Least squares approximation

1. The matrix [x,y] in an ascii file `xy_data` contains measurements of y for different values of x. Save this file to your directory. To use the data in Matlab computations you can load the data as follows:
   `>>load xy_data -ascii; x=xy_data(:,1); y=xy_data(:,2);`
   Which of the two nonlinear models provides for a better least squares fit to these data?

   $$(a)\quad y \approx \tan(a\exp(-x^2)+b), \qquad (b)\quad y \approx a\exp(b/(x+0.5)).$$

   To answer these questions you should approximate the data by each of the models (use Matlab) and compare the values of $\sum(\Delta y_i)^2$. Use data transformations to simplify the problem. In your report describe the algorithm and present graphs showing the data and the best fit curve for each model.

2. Let $Q_1, ..., Q_N$ be given points in R^2. It is needed to find a straight line l minimizing the value of

   $$r = \sum_{j=1}^{N} \mathrm{dist}^2(Q_j, l),$$

   where $\mathrm{dist}(Q_i, l)$ is the distance from the point Q_i to line l. Write an m-function `[alpha,c,r]=lsq_line(X,Y)` which computes the optimal values of α, c, and the residual r. Here X, Y are arrays of x- and y-coordinates, correspondingly, of the points Q_i. In your report, derive the equations determining the optimal line parameters and find the solution. In which case the solution is not unique?
   Hint: if the equation for l is written as

   $$x\cos(\alpha) + y\sin(\alpha) + c = 0$$

   and $Q_i = (x_i, y_i)$ then $\mathrm{dist}(Q_i, l) = x_i\cos(\alpha) + y_i\sin(\alpha) + c$.

3. You are asked to analyze the fluorescence histograms obtained using the Fluorescence-Activated Cell Sorter (FACS) at the Bone Marrow Transplantation Department of Hadassah hospital.

   The binary file `data.mat` contains five vectors of the same length, $f1$, $f2$, $f3$, $f4$, and fm (to get them, download the file and use Matlab command `load data`). Here $f1, ..., f4$ are histograms characterizing the distribution of fluorescence levels in four different populations of cells stained by a fluorochrome (see figure). The vector fm contains a similar histogram for a mixture of these cell populations. Your goal is to estimate the concentrations $c1$, $c2$, $c3$, $c4$ of each of the populations in the mixture by using the least squares method. No program needs to be submitted. In your report describe the model you used (note that $c1+c2+c3+c4 = 1$) and how you calculated the unknown concentrations. Present the concentration values and a graph showing the histogram of the mixture and its fit by the mixture of histograms.

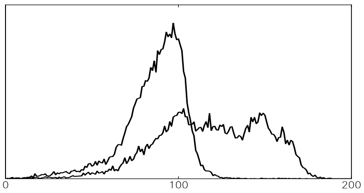

Figure: Histograms $f2$ and fm. To build these histograms, the possible range of fluorescence levels was divided into 200 intervals. FACS measured fluorescence levels of about twenty thousands of cells from each population and calculated, say, $f2(i)$ as the number of cells from the 2-nd population in the i-th interval divided by the total number of analyzed cells from that population.