EE2703: Applied Programming Lab
Assignment 3
Fitting Data to Models

Arjun Menon Vadakkeveedu
EE18B104
Electrical Engineering, IIT Madras

February 9, 2020

Problem Description

The goal is to find the best (linear) fit to a set of noisy data using scipy's least squares regression functions. Here, the form of the data is known- 2nd order Bessel functions with unknown coefficients.

Code Structure

The python code has the following flow of logic:

1. File reading
2. Data Visualisation
   Plots of Raw Data, True Data and Error Bar
3. Least Squares Regression and Solution Visualisation
   Contour Plots of error, Plots of Solution, Error versus Stdev

The function **scipy.linalg.lstsq** solves linear equations of the form $\mathbf{Ax} = \mathbf{b}$ using least squares regression. It is known to us that the data is noisy second order Bessel functions, i.e.-

$$x = p_1 * J_2(t) + p_2 * t + n(t) \tag{1}$$

where, $J_2(t)$ is the second order Bessel function and $n(t)$ is zero-centred Gaussian noise.
Further, the true values of p_1 and p_2 are also known to be: $[p_1, p_2] = [1.05, -0.105]$.
The model that we try to fit the data to is of the form:

$$x_e = \mathbf{M.p} \tag{2}$$

Here, $\mathbf{M}$ is a 2D matrix with the first column holding values of $J_2(t)$ and the second column holding values of time. $\mathbf{p}$ is a column vector holding values of p_1 and p_2. **lstsq** returns the optimal values of $\mathbf{p}$.

Format for running file on Terminal:

```
python3 ee18b104_asgn3.py fitting.dat
```

The program asks for the column number whose solution, MSE contour plot, raw data etc. are to be plotted. Output format:

```
Enter column of fitting.dat that has to be fit: 6
Error between M_p() and g() (SANITY CHECK) =  0.0
Error in the solution w.r.t true value =  3.1033615797370384e-05
Optimal value of A and B =  [ 1.04856529 -0.10492891]
```

Plots

Data Visualisation

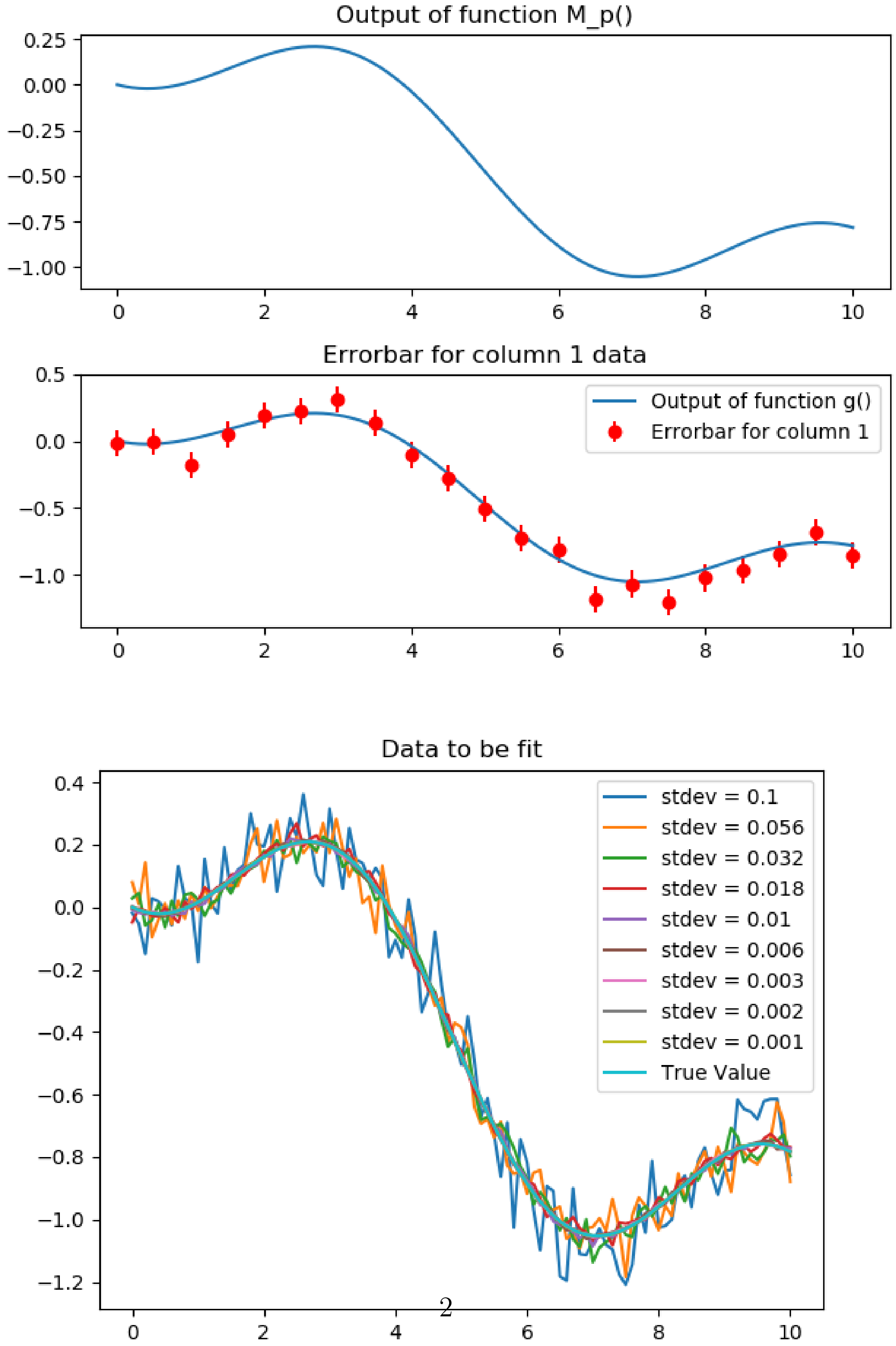
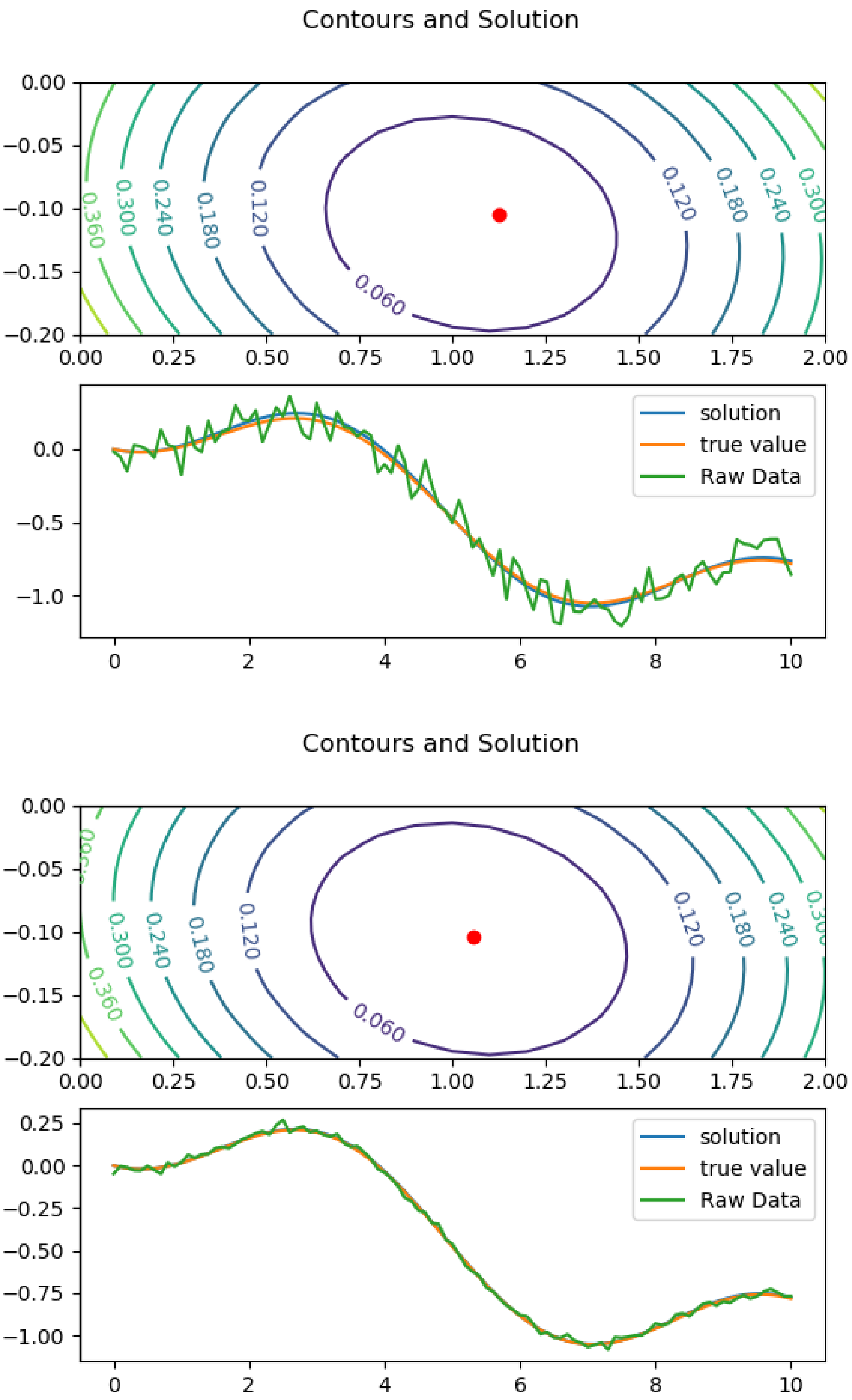


Figure 1: Plots of true value and error bar of data corresponding to column 1 in the file 'fitting.dat'; Raw Data corresponding to all columns along with true value

Plots of Solution


Contours and Solution

Contours and Solution

Figure 2: Plots of contour of MSE when $p_1$and p_2 are varied over the range [0, 2] and [-0.2, 0] respectively with 21 samples- the exact solution has been marked with a red dot; Solution versus true value and raw data; The plots correspond to data from columns 1 and 4 of 'fitting.dat'

Error

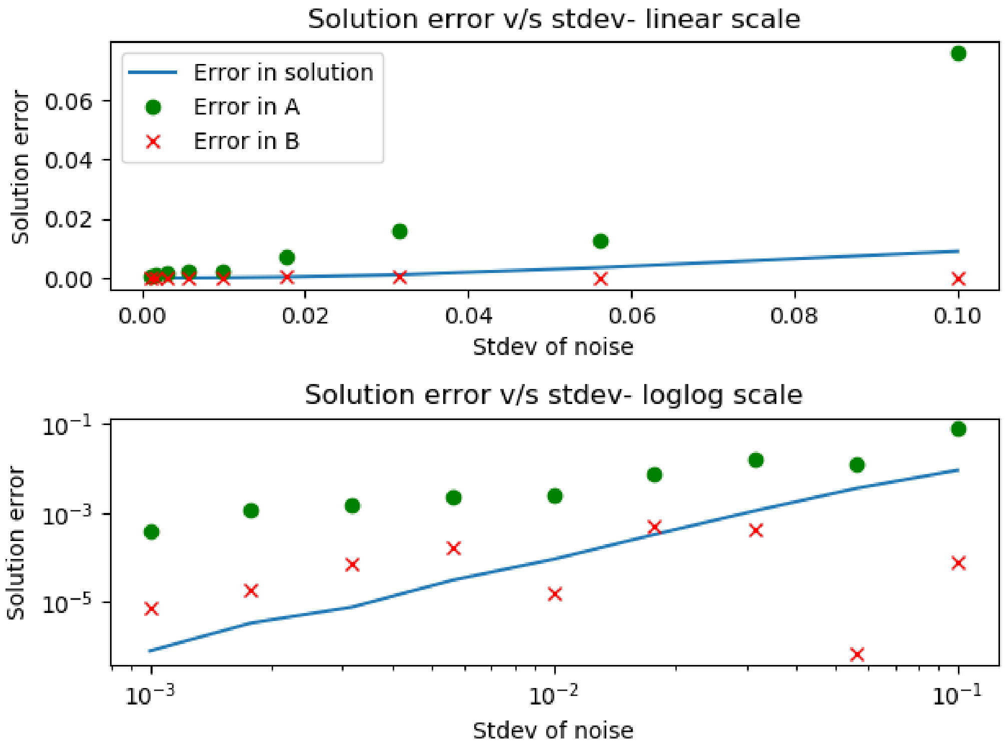


Figure 3: Error versus Standard Deviation of Noise in both linear and log-log scales

Conclusions

1. As expected, the matrix equation and the function g(t, A, B) provide the same function values. This is evident given that the mean squared error between g() and **M.p** is identically 0.

2. The contour plot of errors for different sets of data (which have different σ values) have similar least values (around 0.06). However, the global minimum value of the error is much smaller than this value, for all sets of data. This is so because the contours have been evaluated at large steps, thereby missing points that give lower error value.

3. The linear plot of error v/s standard deviation does not convey much information about the nature of the behaviour. This is because σ is spaced evenly on the log-scale between 1e-3 and 1e-1.

4. The log-log plot of solution error v/s standard deviation shows a linear relation (in the log-scale) between the two. As **lstsq** does not account for noises, it is reasonable that the error varies proportionately with the standard deviation.

5. Error in p_1 is small for small values of standard deviation, but shoots up as σ increases. p_1 error v/s σ can be fit to a straight line in the log-domain. Error in p_2 is smaller and does not indicate a direct linear relationship in the log-domain (due to the presence of outliers).
