MAXIMUM LIKELIHOOD METHOD I

Maximise the likelihood, $P(data \mid model + parameters)$, which if adopt Bayesian philosophy is the same as maximising the posterior probability, $P(model + parameters \mid data)$.

1. Use **all** the raw data if possible
2. Do minimal preprocessing eg. cosmic rays
3. Use MLM to devise optimal solution(s) to the problem
4. Tradeoff optimum method -v- practicalities eg. stacking images
5. Generate error estimates - Minimum Variance Bound (MVB)

$$P(parameters \mid data) \propto P(data \mid parameters)$$

$$L = P(x_1, x_2, x_3.....x_n) = \prod_{i=1}^{n} P(data_i \mid parameters)$$

the latter for independent data points. In practice maximise

$$ln(L) = \sum_{i=1}^{n} ln[P(data_i \mid parameters)]$$

either by direct space searches or by solving (non-linear) equations

$$\frac{\partial ln(L)}{\partial \theta_j} = 0 \; ; \quad j = 1, 2, 3.....m$$

MAXIMUM LIKELIHOOD METHOD II

Desirable properties of estimators

1. Consistency – $\hat{\theta} \rightarrow \theta_{true}$ as $n \rightarrow \infty$
2. Unbiased – $< \hat{\theta} >= \theta_{true}$
3. Robustness – to real noise and imperfect model
4. Information content – make maximal use of information

Fisher information

1. Information should increase with no. of relevant observations
2. Information conditional on what we want from data
3. Information should be related to precision

Information matrix

$$I(\theta)_{i,j} = \int \frac{\partial ln(L)}{\partial \theta_i} \frac{\partial ln(L)}{\partial \theta_j} L(\underline{x} \mid \underline{\theta}) \, d\underline{x}$$

$$= - \int \frac{\partial^2 ln(L)}{\partial \theta_i \partial \theta_j} L(\underline{x} \mid \underline{\theta}) \, d\underline{x}$$

Parameter covariance matrix

$$V(\theta)_{i,j} \geq I^{-1}(\theta)_{i,j}$$

MVB, Cramer–Rao bound

PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

- of all estimators MLE's are generally the ones with minimum parameter error
- in general if you use a MLE no other method can improve upon the estimate
- if the error or residual function is Gaussian MLE and non-linear least squares estimates are the same
- almost all data processing problems admit a ML solution that is relatively straightforward to implement numerically
- analysis of the parameter covariance matrix, or equivalently the Fisher information matrix, can indicate whether or not you have too many parameters, or have chosen the wrong parameters/model
- in model fitting problems always examine the residual function for systematic trends, if they are present improve the model

$$x_i \quad = \quad f_i(\theta) \quad + \quad \epsilon_i$$

data model residual or noise

Likelihood function

$$L = \prod_i P(x_i \mid \theta) = \prod_i P(\epsilon_i)$$

For independent Gaussian errors, ϵ_i,

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\, exp[-\epsilon_i^2/2\sigma_i^2]$$

and hence maximising $ln(L)$ becomes

$$ln(L) = const - \frac{1}{2}\sum_i [x_i - f_i(\theta)]^2/\sigma_i^2$$

and is the same as minimising

$$F(\theta) = \frac{1}{2}\sum_i [x_i - f_i(\theta)]^2/\sigma_i^2$$

Quadratic minimisation/maximisation problem which solve by either direct searches or by solving m non-linear equations

$$\frac{\partial F(\theta)}{\partial \theta_j} = 0 \; ; \quad j = 1, 2, 3.....m$$

Example – Optimal Spectral Extraction

1. pre-process 2D image to remove instrumental signature (bias, trim, flat-field, correct for slit response, "field" distortion)

2. locate spectrum and accurately track position as a function of wavelength – worry about wavelength dependent slit losses due to not observing at parallactic angle

3. estimate sky at all wavelengths along spectrum, "subtract" off and keep – useful to check wavelength calibration

4. optimally extract spectrum using variance weighted profile fitting method

5. primary wavelength calibration via arcs – beware position of object in slit causes wavelength shifts, as does spectrograph flexure

6. flux calibrate extracted spectrum via equivalent observations of spectrographic standards using wide and narrow (normal) slit

7. what is optimum slit width to use for target observations ?

Example – Image Parameter Errors

Consider an idealised photon event distribution $\{x_i, y_i\}$ from an image with normalised profile $\phi(x, y)$ ie. $\int \phi(x, y) dxdy = 1$ and total flux η observed in locally constant background flux density b.

$$L = \prod_{i=1}^{N} \eta \ \phi(x_i, y_i) + b$$

$$lnL = \sum_{i=1}^{N} ln[\eta \ \phi(x_i, y_i) + b]$$

For isolated images the MVB for intensity and position are given by

$$var\{\hat{\eta}\} = [\int\int \frac{\phi^2(x, y)}{\eta\phi(x, y) + b} \ dxdy]^{-1}$$

$$var\{\hat{\theta}_x\} = [\int\int \frac{\eta^2(\partial\phi/\partial\theta_x)^2}{\eta\phi(x, y) + b} \ dxdy]^{-1}$$

For example, for Gaussian images $I(r) = I_p exp(-r^2/2\sigma_G^2)$

faint images intensity MVB $\Rightarrow$ error $= \sqrt{4\pi\sigma_G^2} \ \sigma_{noise}$

bright images error $= \sqrt{2\pi\sigma_G^2 I_p}$

faint images position MVB $\Rightarrow$ error $= \sqrt{2/\pi} \ \sigma_{noise}/I_p$

bright images error $= 1/\sqrt{2\pi I_p}$

Extend to discrete pixels by replacing integrals with summations $\rightarrow$ predict loss of information.

Example – Optimally Combining Images

First align astrometrically apply intensity mapping of the form

$$new = raw \times scale + shift$$

to correct for varying sky levels or atmospheric transmission.

Consider a matched series of scaled values $x_i = \overline{x} + \epsilon_i$ with PDF

$$P(\epsilon_i) = \alpha.N(0, \sigma_i) + \beta.U(-a, a)$$

where $N(0, \sigma_i)$ represents the Gaussian core of the noise distribution, and the uniform distribution, $U(-a, a)$, represents the non-Gaussian extended tail. If the range $[-a, a]$ is large compared to the core size and the fraction of outlying points is low, $\beta << 1$, then to a good approximation this distribution is equivalent to

$$\begin{aligned} P(\epsilon_i) &= \frac{\alpha}{\sqrt{2\pi\sigma_i^2}}.exp(\frac{-\epsilon_i^2}{2\sigma_i^2}), \quad |\epsilon_i| \leq k\sigma_i \\ &= \beta, \quad k\sigma_i < |\epsilon_i| \leq a \\ &= 0, \quad |\epsilon_i| > a \end{aligned} \tag{1}$$

or could have postulated this form for the PDF at the beginning. Assuming independent measurements, x_i, minimising the log-likelihood function then leads to the following estimator for $\hat{x}$,

$$\hat{x} = \frac{\sum_{i=1}^{m'} x_i/\sigma_i^2}{\sum_{i=1}^{m'} 1/\sigma_i^2}$$

where m' denotes the observations within the k-sigma clipped range. Equation solved iteratively as clipping boundary function of current estimates of $\hat{x}$ (and $\hat{\sigma}^2$), $\Rightarrow$ k-sigma clipping $\equiv$ MLE for this PDF form.

MAXIMUM LIKELIHOOD & C-STATISTIC

The C-statistic deals with low count levels and generalises the χ^2 method for situations where event rate per "cell" is 0,1,2.....

Observe n_i events per cell and model predicts m_i

$$P(n_i \mid m_i) = e^{-m_i} \ \frac{(m_i)^{n_i}}{n_i!}$$

Series of N independent cells covering range of model prediction – then likelihood of observations is

$$L = \prod_i \ P(n_i \mid m_i)$$

$$ln(L) = \sum_i -m_i + n_i \ ln(m_i) - ln(n_i!)$$

The last term on the RHS is a constant hence maximising the likelihood is identical to maximising

$$\sum_i -m_i + n_i \ ln(m_i) = C - statistic$$

PRESS & SCHECTER FORMALISM – 1976 ApJ 203 p557

.....or model fitting without binning the data + null results

Example: redshift & column density distribution of LLS – model

$$f(N,z) = k\ N^{-\beta}\ (1+z)^{\gamma}$$

Imagine partioning observable N, z plane into "cells".

Let expected number of observed data points in cell i be ϕ_i

$$\phi_i = f(N,z)_i\ \delta V$$

Probability of observing x_i points in cell i

$$P(x_i) = e^{-\phi_i}\ \frac{\phi_i^{x_i}}{x_i\ !}$$

Let $\delta V \to 0$ then $x_i = 1$ LLS detected $x_i = 0$ none detected.

Therefore the likelihood function for QSO$_j$ is

$$L_j = \prod_i P(x_i) = \prod_i e^{-\phi_i}\ \frac{\phi_i^{x_i}}{x_i\ !}$$

$$= \prod_i e^{-\phi_i} \quad \cdot \quad \prod_i \phi_i\ e^{-\phi_i}$$

empty cells detected cells

$$ln(L_j) = \sum_i -\phi_i \quad + \quad \sum_{i=1}^{m}\ ln(\phi_i)$$

all cells detected

$$ln(L_j) = -\int_{N_j}\int_{z_j} f(N,z)\ dNdz + \sum_{i=1}^{m}\ ln(f(N,z)_i$$

STRUCTURAL ANALYSIS

.....or curve fitting with errors on both variables

Model $Y_i = f(X_i \mid \theta)$

Observe $y_i = Y_i + \epsilon_i$ independent errors

and $x_i = X_i + \delta_i$ with variance $\sigma^2_{\epsilon_i}$ $\sigma^2_{\delta_i}$

Likelihood $L = \Pi_{i=1}^{N} P(x_i, y_i \mid \theta)$

$$ln(L) = -N\ ln(2\pi) - \frac{N}{2}\ \sum_i\ ln(\sigma^2_{\delta_i}.\sigma^2_{\epsilon_i}) - \frac{1}{2}\sum_i \left(\frac{\delta_i^2}{\sigma^2_{\delta_i}} + \frac{\epsilon_i^2}{\sigma^2_{\epsilon_i}}\right)$$

For unknown errors the problem is insoluble – for known errors

$$ln(L) = const - \frac{1}{2}\sum_i \frac{[x_i - X_i]^2}{\sigma^2_{\delta_i}} + \frac{[y_i - Y_i]^2}{\sigma^2_{\epsilon_i}}$$

and the solution effectively solves for X_i and θ_j ie. $N + m$ unknowns

$$\frac{\partial ln(L)}{\partial X_i} = \frac{(x_i - X_i)}{\sigma^2_{\delta_i}} + \frac{\partial f}{\partial X_i}\ \left[\frac{y_i - f(X_i \mid \theta)}{\sigma^2_{\epsilon_i}}\right] = 0$$

$$\frac{\partial ln(L)}{\partial \theta_j} = \sum_i\ \frac{\partial f}{\partial \theta_j}\ \left[\frac{y_i - f(X_i \mid \theta)}{\sigma^2_{\epsilon_i}}\right] = 0$$

(see Hodgkin et al. MNRAS 2009 for an example application of this)

NUMERICAL CONSIDERATIONS

"Numerical Recipes" – Press, Flannery, Teukolsky, Vettering

"Practical Optimisation" – Gill, Murray & Wright

Solve m non-linear partial differential equations or maximise/**minimise** log-likelihood function based on a Taylor expansion

$$F(\theta + \alpha P) = F(\theta) + \alpha P^T g(\theta) + \frac{1}{2}\alpha^2 P^T G(\theta) P + ...$$

$g(\theta)$ is gradient vector, $G(\theta)$ is the Hessian matrix, θ is current parameter vector and αP the update vector.

Descent condition is obviously $F(\theta + \alpha P) < F(\theta)$

Close to the solution $\hat{\theta}$

$$F(\hat{\theta} + \Delta\theta) = F(\hat{\theta}) + \frac{1}{2}\Delta\theta^T G(\hat{\theta})\Delta\theta$$

hence properties of $F(\hat{\theta})$ depend on eigenvalues of Hessian matrix.

1. Algorithmically test for converge of θ the current estimate
2. Compute a search direction P
3. Compute a step length α = univariate minimisation
4. Set $\theta = \theta + \alpha P$ and repeat from step 1.

a. DIRECT SEARCH METHODS

Construct grid of parameter points and calculate $F(\theta)$, look for minimum. Iterate to finer grid if necessary.

Can be awkward and inefficient in m-D problems when m is large.

b. STEEPEST DESCENT METHODS (+ conjugate gradient)

uses 0th and 1st derivative information

$$\theta \quad \rightarrow \quad \theta - \alpha g(\theta)$$

α step length along gradient vector for minimum of $F(\theta)$, repeat. If converges gives optimal solution, can be efficient but problem with linear convergence (slow) and saddle points (may never converge).

c. VARIANTS ON GAUSS-NEWTON METHODS

0th, 1st and 2nd derivative information derive local quadratic model

$$F(\theta + \alpha P) \approx F(\theta) + \alpha P^T g(\theta) + \frac{1}{2}\alpha^2 P^T G(\theta) P$$

finds local minimum with $GP = -g$, solve to give search direction – may need to compute α rather than assume unity, since then allows for perturbations from local quadratic model.

Problems, have to compute Hessian (and gradient), note that for LS problems can approximate well using only 1st derivative information.