STAT 30900 : MATHEMATICAL COMPUTATIONS I

Fall 2019

•

Homework 2

•

Solutions by

Jinhong Du

12243476

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 are in `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 $\operatorname{span}\{\mathbf{v}_j, \mathbf{v}_k\} \cong \mathbb{R}^2$, we may simply plot the n points

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

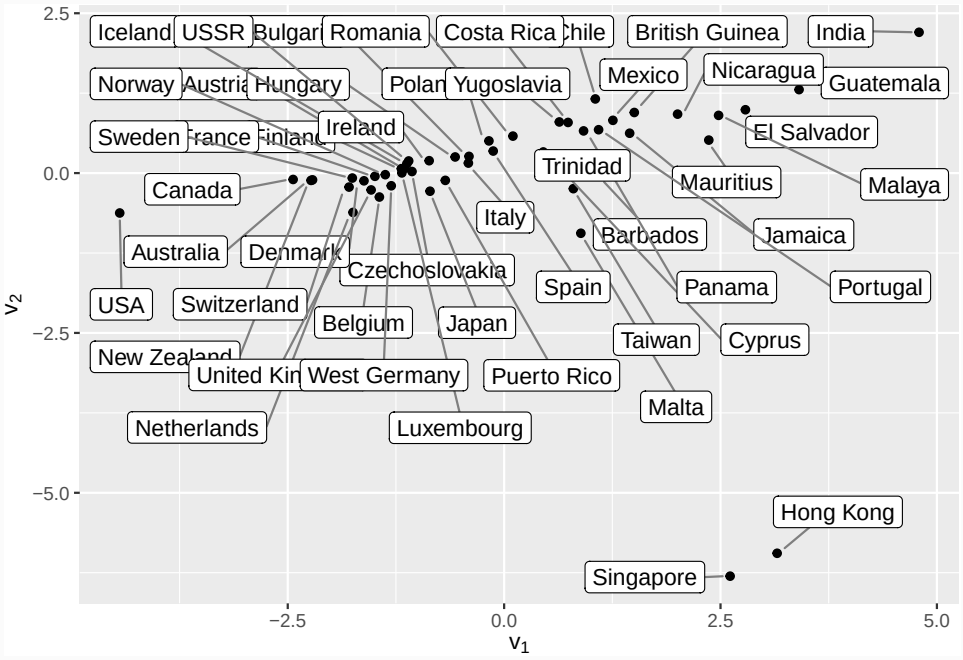
where $U = [u_{ij}] \in \mathbb{R}^{49\times 49}$ and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{49\times 7}$ are the matrix of left singular vectors and matrix of singular values respectively.

Proof. Let $A = U\Sigma V^*$ be the full singular value decomposition, where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$ ($r = \operatorname{rank}(A)$).
Let $\mathbf{v}_j$ be the jth column of V. Since V is unitary, $\operatorname{span}\{\mathbf{v}_j, \mathbf{v}_k\} \cong \mathbb{R}^2$ for $j \neq k$.
Project the rows of A to $\mathbf{v}_j$, we have

$$\begin{aligned} A\mathbf{v}_j &= U\Sigma V^* \mathbf{v}_j \\ &= U\Sigma \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_j \\ &= (U\Sigma)_j \\ &= \sigma_j \mathbf{u}_j \end{aligned}$$

where $\mathbf{u}_j$ is the jth column of U, i.e., $\mathbf{a}_i$ is projected to $\sigma_j u_{ij}$.
Therefore, $\mathbf{a}_i$ is projected to $(\sigma_j u_{ij}, \sigma_k u_{ik})$ in $\operatorname{span}\{\mathbf{v}_j, \mathbf{v}_k\}$. □

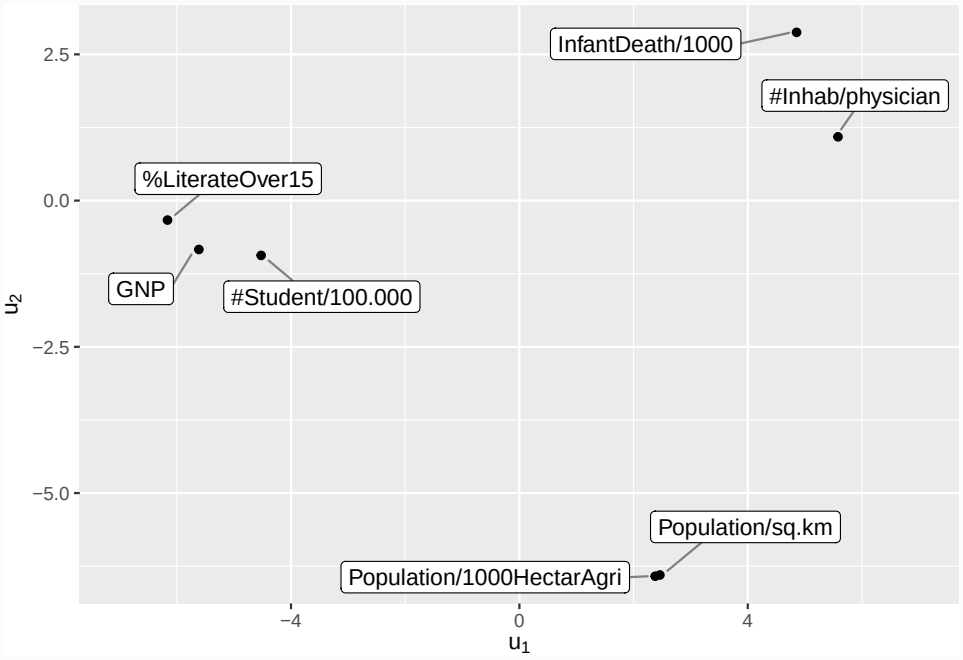
(b) Find the first two right singular vectors of A, $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^7$. Project the data onto the two-dimensional space $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\} \cong \mathbb{R}^2$. Plot this in a graph where the x- and y-axes correspond to $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively and where the points correspond to the countries — label each point by the country it corresponds to. Identify the two obvious outliers.



The two obvious outlier is Hong Kong and Singapore.

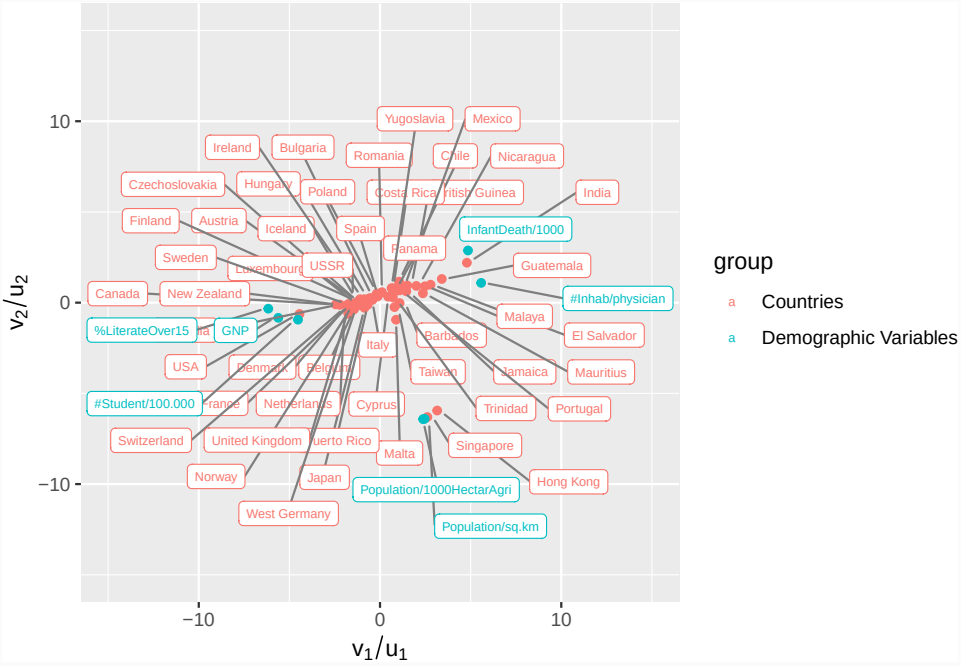
```
library(R.matlab)
library(ggplot2)
library(latex2exp)
library(ggrepel)

# Load data and remove whitespace of strings.
A <- readMat('processed.mat')$data
row_name <- read.table('row.txt', header = FALSE, sep = "\n", dec = ".")$V1
row_name <- trimws(row_name)
col_name <- read.table('column.txt', header = FALSE, sep = "\n", dec = "*",
    comment.char = '*')$V1
col_name <- trimws(col_name)

SVD <- svd(A, nu=nrow(A), nv=ncol(A))
df1 <- data.frame(x=SVD$d[1]*SVD$u[,1], y=SVD$d[2]*SVD$u[,2], label=row_name)
ggplot(df1, aes(x=x, y=y, label=label)) + geom_point() +
    geom_label_repel(aes(label = label),
                     box.padding   = 0.35,
                     point.padding = 0.5,
                     segment.color = 'grey50') +
    xlab(TeX("$v_1$")) + ylab(TeX("$v_2$"))
```

(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) $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_7 \in \mathbb{R}^{49}$ 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.




```
df2 <- data.frame(x=SVD$d[1]*SVD$v[,1], y=SVD$d[2]*SVD$v[,2], label=col_name)
ggplot(df2, aes(x=x, y=y, label=label)) + geom_point() +
    geom_label_repel(aes(label = label),
                     box.padding   = 0.35,
                     point.padding = 0.5,
                     segment.color = 'grey50') +
    xlab(TeX("$u_1$")) + ylab(TeX("$u_2$"))  + xlim(c(-7,7))
```

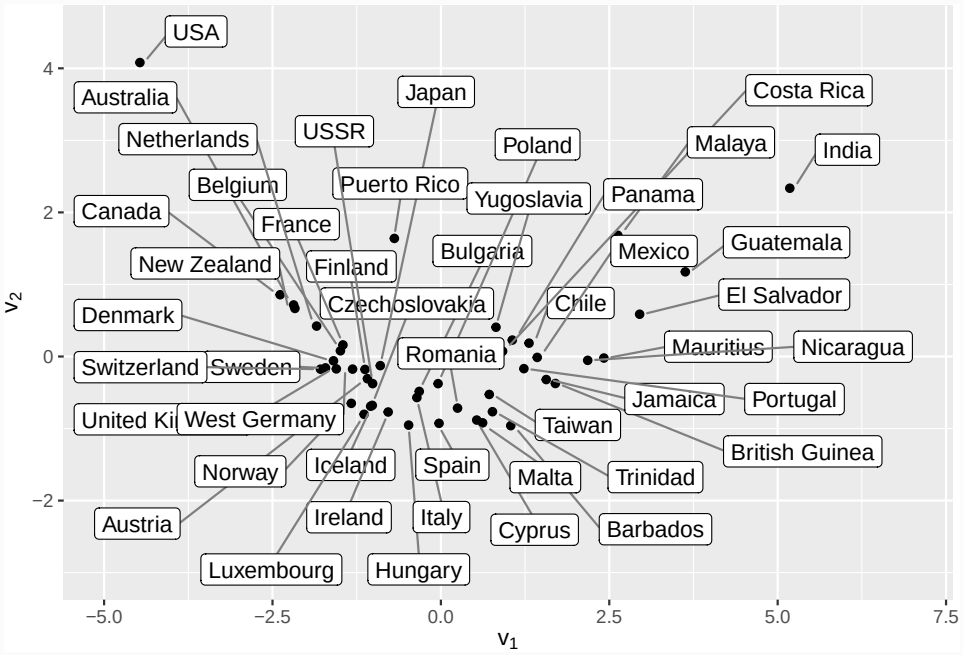
(d) Overlay the two graphs in (b) and (c). Identify the two demographic variables near the two outlier countries — these explain why the two countries are outliers.



The two demographic variables near the two outlier countries are "Population/100HectarAgri" and "Population/sq.km". The two countries are outliers because they are not agricultural countries, so "Population/100HectarAgri" is larger than other countries in the data set. Also, they have higher-density population than other countries.

```
df1['group']='Countries'
df2['group']='Demographic Variables'
df <- rbind(df1,df2)
ggplot(df, aes(x=x, y=y, label=label, group=group, color=group)) + geom_point() +
    geom_label_repel(aes(label = label, group=group),
                     box.padding   = 0.2,
                     size          = 2,
                     point.padding = 0.5,
                     segment.color = 'grey50') +
    xlab(TeX("$v_1$/$u_1$")) + ylab(TeX("$v_2/$u_2$$")) +
    xlim(c(-15,15)) + ylim(c(-15,15))
```

(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?



Finland and Austria are two European countries most alike Japan.

```
A_1 <- A[-which(row_name %in% c('Hong Kong', 'Singapore')),]
row_name_1 <- row_name[-which(row_name %in% c('Hong Kong', 'Singapore'))]
SVD <- svd(A_1, nu=nrow(A_1), nv=ncol(A_1))
df <- data.frame(x=SVD$d[1]*SVD$u[,1], y=SVD$d[2]*SVD$u[,2], label=row_name_1)
ggplot(df, aes(x=x, y=y, label=label)) + geom_point() +
    geom_label_repel(aes(label = label),
                     box.padding   = 0.35,
                     point.padding = 0.5,
                     segment.color = 'grey50') +
    xlab(TeX("$v_1$")) + ylab(TeX("$v_2$")) +
    xlim(c(-5,7)) + ylim(c(-3,4.5))
```

The graphs[1] in (b) and (c) are called *scatter plots* and the overlayed 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.

[1] One 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.

2

Let $A, B \in \mathbb{C}^{m\times n}$ with $n \le m$. In the lectures, we claim that the solution to

$$\min_{X^*X=I} \|A - BX\|_F$$

is given by $X = UV^*$ where $B^*A = U\Sigma V^*$ is its singular value decomposition. Here we will prove it and consider some variants.

(a) Show that

$$\|A - BX\|_F^2 = \operatorname{tr}(A^*A) + \operatorname{tr}(B^*B) - 2\operatorname{Re}\operatorname{tr}(X^*B^*A)$$

and deduce that the minimization problem is equivalent to

$$\max_{X^*X=I} \operatorname{Re}\operatorname{tr}(X^*B^*A).$$

Proof. Since $XX^* = (X^*X)^* = I$, we have

$$\begin{aligned}
\|A - BX\|_F^2 &= \operatorname{tr}((A - BX)^*(A - BX)) \\
&= \operatorname{tr}(A^*A + X^*B^*BX - A^*BX - X^*B^*A) \\
&= \operatorname{tr}(A^*A) + \operatorname{tr}(X^*B^*BX) - [\operatorname{tr}((X^*B^*A)^*) + \operatorname{tr}(X^*B^*A)] \\
&= \operatorname{tr}(A^*A) + \operatorname{tr}(B^*BXX^*) - 2\operatorname{Re}\operatorname{tr}(X^*B^*A) \\
&= \operatorname{tr}(A^*A) + \operatorname{tr}(B^*B) - 2\operatorname{Re}\operatorname{tr}(X^*B^*A).
\end{aligned}$$

Since A and B are given matrix, to minimize $\|A - BX\|_F^2$ such that $X^*X = I$, is equivalent to

$$\max_{X^*X=I} \operatorname{Re}\operatorname{tr}(X^*B^*A).$$

□

(b) Show that

$$\operatorname{Re}\operatorname{tr}(X^*B^*A) \le \sum_{i=1}^{n} \sigma_i(B^*A)$$

for any unitary X. When is the upper bound attained?

Proof. Let $B^*A = U\Sigma V^*$ be the singular value decomposition where $U, V \in \mathbb{C}^{n\times n}$ are unitary matrices and $\Sigma \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal entries $\sigma_1(B^*A) \ge \cdots \ge \sigma_n(B^*A) \ge 0$.
For any unitary $X \in \mathbb{C}^{n\times n}$, since $X^*X = I$, we have the ith column of X satisfies $\mathbf{x}_i^*\mathbf{x}_i = \|\mathbf{x}_i\|_2^2 = 1$, which implies that $|X_{ij}| \le 1$ for all i, j.
For any unitary $X \in \mathbb{C}^{n\times n}$,

Solution (cont.)

$$\begin{aligned}
\operatorname{Re}\operatorname{tr}(X^*B^*A) &= \operatorname{Re}\operatorname{tr}(X^*U\Sigma V^*) \\
&\overset{Y=U^*X}{=\!=\!=} \operatorname{Re}\operatorname{tr}(Y^*\Sigma V^*) \\
&= \operatorname{Re}\operatorname{tr}(V^*Y^*\Sigma) \\
&\overset{Z=YV}{=\!=\!=} \operatorname{Re}\operatorname{tr}(Z^*\Sigma) \\
&= \operatorname{Re}\left(\sum_{i=1}^{n}\sigma_i(B^*A)Z_{ii}\right) \\
&= \sum_{i=1}^{n}\sigma_i(B^*A)\operatorname{Re}Z_{ii} \\
&\le \sum_{i=1}^{n}\sigma_i(B^*A)|Z_{ii}| \\
&\le \sum_{i=1}^{n}\sigma_i(B^*A)
\end{aligned}$$

since $Z = U^*XV$ is also an unitary matrix. The equality holds when

$$Z_{ii} = 1 \text{ for all } i = 1, \dots, n. \tag{1}$$

Since Z is an unitary matrix, $\sum_{j=1}^{n}|Z_{ij}|^2 = 1$, (1) equals to

$$Z_{ij} = \delta_{ij} \text{ for all } i, j = 1, \dots, n. \tag{2}$$

i.e., $Z = U^*XV = I$, i.e. $X = UV^*$.

□

(c) Show that

$$\min_{X^*X=I}\|A - BX\|_F^2 = \sum_{i=1}^{n}(\sigma_i(A)^2 - 2\sigma_i(B^*A) + \sigma_i(B)^2).$$

Proof. From (b), we have

$$\max_{X^*X}\operatorname{Re}\operatorname{tr}(X^*B^*A) = \sum_{i=1}^{n}\sigma_i(B^*A) \tag{1}$$

Let $A = U_A\Sigma_A V_A^*$ be the singular value decomposition of A where $\Sigma \in \mathbb{R}^{m\times n}$ with diagonal entries $\sigma_1(A) \ge \cdots \ge \sigma_n(A) \ge 0$, then

$$\begin{aligned}
\operatorname{tr}(A^*A) &= \operatorname{tr}(V_A\Sigma^*U_A^*U_A\Sigma V_A^*) \\
&= \operatorname{tr}(V_A\Sigma\Sigma V_A^*) \\
&= \operatorname{tr}(\Sigma^2V_A^*V_A) \\
&= \operatorname{tr}(\Sigma^2) \\
&= \sum_{i=1}^{n}\sigma_i(A)^2.
\end{aligned}$$

Solution (cont.)

Analogously, $\operatorname{tr}(B^*B) = \sum_{i=1}^n \sigma_i(B)^2$. So

$$\begin{aligned}\max_{X^*X=I} \|A - BX\|_F^2 &= \operatorname{tr}(A^*A) + \operatorname{tr}(B^*B) - 2\max_{X^*X} \operatorname{Re}\operatorname{tr}(X^*B^*A) \\ &= \sum_{i=1}^n \sigma_i(A)^2 + \sum_{i=1}^n \sigma_i(B)^2 - 2\sum_{i=1}^n \sigma_i(B^*A) \\ &= \sum_{i=1}^n [\sigma_i(A)^2 + \sigma_i(B)^2 - 2\sigma_i(B^*A)].\end{aligned}$$

□

(d) Suppose $A \in \mathbb{R}^{m\times n}$ has full column rank. Show that the following method produces a symmetric matrix $X \in \mathbb{R}^{n\times n}$ that solves

$$\min_{X^\top = X} \|AX - B\|_F. \tag{1}$$

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

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

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

Proof. Since A has full column rank, $n = \operatorname{rank}(A) \le m$. Let $A = U_c \Sigma V_c^\top$ be the condensed singular value decomposition where $U_c \in \mathbb{R}^{m\times n}$ and $V_c \in \mathbb{R}^{n\times n}$ with orthogonal columns, and $\Sigma \in \mathbb{R}^{n\times n}$ is the diagonal martix with entries $\sigma_1 \ge \cdots \ge \sigma_n > 0$.
Let $\mathbf{u}_i \in \mathbb{R}^m$ be the ith column of U_c. Expand the column space $\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ with vectors $\mathbf{u}_{n+1}, \ldots . \mathbf{u}_m$, such that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are orthonormal and $\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_m\} = \mathbb{R}^m$. Let $U = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_m\end{bmatrix}$, $V = V_c$, then

$$\begin{aligned}A &= U_c \Sigma V_c^\top \\ &= U \begin{bmatrix}\Sigma \\ 0\end{bmatrix} V^\top\end{aligned}$$

□

(ii) Show that

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

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

Proof. Let $A = U \begin{bmatrix}\Sigma \\ 0\end{bmatrix} V^\top$ be the full singular value decomposition of A with full column rank. So

Solution (cont.)

$U^\top U = I_m$ and $V^\top V = I_n$.

$$\begin{aligned}
\|AX - B\|_F^2 &= \left\| U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top X - B \right\|_F^2 \\
&= \operatorname{tr}\left(U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top X - B \right) \\
&= \operatorname{tr}\left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top X - U^\top B \right) \\
&= \operatorname{tr}\left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top X V - U^\top B V \right) \\
&\overset{Y = V^\top X V}{=\!=\!=} \left\| \begin{bmatrix} \Sigma Y \\ 0 \end{bmatrix} - U^\top B V \right\|_F^2 .
\end{aligned}$$

If we let $U^\top B V = C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ where $C_1 \in \mathbb{R}^{n \times n}$ and $C_2 \in \mathbb{R}^{(m-n) \times n}$. Then

$$\|AX - B\|_F^2 = \left\| \begin{bmatrix} \Sigma Y - C_1 \\ -C_2 \end{bmatrix} \right\|_F^2 = \|\Sigma Y - C_1\|_F^2 + \|C_2\|_F^2.$$

□

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

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

and deduce that the minimum value of (1) is attained when

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

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

Proof. If X is symmetric, then

$$\begin{aligned}
Y^\top &= (V^\top X V)^\top \\
&= V^\top X^\top V \\
&= V^\top X V \\
&= Y,
\end{aligned}$$

Solution (cont.)

i.e., Y is symmetric. So $y_{ij} = y_{ji}$ and

$$\begin{aligned}
\|\Sigma Y - C_1\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\sigma_i y_{ij} - c_{ij})^2 \\
&= \sum_{i=1}^{n}(\sigma_i y_{ii} - c_{ii})^2 + \sum_{j>i}[(\sigma_i y_{ij} - c_{ij})^2 + (\sigma_j y_{ij} - c_{ji})^2] \\
&\geq 0 + \sum_{j>i}[(\sigma_i^2 + \sigma_j^2)y_{ij}^2 - 2(\sigma_i c_{ij} + \sigma_j c_{ji}) + (c_{ij}^2 + c_{ji}^2)] \\
&= \sum_{j>i}\left[(\sigma_i^2 + \sigma_j^2)\left(y_{ij} - \frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}\right)^2 + (c_{ij}^2 + c_{ji}^2) - \left(\frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}\right)^2\right] \\
&\geq \sum_{j>i}\left[(c_{ij}^2 + c_{ji}^2) - \left(\frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}\right)^2\right],
\end{aligned}$$

the equality holds when $y_{ii} = \frac{c_{ii}}{\sigma_i}$ and $y_{ij} = \frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}$ $(j > i)$, i.e., $y_{ij} = \frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}$ for all $i, j = 1, \ldots, n$.

Therefore,

$$\min_{X^\top = X} \|AX - B\|_F^2 = \min_{X^\top = X} \|\Sigma Y - C_1\|_F^2 = \sum_{j>i}\left[(c_{ij}^2 + c_{ji}^2) - \left(\frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}\right)^2\right],$$

and the minimum value is attained when $y_{ij} = \frac{\sigma_i c_{ij} + \sigma_j c_{ji}}{\sigma_i^2 + \sigma_j^2}$ for all $i, j = 1, \ldots, n$. □

(e) Let $A \in \mathbb{R}^{m\times n}$. Show that

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

has a unique solution when A has full column rank. In general, what is the minimum length solution (i.e., where $\|X\|_F$ is minimum) to this problem?

Proof. Let $\mathbf{x}_i$ and $\mathbf{e}_i$ be the ith column of X and I_m, respectively. Then,

$$\|AX - I_m\|_F^2 = \sum_{i=1}^{m} \|A\mathbf{x}_i - \mathbf{e}_i\|_2.$$

Since A has full column rank, $\min_{\mathbf{x}_i \in \mathbb{R}^n} \|A\mathbf{x}_i - \mathbf{e}_i\|_2 = (A^\top A)^{-1}A^\top \mathbf{e}_i$ for all $i = 1, \ldots, m$. So the solution of $\min_{X \in \mathbb{R}^{n\times m}} \|AX - I_m\|_F^2$ is given by $(A^\top A)^{-1}A^\top$ uniquely. □

3

In the following, $\kappa(A) \coloneqq \|A\|\|A^\dagger\|$ 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 $A \in \mathbb{C}^{m\times n}$,

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

Proof. Since $A^\dagger$ satisfies $AA^\dagger A = A$, we have

$$\|A\| = \|AA^\dagger A\| \leq \|A\|\|A^\dagger\|\|A\|.$$

Since A is nonzero, $\|A\| > 0$. Deviding $\|A\|$ at the above inequality, we have

$$\|A^\dagger\|\|A\| \geq 1.$$

□

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

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

but that in general

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

Proof. Suppose that $\operatorname{rank}(A) = r$. Let $A = U\Sigma V^*$ be the full singular value decomposition, where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r \geq 0 = \cdots = 0$. Then $A^\dagger = V\Sigma^\dagger U^*$, where $\Sigma^\dagger \in \mathbb{R}^{n\times m}$ is a diagonal matrix with diagonal entries $\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_r}, 0, \cdots, 0$. Then,

$$A^*A = V\Sigma^* U^* U\Sigma V^* = V\Lambda V^*$$

where $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal entries $\sigma_1^2 \geq \cdots \geq \sigma_r^2 > 0 = \cdots = 0$. Then $(A^*A)^\dagger = V\Lambda^\dagger V^*$ where $\Lambda^\dagger \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal entries $\frac{1}{\sigma_1^2}, \cdots, \frac{1}{\sigma_r^2}, 0, \cdots, 0$. Since $\|\cdot\|_2$ is unitarily invariant, we have that

$$\begin{aligned}
\kappa_2(A^*A) &= \|A^*A\|_2\|(A^*A)^\dagger\|_2 \\
&= \|V\Lambda V^*\|_2\|V\Lambda^\dagger V^*\|_2 \\
&= \|\Lambda\|_2\|\Lambda^\dagger\|_2 \\
&= \frac{\sigma_1^2}{\sigma_r^2} \\
\kappa_2(A) &= \|A\|_2\|A^\dagger\|_2 \\
&= \|U\Sigma V^*\|_2\|V\Sigma^\dagger U^*\|_2 \\
&= \|\Sigma\|_2\|\Sigma^\dagger\|_2 \\
&= \frac{\sigma_1}{\sigma_r}
\end{aligned}$$

i.e., $\kappa_2(A^*A) = \kappa_2(A)^2$. But in general, it is not true for any norm. For example, Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, then $A^\dagger = A^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$, $A^*A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ and $(A^*A)^\dagger = (A^*A)^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$. So $\kappa_\infty(A^*A) =$

Solution (cont.)

$\|A^*A\|_\infty \|(A^*A)^\dagger\|_\infty = 9 \neq \kappa_\infty(A)^2 = \|A\|_\infty^2 \|A^\dagger\|_\infty^2 = 16.$ □

(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?

Proof. For nonsingular $A, B \in \mathbb{C}^{n\times n}$, let $A = U_A \Sigma_A V_A^*$ be the full singular decomposition, where $U_A, V_A \in \mathbb{C}^{n\times n}$ are unitary matrices and $\Sigma_A \in \mathbb{R}^{n\times n}$ is diagonal matrix with diagonal entries $\sigma_{A,1} \geq \cdots \geq \sigma_{A,n} > 0$. So $A^\dagger = V_A \Sigma_A^{-1} U_A^* = A^{-1}$. Analogously, $B^\dagger = B^{-1}$. Since A and B are nonsingular, AB is also nonsingular($\det(AB) = \det(A)\det(B) > 0$), so that $(AB)^\dagger = (AB)^{-1}$.

$$\begin{aligned}
\kappa(AB) &= \|AB\|\|(AB)^\dagger\| \\
&= \|AB\|\|B^{-1}A^{-1}\| \\
&\leq \|A\|\|A^{-1}\|\|B\|\|B^{-1}\| \\
&= \|A\|\|A^\dagger\|\|B\|\|B^\dagger\| \\
&= \kappa(A)\kappa(B)
\end{aligned}$$

If A or B is singular, this is not true in general. For example, let $A = \begin{bmatrix} 99 & -99.001 & -25 & 76 \\ 99 & -99 & -25 & 76 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 10 & -62 & 0 & 0 \\ -44 & -83 & 0 & 0 \\ 26 & 9 & 0 & 0 \\ -3 & 88 & 0 & 0 \end{bmatrix}$, then $AB = \begin{bmatrix} 4292.044 & 8210.083 & 0 & 0 \\ 4292 & 8210 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $\kappa(AB) \approx 3.4303 \times 10^7 > \kappa(A)\kappa(B) \approx 1.1759 \times 10^6$. □

(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?

Proof. (1) Let $Q = U\Sigma V^*$ be the full singular decomposition, where $U \in \mathbb{C}^{m\times m}, V \in \mathbb{C}^{n\times n}$ are unitary matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal matrix with diagonal entries $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$. Since Q has orthonormal columns, we have $Q^*Q = I_n$, i.e.

$$\begin{aligned}
I_n &= (U\Sigma V^*)^* U\Sigma V^* \\
&= V\Sigma^*\Sigma V^*
\end{aligned}$$

i.e.,

$$\Sigma^*\Sigma = I_n,$$

which implies that $\sigma_1 = \cdots = \sigma_n = 1$. Then $\|Q\|_2 = 1$. Since $Q^\dagger = V\Sigma^\dagger U^* = V \begin{bmatrix} I_n & 0_{n\times(m-n)} \end{bmatrix} U^*$,

Solution (cont.)

we also have $\|Q^\dagger\|_2 = 1$. Therefore

$$\kappa_2(Q) = \|Q\|_2\|Q^\dagger\|_2 = 1$$

(2) If Q has orthonormal rows, then $QQ^* = I_m$. Analogously, $\Sigma\Sigma^* = I_m$, $\sigma_1 = \cdots = \sigma_m = 1$ and $\|Q\|_2 = \|Q^\dagger\|_2 = 1$. So $\kappa_2(Q) = 1$.

(3) It is not true for κ_1 and κ_∞. To see this, let $Q = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$, then Q has orthonormal columns and rows, $Q^\dagger = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ and $\kappa_1(Q) = \kappa_\infty(Q) = 2$. □

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

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

Proof. Since R is nonsingular, the diagonal entries $r_{11}, \ldots, r_{nn}$ are eigenvalues and nonzero, and $R^\dagger = R^{-1}$. First we have to show that R^{-1} is also upper-triangular. Write R as $R = D(I + N)$ where $D = \mathrm{diag}(r_{11}, \ldots, r_{nn})$, and N satisfies $N^m (m = 1, \ldots, n-1)$ is upper-triangular with zero diagonal entries and $N^n = 0$, then $R^{-1} = (I+N)^{-1}D^{-1} = [I - N + N^2 - \cdots + (-1)^{n-1}N^{n-1}]D^{-1}$ is also upper-triangular and its diagonal entries are $\frac{1}{r_{ii}}$ $(i = 1, \ldots, n)$.
Let r_{ij} and r'_{ij} $(i, j = 1, \ldots, n)$ denote the elements of R and R^{-1} respectively, then

$$\begin{aligned}
\kappa_\infty(R) &= \|R\|_\infty\|R^{-1}\|_\infty \\
&= \left(\max_i \sum_{j=1}^n |r_{ij}|\right)\left(\max_i \sum_{j=1}^n |r'_{ij}|\right) \\
&\geq \max_i |r_{ii}| \cdot \max_j |r'_{jj}| \\
&= \max_i |r_{ii}| \cdot \max_j \frac{1}{|r_{jj}|} \\
&= \frac{\max_i |r_{ii}|}{\max_j |r_{jj}|}.
\end{aligned}$$

□

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

Proof. Since $AX - I = A(XA - I)A^{-1}$, we have

$$\begin{aligned}\frac{\|AX-I\|}{\|XA-I\|} &= \frac{\|A(XA-I)A^{-1}\|}{\|XA-I\|}\\ &\leq \frac{\|A\|\|XA-I\|\|A^{-1}\|}{\|XA-I\|}\\ &= \|A\|\|A^{-1}\|\\ &= \kappa(A)\end{aligned}$$

Since $XA - I = A^{-1}(AX - I)A$, we have

$$\begin{aligned}\frac{\|XA-I\|}{\|AX-I\|} &= \frac{\|A^{-1}(AX-I)A\|}{\|AX-I\|}\\ &\leq \frac{\|A^{-1}\|\|AX-I\|\|A\|}{\|AX-I\|}\\ &= \|A\|\|A^{-1}\|\\ &= \kappa(A)\end{aligned}$$

So

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

□

4

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

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

Proof.

$$\begin{aligned}
\left\| A\left(I - \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right)\right\|_F^2 &= \operatorname{tr}\left(\left(I - \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right)^* A^*A\left(I - \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right)\right) \\
&= \operatorname{tr}\left(A^*A - \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}A^*A - A^*A\frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}} + \frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}A^*A\frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right) \\
&= \operatorname{tr}(A^*A) - \operatorname{tr}\left(\frac{\mathbf{x}(A\mathbf{x})^*A}{\mathbf{x}^*\mathbf{x}}\right) - \operatorname{tr}\left(\frac{A^*(A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right) + \operatorname{tr}\left(\frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}A^*A\frac{\mathbf{x}\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}\right) \\
&= \|A\|_F^2 + \frac{1}{\mathbf{x}^*\mathbf{x}}\left[-\operatorname{tr}((A\mathbf{x})^*(A\mathbf{x})) - \operatorname{tr}((A\mathbf{x})^*(A\mathbf{x})) + \operatorname{tr}((A\mathbf{x})^*(A\mathbf{x}))\right] \\
&= \|A\|_F^2 - \frac{\|A\mathbf{x}\|_2^2}{\mathbf{x}^*\mathbf{x}}
\end{aligned}$$

where the second to the last quality comes from the invariant property under cyclic permutations of trace. □

(b) Show that the matrix

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

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

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

Proof. Substitue $E_0 = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$ into $(A + E)\mathbf{x}$, we have

$$A\mathbf{x} + \frac{(\mathbf{b} - A\mathbf{x})\mathbf{x}^*\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \mathbf{b}$$

i.e., $E_0 = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$ satisfies $(A + E)\mathbf{x} = \mathbf{b}$.
Since $(A + E)\mathbf{x} = \mathbf{b}$ equals to $E\mathbf{x} = \mathbf{b} - A\mathbf{x}$ for all E satisfying the equation, we have

$$\|E\|_2\|\mathbf{x}\|_2 \geq \|E\mathbf{x}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2,$$

i.e.,

$$\|E\|_2 \geq \frac{\|\mathbf{b} - A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{1}$$

Solution (cont.)

For $E_0 = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$, we have

$$\begin{aligned}\|E_0\|_2 &= \frac{1}{|\mathbf{x}^*\mathbf{x}|}\|(\mathbf{b}-A\mathbf{x})\mathbf{x}^*\|_2 \\ &\leq \frac{1}{\|\mathbf{x}\|_2^2}\|\mathbf{b}-A\mathbf{x}\|_2\|\mathbf{x}^*\|_2 \\ &= \frac{\|\mathbf{b}-A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \end{aligned} \tag{2}$$

From (1) and (2), we have $\|E_0\|_2 = \frac{\|\mathbf{b}-A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ and $E_0 = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$ 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

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

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

From the proof of (b), we have that

$$\min_{(A+E)\mathbf{x}=\mathbf{b}} \|E\|_2 = \|E_0\|_2 = \frac{\|\mathbf{b}-A\mathbf{x}\|_2}{\|\mathbf{x}\|_2},$$

and the minimum is taken at $E = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$.
Similar results can be attained for $\|\cdot\|_F$. By submultiplicative property of F-norm,

$$\|E\|_F\|\mathbf{x}\|_F \geq \|E\mathbf{x}\|_F = \|\mathbf{b}-A\mathbf{x}\|_F,$$

i.e.,

$$\|E\|_F \geq \frac{\|\mathbf{b}-A\mathbf{x}\|_F}{\|\mathbf{x}\|_F}.$$

Also,

$$\begin{aligned}\|E_0\|_F &= \frac{1}{|\mathbf{x}^*\mathbf{x}|}\|(\mathbf{b}-A\mathbf{x})\mathbf{x}^*\|_F \\ &\leq \frac{1}{\|\mathbf{x}\|_F^2}\|\mathbf{b}-A\mathbf{x}\|_F\|\mathbf{x}^*\|_F \\ &= \frac{\|\mathbf{b}-A\mathbf{x}\|_F}{\|\mathbf{x}\|_F}. \end{aligned}$$

Therefore,

$$\min_{(A+E)\mathbf{x}=\mathbf{b}} \|E\|_F = \|E_0\|_F = \frac{\|\mathbf{b}-A\mathbf{x}\|_F}{\|\mathbf{x}\|_F},$$

and the minimum is taken at $E = \frac{(\mathbf{b}-A\mathbf{x})\mathbf{x}^*}{\mathbf{x}^*\mathbf{x}}$.

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

$$\min_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2 \qquad \text{and} \qquad \max_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2$$

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

(1) If $\mathbf{a} = \mathbf{0}$, then

$$\min_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2 = \max_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2 = \|\mathbf{b}\|_2$$

where the minimum and the maximum are achieved for all $E \in \mathbb{C}^{m \times n}$ such that $\|E\|_F \le \Delta$.
If $\mathbf{a} \neq \mathbf{0}$, from (c), we know that

$$\arg\min_E \|E\mathbf{a} - \mathbf{b}\|_2 = E_0 \triangleq \frac{\mathbf{b}\mathbf{a}^*}{\mathbf{a}^*\mathbf{a}}$$

and

$$\begin{aligned} \|E_0\|_F &= \left\| \frac{\mathbf{b}\mathbf{a}^*}{\mathbf{a}^*\mathbf{a}} \right\|_F \\ &= \frac{\sqrt{\operatorname{tr}(\mathbf{a}\mathbf{b}^*\mathbf{b}\mathbf{a}^*)}}{\|\mathbf{a}\|_2^2} \\ &= \frac{\sqrt{\operatorname{tr}(\mathbf{b}^*\mathbf{b}\mathbf{a}^*\mathbf{a})}}{\|\mathbf{a}\|_2^2} \\ &= \frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2}. \end{aligned}$$

(2) If $\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} \le \Delta$, then E_0 is the solution to $\min_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2$.
(i) Consider the first optimization problem. If $\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} > \Delta$, let $E_0 = \frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2}\mathbf{u}\mathbf{v}^*$ be the singular value decomposition of rank-one matrix E_0 where $\mathbf{u} \in \mathbb{C}^m$ and $\mathbf{v} \in \mathbb{C}^n$ are unit vectors. Let $E_1 = \Delta\mathbf{u}\mathbf{v}^*$, then $\|E_1\|_F = \Delta$. Notice that when $\|E\|_F \le \Delta$,

$$\begin{aligned} \|E\mathbf{a} - \mathbf{b}\|_2 &\ge \|\mathbf{b}\|_2 - \|E\mathbf{a}\|_2 \\ &\ge \|\mathbf{b}\|_2 - \|E\|_F\|\mathbf{a}\|_2 \\ &\ge \|\mathbf{b}\|_2 - \Delta\|\mathbf{a}\|_2 \\ &> 0 \end{aligned}$$

and

$$\begin{aligned} \|E_1\mathbf{a} - \mathbf{b}\|_2 &= \|(E_1 - E_0)\mathbf{a}\|_2 \\ &= \left\| \left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} - \Delta \right) \mathbf{u}\mathbf{v}^*\mathbf{a} \right\|_2 \\ &= \left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} - \Delta \right) \|\mathbf{u}\mathbf{v}^*\mathbf{a}\|_2 \\ &= \left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} - \Delta \right) \|\mathbf{a}\|_2 \\ &= \|\mathbf{b}\|_2 - \Delta\|\mathbf{a}\|_2, \end{aligned}$$

so $\min_{\|E\|_F \le \Delta} \|E\mathbf{a} - \mathbf{b}\|_2 = \|\mathbf{b}\|_2 - \Delta\|\mathbf{a}\|_2$ which can be attained at E_1.

Solution (cont.)

(ii) Consider the second optimization problem. Let $E_2 = -\Delta \mathbf{u}\mathbf{v}^*$, then $\|E_2\|_F = \Delta$. Notice that when $\|E\|_F \leq \Delta$,

$$\begin{aligned}\|E\mathbf{a} - \mathbf{b}\|_2 &\leq \|\mathbf{b}\|_2 + \|E\mathbf{a}\|_2 \\ &\leq \|\mathbf{b}\|_2 + \|E\|_F\|\mathbf{a}\|_2 \\ &\leq \|\mathbf{b}\|_2 + \Delta\|\mathbf{a}\|_2\end{aligned}$$

and

$$\begin{aligned}\|E_2\mathbf{a} - \mathbf{b}\|_2 &= \|(E_2 - E_0)\mathbf{a}\|_2 \\ &= \left\| \left(-\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} - \Delta \right) \mathbf{u}\mathbf{v}^*\mathbf{a} \right\|_2 \\ &= \left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} + \Delta \right) \|\mathbf{u}\mathbf{v}^*\mathbf{a}\|_2 \\ &= \left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{a}\|_2} + \Delta \right) \|\mathbf{a}\|_2 \\ &= \|\mathbf{b}\|_2 + \Delta\|\mathbf{a}\|_2,\end{aligned}$$

so $\min\limits_{\|E\|_F \leq \Delta} \|E\mathbf{a} - \mathbf{b}\|_2 = \|\mathbf{b}\|_2 + \Delta\|\mathbf{a}\|_2$ which can be attained at E_2.

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 $1, 2, \ldots, n^2$ arranged in a way that A has equal row, column, and diagonal sums [`magic(n)`].

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

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) $\mathbf{x} = [1, \ldots, 1]^\top$ [`ones(n,1)`];

(ii) $\mathbf{b} = A\mathbf{x}$ [`b = A*x`].

(c) For each A generated as above, perform the following for $n = 5, 10, 15, \ldots, 500$.

(i) Solve $A\mathbf{x} = \mathbf{b}$ using your program to get $\widehat{\mathbf{x}}$ [`xhat = A\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 $\Delta\mathbf{b} = A\widehat{\mathbf{x}} - \mathbf{b}$ and record the values of $\|\mathbf{x} - \widehat{\mathbf{x}}\|/\|\mathbf{x}\|$, $\kappa(A) = \|A\|\|A^{-1}\|$ 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.

The results for $n = 5$ is presented in Table 1. And the relevant trend for general n is presented in Fig 1-3 for $\|\mathbf{x} - \widehat{\mathbf{x}}\|/\|\mathbf{x}\|$, $\kappa(A) = \|A\|\|A^{-1}\|$ and $\kappa(A)\|\Delta\mathbf{b}\|/\|\mathbf{b}\|$, respectively.

Solution (cont.)

	Matrix	1 norm	2 norm	∞ norm
$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|}$	Random Normal	0.2931*1.0e-14	0.3225*1.0e-14	0.4885*1.0e-14
	Hilbert	0.1894*1.0e-11	0.2368*1.0e-11	0.3992*1.0e-11
	Pascal	0	0	0
	Magic	0.1554*1.0e-15	0.2483*1.0e-15	0.4441*1.0e-15
$\kappa(A)$	Random Normal	228.5836	124.9406	179.1639
	Hilbert	9.4366*1.0e+05	4.7661*1.0e+05	9.4366*1.0e+05
	Pascal	1.5624 *1.0e+04	0.8518*1.0e+04	1.5624*1.0e+04
	Magic	6.8500	5.4618	6.8500
$\kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$	Random Normal	0.1706*1.0e-13	0.1464*1.0e-13	0.2882*1.0e-13
	Hilbert	0.3245*1.0e-10	0.3372*1.0e-10	0.9177*1.0e-10
	Pascal	0	0	0
	Magic	0.0300*1.0e-14	0.0534*1.0e-08	0.1498*1.0e-10

Table 1. Results for $n = 5$.

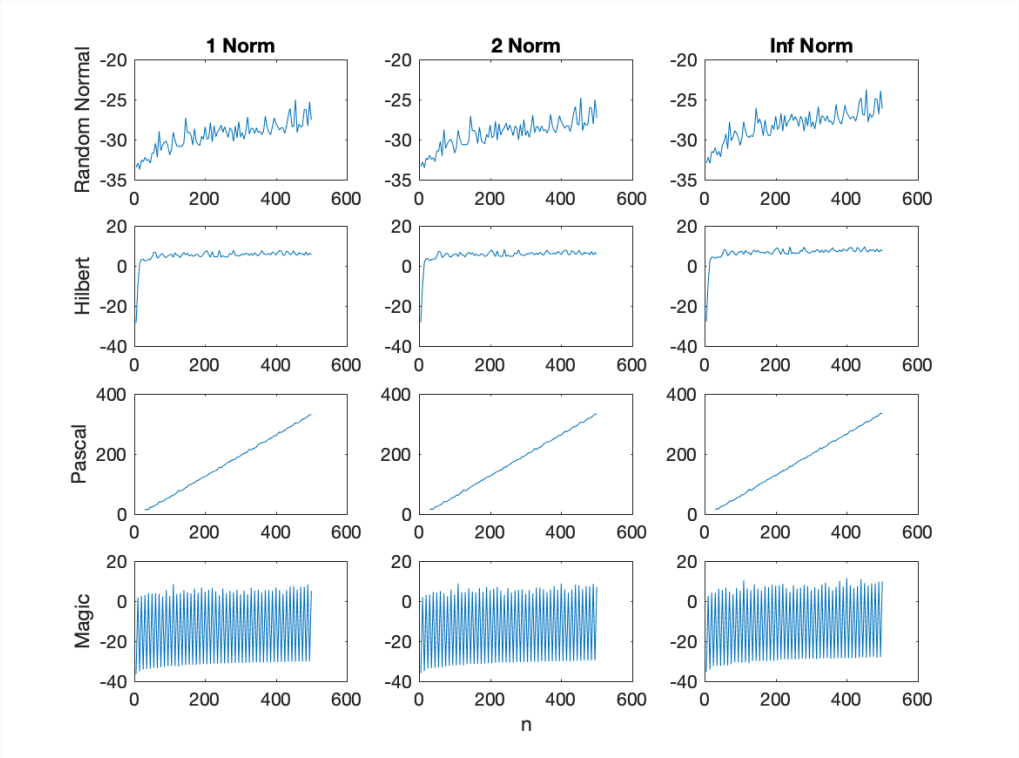


Fig 1. Result for $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|}$ (base-e log scale on y-axis).

Solution (cont.)

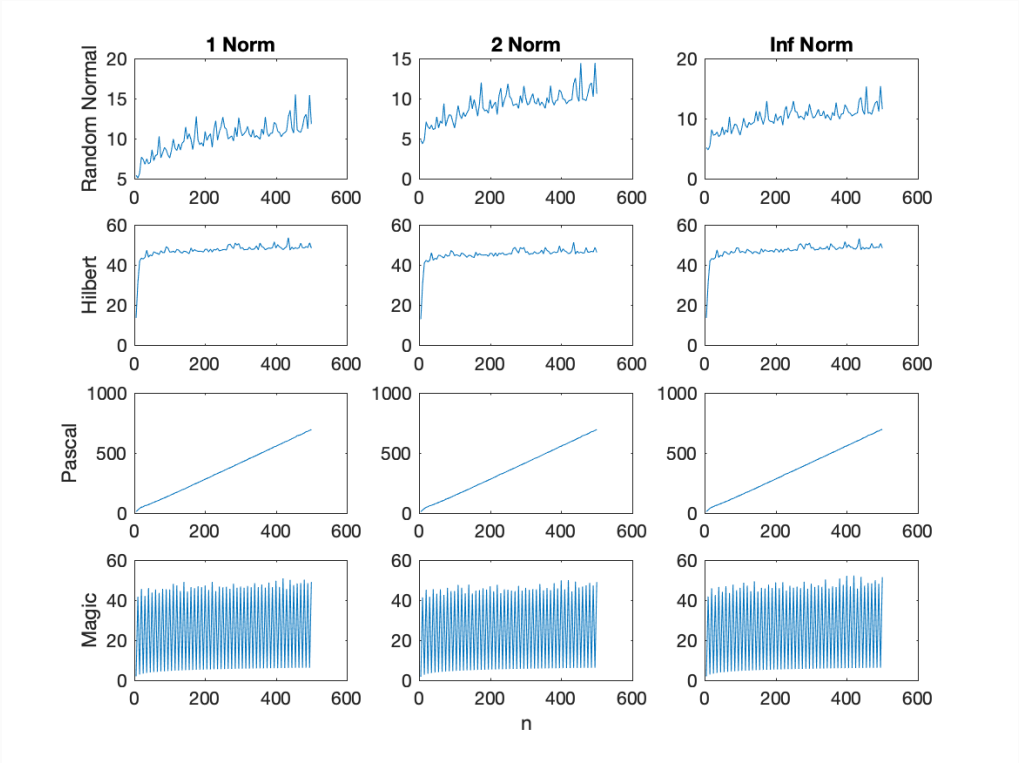


Fig 2. Result for $\kappa(A)$ (base-e log scale on y-axis).

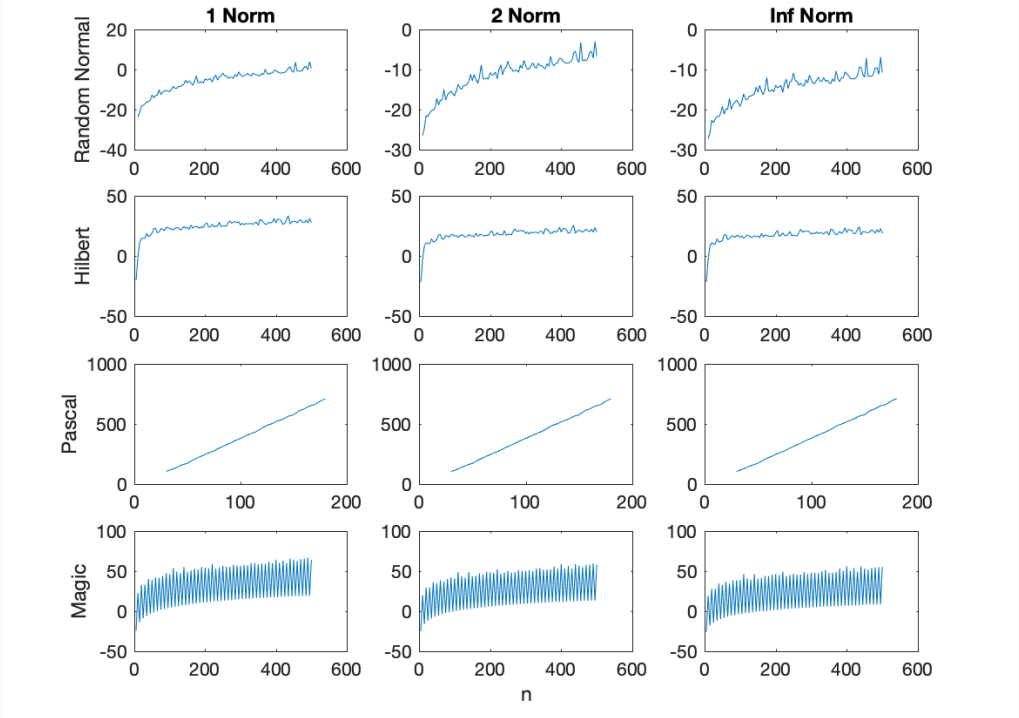


Fig 3. Result for $\kappa(A)\frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}$ (base-e log scale on y-axis).

Solution (cont.)

MATLAB Codes:

```
delta_x = zeros(100,3,4);
kappa = zeros(100,3,4);
kappa_delta_b = zeros(100,3,4);
norm_list = [1,2,Inf];
A_list = {'Random Normal', 'Hilbert', 'Pascal', 'Magic'};
for n=5:5:500
   x = ones(n,1);
   for i=1:4
        A = generateMatrix(i, n);
        b = A*x;

        x_hat = A\b;
        delta_b = A*x_hat-b;
        for j=1:3
            nor = norm_list(j);
            delta_x(n/5,j,i) = norm(x-x_hat,nor)/norm(x,nor);
            kappa(n/5,j,i) = cond(A,nor);
            kappa_delta_b(n/5,j,i) = kappa(n/5,j)*norm(delta_b,nor)/norm(b,nor);
        end
   end
end

delta_x(1,:, :)
figure();
for i=1:4
    for j=1:3
        subplot(4,3,3*(i-1)+j);
        plot(5:5:500, delta_x(:, j,i));
        if i==1
            title([num2str(norm_list(j)), ' Norm']);
        end
        if j==1
            ylabel(A_list(i));
        end
        if i==4 && j==2
            xlabel('n')
        end
    end
end
saveas(gcf,'result1.png')

kappa(1,:, :)
```

Solution (cont.)

```
figure();
for i=1:4
    for j=1:3
        subplot(4,3,3*(i-1)+j);
        plot(5:5:500, kappa(:, j,i));
        if i==1
            title([num2str(norm_list(j)), ' Norm']);
        end
        if j==1
            ylabel(A_list(i));
        end
        if i==4 && j==2
            xlabel('n')
        end
    end
end
saveas(gcf,'result2.png')

kappa_delta_b(1,:,:)
figure();
for i=1:4
    for j=1:3
        subplot(4,3,3*(i-1)+j);
        plot(5:5:500, kappa_delta_b(:, j,i));
        if i==1
            title([num2str(norm_list(j)), ' Norm']);
        end
        if j==1
            ylabel(A_list(i));
        end
        if i==4 && j==2
            xlabel('n')
        end
    end
end
saveas(gcf,'result3.png')

function A = generateMatrix(i, n)
    switch i
        case 1
            A = rand(n);
        case 2
            A = hilb(n);
        case 3
```

Solution (cont.)

```
            A = pascal(n);
        case 4
            A = magic(n);
    end
end
```

(d) Discuss and explain the effects of different choices of A, $\mathbf{b}$, $\|\cdot\|$, and n have on the accuracy of the computed solution $\widehat{\mathbf{x}}$.

Among different choices of A, random normal matrix has least error on $\hat{\mathbf{x}}$. As n increases, the error increases slowly.
Hilbert matrix has small error when n is very small, and has large error ($\approx 10^{10}$) when n increases.
Pascal matrix has very large error, and as n increases, the error increases exponentially.
Magic matrix has fluctuant error. As n increases, the error seems to increase on average.

(e) Instead of solving the linear system directly, compute A^{-1} and then $\widehat{\mathbf{x}} := A^{-1}\mathbf{b}$ [`xhat = inv(A)*b`]. Comment on the accuracy of this approach. Provide numerical evidence to support your conclusion.

This approach causes more errors for Hilbert matrices and espiecially Pascal matrices. While for Random Normal matrices and Magic matrices, the accuracy of these two methods is close to each other.

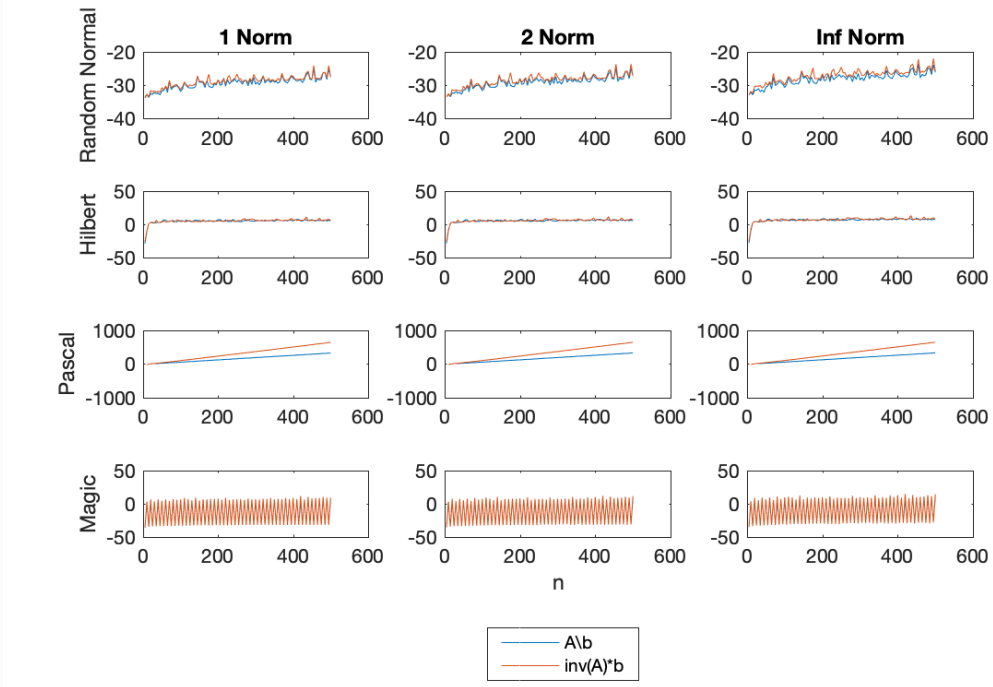


Fig 4. Result for $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|}$ with different solving methods (base-e log scale on y-axis).

Solution (cont.)

Matrix	1 norm	2 norm	∞ norm
Random Normal	0.0126*1.0e-11	0.0210*1.0e-11	0.1122*1.0e-11
Hilbert	0.0511*1.0e+04	0.1707*1.0e+04	1.3219*1.0e+04
Pascal	0.0171*1.0e+146	0.2356*1.0e+146	3.7433*1.0e+146
Magic	-0.0184*1.0e-13	-0.00127*1.0e-13	0.11149*1.0e-13

Table 2. Results for median difference of $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|}$ between $inv(A)*b$, the method using inverse, and $A\backslash b$, the mothed solves the linear system directly.

```
rand('seed',0)
delta_x = zeros(100,2,4);
delta_x_inv = zeros(100,2,4);
for n=5:5:500
   x = ones(n,1);
   for i=1:4
        A = generateMatrix(i, n);
        b = A*x;

        x_hat = A\b;
        x_hat_inv = inv(A)*b;
        for j=1:3
            nor = norm_list(j);
            delta_x(n/5,j,i) = norm(x-x_hat,nor)/norm(x,nor);
            delta_x_inv(n/5,j,i) = norm(x-x_hat_inv,nor)/norm(x,nor);
        end
   end
end

figure();
for i=1:4
    for j=1:3
        subplot(5,3,3*(i-1)+j);
        h1 = plot(5:5:500, log(delta_x(:, j,i)));
        hold on
        h2 = plot(5:5:500, log(delta_x_inv(:, j,i)));
        if i==1
            title([num2str(norm_list(j)), ' Norm']);
        end
        if j==1
            ylabel(A_list(i));
        end
```

Solution (cont.)

```
        if i==4 && j==2
            xlabel('n')
        end
    end
end

% Construct a Legend with the data from the sub-plots
hL = legend([h1, h2],{'A\b','inv(A)*b'});
% Programatically move the Legend
newPosition = [0.45 0.15 0.14 0.06];
newUnits = 'normalized';
set(hL,'Position', newPosition,'Units', newUnits);
saveas(gcf,'result4.png')
median(delta_x_inv - delta_x)
```

(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}.

We can solve the linear system $A\mathbf{b} = \mathbf{e}_1$ where $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{e}_1 \in \mathbb{R}^n$ is the unit vector with first entry being 1. Since A is invertible, this equation has unique solution $\mathbf{b}_1 = A^{-1}\mathbf{e}_1$, i.e., the first column of A^{-1}.

```
function b11 = firstEntry(A)
    n = length(A);
    e = zeros(n);
    e(1) = 1;
    b = A\e;
    b11 = b(1);
end
```

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

$$\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\kappa(A)}, \tag{2}$$

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

Proof. If $A + \Delta A$ is singular, then exists nonzero $\mathbf{v}$ such that $(A + \Delta A)\mathbf{v} = \mathbf{0}$. Since A is nonsingular, we have $A^{-1}\Delta A\mathbf{v} = -\mathbf{v}$. So

$$\begin{aligned}\|\mathbf{v}\| &= \|A^{-1}\Delta A\mathbf{v}\| \\ &\leq \|A^{-1}\|\|\Delta A\|\|\mathbf{v}\|\end{aligned}$$

i.e.,

$$\|\Delta A\| \geq \frac{1}{\|A^{-1}\|}.$$

Since $\kappa(A) = \|A\|\|A^\dagger\| = \|A\|\|A^{-1}\|$, so

$$\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\|A\|\|A^{-1}\|},$$

i.e.,

$$\|\Delta A\| < \frac{1}{\|A^{-1}\|}.$$

Contradiction. Therefore, $A + \Delta A$ must be nonsingular. □

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

$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\frac{\|\Delta A\|}{\|A\|}. \tag{3}$$

Proof. Since $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$ and $A\mathbf{x} = \mathbf{b}$, we have $A\Delta\mathbf{x} + \Delta A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{0}$. So

$$\Delta x = -A^{-1}\Delta A(\mathbf{x} + \Delta\mathbf{x}).$$

$$\begin{aligned}\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} &= \frac{\|A^{-1}\Delta A(\mathbf{x} + \Delta\mathbf{x})\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \\ &\leq \frac{\|A^{-1}\|\|\Delta A\|\|\mathbf{x} + \Delta\mathbf{x}\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \\ &= \|A^{-1}\|\|\Delta A\| \\ &= \|A^{-1}\|\|A\|\frac{\|\Delta A\|}{\|A\|} \\ &= \kappa(A)\frac{\|\Delta A\|}{\|A\|}.\end{aligned}$$

□

(c) Suppose $(A+\Delta A)(\mathbf{x}+\Delta\mathbf{x})=\mathbf{b}$ and $\hat{\mathbf{x}}=\mathbf{x}+\Delta\mathbf{x}$ and (2) is satisfied. Show that

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

You may like 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 (2).

Proof. (i) From (b), have

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

(ii) Rearrange the above inequality, we have

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

Since (6.2) is satisfied, we have $1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}>0$. Since A is nonsingular and $\mathbf{b}\neq 0$, $\mathbf{x}=A^{-1}\mathbf{b}\neq 0$ and so dose $\|\mathbf{x}\|$. Therefore,

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

□

(d) *Bonus*: 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

$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|} + \frac{\|\Delta A\|}{\|A\|}\frac{\|\Delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|}\right). \tag{4}$$

You may like use the following outline:

(i) Show that

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

and so

$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right). \tag{5}$$

(ii) Show that

$$\frac{1}{\|\hat{\mathbf{x}}\|} \leq \frac{\|A\|+\|\Delta A\|}{\|\hat{\mathbf{b}}\|}. \tag{6}$$

(iii) Combine (5) and (6) to get (4).

Proof. (i) Since $(A+\Delta A)\hat{\mathbf{x}} = \mathbf{b}+\Delta\mathbf{b}$ and $A\mathbf{x} = \mathbf{b}$, we have $A\Delta\mathbf{x} + \Delta A\hat{\mathbf{x}} = \Delta\mathbf{b}$. So $\Delta\mathbf{x} = A^{-1}(\Delta\mathbf{b} - \Delta A\hat{\mathbf{x}})$. Then

$$\begin{aligned}
\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} &= \frac{\|A^{-1}(\Delta\mathbf{b}-\Delta A\hat{\mathbf{x}})\|}{\|\hat{\mathbf{x}}\|} \\
&\leq \frac{\|A^{-1}\|(\|\Delta A\|\|\hat{\mathbf{x}}\| + \|\Delta\mathbf{b}\|)}{\|\hat{\mathbf{x}}\|} \\
&= \|A^{-1}\|\|A\|\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right) \\
&= \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right)
\end{aligned}$$

(ii) Since $(A+\Delta A)\hat{\mathbf{x}} = \hat{\mathbf{b}}$, we have

$$\begin{aligned}
\|\hat{\mathbf{b}}\| &= \|(A+\Delta A)\hat{\mathbf{x}}\| \\
&\leq (\|A\| + \|\Delta A\|)\|\hat{\mathbf{x}}\|
\end{aligned}$$

Since $\hat{\mathbf{x}} \neq \mathbf{0}$, $\|\hat{\mathbf{x}}\| \neq \mathbf{0}$ and

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

(iii) Substitue (6.6) to the right side of (6.5), we get

$$\begin{aligned}
\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} &\leq \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|A\|}\cdot\frac{\|A\|+\|\Delta A\|}{\|\hat{\mathbf{b}}\|}\right) \\
&= \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right).
\end{aligned}$$

□

(e) *Bonus*: 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 (2) is satisfied. Use the same ideas in (b) to deduce that

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

Proof. From (d) (i), we have

$$\begin{aligned}
\|\Delta\mathbf{x}\| &\leq \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right)\|\hat{\mathbf{x}}\| \\
&= \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|}\|\hat{\mathbf{x}}\|+\frac{\|\Delta\mathbf{b}\|}{\|A\|}\right) \\
&\leq \kappa(A)\frac{\|\Delta A\|}{\|A\|}(\|\mathbf{x}\|+\|\Delta\mathbf{x}\|)+\kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|A\|}
\end{aligned}$$

Rearrange the above inequality, we have

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

Then

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

where the last inequality holds since $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$. □