02610 Optimization and Data Fitting Autumn 2015

Assignment 2

List all your Matlab code in an appendix!

The report must be handed in no later than monday december 07, 2014 at 13:00 to secretary Mette E. Larsen, office 105, building 303B. You must also upload a version to Campusnet. Write your group number (you will get a group number when you get report 1 back) on the front page of your report. Also your study numbers and names should appear on the frontpage of your report.

Problem 1: Data Fitting with Linear Models

In this problem we consider data fitting in the linear model

$$y_i = \alpha t_i + \beta + e_i \qquad e_i \sim N(0, \sigma^2) \tag{1}$$

The data for this problem is given in the file `Problem1Data.mat`. Note that these data contain outliers.

1.1 Plot the data

Plot the data in the file `Problem1Data.mat`. The true model for these data points are $y = t$ (i.e. $\alpha = 1.0$ and $\beta = 0.0$). Plot the true model in the same plot as the data.

1.2 Least Squares Estimation (ℓ_2-estimation)

- Formulate the least squares problem for this data fitting problem. Show that the solution corresponds to solution of an unconstrained quadratic program. Formulate the equations used for solving the problems.
- Compute the least squares solution.
- Make a new figure in which you plot the least squares solution in the same plot as the data and the true solution.
- Make a histogram of the errors.
- Make a histogram of the errors in which you have removed the points that you believe are outliers.
- Using the remaining points, estimate the standard deviation of the noise, the covariance matrix of the parameters, the standard deviation of the parameters, the 95% confidence interval of the parameters, and the 95% confidence interval of the predictions. Your results must be reported in appropriate tables and as plots in figures.

1.3 ℓ_1-estimation

- Formulate the ℓ_1-estimation problem for this data fitting problem. Show that the solution corresponds to solution of a linear program. Formulate the equations used for solving the problems.
- Compute the solution of this ℓ_1-estimation.
- Make a new figure in which you plot the ℓ_1-estimate in the same plot as the data and the true solution.
- Make a histogram of the errors.
- Make a histogram of the errors in which you have removed the points that you believe are outliers.
- Using the remaining points, estimate the standard deviation of the noise, the covariance matrix of the parameters, the standard deviation of the parameters, the 95% confidence interval of the parameters, and the 95% confidence interval of the predictions. Your results must be reported in appropriate tables and as plots in figures.

1.4 ℓ_∞-estimation

- Formulate the ℓ_∞-estimation problem for this data fitting problem. Show that the solution corresponds to solution of a linear program. Formulate the equations used for solving the problems.
- Compute the solution of this ℓ_∞-estimation.
- Make a new figure in which you plot the ℓ_∞-estimate in the same plot as the data and the true solution.
- Make a histogram of the errors.
- Make a histogram of the errors in which you have removed the points that you believe are outliers.
- Using the remaining points, estimate the standard deviation of the noise, the covariance matrix of the parameters, the standard deviation of the parameters, the 95% confidence interval of the parameters, and the 95% confidence interval of the predictions. Your results must be reported in appropriate tables and as plots in figures.

1.5 Huber-estimation

- Formulate the Huber-estimation problem for this data fitting problem. Show that the solution corresponds to solution of a quadratic program. Formulate the equations used for solving the problems.
- Compute the solution of this Huber-estimation.
- Make a new figure in which you plot the Huber-estimate in the same plot as the data and the true solution.
- Make a histogram of the errors.
- Make a histogram of the errors in which you have removed the points that you believe are outliers.
- Using the remaining points, estimate the standard deviation of the noise, the covariance matrix of the parameters, the standard deviation of the parameters, the 95% confidence interval of the parameters, and the 95% confidence interval of the predictions. Your results must be reported in appropriate tables and as plots in figures.

1.6 Comparison of Methods

Discuss the estimation methods that you have investigated in terms of the quality of the estimate (when there is outliers as in the case considered) and computational speed. You can for instance make a table in which you summarize the properties of each estimation method.

Problem 2: Enzyme Catalyzed Reaction

We consider parameter estimation in an enzyme catalyzed reaction. Enzyme catalyzed reactions are important in biochemistry and biotechnology.
Simple enzyme catalyzed conversion of a substrate, S, to a product, P, may be denoted

$$S + E \rightleftarrows ES \rightarrow P + E$$

E denotes enzymes and ES denotes the enzyme-substrate complex.
The rate of production formation may be described by the Michaelis-Menten kinetics

$$r = V_{\max} \frac{C_S}{K_M + C_S} \tag{2}$$

r is the reaction rate and C_S is the substrate concentration. The parameters in this expression are $V_{\max}$ and K_M. The value of these parameters depends on the reaction and the enzyme.
To simply notation we denote the model as

$$\hat{y} = f(\theta) = f(\theta; x) = \frac{\theta_1 x}{\theta_2 + x} \tag{3}$$

where $\hat{y}$ is the reaction rate predicted by the model, x is the substrate concentration, $\theta_1 = V_{\max}$ and $\theta_2 = K_M$. Let $\theta = [\theta_1; \theta_2]$.
Corresponding values of measured reaction rates, y, and enzyme concentration, x are given in the table below:

x	y
2.0000	0.0615
2.0000	0.0527
0.6670	0.0334
0.6670	0.0334
0.4000	0.0138
0.4000	0.0258
0.2860	0.0129
0.2860	0.0183
0.2220	0.0083
0.2200	0.0169
0.2000	0.0129
0.2000	0.0087

We want to use these data points to estimate the parameters θ.
Assume that the measured reaction rates, y, and the predicted reaction rates, $\hat{y}$, are related by

$$y = \hat{y} + e = f(\theta; x) + e \qquad e \sim N(0, \sigma^2) \tag{4}$$

e is measurement noise. We assume that the measurement noise is normally distributed. The variance, σ^2, of the measurement noise is unknown.

Problem 2.1

- Plot the experimental data (x, y).
- Plot the experimental data $(1/x, 1/y)$.
- Describe a method using *linear* least squares estimation of the parameters, θ. Write a Matlab program that implements your method.
- Estimate the parameters using the linear least squares method. Denote this estimate as θ^*_{LS}.
- Plot the predictions by the model along with your data.

Problem 2.2
In this problem we estimate the parameters using nonlinear optimization. In this method we determine the parameters, θ, by solution of

$$\min_{\theta} \quad \phi(\theta) \tag{5}$$

in which

$$\phi(\theta) = \frac{1}{2}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 = \frac{1}{2}\sum_{i=1}^{n} \|y_i - f(\theta; x_i)\|_2^2 \tag{6}$$

- Make a contour plot of $\phi(\theta)$. Plot θ^*_{LS} in this contour plot. Comment on your observations.
- Which numerical algorithm will you use for computation of the parameters, θ, and why? Describe advantages and disadvantages of your chosen method compared to other methods.
- Compute the parameters and denote the result, θ^*. Compute an estimate of σ^2. Compute the covariance-matrix and the 95% confidence interval for your estimate, θ^*.
- Plot the optimal estimate, θ^*, in your contour plot. Also plot the predicted reaction rates by the Michaelis-Menten model for the parameter estimate, θ^*. Plot your experimental data and the prediction from problem 2.1 in the same plot. Comment on your observations.
- Describe how you will choose a starting guess for nonlinear optimization algorithm. Try this starting guess and determine the number of iterations used to compute the optimal estimate.

Problem 2.3

In practice, it is often difficult to measure the reaction rate. Instead, the substrate concentration, x, is measured as function of time, t, for an experiment in which a given starting concentration, $x(t_0) = x_0$ with $t_0 = 0$, of substrate reacts in a batch reactor and is converted to product, P.

Let $x(t)$ denote the substrate concentration in the reactor at time t. The measured substrate concentration is now denoted $y(t)$. The measured substrate concentration, $y(t)$, at time $\{t_i\}_{i=0}^n$ is

$$y(t_i) = \hat{y}(\theta; t_i) + e(t_i) \qquad e(t_i) \sim N(0, \sigma^2) \tag{7}$$

where $\hat{y}(\theta; t_i) = x(t_i; \theta)$ is the substrate concentration in the reactor predicted by the model

$$\frac{dx(t)}{dt} = -\frac{\theta_1 x(t)}{\theta_2 + x(t)} \qquad x(t_0) = x_0 \tag{8}$$

The substrate concentration is $x(t) = x(t; \theta)$.

In this setup we determine the parameters by solution of

$$\min_{\theta} \quad \phi(\theta) = \frac{1}{2} \sum_{i=1}^{n} \|y(t_i) - \hat{y}(\theta; t_i)\|_2^2 \tag{9}$$

In the following we let $\frac{\partial}{\partial \theta}$ denote the operator

$$\frac{\partial}{\partial \theta} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} & \frac{\partial}{\partial \theta_2} \end{bmatrix}$$

The data $\{(t_i, y_i)\}_{i=0}^n$ for the experiment are provided in `MMBatchData.mat`. The initial substrate concentration at time $t_0 = 0$ is $x_0 = 10.0$.

- Plot the data in `MMBatchData.mat`
- Program a Matlab function that can compute $\hat{y}(\theta; t_i)$ for $\{t_i\}_{i=1}^n$. You must describe your function. Hint: You can use `ode45` in making this function.
- Make a contour plot of $\phi(\theta)$ using the data in the file `MMBatchData.mat`.
- Make a Matlab function that can compute, $\hat{y}(\theta; t_i)$ and $\frac{\partial}{\partial \theta}\hat{y}(\theta; t_i)$ for $\{t_i\}_{i=1}^n$. You must describe your function. Hint: You can use `ode45` in making this function.
- Estimate the parameters, θ, by solving the unconstrained optimization problem. You should argue for your choice of algorithm and comment on its performance. In particular you should comment on your choice of tolerances as well as any other option you set for the optimization algorithm and other numerical algorithms used in solving this parameter estimation problem. Estimate σ^2. Compute the covariance matrix of the parameter estimate and 95% confidence intervals for your parameters.