Numerical Linear Algebra

Least squares problems

Gerard Sleijpen and Martin van Gijzen

November 4, 2009

TU Delft

Delft University of Technology

Program November 4

- Least squares problems
- The SVD
- Regularisation
- CG for the normal equations
- LSQR and Bi-diagonalization

Least squares problems

In this lesson we consider the problem

$$Ax = b$$

with $A \in \mathbb{C}^{m\times n}$, $x \in \mathbb{C}^n$ and $b \in \mathbb{C}^m$.
Furthermore,

- The system may be inconsistent ($b \notin \mathcal{R}(A)$).
- Usually $m >> n$.
- The rank of A may be smaller than n.

Least squares problems (2)

The system $Ax = b$ may be inconsistent. We therefore solve it in the sense of least squares, meaning that we solve the minimisation problem

$$\min_x \|Ax - b\|_2$$

Solutions to this problem satisfy the normal equations

$$A^H A x_{LS} = A^H b$$

and hence

$$r_{LS} = b - Ax_{LS} \perp \mathcal{R}(A)$$

If $rank(A) < n$ the least squares solution is not unique.

Least squares problems (3)

Suppose $rank(A) < m$ and x_{LS} is a least-squares solution. Then

$$\hat{x} = x_{LS} + y \text{ with } y \in \mathcal{N}(A)$$

is also a least squares solution.

A unique least squares solution x_{LSMN} is the one with minimum norm, which is the solution of the constrained problem

$$\min_x \|Ax - b\|_2 \text{ subject to } x \perp \mathcal{N}(A) \text{ .}$$

The Singular Value Decomposition

Let $A \in \mathbb{C}^{m\times n}$ be a matrix of rank r. Then there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that

$$A = U\Sigma V^H, \quad \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$$

where $\Sigma \in \mathbb{R}^{m\times n}$ and $\Sigma_r = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$, and

$$\sigma_1 \geq \sigma_1 \geq \cdots > 0.$$

The σ_i are called the singular values of A.

The SVD and the LSMN solution

The least-squares minimum norm solution can be computed using the SVD by

$$x_{LSMN} = V \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H b$$

The matrix

$$A^+ = V \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H$$

is called the pseudoinverse or the Moore-Penrose inverse of A.

The SVD and the LSMN solution (2)

Proof that $x_{LSMN} = A^+ b$:

$$z = V^H x = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad c = U^H b = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

where $z_1, c_1 \in \mathbb{C}^r$. Then

$$\|b - Ax\|_2 = \|U^H(b - AVV^H x)\|_2 =$$

$$= \| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \|_2 = \| \begin{pmatrix} c_1 - \Sigma_r z_1 \\ c_2 \end{pmatrix} \|_2$$

Hence $\|b - Ax\|_2$ is minimized by $z_1 = \Sigma_r^{-1} c_1$ and $\|x\|$ by $z_2 = 0$.

Noisy problems

In least-squares problems b often corresponds to measured date, which means that we are actually solving the noisy problem

$$Ax = b + \delta b\ .$$

Moreover, small singular values typically correspond to the noise.

These small singular values have a dramatic effect on the LSMN-solution (why?)!!!

This is an example of a so-called ill-posed problem: small perturbations in the data give a large perturbation in the solution.

Regularization

Limiting this effect is called regularization. Several regularization methods have been proposed:

- Set small singular values to 0. This requires the explicit calculation of the SVD, which is not possible for large scale problems.

- (Tykhonov regularisation) Solve the damped least squares problem:

$$\min_x \left\| \begin{pmatrix} A \\ \tau I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2$$

- Use an iterative method (reason: convergence to small singular values is slow)

CGLS (1)

CG can always be applied to the normal equations

$$A^H A x = A^H b$$

since $A^H A$ is Hermitian positive semi-definite.

The stability can be improved by replacing inner products

$$p^H (A^H A p)$$

by inner products

$$(Ap)^H Ap$$

which leads to the algorithm CGLS.

CGLS (2)

$r_0 = b - Ax_0;$	$z_0 = A^H r_0, p_0 = z_0$	initialization
FOR	$k = 0, 1, \cdots ,$ DO	
	$w_k = Ap_k$	
	$\alpha_k = \frac{z_k^H z_k}{w_k^H w_k}$	
	$x_{k+1} = x_k + \alpha_k p_k$	update iterate
	$r_{k+1} = r_k - \alpha_k w_k$	update residual
	$z_{k+1} = A^H r_{k+1}$	residual normal equati
	$\beta_k = \frac{z_{k+1}^H z_{k+1}}{z_k^H z_k}$	
	$p_{k+1} = z_{k+1} + \beta_k p_k$	update direction vector
END FOR		

CGLS (3)

CGLS can also be used for solving nonsymmetric square systems. However, this has two important disadvantages:

- The work per iteration is twice as much as in CG;
- $K_2(A^H A) = K_2(A)^2$, which means that convergence is often very slow.

CGLS (4), Assignment

Assuming that $K_2(A) = 100$:

1. Give an upper bound on the number of CG iterations required to satisfy $\frac{\|x-x_k\|_A}{\|x-x_0\|_A} < 10^{-6}$.
   Hint: use the upper bound

$$\|x - x_k\|_A \leq 2 \left(\frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} \right)^k \|x - x_0\|_A.$$

2. Answer the same question for CGLS.

CGLS (4), Assignment

Assuming that $K_2(A) = 100$:

1. Give an upper bound on the number of CG iterations required to satisfy $\frac{\|x-x_k\|_A}{\|x-x_0\|_A} < 10^{-6}$.
   Hint: use the upper bound

$$\|x - x_k\|_A \leq 2 \left(\frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} \right)^k \|x - x_0\|_A .$$

2. Answer the same question for CGLS.

Answer: CG: 73, CGLS 726

TU Delft

LSQR

LSQR (Paige and Saunders) is derived by applying Lanczos to

$$\begin{pmatrix} I & A \\ A^H & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} .$$

with starting vector $u_1 = \frac{1}{\|b\|} \begin{pmatrix} b \\ 0 \end{pmatrix}$

LSQR (2)

The second vector in the Krylov subspace becomes

$$\frac{1}{\|b\|}\begin{pmatrix} b \\ A^H b \end{pmatrix}$$

After orthonormalisation we obtain

$$\frac{1}{\|A^H b\|}\begin{pmatrix} 0 \\ A^H b \end{pmatrix}$$

Repeating this procedure shows that we get alternatingly

orthogonal vectors $\begin{pmatrix} u \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ v \end{pmatrix}$

LSQR (3)

This observation leads to the following

Bidiagonalisation algorithm (Golub and Kahan)

$\beta_1 u_1 = b \quad \alpha_1 v_1 = A^H u_1$

FOR $i = 1, \cdots$ DO

$$\beta_{i+1} u_{i+1} = A v_i - \alpha_i u_i$$
$$\alpha_{i+1} v_{i+1} = A^H u_{i+1} - \beta_{i+1} v_i$$

END FOR

with $\alpha_i > 0$ and $\beta_i > 0$ such that $\|u_i\| = \|v_i\| = 1$.

LSQR (4)

With $U_k = [u_1, u_2, \cdots, u_k], \quad V_k = [v_1, v_2, \cdots, v_k]$ and

$$B_k = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{bmatrix},$$

it follows that

$$\begin{aligned} \beta_1 U_{k+1} e_1 &= b \\ A V_k &= U_{k+1} B_k \\ A^H U_{k+1} &= V_k B_k^H + \alpha_{k+1} v_{k+1} e_{k+1}^T \end{aligned}$$

LSQR (5)

Now construct solution vectors $x_k = V_k y_k$. Then we get for $r_k = b - Ax_k$:

$$\begin{aligned} r_k &= \beta_1 U_{k+1} e_1 - AV_k y_k \\ &= \beta_1 U_{k+1} e_1 - U_{k+1} B_k y_k \\ &= U_{k+1}(\beta_1 e_1 - B_k y_k) \\ &= U_{k+1} t_k \end{aligned}$$

LSQR (6)

Substitution in the augmented system and using the Gallerkin condition gives

$$\begin{pmatrix} U_{k+1}^H & 0 \\ 0 & V_k^H \end{pmatrix} \begin{pmatrix} I & A \\ A^H & 0 \end{pmatrix} \begin{pmatrix} U_{k+1}t_{k+1} \\ V_k y_k \end{pmatrix} = \begin{pmatrix} U_{k+1}^H b \\ 0 \end{pmatrix},$$

which leads to the reduced system

$$\begin{pmatrix} I & B_k \\ B_k^H & 0 \end{pmatrix} \begin{pmatrix} t_{k+1} \\ y_k \end{pmatrix} = \begin{pmatrix} \beta_1 e_1 \\ 0 \end{pmatrix}.$$

LSQR (7)

This last equation is equivalent to the least squares problem

$$\min \|\beta_1 e_1 - B_k y_k\|_2$$

In LSQR this problem is solved using the QR-algorithm.

LSQR is famous for its robustness.

Final remarks

Today we have seen CG-type methods for the normal equations.

These methods can also be applied to nonsymmetric systems.

The disadvantage of this approach is that the condition number is squared compared to the original system. This may lead to slow convergence and/or an inaccurate solution.

However, there are also classes of problems for which the normal equations approach works quite well, in particular if A is close to an orthogonal matrix.