TANA09 Datatekniska beräkningar

Laboration 2. Linear algebra

Name:_____

LiU-Id:_____

Email:_____

Name:_____

LiU-ID:_____

Email:_____

Approved:_____ Sign:_____

Corrections:_____

1 Introduction

In Scientific Computing the most common operations are those from linear algebra. Solving partial differential equations and several image processing applications at its core means solving linear systems of equations. Many statistical data analysis applications involves solving linear least squares problems. Data mining or clustering applications often involve the singular value decomposition. Thus it is important to be familiar with numerical software for solving such problems and to be able to estimate time requirements and accuracy in the results.

2 LU-decomposition

The LU decomposition of a matrix can be written

$$PA = LU,$$

where both L and U are triangular and P is a permutation. In this exercise we will use the MATLAB function `lu` to investigate the properties and uses of said decomposition.

Exercise 2.1 Use the MATLAB function **lu** and compute the matrices L, R, and P corresponding to the matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 4 & 5 & 7 \end{pmatrix}$.

L= **U**= **P**=

Exercise 2.2 Compute the product $L \cdot U$ and describe the result.

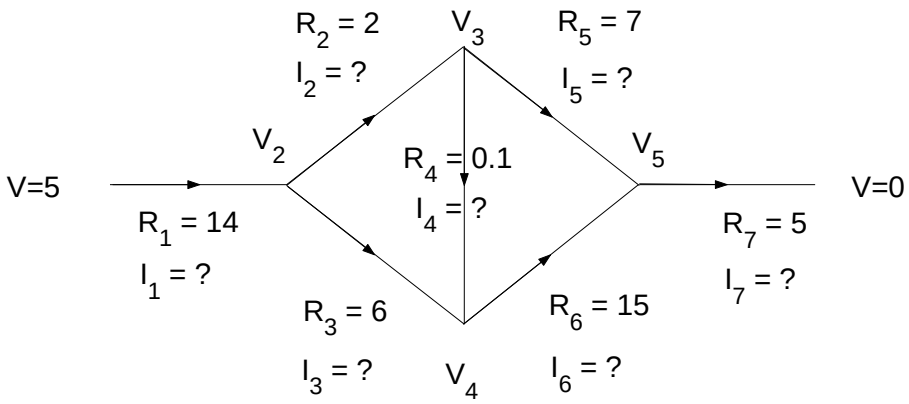
Answer:_____

Exercise 2.3 Select an exact solution `x=ones(3,1)` and compute the right hand side `b=A*x`. Solve the system of equations $Ax = b$ using the LU decomposition. The triangular equation systems should be solved using MATLABs \–operator (`help slash`).

Write down the MATLAB commands and also the numerical solution $\bar{x}$. How big is the difference between the numerical solution and exact solutions?

Svar:_____

3 An Electric Circuit



Figur 1: A simple electric circuit.

In this exercise we shall study the simple electric curcuit shown in Figure 1. Before manufacturing a circuit it is tested numerically. The objective is to calculate currents in the circuit. We will do this by formulating a linear system of equations.

The resistances $R_1, \ldots, R_7$ are given. We seek to compute the potentials $V_2, \ldots, V_5$, and the currents $I_1, \ldots, I_7$.

The physical laws that govern the circuit are Kirchhoffs Law, wich states that the sum of all currents at each node is zero, and Ohms law which says that the difference in potential between two nodes is $V_a - V_b = R_{ab} \cdot I_{ab}$.

Exercise 3.1 Use Kirchhoffs Law to formulate an equations for the currents I_2, I_4 and I_5. Also use Ohms Law to formulate an equation for the quantities V_3, V_4, I_4 and R_4? Are these equations linear?

Answer:_____

Exercise 3.2 Using Kirchhoffs and Ohms Laws we obtain a linear system of equations, $Ax = b$, where the unknowns are

$$x = (\, V_2\,,\, V_3\,,\, V_4\,,\, V_5\,,\, I_1\,,\, I_2\,,\, I_3\,,\, I_4\,,\, I_5\,,\, I_6\,,\, I_7\,)^T.$$

The Matlab program `ElectricCircuit.m` calculates the matrix and right hand side. Look at the matrix. Each row of the matrix corresponds to an equation. Which rows correspond to the equations you derived in the previous exercise? Also what is the dimension of the matrix A?

Answer:

Exercise 3.3 Solve the system of equations $Ax = b$ using Matlab. What are the computed values for the potential V_4 and the current I_5?

Answer:

Exercise 3.4 How is the condition number $\kappa(A)$ of a matrix A defined?

Answer:

Exercise 3.5 Compute the condition number of the matrix A in both 2-norm and max-norm. The same matrix A as previously should be used. Do you belive the computed solution is accurate?

$\kappa_2(A) =$ $\kappa_\infty(A) =$

Exercise 3.6 In this example we need to measure the two potentials $V = 5$ and $V = 0$. These go into the right hand side vector b. Suppose we have measurement errors and these quantities are only known with two correct digits. Give a bound for $\|\Delta b\|_\infty$ in this case. Also compute the corresponding bound for the error in the solution. Give both the formula and the value.

$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq$

Exercise 3.7 Suppose the correct values are $V = 5.0023$ and $V = 0.0047$. Change the right handside b accordingly and compute the new solution x_{new}. Compute $\|x_{new} - x\|_\infty$ and verify that the bound given by the previous error estimate holds.

Answer:

Exercise 3.8 (optional) The resistances $R_1, \ldots, R_7$ also have to be measured and will contain errors. Assume these are known to three correct digits. First compute the corresponding error $\|\Delta A\|_\infty$ and also the resulting error bound for the solution. Give both the formula and the value.

$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq$

Secondly edit the file `ElectricCircuit.m` and change the values for the resistances (make sure the new values are within the error range). Again compute the new solution and verify that $\|x_{new} - x\|_\infty$ is within the bound given by the error estimate.

Answer:

4 Least Squares and the Normal Equations

The datafile `EluxB.mat` contains the price for the Electrolux B stock from March 13th to January 2006. The stock price was recorded almost every day. The time vector $t \in [1, 2118]$, is the day.

Load the data and look at a plot of the data using `plot(t,EluxB,'-b')`.

The goal of this exercise is to fit a continuous function to the data and use the results to estimate future stock prices. The standard method to use is the least squares method.

Exercise 4.1 First we will attempt to model the data using a quadratic polynomial,

$$p(t) = c_0 + c_1(t - m) + c_2(t - m)^2.$$

The matrix can be created in Matlab using

```
>>A=[(t-m).^0 (t-m)  (t-m).^2];
```

where m has been chosen appropriately.

What is a good value for m? Either motivate your choice theoretically of experiment and form the matrix A and the normal equations A^TA for different values of m. The goal is to pick m so the condition number of the normal equations is kept at a minimum.

Answer:_____

Exercise 4.2 Form the normal equations and solve them. Give the values for the coefficients $c = (c_0\, c_1\, c_2)^T$. Also plot the data and the polynomial in the same graph. You evaluate the polynomial for an appropriate ranges of times using the Matlab commands:

```
>>tt=(1:2600)';
>>pp=[(tt-m).^0 (tt-m) (tt-m).^2]*c;
```

Note that the polynomial is computed for "future" values. Is the fit good? Is the progonosis good?

Answer:_____

Exercise 4.3 (optional) See what happens if you increase the degree of the polynomial to, e.g., $n = 6$. Does the polynomial fit the data better? Do you think the prognosis is improved?

Answer:_____

Hint Since the data is old the "future" stock price values are also available. The data file also contains data up to the november 2013. Plot these values in the same graph and you can see how good the prognosis is in practice.

There is no real reason to assume that a polynomial can accurately describe stock prices. A better model would be to combine a straight line, for the possible long term growth, and combinations of

sines and cosines for periodic variations. A period of one year leads to an angle $v(t) = \pi(t-m)/365$, with the previously defined m. The resulting model is

$$f(t) = c_1 + c_2(t-m) + c_3 \sin(\, v(t)\,) + c_4 \cos(\, v(t)\,) + c_5 \sin(\, 2v(t)\,) + c_6 \cos(\, 2v(t) +$$

$$c_7 \sin(3v(t)) + c_8 \cos(3v(t)).$$

Exercise 4.4 Determine the coeffients $c_1, \ldots, c_8$ using the least squares method. Plot the function $f(t)$, for $1 \leq t \leq 2600$, and the given data in the same graph. Does the prognosis look resonable?

Answer:

Remark This is actually a method that is used for analysing stock prices. The trick is to find a resonable model. Typically the data fit always looks good. Most models captures the data used to fit the model parameters very well. For a prognosis to be successful there has to be some additional reason why the model is good for this particular data set.

5 Real time tracking of a comet trajectory

Most celestial objects, e.g. planets or comets, moves along an elliptial orbits. When a new comet is first observed one wants to estimate its path through the solar system to make sure it will not crash into the earth. This is done by collecting a set of positional data and finding an ellipse that fit the data in the least squares sense.

A point $(x, y)^T$ on an ellispe satisfies an equation

$$c_1 x^2 + c_2 xy + c_3 y^2 + c_4 x + c_5 y + 1 = 0.$$

for a specific set of parameters $c = (c_1, c_2, c_3, c_4, c_5)^T$. Not every such equation describes an ellipse. Depending on the values of the coefficients you may get a hyperbola or parabola.

Exercise 5.1 Suppose a set of observed locations (x_k, y_k), $k = 1, .., n$, for the comet is given. Write down an over determined linear system $Ac = b$ that can be solved to find the shape of the ellipse.

Answer:

Exercise 5.2 On the course library there is a file `CometTracking.m`. Open that file in the editor and add code that creates the matrix A and right hand side b. Also add code for solving the problem using the QR decomposition.

Run the program by tying

```
>> CometTracking();
```

The program attempts to use $n = 10$ observations of the comet to find the elliptic orbit. Note that by only using observations that are close together and that only covers a small part of the ellipse we get an ill–conditioned problem. Was the attempt to find an ellipse successful? What is the condition number of the matrix R?

Answer:_____

Exercise 5.3 We will improve the ellipse fitting by adding more observations of the comet position. Change the obswervation times so that

```
Times =  0:0.05:3.0;
```

How many observations are used now? What is the condition number of the R matrix? Was the fit successful?

Answer:_____

Exercise 5.4 Instead of using more points we will spread out the observations more in time. Change the code to

```
Times =  0:0.5:5.0;
```

How many observations are used now? What is the condition number of the R matrix? Was the fit successful?

Answer:_____

Exercise 5.5 From the exercises above. What is more important: Having alot of observations or having well spread out observations?

Answer:_____

Remark This is the background to the periodically occuring disaster warnings in newspapers: If you spot a new comet initially only observations during a short time window are available. This means that the least squares problem for the elipse fit is ill-conditioned and the coefficients can't be determined very well. That leads to a huge number of possible trajectories for the comet and you can't rule out that one of these possible trajectories might hit the earth. After a few days you have observations better spread out in time and the least squares problem is well-conditioned and the elipse fit is much better.

6 Automatic Character Recognition

Often in applications you study objects that occur in various types, or classes. One is then intressted in determining the particular type, or class, of a certain unknown object. This is called the *Classification problem* and is used in many technical applications. Examples include mail filters where a particular email is either junk or it is not. Another application is the automatic sorting of mail done by the postal services. Each envelope is photographed and the postal code identified. The individual digits each has a type, i.e. one of the characters 0,1,...,9. An automatic algorithm is then used to identify each digit; and the mail can be sent to the correct post office for distribution.

In this exercise we will study the problem of automatically reading handwritten digits. We are given a number of 16×16 pixel bitmap images that were all scanned from actual mail sent through the US Postal Service. Each image represents one handwritten digit. The images are divided into two sets. First there is a *Reference set* which is used to create the digit recognition algorithm. Second there is the *Test set* that is used to evaluate the algorithm.

Exercise 6.1 The images are available as a file `DataSet.mat`. Load the data into Matlab. There are two matrices `RefSet` and `TestSet` that contain the reference digits and the test digits respectively. Each column in the matrices represents one indivudual digit.

Use the program `DisplayDigit` to look at a couple of images. The vectors `RefAns` and `TestAns` contain the actual digit that the images are supposed to represent. So that `RefAns(k)` tells what digit is represented by the image `RefSet(:,k)`.

The *Reference set* is used to create the character recognition algorithm. The idea is to identify common traits for the different classes, i.e. the different digits. The basic idea is that we split the reference set into smaller subsets that collect digits of a specific type. Then we implement a function that measures how much a given unknown digit has with the digits in the different subsets.

The character recognition algorithm will be implemented using the following steps. Suppose that the matrix `Rj` collects all digits from the reference set that represent the digit j. We then compute the SVD,

```
>> [Uj,Sj,Vj]=svd(Rj);
```

Most of the digits in the reference set look very similar. There are a few different styles of writing but mostly all the handwritten digits are only small variations of others. This means that the collection of digits of a certain type can be well approximayed by a low rank approximation. The following idea can be used to compute the distance from an unknown digit D to one of the subspaces that collect digits of a single type.

- Approximate the space $\text{span}(R_j)$ by a subspace $\text{span}(Uj(:, 1:k)$ for a fixed number k.
- Compute the orthogonal distance from D to the linear subspace $\text{span}(Uj(:, 1:k)$.
- The unknown digit D is of the same type as the closest subspace.

Exercise 6.2 Implement a function

```
>> TheSubspaces = CreateSubspace( RefSet, RefAns , k );
```

that computes the subspaces span($Uj(:, 1:k)$, for $j = 0, 1, 2 \ldots, 9$. This function may be extremely costly to run but since it can be precomputed and reused for each classification this is not an issue.

Hint You need 10 different matrices U_j; each with a subspace corresponding to one of the digits 0-9. Its convinient to let the output parameter `TheSubspaces` be a `cell`-array and each cell contain one of the matrices $U_j(:, 1:k)$. Read the documentation about cell arrays.

Exercise 6.3 Implement a function

```
>> d = DistanceFromSubspace( Uj(:,1:k) , TestDigit );
```

that computes the distance as detailed above. What is the appropriate formula that should be implemented?

Answer:_____

Hint: The orthogonal distance between a vector and a linear subspace can be formulated as a least squares problem.

Exercise 6.4 Write a Matlab function `ClassifyDigit` That uses the rank k subspaces created above, and uses the orthogonal distance from an unknown digit to the respektive subspaces to classify an unkown digit from the *test set*, i.e. a function

```
>> Type = ClassifyDigit( S , TheSubspaces );
```

Exercise 6.5 Use the `ClassifyDigit` function to determine the type of all the unknown digits in the matrix `TestSet`. Experiment to find a good value for the dimension k of the subspaces used. How many digits from the test set are classified correctly?

Answer:_____

Remark The test set was collected from actual mail. But its not realistic in the sense that many of the "easy" cases were removed thus making this test set much more difficult than what you'd expect from the application. This is to make the difference in performance bigger between different algorithm.

Exercise 6.6 (**Optional**) Pick a class of digits and plot the corresponding singular values. Measure the "energy" contained in a subspace span($Uj(:, 1:k)$ as

$$E_k = \frac{\sqrt{\sigma_1^2 + \ldots + \sigma_k^2}}{\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}}.$$

This can roughly be interpreted as follows: During the classification we will use the digits from the *reference set* as a basis and approximate an unknown digit from the *test set*. If the energy of the k–dimensional subspace is 90% then the subspace is about 90% as effective for representing digits from the testset as the whole space.

Suppose we want each subaspace to have energy $E_k \geq 0.9$. How many basis vectors do we need for each type of digit? Are some digits significantly harder to classify than others?

Answer: