Computational Exercise #1

Type of Activity: Illustrate the stage of the inverse process

Geophysical problem: Let us assume a set of N observed magnetic field, $y_1^o, y_{2,}^o \dots, y_N^o$, measured along a profile at the horizontal coordinates $x_1, x_2, \dots, x_N$, respectively.

We want to estimate the geomagnetic field (IGRF) from measurements of the magnetic field along a profile.

Simplifications: We assume that the geomagnetic field (IGRF) can be approximated by a first-order polynomial. This simplification is valid for small areas where the geomagnetic field varies linearly.

Forward problem: The mathematical model that describes a first-order polynomial is given by:

$$y_i = ax_i + b, \qquad i = 1, \dots, N, \tag{1}$$

where y_i is the ith "theoretical" geomagnetic field computed at the ith horizontal coordinate x_i (the ith observation station). In equation 1, a and b are the coefficients describing the first-order polynomial, where a is the slope.

Mathematical problem: We want to estimate a and b from a set of N observed magnetic field, $y_i^o \equiv y_i^o(x_i)$, $i = 1, \dots, N$. Let us assume that the ith observed magnetic field, $y_i^o \equiv y_i^o(x_i)$ can be write as:

$$y_i^o = ax_i + b + \varepsilon_i \,, \tag{2}$$

where ε_i is the ith residual. Hence, for each observation one defines an error or a misfit defined as the difference between the observed data and the predicted data, i.e.,

$$\varepsilon_i = y_i^o - (ax_i + b). \tag{3}$$

We wish to find (estimate) the unknown coefficients a and b that minimize ε_i.

Because we have a set of N measurement points ($x_1, x_2, \dots, x_N$), we have a set of N residuals ($\varepsilon_i(x_i) \equiv \varepsilon_{i\ ,}$ $i = 1, \dots N$. Here, we will use a traditional strategy which consists in minimizing the function given by:

$$Q = \sum_{i=1}^{N} \varepsilon_i^2. \qquad (4)$$

The total error Q (the sum of the squares of the individual errors) is exactly the squared Euclidean length of the errors. Usually, this function is called *data-misfit function*.

By minimizing the function Q (equation 4), we obtain the estimators for parameters a and b given by:

$$\hat{a} = \frac{\sum_{i=1}^{N} y^o{}_i (x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad (5a)$$

and

$$\hat{b} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} y^o{}_i - \hat{a} \sum_{\mathrm{i}=1}^{\mathrm{N}} x_i}{N} \qquad (5b)$$

where $\bar{x}$ in equation 5a is the mean of N horizontal coordinates, i.e.:

$$\bar{x} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} x_i}{N} \qquad (6).$$

Let us assume statistic assumptions such as: 1) additive errors; 2) the errors are independent (the errors are uncorrelated); 3) the expected value of the errors are null; 4) the errors have constant variance equal to σ^2; 5) the independent variables of the forward problem ($x_1, x_2, ..., x_N$, in equation 1) are not a random variable; 6) the parameters of the forward problem (a and b, in equation 1) are not random variables.

Under these assumptions, we can calculate the variance of the estimate $\hat{a}$ and $\hat{b}$, i.e.:

$$V[\hat{a}] = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \qquad (7a)$$

and

$$V[\hat{b}] = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad (7b)$$

where σ^2 is a constant and known variance of the errors.

In this exercise we need to produce:

I - THE FIRST SYNTHETIC TEST

(1) Generate the first synthetic test:

1.1.1) Assign the number of observations N (e.g., N = 3),

1.1.2) Assign numeric values to a and b (true parameters).

1.1.3) Generate the N horizontal coordinates $x_1, x_2, ..., x_N$ in the following way:

- In the first synthetic test – Generate horizontal coordinates such that x_1 and x_N are very distant to $\bar{x}$ (equation 6). For example, if N=3, the horizontal coordinates can be $x_1 = 0$, $x_2 = 10$ and $x_3 = 20$.

(1.2) Compute a noise-free data (exact data) by using equation 1 and the true parameters.

(1.3) Generate the noise-corrupted data (observed data set) $y_1^o, y_{2,}^o ..., y_N^o$:

1.3.1) Generate one pseudorandom sequence of noise with Gaussian distribution, null mean and standard deviation of 3% of the absolute value of the average of the noise-free data. To generate a pseudorandom sequence of noise you need to set a **SEED.**

1.3.2) After generating the pseudorandom sequence of noise, you can generate the noise-corrupted data (equation 2)

1.3.3) By using equations 6a and 6b and the noise-corrupted data obtained in the item 1.3.2) , obtain the estimate of a and b , respectively.

1.3.4) After estimating the parameters ($\hat{a}$ and $\hat{b}$), compute the predicted data.

(2) In the same Figure, you can plot:

2.1) Plot the noise-free data (see item 1.2) versus the horizontal coordinates $x_1, x_2, ..., x_N$

2.1) Plot the noise-corrupted data (see item 1.3) versus the horizontal coordinates $x_1, x_2, ..., x_N$

2.2) Plot the predicted data by using the estimated parameters (see item 1.3.4) versus the horizontal coordinates $x_1, x_2, ..., x_N$

(3) Compute the variances of the estimate of a and b by using equations 7a and 7b, , respectively.

(4) Compute the residuals;

$$\varepsilon_i = y_i^o - y_i^p, \qquad i = 1, \dots, N \tag{8}$$

where y_i^p is the ith predicted data.

(5) Compute the data-misfit function (equation 4).

(6) Compute the estimated noise standard deviation for the noise corrupted data by:

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{N} \varepsilon_i^2}{N-M}} \quad , \tag{9}$$

where M is the number of parameters (here M=2). Compare $\hat{s}$ with the true standard deviation of the noise (item 1.3.1).

I I – THE SECOND SYNTHETIC TEST

Repeat all the items (1)-(3) of the first synthetic test keeping the same number of observations (1.1.1), the assigned true parameters (1.1.2), the same seed and the same standard deviation of the noise. In this second test the horizontal coordinates will be different. Here, the item 1.1.3 will be:

1.1.3) Generate the N horizontal coordinates $x_1, x_2, ..., x_N$ in the following way:

- In the second synthetic test – Generate horizontal coordinates such that all of them, x_i, $i = 1, \dots, N$, are very close to $\bar{x}$ (equation 6). For example, if N=3, the horizontal coordinates can be $x_1 = 9.999$, $x_2 = 10$ and $x_3 = 10.001$.

I I I – VIEW THE INSTABILITY IN THE DATA SPACE:

Repeat the first and second synthetic tests changing the **SEED** (item 1.3.1)**.** Compare the estimated parameters and the predicted data in this topic with those obtained in the above-described items (I and II).