12.515 Data and Models
Theory and Computer Practice

Dale Morgan, Rama Rao
Earth Resources Laboratory
Department of Earth, Atmospheric and Planetary Sciences, MIT

Handed out: 11 Sep 03 (thursday) , **Due: 18 Sep 03 (thursday)**

Problem Set 1

Be creative in visualizing your results and create plots and keep records of all your trials.

NOTE: When drawing conclusions about the results of the random search it is most reliable if you repeat the search a few times (say thrice) with all settings being unchanged to ensure that you are not tricked by serendipity

1) Download the data set to be used for inversion from the 12.515 course locker. In athena, type

% `add 12.515`
cd to your directory of choice, and type
% `cp /mit/12.515/data1.mat .`
to get a matlab format data file
or type
% `cp /mit/12.515/data1.txt .`
to get an ascii data file
Note: Include the period at the end of the command line above

The data is two columns with z in the first column and T(z) in the second column.

2) A scientist has identified the model underlying the data as the following.

$$T(z) = \beta_1+\beta_2 z +\beta_3 z^2, \quad 0<=z<=5$$

- Create a Monte-Carlo (Random) search scheme to identify parameters β_1 β_2, β_3 using the above model and the downloaded data. Code your algorithm to save the 100 best solutions.

- Add 25% *gaussian* noise, as a% of the mean of the data, to the data and run the monte-carlo search and identify the parameters. Repeat thrice using three termination criteria (max number of iterations or RMS error threshold or both). How does this affect your parameter estimates? Which would you choose?

- Plot the 100 best solutions. What can you infer from the solution cloud? Comment about the errors in the three parameter estimates and the correlation between the parameters

- Add 25% *gaussian* noise, as % of the data value, to the data and run the random search. What is the difference in the parameter estimates between this attempt and one with data that had 25% *gaussian* noise, as a % of the mean of the data. Are some parameters affected more or less by the type of noise? If yes, what is the cause. You might have to run the searches multiple times with each noise type to get consistent parameter estimates.

Another scientist challenges the above model and proposes the following:

$$T(z) = \beta_1 z + \beta_2 \exp(- \beta_3 z), \quad 0<=z<= 5$$

- Modify the Monte-Carlo (Random) search scheme to use this model
- Add 25% *gaussian* noise, of the mean of the data, to the data and run the monte carlo search again. Which model do you think is a more accurate reflection of the data ? Why ?

3) Pick the model that you think is most appropriate, for the following exercises.

- Which data in the interval 0<=z<=5 contributes most to the determination of the model parameters? How does this sensitivity of the data to the parameter values vary with the different parameters?

- Run the search using RMS error and absolute error as the "goodness-of-fit" metric. Run with *uniformly distributed noise* levels of 10%, 40% and 70%, of the mean of the data. What differences do you see in the parameter estimates in using the two metrics (Try plotting the 100 best solutions)? Which metric would you choose for each of the three noise cases. Create a plot of error (RMS and absolute) vs noise magnitude (Run additional trials to get at least 6 points for each curve). From this, what general guideline can you extrapolate for the choice of the metric as a function of noise magnitude?

4) Create a grid search scheme to identify the parameters using the data and your chosen model

- Do a grid search by partitioning the initial b ranges into 999 intervals and find the best estimate of the parameters, adding 25% *gaussian* noise to the data.

- Repeat by partitioning the b ranges into 99 intervals and 9 intervals and estimate the parameters

- Quantify the trade-off between RMS error and the effort expended by plotting the final RMS error vs the resolution of your search. Partition the b ranges into additional interval sizes (apart from the three indicated above) and create the trade-off curve with at least 6 points. What do you learn from this? What is the 'optimal resolution' for your problem?

5) Modify the Grid and Random search schemes to concatenate the two of them to run two-step searches. Use 25% *gaussian* noise.

- Run a two-step search by
  - Identifying a wide parameter range initially
  - Running a coarse search on this range to identify a 'promising volume' to be searched further
  - Run a high-resolution search on this volume to refine your parameter estimates further.
  - Compare the Grid-Grid, Random -Grid, Random-Random and Grid-Random searches. What are the differences between these approaches? Which would you choose and why?

- Compare this to a one-step high resolution search run on the same initial parameter range. Did you gain anything by the two-step approach?