

STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2022
PROBLEM SET 2

For the coding problems, use any program you like but present your codes and results in a way that is comprehensible to someone who is unfamiliar with that program (e.g. comment your codes appropriately).

1. The files required for this problem can be found in the subfolder `hw2` under ‘Files’ in Canvas or at <http://www.stat.uchicago.edu/~lekheng/courses/309/stat309-hw2/>. The matrix in `processed.mat` (Matlab format) or `processed.txt` (comma separated, plain text) is a 49×7 matrix where each row is indexed by a country in `row.txt` and each column is indexed by a demographic variable in `column.txt`, ordered as in the respective files. So for example, if we denote the matrix by

$$A = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_{49}^\top \end{bmatrix} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_7] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{17} \\ a_{21} & a_{22} & \cdots & a_{27} \\ \vdots & \vdots & \ddots & \vdots \\ a_{49,1} & a_{49,2} & \cdots & a_{49,7} \end{bmatrix} \in \mathbb{R}^{49 \times 7},$$

then $a_{23} = -0.2743$ is Austria’s population per square kilometers (row index 2 = Austria, column index 3 = population per square kilometers). As you probably notice, this matrix has been slightly preprocessed. If you want to see the raw data, you can find them in `raw.txt` (e.g. the actual value for Austria’s population per square kilometers is 84) but you don’t need the raw data for this problem.

- (a) Show that to plot the projections of the row vectors (i.e., samples) $\mathbf{a}_1, \dots, \mathbf{a}_{49} \in \mathbb{R}^7$ onto the two-dimensional subspace $\text{span}\{\mathbf{v}_j, \mathbf{v}_k\} \cong \mathbb{R}^2$, we may simply plot the n points

$$\{(\sigma_j u_{ij}, \sigma_k u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, 49\}$$

where $U = [u_{ij}] \in \mathbb{R}^{49 \times 49}$ and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{49 \times 7}$ are the matrix of left singular vectors and matrix of singular values respectively.

- (b) Find the first two right singular vectors of A , $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^7$. Project the data onto the two-dimensional space $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\} \cong \mathbb{R}^2$. Plot this in a graph where the x - and y -axes correspond to \mathbf{v}_1 and \mathbf{v}_2 respectively and where the points correspond to the countries — label each point by the country it corresponds to. Identify the two obvious outliers.
- (c) Now do the same with the two left singular vectors of A , $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^{49}$. Project the column vectors (i.e., variables) $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_7 \in \mathbb{R}^{49}$ onto the two-dimensional space $\text{span}\{\mathbf{u}_1, \mathbf{u}_2\} \cong \mathbb{R}^2$ and plot this in a graph as before. Note that in this case, the points correspond to the demographic variables — label them accordingly.
- (d) Overlay the two graphs in (b) and (c). Identify the two demographic variables near the two outlier countries — these explain why the two countries are outliers.
- (e) Remove the two outlier countries and redo (b) with this 47×7 matrix. This allows you to see features that were earlier obscured by the outliers. Which two European countries are most alike Japan?

The graphs¹ in (b) and (c) are called *scatter plots* and the overlaid one in (d) is called a *biplot*. See <http://en.wikipedia.org/wiki/Biplot> for more information. The reason we didn't need to adjust the scale of the axes using the singular values of A like in the Wikipedia description is because the preprocessing has taken care of the scaling; if we had started from the raw data, then we would need to deal with this complication.

2. Let $A, B \in \mathbb{R}^{m \times n}$ where A has full column rank.

(a) Show that

$$\min_{X \in \mathbb{R}^{n \times m}} \|AX - I_m\|_F$$

has a unique solution. What is the minimum length solution, i.e., where $\|X\|_F$ is minimum?

- (b) Show that the following method produces a symmetric matrix $X \in \mathbb{R}^{n \times n}$ that solves

$$\min_{X^T=X} \|AX - B\|_F. \quad (2.1)$$

- (i) Show that the SVD of A takes the form

$$A = U \begin{bmatrix} \Sigma \\ O \end{bmatrix} V^T$$

where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are unitary matrices and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n \times n}$ is a diagonal matrix.

- (ii) Show that

$$\|AX - B\|_F^2 = \|\Sigma Y - C_1\|_F^2 + \|C_2\|_F^2$$

where $Y = V^T X V$ and $C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = U^T B V$.

- (iii) Note that Y must be symmetric if X is. Show that

$$\|\Sigma Y - C_1\|_F^2 = \sum_{i=1}^n |\sigma_i y_{ii} - c_{ii}|^2 + \sum_{j>i} (|\sigma_i y_{ij} - c_{ij}|^2 + |\sigma_j y_{ji} - c_{ji}|^2)$$

and deduce that the minimum value of (2.1) is attained when

$$y_{ij} = \frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}$$

for all $i, j = 1, \dots, n$.

- (c) Now emulate the previous part to find a rank- r matrix $X \in \mathbb{R}^{m \times n}$ that solves

$$\min_{\text{rank}(X) \leq r} \|AX - B\|_F.$$

3. Let $A \in \mathbb{C}^{m \times n}$ and $\mathbf{b} \in \mathbb{C}^m$. We will discuss a variant of $A\mathbf{x} \approx \mathbf{b}$ where the error occurs only in A . Note that in ordinary least squares we assume that the error occurs only in \mathbf{b} while in total least squares we assume that it occurs in both A and \mathbf{b} .

- (a) Show that if $0 \neq \mathbf{x} \in \mathbb{C}^n$, then

$$\left\| A \left(I - \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}} \right) \right\|_F^2 = \|A\|_F^2 - \frac{\|A\mathbf{x}\|_2^2}{\mathbf{x}^*\mathbf{x}}.$$

- (b) Show that the matrix

$$E = \frac{(\mathbf{b} - A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}} \in \mathbb{C}^{m \times n}$$

¹One point to observe is that all the information needed for all three plots are already contained in the SVD of A , i.e., in U , Σ , and V ; it is just a matter of deciding which numbers to plot against which numbers.

has the smallest 2-norm among all $E \in \mathbb{C}^{m \times n}$ that satisfy

$$(A + E)\mathbf{x} = \mathbf{b}.$$

- (c) Let A , \mathbf{b} , and \mathbf{x} be given and fixed. What are the solutions of

$$\min_{(A+E)\mathbf{x}=\mathbf{b}} \|E\|_2 \quad \text{and} \quad \min_{(A+E)\mathbf{x}=\mathbf{b}} \|E\|_F$$

where the minimum is taken over all $E \in \mathbb{C}^{m \times n}$ such that $(A + E)\mathbf{x} = \mathbf{b}$?

- (d) Given $\mathbf{a} \in \mathbb{C}^n$, $\mathbf{b} \in \mathbb{C}^m$, and $\delta > 0$. Show how to solve the problems

$$\min_{\|E\|_F \leq \delta} \|E\mathbf{a} - \mathbf{b}\|_2 \quad \text{and} \quad \max_{\|E\|_F \leq \delta} \|E\mathbf{a} - \mathbf{b}\|_2$$

over all $E \in \mathbb{C}^{m \times n}$.

4. In the following, $\kappa(A) := \|A\| \|A^\dagger\|$ for $A \in \mathbb{C}^{m \times n}$ where $\|\cdot\|$ denotes a submultiplicative matrix norm. We will write $\kappa_p(A)$ if the norm involved is a matrix p -norm.

- (a) Show that for any nonzero $A \in \mathbb{C}^{m \times n}$,

$$\kappa(A) \geq 1.$$

- (b) Show that for any $A \in \mathbb{C}^{m \times n}$,

$$\kappa_2(A^* A) = \kappa_2(A)^2$$

but that in general

$$\kappa(A^* A) \neq \kappa(A)^2.$$

- (c) Show that for nonsingular $A, B \in \mathbb{C}^{n \times n}$,

$$\kappa(AB) \leq \kappa(A)\kappa(B).$$

Is this true in general without the nonsingular condition?

- (d) Let $Q \in \mathbb{C}^{m \times n}$ be a matrix with orthonormal columns. Show that

$$\kappa_2(Q) = 1.$$

Is this true if Q has orthonormal rows instead? Is this true with κ_1 or κ_∞ in place of κ_2 ?

- (e) Let $R \in \mathbb{C}^{n \times n}$ be a nonsingular upper-triangular matrix. Show that

$$\kappa_\infty(R) \geq \frac{\max_{i=1,\dots,n} |r_{ii}|}{\min_{i=1,\dots,n} |r_{ii}|}.$$

- (f) Show that for any nonsingular $A \in \mathbb{C}^{n \times n}$,

$$\kappa(A) \geq \max \left\{ \frac{\|AX - I\|}{\|XA - I\|}, \frac{\|XA - I\|}{\|AX - I\|} \right\}.$$

(Hint: $AX - I = A(XA - I)A^{-1}$.)

5. We will examine the effect of various parameters on the accuracy of a computed solution to a nonsingular linear system. Relevant commands in Matlab syntax are given in brackets.

- (a) Generate $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ as follows:

- (i) a_{ij} randomly generated from a standard normal distribution [`randn(n)`];
- (ii) a Hilbert matrix, i.e., $a_{ij} = 1/(i + j - 1)$ [`hilb(n)`];
- (iii) a Pascal matrix, i.e., the entries $a_{ij} = \binom{i+j}{i}$ [`pascal(n)`];
- (iv) a magic square, i.e., the entries a_{ij} 's are the integers $1, 2, \dots, n^2$ arranged in a way that A has equal row, column, and diagonal sums [`magic(n)`].

$$\text{hilb}(4) = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix}, \quad \text{pascal}(4) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}, \quad \text{magic}(4) = \begin{bmatrix} 16 & 2 & 3 & 13 \\ 5 & 11 & 10 & 8 \\ 9 & 7 & 6 & 12 \\ 4 & 14 & 15 & 1 \end{bmatrix}$$

For simplicity, we will assume that A is stored exactly with no errors even though this is only true for those matrices with integer-valued entries.

- (b) Generate \mathbf{x} and $\mathbf{b} \in \mathbb{R}^n$ as follows:
 - (i) $\mathbf{x} = [1, \dots, 1]^\top [\text{ones}(n, 1)]$;
 - (ii) $\mathbf{b} = A\mathbf{x}$ [$\mathbf{b} = A*\mathbf{x}$].
- (c) For each A generated as above, perform the following for $n = 5, 10, 15, \dots, 500$.
 - (i) Solve $A\mathbf{x} = \mathbf{b}$ using your program to get $\hat{\mathbf{x}}$ [$\mathbf{xhat} = A \setminus \mathbf{b}$]. Note that in general the result computed by your program will not be exactly the true solution $\mathbf{x} = A^{-1}\mathbf{b}$ because of roundoff errors that occurred during computations.
 - (ii) Compute $\Delta\mathbf{b} = A\hat{\mathbf{x}} - \mathbf{b}$ and record the values of $\|\mathbf{x} - \hat{\mathbf{x}}\|/\|\mathbf{x}\|$, $\kappa(A) = \|A\|\|A^{-1}\|$ and $\kappa(A)\|\Delta\mathbf{b}\|/\|\mathbf{b}\|$ for $\|\cdot\| = \|\cdot\|_1, \|\cdot\|_2$, and $\|\cdot\|_\infty$.
 - (iii) Present everything for the $n = 5$ case but only tabulate the relevant trend for general $n > 5$ in a graph.
- (d) Discuss and explain the effects of different choices of A , \mathbf{b} , $\|\cdot\|$, and n have on the accuracy of the computed solution $\hat{\mathbf{x}}$.
- (e) Instead of solving the linear system directly, compute A^{-1} and then $\hat{\mathbf{x}} := A^{-1}\mathbf{b}$ [$\mathbf{xhat} = \text{inv}(A)*\mathbf{b}$]. Comment on the accuracy of this approach. Provide numerical evidence to support your conclusion.
- (f) Write a program that computes the $(1, 1)$ -entry of the matrix A^{-1} that does not involve computing A^{-1} , i.e., if $A^{-1} = [b_{ij}]$, you want the value b_{11} but you are not allowed to compute A^{-1} .

6. Let $A \in \mathbb{R}^{n \times n}$ be nonsingular and let $\mathbf{0} \neq \mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{x} = A^{-1}\mathbf{b} \in \mathbb{R}^n$. In the following, $\Delta A \in \mathbb{R}^{n \times n}$ and $\Delta\mathbf{b} \in \mathbb{R}^n$ are some arbitrary matrix and vector. We assume that the norm on A satisfies $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$ for all $A \in \mathbb{R}^{n \times n}$ and all $\mathbf{x} \in \mathbb{R}^n$.

- (a) Show that if $\Delta A \in \mathbb{R}^{n \times n}$ is any matrix satisfying

$$\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\kappa(A)}, \quad (6.2)$$

then $A + \Delta A$ must be nonsingular. (*Hint*: If $A + \Delta A$ is singular, then there exists nonzero \mathbf{v} such that $(A + \Delta A)\mathbf{v} = \mathbf{0}$).

- (b) Suppose $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$. Show that

$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A) \frac{\|\Delta A\|}{\|A\|}. \quad (6.3)$$

- (c) Suppose $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ and (6.2) is satisfied. Show that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A) \frac{\|\Delta A\|}{\|A\|}}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}}.$$

You may like to use the following outline:

- (i) Show that

$$\Delta\mathbf{x} = -A^{-1}\Delta A\hat{\mathbf{x}}$$

and so

$$\|\Delta\mathbf{x}\| \leq \kappa(A) \frac{\|\Delta A\|}{\|A\|} (\|\mathbf{x}\| + \|\Delta\mathbf{x}\|).$$

(ii) Rewrite this inequality as

$$\left(1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}\right) \|\Delta \mathbf{x}\| \leq \kappa(A) \frac{\|\Delta A\|}{\|A\|} \|\mathbf{x}\|$$

and use (6.2).

(d) Suppose $(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b} + \Delta \mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b} + \Delta \mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta \mathbf{x} \neq \mathbf{0}$. Show that

$$\frac{\|\Delta \mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \mathbf{b}\|}{\|\hat{\mathbf{b}}\|} + \frac{\|\Delta A\|}{\|A\|} \frac{\|\Delta \mathbf{b}\|}{\|\hat{\mathbf{b}}\|} \right). \quad (6.4)$$

You may like to use the following outline:

(i) Show that

$$\Delta \mathbf{x} = A^{-1}(\Delta \mathbf{b} - \Delta A \hat{\mathbf{x}})$$

and so

$$\frac{\|\Delta \mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \mathbf{b}\|}{\|A\| \|\hat{\mathbf{x}}\|} \right). \quad (6.5)$$

(ii) Show that

$$\frac{1}{\|\hat{\mathbf{x}}\|} \leq \frac{\|A\| + \|\Delta A\|}{\|\hat{\mathbf{b}}\|}. \quad (6.6)$$

(iii) Combine (6.5) and (6.6) to get (6.4).

(e) Suppose $(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b} + \Delta \mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b} + \Delta \mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta \mathbf{x} \neq \mathbf{0}$ and (6.2) is satisfied. Use the same ideas in (b) to deduce that

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|} \right)}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}}.$$