Geophysical Inversion

Vector Space:

It is a mathematical structure formed by a collection of elements called 'vector'. Vector may be added and multiplied by scales.

For example, $\vec{F} = \vec{x} + \vec{y}$

Where, F, x, y are the vectors.

There are also vector spaces with scalar multiplication by complex number, rational number.

From the linear algebra point of view, vector spaces are characterised by their dimension. Dimension is defined as the number of independent directions in spaces.

Example:

If we record one seismic trace, one second in length at a sample rate of 1000 samples per second, and let each sample be defined by one byte, then we can put these 1000 bytes of information in 1000

$(s_1, s_2, s_3, ..., s_{1000})$

s_i is the i^{th} sample. Consider that it is a 3- component physical vector. While stacking seismic traces, we just add these n-dimensional vectors component, say,

$s+t = (s_1+t_1, s_2+t_2, ..., s_{1000}+t_{1000})$

Note: Mathematical definition of vector space is sufficiently general to incorporate objects like, functions, matrices.

Hilbert space:

Hilbert space is names after David Hilbert. A Hilbert space is an abstract vector space possessing the structure of an inner product that allows length and angle to be measured. One of the most familiar examples of Hilbert space is the Euclidean space consisting of three- dimensional vectors, defined by R^3, equipped with the dot product.

Example:

The dot product takes two vectors x & y and produces a real number x .y

In Cartesian coordinates, the dot product is defined by

$(x_1 , x_2 , x_3) . (y_1, y_2, y_3) = x_1y_1 + x_2y_2 + x_3y_3$

The dot product satisfy the following properties

(1). It is symmetric in x & y (i.e. x . y = y . x)

(2). It is linear in its first arguments (i.e. $(ax_1 + bx_2) y = ax_1. y + bx_2. y$; where a, b are the scalars and x_1 , y_1 are the vectors)

(3). It is positive defined for all vectors x ; $x \geq 0$ with equality if and only if x=0.

Note:

Dot product satisfies these three properties is known as a inner product. Every finite – dimensional inner product space is also a Hilbert space.

In order to define the relative 'size' of vectors matrices a generalized concept of length is introduced which is called norm.

Norm:

A norm is a function form the space of vectors onto the scalar, defined by $\|\bullet\|$ satisfying the following properties for any two vectors

$\vec{u}$ & $\vec{v}$ and any scalar α

(1).$\| v\| > 0$ for any $v \neq 0$ and $| v |=0 \leftrightarrow v=0$

(2). $\| \alpha v \| = | \alpha | \| v \|$

(3). $\|v +u \| \leq \|v\| + \|u\|$ [Triangle property]

The most useful class of norms for vectors in R^n is the l_p norm defined for $p \geq 1$

$$\|x\|_{l_p} = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

For p= 2 this is just the ordinary Euclidean norm: $\|x\|_2 = \sqrt{x^T x}$

l_p norm exists as p→ ∞ called l_∞ norm

$\|x\|_{l_p} = \max|x_i|$; 1≤ i ≤ n

Note:

A matrix norm that is not induced by any vector norm is the Frobenius norm defined for all A ε R^{n*m}

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2}$$

<u>Interpretation:</u>

Which p is best for optimization????

It is observed that p values near 1 are more stable than p- values near 2. In inversion, if our data have say, Gaussian distribution, the l_2 is optimal. If our data have double- exponential distribution, the l_p is optimal.

Figure

$$\rho_p(x) = \frac{p^{1-1/p}}{2\sigma_p \Gamma(1/p)} \exp\left(\frac{-1|x - x_0|^p}{p(\sigma_p)^p} \right)$$

Γ = Gamma function

ρ(x) = Probability density

σ_p = Dispersion

$$(\sigma_p)^p \equiv \int_{-\infty}^{\infty} |x - x_0|^p \rho(x) dx$$

<u>Dimension:</u>

The dimension of vector space V is the cardinality (ie. number of vector / size of the sets) of a basis of V.

Basis is a set of linearly independent vector which can represent every vector in a given space / (or in coordinate system)

Note:

In physics / mathematics the dimension of a space is defined as the minimum number of co- ordinate needed to specify any point without Gaussian distribution.

$$P(d) = \frac{1}{\sqrt{2\pi}} \sigma \exp\left[- \frac{(d - \langle d \rangle)^2}{2\sigma^2} \right]$$

d = variable and <d> is the mean variable.

Matrices:

A matrix is a rectangular array of numbers, symbols, or expression, arranged in rows and columns, the individual items in a matrix are called its elements or entries.

$$A = \begin{bmatrix} 2 & 5 \\ 3 & 8 \\ 1 & 0 \end{bmatrix}$$

The components are denoted by A_{ij}. The transpose of a matrix, denoted by A^T

$$A^T = \begin{bmatrix} 2 & 3 & 1 \\ 5 & 8 & 0 \end{bmatrix}$$

Prove that $(AB)^T = B^T A^T$, $AB \neq BA$ for non-square matrices

Symmetric matrix:

A matrix which equals its transpose i.e. $A^T = A$ is said to be symmetric.

Skew – symmetric:

If $A^T = -A$, the matrix is said to be skew- symmetric.

Split / Partition:

Any square matrix A can be portioned into a sum of a symmetric and a skew-symmetric part via

$$A = \frac{1}{2}\left(A + A^T\right) + \frac{1}{2}\left(A - A^T\right)$$

Hermitian matrix:

Hermitian matrix is a square matrix with complex entries such that

$A_{ij} = A_{ji}$

$$A = \begin{bmatrix} 3+i & 5 & -2i \\ 2-2i & i & -7-13i \end{bmatrix}$$

$$A^T = \begin{bmatrix} 3+2i & 2-2i \\ 5 & i \\ -2i & -7-13i \end{bmatrix}$$

$$A^* or A^H = \begin{bmatrix} 3-i & 2+2i \\ 5 & -i \\ 2i & -7+13i \end{bmatrix}$$

Note:

Hermitian: $A = A^*$

Skew- Hermitian: $A = -A^*$

Normal: $A^*A = AA^*$

Unitary: $A^* = A^{-1}$

Orthogonal matrix:

$A^T = A^{-1}$

i.e. $A^TA = AA^T = I$

I = identity matrix

Diagonal matrix:

If all entries outside the main diagonal are zero, A is called diagonal matrix.

$$A=\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

Square matrix:

Square matrix is matrix with same number of rows and columns.

Lower triangular matrix:

Entries above the main diagonal are zero.

$$A=\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Upper triangular matrix:

Entries below the main diagonal are zero.

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

Invertible matrix:

A square matrix a is called invertible or non- singular if there exists a matrix B such that $AB = BA = I_n$

If B exists, it is unique and is called the inverse matrix of A, denoted by A^{-1}.

Note:

A matrix is invertible if and only if its determinant is non- zero.

$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

Rank of a matrix:

The rank of a matrix A is the maximum number of linearly independent row vectors of the matrix, which same as the maximum number of linearly in dependent column vectors.

Example:

$$A = -\begin{bmatrix} 1 & 2 & 2 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}$$

Rank = 2

First two rows are linearly independent. So that the rank is at least 2 but all three rows are linearly dependent. (First is equal to the sum of the second & rank must be less than 3).

$$A = \begin{bmatrix} 1 & 1 & 0 & 2 \\ -1 & -1 & 0 & -2 \end{bmatrix}$$ Has rank 1.

Eigen values and Eigenvectors:

A number λ and a non-zero vector x satisfying

$A x = \lambda x$

Where number λ must be chosen such that $(A - \lambda I)$ has a null space so that

$Det(A - \lambda I) = 0$ [we must choose x so that it lies in that null space]

This determinant is a polynomial in λ, called the characteristics polynomial.

Example:

$$A = \begin{bmatrix} 5 & 3 \\ 4 & 5 \end{bmatrix}$$

Characteristics polynomial is

$\lambda^2 - 10\lambda + 13 = 0$

Where roots are $\lambda = (5 + 2\sqrt{3})$ & $(5-2\sqrt{3})$

This leads to solve two homogeneous systems

$$\begin{bmatrix} 2\sqrt{3} & 3 \\ 4 & 2\sqrt{3} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -2\sqrt{3} & 3 \\ 4 & -2\sqrt{3} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

From which we arrive at two Eigen vectors,

$$\begin{bmatrix} \sqrt{3}/2 \\ 1 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ 1 \end{bmatrix}$$

Moore – Penrose Matrix Inverse:

Given an m by n matrix A, the Moore – Penrose generalized matrix inverse is a uniquely n by m matrix pseudo inverse A^+. This matrix was independently defined by Moore in 1920 and Penrose (1955), and variously known as the generalized inverse, pseudo inverse, or Moore- Penrose.

The Moore- Penrose inverse satisfies

(1). $AA^+A = A$

(2). $A^+AA^+ = A^+$

(3). $(AA^+)^H = AA^+$

(4). $(A^+A)^H = A^+A$

Where A^H is the conjugate transpose.

It is also true that

$$\vec{y} = A^+b$$

Is the shortest least square solution to the problem Ay = b ..... (6)

If the inverse of $A^+ = (A^HA)^{-1}A^H$

As we can seen by premultiplying both sides of (6) by A^H to create a square matrix which can then be interval?

$A^HAy = A^H b$

Giving $y = (A^HA)^{-1} A^H b$

$= A^+ b$

Orthogonal decomposition of a real symmetric matrix:

A real symmetric matrix G can be factored into

$$G = UQU^T$$

Where orthogonal eigenvectors in U and real Eigen values in Q.

Note:

For dimensional reasons there is clearly no hope of the kind of eigenvector decomposition discussed above being applied to rectangular matrix. If we allow different orthogonal matrix on each side of G.

Geophysical Inversion:

Introduction:

In the geophysical and the related science, experiments are performed under controlled conditions (ie. in a systematic manner), the outcome may be numerical values: that represent our observations at fixed (predetermined intervals) say, so the observation of the some physical properties of the physical world are commonly referred to as the experimental / observational data. In order to explain the observational data, it is required to understand the relationship between the distribution of properties of the physical system under study (e.g. earth) and observable geophysical response.

System of equations → described the relationship is FORWARD THEORY

Inverse theory is an organised set of mathematical statistical techniques (e.g. calculus, matrix algebra, statistical estimation & inference etc) for retrieving useful information about the physical system (physical world) from controlled observations on the system. It is directly relate to:

(1). Analysis of experimental data.

(2). Fitting of mathematical model to estimate the model parameter.

(3). Optimal experimental design.

Examples of problems where inverse theory is used:

(i). Curve fitting.

(ii). Digital filter design / De-convolution of seismogram.

(iii). Determination of earth structure / ore deposits / energy resource form geophysical information.

(iv). Determination of earthquake location wave arrival times.

Geophysical processes & Systems:

Geophysical Processes:

Seismic and EM wave propagation through the earth & current or fluid flow (in porous) rocks.

Geophysical systems:

(1). Density distribution within the Earth.

(2). Velocity distribution within the Earth.

(3). Temperature distribution within the Earth.

(4). Resistivity distribution within the Earth.

(5). Distribution radioactive materials within the Earth.

(6). Magnetic susceptibility variation within the Earth.

Geophysical explanations philosophy & inverse theory:

Goal of explanation of geophysics:

Understand / reconstruct the structure of the earth from recorded data. (above / below the earth).

Need to data:

Make observation on the various geophysical processes.

Often data may be (i) Noisy

(ii) Incomplete

(iii) Insufficient

We need to get something out of the data to proceed with processing data.

Inverse theory provides a formalism by which many of the questions fundamental to geophysics / geophysical data processing may be entertained. (e.g. optimum sampling rate, how many more data are nodded / desired accuracy).

Theoretical modelling technique helps to improve the understanding the relation between observed data (due to earth response to some excitation) & various subsurface physical property changes or discontinuity that may have generated.

Types of geophysical data:

(1). Data: Mass / moment of inertia of the Earth.

(2). Measurement of travel time and seismic wave (earthquake and natural explosions).

(3). Gravity / Magnetic anomaly.

(4). Apparent resistivity of the ground.

(4). Well draw- down data.

Data → Field / Lab based.

Physical properties:

(i). Seismic velocity.

(ii). Bulk & shear moduli.

(iii). Ground resistivity.

(iv). Magnetic susceptibility.

Scaled down physical models of the earth which are useful when & where mathematical models are very complicated & different to work.

Gathering data

Field experiment:

Controlled excitation ⟶ Unknown earth system ⟶ Observed response

(e.g. inductive excitation) (e.g. resistive data)

Laboratory experiment:

Systematic input → Known physical scaled model → Observed response

(e.g. seismic model) (Seismic data)

d_i sample drawn from a set of equally likely events / values.

Experiment → Outcome may vary

Under some condition

Instrument error Human error

We have to find

(i). Mean of the data.

(ii). STD / uncertainty.

(iii). Modelling of lithosphere's response to loading / strain rate variations in sedimentary basins.

(iv). Well (pump) test analysis in hydrology.

(v). Factor analysis in geology.

(vi). Geochronology determination & geomagnetic reversal data.

(vii). Satellite navigation.

(viii). Optimal control of engineering system

(ix). Medical tomography.

(x). Decisions making / operational research in management and mineral economics.

(1). Curve fitting:

FORWARD THEORY = Mathematical expression to represent or reproduce "data"

$$d_i = \sum_{j=1}^{p} G_{ij} m_j$$

FORWARD THEORY = y = F(x)

Where y_i = data $\approx ax_i^2 + bx_i + c$

Let, x = 1, ..., 10 km (1by 10)

You have say y = 5, 20, 7, 15, ... (1by 10)

Directly determine a, b, c by least square data fitting / minimum of error between y predicted from forward theory and observed response.

$$s = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i + c)$$

$$\frac{\partial s}{\partial a} = 0, \frac{\partial s}{\partial b} = 0, \frac{\partial s}{\partial c} = 0$$

(2). Digital filter design / deconvolution of seismogram:

Let two signals a(t) & b(t) they may be related by a filter function f(t)

$$a(t) = f(t) * b(t) = \int f(\tau) b(t - \tau) d\tau$$

Given / knowing a(t) & b(t), finding f(t) in signal analysis.

If the time series of length n, filter function of length p, the integral equation may written as

$$a_i = \Delta t \sum_{j=1}^{p} f_j b_{i-j+1}$$

$b_i = 0$, $i < 1$ or $i > n$ and Δt = sampling interval

The equation is a linear and the unknown filter coefficients f_i can be react in the form

D = Gm

Where, m=sought filter and d = time series data

$$G = \begin{bmatrix} b_1 & 0 & 0 & 0 & 0 & .... & 0 \\ b_2 & b_1 & 0 & 0 & 0 & .... & . \\ b_3 & b_2 & b_1 & 0 & 0 & .... & . \\ . & . & . & b_1 & 0 & .... & . \\ . & . & . & . & . & .... & . \\ b_n & b_{n-1} & b_{n-2} & b_{n-3} & b_{n-4} & .... & b_k \\ & & & & & & \end{bmatrix}$$

Where k= n – p +1, the above system is over determined.

Description / Characterization of Geophysical process: Mathematical models:

Most of the geophysical processes can be described mathematically. As mentioned earlier set of equations that characterise each process or geophysical system is known as FORWARD THEORY or Model.

The word 'Model' has various connotations in geoscientific community. It may refer to conceptual Model by geologist or physical (Lab scale) and mathematical models as is common in geophysics. Here mostly we will use mathematical model but may refer to conceptual model when if required.

A number of geophysical processes can be expressed by integral equation of the form

$$d_i = \int_0^z k_i(z)\, p(z)\, dz \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1)$$

Where d_i = measurable or observable response of system to an i^{th} external input or excitation (e.g. explosions or electrical current injection into the ground.)

p(z) is the function relate to the physical properties of the earth (e.g. density / velocity distribution of the earth & expressed as function of depth / laterally homogeneous earth) / Model parameter & k_i are called data kernels. The data

kernels describe the relation between data & earth model function p(z). The model parameter may be continuous function of radius & position.

Example: Travel time seismic source & receiver along a ray path for a continuous velocity field v(x,z) is given by

$$t = \int \frac{1}{v(x,z)} dt$$

Mathematical description & physical system refer to Forward Theory.

Forward theory has been developed to predict the data or observed response that we would record over a hypothetical Earth- type structure. These data are therefore variously called Synthetic or predicted data.

Discretization & Linearization:

In many case earth model is continuous function of depth or radius consider, for example, the mass and moment of inertia of the earth. Both are related to density within the earth by the formula

$$Mass = 4\pi \int_0^R r^2 \rho(r) dr \ldots\ldots\ldots\ldots (1.3a)$$

$$MI = \frac{8\pi}{3} \int_0^R r^4 \rho(r) dr \ldots\ldots\ldots\ldots (1.3b)$$

Where R is the earth's radius and ρ(r) corresponds to p(z) in eq. (1.1) and is the density at radial distance r.eq^n (1.3a & 1.3b) may be combined to give the general expression

$$d_i = \int_0^R k_i(r) \rho(r) dr \ldots\ldots\ldots\ldots (1.4)$$

This equation can be solved by digital computer. The computational formula

$$d_i = \sum G_{ij} m_{ij} \ldots\ldots\ldots\ldots (1.5)$$

Here m_j set to $\rho(r)\,dr$; G_{ij} set to $k_i(r)$

In this equation we can say that theoretical problem is discretized.

For technical reasons our field or experimental observation are recorded over finite intervals (e.g., discrete frequencies or fixed bandwidth) instead of all in the range $[0, \infty]$ required to uniquely characterise the earth- system. For computational simplicity we often express the continuous distribution of earth's physical properties p(z) by finite set of parameters.

e.g., layered earth structure with each layer having a specific density and thickness. This practice is referred to as parameterization. For convenience we will be considering only discrete models & discrete parameter which are easier to handle than continuous distributions.

Let, $$t_i = \sum_{j=1}^{p} \frac{L_{ij}}{v_j} \ldots\ldots\ldots\ldots (1.6)$$

Notice that travel time is not directly proportional to the model parameter v but to its inverse. The relation is said to be non-linear in v.

However if we define c = 1/v, where c is the slowness of the seismic wave then

$$t = .\sum_{j=1}^{p} L_{ij} c_j \ldots\ldots\ldots (1.7)$$

Which is the form d = Gm. The relationship is said to be linear. Such transformation is called "Linearising parameterization".

<u>Two layer earth model:</u>

$$\rho_a(L) = \rho_1\left[1 + 2L^2 \int_0^{\infty} K(\lambda) J_1(\lambda L) \lambda d\lambda\right] \ldots\ldots\ldots\ldots (1.8)$$

L = AB/2 , J_1 first order Basel

K(λ) is the function of ρ_1, ρ_2 , t and λ is integration variable.

$$K(\lambda) = \frac{-K_{12}\exp(-2\lambda t)}{1+K_{12}\exp(-2\lambda t)} \ldots.(1.9)$$

$$K_{12} = \frac{\rho_1 - \rho_2}{\rho_1 + \rho_2} \ldots.(2.0)$$

We cannot put equation (1.8) in simple form d= Gm as we did in equation (1.2). The resistivity depth sounding problem is highly nonlinear. The usual method of dealing with such problem involves using Taylor's theorem, a procedure termed Linearization.

Meaning of inverse problem:

Forward problem: T(z) = a + b(z)

Given: Estimated or values of the model parameters (a, b)

Determine: Theoretical response (data)

Model parameter → Numerical representation of system dynamic i.e. Forward Theory → Computed response

Input **Operators** **Output**

Classification of inverse problem:

(a). Over determined problem:

The sought model consists of fewer parameters than the number of field data. It is solved by best fit to the data. [Least – square fitting method]

(b) Underdetermined problem:

The sought model consists of more parameter than the number of field data. In this case many solutions / model can explain the field data non-unique solution.

In that case it is possible to use the method originally devised for over determined problem to derive the smooth model.

(c) Even / exact determined problem:

The model parameter sought is exactly equal to the number of observation of the field data.

Discretization and parameterization:

Geophysical measurements are usually made to determine the subsurface properties or structure. The properties be uniquely determined if the measurement span the observational band- width $[0 \infty]$. However, this is not possible owing to technical limitations; we typically conduct our field experiments over a finite observation interval. The outcomes are discrete numerical values called field data which are in complete and often inconsistent. For computation simplicity we often tend to seek the minimum set of parameters that describe our observations or earth structure. The inverse problem is therefore discretized and our hypothetical Earth- model is parameterized into a finite number parameters. Geophysical inverse theory is thus concerned with the approximation of otherwise continuous functions with a finite number of parameters.

Weighted measure of length as a type of A priori solution:

There are many instances in which $L = m^T m$ is not a good measure of solution of simplicity. Let us consider the inverse problem for finding density fluctuations in the ocean. One may not look for solution that is smallest in the sense of closest to zero but one that is smallest in the sense that it is closest to some other values such as average density of sea water. The obvious generalized solution,

$L = (m-m^2)^T (m-m^2)$

Where m^2, a priori value of the model parameters / average model parameters.

Sometimes, the whole idea of length is a measure of simplicity is not appropriate. One may feel that a solution is simple if it is smooth or if it is in some sense flat. This measure may be particularly appropriate, when the model parameters represent a ‘discretized continuous function such as density / opacity properties such as flatness of continuous function can easily be quantified by norm of its first derivative.

The flatness of a vector m is

$$l = \begin{bmatrix} 1 & -1 & 0 & . & . & . & 0 \\ 0 & 1 & -1 & 0 & . & . & 0 \\ 0 & 0 & 1 & -1 & 0 & . & 0 \\ . & . & . & . & . & 0 & . \\ . & 0 & . & . & . & . & . \\ 0 & 0 & . & . & . & -1 & 1 \\ & & & & & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ . \\ . \\ . \\ m_M \end{bmatrix} = D_1 m$$

Where D_1 = flatness / roughness matrix.

So, solution roughness can be quantified by the second derivative,

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & . & . & 0 \\ 0 & 1 & -2 & 1 & 0 & . & . \\ . & . & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & 1 & 2 \\ 0 & . & . & . & . & . & 1 \\ & & & & & & \end{bmatrix}$$

$$L = l^T l = [Dm]^T [Dm] = m^T D^T D m = m^T w_m m$$

Where $w_m = D^T D$ can be interpreted as weighting factor.

The generalised solution,

$$L = \left[m - m^2\right]^T w_m \left[m - m^2\right]$$

By suitable choosing the priori model m and weighting matrix w_m , we can quantify a wide variety of measure of simplicity.

Similarly, weighted measure of prediction error can be written as,

$$E = e^T w_e e$$

Where w_e denotes the relative contribution of each individual error to the total prediction error.

Weighted least- square solution of over – determined problem:

Letting d = Gm

$$q = e^T w_e e$$
$$= (d - Gm)^T w_e (d - Gm)$$
$$= (d^T - m^T G^T) w_e (d - Gm)$$
$$= \left(d^T w_e - m^T G^T w_e\right) \times (d - Gm)$$
$$= d^T w_e d - d^T w_e G m - m^T G^T w_e d + m^T G^T w_e G m$$

$$\left(\frac{\partial q}{\partial m^T}\right) = 0 - 0 - G^T w_e d + G^T w_e G m = 0$$

$$G^T w_e G m = G^T w_e d$$

$$m^{est} = \left(G^T w_e G\right)^{-1} G^T w_e d$$

Weighted minimum length solution of completely under- determined problem:

$$L = \left(m - m^2\right)^T w_m \left(m - m^2\right)$$
$$L = \left(m^T - m^{2^T}\right) w_m \left(m - m^2\right)$$
$$L = \left(m^T w_m - m^{2^T} w_m\right)\left(m - m^2\right)$$
$$= m^T w_m m - m^T w_m m^2 - m^{2^T} w_m m + m^{2^T} w_m m^2$$
$$Let, m^T = m \;\&\; m^2 = m^{2^T}$$
$$= m^T w_m m^T - m^T w_m m^2 - m^{2^T} w_m m^T + m^{2^T} w_m m^2$$

Total error / misfit

$$q = L + \lambda[d - G(m)]$$
$$\left(\frac{\partial q}{\partial m^T}\right) = 2w_m m^T - w_m m^2 - m^2 w_m + 0 - \lambda G = 0$$
$$m^T = m^2 + \frac{\lambda}{2} G \times \frac{1}{w_m}$$
$$\therefore, m = m^2 + \frac{1}{2}\left(\frac{G^T \lambda}{w_m}\right)$$
$$Now,$$
$$d = Gm$$
$$d = G\left(m^2 + \frac{G^T \lambda}{2w_m}\right)$$
$$d = Gm^2 + \frac{\lambda G G^T}{2w_m} = Gm^2 + \frac{\lambda}{2}.Gw_m^{-1}G^T$$
$$\lambda = 2\left(Gw_m^{-1}G^T\right)\left(d - Gm^2\right)$$
$$m^{est} = m^2 + \frac{1}{2}.\left(\frac{G^T \lambda}{w_m}\right)$$

$$m^{est} = m^2 + \frac{1}{2} w_m^{-1} G^T .2\left(G w_m^{-1} G^T\right)^{-1}(d - Gm^2)$$

$$m^{est} = m^2 + w_m G^T (G w_m G^T)^{-1}(d - Gm^2)$$

Weighted damped least – square:

If the equation

Gm = d is slightly under-determined it can be solved by minimizing a combination of prediction error and solution length,

$E + \beta^2 L$.

The solution is then

$$m^{est} = m^2 + \left[G^T w_e G + \beta^2 w_m\right]^{-1} G^T w_e \left[d - Gm^2\right]$$

Which is equivalent to

$$m^{est} = m^2 + \left[G\ w_m^{-1} G^T + \beta^2 w_e^{-1}\right]^{-1}\left[d - Gm^2\right]$$

In both instance, one must take care to ascertain whether the inverse actually exist. Depending on the choice of the weighting matrices, sufficient a priori information may be or may not have been added to damp the indeterminacy.

Inverse Problem

Given: Field observation (Earth system response), T(z) data

$$T(z) = a + bz$$

Determine: Parameters of the earth- model (a, b)

The inverse process

Observe data ⟶ Mathematical tool / Inverse theory ⟶ Model parameter

1/p Operator output

Generalised least square solution:

The linear problem is posed in a matrix form d = Gm. We now want to solve for m.

For perfect data: there are no experimental errors.

$m = G^{-1} d$

However gauss (1809) suggested that due to experimental errors, practical data d_i would not fit the model exactly, i.e., d = Gm+ e

The best way to get unique solution for the model parameters is to minimize the sum of squares of the residual, i.e.,

$$q = e^T e = \sum_{j=1}^{n} \left(d_i - \sum_{j}^{p} G_{ij} m_j \right)^2, \; j = 1, \; p \ldots\ldots\ldots\ldots (1)$$

We can re- write the equation as

$$q = (d - Gm)^T (d - Gm)$$

$$Expansion$$

$$= \left[d^T d - d^T Gm - m^T G^T d + m^T G^T Gm \right]$$

$$\left(\frac{\partial q}{\partial m_j} \right) = 0$$

$$or, -d^T G - G^T d + G^T Gm + m^T G^T G = 0$$

Giving

$$2G^T Gm = 2G^T d$$

The generalised least square solution

(Unconstraint solution)

$$m^2 = \left[G^T G\right]^{-1} G^T d$$

The damped – least square solution

$$m^2 = \left[G^T G + \beta^2 I\right]^{-1} G^T d$$

B is the Lagrange multiplier.

Alternative (method):

$$q = (d - Gm)^T (d - Gm)$$
$$= d^T d - d^T Gm - m^T G^T d + m^T G^T Gm$$
$$\frac{\partial q}{\partial m^T} = 0 - 0 - G^T d + G^T Gm = 0$$
$$G^T Gm = G^T d$$
$$m^2 = (G^T G)^{-1} G^T d$$

<u>Orthogonal decomposition of real symmetric matrix:</u>

If G is real symmetric matrix can be factored into

$$G = UQU^T$$

With orthonormal Eigen vector in U and real Eigen values in Q

But for dimensional / reason / non- symmetric / rectangular matrix this eigen vector decomposition is not useful. In that case singular value decomposition is the alternative.

Singular value decomposition (SVD):

Using SVD we can factorized an n by n or n by m Jacobian matrix G, such that

$$G = UQL^T$$

Where n=data & m = model parameter.

$U(n \times m)$ & $L(m \times m)$

are two orthogonal matrixes, containing respectively the data space / parameter space Eigen vectors.

$Q(m \times m)$

Q is diagonal matrix containing at most real non-zero Eigen values of G with a condition $r \leq m$. These diagonal entities in matrix Q (α_1, α_2, ...α_p) are called singular value G.

Columns of U are Eigen vector of G G^T, columns of L are Eigen vector of G^T G and Q is rectangular matrix with singular value in its main diagonal.

Application:

$$\Delta m = \left(G^T G + \beta^2 I\right)^{-1} G^T \Delta d$$

Δm is the parameter correction vector and Δd is the data correction vector.

G= Jacobian matrix containing partial derivative of data with respect to the initial model parameter.

$$\Delta m = \left(LQ^2L + \beta^2 I\right)^{-1} LQU^T \Delta d$$
$$Now, \left(LQ^2L^T + \beta^2 I\right)$$
$$= \left(Ldiag\left\{\alpha_{ij}^2\right\}L^T + \beta^2 I\right)$$
$$= Ldiag\left(\alpha_j^2 + \beta^2\right)L^T$$

Now,

$$Ldiag\left(\alpha_j^2 + \beta^2\right)L^T = Ldiag\left\{\frac{1}{\alpha_j^2 + \beta^2}\right\}L^T$$

$$\Delta m = Ldiag\left\{\frac{1}{\alpha_j^2 + \beta^2}\right\}L^T.LQU^T\Delta d$$

$$\Delta m = Ldiag\left\{\frac{\alpha_j}{\alpha_j^2 + \beta^2}\right\}U^T\Delta d$$

↓

Damped least square solution

<u>Damped least square solution of VES data (1-D) data:</u>

<u>Forward modelling:</u>

Following Koeford 1970, earth layer consisting of homogeneous / isotropic layers

$$\rho_a(s) = s^2\int_0^\infty T(\lambda)J_1(\lambda s)\lambda d\lambda \ldots\ldots\ldots(1)$$

S = AB/2, J_1 is the first order Bessel function and λ is the integration variable.

T(λ) Resistivity transfer function.

$$T_i(\lambda) = \frac{T_{i+1}(\lambda) + \rho_i \tanh(\lambda h_i)}{1 + T_{i+1}(\lambda)\tanh(\lambda h_i) \big/ \rho_i} \quad \text{.....(2)}$$

n= no. of layer

ρ_i & h_i true resistivity and thickness of i^{th} layer.

$$T_i(\lambda) = \frac{T_{i+1}(\lambda) + \rho_i \tanh(\lambda h_i)}{1 + T_{i+1}(\lambda)\tanh(\lambda h_i) \big/ \rho_i} \quad \text{.....(2)}$$

Inversion:

$$\Delta m = \left(G^T G + \beta^2 I\right)^{-1} G^T \Delta d$$

Using SVD:

$$\Delta m = L\,diag\left\{\frac{\alpha_j}{\alpha_j^2 + \beta^2}\right\} U^T \Delta d$$

Initially damping factor is said to be large positive value while making the full use of steepest descent method. Subsequently at each iteration the damping factor is multiplied by a factor less than unity so that least square method dominates near the solution.

Arinason & Hersir (1988)

$$\rho = \alpha_w \Delta c^{1/w}$$

$$\Delta c_r = \frac{c_{r-1} - c_r}{c_{r-1}}$$

w= Test number.

A = Parameter eigen value.

c_r = Misfit value at current iteration.

c_{r-1} = Misfit value at previous iteration.

Year 1760 – 1810.

Boscovich & Laplace: Minimizing the sum of the values of the misfit function.

Least – absolute – values method.

Legendre & Gauss: Minimizing the sum of the squared values of misfits.

Least square method.

Two methods follow two different hypothesis statistical distribution of error in data follows.

Laplacian distribution

$$f(x) \approx \exp(-|x|)$$

Gaussian distribution

$$f(x) \approx \exp(-x^2)$$

Least- absolute- values method follows $\longrightarrow$ Linear programming (LPP)

Least- square method follows $\longrightarrow$ Linear algebra.

Over whelming popularity of least square method is due to the use of linear algebra which is simpler than LPP.

It is widely recognized that the least absolute criteria is less sensitive than the least-square to the presence of large uncontrolled errors in the data. (Property called robustness).

Inverse problem may be posed optimization problem

$$\min\|\Delta x - y\|$$

(Similarly ||Gm – d||)

It can be shown that for the l_p family of norms, if this optimization problem has a solution, then it is unique, provided the matrix has full column rank and p > 1.

For p = 1, the norm loses in the technical jargon, strict convexity.

Let us consider one- parameter linear system,

$$\begin{bmatrix} 1 \\ \lambda \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Let $\lambda \geq 0$. If we solve the problem on the open interval $x \leq 0,1$

l_p error function

$$E_p(x) = \left[|x-1|^p + \lambda^p |x|^p \right]^{1/p}$$

Remember λ is just parameter.

This has unique solution for any λ, p > 1

p = 1(multi-valued function / not a single valued)⟶ Uniqueness theorem is only valid p >1.

<u>Describing & formulating inverse problem:</u>

Key questions:

1. what is applicable parameterization?

(a) discrete (b) continuous ?

2. What is the number of the geophysical data?

≈ what are the errors in the observation?

3. Can we pose the problem mathematically? Ill posed / well posed??

4. Are there any physical constraints on the problem?

5. What types of solution to the problem are describe; to what accuracy?

Are we looking for exact / approximate solution?

6. Is the problem is non- linear / linear?

7. What are the confidence limits of the solution? Can it be appraised by other means?

Types of solution to the inverse problem:

What do we ask of a given data set?

Depending on the problem in hand, we may have variety of solution types.

If we are analysing geophysical time series, we might have interested in finding

(a) Optimum sampling rate.

(b) Suppressing noise / unwanted signal.

(c) Remember the trend from the time series.

If we deal with finding physical properties distribution over subsurface such as

(Layer resistivity, layer velocity , layer density, thickness etc.)

However, owing to the fact that some of the solutions are inherently non-unique. e.g., Conductivity thickness (σt).

Geophysical technique can resolve (σt) productivity uniquely rather than the individual parameter. We may prefer slowness (1/v) to the velocity (v) of acoustic waves, due to the advantage of such parameterization. In some situation we may be of interest to seek the non- uniqueness bounds defined by our data or a suite of extreme models that define a particular aspect of the model or even the model space rather a single model for the subsurface.

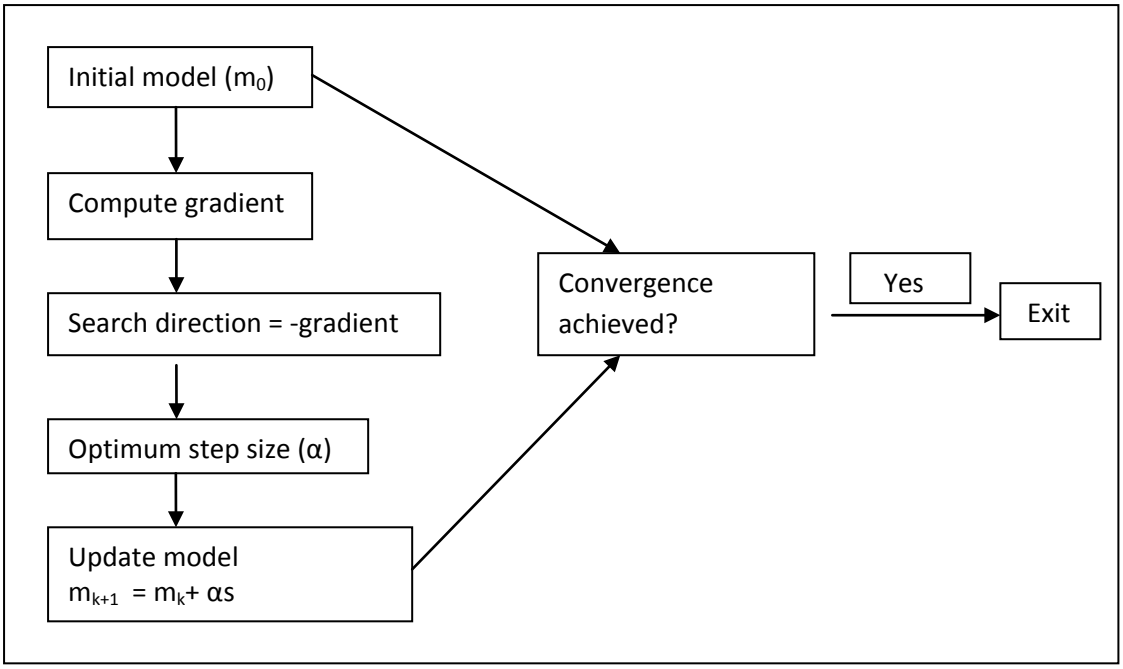
Steepest-decent (Gradient method):

A linear inverse problem can be posed as optimization problem where the cost function is surface of the cost function is quadratic. The cost function is a parabolic with a single minimum.

$$f(m) = \frac{1}{2} m^T G m - d^T m$$

Will have single minima with respect to the model parameter if the second order partial determine matrix G is positive definite.

Though the negative gradient provides the direction of the maximum decrease in the cost function if does not provide the step size. One way to keep the step size constant. However larger step size may miss the minimum point, so the optimization becomes oscillatory. Step size may be calculated at every iteration by computing the first order derivative with respect to the step size and equating to zero.



Flowchart

Code:

1. Given m_0

2. Set k ← 0

3. While (convergence criteria not satisfied) d_0

$$4. S^k = -\nabla f(m^k)$$

5. If $S^k = 0$ then

Stop

6. End if

$$7. \alpha^k \leftarrow \min f(m^k + \alpha s^k)$$
$$8. m_{k+1} \leftarrow m^k + \alpha^k s^k$$
$$9. k \leftarrow k + 1$$

10. end while.

$$q = (d - Gm)^T (d - Gm) \approx |d - f(m)|^2$$

$$\Delta m = -k \frac{\partial q}{\partial m} = -k\left\{-2(d - f(m)).\frac{\partial f(m)}{\partial m}\right\}$$

$$\frac{\partial q}{\partial m} = -2(d - f(m)).\frac{\partial f(m)}{\partial m} \equiv -2G^T(d - f(m))$$

$$\Delta m = -k.\{-2G^T(d - f(m))\} = [2k].G^T(d - f(m))$$

$$m^{k+1} = m^m + \Delta m$$

Where k= constant.

Conjugate gradient (CG): (Fletcher – reeves Method):

$$f(m) = \frac{1}{2} m^T G m - d^T m \ldots\ldots\ldots(1)$$

is a quadratic equation of m.

Where d & G are constant, G is positive definite.

Local gradient of function

$$\nabla f(m) = Gm - d \ldots\ldots\ldots(2)$$

Let m_0 is initial model.

The direction of steepest –descent

$$S_o = -\nabla f_o$$

$$= Gm_o + d \ldots\ldots\ldots(3)$$

Minimization of equation (1) at m_1

$$Gm_1 - d = 0 \ldots\ldots\ldots\ldots\ldots(4)$$

Suppose we can find a set of w vectors (where w is the dimensionality of the parameter space) which are mutually conjugate with respect to G so that

$$S_j^T G S_i = 0 \ldots\ldots\ldots\ldots\ldots (5)$$
$$j \neq i$$

Then it is easily shown that these vectors will be linearly independent if G is positive definite. Such vectors therefore form a complete, but non- orthogonal, basis set in parameter space. Say, we are starting from some point m_0, we wish to get to the minimum m_1 of the function. The difference between vectors m_0 & m_1 can be written

$$m_1 - m_0 = \sum_{j=1}^{w} \alpha_i S_i \ldots\ldots\ldots\ldots (6)$$

$$m_j = m_0 + \sum_{i=1}^{j-1} \alpha_i S_i \ldots\ldots\ldots\ldots (7)$$

In the iterative form we can write

$$m_{j+1} = m_j + \alpha_j S_j \ldots\ldots\ldots\ldots (8)$$

This represents a succession of steps parallel to the conjugate directions, with step length controlled by the parameters α_j.

If we multiply S_j^T G with equation (6) & using eqn (4)

$$S_j^T G m_1 - S_j^T G m_0 = \sum_{i=1}^{w} \alpha_i S_j^T G S_i \text{.....(9)}$$

$$S_j^T (d - G m_0) = \sum_{i=1}^{w} \alpha_i S_j^T G S_i$$

$$\alpha_i = \frac{S_j^T (d - G m_0)}{S_j^T G S_i} \text{.....(10)}$$

In simple form, we multiply S_j^T G to equation (7)

$$S_j^T G m_j = S_j^T G m_0 + \sum_{i=1}^{j-1} \alpha_j S_j^T G S_i$$

$$S_j^T G m_j = S_j^T G m_0 \text{.....(11)}$$

Let S_j^T G S_i = 0 ; i ≠ j

$$\alpha_j = \frac{S_j^T d - S_j^T G m_0}{S_j^T G S_i} = \frac{S_j^T d - S_j^T G m_j}{S_j^T G S_i} = \frac{S_j^T (d - G m_j)}{S_j^T G S_i} = -\frac{S_j^T \nabla f}{S_j^T G S_i}$$

.....(12)

For general k^{th} iteration we can write

$$\alpha_k = -\frac{S_k^T \nabla f_k}{S_k^T G S_k} \text{.....(13)}$$

$$\alpha_0 = -\frac{S_0^T \nabla f_0}{S_0^T G S_0} \text{.....(14)}$$

Sine residual vector at the k^{th} iteration is

$$r_k = G m_k - d = \nabla f_k \text{.....(15)}$$

$$so, \alpha_k = -\frac{S_k^T r_k}{S_k^T G S_k} \text{.....(16)}$$

The current search direction S_1 is given as a linear combination of the previous search direction and the current gradient vector.

The second search direction,

$$S_1 = -\nabla f(m_1) + \beta_1 S_0 \ldots\ldots\ldots (17)$$

Where β_1 = scalar is chosen such that the search direction S_1 & previous search direction S_0 are conjugate implying $S_0^T G S_1 = 0$

Multiply $S_0^T G$ in both sides

$$S_0^T G S_1 = S_0^T G\{-\nabla f(m_1) + \beta_1 S_0\} = 0$$

Since,

$$m_1 = m_0 + \alpha_0 S_0$$

$$S_0 = \frac{m_1 - m_0}{\alpha_0}$$

$$\left(\frac{m_1 - m_0}{\alpha_0}\right)^{-1} G S_1 = S_0^T G\left(-\nabla f(m_1) + \beta_1 S_0\right) = 0$$

$$\nabla f(m_1) - \nabla f(m_0) = G(m_1 - m_0)$$

We can write,

$$\{\nabla f(m_1) - \nabla f(m_0)\}\{\nabla f(m_1) - \beta_1 S_0\} = 0$$

$$\nabla f(m_1)^T \nabla f(m_1) - \nabla f(m_1)^T \beta_1 S_0 - \nabla f(m_0)^T \nabla f(m_1) + \nabla f(m_0)^T \beta_1 S_0 = 0$$

$$\beta_1 \times \{\nabla f(m_0)^T S_0 - \nabla f(m_1)^T S_0\} = \nabla f(m_0)^T \nabla f(m_1) - \nabla f(m_1)^T \nabla f(m_1)$$

$$\beta_1 = \frac{\nabla f(m_1)^T \nabla f(m_1) - \nabla f(m_0)^T \nabla f(m_1)}{\nabla f(m_1)^T S_0 - \nabla f(m_0)^T S_0}$$

$$\beta_1 = -\frac{\nabla f(m_1)^T \nabla f(m_1)}{\nabla f(m_0)^T S_0}$$

$$\beta_1 = \frac{\nabla f(m_1)^T . \nabla f(m_1)}{\nabla f(m_0)^T . \nabla f(m_0)}$$

$$Since, S_0 = -\nabla f(m_0)$$

And taking conjugate condition equation (3).

Similarly, we express third direction as the linear combination of the current gradient & all past search direction.

$$S_2 = -\nabla f(m_2) + \beta_2 S_1 + \gamma_2 S_0$$

S_2 is the current search direction at the updated model m_2 , β_2 & γ_2 are two scalars that ensure conjugacy among the current and past search directions.

The condition of conjugacy between S_0 and S_2 requires that γ_2 be zero. From the condition of conjugacy between S_1 & S_2 , we obtain

$$\beta_1 = \frac{\nabla f(m_2)^T . \nabla f(m_2)}{\nabla f(m_1)^T . \nabla f(m_1)}$$

Hence the current search direction is given by,

$$S_2 = -\nabla f(m_2) + \frac{\nabla f(m_2)^T . \nabla f(m_2)}{\nabla f(m_1)^T . \nabla f(m_1)} \times S_1$$

The generalised expression for the k^{th} iteration is written as,

$$S_k = -\nabla f(m_k) + \beta_k S_{k-1}$$

$$where, \beta_k = \frac{\nabla f(m_k)^T . \nabla f(m_k)}{\nabla f(m_k)^T . \nabla f(m_k)}$$

The residual vector r_k at the k^{th} iteration is given by,

$$r_k = \nabla f(m_k)$$

$$\beta_k = \frac{r_k^T r_k}{r_{k-1}^T r_k}$$

Code:

1. Given m_0

$$2. Set, k \leftarrow 0; r_0 \leftarrow Gm_0; S_0 \leftarrow -r_0$$

3. While (convergence criteria is not satisfied) d_0

$$4.\alpha_k \leftarrow -\frac{S_k^k r_k}{S_k^T G S_k}$$
$$5.m_{k+1} \leftarrow m_k + \alpha_k S_k$$
$$6.r_{k+1} \leftarrow Gm_{k+1} - d$$
$$7.\beta_{k+1} \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$$
$$8.S_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} S_k$$
$$9.k \leftarrow k+1$$

10. end while

Non-linear Conjugate Gradient:

If the function to be minimized quadratic example

$$f(m) = \frac{1}{2} m^T G m - d^T m$$

The step length α_k along the direction S_k for which the function $f(m_k + \alpha_k S_k)$ is minimum can be analytically calculated by minimizing the function $f(m_k + \alpha_k S_k)$ with respect to α_k .

However for non-linear function in general there does not exist analytic expression to determine the optimum step length α_k . The non-linear function is minimized along the direction of S_k and the residual

$r_k (= \nabla f)$

is replaced by the gradient of the non-linear function.

Code:

1. Given m_0

$$2.Compute,\ f_0 \leftarrow f(m_0); \nabla f_0 \leftarrow \nabla f(m_0)$$
$$3.Set, k \leftarrow 0; S_0 \leftarrow -\nabla f_0$$

4. While (convergence criteria is not satisfied) d_0

5. Calculate α_k by line search

$$6. m_{k+1} \leftarrow m_k + \alpha_k S_k$$

$$7. Calculate, \nabla f(k+1)$$

8. If (Fletcher- Reeves), then

$$9. \beta_{k+1} \leftarrow \frac{\nabla_f^T (k+1) \nabla f(k+1)}{\nabla f_k^T \nabla f_k}$$

10. End if.

11. If (Polak- Ribiere) then

$$12. \beta_{k+1} \leftarrow \frac{\nabla f^T (k+1)(\nabla f_{k+1} - \nabla f_k)}{\|\nabla f_k\|^2}$$

13. end if

$$14. S_{k+1} \leftarrow -\nabla f_{k+1} + \beta_{k+1} S_k$$

$$15. k \leftarrow k+1$$

16. end while

Numerical studies shows that Polak – Ribiere method is generally more robust then the Fletcher- Reeves method (Nocedal & wright, 1999).

Non-linear CG is the generalization of CG to optimize non-linear function. While CG aims at finding the solution of the non-linear equation $G^T G\, m = G^T d$. The success of non-linear CG lies in the fact that paraboloid approximation of the function (cost) in the vicinity of the initial model encompasses the global minimum of the cost function. When this condition fails, non-linear CG may not Necessarily converge to the global minimum of the cost function.

Non-Linear Inverse problems:

Non- linear inverse problem belong to a class of inverse theory where there exists non- linearity in the model data relationship. Non- linear model data relationship leads to non-quadratic cost function as opposed to the linear model-data relationship where the cost function is quadratic. The cost function topology is likely to be multimodal in case of non- quadratic cost function. Optimization of such cost function is complex because of the presence of several minima in the cost function surface.

Example:

Certain inverse problems are inerasable via approximations of the forward operator. For example estimation of earth elastic parameters from AVO data using a forward operator that involves. Computation of reflection coefficients from Zeoppritz equation is a non-linear problem. However, for a reasonable angle of reflection, The Zeoppritz equation can be approximated to obtain the linear form. Aki-Richards equation (Aki and Richardss, 1980) is one such linear representations of the Zeoppritz equation.

Structure of an inverse problem:

The inverse solution of an equation

$d = Gm$ is

$m = G^{-1} d$

This leads to a following question of interest.

(a). Existence:

Given observed data for the system, is there some value for the unknown parameters that actually yields the observed data? If not the inversion problem has no solution.

(b). Uniqueness:

Can the unknown parameters in principle be uniquely determined from the measured data? Or could two different sets of values for the unknown parameters give rise to the same observation?

A solution is said to be unique if changing a model from m_1 to m_2, the data will also change from d_1 to d_2 such that $d_1 \neq d_2$.

(c). Stability:

If the measured data contains small errors, will the error in the resulting estimates of the unknowns be correspondingly small? Or could small measurement error lead to huge errors in our estimates?

(d) Robustness:

Robustness indicates the level of insensitivity in presence of large uncontrolled error i.e. (outlier in the data). An inverse problem for which existence, uniqueness, and stability hold is said to be well-posed. The alternative is an ill-posed inverse problem. Ill-posed problem tend to be the most common.

Steepest- Descent Algorithm:

A linear / literalized problem can be posed as an optimization problem where the cost function is quadratic. This means that the surface of the cost function is a paraboloid with a single minimum.

A quadratic function

$$f(m) = \frac{1}{2} m^T G m - d^T m$$

Will have a single minimum with respect to the model parameters if the second order partial derivative matrix G is positive definite.

To the find the minimum of the cost function then the algorithm proceeds along a negative gradient direction calculated at each iteration.

Though the negative gradient provides the direction of maximum decrease in the cost function, it does not provide the step size. One way to obtain the step size. One way to obtain the step size is to keep it constant. However, if the step size is to keep is large, there is a possibility that the algorithm may miss the minimum point and become oscillatory. In order to avoid, step size is calculated.

Newton Optimization method:

The idea of well known Newton's method of root finding for an invariate function. Let f (m) is a multivariate function whose Taylor series expansion at $m = m_i$ is given by,

$$f(m) = f(m) + \nabla f(m_i)(m - m_i) + \frac{1}{2}(m - m_i)^T H_i (m - m_i)$$

$$where, H_i = \nabla^2 f(m)\Big|_{m_i}$$

Is the Hessian matrix evaluated at $m = m_i$ .The first derivative is zero at the point where the function is minimum. So equating

$$\left(\frac{\partial f(m)}{\partial m}\right) = 0$$

We obtain

$$\nabla f(m) = \nabla f(m_i) + H_i(m - m_i) = 0$$

Thus m_{i+1} for the updated model is given by,

$$m_{i+1} = m_i - H_i^{-1}\nabla f(m_i)$$

This is analogous to the expression for the Newton's root finding method for an unvariate function which is given by

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

Convergence is achieved by the Newton's method provided that the Hessian matrix is non- singular. For quadratic cost function Newton's method will find the minimum in one step.

Marquardt Optimization method:

Marquardt's method for optimization (Marquardt 1963) uses the benefits of both Newton's technique to obtain a faster convergence. Marquardt's method Modifies the diagonal terms of the Hessian matrix is given by,

$$\widetilde{H} = H_i + \lambda_i I$$

Where I is an identity matrix and λ_i is a scalar to modify the diagonal elements of the Hessian matrix.

It is evident from the equation that for a very large λ_i the term λ_i I dominates the Hessian matrix H_i. In such a case

$$\widetilde{H}_i^{-1} = (H_i + \lambda_i I)^{-1} \cong \frac{1}{\lambda_i} I$$

So in that condition model is updated in a gradient descent approach. It is obvious that when λ_i

Is reduces to a small number, the model is updated with a Newton's model approach.

The Marquardt algorithm provides a faster convergence when λ_i is a set to a large parameter during the initial iterations and then gradually reduced to a small number during the later iterations as the updated model approaches the optimum point.

Data Resolution matrix:

Linear inverse problem takes the form Gm = d. Using the generalized theory, we get an estimate of model parameters,

$$m^{est} = G^{-g} d^{obs}$$

[For the sake of simplicity we assume that there is no additive vector / noise.]

$$d^{pred} = Gm^{est}$$
$$d^{pred} = G\left[G^{-g} d^{obs}\right]$$
$$d^{pred} = \left[GG^{-g}\right] d^{obs} = Nd^{obs}$$

Here the subscript 'pred' and 'obs' means predicted & observe respectively. The (N by N) square matrix $N = G\,G^{-g}$ is called the data resolution matrix. Data resolution matrix characteristics whether data can be independently predicted / resolved.

Model Resolution Matrix:

We imagine that there is true but unknown set of model parameters m^{true} that solve $G\,m^{true} = d^{obs}$

$$m^{est} = \left[G^{-g} d^{obs}\right]$$
$$m^{est} = G^{-g}\left[Gm^{true}\right]$$
$$m^{est} = \left[G^{-g}G\right] m^{true} = Rm^{true}$$

Where R is (M by M) model resolution matrix. If R = I the each model parameter is uniquely determined.

Determination of the damping factor in ridge regression / Marquardt Lavenberg:

For automated inversion, the common practice is to set β first to large positive value thus taking advantage of the good initial convergence properties of the steepest descent method thereafter β is multiplied by a factor less than unity after each iteration, so that the linear least- squares method predominates near the solution.

A variant of the procedure (Johansen, 1977) assumes as β the smallest Eigen value of G^TG matrix. If divergence occurs, it is replaced by the next largest Eigen value until the solution is obtained.

A more sophisticated method of damping has been developed by Meju (1988, 1992) which is in effect a hybrid of the two methods highlighted.

(1). Damping factor is determined empirically which is linked to approximate derivatives of a Lagrangian function (Herskovits, 1986) and is used in a minimization of sub- problem at each iteration.

Find largest & smallest values of G^T G.

Operationally, the largest & smallest Eigen values of the problems are multiplied by 10 & 0.1 respectively. Giving λ_l & λ_s that are used to determine the coefficients of a parabola from which ten samples the auxiliary factors λ_k are obtained using the formula (Meju, 1992)

$$\lambda_k = \left(\{100\lambda_s - \lambda_l\} + \{\lambda_l - \lambda_s\} k^2 \Big/ 99 \right)$$

$$where, k = 1.....10$$

$$hence,$$

$$\beta_k = \lambda_k^2$$

β_k is required in a line search procedure.

Algorithm by Meju (1992): / Ridge- regression:

Step:

1. Select a starting model m_0

2. Calculate

$$Misfit, q_i^0 = \sum_{i=1}^{n} (w_y)_i^2$$
$$where, w_y = wd - wf(m)$$

3. If the misfit function is satisfied, stop the program.

4. Obtain the weighted partial derivative WG = G_*

5. Calculate SVD of G_*

6. Obtain most feasible solution for the line search

a. Calculate damping factors for the

$$\lambda_k = \left(\{100\lambda_s - \lambda_l\} + \{\lambda_l - \lambda_s\} k^2 \Big/ 99 \right)$$
$$where, k = 1.....10$$
$$Hence,$$
$$\beta_k = \lambda_k^2$$

Where λ_s & λ_l are the smallest and largest singular values of G_* , multiplied respectively by 0.1 & 10.

b. Set $\beta_0 = 0$

c. perform line search with ridge regression.

Loop (j =1 to 11 & k = 11-j)

Get,

$$Q_i^{-1} = \frac{Q_i}{(Q_i + \beta_k)^2}$$
$$i = 1, p$$
$$calculate, m_j = nm_0 + LQ_d^{-1}U^T y_*$$
$$compute, q_1^j = \sum_{i=1}^{n} \left| wd - wf(m_j) \right|^2$$

$$If, \left(q_1^j > q_1^{j-1}\right)$$

Set optimal solution to $m_j - 1$quit else.

Set optimal solution to m_j .

end loop

7. Set the optimal model from step 6. As the new iterate (i.e. m_0)

8. Go to step 2.

GOODNESS OF FIT:

Assuming that our data d_i are normally distributed about their expected values and with known uncertainties σ_i (experimental error)

$$q = fit = \sum_{i=1}^{n} \left(\frac{d_i^{obs} - G_{ij} m_{ij}}{\sigma^2} \right)^2, \; j = 1, p$$

$$or, q = \sum_{i-1}^{n} \left\| W d^{obs} - WGm \right\|^2$$

For n – independent observations

p- Independent parameters

q is distributed according to χ^2 (chi –square fit) with (n-p) degrees of freedom.

In geophysical inversion, we rejected or accept the solution to the problem being considered based on the value of q.The expected value of q is say n,

The model is acceptable, (from chi square statistics) with

$$n - p \leq n + \sqrt{2n}$$

If q ≤ n ,–model is said to be over fit.

q ≥ n, – model is said to be under fit.

If experimental error are not available an unbiased estimate of $\sigma^2$is

$$\Delta^2 = \frac{\left(d^T d - m^T G^T G m\right)}{n - p}$$

$$\equiv \quad \frac{sum\ of\ square\ of\ residuals}{n-p}$$

$$RMS = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(d_i^{obs} - G_{ij}m_{ij}\right)^2}{\sigma^2}$$

$$RMS = \frac{1}{n}\sum_{i=1}^{n}\left\|Wd^{obs} - WGm\right\|^2$$

for weighted solution

$$q = (d - Gm)^T (d - Gm) + \beta^2 (Dm - h)^T (Dm - h)$$

$$q = (d^T - m^T G^T)(d - Gm) + \beta^2 (m^T D^T - h^T)(Dm - h)$$

$$q = (d^T d^T - d^T Gm - m^T G^T d + m^T G^T Gm) + \beta^2 (m^T D^T Dm - m^T D^T h$$

$$- h^T Dm + h^T h)$$

$$\left(\frac{\partial q}{\partial m^T}\right) = 0 - 0 - G^T d + G^T Gm + \beta^2 D^T Dm - \beta^2 D^T h - 0 + 0$$

$$Set, \left(\frac{\partial q}{\partial m^T}\right) = 0$$

$$G^T Gm + \beta^2 D^T Dm = G^T d + \beta^2 D^T h$$

$$(G^T G + \beta^2 D^T D)m = G^T d + \beta^2 D^T h$$

$(G^T G + \beta^2 D^T D)$ – Normal equations

Or if D is identity matrix

$$(G^T G + \beta^2 I)m = (G^T d + \beta^2 h)$$

$(G^T d + \beta^2 h)$ - Normal equations

$$m^{est} = (G^T G + \beta^2 I)^{-1}(G^T d + \beta^2 h)$$

This is constrained linear inversion formula. It is also known as biased linear estimation technique.

INVERSION WITH PRIOR INFORMATION

These previous information obtained the sought model parameter can be incorporated in our present formulation. The previous information may be the information from the previous experiments /quantified expectations dictated by physics of the problem .Generally these external data help to single out a unique solution from among all equivalent ones .The constraining equations (data) are arranged to form an expression of form

Dm = h

Where D is a matrix (with all the off-diagonal element equal to zero) that operates on the model parameters m to yield or preserve the prior values of m that are contained in the vector h.

Dm=h means that we are employing linear equality constraints that are to be satisfied exactly .The mathematical development is straight forward .we wish to bias m_j towards h_j .

APPLICATION OF CONTRAINED INVERSION

Fitting a straight line problem

$d_i = m_1 + m_2 x_i$

In the form d = Gm, where m =$(m_1, m_2)^T$

Data pairs ($\{x_i, t_i\}$, i=1, 2.....n)

The a prior information is the t_i fitted line should pass through (x_c, t_i) . So we have one constraint (you may impose a number of constraints on the problem).

Dm = h takes the form,

$$\begin{bmatrix}1 & x_c\end{bmatrix}\begin{bmatrix}m_1\\ m_2\end{bmatrix} = \begin{bmatrix}t_c\end{bmatrix}$$

$$\begin{bmatrix}1 & x_c\end{bmatrix} = D, \begin{bmatrix}m_1\\ m_2\end{bmatrix} = m, \begin{bmatrix}t_c\end{bmatrix} = n$$

Let β= 1.0

$$(G^T G + \beta^2 I) = \begin{bmatrix} n & \sum x_i & 1 \\ \sum x_i & \sum x_i^2 & x_c \\ 1 & x_c & 0 \end{bmatrix} \leftarrow augmenting\ equations$$

$$(G^T d + \beta^2 H) = \begin{bmatrix} \sum t_i \\ \sum x_i t_i \\ . \\ . \\ t_c \end{bmatrix} \leftarrow augmentary\ equations$$

FORMULATING CONSTRAINING EQUATIONS

The equations Dm = h in the general form,

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ . \\ m_p \end{bmatrix} = \begin{bmatrix} n_1 \\ n_2 \\ . \\ . \\ n_p \end{bmatrix}$$

One parameter is known. Then we can modify it as,

$$\begin{bmatrix} 1 & 0 & 0.... & 0... & 0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ . \\ . \\ m_p \end{bmatrix} = \begin{bmatrix} h_{known} \end{bmatrix}$$

$\leftarrow just\ a\ number\ not\ a\ vector.$

If two parameters (say, first & fourth) values are known,

$$\begin{bmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} = \begin{bmatrix} h_1 \\ 0 \\ 0 \\ h_4 \end{bmatrix}$$

Operationally, we need to add the row [1 0 ... 0] on to add the bottom of the G matrix and the known values of the parameters $[h_{khown}]$ onto the bottom of the actual field data d.

Where desired, both d and h are multiplied by β (usually chosen to be less than or equal to unity.)

The constrained least square solution for the straight line through (x_c, t_c) is therefore,

$$m_c = \begin{bmatrix} m_1 \\ m_2 \\ \beta \end{bmatrix} = \begin{bmatrix} n & \sum x_i & 1 \\ \sum x_i & \sum x_i^2 & x_c \\ 1 & x_c & 0 \end{bmatrix}^{-1} \begin{bmatrix} \sum t_i \\ \sum x_i t_i \\ t_c \end{bmatrix}$$

Let, x_c = 8

y_c =14.9

Example: 1. Constrained seismic refraction time –term analysis.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \gamma_1 \\ \lambda \quad \gamma_2 \\ \gamma_3 \\ 1/v_1 \end{bmatrix} = [\delta_1] = 0.433$$

$$G = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 6.708 \\ 1 & 0 & 0 & 0 & 1 & 8.485 \\ 0 & 1 & 1 & 0 & 0 & 7.616 \\ 0 & 1 & 0 & 1 & 0 & 7.0 \\ 0 & 1 & 0 & 0 & 1 & 7.616 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \leftarrow \beta D$$

$$d = \begin{bmatrix} 2.323 \\ 2.543 \\ 2.857 \\ 2.640 \\ 2.529 \\ 2.553 \\ 0.433 \end{bmatrix} \leftarrow \beta h$$

INVERSION WITH SMOOTHNESS MEASURES

Suppose we do not have any known estimates of delay time .The simplest & cheapest remedy for such problems or a prescription for indeterminably or non – uniqueness in inversion,

SMOOTH ← $remedy\ for\ non\ uniqueness.$

Problem Formulation:

if it is desired that the model parameters vary slowly with positions say, we may choose to minimize the difference between physically adjacent parameters $(m_1 - m_2)$,$(m_2 - m_3)$... $(m_{p-1} - m_p)$

These first differences can be written as constraining equations in the form Dm =h

$$\begin{bmatrix} 1 & -1 & . & . & . & . \\ & 1 & -1 & & & \\ & & . & . & . & \\ & & & . & 1 & -1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ . \\ m_p \end{bmatrix} = \begin{bmatrix} 0 \\ . \\ . \\ 0 \end{bmatrix}$$

D m h

D =Flatness/ smoothness matrix of solution vector m

If the model parameters do not vary smoothly with position, then the use of constraining equations of the form

$$\begin{bmatrix} 1 & \cdots & \cdots\cdots \\ \cdots & 1 & \cdots\cdots \\ & & 1 \quad \cdots \\ \cdots & \cdots & \cdots \quad 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

D m h

is recommended .Then D→identing matrix. Then it is biased estimation with non – information .The operation effectively damps the length of the solution (by forcing it into conforming to h) leading to the stable inverse process.

The quadratic $q_2(m)$ is given by,

$$q_2(m) = (Dm - h)^T (dm - h) = (m^T D^T - h^T)(Dm - h)$$
$$q_2(m) = m^T D^T Dm - m^T D^T h - h^T Dm + h^T h$$
$$\cong m^T D^T Dm$$
$$\cong m^T Hm$$
$$where, H = D^T D$$
$$q = (d - Gm)^T (d - Gm) + \beta^2 (m^T D^T Dm)$$
$$\therefore, q = q_1 + \beta^2 q_2$$

We minimize q .

Example :

$$G = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 6.708 \\ 1 & 0 & 0 & 0 & 1 & 8.485 \\ 0 & 1 & 1 & 0 & 0 & 7.616 \\ 0 & 1 & 0 & 1 & 0 & 7.0 \\ 0 & 1 & 0 & 0 & 1 & 7.616 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & -0.01 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.01 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & -0.01 \end{bmatrix} \} \beta D$$

$$d = \begin{bmatrix} 2.322 \\ 2.543 \\ 2.857 \\ 2.640 \\ 2.529 \\ 2.553 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix} \} \beta h$$

Input data structure with first difference operator for β= 0.01.

Error/Bounds on the parameter estimates:

This is an important aspect of geophysical data analysis (or interpretation) that gives the determination of bounds(confidence limits) on the various model parameters that are consistent with the experimental data and their associated error.

Parameter Covariance matrix:

The error bound/ confidence limit can be calculated from covariance matrix, cov(m). Covariance matrix depends on the co-variance of the experimental errors and the way in which we map the data errors into parameter errors(Menke, 1984).

The least square solution

$$m^{est} = (G^T G)^{-1} G^T d = G^{-g} d$$

G^{-g} = Generalised inverse

Expectation (E) of m^{est}

$$E(m^{est}) = E(G^{-g} d) = G^{-g} E(d)$$

If the experimental data are uncorrelated and of equal variance and usind law of propagation of errors

$$Cov(m^{est}) = G^{-g}[Cov(d)(G^{-g})^T]$$

$$Cov(m^{est}) = \{(G^T G)^{-1} G^T\}\{\sigma^2 I\}\{(G^T G)^{-1} G^T\}^T$$

$$= \{(G^T G)^{-1} G^T\}\{\sigma^2 I\}\{G(G^T G)^{-T}\}\rho_a(s)$$

$$= \{(G^T G)^{-1}\}\{\sigma^2 I\}\{\frac{G}{(G^T G)^T}\}$$

$$= \{(G^T G)^{-1} G^T\}\{\sigma^2 I\}\{\frac{G}{(G^T G)}\}$$

$$= \{(G^T G)^{-1} G^T\}\{\sigma^2 I\}\{G(G^T G)^{-1}\}$$

$$= \sigma^2 (G^T G)^{-1} \rho_a(s)$$

For Marquart-type damped least square solution,

$$m^{est} = (G^T G + \beta^2 I)^{-1} G^T d = G^{-g} d$$

$$Con(m^{est}) = (G^T G + \beta^2 I)^{-1} G^T [\sigma^2 I][(G^T G + \beta^2 I)^{-1} G^T]$$

$$= \sigma^2 (G^T G + \beta^2 I)^{-1} G^T G (G^T G + \beta^2 I)^{-1}$$

Noting:

$$m_c = \{(G^T G + \beta^2 D^T D)^{-1} G^T\} d + \{(G^T G + \beta^2 D^T D)^{-1} \beta D^T\} \beta h$$

Parameter resolution

$$R = (G^T G + \beta^2 D^T D)^{-1} G^T G + (G^T G + \beta^2 D^T D)^{-1} \beta D^T \beta D = I$$

Incorporating a priori information:

$$Cov(m^{est}) = (G^T G + \beta^2 D^T D)^{-1} G^T (\sigma^2 I) G (G^T G + \beta^2 D^T D)^{-1} +$$
$$(G^T G + \beta^2 D^T D)^{-1} \beta^2 D^T (\beta^2 I)^{-1} \beta D (G^T G + \beta^2 D^T D)^{-1} \rho_a(s)$$

Or

$$Cov(m^{est}) = (G^T G + \beta^2 D^T D)^{-1} \{\sigma^2 G^T G + D^T D)^{-1} + D^T D\} \{G^T G + \beta^2 D^T D)^{-1}$$

Interpretation:

Cov(m) is a parameter by parameter matrix whose ith diagonal elements is the statistical variance of the ith parameter mi, and whose off-diagonal elements indicates the correlation between the model parameters. Large off-diagonal elements cov ij mean that ith and jth model parameters are highly correlated. The square roots of the diagonal elements of cov(m) are generally referred to as the standard deviations of the least-squares parameter estimates and may be used to estimate the bounds of the model parameters.

Bayes' theorem:

The conditional probability

$$P(B/A) = \frac{P(AB)}{P(A)}$$

$$P(AB) = P(BA)$$

The intersection of A and B be the same as the intersection of B and A.

$$P(AB) = P(B/A)P(A) = P(BA) = P(A/B)P(B)$$

So, we have the following relations between the two different conditional probabilities.

$P(A/B)$ & $P(B/A)$

$$P(B/A) = \frac{P(A/B).P(B)}{P(A)}$$

Or

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Posterior distribution = $\frac{likelihood \times priordist}{Scale}$

N.B:

Probability is fundamentally about measuring sets. The sets can be finite.

The space of all possible outcomes of a given experiment is called "sample space" denoted by Ω.

Sample space={1,2,3,4,5,6}

Equally likely

The probability associated with A we call

$$P(A) = \frac{N(A)}{N(\Omega)} = \frac{1}{6}$$

= No of possible way of achieving event A/size of sample space

Random Variables:

A variable is denoted the outcome of a random trial.

A given outcome of a random trial is called a realization.

Let d denote the outcome of a flip of a fair coin, d is a random with two possible heads and tails.

Solution of a general inverse problem in Bayesian framework(Tarantola and Valette , 1982):

According to Bayes theorem:

$$\sigma(d,m) = K\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Where, K= normalization constant

$\rho(d,m) =$ Priori knowledge on data d and model parameter m(prior)

$\theta(d,m) =$ Physical theory relating model parameters m to the observable parameters d, (likelihood)

$\mu(d,m) =$ Objective reference state of minimum information (Scale)

The solution of general inverse problem is given by marginal posterior distribution

$$\sigma(m) = \int_D \sigma(d,m)dd$$

Which is the classical Bayesian framework is a conditional probability density conditional on the observed data (Tarantola,2005)

Let $d = g(m) + t$

$d =$ data

$t =$ Gaussian noise

$g(m) =$ function

So, conditional probability distribution of d given m is Gaussian

$$P(d/m) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\exp(-\frac{((d-g(m))^2}{2\sigma^2}$$

Where g(m) = mean

$\sigma =$ standard deviation of E

The joint probability density may be decomposed to be Gaussian.

$$P(d) = \frac{1}{(2\pi)^{\frac{1}{2}}\left|C_D\right|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(d_{obs}-d)^T C_D{}^{-1}(d_{obs}-d)\}$$

Where C= dimensionality of d

$d_{obs} =$ observation data

$d =$ mean value of data/noise less data

$C_D =$ Covariance matrix

C_D = to be diagonal for uncorrelated noise

Likelihood:

$\theta(d,m) = \delta(d - G(m)).\mu(m)$

Where δ = the delta function

G= non-linear forward operator

$\mu(m)$ = Scale

So, $\sigma(m)$ = solution to the inverse problem

$$= Kp(m).\exp\{-\frac{1}{2}(d_{obs} - G(m))^T C_D^{-1}(d_{obs} - G(m))\}$$

If d represents a set of distinct measurement m is the set of model parameter

In Bayesian framework, we can write

$$P(m/d) = \frac{P(d,m)P(m)}{P(d)}$$

P()= probability

P(d/m)=pdf of observed data given the model m (likelihood)

P(d)=pdf of data d (scale factor in inversion represents limitation on data space imposed by the physics and prior constants on the model space.

P(m)=pdf of the model parameter, m independent of the data (prior information on model)

The solution of inverse problem may be approximated by the sampling based method according to p(m,d).

Answer is given by constraining some model vector m. Model m might represent distribution of fluids in a sub-surface reservoir.

The examples of the type of pertinent background knowledge might include

1) Knowledge of regional geologic, tectonic, inferencesfrom previously acquired data; typical problems experienced when collecting data.

We represent all such background knowledge by symbol B

Using Bayes rule, we can incorporate information B such that

$$P(m/d,B) = \frac{P(d/m,B)}{P(d/B)}P(m/B)$$

P(m/d, B)= posterior distribution

$\frac{P(d/m,B)}{P(d/B)}$ = Relative likelihood, is the ratio of the probability of measuring data d if model m was true(numerator) to the probability when no knowledge about m is available(denominator)

The second term on the right hand side is called the prior distribution of model m, and describes all information about m that existed prior to data d having been acquired(since data d does not features in this term).

$m^2 = (G^T G + \lambda D_1^T D_1)^{-1} G^T d$

λ is termed as regularization(Tikhonov, Arsenin 1977)

λ depends on error in the observed data.

If λ =1 –It is called generalised least-square solution(Tarantola and Vallette 1982)

The role of prior information:

An underdetermined inverse problem requires regularization. The idea is to select a particular suite of solutions from infinitely possible solutions that honour the data. For example, in case of fitting line to a single datum, it might be known a priori that the straight line passes through a known point, say the origin such as a priori information is enough to solve the problem. Incorporation of a priori knowledge about the model space help in reducing the domain of the solution such that a solution is obtained that belong to a particular class of solutions that fit the data and at the same time honour the a priori information.

Say, we would like to answer a geoscientific question upon which geological information might have some bearing and assume further that the question can be posed in such a way that its.

Hence, this term includes all information about m from background information B alone.

Priori information concerns expectations that the model parameters possess in the given range or possess a given sign.

For example, even without making any measurements one can state with certainty that density is everywhere positive, Since density inherently is a positive quantity.If one used prior information when solving inverse problem, it may greatly reduce the range of possible solutions or even cause the solution to be unique.

From Bayes to weighted least-squares:

Denote by f(m,d) the joint distribution on models and data, using Bayes theory, the conditional probability on m given d is

$$P(m/d) = \frac{f(d.m)\,p(m)}{h(d)}$$

Where f(d/m) measures how well a model fits the data, p(m) is the prior model distribution, h(d) is the marginal density of d. The conditional probability p(m/d) is the so-called Bayesian posterior probability, expressing the idea that p(m/d) assimilates the data and prior information.

We assume that all uncertainties (model and data) can be described by Gaussian distributions. Since Gaussian distribution can be characterized by its mean and covariance, this means that we must specify a mean and covariance for both the a priori distribution and the sata uncertainties.

Bayesian posterior probability is the normalized product of the following two functions

$$\sqrt{\frac{(2\pi)^{-n}}{\det C_D}}\exp[-\frac{1}{2}(G(m)-d_{obs})^T C_D^{-1}(G(m)-d_{obs})$$

d_{obs} is the vector of observed data which dimension is n, C_D is the data covariance matrix , G_m is the forward operator and

$$\sqrt{\frac{(2\pi)^{-m}}{derC_m}}\exp[-\frac{1}{2}(m-m_{prior})^T C_m^{-1}(m-m_{priori})]$$

Where m is the number of model parameters and C_m is the covariance matrix describing the distribution of models about the priori mode m_{priori}.

If the forward operator is linear, then the posterior distribution is itself a Gaussian. If the forward operator is non-linear, then the posterior is non-Gaussian.

If the forward operator is linear, normalized product

$$\sigma(m)\alpha\exp-\frac{1}{2}[(G(m)-d_{obs})^T C_D^{-1}(G(m)-d_{obs})+(m-m_{priori})^T C_m^{-1}(m-m_{priori})]$$

Can be written as

$$\sigma(m)\alpha\exp[(m-m_{map})C_m^{-1}(m-m_{map})]$$

Where

$$C_{M'} = [G^T C_D^{-1} G + C_m^{-1}]^{-1}$$

Is the covariance matrix of the posterior probability. This approximation is true even when g is non-linear(weakly)

m_{map} is the maximum of the posterior distribution, which for a Gaussian is also the mean.

So to find out estimator are need to optimize the following

$$\min[(G(m)-d_{obs})^T C_D^{-1}(G(m)-d_{obs})+(m-m_{prior})C_m^{-1}(m-m_{prior})]$$

This is nothing but weighted least square problem.

$$((C_D^{-\frac{1}{2}})^T(G(m)-d_{obs})C_D(C(m)-d_{obs}))=\left\|C_D^{-\frac{1}{2}}(G(m)-d_{obs})\right\|^2$$

While the second term is

$$\left\|C_m^{-\frac{1}{2}}(m-m_{prior})\right\|^2$$

Occam's inversion:

The general regularized solution to find out model through solving the optimization

$$\min\{J(m)\}=\min\{J_d(m)+\lambda J_m(m)\}$$

Occam inversion involves

$$\min\{J_m(m)\}J_d(m)=X_d$$

Where X_d is the desired tolerance of $J_d(m)$

The basic motivation for seeking the smoothest model is that one does not wash to be misled by features that appear in the model but are not essential in matching the noise contaminated observations. In other words, of all the possible solutions we seek the simplest(smoothest) one in the sense that it requires the least spurious features not required by the data. This is the principle of parsimony(Occam's razor).

Occam Factor:

Assume two different models m1 and m2 of similar type but where the model m2 has more parameter than m1 using Bayes theorem , posterior probability of each model

$$p(m_i/d)=\frac{p(d/m_i)p(m_i)}{p(d)}$$

Where d=data

M= model

The denominator does not depend on the model and can be ignored for the purpose of model comparison. If there is no reason to prefer m1 or m2, the priors p(mi) should be same, and model can be compared on the basis of likelihood function $p(d/m_i)$ which can be written in the form

$$p(d/m_i) = \int p(d/w, m_i)\, p(w/d, m_i)\, dw$$

Where $p(d/w, m_i)$ = Likelihood of the data

$p(w/d, m_i)$ = Posterior distribution of model parameters

W=model parameters

Now, let us consider a single model parameter w. If the posterior distribution of the parameter is sharply peaked in model parameter space around the most probable parameter space around the most probable value W_{MP}, then the integral can be approximately by the value at the maximum times the width $\Delta w_{posterior}$ of the peak.

$$p(d/m_i) = p(d/w_{mp}, m_i)\, p(w_{mp}/m_i)\, \Delta w_{posterior}$$

The prior probability $p(w_{mp}/m_i)$ is taken to be uniform over some large region Δw_{prior}, when the data arrives, this collapses to a posterior distribution $p(w/d, m_i)$ with a width $\Delta w_{posterior}$. The ratio $\Delta w_{posterior} / \Delta w_{prior}$ represents a factor which penalizes the model for having particular posterior distribution of model parameter.

As the model parameter prior $p(w_{mp}/m_i)$ is considered to be uniform over some large interval Δw_{prior} then

$$p(d/m_i) = p(d/w_{mp}, m_i)\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

The first term on the right hand side is the likelihood of the data evaluated at the most probable model parameter values, while the second term known as Occam factor(with value<1) penalized the model for having this particular posterior distribution of model parameters.

If the model is very complex the term(Occam factor) will be very small as well. That is, it penalizes m2 more than m1.

Bayesian Inversion:

In this framework the solution to an inverse problem is represented by a posterior (post inversion) probability distribution function (pdf), $\sigma(m)$, over the model parameters m(eg. Tarantola 2005.

The posterior pdf is related to the so-called prior pdf p(m) which reprents pre-inversion knowledge about parameters m through

$$\sigma(m) = CL(m)\,p(m)$$

Where C= Constant

L(m)= Likelihood function which describes how well the synthetic or modelled data corresponding to each model m match the recorded data and here is defined to be a multivariate Gaussian in the difference between observed data d^i_{obs} and synthetic data $d^i(m)$

$$L^i(m) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp(-\frac{1}{2}\varepsilon^T\sum\varepsilon^{-1})$$

Where Σ is the covariance and ε is the error vector difference between observed data and synthetic data.

Backus-Gilbert method:

Backus and Gilbert [1968, 1970] formulated an approach to the inverse problem that is especially well-suited to problems in geophysics and imaging of the earth.

The mathematical formulation of their approach assumes that the discrete data di are related to an Earth model m(r) and some set of function Gi(r) that characterize the interaction between the earth to an integral relation of the form

$$\int_0^1 G_i(r)m(r)dr = d_i \text{ for i= 1... N}$$

The function Gi(r) are assumed to be known and result from a forward computation using the "good" earth model.

Backus and Gilbert introduce a function $A(r, r_0)$ with the desired properties that

$$\int_0^1 A(r, r_0)dr = 1$$

And $\int_0^1 A(r, r_0)m(r)dr = m(r_0)$

So, A is intended to act much like a dirac delta function $\delta(r - r_0)$. The point of the method is to try to construct such an A from the known functions Gi in the form of an expansion,

$$A(r, r_0) = \sum_{i=1}^{N} a_i(r_0) G_i(r)$$

We must have one constraint on the ai's

$$\sum_{i=1}^{N} a_i(r_0) \int_0^1 G_i(r) dr = 1$$

And an approximate formula for the Earth model in terms of the di's

$$\sum_{i=1}^{N} a_i(r_0) d_i = I_m(r_0)$$

Making use of the idea that A should resemble a delta function, Backus and Gilbert introduce a spread function

$$S(A, r_0) = 12 \int_0^1 (r - r_0)^2 A^2(r; r_0) dr$$

If A is a delta function $\delta(r - r_0)$, this integral vanishes. So, S is a measure of the deviation of A from a delta function.

$$S(A; r_0) = \sum_{i=1}^{M} \sum_{j=1}^{N} a_i(r_0) a_j(r_0) K_{ij}(r_0) = a^T K a$$

Where matrix $K_{ij}(r_0) = 12 \int_0^1 (r - r_0)^2 G_i(r) G_j(r) dr$

If we define $C_i = \int_0^1 G_i(r) dr$

We can write $a(r_0) = (C^T K^{-1}(r_0) C)^{-1} K^{-1}(r_o) C$

K,a,C are matrix, we assume that K is invertable.

Local Optimization :

1-Stepest descent method

2-Conjugatw gradient method.

3- Nonlinear conjugate gradient method.

4-Newton's method.

5-Marquardt method

Global Optimization

1-Monte carlo method

2-Genetic algorithm

3-Simulated annealing method

Monte Carlo Method:

The phrase" Monte carlo method" was first used by metrolis and Ulam(1949). During American civil war, numerical procedure was used to determine the value of π by injured officers.

This procedure consisted of throwing a needler on to board containing parallel straight lines. The statistics of number of number of times the needle intersected each time could be used to estimate π . The usefulness of Monte carlo type of numerical experiments was therefore known well before the beginning of the century; However there systematic development and widespread use had to wait for the arrival of the electronic computer.

The idea behind the metropolis-Hastings algorithm and the Gibbs sampler is the same. They are both so-called Markev chain Monte-carlo algorithm. The algorithm allows us to calculate the values of p at a given point in the space is available. In geophysics p is a probability density derived from a misfit function J through , for example,

$$p(m_k) = A\exp(-BJ(mk))$$

Where mk is a model and A and B are constants.We wish to sample a probability distribution p in a discretized model space M. Sampling from the distribution p means that the probability of visiting model m is proportional to p(m). We assume the following criteria to sample from target distribution.

1. The probability of visiting a point mi in model space, given that the algorithm currently is at point mj depends only on mj and do not on previously visited oints. This is called Markov-property. This property means the algorithm is completely described by a transition probability matrix pij whose ijth component is the conditional probability of going to point mi, given the algorithm currently visits mj.
2. To sample target distribution , an algorithm that satisfy microscopic reversibility.

   $P_{ij}\,p(m_j) = P_{ji}\,p(m_i)$

   Where $P_{ij} =$ transition probability matrix.

   $P_{ij}(m_j) =$ Probability that a transition from mj to mi.

   $P_{ji}\,p(m_i) =$ Probability that a transition from mi to mj.

   Microscopic reversibility means that the probability of these two transitions is the same at all times and the each pair of points in M maintains mutual equilibrium.

3. In Metropolis-Hasting algorithm the transition probability pij are given b

$$P_{ij} = \frac{1}{N}\min(1, \frac{p(m_i)}{p(m_j)}$$

Assume that current point visited by the algorithm id mj. We choose (or, rather propose) one of its N neighbours mi with probability

$$P_{proposal} = \frac{1}{N}$$

Finally, we accept mi only with probability

$$P_{accept} = \min(1, \frac{p(m_i)}{p(m_j)})$$

If mi is accepted, the algorithm goes to mi in this iteration, but if mi is rejected, the algorithm stays in mj(mj is visited once again).

It is possible for algorithm to go from any point mj to any other point mi, give enough steps. An algorithm satisfying the property is called irreducible .

So, <u>In Bayesian Framework:</u>

Posterior probability density function(pdf)

$\sigma(m) = CL(m)\,p(m)$

Where C= is a constant

L(m)= likelihood function which describes how well the synthetic or model data corresponds to each model m match. The reconded data

$$L^i(m) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp(-\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon$$

Where Σ is the covariance matrix.

$\varepsilon = d_{obs}{}^i - d_{syn}{}^i(m)$

<u>Steps:</u>

1.Start with initial values for mj and calculate corresponding geophysical parameters d(mj) by forward modelling calculate likelihood L(mj).

2.Define a new candidate parameter vector m_{j+1} by randomly selecting candidate geophysical values from the prior distribution.

Calculate the corresponding geophysical parameters $d(m_{j+1})$ and likelihood $L(m_{j+1})$.

3.Use the Metropolis rule to accept or reject the new candidate model by calculating the ratio of the current and candidate likelihood $\frac{L(m_{j+1})}{L(m_i)}$

The acceptance probability is $P = \min[1, \frac{L(m_{j+1})}{L(m_j)}]$

4.If the candidate model configuration is rejected, the current model remains for the next iteration, otherwise m_{j+1} is accepted as the next model sample.

5.Repeat 2 to 3 untill the required number of samples in the set S={m1,m2 ...mN} are obtained.

6.By calculating sample density in S, we obtain an estimate of the posterior pdf $\sigma(m)$ for any new measured data d_{obs} .

Simulated Annealing:

It is an empirical fact that the process of chemical annealing, where a crystalline material is slowly cooled through its melting point results in formation of highly ordered low energy crystals. The slower the coolong the more perfect is the crystal growth and the lower is the lattice energy. This process can be viewed as a physical optimization method, which the objective function is the lattice energy J.

The algorithm runs as follows:

1.In each step a random perturbation of the model parameters mj of the numerical system is attempted.

The new set of model parameters mi are accepted if the value of the objective function J decreases. However if J increases, the new parameters may be accepted with probability

$$P_{accept} = \exp(-\frac{\Delta J}{T})$$

Where $\Delta J =$ Change in the objective function.

If the new model is rejected, a new perturbation is attended in the next move, and the above process of decision is repeated.

2.A close inspection of the above algorithm reveals that for constant temperature parameter T , it is actually a Metropolis-Hasting algorithm designed to sample the probability distribution.

$$P_B(m) = \frac{\exp(-\frac{J(m)}{T})}{Z(T)}$$

Which is known in statistical physics as the Gibbs-Boltzmann distribution. Here $\frac{1}{Z(T)}$ is a normalization constant.

In SA, the temperature parameter is gradually decreased from a high value;allowing large thermal fluctuation, down to zero, where only decreasing values of the objective function are allowed. For decreasing temperature T the Gibbs-Boltzmann distribution converges towards a distribution having all its probability mass in the global minimum for J. In other words, as the temperature gradually approaches zero, the probability that our system is close to the global minimum for its objective function approaches 1.

Genetic Algorithm:

The genetic algorithm works with a population of Q models simultaneously. Usually the population is initially generated randomly, but at each iteration it is altered by the action of three operators. The fitness (objective function) for each model in the starting population is evaluated by solving the forward problem for each of the Q models. The purpose of the GA is then to seek out fitter models in parameter space.The three operator are known as

(1) Selection (2)Cross-over (3) Mutation

1. Selection:

From the initial population of Q bit-strings an iteration population of Q parents is generated by selecting models from the original group with the probability of selection determined by the fitness value say,

$P(m_k) = A\exp[BJ(m_k)]$

Aand B are problem specific constants.

The higher the fitness of each model, the more likely that it will pass into the next population. Since, the population size does not change, multiple copies of the fitter models will survive at the expense of the less fit models. This operator introduces the element of survival of the fittest into the algorithm.

2.Cross-over:

This operator cuts and mixes pairs of randomly chosen bit strings together. All Q parent strings are randomly paired to produce Q/2 couples. A crossover probability Pc is assigned, and if a random number between 0 and 1 is less than Pc, parts of the two strings are

interchanged. If a crossover is selected, the location at which the strings are cut is determined randomly, otherwise the two parent strings are passed unscatted to the next generation. In this way it is hoped that information is passed on to subsequent generations.

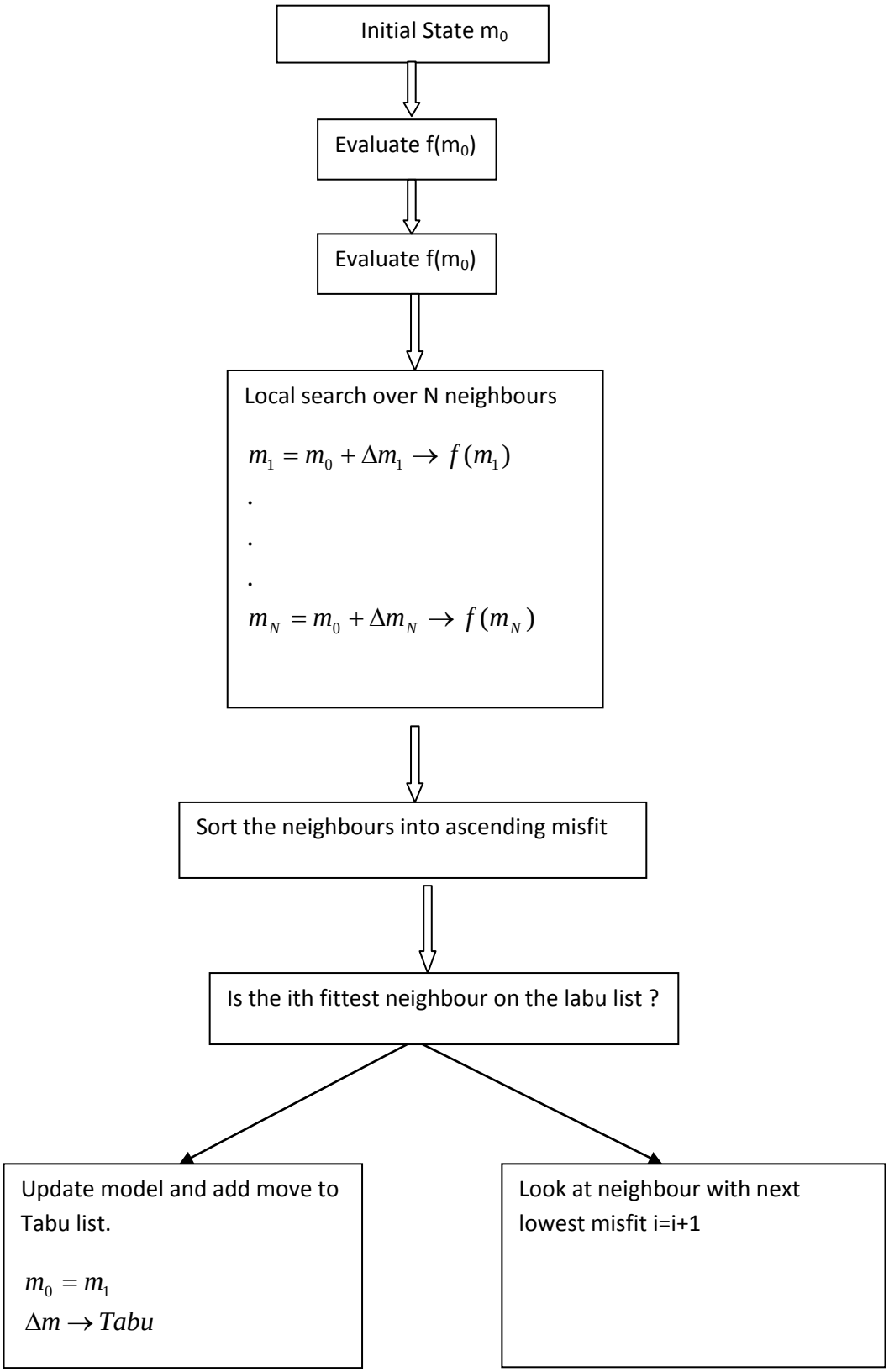
3.Mutation:

The purpose of the mutation operator is to maintain a degree of diversity in the population.

Note that selection operator acts to remove diversity of bit patterns from the population. In mutation, any bit in an individual string is allowed to flip between 0 and 1. This is usually performed with a relatively low permeability Pm.

Overall, the action of the three operator is to produce a new population of models for which the forward problem must be solved. After many iterations, the population has the potential to evolve towards a fitter on averaging state. Each state involves randomly decisions which are influenced by the control parameters. Even in this basic parameters(Q, Pc, Pm, A nad B), which usually must be chosen (termed)for each application.

Tabu Search:


Initial State m_0
Evaluate f(m_0)
Evaluate f(m_0)
Local search over N neighbours
$m_1 = m_0 + \Delta m_1 \rightarrow f(m_1)$
.
.
.
$m_N = m_0 + \Delta m_N \rightarrow f(m_N)$
Sort the neighbours into ascending misfit
Is the ith fittest neighbour on the labu list ?
Update model and add move to Tabu list.
$m_0 = m_1$
$\Delta m \rightarrow Tabu$
Look at neighbour with next lowest misfit i=i+1

Flow chart of general Tabu Search Scheme.

To begin a TS a model is selected at random. A local search is then performed over the adjacent models. From these neighbours the best model is selected, note that this model need not be better than the starting model. This relative move from the starting model to the best neighbouring model is added to the tabu list. From this new model a local search is again performed. Once again the best neighbouring model is selected as a candidate for the next move. However the move must not return the search to the previous starting model. To prevent such cycling the move is checked on the Tabu list of previous moves. If the move is not on the list it is allowed otherwise the next best model is checked for its acceptability.

This checking is repeated until an acceptance neighbour is found which does not lie in a Tabu direction. This process of local search and selection of best non-tabu neighbour is then repeated until some termination criteria is satisfied.

Neural Network:

Neural network form a class of non-linear computational systems that attempt to mimic the natural behaviour of biological neurons. It is popular in geophysics because it can able to map between two domain domain say input and output even from a set of training samples even if there is no deterministic or linear relationship between the two domain. In general architecture of a MLP(Multilayer Perception) consists of one input layer, one output layer and atleast one intermediate hidden layer. Neurons of each layer are connected to the neurons of the next layer through weight.

Sigmoid function

$$f_i(netj) = \frac{e^{\beta(netj)} - e^{-\beta(netj)}}{e^{\beta(netj)} + e^{-\beta(netj)}}$$

E= basis of the natural logarithm.

The principle goal of neural network approach is to learn the relationship between an input and an output in spaced domain from the finite data set.

$$S = \{d_k, x_k; k = 1,2...N\}$$

By adjusting neural network parameter (weight and biases). This is done by maximizing the likelihood of the data set S (or equivalently by minimizing its negative logarithm) which forms a conventional least square error measure in the form

$$E = \frac{1}{2}\sum_{k}^{N}\{x_k - o_k(d_k; w_k)\}$$

Where x_k = target

o_k = Network output

The network function can be sigmoid linear or logistic. The network can be trained on supervised or unsupervised way. In unsupervised (self-organizing map) the bearing took place based on different statistical distance function.

Genetic algorithm:

Genetic algorithm (GA) is based on the analogy that take place in the living species work towards making the species more intelligent and adaptive to the changing natural surroundings. The process of biological evolution is mimicked in genetic algorithm where a pool of possible solutions is updated every iterations so that the pool contain better candidate models as the iterations proceed. The salient features of genetic algorithm are coding, fitness function, selection, cross-over and mutation.

Coding:

Coding is a method to digitally represent the model space. A commonly used coding method is the conversion of numerical model parameters to binary strings. In GA literatures such individual binary strings representing individual model parameters are referred to as the chromosomes. Every bit in such a binary string is referred to as gene. Thus a gene can either be one or zero.

Fitness function:

In GA literature, the function that defines how closely a model fits the data and constraints is referred to as the fitness. As opposed to other optimization algorithms, the aim of the GA is to increase the fitness function . so that the optimum model is the one that has the maximum fitness. Thus, the usual practice is to use either a negative or inverse of a cost function to define the fitness function.

Selection:

Initially the model parameters are randomly selected within a user defined upper and lower bounds. Each random selection is then converted to chromosomes within a user defined resolution. A pool of such possible solution is referred to as the population pool. The next step is selection. The selection is the procedure to update the population pool with fitter models as the algorithm proceeds through generations (the updating of the population pool through iterations is termed as generations).

Selection is done by two procedures a)proportional selection b)tournament selection. In proportional selection method, probability of selection is computed based on the fitness value of each candidate models. The next generation population pool is created by replicating the candidate models of the current pool in direct proportion to their fitness probabilities.

In the tournament selection method, population pool is randomly paired and their fitness values are compared. One, out of the two pairing candidate models is selected depending on a user defined probability value.

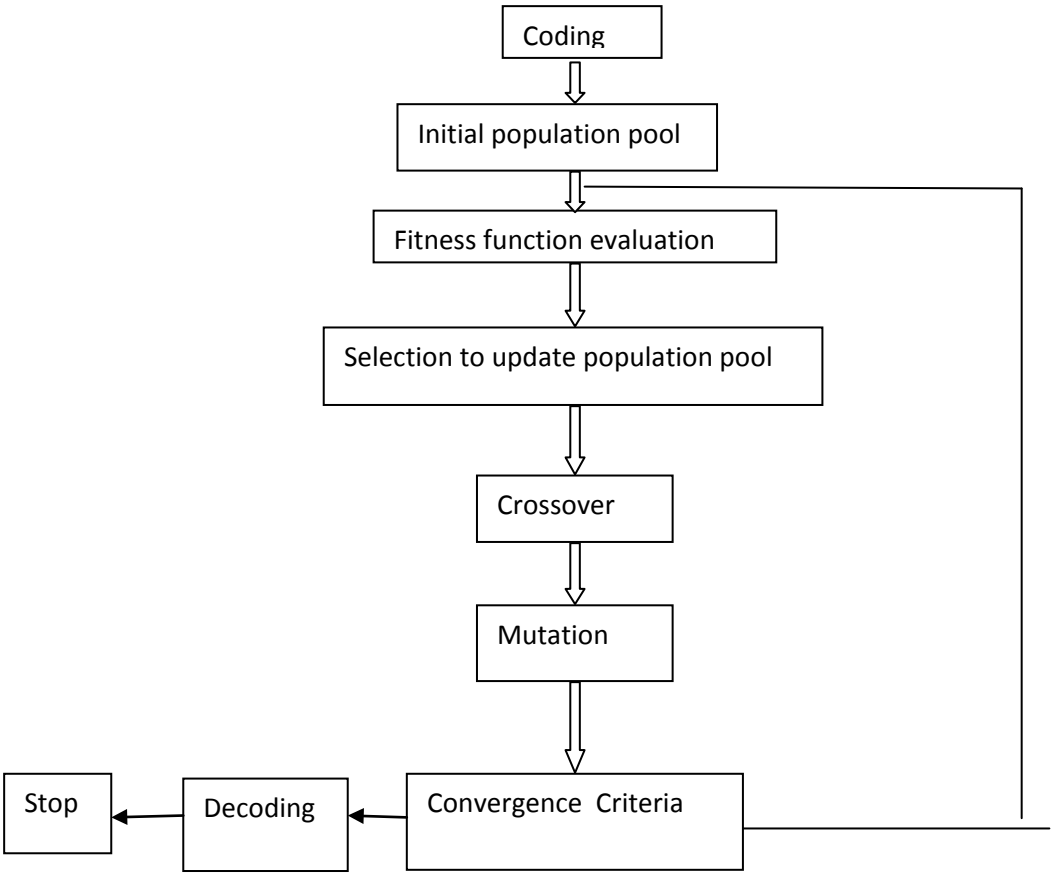
Cross-over:

Cross over operation allows transferring and sharing of genetic formation among chromosomes. Cross over can be single point or multi-point. In a single-point crossover, one point and a pair of concatenated binary strings representing chromosomes is also randomly selected. All binary bits to the right of the randomly chosen point are swapped between the two chromosomes pairs. In case of multipoint crossover, several crossover points are chosen representing each of the concatenated model parameters. All the bits to the right of the chosen crossover points but within the binary string representing a model parameter are swapped between the pairing chromosomes. The number of cross over operations inside a population pool is controlled by the user defined crossover probability.

Mutation:

Mutation mimics the process of genetic mutation where a particular gene undergoes a change. In GA mutation operation is performed by randomly picking a bit from a chromosome (binary string) and changing the binary value. The number of mutation operations is controlled by the mutation probability.

Higher mutation probability means greater diversity. After several generations of updating the population pool, when the convergence is achieved as per the user defined criteria, one candidate model is selected from the optimum population pool. The candidate model is decoded to obtain the numerical values for the unknown model parameters.



Flowchart of Genetic algorithms

Simulated Annealing(SA):

Many approaches of SA are available

1) Metropolis algorithm
2) Heat Bath
3) Fast simulated Annealing (FSA)
4) Very fast simulated annealing (VFSA)

In different approaches the cooling schedule is different for lowing the temperature. The different format for lowering the temperature are

1) $T_k = T_0(0.99)^k$ or $T_0(0.98)^k$
2) $T = \frac{T_0}{k}(FSA)$
3) $T = \frac{T_0}{\ln k}(Heatbata)$

   Here, k= iteration.

   $T_0 =$ initial temperature

1) Metropolis algorithm: Given a starting model mi with energy E(mi), a small perturbation to mi is made to obtain a new model mj by

   $m_j = m_i + \Delta m_i$.....(1)

If ΔE_{ij} is the difference in energy between two states, i.e $\Delta E_{ij} = E(m_j) - E(m_i)$

The acceptance and rejection of new model is depends on the value of ΔE_{ij}. If $\Delta E_{ij} \le 0$, then the new model will be accepted. However if $\Delta E_{ij} > 0$, then the new model will be accepted with the probability

$$P = \exp\left(-\frac{\Delta E_{ij}}{T}\right)$$

ii) Heat Bath Algorithm:

Metropolis algorithm is a two step procedure. (1) First a random move is made (ii) It is decided the move should be accepted or rejected. But at lower temperature, the rejection to acceptance ratio is very high. To overcome this, heat bath algorithm is proposed which attempts to avoid a high rejection to acceptance ratio by computing the relative probability of acceptance of each trial move before any random guesses are made. It is done in one step.

Consider a model vector m consisting of N model parameters. Next we assume that each mi can take M possible values. This is obtained by assigning some lower $\left(m_i^{\min}\right)$ and upper $\left(m_i^{\max}\right)$ bounds and a search increment Δm_i for each model parameter such that

$$M = \frac{m_i^{\max} - m_i^{\min}}{\Delta m_i}$$

Geman and Geman(1984) showed that necessary and sufficient condition for convergence to global minimum level for SA is given by following cooling schedule.

$$T(k) = \frac{T_0}{\ln k}$$

Where T(k)=Temperature at iteration k

To= Sufficiently high starting temperature.

- Heat bath takes 2000to 10000 iterations for convergence. This remains a problem.

Fast Simulated annealing (FSA):

It is found that logarithmic cooling schedule is very slow. In order to overcome this fast simulated annealing is proposed. FSA is very similar to metropolis SA except that unlike metropolis SA, it uses Cauchy-like distribution rather than a flat distribution for the model parameters to generate the models for testing and Cauchy like distribution is also a function of temperature and is given by

$$f(\Delta m_i) \propto \frac{T}{\sqrt{\left(\Delta m_i^{\,2} + T^2\right)}}$$

Where $\Delta m_i =$ Perturbation in model parameter w.r.to the current value of a model parameter.

In this , the cooling schedule required is no longer logarithmic and the temperature schedule is now inversely proportional to the iteration number.

$$T(k) = \frac{T_0}{k}$$

Reference:

Mckenzic and Sclater, 1971

Very Fast simulated Annealing(VFSA):

Ingbar(1989) introduced the concept of model parameter and a Cauchy-like model generation scheme for each of the model parameters. The VFSA algorithm provides a means to regulate the expansion or contraction of the model generationg pdf depending on the sensitivity of the model parameters to the cost function.

Let us consider a particular model parameter m_i^k in the model vector m at an iteration k. Let the upper and lower bounds in the model parameters search space be Bi and Ai such that $A_i \le m_i^k \le B_i$. The new model parameter generated in the (k+1)th iteration is given by

$$m_i^{(k+1)} = m_i^k + y_i(B_i - A_i)$$

Where yi is a random number between [-1,1]. The model parameter $m_i^{(k+1)}$ is generated such that $A_i \le m_i^{(k+1)} \le B_i$. The model generating function for the kth iteration is given by the expression

$$g_T(y) = \prod_{i=1}^{M} \frac{1}{(2|y_i| + T_i)\ln(1+\frac{1}{T_i})}$$

Where M is the model dimension and Ti is the model parameter temperature for the ith mode. The cumulative probability is given by

$$G_{Ti} = \frac{1}{2} + \frac{\text{sgn}(y_i) lm(1+\frac{|y_i|}{T_i})}{2\ln(1+\frac{1}{T_i})}$$

Yi can be written as

$$y_i = \text{sgn}\left(u_i - \frac{1}{2}\right) T_i \left[\left(1+\frac{1}{T}\right)^{|2u-1_i|} - 1\right]$$

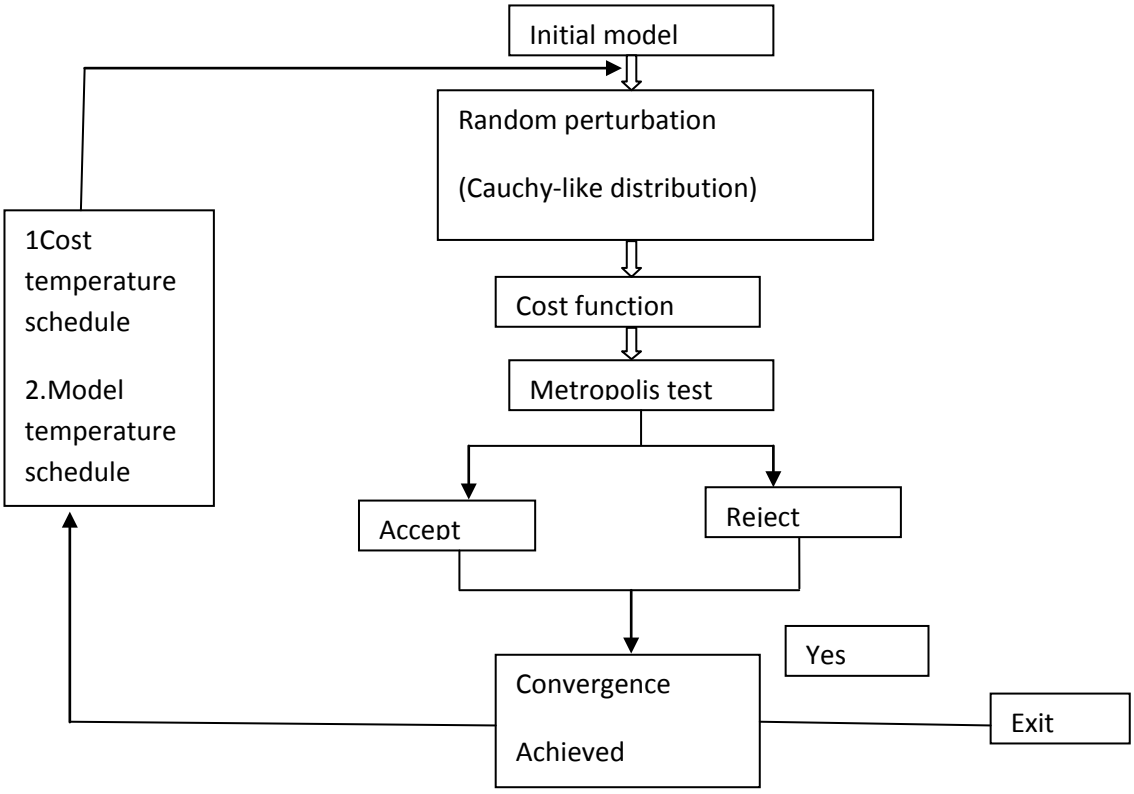
Where $u_i \varepsilon U[0,1]$ is a uniform distribution. With the model generating pdf as defined above, the cooling schedule for the model parameter temperature is given by

$$T_i = T_{0i} e^{-cik^{\frac{1}{M}}}$$

Where ti is the model parameter temperature for the ith model at the kth iteration and ci is a constant that is used to attain a particular final model temperature at a given final iteration. The initial model parameter temperature is given by Toi. The ci is computed by

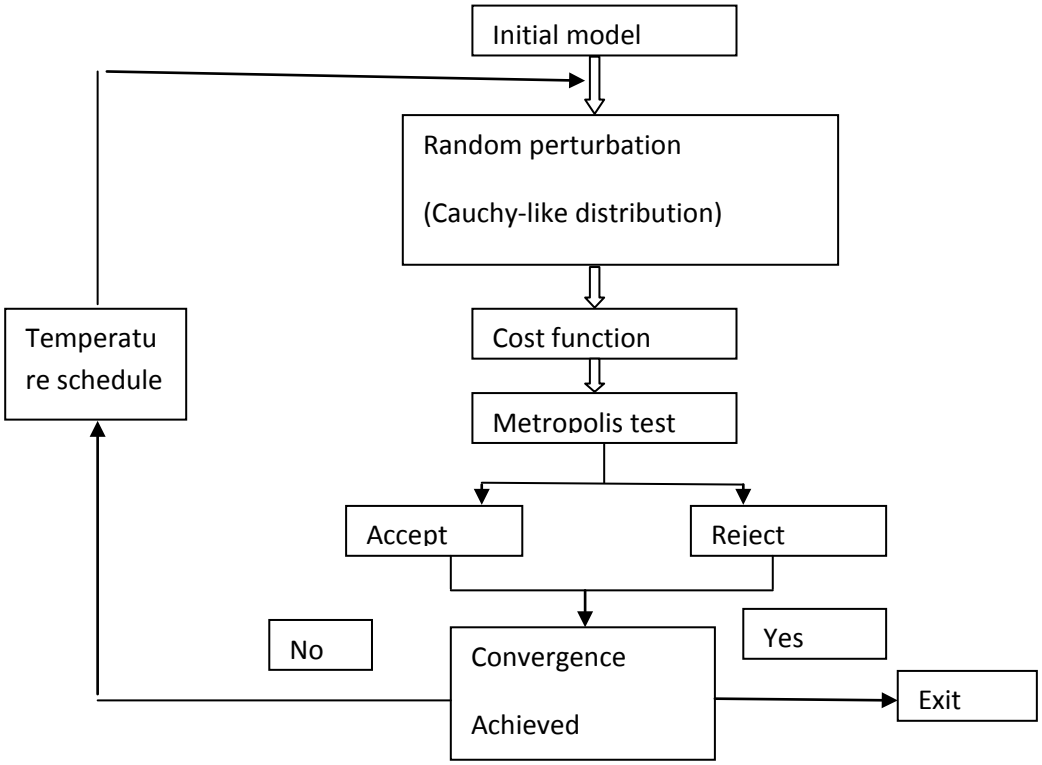
$$c_i = Kfi^{\frac{-1}{M}} \log\frac{T_{0i}}{T_{fi}}$$

Where kfi is the final iteration number for the ith model parameter.



Flow chart for VFSA algorithm

Flowchart of Metropolis algorithm:



DC Resistivity Inversion: 1D case

Forward Modelling:

For schlumberger soundings, the relationship between apparent resistivity ρa and the geological parameters(e.g. layer thickness and true resistivity) is expressed by an integral equation(Koefoed 1970)

$$\rho_a(s) = s^2 \int_0^\infty T(\lambda) J_1(\lambda s) \lambda d\lambda \text{.....(1)}$$

Where S= half the current electrode spacing(AB/2) in the schlumberger electrode configuration.

J1=the first order Bessel function of the first kind and λ denotes the integral variables.

The recurrence relationship of the resistivity transform function for the ith layer Ti(λ)

$$T(\lambda) = \frac{T_{i+1}(\lambda) + \rho_i \tanh(\lambda h_i)}{[1 + T_{i+1}(\lambda) \frac{\tanh(\lambda h_i)}{\rho_i}} \qquad \text{i=n-1, ... 1.....(2)}$$

Where n denotes the number of layers, ρi and hi are the true resistivity and thickness of the ith layer respectively.

Inversion scheme:

The damped least-square inversion solution can be written as (Menke 1920)

$$\Delta m = (G^T G + \beta^2 I)^{-1} G^T \Delta d \text{.....(3)}$$

Where $\Delta m =$ parameter correction vector

$\Delta d =$ data difference vector

G= Jacobian matrix containing partial derivative of data with respect to the initial model parameter.

I=Identity matrix

B=damping factor

Let use say, $G = UQL^T$ .....(4)

Where n=data

M=parameter

$U(n \times m) =$ Orthogonal matrix

$L(m \times m) =$ Orthogonal matrix

$Q = (m \times m) =$ diagonal matrix containing at most r non-zero eigen values of G. The diagonal entities $Q = (\alpha_1, \alpha_2 ... \alpha_p)$ are the singular values of G.

Using SVD, we obtain the parameter correction vector,

$$\Delta m = L diag\{\frac{\alpha_i}{\alpha_i^{\ 2} + \beta^2}\} U^T \Delta d \text{.....(5)}$$

This is the solutions of the inverse problem.

The damping factor is,

$$\beta = \alpha_w \Delta c^{\frac{1}{w}} \text{.....(6)}$$

Where w is the number for the damping factor at any iterations, α is the parameter eigen value.

The term Δc is given by

$$\Delta c_r = \frac{c_{r-1} - c_r}{c_{r-1}} \text{.....(7)}$$

Where c_{r-1} is the misfit value obtained at previous iteration and c_r is the misfit found at the current iteration.

Resistivity inversion scheme:

In the case of two-dimensional earth model resistivity varies along x and t-axis and remains invariant along the y-axis. Hence response for the case of a 2D earth model can be given in the form of poisons equation.

$$-\nabla[\sigma(x,z)\nabla v(x,y,z) = I(x,y,z) \text{.....(1)}$$

Here $\sigma(x,z)$ represents the conductivity, v(x,y,z) is the electric potential and I(x,y,z) denotes the current source intensity. The Fourier transform of eq(1) with respects the y-co-ordinates takes the following form

$$-\nabla[\sigma(x,z)\nabla v(x,k_y,z)] = k_y\sigma(x,z)v(x,k_y,z) = I(x,k_y,z) \text{.....(2)}$$

Here v denotes the fourier transform and ky is the fourier transform variable. Equation (2) can be solved by numerically using finite element and finite differences. The potential in real 3D domain can be obtained by solving eq(2) and appliying inverse fourier transform

$$\Delta v(x,o,z) = \frac{1}{\pi}\int_0^\infty v(x,k_y,z)dk_y$$

The apparent resistivity for schlumberger can be calculated as $\rho_a = \frac{G\Delta v}{I}$

G=Geometrical factor

Δv = Calculated potential difference

Resistivity Inversion 2D/3D Case:

Forward Modelling:

The equation governing the DC response due to a point source $I_s(x,y,z)$ are given by (e.g. Telford et al , 1976)

$$\sigma\Phi(x,y,z) = -\rho(x,y,z)J(x,y,z) \text{.....(1)}$$

And $\sigma J(x,y,z) = I_s(x,y,z)$.....(2)

Where $\Phi(x,y,z)$ is the electrical potential, J(x,y,z) is the current density and $\rho(x,y,z)$ is the 3D resistivity distribution. The basic equations(1) and (2) governing the potential Φ and

current density J around a point source I can be re-written as an operator D acting on v with source β

$DV = \beta$.....(3)

Where

$$D = \begin{bmatrix} \rho & \nabla \\ \sigma & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} J \\ \Phi \end{bmatrix}$$

And $\beta = \begin{bmatrix} 0 \\ I \end{bmatrix}$.....(4)

A perturbation analysis can be done to determine the sensitivity of the potential field to changes in resistivity at depth. Perturbing the resistivity gives

$(D + \delta D)(V + \delta V) = \beta$.....(5)

Where

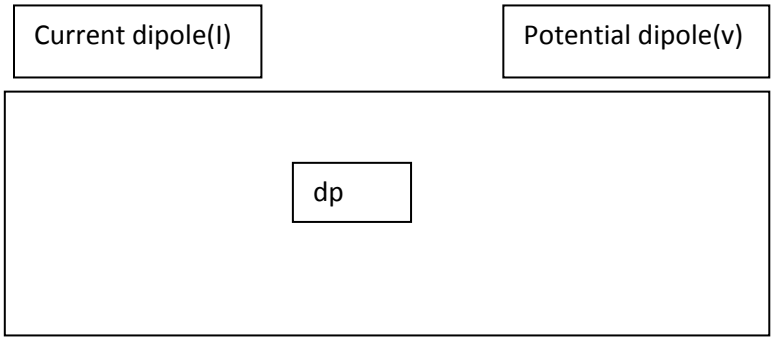
$\delta D = \begin{bmatrix} \Delta\rho & 0 \\ 0 & 0 \end{bmatrix}$ and

$\delta v = \begin{bmatrix} \delta J \\ \delta\Phi \end{bmatrix}$.....(6)

Expanding this equation and neglecting second order terms yields

$D.\delta V = -\delta D.V$.....(7)

Comparison of equations(3) and (7) reveals changes in conductivity act as equivalent sources of current for changes in potential fields at the surface.



Vertical cross-section of the 3D body at depth is shown here. Reciprocity can be established by interchanging the current and potential dipoles.

Using the above relation, the concept of reciprocity, and the generalized Green's identity(Lancz 1961), the changes in the potential field at the surface due to changes in the conductivity of each discretized block is given by

$$\delta\Phi = \int_J \delta\rho J.J'dJ \text{.....(8)}$$

In the above relation, J' is the current density osf a discretized block after interchanging the current and potential dipole. Using equation(8), the sensitivity of a surface measurement to the changes in resistivity of a 3D structure discretized into small blocks can be expressed by the algebraic sum of sensitivities to individual blocks. For a particular set of measurements the sensitivity of the potential field over a 3D discretized body can be written as

$$\delta\Phi = \sum_{i=1}^{N} \delta p_i \int_0^{\Delta x}\int_0^{\Delta y}\int_0^{\Delta z} \left(J_x J_x' + J_y J_y' + J_z J_z'\right) dxdydz \text{.....(9)}$$

The current density can be obtained for each block by a forward scheme, so that the above equation reduces to

$$\delta\Phi = \sum_{i=1}^{N} \delta p_i (J_x J_x' + J_y J_y' + J_z J_z')_i \Delta x\Delta y\Delta z \text{.....(10)}$$

This relationship can be reduced for 2D and 1D structures by assuming that the current density in the strike direction of the 2D body is negligible(i.e the average current density in the strike direction is zero), in which case $\delta\Phi$ for 2D bodies is

$$\delta\Phi = \sum_{i=1}^{N} \delta p_i (J_x J_x' + J_z J_z')_i \Delta x\Delta z \int_{-\infty}^{\infty} dy \text{.....(11)}$$

$\delta\Phi$ for 1D bodies is

$$\delta\Phi = \sum_{i=1}^{N} \delta\rho_i (J_z J_z')_i \Delta z \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy + \sum_{i=1}^{N} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta\rho_i (J_x J_x' + J_y J_y')_i dxdy \text{.....(12)}$$

Inversion:

Non-linear resistivity problem can be linearized using a Taylor series expansion and written in a matrix form as

$$\Delta m = G\Delta p$$

Where Δm (i.e $\delta\Phi$) is the difference between field and calculated responses.

G= is the derivative of data with respected to the model parameters(known as the system or co-efficient matrix, equal to the sum of the current densities in each block).

$\Delta p =$ is the difference between the initial guess and the calculated model parameters(i.e δp)

It is better to deal with logarithmic variations as the data and model parameters are weighted uniformly. The damped least-squares solution is given by

$$\Delta \ln(\rho) = (G^T W G + \alpha I)^{-1} G^T W \Delta \ln(\Phi)$$

Where W is a square diagonal matrix containing the reciprocal of the data variances. The choices of α is critical because it controls both the speed of convergence and the final solution. A damping which is the product of the misfit error and an estimate of a cut-off eigen value λ is used. The trace of $G^T W G$ is the sum of eigen values, and from the trace, the average eigen value is calculated as $\lambda_{avg} = Trace / N$

This gives a damping factor $\alpha = f \lambda_{avg} \varepsilon$

Where ε is the RMS error, f is the fraction of the average eigen value, and may range from 0 to 1.

The resolution of model parameters can be studied by computing the resolution matrix(R) as follows:

$$R = \left(G^T W G + \alpha I\right)^{-1} G^T W G \text{.....(14)}$$

If the resolution matrix is an identity matrix(I) or close to a identity matrix, parameters are well resolved,Otherwise, the model parameters are poorly resolved.

Flow-Diagram:

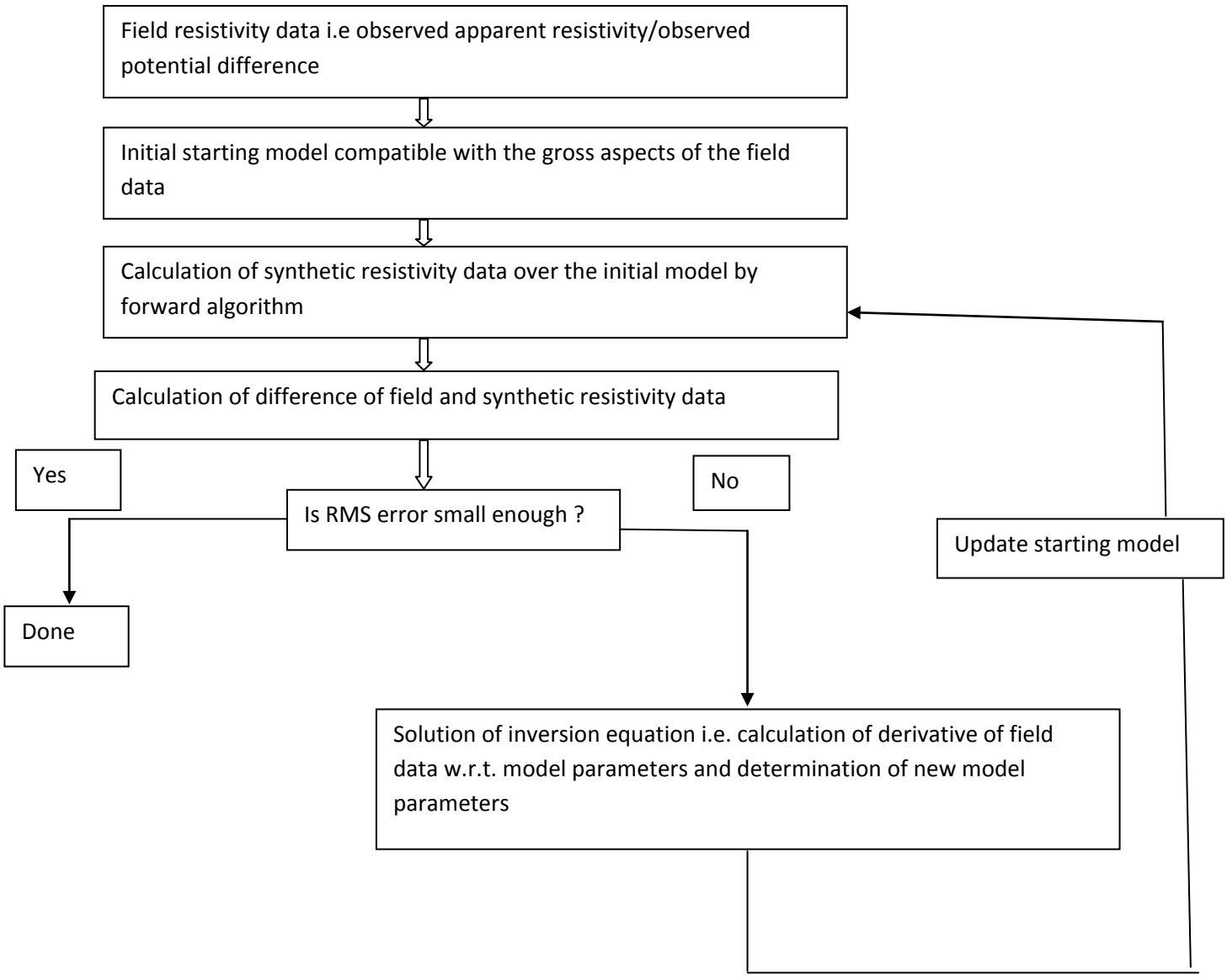


Fig: Flowchart for a typical resistivity inversion algorithm

Gravity Modelling and Inversion:

Following parker (1973) the gravity anomaly caused by an uneven, uniform layer of material by means of a series of Fourier Transform.

$$F(\Delta g) = -2\pi G\rho e^{-kzo}\sum_{n=1}^{\infty}\frac{k^{n-1}}{n}F\left[h^n(x)\right] \quad \text{.....(1)}$$

$F(\Delta g) =$ Fourier transform of the gravity anomaly

G= Gravitational constant
ρ=density contrast across the interface
K=wave number
h(x)=depth to the interface(positive downwards)
Z_0=mean depth of the horizontal interface

Oldenburg (1974) rearranged this equation to compute the depth to the undulating interface from the gravity anomaly profile by means of an iterative process, given by

$$F\left[h(x)\right] = -\frac{F\left[\Delta g(x)\right]e^{-kzo}}{2\pi G\rho} - \sum_{n=2}^{\infty}\frac{k^{n-1}}{n}F\left[h^n(x)\right] \quad \text{.....(2)}$$

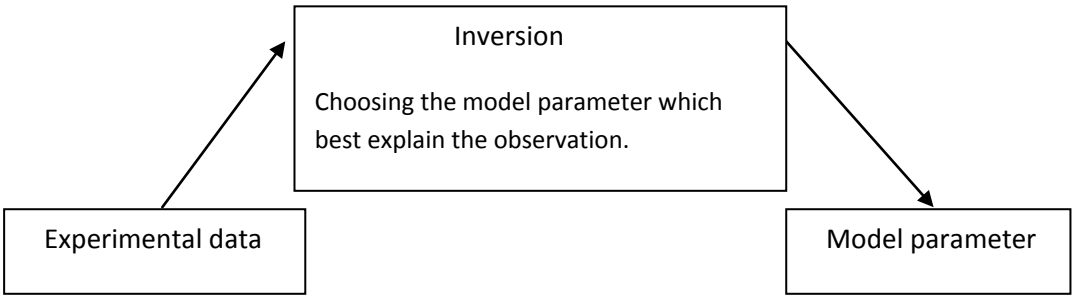
Steps:

1. Read the gravity data.
2. The gravity data is de-meaned prior to the calculation of the Fourier Transform.
3. The first term of equation (2) is computed by assuming that h(x)=0 (Oldenburg, 1974).
4. Inverse Fourier Transform provides the first approximation of the topography interface, h(x).
5. The value of h(x) is then used in equation (2) to evaluate a new estimation of h(x). This process is continued until a reasonable solution is achieved.
6. In order to gain the stability of the inversion a high cut filter HCF(k) is included in the inversion such that

$$HCF(k) = \frac{1}{2}\left[1+\cos\left(\frac{k-2\pi\omega H}{2(SH-\omega H)}\right)\right]$$

For ωH<k<SH
HCF(k)=0 for k>SH
HCF(k)=1 for k< ωH
[SH=0.012 and ωH=0.01]
The frequency, k can be expressed as 1/λ , where λ= wavelength in kilometres.

Problem Formulation



The process of selecting variable to represent data and model parameters may be broadly referred to as parameterization (many workers use this term for parameter selection only).

We need to pose the problem as

Discrete system

Earth can be parameterised into number of discretised layer each with own, density (ρ_i) or seismic velocity (v_j), electrical resistivity .

Instead of expressing density as a function of radius we may be interested in determining the average density of core/mantle.

EXAMPLE: Borehole temperature measurement

Suppose that we made a temperature measurement T_i at n depths Z_i and want to fit a straight line to the data.

Here, $d = [T1, T2, T3, ...Tn]^T$

And Intercept a and b are two model parameter i.e. $m = [a, b]^T$

By FORWARD THEORY the data T must satisfy the relation

$T_i = a + bz_i$

$T_1 = a + bz_2$

.

.

.

$T_{n-1} = a + bz_{n-1}$

$T_n = a + bz_n$

$$\begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ . \\ T_{n-1} \\ T_n \end{bmatrix} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ . & . \\ . & . \\ . & . \\ 1 & z_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Which in the matrix form d=Gm which is an over determined problem.

Example: Density distribution within the Earth

Consider the problem of determining the average density of core (ρ1) and that of mantle(ρ2) from the measurement of Earth's mass and moment of inertia.

In this case, we have two data values(d1=mass, d2=moment of inertia)

Two model parameters ($\begin{matrix} \rho 1 = m1 \\ \rho 2 = m2 \end{matrix}$).

d1=mass of the earth= $\frac{4}{3}\pi\rho 1 c^3 + \frac{4}{3}\pi\rho 2\left(a^3 - c^3\right)$

d2=moment of inertia= $\frac{8}{15}\pi\rho 1 c^5 + \frac{8}{15}\pi\rho 2\left(a^5 - c^5\right)$

We can write in short hand notations

$di = \sum_{j=1}^{2} G_{ij} m_j$ i=1,2

In complete form

$$\begin{bmatrix} d1 \\ \dots \\ d2 \end{bmatrix} \begin{bmatrix} \frac{4}{3}\pi c^3 \\ \dots \\ \frac{4}{3}\pi c^5 \end{bmatrix} \begin{bmatrix} \frac{4}{3}\pi\left(a^3 - c^3\right) \\ \dots \\ \frac{8}{15}\pi\left(a^5 - c^5\right) \end{bmatrix} \begin{bmatrix} \rho 1 \\ \dots \\ \rho 2 \end{bmatrix}$$

This is an even determine problem.

Example: Digital filter design in seismic de-convolution:

Two signals a(t) and b(t) may be related by convolution with a filter f(t) in the form

A(t)=f(t)*b(t)

Given: a(t) and b(t)

To find out: f(t)

Let us discretize the problem

Let time series length =n

Filter length=p

$$ai = \Delta t \sum_{j=1}^{p} f_i b_{i-j+1}$$

Where bi=0 if i<1 or i>1

$\Delta t =$ sampling interval

The equation is linear in unknown filter coefficient fj

We can write, d=Gm

Where d=a[in time series data]

M=f[sought filter]

$$G = \Delta t \begin{bmatrix} b1 & 0 & 0 & . & . & 0 \\ b2 & b1 & 0 & . & . & 0 \\ b3 & b2 & b1 & . & . & 0 \\ . & . & . & . & . & . \\ bn & b_{n-1} & b_{n-2} & . & . & b_k \end{bmatrix}$$

Where k=n-p+1

The system is over determined i.e. n>p

Solving over-determined linear inverse problem:

Simple linear regression:

For a collection of n data pairs {(x1 y1),(x2 y2),...(xn yn)}, the fitted line known as the regression line is described by the equation.

Y=a+bx

Each data pair satisfy the relation

Yi=a+bxi+ei

[ei the vertical distance between ith data point regression line]

The ei is the vertical distance between the ith data point and regression line. The solution to the straight line problem in this case is not exact solution since the relation yi=a+bxi can not be satisfied for every i and problem is also over-determined. This type of problem is solved by least squares method.

In the least square method we try to minimize the error 'e' by determining those parameters a,b, such that the sum of squares of the error(s) is minimal is minimize.

$$S = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (yi - a)$$

Minimization is accomplished by differentiating S with respect to model parameters.

$$\frac{ds}{da} = 2\sum_{i=1}^{n} (yi - a - bxi)(-1) = 0$$
$$\frac{ds}{db} = 2\sum_{i=1}^{n} (yi - a - bxi)(-xi) = 0$$

Giving $\sum a + \sum bxi = \sum yi$

And

$$\sum axi + \sum bxi^2 = \sum xiyi$$
$$\sum y = na + (\sum xi)b$$
$$n\sum xy = n\left(\sum x\right)a + n\left(\sum x^2\right)b$$
$$\sum x \sum y = n\left(\sum x\right)a + \left(\sum xi\right)\left(\sum x\right)b$$
$$n\sum xy = n\left(\sum x\right)a + n\left(\sum x^2\right)b$$
$$\sum x \sum y - n\sum xy = \left(\sum x\right)\left(\sum x\right)b - n\left(x^2\right)b$$
$$b = \frac{n\sum xy - \sum x \sum y}{x\sum x^2 - \left(\sum x\right)^2}$$

Is the slope of the fitted line

$a = \frac{\sum y - b\sum x}{n} = \bar{y} - b\bar{x} \rightarrow$ intercept on the y-axis.

The above concept are used routinely in geophysical data analysis and especially when dealing with problems with me or two parameters(e.g. a simple straight line fitting) and technique is termed linear regression analysis or classical least square fitting.

The method was originally formulated to provide a solution to the over determined problem but the same approach can be adopted for under determined problems. The solution was originally given by gauss in 1809. When we have more than two model parameters then we require a simple extension of the above method referred to as multiple-regression analysis. However it is possible to formulate a generalized relationship that will be applicable to any dimension of data and model parameters. This approach is commonly adopted in geophysics and the procedure uses matrix formulations instead and is aptly dubbed generalized least squares or matrix inversion(GMI).

Solution of purely under determine problem:

To obtain a solution mest to the inverse problem, we must have some means of singing out precisely one of the infinite number of solutions with zero prediction error, E, To do this, we must add to the problem some information not contained in the equation.

Gm=d, this extra information is called a priori information. A priori information is can take many forms, but in each case, it quantifies expectations about the character of the solution that are not based on the actual data. For instance, in the case of fitting a straight line through a single data point, one might have the expectation that the line also passes through the origin. This is a priori information now provides enough information to solve the average problem uniquely, since two points(one datum, one priori) determine a line.

Example of a priori information:

1.Density is everywhere inside the earth positives

2. Sine the interior of the earth . can reasonable be assumed to be rock, its density must have values in some range known to characterize rock, between 1D 100gm/cc. If one can use this a priori information when solving the inverse problem, it may greatly reduce the range of possible solutions-or even cause the solution to be unique.

Choices of a priori?

Where does this information come from, How certain is it?

The first kind of a priori assumption, we shall consider is the expectation that the solution to the inverse to the inverse problem is simple, where the notion of simplicity is quantified by some measure of the length of the solution. One such measure is simple the Euclidean length of the solution.

$$L = m^T m = \sum mi^2$$

A solution is therefore defined to be simple if its is small when measured under the L2 norm.

We define constrained optimization function

$$q^2 = m^T m + \lambda(d - Gm)$$

$m^T m =$ measure of model simplicity

$\lambda(d - Gm) =$ measure of the data misfit

$\lambda =$ Lagrange multiplier

$$\frac{dq}{dm} = 2m - \lambda G$$

$$m = \frac{\lambda}{2} G$$

$$m^T = \frac{1}{2} G^T \lambda$$

$$m = \frac{1}{2} G^T \lambda$$

We know d=Gm

$$d = GG^T \frac{\lambda}{2}$$

That mean $\lambda = 2(GG^T)d$

We note that $G^T G$ is a measure $(N \times N)$ matrix if its inverse exists, we can solve the equation for lagrange multipliers.

$$m_{est} = \frac{1}{2} G^T \lambda = \frac{1}{2} G^T 2(GG^T)^{-1} d$$

$$m_{est} = G^T (GG^T)^{-1} d$$

$$= G^{-g} d$$

Here, generalized inverse operator

$$G^{-g} = G^T (GG^T)^{-1}$$

<u>Damped least square solution:</u>

Cost function containing data misfit and model misfit

$$q = \beta^2 m^T m + (Gm - d)^T (Gm - d)$$

Error=solution error + prediction error

$\beta =$ Trade off parameter between data norm and model norm in the optimization process.

$$q = \beta^2 m^T m + d^T d - d^T Gm - m^T G^T d + m^T G^T Gm$$

$$\frac{dq}{dm^T} = \beta^2 \text{ Im} + (G^T Gm - G^T d) = 0$$

$$m(G^T G + \beta^2 I = G^T d$$

$$m = (G^T G + \beta^2 I)^{-1} G^T d$$

$$m_{est} = (G^T G + \beta^2 I)^{-1} G^T d$$

Where I= Identity matrix

$\lambda =$ damping factor

This is a damped least square solution.

If β is made large minimization procedure will clearly minimize the underground part of the solution. Unfortunately , it also tends to minimize the over determined part of the solution.

β = must be estimated by trial and error procedure.

Data resolution matrix:

Linear inverse problem takes the form d=Gm

Using the generalized inverse theory , we get an estimate of model parameter

$$m^{est} = G^{-g} d^{obs}$$

For the sake of simplicity, we assume that there is no additional vector space.

Data predicted

$$d^{pred} = Gm^{est}$$
$$= GG^{-g} d^{obs}$$
$$= [GG^{-g}] d^{obs}$$
$$= Nd^{obs}$$

$N = (N \times N)$ square matrix= GG^{-g} is called data resolution matrix.

Date resolution matrix characterize whether data can be independently predicted/ resolved.

When N=I, $d^{pred} = d^{obs}$ i.e prediction error is zero if $N \neq 0$, prediction error is non zero.

Interpretation:

Consider a problem of fitting a straight line to (Z,T) points, where data T have been ordered according to value of auxiliary variable Z. If $N \neq I$ but is close to an identity matrix(in the sense that its largest elements are near its main diagonal), then the configuration of the matrix signifies that averages of neighbouring data can be predicted, where as individual data can not.

If i[th] term of N is $[...0,0,0,.1,.8,.1,0,0,0,...]$

$$T^{pred} = \sum_{j=1}^{N} N_{ij} T_j^{obs} = .1T_{i-1}^{obs} + .8T_i^{obs} + .1T_{i+1}^{obs}$$

Predicted value is weighted average of three neighbouring observed data.

The row of data resolution matrix N describes how well neighbour observed data.

x=dia[N]

n=impedance of data

Model resolution matrix:

We imagine that there is true but unknown set of model parameters m that solve,

$$G_m^{true} = d^{obs}$$

Now,

$$\begin{aligned} m^{est} &= G^{-g} d^{obs} \\ &= G^{-g} G m^{true} \\ &= \left[G^{-g} G\right] m^{true} \\ &= R m^{true} \end{aligned}$$

Here R is $(M \times M)$ model resolution matrix. If R=I the each model parameter uniquely determined.

Then $m^{est} = m^{true}$

If $R \neq I$, estimates of the model parameters are really weighted everages of the true model parameters.

If the model parameters have a natural ordering(as they would if they represented a discretized version of a continuous function), then plots of the rows of the resolution matrix can be useful in determining to what scale features in the model can actually be resolved.

Null vectors:

Suppose that the inverse problem has two distinct solutions m1 and m2

So, Gm1=d

Gm2=d

Subtracting these two equations yiels

G(m1-m2)=0

Since, the two solutions are by assumptions distinct, their difference

$n^{null} = (m1 - m2)$ is non-zero.

The converse is also true, any linear inverse problem that has null vectors is non-unique. If m^{par} (par-particular) is ant non-null solution to Gm=d for instance minimum length solution, then $m^{par} + \alpha m^{null}$ is also a solution any choice of α. Note that since αm^{null} ia a null vector for any non-zero α, null vectors are only distinct if they are linearly independent. If a given inverse problem has a q distinct null solutions, then most general solution

$$m^{gen} = m^{par} + \sum_{i=1}^{q} \alpha_i m^{null}$$

Note that we have seen, we could able to estimate model parameter of linear inverse problem . Sometimes we had to add a priori information to obtain solution. In the case of non-uniqueness, it is possible to devise a solution that does not contain a priori information.

<u>Null vector of simple inverse problem:</u>

$$Gm = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} m1 \\ m2 \\ m3 \\ m4 \end{bmatrix} = \begin{bmatrix} d1 \end{bmatrix}$$

One obvious solutions to this equation is $m = \begin{bmatrix} d1 & d2 & d3 & d4 \end{bmatrix}^T$ (In fact this is a minimum length solution)

The null solutions can be determined by

$$m1^{null} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

$$m2^{null} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

$$m3^{null} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

The most general solution

$$m^{gen} = \begin{bmatrix} d1 \\ d2 \\ d3 \\ d4 \end{bmatrix} + \alpha 1 \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} + \alpha 2 \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} + \alpha 3 \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

Finding the particular solution to this problem , now consists of choosing values for α , if one chooses the parameters so that $\|m\|_2$ is minimized, one obtain minimum length solution, which occurs $\alpha i = 0[i = 1,2,3]$. Note that this is a general solution, minimum length solution never contains any null vectors. But if we use other solution simply(flatness/roughness), those solutions will contain null vectors.

Least squares:

$$G^{-g} = \left(G^T G\right)^{-1} G^T$$
$$N = GG^{-g} = G\left(G^T G\right)^{-1} G^T$$
$$R = G^{-g} G = \left(G^T G\right)^{-1} G^T G = I$$
$$[Conm] = G^{-g} G^{-gT} = \left[G^T G\right]^{-1} G^T G\left[G^T G\right]^{-1} = \left[G^T G\right]^{-1}$$

Minimum length:

$$G^{-g} = G^T \left[GG^T\right]^{-1}$$
$$N = GG^{-g} = GG^T \left[GG^T\right]^{-1}$$
$$R = G^{-g} G$$
$$= G^T \left[GG^T\right]^{-1} G$$
$$[Conm] = G^{-g} G^{-gT} = G^T \left[GG^T\right]^{-1} \left[GG^T\right]^{-1} G^T$$
$$= G^T \left[GG^T\right]^{-2} G^T$$

Null space:

The concept of vectors space is particularly helpful in understanding the mixed determined problem, in which some linear combinations of the model parameters are overdetermined and some are underdetermined. If the problem is to some degree underdetermined, then the equation Gm=d contains information about only some of the model parameters. We can express this combinations lying in a subspace, Sp(m) of the model parameters space. No information is provided the part of the solution that lies in the rest of the space, which is called as null space so(m). The part of the m, that lies in the null space is completely 'unilluminated' by the Gm=d, Since the equation contains no information about these linear combinations of the model parameters.

$$Mixed\ \det er\ \min edproblem \rightarrow over\ \det er\ \min edpart + underdter\ \min edpart$$

If the model parameters and data are divided into parts with subscript p that lie in the p space and parts with subscript 'o' that lie in the null space.

So, we can write Gm=d as $G\lfloor m_p + m_o \rfloor = \lfloor d_p + d_o \rfloor$

The solution length is then

$$L = m^T m = \left[m_p + m_o\right]^T \left[m_p + m_o\right]$$
$$= \left[{m_p}^T + {m_o}^T\right]\left[m_p + m_0\right]$$
$$= {m_p}^T m_p + {m_p}^T m_0 + {m_0}^T m_p + {m_0}^T m_0$$
$$= {m_p}^T m_p + {m_0}^T m_0$$

[The cross term ${m_p}^T m_0$ and ${m_0}^T m_p$ are zero since the vector lie in different spaces]

The prediction error

$$I = \left[d_p + d_0 - Gm_p\right]^T \left[d_p + d_0 - Gm_p\right]$$
$$= \left[{d_p}^T + {d_0}^T - {m_p}^T G^T\right]\left[d_p + d_0 - Gm_p\right]$$
$$= {d_p}^T d_p + {d_p}^T d_0 - {d_p}^T Gm_p + {d_0}^T d_p + {d_0}^T d_0 - {d_0}^T Gm_p - {m_p}^T G^T d_p - {m_p}^T G^T d_0 + {m_p}^T G^T Gm_p$$
$$= \left[d_p - Gm_p\right]^T \left[d_p - Gm_p\right] + {d_0}^T d_0$$

A priori information is added to specify only those linear combinations of the model parameters that reside in the null space So(m), and the prediction error is reduced to only the portion in the null space So(d) by satisfying $e_p = \lfloor d_p - Gm_p \rfloor$=0 exactly.

One possible choice of a priori information is $m^{est} = 0$, which is sometimes called the natural solution of the mixed-determined problem.

When Gm=d is purely underdetermined the natural solution is just minimum length, and when Gm=d is purely over-determined, it is just the least-square solution.

Cross-hole tomography:

1	2	3	4
5	6	7	8
...	...	...	...
13	14	15	16

$d = [T1, T2 ... T8]^T$

Let each brick =4(height=h, width=n)

S=slowness(inverse of velocity)

$m = [S1, S2, ... S16]^T$

Row1: T1=hs1+hs2+hs3+hs4

Row2: T2=hs5+hs6+hs7+hs8

.

.

.row4: T4=hs13+hs14+hs15+hs16

Col1:T5=hs1+hs5+hs9+hs13

.

.

.col 4: T8=hs4+hs8+hs12+hs16

$$\begin{bmatrix} T1 \\ T2 \\ . \\ . \\ . \\ T8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & . & . & . & . & . & . & . & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S1 \\ S2 \\ . \\ . \\ . \\ S6 \end{bmatrix}$$

Solution:

$$\Delta m = \left(G^T G\right)^{-1} G^T \Delta d$$

Methods:

$$t = \int \frac{ds}{v(x,y)} \rightarrow Fourier\,\mathrm{Pr}\,ojection \rightarrow V(x,y)$$
$$\rightarrow FilteredBackprojection \rightarrow$$
$$\rightarrow ART, SIRT \rightarrow$$
$$\rightarrow MatrixInversion \rightarrow$$