

CAAM 453 · NUMERICAL ANALYSIS I

Problem Set 4

Posted Thursday 8 October 2009. Due Monday, 19 October 2009. [Minor typos corrected 16 October 2009]
CAAM 453 students should complete 4 problems (including problem 5).
CAAM 553 students should complete 5 problems (including problem 5).
Students are welcome to attempt more problems if they wish.

1. [25 points]

Determine *by hand calculation*, the singular value decompositions of the matrices

$$(a) \begin{pmatrix} 3 & 0 \\ 0 & -2 \end{pmatrix} \quad (b) \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \quad (c) \begin{pmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad (d) \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad (e) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

and, for each matrix, write down the optimal (2-norm) rank-1 approximation.

[Trefethen and Bau, problem 4.1]

2. [25 points]

(a) Suppose $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$, has full rank. The exact solution to the least squares problem

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$$

is given by $\mathbf{x} = \mathbf{A}^+\mathbf{b}$, where the *pseudoinverse* \mathbf{A}^+ is defined, for full rank \mathbf{A} , by

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

Show that the matrix $\mathbf{\Pi} = \mathbf{A}\mathbf{A}^+$ is an *orthogonal projector* onto $\text{Ran}(\mathbf{A})$.

(b) If $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$, has full-rank, develop an expression for \mathbf{A}^+ in terms of the singular value decomposition of \mathbf{A} . (You may find it most convenient to work with the dyadic form of the SVD, $\mathbf{A} = \sum_{j=1}^n \sigma_j \mathbf{u}_j \mathbf{v}_j^*$.)

(c) From (b), deduce a formula for \mathbf{A}^+ that is suitable when $\text{rank}(\mathbf{A}) = r < n$ for $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$. Your formula should satisfy two properties: $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ should be a vector that minimizes $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$, and $\mathbf{A}\mathbf{A}^+$ should be an orthogonal projector onto $\text{Ran}(\mathbf{A})$. Confirm that both hold.

3. [25 points]

Prove the following three pseudoinverse identities for arbitrary $\mathbf{A} \in \mathbb{C}^{m \times n}$.

$$(a) \mathbf{A}^+ = \lim_{t \rightarrow 0} (\mathbf{A}^*\mathbf{A} + t\mathbf{I})^{-1}\mathbf{A}^*.$$

$$(b) \mathbf{A}^+ = \int_0^\infty e^{-\mathbf{A}^*\mathbf{A}t} \mathbf{A}^* dt.$$

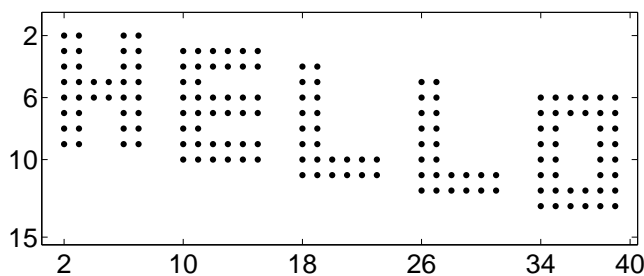
(c) Let Γ be a closed contour in the complex plane that encloses all nonzero eigenvalues of $\mathbf{A}^*\mathbf{A}$ but does not enclose the origin. Then

$$\mathbf{A}^+ = \frac{1}{2\pi i} \int_\Gamma \frac{1}{z} (z\mathbf{I} - \mathbf{A}^*\mathbf{A})^{-1} \mathbf{A}^* dz.$$

[Stewart; Campbell & Meyer]

4. [25 points] Recall the example of image compression using the singular value decomposition that we saw in class. Now it is your turn to compress an image, this time a simple bitmap.

- (a) Write a MATLAB routine to construct the 15×40 matrix \mathbf{A} that is zero everywhere except for ones in the positions marked in the figure below. The upper-left point of the ‘H’ is in the (2,2) entry, and the bottom-right point of the ‘O’ is in the (13,39) entry.



- (b) What are the singular values of \mathbf{A} ? (Use MATLAB’s `svd` and `format long` to print 14 digits after the decimal point.) By counting the number of independent rows and columns of \mathbf{A} , determine the exact rank of \mathbf{A} . Does this agree with your output from `svd`?
- (c) For each k from 1 to $\text{rank}(\mathbf{A})$, compute the rank- k matrix \mathbf{A}_k that best approximates \mathbf{A} in the 2-norm. Visualize this matrix using the commands:

```
imagesc(Ak)
colormap(flipud(gray))
```

[Adapted from Trefethen and Bau, problem 9.3]

5. [25 points] Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ be a full rank matrix, with $m > n$. In general, $\mathbf{A}\mathbf{x} \neq \mathbf{b}$ for all $\mathbf{x} \in \mathbb{C}^n$. The least squares problem amounts to finding the optimal approximation to $\mathbf{b} \in \mathbb{C}^m$ from $\text{Ran}(\mathbf{A})$:

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \min_{\widehat{\mathbf{b}} \in \text{Ran}(\mathbf{A})} \|\mathbf{b} - \widehat{\mathbf{b}}\|_2.$$

In other words, the standard least squares problem seeks the smallest perturbation $\delta\mathbf{b}$ such that there exists some \mathbf{x} for which $\mathbf{A}\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$. Implicitly, we are thus assuming that the matrix \mathbf{A} is exact, but the data \mathbf{b} has some errors.

An alternative approach, called *total least squares*, allows for errors in both \mathbf{A} and \mathbf{b} . Now we look for the smallest $\delta\mathbf{A}$ and $\delta\mathbf{b}$ such that there exists some \mathbf{x} for which $(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$, i.e.,

$$[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}] \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \quad (*)$$

This equation implies that the matrix $[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}] \in \mathbb{C}^{m \times (n+1)}$ has rank less than $n + 1$. (Recall that $m > n$.)

- (a) Use the singular value decomposition of the matrix $[\mathbf{A} \quad \mathbf{b}]$ to describe how to compute the matrix $[\delta\mathbf{A} \quad \delta\mathbf{b}]$ that makes $[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}]$ rank-deficient and minimizes $\|[\delta\mathbf{A} \quad \delta\mathbf{b}]\|_2$.

- (b) Use the optimal $[\delta\mathbf{A} \ \delta\mathbf{b}]$ in (b) to write a simple formula for the solution \mathbf{x} in (*) in terms of appropriate singular values and/or vectors of $[\mathbf{A} \ \mathbf{b}]$.
- (c) Explain why $\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$ cannot be smaller than the smallest singular value of $[\mathbf{A} \ \mathbf{b}]$.
- (d) Compute (in MATLAB) the solution \mathbf{x} produced by (i) standard least squares and (ii) total least squares for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}.$$

For (i), also report $\delta\mathbf{b} = \mathbf{b} - \mathbf{Ax}$ and $\|\delta\mathbf{b}\|$; for (ii), report $\delta\mathbf{A}$, $\delta\mathbf{b}$, and $\|[\delta\mathbf{A} \ \delta\mathbf{b}]\|$ as in part (b).

6. [25 points]

In a celebrated 2003 article in the *Proceedings of the National Academy of Sciences*, Lawrence Sirovich analyzed US Supreme Court voting patterns using the singular value decomposition. Thanks to your skills with the SVD, and data kindly supplied by Prof. Keith Poole at the University of Houston, you can do the same analysis for yourself.

The data is found in `supreme_court.mat`, available from the course web site. With this file in your working directory, the command `load supreme_court` will create an 493-by-9 matrix \mathbf{A} . Each row of this matrix corresponds to a decision made by the between the period that Justice Stephen Breyer joined the court in 1994, and approximately 2002. (The period from Breyer's appointment to Renquist's death in 2005 was one of the longest intervals over which the same nine justices served the court.) Each of the nine columns of \mathbf{A} corresponds to one of the justices. The (j, k) entry of \mathbf{A} corresponds to the opinion of justice k on case j : agreement with the majority is denoted by 1, dissent by 0.

Here is how four famous recent Supreme Court cases would be encoded. The *Bush v. Gore* decision ended the Florida vote recount after the 2000 presidential election; the *Lawrence v. Texas* decision overturned the Texas sodomy law in 2003; the 2004 *Hamdi v. Rumsfeld* decision stated that Yaser Hamdi could challenge his 'enemy combatant' status in court; the *Kelo v. New London* decision, handed down in summer 2005, allows municipalities to seize the land of private citizens for the municipality's economic benefit. These results would contribute the following rows:

	Rehnquist	Stevens	O'Connor	Scalia	Kennedy	Souter	Thomas	Ginsberg	Breyer
<i>Bush v. Gore</i>	1	0	1	1	1	0	1	0	0
<i>Lawrence v. Texas</i>	0	1	1	0	1	1	0	1	1
<i>Hamdi v. Rumsfeld</i>	1	1	1	1	1	1	0	1	1
<i>Kelo v. New London</i>	0	1	0	0	1	1	0	1	1

- (a) Compute the singular values of \mathbf{A} .
- (b) From your answer to part (a), explain why \mathbf{A} might be well approximated by a rank-2 matrix. Report the value of $\|\mathbf{A} - \mathbf{A}_2\|_2$, where \mathbf{A}_2 is that optimal rank-2 approximation to \mathbf{A} .

- (c) The leading right singular vectors \mathbf{v}_1 and \mathbf{v}_2 indicate properties of the most common rows of \mathbf{A} (though the entries will be positive and negative entries and the vectors will be scaled to have norm one).

What are the first two right singular vectors, \mathbf{v}_1 and \mathbf{v}_2 ?

What voting patterns do these two vectors correspond to? That is, for each vector, list the justices most likely to side with the majority, and those that dissent.

From \mathbf{v}_2 , can you deduce the two most frequent ‘swing votes’ — that is, those least tied to the other members of the \mathbf{v}_2 majority?

[Here is an example of how you would interpret the *last* singular vector,

$$\mathbf{v}_9 \approx (0.02, -0.02, -0.03, -0.70, -0.04, 0.02, 0.70, 0.10, -0.03)^T.$$

There are two large entries, -0.70 corresponding to Scalia, and 0.70 corresponding to Thomas. As the *signs* of these entries are opposite, we would interpret this vector as representing decisions in which Scalia and Thomas *disagreed*. (Such cases are quite rare, which explains why this is the least significant singular vector!) The other components, be they positive or negative, are small, so we would regard those justices as ‘swing voters’.]

N.B. For those unfamiliar with the US Supreme Court: the Renquist Court was said to consist of a conservative wing (Rehnquist, Scalia, Thomas), a liberal wing (Stevens, Souter, Ginsberg, Breyer), and two swing voters (O’Connor and Kennedy) who usually aligned with the conservatives but broke from them on key votes, such as *Lawrence v. Texas* above. The last part of question (c) is essentially asking you to assess how well the singular values confirm this conventional wisdom.

Supplemental Problems

S1. Recall that the Frobenius norm is defined as

$$\|\mathbf{A}\|_F = \left(\sum_{j=1}^m \sum_{k=1}^n |a_{jk}|^2 \right)^{1/2}.$$

- (a) Prove that if $\mathbf{Q}_1 \in \mathbb{C}^{m \times m}$ and $\mathbf{Q}_2 \in \mathbb{C}^{n \times n}$ are unitary, then $\|\mathbf{Q}_1 \mathbf{A} \mathbf{Q}_2\|_F = \|\mathbf{A}\|_F$.
 (b) Determine a formula for $\|\mathbf{A}\|_F$ in terms of the singular values of \mathbf{A} , and use this to conclude that

$$\|\mathbf{A}\|_F \leq \sqrt{\text{rank}(\mathbf{A})} \|\mathbf{A}\|_2.$$

- (c) What is the nearest rank-1 matrix to

$$\mathbf{A} = \begin{pmatrix} 1 & M \\ 0 & 1 \end{pmatrix}$$

in the Frobenius norm, where $M \in \mathbb{R}$?

[Golub and Van Loan]

S2. Suppose that $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$.

- (a) Suppose $r < n$, and let \mathcal{Z} be any subset of r distinct integers from $1, \dots, m$.
 Prove that there always exists some nonzero $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{q} := \mathbf{A}\mathbf{z}$ is zero in each entry in \mathcal{Z} , i.e., $q_j = 0$ for all $j \in \mathcal{Z}$.

The rest of this problem walks through a proof (adapted from Powell's *Approximation Theory and Methods*) of the following remarkable fact: there exists some vector $\mathbf{w} \in \mathbb{R}^n$ for which

$$\|\mathbf{b} - \mathbf{A}\mathbf{w}\|_1 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$$

and the residual $\mathbf{b} - \mathbf{A}\mathbf{w}$ has at least n zero entries, i.e., an optimal 1-norm approximation exactly satisfies at least n of the equations in $\mathbf{A}\mathbf{x} \approx \mathbf{b}$.

Suppose that $\|\mathbf{b} - \mathbf{A}\mathbf{w}\|_1 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$, and set $\mathbf{y} := \mathbf{A}\mathbf{w}$. Let \mathcal{Z} denote the subset of the indices $1, \dots, m$ that correspond to zero entries of the vector $\mathbf{b} - \mathbf{A}\mathbf{w}$. We wish to show that \mathcal{Z} must contain at least n distinct indices.

If \mathcal{Z} contains $r < n$ distinct entries, then, from part (a), we can construct some vector $\mathbf{q} = \mathbf{A}\mathbf{z}$ that is zero in every entry in \mathcal{Z} , i.e., $q_j = 0$ for all $j \in \mathcal{Z}$.

Define the function

$$\gamma(\theta) := \|\mathbf{b} - \mathbf{A}(\mathbf{w} + \theta\mathbf{z})\|_1 = \sum_{j=1}^n |b_j - (y_j + \theta q_j)|.$$

- (b) Explain why γ is a continuous and piecewise linear function of θ , why γ must have a minimum at $\theta = 0$, and why $\gamma(\theta) \rightarrow \infty$ as $|\theta| \rightarrow \infty$.
 (c) Explain why γ is constant in a neighborhood about $\theta = 0$. (Hint: use the zero structure in \mathbf{q} .)
 (d) Since $\gamma(\theta) \rightarrow \infty$ as $|\theta| \rightarrow \infty$, $\gamma(\theta)$ cannot be constant for all θ . Explain why, as θ increases from $\theta = 0$, the value of $\gamma(\theta)$ must remain constant until θ reaches some distinguished value $\hat{\theta}$, where $b_j - (y_j + \hat{\theta}q_j) = 0$ for some value of $j \in \{1, \dots, m\}$ for which $q_j \neq 0$, i.e., $j \notin \mathcal{Z}$.
 (e) Explain why this implies that $\mathbf{w} + \hat{\theta}\mathbf{z}$ must be as good an approximation as \mathbf{w} itself was, but that $\mathbf{b} - \mathbf{A}(\mathbf{w} + \hat{\theta}\mathbf{z})$ has (at least) $r + 1$ zeros.

Repeatedly applying this argument, we see that there must exist some best approximation, say $\hat{\mathbf{w}}$ for which $\mathbf{b} - \mathbf{A}\hat{\mathbf{w}}$ has at least n zero entries.