MATH 5524 · MATRIX THEORY

Problem Set 4

Posted Tuesday 28 March 2017. Due Tuesday 4 April 2017. [Corrected 3 April 2017.]
[Late work is due on Wednesday 5 April.]

Complete any four problems, 25 points each.

1. Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a full rank matrix, with $m > n$. In general, $\mathbf{Ax} \neq \mathbf{b}$ 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 $\mathcal{R}(\mathbf{A})$:

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

In other words, the standard least squares problem seeks the smallest perturbation $\delta\mathbf{b}$ such that there exists some $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{b} + \delta\mathbf{b}$. 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.,

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

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

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

   (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 $[\mathbf{A}\ \mathbf{b}]$. (You might note when this construction breaks down, as a unique solution need not always exist.)

   (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

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

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

2. (a) Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a rank-r matrix whose singular value decomposition can be expressed as

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*.$$

   We defined the *pseudoinverse* of $\mathbf{A}$ to be

$$\mathbf{A}^+ = \sum_{j=1}^{r} \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^*.$$

Show that $\mathbf{X} = \mathbf{A}^+$ satisfies the four *Penrose conditions*:

$$\text{(i) } \mathbf{AXA} = \mathbf{A}; \quad \text{(ii) } \mathbf{XAX} = \mathbf{X}; \quad \text{(iii) } (\mathbf{AX})^* = \mathbf{AX}; \quad \text{(iv) } (\mathbf{XA})^* = \mathbf{XA}.$$

(b) Conditions (iii) and (iv) in part (a) might seem trivial, but they are important. Suppose we write the full singular value decomposition of the rank-r matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^*,$$

where $\mathbf{\Sigma}_r = \operatorname{diag}(s_1, \ldots, s_r) \in \mathbb{C}^{r\times r}$, the zero blocks have appropriate dimension (e.g., in the (1,2) entry, $\mathbf{0} \in \mathbb{C}^{r\times(n-r)}$), and $\mathbf{U} \in \mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{C}^{n\times n}$ are unitary. Show that

$$\mathbf{X} = \mathbf{V} \begin{bmatrix} \mathbf{\Sigma}_r^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{L}\mathbf{\Sigma}_r\mathbf{K} \end{bmatrix} \mathbf{U}^*$$

satisfies Penrose conditions (i) and (ii) for any $\mathbf{K} \in \mathbb{C}^{r\times(m-r)}$ and $\mathbf{L} \in \mathbb{C}^{(n-r)\times r}$. Does $\mathbf{X}$ satisfy (iii) and (iv) when $\mathbf{L}$ and $\mathbf{K}$ are nonzero?

(c) For arbitrary $\mathbf{A} \in \mathbb{C}^{m\times n}$, show $\mathbf{A}^+ = \lim_{t\to 0}(\mathbf{A}^*\mathbf{A} + t\mathbf{I})^{-1}\mathbf{A}^*$.

(d) For arbitrary $\mathbf{A} \in \mathbb{C}^{m\times n}$, show $\mathbf{A}^+ = \displaystyle\int_0^\infty \mathrm{e}^{-\mathbf{A}^*\mathbf{A}t}\mathbf{A}^*\,\mathrm{d}t$.

(e) [optional] For arbitrary $\mathbf{A} \in \mathbb{C}^{m\times n}$, show 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

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

[Stewart; Campbell & Meyer]

3. (a) Given the matrix

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

construct a cubic polynomial p such that $p(\mathbf{A}) = (\mathbf{I} + \mathbf{A})^{-1}$.

(b) Given the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

construct a cubic polynomial p such that $p(\mathbf{A}) = \mathrm{e}^{\mathbf{A}}$.

(c) The *Drazin inverse* is an alternative to the pseudoinverse for square matrices; it is defined as follows. Suppose the Jordan canonical form of $\mathbf{A} \in \mathbb{C}^{n\times n}$ can be written as

$$\mathbf{A} = [\,\mathbf{V}_\lambda \quad \mathbf{V}_0\,] \begin{bmatrix} \mathbf{J}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_0 \end{bmatrix} [\,\mathbf{V}_\lambda \quad \mathbf{V}_0\,]^{-1},$$

where $0 \notin \sigma(\mathbf{J}_\lambda)$ and $\{0\} = \sigma(\mathbf{J}_0)$. Then the Drazin inverse can be written as

$$\mathbf{A}^D = [\,\mathbf{V}_\lambda \quad \mathbf{V}_0\,] \begin{bmatrix} \mathbf{J}_\lambda^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\,\mathbf{V}_\lambda \quad \mathbf{V}_0\,]^{-1}.$$

Construct a degree $n-1$ polynomial p such that $\mathbf{A}^D = p(\mathbf{A})$.

(The Drazin inverse plays an important role in the solution of *differential–algebraic equations*. Stewart and Sun write, "The clear winner in the generalized inverse sweepstakes is the pseudoinverse applied to full rank problems. …A distant second is the Drazin generalized-inverse.")

4. In class we will claim that $e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{\mathbf{A}+\mathbf{B}}$ in general. This question investigates some of the subtleties involved in this statement.

   (a) Prove that if $\mathbf{A}$ and $\mathbf{B}$ commute ($\mathbf{AB} = \mathbf{BA}$), then $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$.

   (b) Consider the matrices
$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 2\pi i \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 2\pi i \end{bmatrix}.$$
   Show (by hand) that $\mathbf{A}$ and $\mathbf{B}$ do not commute, yet $e^{\mathbf{A}} = e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}} = \mathbf{I}$. (Hence $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$.)
   [Horn and Johnson]

   (c) Consider the matrices
$$\mathbf{A} = \begin{bmatrix} \pi i & 0 \\ 0 & -\pi i \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$
   Show (by hand) that $\mathbf{A}$ and $\mathbf{B}$ do not commute, but $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{B}}e^{\mathbf{A}} \neq e^{\mathbf{A}+\mathbf{B}}$.
   [Horn and Johnson]

5. This problem concerns the *matrix sign function*. For scalar z, define
$$\operatorname{sign}(z) = \begin{cases} -1, & \operatorname{Re} z < 0; \\ 1, & \operatorname{Re} z > 0; \end{cases}$$
   ($\operatorname{sign}(z)$ is not defined for z on the imaginary axis). The matrix sign function $\operatorname{sign}(\mathbf{A})$ is useful tool in control theory and quantum chromodynamics.

   (a) Let $\mathbf{A} = \mathbf{VJV}^{-1}$ denote the Jordan canonical form of a matrix $\mathbf{A}$ with no purely imaginary eigenvalues. Suppose that $\mathbf{V}$ and $\mathbf{J}$ are partitioned in the form
$$\mathbf{V} = [\,\mathbf{V}_1 \quad \mathbf{V}_2\,], \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{bmatrix},$$
   where all eigenvalues associated with the Jordan blocks in $\mathbf{J}_1$ are in the left half of the complex plane, while those associated with the blocks in $\mathbf{J}_2$ are in the right half plane.
   Use one of our usual approaches for defining $f(\mathbf{A})$ to write down a concise expression for $\operatorname{sign}(\mathbf{A})$ in terms of the Jordan form, and confirm that $\operatorname{sign}(\mathbf{A})^2 = \mathbf{I}$.

   (b) Consider a generic matrix-valued function $\mathbf{F}(\mathbf{X}) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$. Newton's method attempts to compute a solution of the equation $\mathbf{F}(\mathbf{X}) = \mathbf{0}$ via the iteration
$$\mathbf{X}_{k+1} = \mathbf{X}_k - \mathbf{G}(\mathbf{X}_k)^{-1}\mathbf{F}(\mathbf{X}_k),$$
   where $\mathbf{G}(\mathbf{X}_k) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ denotes the *Fréchet derivative* of $\mathbf{F}$ evaluated at $\mathbf{X}_k$; often this is easier to view as solving
$$\mathbf{G}(\mathbf{X}_k)(\mathbf{X}_k - \mathbf{X}_{k+1}) = \mathbf{F}(\mathbf{X}_k).$$
   To compute the matrix sign function, we will show how this method works for $\mathbf{F}(\mathbf{X}) = \mathbf{X}^2 - \mathbf{I}$. We can compute $\mathbf{G}(\mathbf{X})\mathbf{E}$, the Fréchet derivative of $\mathbf{F}$ applied to the matrix $\mathbf{E}$, as the linear term (in $\mathbf{E}$) in the expansion
$$\mathbf{F}(\mathbf{X}+\mathbf{E}) = (\mathbf{X}+\mathbf{E})^2 - \mathbf{I} = (\mathbf{X}^2 - \mathbf{I}) + (\mathbf{XE} + \mathbf{EX}) + \mathbf{E}^2,$$
   i.e., $\mathbf{G}(\mathbf{X})\mathbf{E} = \mathbf{XE} + \mathbf{EX}$. Newton's method seeks the $\mathbf{E}$ that makes $\mathbf{F}(\mathbf{X}+\mathbf{E}) = \mathbf{0}$, neglecting the quadratic $\mathbf{E}$ term, i.e.,
$$\mathbf{XE} + \mathbf{EX} = \mathbf{I} - \mathbf{X}^2.$$

Since we neglected $\mathbf{E}^2$, we do not exactly have $\mathbf{F}(\mathbf{X}+\mathbf{E}) = \mathbf{0}$, so instead we iterate. Given $\mathbf{X}_k$, solve

$$\mathbf{X}_k\mathbf{E}_k + \mathbf{E}_k\mathbf{X}_k = \mathbf{I} - \mathbf{X}_k^2 \tag{$**$}$$

for $\mathbf{E}_k$, and then set $\mathbf{X}_{k+1} = \mathbf{X}_k + \mathbf{E}_k$, and repeat.
All you need to do for part (b) of this problem is to show that

$$\mathbf{E}_k = \frac{1}{2}(\mathbf{X}_k^{-1} - \mathbf{X}_k)$$

satisfies $(**)$ and hence Newton's method yields the iteration

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{X}_k^{-1}).$$

(c) Write a MATLAB code to implement the iteration in part (b), using $\mathbf{X}_0 = \mathbf{A}$.
Test your code out on the matrix in part (e) below.
(Higham observes that "this is one of the rare circumstances in numerical analysis where explicit computation of a matrix inverse is required." So, use `inv` with abandon!)

(d) Suppose $\mathbf{A}$ is Hermitian, and that you have a black box for computing $\operatorname{sign}(\mathbf{A})$.
Describe a (not necessarily efficient!) numerical algorithm for computing all the eigenvalues of $\mathbf{A}$.

(e) Implement the eigenvalue algorithm in part (d), and test it on the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix} \in \mathbb{C}^{16\times 16}.$$

(Within your algorithm, use your Newton-based algorithm from part (c) to compute $\operatorname{sign}(\mathbf{A})$.)

[adapted from Higham]

6. Recall that our earlier block diagonalization of a square matrix required the solution of a *Sylvester equation* of the form $\mathbf{AX}+\mathbf{XB} = \mathbf{C}$. The same equation arises in control theory, where it is common for $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{B} \in \mathbb{C}^{m\times m}$ to both be *stable*, meaning that all of their eigenvalues have negative real part. Make that assumption about $\mathbf{A}$ and $\mathbf{B}$ throughout this problem.

(a) Show that

$$\mathbf{X} = -\int_0^\infty \mathrm{e}^{t\mathbf{A}}\mathbf{C}\,\mathrm{e}^{t\mathbf{B}}\,\mathrm{d}t$$

solves the equation $\mathbf{AX}+\mathbf{XB} = \mathbf{C}$.

(b) Let $\mu \in \mathbb{R}$ be positive. Manipulate $\mathbf{AX}+\mathbf{XB} = \mathbf{C}$ to show that

$$\mathbf{X} = -(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{X}(\mathbf{B}+\mu\mathbf{I}) + (\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{C}$$

and

$$\mathbf{X} = -(\mathbf{A}+\mu\mathbf{I})\mathbf{X}(\mathbf{B}-\mu\mathbf{I})^{-1} + \mathbf{C}(\mathbf{B}-\mu\mathbf{I})^{-1}$$

and hence conclude that $\mathbf{X}$ satisfies the *Stein equation*

$$\mathbf{X} = \mathbf{A}_\mu\mathbf{X}\mathbf{B}_\mu + \mathbf{C}_\mu,$$

for $\mathbf{A}_\mu := (\mathbf{A}+\mu\mathbf{I})(\mathbf{A}-\mu\mathbf{I})^{-1}$, $\mathbf{B}_\mu := (\mathbf{B}+\mu\mathbf{I})(\mathbf{B}-\mu\mathbf{I})^{-1}$, $\mathbf{C}_\mu := -2\mu(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{C}(\mathbf{B}-\mu\mathbf{I})^{-1}$.

(c) Explain why the series

$$\mathbf{X} = \sum_{j=0}^{\infty} \mathbf{A}_\mu^j \mathbf{C}_\mu \mathbf{B}_\mu^j$$

converges, and show that it solves the Stein equation $\mathbf{X} = \mathbf{A}_\mu \mathbf{X} \mathbf{B}_\mu + \mathbf{C}_\mu$.

(d) In many control theory applications, $\mathbf{C}$ has low rank. Suppose that $\mathbf{C}$ has rank-1 and that $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable: $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ and $\mathbf{B} = \mathbf{Y}\mathbf{\Phi}\mathbf{Y}^{-1}$. Using the partial sum

$$\mathbf{X}_k = \sum_{j=0}^{k-1} \mathbf{A}_\mu^j \mathbf{C}_\mu \mathbf{B}_\mu^j$$

to develop an upper bound on the singular values of $\mathbf{X}$:

$$s_{k+1}(\mathbf{X}) \leq \gamma \rho^k$$

for constants $\gamma \geq 1$ and $\rho \in (0,1)$ that you should specify.

By constructing the partial sum $\mathbf{X}_k$, you have an algorithm for constructing the solution $\mathbf{X}$, called Smith's method or the Alternating Direction Implicit (ADI) method. The bound on $s_{k+1}(\mathbf{X})$ proves the singular values of $\mathbf{X}$ decay exponentially at the rate ρ, a fact with deep implications.

7. Suppose the columns of $\mathbf{A} \in \mathbb{C}^{m\times n}$ for $m \geq n$ are *approximately* orthonormal. In many situations we must assess the *departure* of these columns *from orthonormality*. The quantity $\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\|$ provides one natural way to measure this departure. Another approach comes from the polar decomposition $\mathbf{A} = \mathbf{Z}\mathbf{R}$, where $\mathbf{Z} \in \mathbb{C}^{m\times n}$ is a subunitary matrix (i.e., $\mathbf{Z}^*\mathbf{Z} = \mathbf{I}$) with $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{Z})$. Then one could use $\|\mathbf{A} - \mathbf{Z}\|$ to gauge the departure of $\mathbf{A}$ from orthonormality.

   Show that

$$\frac{\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\|}{1 + s_1(\mathbf{A})} \leq \|\mathbf{A} - \mathbf{Z}\| \leq \frac{\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\|}{1 + s_n(\mathbf{A})},$$

   where $s_1(\mathbf{A})$ and $s_n(\mathbf{A})$ denote the largest and smallest singular values of $\mathbf{A}$. (This bound implies that these two measures of the departure from orthonormality are essentially equivalent.)

   [Higham]