Math 128A, Summer 2019
Lecture 28, 8/8/2019

1 Review: Homework 6

1.1 Problem 5

Consider the Fredholm equation:

$$u(x) + \int_0^1 \underbrace{K(x,y)}\, u(y)\ dy = g(x)$$

which is a rank 1 perturbation of the identity.
versus the Volterra equation:

$$y(t) = y(0) + \int_0^t \underbrace{f(s, y(s))}_{\text{causal}}\ ds$$

and $x_i + \sum_{j=1}^n A_{ij}x_j = b_i$ which is $(I + A)x = b$ so $x = b - Ax$.
Fredholm are generally harder than Voltera, but in our problem we have a special kernel (K for kernel),

$$K(x, y) = \cos(x)\sin(y)$$

with

$$\int_0^1 \cos(x)\sin(y)u(y)\ dy = \cos(x)\int_0^1 \sin(y)u(y)\ dy$$

We need to solve forward and backward simultaneously. Recall that any rank-1 matrix, say A, is a row vector left-multiplied by a column vector:

$$A = ab^T = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix} = \begin{bmatrix} a_1b_1 & \cdots & a_1b_n \\ \vdots & & \vdots \\ a_nb_1 & \cdots & a_nb_n \end{bmatrix}$$

Then let $(I + A)x = v$, so that:

$$\begin{aligned} x + a(b^Tx) &= v \\ x &= v - \underbrace{(b^Tx)}_{\text{scalar}}\, a \end{aligned}$$

We have an idea: let's take the dot product across the equation with b.

$$\begin{aligned} b^Tx &= b^Tv - (b^Tx)(b^Ta) \\ (1 + b^Ta)b^Tx &= b^Tv \\ b^Tx &= \frac{b^Tv}{1 + b^Ta}, \end{aligned}$$

if $b^Ta \neq -1$. This is how we solve this rank-1 perturbation problem. In this particular problem, we have a rank-1 kernel, and thus our problem has a rank-1 perturbation.
In our problem, once we discretize the equation via gaussian quadrature, we take on the linear system to solve via Gaussian elimination with partial pivoting (GEPP).

2 Review

Yesterday, we took A symmetric positive definite, with $x^TAx > 0$ for $x \neq 0$ and $A = A^T$. We computed the Cholesky factorization,

$$A = R^TR = LL^T,$$

where R is (right) upper-triangular with $r_{ii} > 0$, we check:

$$\begin{aligned} x^TAx &= (Rx)^T(Rx) \\ &= \|Rx\|^2 > 0, \end{aligned}$$

if $x \neq 0$, because R is invertible.
One of the checks if A is symmetric positive definite is for us to compute the cholesky factorization, and if multiplying them together doesn't return A, then A is not symmetric positive definite (we have nothing to lose).
Recall that we have three ways to compute the Cholesky factorization:
(1) $LDL^T = \underbrace{(\sqrt{D}L^T)^T}_{R^T}\underbrace{(\sqrt{D}L^T)}_{R}$
(2) Direct calculation, solving for one element at a time (and using symmetry). We specify an algorithm to do this. First we write the underlying math:

$$i \geq i: \quad a_{ij} = \sum_{k=1}^{i} r_{ki}\, \underbrace{r_{kj}}$$

where $r_{kj} = r_{1j}r_{1j} + r_{2i}r_{2j} + \cdots + r_{ii}\overbrace{r_{ij}}$ where r_{ij} is what we solve for first, then the rest one at a time.
Additionally,

$$a_{ii} = \sum_{k=1}^{1} r_{ki}^2$$

so we find a_{11} first: $r_{11} = \sqrt{a_{11}}$.
and note $r_{kj} = 0$ when $k > j$ and $r_{ki} = 0$ when $k > i$. Our pseudocode is as follows: Find column j of $\mathbb{R}$.

```
for j = 1:n
  for i = 1:j-1
    r_{ij} = ( a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} ) / r_{ii
    }
  end
  r_{jj} = \sqrt{ a_{jj} - \sum_{k=1}^{j-1} r_{kj}^2 }
end
```

(3) Use Newton's Method for the function we are trying to solve:

$$f(R) = R^TR - A = 0.$$

and of course, R is an upper-triangular matrix with nonnegative diagonal entries. We can say we want to find a correction E such that:

$$(R+E)^T(R+E) - A = E^TE,$$

where we usually set this equal to 0, but instead we cancel out the quadratic term. We check the case that if this is exact, then we would have to deal with a quadratic term, so we insert (recall from yesterday):

$$E := \text{uph}(R^TAR^{-1} - I)R$$

On the other hand, if F is the exact correction to the Cholesky factorization (and E is only what we can get),

$$(R+F)^T(R+F) - A = 0$$

Comparing these,

$$\begin{aligned} R^TR - A + E^TR + R^TE + E^TE &= E^TE \\ R^TR - A + F^TR + R^TF + F^TF &= 0 \end{aligned}$$

where subtracting gives

$$(E-F)^TR + R^T \underbrace{(E-F)}_{(R+E)-(R+F)} = F^TF$$

and rewriting,

$$\underbrace{R^T}_{low.tri}\underbrace{(E-F)^T}_{low.tri} + \underbrace{(E-F)}_{up.tri}\underbrace{R^{-1}}_{up.tri} = R^{-T}F^TFR^{-1},$$

where a lower-triangular times a lower-triangular matrix gives a lower triangular matrix and accordingly for upper-triangular matrices. Hence this equation takes the form of a lower-triangular matrix plus an upper-triangular matrix equals a full matrix.
Hence the error in the correction, the error in $(R+E)$ since $(R+F)$ is exactly Cholesky factor.
We have, for the error relative to the solution we started with,

$$\begin{aligned} (E-F)R^{-1} &= \text{uph}\left((FR^{-1})^T(FR^{-1})\right), \\ E-F &= \text{uph}\left(R^{-T}\underbrace{F^TF}_{(\text{err in } R)^T\cdot(\text{err in } R)}R^{-1}\right)R \end{aligned}$$

which is in terms of the relative error of the approximate old solution, so is **quadratic convergence** a la **Newton's**.

3 Floating Point Errors in Cholesky Factorization

When we compute the Cholesky factorizations,

$$fl(r_{11}) = fl(\sqrt{a_{11}}) = \sqrt{a_{11}}(1+\delta_{11}),$$

where $|\delta_{11}| \le \varepsilon$ is the exact result, correctly rounded. This gives:

$$\begin{aligned} r_{11}^2 &= a_{11}(1+\delta_{11})^2 \\ &= a_{11}(1+2\delta_{11}) + \cancel{O(\varepsilon^2)} \\ &= \hat{a}_{11}, \end{aligned}$$

where $|\hat{a}_{11} - a_{11}| \le 2\varepsilon|a_{11}|$. Our hope is that

$$\hat{R} = fl(R)$$

is the **exact** Cholesky factor of

$$\left|\hat{A} - A\right| \le O(n)\varepsilon|A| = [|a_{ij}|].$$

Because we don't yet have the notion of a norm, we check each entry value and check that each can be bounded.
Remember from our algorithm earlier for directly computing the Cholesky factorization,

$$r_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} \right) / r_{ii}$$

where we can only work with what we know ('the guys with hats'). Recalling $fl(a-b) = a(1+\delta) - b(1+\delta)$, what actually happens is:

$$\hat{r}_{ij} = \left(a_{ij}(1+\delta_{ij}) - \sum_{k=1}^{i-1} \hat{r}_{ki}\hat{r}_{kj}(1+i\delta_{ij}) \right) / (r_{ii}(1+\delta_{ii}))$$

We tend to not like forward error analysis (as the above). We prefer backwards error analysis as the computed results almost satisfying the exact equation it is trying to solve.
Multiplying across by the denominator gives:

$$\hat{r}_{ii}\hat{r}_{ij}(1+\delta''_{ii}) = a_{ij}(1+\delta_{ij}) - \sum_{k=1}^{i-1} \hat{r}_{ki}\hat{r}_{kj} - \sum_{k=1}^{i-1} \hat{r}_{ki}\hat{r}_{kj} i d_{ij}$$

What we have is doing exactly what we want, if we interpret everything a little differently.

$$\hat{r}_{ii}\hat{r}_{ij} = \underbrace{\left[a_{ij}(1+\delta_{ij}) - \sum_{k=1}^{i} \hat{r}_{ki}\hat{r}_{kj} i\delta_{ij} \right]}_{\text{scapegoat}} - \sum_{k=1}^{i-1} \hat{r}_{ki}\hat{r}_{kj}$$

and this is the exact Cholesky factor of our matrix, where we labeled the bracketed quantity as the 'scapegoat' as the source of our error.
So we conclude:

$\hat{R}$ is the exact Cholesky factor of $\hat{A} = \left[a_{ij}(1+\delta_{ij}) - \sum_{k=1}^{i} \hat{r}_{ki}\hat{r}_{kj} i\delta_{ij} \right]$, and we write the error as a good quantity plus an uncertain quantity:

$$|\hat{A} - A| \leq \varepsilon |A| + (n\varepsilon)|\hat{R}^T||\hat{R}|,$$

where $|M|$ denotes taking the absolute value of every element.

This means that all we need to do is take the absolute value of all entries, place them into a matrix R, multiply it by its transpose, then we get close to the exact Cholesky factors.
We may also want to check how close is $\hat{R}$ to **the exact** cholesky factor? Then, what can we say about how close $\hat{R}$ is to R. This gives a measure of **how sensitive** the Cholesky factors are to **perturbations** in A.

4 Sensitivity to Perturbations

Generally we look at the sensitivity of the inverse A^{-1} by perturbations in A, say $A \mapsto A + E$. We write per element, via geometric series:

$$\frac{1}{a+e} = \frac{1}{a\left(1+\frac{e}{a}\right)} = \frac{1}{a}\left(1 - \frac{e}{a} + O(\frac{e}{a})^2\right)$$

Now how do we do this for matrices?

$$(A+E)^{-1} = \left(A(I+A^{-1}E)\right)^{-1}$$

We need a geometric series for matrices. Recall the derivation for the 1-dimensional case:

$$\begin{aligned}
\frac{1-x^{n+1}}{1-x} &= 1+x+x^2+\cdots+x^n \\
1-x^{n+1} &= (1-x)(1+x+\cdots+x^n) \\
&= (1+x+\cdots+x^n) - \left(x+x^2+\cdots+x^{n+1}\right) \\
&= 1-x^{n+1}.
\end{aligned}$$

and in the matrix case,

$$\begin{aligned}
\left(I-A^{n+1}\right)(I-A)^{-1} &= I+A+\cdots+A^n \\
\left(I-A^{n+1}\right) &= (I+A+\cdots+A^n)(I-A) \\
&= I+A+\cdots+A^n - \left(A+\cdots+A^{n+1}\right)
\end{aligned}$$

The first condition for an infinite geometric series of matrices is that $(I-A)^{-1}$ must **exist**, and the second is that $A^n \xrightarrow{n\to\infty} 0$. It turns out this is equivalent ($\iff$) all eigenvalues of A are < 1 in absolute values: $|\lambda_j(A)| < 1, \forall j$.

Lecture ends here.

Next time, we'll think of a way to prove this **without** using Jordan Normal Form. We'll talk about **perturbation theory** and perhaps **norms**.