5.2 Show from equation (1.44),

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial^2}{\partial \alpha^2} & \dfrac{\partial^2}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2}{\partial \alpha \partial \beta} & \dfrac{\partial^2}{\partial \beta^2} \end{pmatrix} \ln \mathcal{L}(\alpha, \beta, \sigma^2) = -\frac{1}{\sigma^2}\begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix} \tag{1.50}$$

This is the *Hessian* of the log-likelihood.

[5] There is a complete discussion of this from the point of view of χ^2 fitting in Press et al. (2007, Section 15.2), where the discussion allows each measurement to have its own variance. Their matrix S in that section is the normalised information matrix of equation 1.49 .

1.6 The Inverse Gamma distribution

The Inverse Gamma distribution plays a special role in Bayesian inference. It is a function that can conveniently be used as a prior when estimating of the variance of a Gaussian whose mean is known: it yields analytic expressions for the estimates. Of course, there has to be a judgement of the appropriateness of such a choice. Such convenient pairs of distributions are referred to as *conjugate prior distributions*.

The *Inverse Gamma Distribution* of a variate x is defined by two parameters: the *shape parameter* $\alpha > 0$ and the *scale parameter* $\beta > 0$:

$$g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{-\alpha-1}\, e^{-\beta/x} \tag{1.51}$$

where $\Gamma(\alpha)$ is the standard Gamma function. The inverse gamma distribution is the distribution of a variate whose inverse is itself Gamma distributed.

The following example shows how to estimate the parameters of the Inverse Gamma function, given a set of observations drawn from that distribution.

Ex 1.6.1 We wish to provide estimators for α and β given a set of N observations X_i from this distribution. Show that the log-likelihood L is

$$L = \sum_{i=1}^{N}\left[\ln\frac{\beta^\alpha}{\Gamma(\alpha)} - (\alpha+1)\ln X_i - \frac{\beta}{X_i}\right] \tag{1.52}$$

Ex 1.6.2 By differentiating L with respect to each of α and β, show that the likelihood estimators α and β are given by

$$\frac{\alpha}{\beta} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i}, \quad \ln\beta = \frac{1}{\Gamma(\alpha)}\frac{d\Gamma(\alpha)}{d\alpha} + \frac{1}{N}\sum_{i=1}^{N}\ln X_i \tag{1.53}$$

Ex 1.6.3 Hence show that α is given by the solution of

$$\ln\alpha - \psi(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\ln X_i + \ln\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i}\right] \tag{1.54}$$

where $\psi(\alpha)$ is the *Digamma* or *Psi* function defined as $\psi(x) = \Gamma'(x)/\Gamma(x)$.

Ex 1.6.4 How would you propose to solve equation (1.54)?

1.7 A simple astronomical example

Also need a more substantial example

It is instructive to give a simple example of how Maximum Likelihood is used in practise. We will show how to estimate the velocity **U** of the Galaxy relative to a sample of galaxies distributed around the sky. It will be supposed that we have redshift independent distance estimates d_l for the galaxies in the sample. We can therefore estimate the radial component of the peculiar velocity of each galaxy, u_l, over and above the Hubble flow: $u_l = cz - Hd_l$ (cz is the observed recession velocity).

Suppose that in a sample of galaxies, galaxy l is observed in direction $\hat{\mathbf{r}}_l$ and that it is assigned a radial component of peculiar velocity u_l. Suppose further that the probable error in measuring u_l is σ_l (the error is a consequence of the uncertainty in the distance estimate).

The component of our velocity **U** relative to the sample in the direction of galaxy l is **U**, is $\hat{\mathbf{r}}.\mathbf{U}$. The velocity of galaxy l relative to the sample is therefore $u_l - \hat{\mathbf{r}}.\mathbf{U}$. Hence on the assumption that the errors are Gaussian, the likelihood of the entire data set is

$$\mathcal{L}(U_1, U_2, U_3) = \prod_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left[-\frac{(u_l - \hat{\mathbf{r}}_l.\mathbf{U})^2}{2\sigma_l^2}\right] \tag{1.55}$$

We wish to chose the components of **U** that maximize this. To this end we take the logarithm of this expression, thus turning the product into a sum, and then differentiate with respect to the components U_i of **U**. This gives

$$\mathbf{U} = \underline{\mathbf{A}}^{-1}.\sum_l \frac{u_l \hat{\mathbf{r}}_l}{\sigma_l^2}, \qquad A_{ij} = \sum_l \frac{\hat{r}_i \hat{r}_j}{\sigma_l^2} \tag{1.56}$$

The matrix $\underline{\mathbf{A}}$ contains only information about the directions in which the galaxies are observed and the errors in measuring a radial velocity.

Since it is harder to measure the distances of the furthest galaxies in the sample, the error σ_l increases with distance. The most distant galaxies in the sample therefore have the least weight. The error analysis is, however, very complicated in part because the vectors $\hat{r}_i$ are not in fact randomly distributed on the sky: there is a zone of avoidance to contend with, and we know the vectors are correlated since galaxies lie in clusters and the clusters themselves are correlated.

This is an important problem in cosmology because we would like to know what our motion is relative to the most distant systems of galaxies. We can obtain an alternative measure of this by observing the dipole anisotropy of the cosmic microwave background radiation. The two estimates should agree in magnitude and direction.

It is the inverse of $\underline{\mathbf{A}}$ which comes into the solution for **U**. In this example, $\underline{\mathbf{A}}$ is a 3x3 matrix when there are 3 velocity components to determine. However, for a problem in which there were 1000 parameters to determine, this would mean inverting a 1000-square matrix. In many cases, $\underline{\mathbf{A}}$ would be an almost diagonal matrix in which case there are special techniques to deal with that efficiently. In the general case special techniques are required, see for example Golub and van Loan (1996).

References

Aban, I.B., Meerschaert, M.M., and Panorska, A.K. 2006. Parameter Estimation for the Truncated Pareto Distribution. *Journal of the American Statistical Association*, **101**(473), 270–277.

Beg, M. A. 1983. Unbiased Estimators and Tests for Truncation and Scale Parameters. *American Journal of Mathematical and Management Sciences*, **3**, 251–274.

Crawford, D. F., Jauncey, D. L., and Murdoch, H. S. 1970. Maximum-Likelihood Estimation of the Slope from Number-Flux Counts of Radio Sources. ApJ, **162**(Nov.), 405–+.

Golub, G. H., and van Loan, C. F. 1996. *Matrix Computations*. 3rd edn. Johns Hopkins University Press.

Gradshteyn, I.S., and Ryzhik, I.M. 2007. *Table of Integrals, Series, and Products*. Academic Press; 7 edition.

Maschberger, T., and Kroupa, P. 2009. Estimators for the exponent and upper limit, and goodness-of-fit tests for (truncated) power-law distributions. MNRAS, **395**(May), 931–942.

Press, W.H., Teukolsky, S., Vetterling, W.T., and Flannery, B.P. 2007. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press; 3rd. edition.