Inversion with inaccurate data.

Solving Linear Inverse Problems

Thus far we have been considering exact solutions to linear inverse problems and we haven't looked at the practical aspects. Now turn our attention to these.

I: Smallest model calculations.

$$d_j = (g_j, m) \qquad j = 1, N$$

$$m = \sum \alpha_j g_j$$

solve $\quad \Gamma \bar{\alpha} = \bar{d} \qquad\qquad \Gamma_{ij} = (g_i, g_j)$

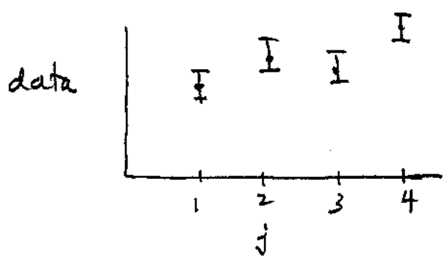
We stated that if g_j's were linearly independent, then Γ was invertible. In fact Γ is symmetric and positive definite. The positive definitness states that all of the eigenvalues are > 0 and hence

$$\det[\Gamma] = \prod_{i=1}^{N} \lambda_i > 0$$

So Γ^{-1} exists and we should be able to calculate it. However a computer is capable of only finite precision in its arithmetic computations and so it may not be possible to get out a reliable answer, especially as N becomes large.

II: Even if we could get an arbitrarily accurate representation of Γ^{-1}, do we really want to solve the equations $\Gamma \bar{\alpha} = \bar{d}$ exactly.

All measured geophysical data are uncertain. Such data may be presented as



If the data are uncertain then we do not want to find a model which reproduces these data exactly. Doing so would ensure that we do not have a valid solution.

These two points mean that we cannot, nor do we want to, solve our matrix equations exactly.

Inaccuracies in geophysical measurements.

All geophysical data are inaccurate. Each observation d_j^{obs} is really

$$d_j^{obs} = d_j^t + \delta d_j$$

where d_j^t is the true value and δd_j is an error. The primary difficulty lies in evaluating δd_j.

Certainly δd_j is not known; if it were, we could simply subtract δd_j from the data to obtain d_j^t.

The quantity δd_j is a statistical quantity. Its value is controlled by a probability density function, and δd_j is properly referred to as a random variable.

Consider, for the moment a random variable X. Let the pdf for X be given by $p_X(x)$.

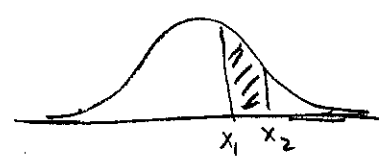
The pdf has certain properties:

$$\int_{-\infty}^{\infty} p_X(x)\,dx = 1$$

unit area

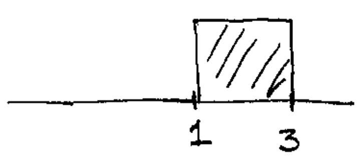
$$p_X(x) \geq 0 \qquad \text{everywhere.}$$

$$\text{Probability } (x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(x)\,dx$$



The pdf therefore quantifies the chance, or probability that a rv will take on a certain value.

eg

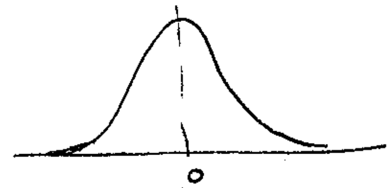


$\Rightarrow$ equal liklihood that X will take on a value between 1 and 3 and zero liklihood that it will have a value outside that region.

Think of an experiment in which one is pulling numbers out of a box.

A particularly useful pdf is the Gaussian pdf

$$p_X(x) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma}$$



Suppose X was a Gaussian r.v. What is its expected value or most likely value.

$$\mathcal{E}[X] = \int_{-\infty}^{\infty} x\, p_X(x)\, dx = 0$$

(not hard to figure out since $p_X(x)$ is even!)

But what about deviations away from this expected value. Effectively, what is the width of pdf

$$\mathrm{Var}[X] = \mathcal{E}\left[(X - \mathcal{E}[X])^2\right] = \int_{-\infty} x^2 p_X(x)\, dx = \sigma^2$$

The square root of this number is the standard deviation

$$SD[X] = (\mathrm{Var}[X])^{1/2} = \sigma$$

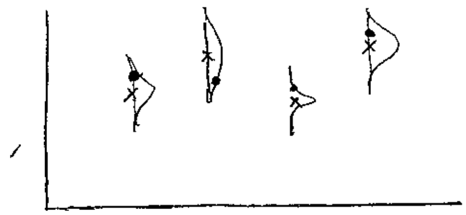
For the Gaussian pdf.

68% of values lies in $(-\sigma, \sigma)$
95% of values lie in $(-2\sigma, 2\sigma)$

So the standard deviation provides a great deal of insight about the spread, or the width, of the pdf about its mean value.

The difficulty with geophysical data uncertainties is that their statistics are unknown! We do not generally even know the form of the pdf, never mind specifying its parameters.

For the geophysical inverse problem, the problem becomes more complicated because we have N data.



x : true data
pdf for the error.
• denote observations.

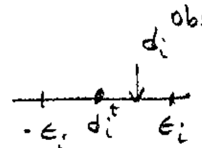
If the inverse problem is solved to find a model m_0 which reproduces the observations exactly then there is (essentially) zero possibility that we have the correct model since there is (essentially) zero probability that

$$d_i^{obs} = d_i^{t} \quad \text{for } i=1,N$$

Therefore, we don't want to fit the data precisely, but how well do we want to fit them?

There are choices:

if $\quad d_i^t - \epsilon_i \le d_i^{obs} \le d_i^t + \epsilon_i$



then we could require that the predicted datum d_i^{pred} (which hopefully will be d_i^{true})

$$\boxed{d_i^{obs} - \epsilon_i \le d_i^{pred} \le d_i^{obs} + \epsilon_i}$$

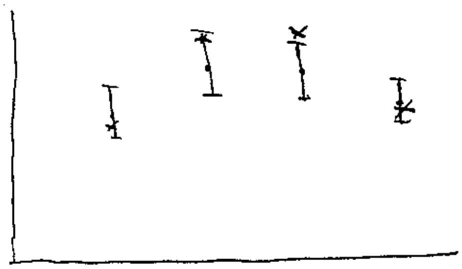
For N data one could require that this hold on an individual basis.

An alternate strategy is that the observed data $\vec{d}^{obs} \in \mathbb{R}^N$ and so does $\vec{d}^{pred}$.

Equip our data space with a norm and then measure the distance between the predicted data and the observations.

ie $\boxed{\text{misfit} = \| \vec{d}^{obs} - d^{pred} \|^2}$

eg. Could use a 2-norm misfit $\boxed{\phi = \sum_i \left(d_i^{obs} - d_i^{pred}\right)^2}$



← we might accept this as a final misfit. ie our criterion is a global one which involves all of the data.

Remark: Our <u>criterion</u> for misfit is then specified by the norm chosen on data space. We are at complete liberty to select different norms. (i) l_p, (ii) weighted.

Remark: In addition to a <u>criterion</u> we also need to define a <u>tolerance</u>. This is required to determine whether the data are

(i) fit too well
(ii) about right
(iii) fit too poorly

To define a tolerance we need to know what the statistics of the data are. Here we assume

(1) Each datum has an error δd_i which is Gaussian with zero mean and variance σ_i^2. (or standard deviation σ_i)

(2) The correlation between the errors on the i^{th} & j^{th} data is quantified by the covariance

$$\text{cov}[\delta d_i, \delta d_j] = C_{ij}$$

We will generally assume that the errors are uncorrelated. Then

$$\boxed{C_{ij} = \sigma_i^2 \delta_{ij}} \qquad \text{is diagonal.}$$

<u>Quantification of misfit.</u>

true data $\qquad d_i = (g_i, m)$

observed data $\qquad d_i^{obs} = (g_i, m) + \epsilon_i \qquad (\equiv d_i + \epsilon)$

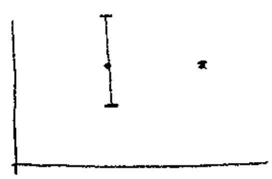
So the error is $\qquad \boxed{\epsilon_i = d_i^{obs} - (g_i, m)}$

If each ϵ_i was Gaussian, zero mean, independent and std dev. σ_i then one could choose to quantify the misfit as

$$\phi = \text{misfit} = \left(\sum \epsilon_i^2\right)^{1/2} = \| \vec{\epsilon} \|_2$$

But this is not reasonable if σ_i's are different. Consider two data

$$d_1^{obs} \pm \sigma_1 = 1.0 \pm .5$$
$$d_2^{obs} \pm \sigma_2 = 1.0 \pm .001$$



Clearly a model that gave predicted data $(.75, .75) = (d_1, d_2)$ is not acceptable for representing the information content in d_2. (probability of misfitting d_2 by 250 standard deviations is infinitesimal.
The appropriate misfit is

$$\boxed{\phi = \left[\sum \left(\frac{\epsilon_i}{\sigma_i}\right)^2\right]^{1/2} = \left[\sum_i \left(\frac{d_i - (g_i, m)}{\sigma_i}\right)^2\right]^{1/2}} \qquad (1)$$

ie effectively each random variable is normalized by its standard deviation so that it has <u>unit</u> standard deviation.
Equivalently the data equations are first normalized by their standard deviations prior to solution.

$$\boxed{\begin{aligned} \phi &= \| W_d (d^{obs} - d^{pred}) \|^2 \\ W_d &= \mathrm{diag}(1/\sigma_1 \; \cdots \; 1/\sigma_N) \end{aligned}}$$

If we adopt (1) as a misfit then it is possible to find an expected value and variance of ϕ.

Note. each r.v. ϵ_j/σ_i is a Gaussian r.v. with zero mean, and unit standard deviation. The statistics of this quantity are well known.

Let $\vec{X}$ be a vector of random variables $\vec{X} = (X_1, X_2, \dots X_N)$ each rv X_i is Gaussian, zero mean, unit standard deviation. Let

$\|X\|_2 = \left(\sum (X_i)^2\right)^{1/2}$ be the 2-norm of the vector X. Then

$$\boxed{\mathcal{E}[\|X\|] = N^{1/2}\left[1 - \frac{1}{4N} + \frac{1}{32N^2} + O(N^{-3})\right]} \qquad (2)$$

$$\boxed{\mathrm{Var}[\|X\|] = \frac{1}{2}\left[1 - \frac{1}{2N} + O(N^{-2})\right]} \qquad (3)$$

Eventhough it makes intuitive sense to measure misfit by use of a length of a vector in data space, it will be more convenient to work the square of misfit length. (This is similar to what was done in model space where we formulated the problem to minimize (m, m) rather than $(m, m)^{1/2}$.

If we work with squares we have

$$\boxed{\begin{aligned} \mathcal{E}[\|X\|^2] &= \mathcal{E}\left[\sum_{i=1}^{N} X_i^2\right] \simeq N \\ \mathrm{Var}[\|X\|^2] &\approx 2N \end{aligned}}$$

← obtained from (2) above.

Actually, the above quantity is very commonly used in statistics

$$\chi_N^2 = \sum_{i=1}^{N} X_i^2$$

Chi-squared variable with N degrees of freedom.

Remark: This now defines what our target misfit should be. Given N data contaminated with Gaussian error that is independent, zero mean, standard deviation σ_i then the desired model is that one which yields a misfit equal to the expected value

$$\boxed{\text{misfit} = \phi_d = \| W_d (d^{obs} - d^{pred}) \|^2 \simeq N} \qquad (4)$$

Now formalize the solution.

Given data equations $\quad d_i = (g_i, m) \quad i = 1, N$

and observations d_i^{obs}, each of which is assumed to be contaminated by a Gaussian r.v. independent, zero mean, standard deviation σ_i. Find a model which minimizes $\|m\|^2$ subject to misfitting the data by a specified amount ϕ_d^*.

$$\min \quad \phi_m = \|m\|^2$$

$$\text{subject to} \quad \phi_d = \phi_d^*$$

To solve this we introduce a Lagrange multiplier μ and form an objective fcn

$$\boxed{\phi(m) = \|m\|^2 + \mu(\phi_d - \phi_d^*)}$$

This is a constrained optimization problem in which the goal is to find m and μ. This is difficult to do rigorously.

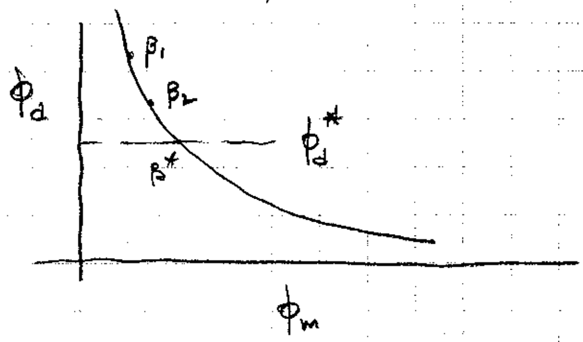
An alternate approach is to cast this problem as an unconstrained optimization problem.

$$\text{minimize} \quad \phi = \phi_d + \beta \phi_m \qquad (\text{ie } \beta = 1/\mu)$$

$$\text{and find } \beta \ni \quad \phi_d = \phi_d^*$$

where $\beta \in (0, \infty)$ is the regularization parameter, tradeoff parameter or Tikhonov parameter.

Both ϕ_d and ϕ_m are quadratic functionals. We will show later that the solution of the minimization yields



So the procedure to get a solution will involve a line search on β:

ie

for $\beta^{trial} = 1, \dots$
    min $\phi = \phi_d + \beta^{trial} \phi_m$
    evaluate ϕ_d
    if $|\phi_d - \phi_d^*| < tol$ exit
    else adjust β^{trial}.