Outline

1. Summary of Linear Least Squares Problem
2. Singular Value Decomposition
3. Non-Linear Least-Squares Problems
4. Levenberg-Marquart Algorithm
5. Errors with Poisson Distribution
6. Downhill Simplex Method
7. Simulated Annealing
8. Genetic Algorithms

Summary of Linear Least Squares Problem

Linear Models

- model y given by linear combination of M functions of x
- general form

$$y(x) = \sum_{k=1}^{M} a_k X_k(x)$$

- $X_1(x), \ldots, X_M(x)$ arbitrary (non-linear!) fixed functions of x
- minimize

$$\chi^2 = \sum_{i=1}^{N} \frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i^2}$$

- *design matrix*

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i}$$

Solution for Linear Models

- A typically has more rows than columns ($N > M$)
- vector $\vec{b}$ of length N:

$$b_i = \frac{y_i}{\sigma_i}$$

- minimum of χ^2 where all M partial derivatives with respect to parameters are zero leads to *normal equations*

$$(A^T A)\vec{a} = A^T \vec{b}$$

- inverse matrix of positive definite matrix $A^T A$

$$C = (A^T A)^{-1}$$

- errors in parameters then given by

$$\sigma^2(a_j) = C_{jj}$$

- off-diagonal elements C_{jk} are covariances between a_j and a_k

Singular Value Decomposition

Singular Fitting Problems

- normal equations often very close to singular
- in Gauss elimination, zero or very small pivot element $\Rightarrow$ large values for a_k that largely cancel each other in fitted function
- observations often do not clearly distinguish between two or more basis functions
- if two functions, or two different combinations of functions, fit data about equally well, design matrix A becomes singular
- least-squares problems are both
  - overdetermined (number of data points greater than number of parameters)
  - underdetermined (ambiguous combinations of parameters exist)
- complicated problems $\Rightarrow$ extremely hard to notice ambiguities a priori

Minimization Problem

- overdetermined system ⇒ SVD produces solution that is best approximation in least-squares sense
- underdetermined system ⇒ SVD produces solution where a_k are smallest in least-squares sense
- combination of basis functions is irrelevant to fit ⇒ corresponding combination of basis functions will be driven down to small value, rather than pushed up to delicately canceling infinities
- using design matrix A and vector $\vec{b}$ minimization of χ^2 can be written as:

$$\text{find } \vec{a} \text{ that minimizes } \chi^2 = |A\vec{a} - \vec{b}|^2$$

Singular Value Decomposition

- any $M \times N$ matrix A with $M \geq N$ can be written as product of $M \times N$ column-orthogonal matrix U, $N \times N$ diagonal matrix W with positive or zero elements (*singular values*), and transpose of $N \times N$ orthogonal matrix V

$$\begin{pmatrix} & & \\ & & \\ & \mathbf{A} & \\ & & \\ & & \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & \mathbf{U} & \\ & & \\ & & \end{pmatrix} \cdot \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \cdots & \\ & & \cdots & \\ & & & w_N \end{pmatrix} \cdot \begin{pmatrix} & & \\ & \mathbf{V}^T & \\ & & \end{pmatrix}$$

- orthonormal columns of U, V: for all $1 \leq k \leq N$, $1 \leq n \leq N$

$$\sum_{i=1}^{M} U_{ik} U_{in} = \delta_{kn} \qquad \sum_{j=1}^{N} V_{jk} V_{jn} = \delta_{kn}$$

- in matrix form: $U^T U = V^T V = V V^T = 1$
- problem with infinite number of solutions $\vec{x} \Rightarrow$ SVD returns solution with smallest $|\vec{x}|^2$

SVD Applied to Fitting

- U, V, w_i from SVD of A
- vectors $U_i, i = 1, ..., M$ are columns of U (vector of length N)
- vectors $V_i; i = 1, ..., M$ are columns of V (vector of length M)
- solution of least-squares problem

$$\vec{a} = \sum_{i-1}^{M} \left(\frac{\vec{U}_i \cdot \vec{b}}{w_i} \right) \vec{V}_i$$

- fitted parameters $\vec{a}$ are linear combinations of columns of V, coefficients obtained from scalar products of columns of U with weighted data vector

SVD and Errors

- errors in fitted parameters also linear combinations of columns of V
- standard deviations are all mutually independent (uncorrelated)
- vectors $\vec{V}_i$ are principal axes of error ellipsoid of fitted parameters $\vec{a}$
- variance in estimate of parameter a_j:

$$\sigma^2(a_j) = \sum_{i=1}^{M} \left(\frac{V_{ji}}{w_i} \right)^2$$

- covariances given by

$$\mathrm{Cov}(a_j, a_k) = \sum_{i=1}^{M} \left(\frac{V_{ji} V_{ki}}{w_i^2} \right)$$

SVD Avoids Singularities

- SVD overcomes singularities: if $w_i = 0$, $\frac{1}{w_i}$ should be set to zero
- adds zero multiple of any linear combination of basis functions that are degenerate in the fit
- if singular value w_i is nonzero but very small, set its reciprocal to zero, asit is probably an artifact of roundoff error
- typically: remove all singular values whose ratio to the largest singular value is less than N times the machine precision
- also: SVD identifies linear combinations of variables that just happen not to contribute much to reducing the χ^2
- can sometimes reduce the probable error on coefficients quite significantly, while increasing minimum χ^2 only negligibly
- always use SVD: great advantage, that it (theoretically) cannot fail, more than makes up for speed disadvantage

Multidimensional Models

- model y is function of vector $\vec{x} \Rightarrow$ basis functions will be functions of a vector $X_1(\vec{x}), ..., X_M(\vec{x})$
- χ^2 merit function becomes

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(\vec{x})}{\sigma_i} \right]^2$$

- repeat same procedure as before with x replaced by $\vec{x}$
- x_i only used to calculate values of basis functions at $\vec{x}_i$

Non-Linear Models

- model with parameter a that enters model y_m and χ_i non-linearly $\Rightarrow$ cannot execute summation without having (an estimate for) value of a
- example: $y = \sin(ax)$ with partial derivative

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{[y_i - \sin(ax_i)]x_i \cos(ax_i)}{\sigma_i{}^2}$$

- no summation without value for a
- model $y(x, \vec{a})$ that is non-linear in $\vec{a}$ can be fitted to a set of data only iteratively
- need a first set of values for $\vec{a}$, and then find successive improvements of these values
- first a one-dimensional case with single parameter a, $\chi^2 = \chi^2(a)$
- find a such that $\chi^2(a)$ is minimized

Taylor Approximations

- far from minimum ⇒ use derivative $\partial\chi^2/\partial a$ to decide in which direction to look for the improved value:

$$a_{n+1} = a_n - K\frac{\partial\chi^2}{\partial a}$$

- K is a constant
- close to minimum of χ^2 first derivative approaches zero
- close to minimum approximate χ^2 as quadratic function of a: $\chi^2(a) = p + q(a - a_{min})^2$, where p is the minimum value of χ^2, reached at $a = a_{min}$
- combine derivatives $\partial\chi^2/\partial a = 2q(a - a_{min})$, $\partial^2\chi^2\partial a^2 = 2q$

$$a - a_{min} = \frac{\partial\chi^2/\partial a}{\partial^2\chi^2/\partial a^2} \Rightarrow a_{n+1} = a_n - \frac{\partial\chi^2/\partial a}{\partial^2\chi^2/\partial a^2}$$

Multi-Dimensional Non-Linear Problem

- more-dimensional case (more than one parameter)

$$\chi^2(\vec{a}) \simeq p - \vec{q}\cdot\vec{a} + \frac{1}{2}\vec{a}\cdot\vec{\vec{D}}\cdot\vec{a}$$

- model $y_m = y_m(x, \vec{a})$:

$$q_k \equiv \frac{\partial\chi^2}{\partial a_k} = -2\sum_{i=1}^{N}\frac{[y_i - y_m]}{{\sigma_i}^2}\frac{\partial y_m}{\partial a_k} \equiv -2\beta_k$$

- and

$$D_{kl} \equiv \frac{\partial\chi^2}{\partial a_k\partial a_l} = 2\sum_{i=1}^{N}\frac{1}{{\sigma_i}^2}\left[\frac{\partial y_m}{\partial a_k}\frac{\partial y_m}{\partial a_l} - [y_i - y_m]\frac{\partial^2 y_m}{\partial a_k\partial a_l}\right] \equiv 2\alpha_{kl}$$

(continued)

- from before

$$D_{kl} \equiv \frac{\partial \chi^2}{\partial a_k \partial a_l} = 2 \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} \left[\frac{\partial y_m}{\partial a_k} \frac{\partial y_m}{\partial a_l} - [y_i - y_m] \frac{\partial^2 y_m}{\partial a_k \partial a_l} \right] \equiv 2\alpha_{kl}$$

- second term on RHS small with respect to first one because $y_i - y_m$ will almost equally often be positive as negative $\Rightarrow$ subsequent terms in summation will almost cancel
- dropping second term reduces computing time and makes iteration to best solution more stable
- rewrite first and second-order approximations:

$$\beta_k = \lambda \alpha_{kk} \delta a_k \qquad \beta_k = \sum_{l=1}^{M} \alpha_{kl} \delta a_l$$

- proportionality constant K is scaled with second derivative, where λ is constant

Levenberg-Marquardt

- from before:

$$\beta_k = \lambda \alpha_{kk} \delta a_k \qquad \beta_k = \sum_{l=1}^{M} \alpha_{kl} \delta a_l$$

- Levenberg-Marquardt method adds two equations:

$$\beta_k = \sum_{l=1}^{M} \alpha'_{kl} \delta a_l \qquad \text{where}$$

$$\alpha'_{kl} = \alpha_{kl} \quad (\text{if } k \neq l);$$

$$\alpha'_{kl} = \alpha_{kl}(1+\lambda) \quad (\text{if } k = l)$$

- approaches linear descent for large λ
- approaches quadratic approximation for small λ

Levenberg-Marquardt Algorithm

1. pick initial solution
2. pick small value for λ (i.e. one hopes that the solution is already close enough for the quadratic approximation)
3. compute χ^2 for initial solution,
4. compute new value for $\vec{a}$
5. if χ^2 for the new solution is smaller (larger) than for the old one, then the quadratic approach does (doesn't) work, and one should decrease (increase) λ to get closer to the purely quadratic (linear) method
6. iterated until minimum χ^2 is found

Parameter Errors

- errors on best-fit parameters $\vec{a}$ from general relation (near minimum)

$$\delta\chi^2 = \delta\vec{a} \cdot \vec{\vec{\alpha}} \cdot \delta\vec{a}$$

- if best-fit parameters are not correlated $\Rightarrow$
  - matrix α close to diagonal
  - off-diagonal elements α_{kl} are much smaller than elements α_{kk} on diagonal
- inverse matrix C almost given by $C_{kk} = 1/\alpha_{kk}$
- distance δa_k to best value a_k

$$\delta {a_k}^2 = \frac{\Delta\chi^2}{\alpha_{kk}}$$

- to estimate a 1-*sigma* error in a_k we enter $\Delta\chi^2 = 1$
- correltaed errors: matrix C gives correlation between deviations of parameters from best-fit values

Errors with Poisson Distribution

Overview

- counting of small number of events ⇒ often Poisson distribution of measurements
- least-squares methods do not apply
- here: *maximum-likelihood method for errors with a Poissonian distribution*
- in literature often referred to as *the maximum-likelihood method* where Poissonian error distribution is implied
- abbreviated name misleading as least-squares method is also maximum-likelihood method for errors distributed as Gaussians

Maximum-Likelihood Method

- consider number of counts on photon-counting detector
- number of counts detected in location i is n_i
- number of counts predicted for location by model as m_i
- probability at one location to obtain n_i photons when m_i are predicted is

$$P_i = \frac{{m_i}^{n_i} e^{-m_i}}{n_i!}$$

- if values of m_i (and n_i) are large, probability may be approximated with Gaussian and use least-squares method
- typically value is 20 based on assumption that difference between Poisson and Gaussian distributions for large μ is less important than uncertainties due to systematic effects in measurements
- assumption should be verified in each case

Maximum-Likelihood Method (continued)

- small values of m_i (and n_i): use Possion distribution

$$P_i = \frac{{m_i}^{n_i} e^{-m_i}}{n_i!}$$

- to maximize overall probability, maximize

$$L' \equiv \prod_i P_i$$

- easier to maximize logarithm

$$\ln L' \equiv \sum_i \ln P_i = \sum_i n_i \ln m_i - \sum_i m_i - \sum_i \ln n_i!$$

- last term independent of model ⇒ consider as constant

Maximum-Likelihood Method (continued)

- from before

$$\ln L' \equiv \sum_i \ln P_i = \sum_i n_i \ln m_i - \sum_i m_i - \sum_i \ln n_i!$$

- maximizing L' equivalent to minimizing

$$\ln L \equiv -2\left(\sum_i n_i \ln m_i - \sum_i m_i\right)$$

- compare two models A and B with number of fitted parameters n_A, n_B and likelihoods $\ln L_A$, $\ln L_B$, difference $\Delta L \equiv \ln L_A - \ln L_B$ is χ^2 distribution with $n_A - n_B$ degrees of freedom, for sufficient number of photons

Example 1: Constant Background

- constant background: $m_i = A$
- Z pixels in total ⇒ total number of photons in model is $N_m = ZA$
- total observed number of photons is N_o
- likelihood

$$\ln L \equiv -2\left(\sum_i n_i \ln m_i - \sum_i m_i\right)$$

- minimize with respect to N_m

$$-0.5 \ln L = \sum n_i \ln A - \sum A = N_o \ln A - ZA = N_o \ln(N_m/Z) - N_m$$

- therefore

$$\frac{\partial \ln L}{\partial N_m} = 0 \Rightarrow \frac{N_o}{N_m} - 1 = 0 \Rightarrow N_o = N_m$$

- best solution has equal number of photons in model and observation

Example 2: Constant Background plus One Source

- source with strength B and fraction f_i lands on pixel i
- likelihood

$$\ln L \equiv -2\left(\sum_i n_i \ln m_i - \sum_i m_i\right)$$

- here

$$-0.5 \ln L = \sum_i n_i \ln(A + Bf_i) - \sum_i (A + Bf_i)$$

- find minimum of L for variations in A and B:

$$\frac{\partial \ln L}{\partial A} = 0 \Rightarrow \sum_i \frac{n_i}{A + Bf_i} - \sum_i (1) = \sum_i \frac{n_i}{A + Bf_i} - Z = 0$$

$$\frac{\partial \ln L}{\partial B} = 0 \Rightarrow \sum_i \frac{n_i f_i}{A + Bf_i} - \sum_i f_i = \sum_i \frac{n_i f_i}{A + Bf_i} - 1 = 0$$

Example 2: Constant Background plus One Source (continued)

- two equations for two unknowns A and B

$$\sum_i \frac{n_i}{A+Bf_i} - \sum_i (1) = \sum_i \frac{n_i}{A+Bf_i} - Z = 0$$

$$\sum_i \frac{n_i f_i}{A+Bf_i} - \sum_i f_i = \sum_i \frac{n_i f_i}{A+Bf_i} - 1 = 0$$

- multiply first equation with A, second with B, and add:

$$\sum_i n_i = AZ + B$$

- total number of counts in best model is equal to total number of observed counts
- may used this to fit only one parameter

General Methods

Overview

- several established methods to find best fit for Poissonian error distributions
- methods shown here do not use derivative of function but only function value itself and minimization criterion
- methods are equally suited for
  - least-squares problems with χ^2 as criterion
  - maximum-likelihood problems with Poisson statistics
- tend to be slower than Levenberg-Marquardt method for simple problems
- tend to be more efficient than Levenberg-Marquardt for
  - many variables $\Rightarrow$ matrix α is very big
  - problems in which the χ^2 distribution has many local minima

Downhill Simplex Method

- Numerical Recipes (2nd ed.) chapter 10.4
- often best method for models with small computational burden
- *simplex* is geometrical figure consisting, in N dimensions, of $N + 1$ points (or vertices) and all interconnecting line segments, polygonal faces, etc.
- 2D: triangle, 3D: (irregular) tetrahedron
- non-degenerate simplexes enclose finite inner N-dimensional volume
- one point as origin, other N points define vector directions in N-D vector space

Downhill Simplex Method (continued)

- start with $N + 1$ points in N-D hyperspace, defining initial simplex
- one point is initial starting point P_0
- other N points are

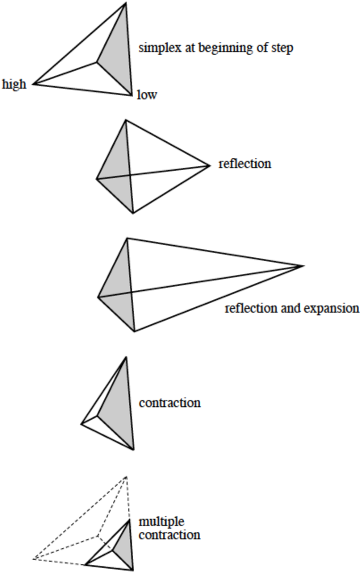
$$P_i = P_0 + \lambda e_i$$

- e_i are N unit vectors
- λ a constant which is your guess of problem's characteristic length scale

Downhill Simplex Method (continued)

- downhill simplex method is series of steps:
- point of simplex where function is largest through opposite face of simplex to lower point
- these reflections are constructed to conserve volume of simplex (maintains nondegeneracy)
- when possible, method expands simplex in one or another direction to take larger steps
- when it reaches valley floor, method contracts itself in transverse direction and tries to ooze down the valley
- when simplex passes through eye of needle, it contracts itself in all directions, pulling itself in around its lowest (best) point

Downhill Simplex Method



- termination criteria can be delicate
- terminate when vector distance moved is fractionally smaller in magnitude than some tolerance
- could also require that decrease in function value in terminating step be fractionally smaller than some tolerance
- might be fooled by single anomalous step that failed to get anywhere
- good idea to restart at point where minimum was found

Simulated Annealing

- Numerical Recipes (2nd ed.) chapter 10.9
- suitable for large-scale problems and global minimum hidden among many, poorer, local minima
- also works well when functions are in discrete space where derivatives are not available
- analogy with thermodynamics: liquids freeze and crystallize, metals cool and anneal:
  - at high temperatures, molecules of liquid move freely with respect to one another
  - if liquid is cooled slowly, thermal mobility is lost; atoms are able to line themselves up and form pure crystal that is completely ordered over distance up to billions of times the size of individual atom in all directions
  - crystal is state of minimum energy for system
  - slowly cooled systems: nature finds minimum energy state

Simulated Annealing (continued)

- if liquid metal is cooled quickly or quenched, it does not reach this state but rather ends up in a polycrystalline or amorphous state having somewhat higher energy
- essence is slow cooling, allowing ample time for redistribution of atoms as they lose mobility
- this is technical definition of annealing; essential for ensuring that a low energy state will be achieved
- analogy not perfect, but provides sense in which all other minimization algorithms correspond to rapid cooling or quenching
- gone greedily for the quick, nearby solution: from the starting point, go immediately downhill as far as you can go
- often leads to local but not necessarily global minimum

Simulated Annealing (continued)

- Boltzmann probability distribution expresses the idea that a system in thermal equilibrium at temperature T has its energy probabilistically distributed among all different energy states E

$$Prob(E) \sim e^{\frac{-E}{kT}}$$

- Even at low temperature, there is a chance, albeit very small, of a system being in a high energy state
- Therefore, there is a corresponding chance for the system to get out of a local energy minimum in favor of finding a better, more global, one.
- quantity k (Boltzmann's constant) relates temperature to energy
- system sometimes goes uphill as well as downhill; but the lower the temperature, the less likely is any significant uphill excursion

Simulated Annealing

- first application by Metropolis et al. (1953)
- simulated thermodynamic system changes configuration from energy E_1 to energy E_2 with probability $p = e^{-\frac{E_2 - E_1}{kT}}$
- if $E_2 < E_1$ probability is greater than unity $\Rightarrow$ change get $p = 1$ (always takes this option)
- general scheme of always taking a downhill step while sometimes taking an uphill step is known as *Metropolis algorithm*

Simulated Annealing (continued)

- to use Metropolis algorithm need following elements:
  1. description of possible system configurations
  2. generator of random changes in the configuration that are presented to system
  3. objective function E (analog of energy) whose minimization is goal of procedure
  4. control parameter T (analog of temperature) and an annealing schedule which tells how it is lowered from high to low values, e.g., after how many random changes in configuration is each downward step in T taken, and how large is that step
- meaning of high and low T and assignment of schedule may require physical insight and/or trial-and-error experiments

Genetic Algorithms

based on Charbonneau (1995, ApJ Supl.Ser. 101, 309-334)

1. construct random initial population and evaluate fitness of members
2. construct a new population by breeding selected individuals from population
3. evaluate fitness of each member of population
4. replace old population with new population
5. test convergence: if best member does not (yet) match required criteria, go to step 2

- breeding includes combination and mutation