Numerical Linear Algebra

Maya Neytcheva, TDB, February-March 2021

Plan of the lecture

- Singular value decomposition - brief recollection
- Pseudoinverses
- Least Squares problems - brief recollection
- Solution methods for LS problems - CGLS

Singular value decomposition

Let $A(m,n)$, $n \leq m$ or $n \geq m$, $rank(A) = rank(A^*)k$.

Definition

If there exist $\mu \neq 0$ and vectors **u** and **v**, such that

$$A\mathbf{v} = \mu\mathbf{u} \quad \text{and} \quad A^*\mathbf{u} = \mu\mathbf{v}$$

then μ is called a singular value of A, and $\mathbf{u}, \mathbf{v}$ are a pair of singular vectors, corresponding to μ.

The existence of singular values and vectors is shown...

via the following construction:

$$A\mathbf{v} = \mu\mathbf{u}, \quad A^*\mathbf{u} = \mu\mathbf{v}$$

can be written as

$$\widetilde{A}\begin{bmatrix}\mathbf{v}\\ \mathbf{u}\end{bmatrix} = \begin{bmatrix}0 & A\\ A^* & 0\end{bmatrix}\begin{bmatrix}\mathbf{v}\\ \mathbf{u}\end{bmatrix} = \mu\begin{bmatrix}\mathbf{v}\\ \mathbf{u}\end{bmatrix}$$

The matrix $\widetilde{A}$ is selfadjoint, has real eigenvalues and a complete eigenvector space.
Furthermore, μ^2 is an eigenvalue of A^*A with eigenvector **u** and of AA^* with eigenvector **v**, because

$$\begin{array}{rcl} A\mathbf{v} = \mu\mathbf{u}, & \rightarrow & A^*A\mathbf{v} = \mu A^*\mathbf{u} = \mu^2\mathbf{v}\\ A^*\mathbf{u} = \mu\mathbf{v}, & \rightarrow & AA^*\mathbf{u} = \mu A\mathbf{v} = \mu^2\mathbf{u}\end{array}$$

Singular Value Decomposition

Theorem (SVD)

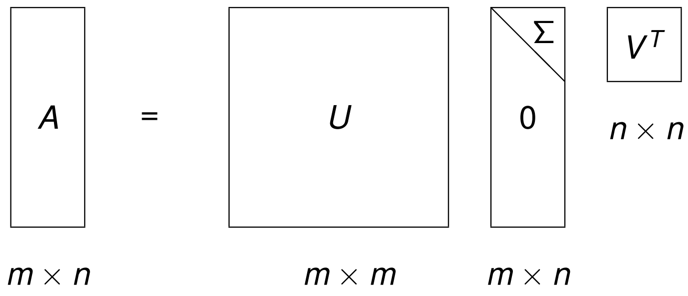
Any $m \times n$ matrix A with dimensions, say, $m \geq n$, can be factorized as

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

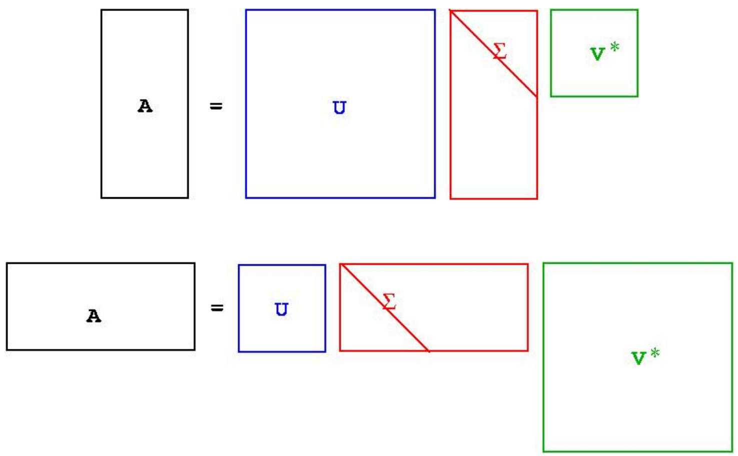
where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal,

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

SVD



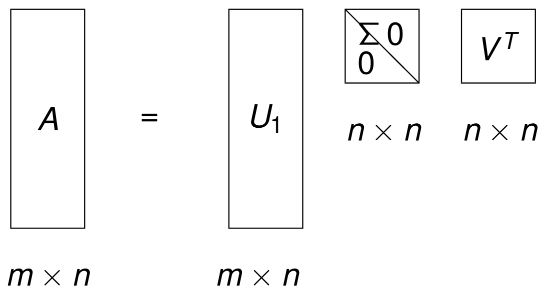
SVD



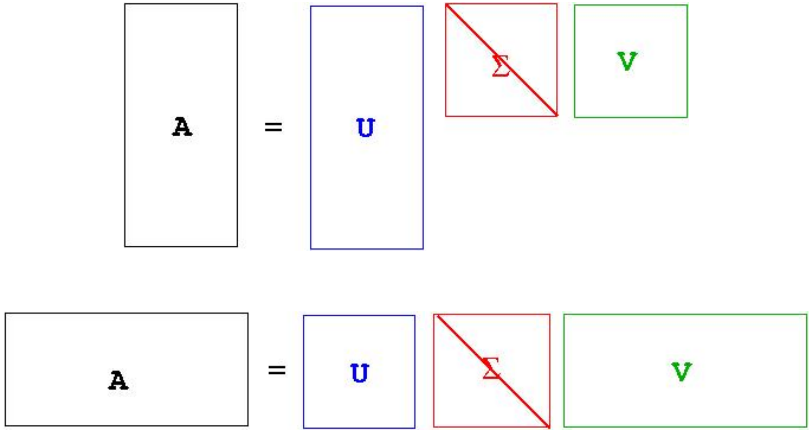
Thin SVD

Partition $U = (U_1\ U_2)$, where $U_1 \in \mathbb{R}^{m\times n}$,

$$A = U_1 \Sigma V^T,$$



Thin SVD



Fundamental Subspaces I

The *range of the matrix A:*

$$\mathcal{R}(A) = \{y \mid y = Ax, \text{ for arbitrary } x\}.$$

Assume that A has rank r:

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Outer product form:

$$y = Ax = \sum_{i=1}^{r} \sigma_i u_i v_i^T x = \sum_{i=1}^{r} (\sigma_i v_i^T x) u_i = \sum_{i=1}^{r} \alpha_i u_i.$$

Fundamental Subspaces II

The *null-space of the matrix A*:

$$\mathcal{N}(A) = \{x \mid Ax = 0\}.$$

$$Ax = \sum_{i=1}^{r} \sigma_i u_i v_i^T x$$

Any vector $z = \sum_{i=r+1}^{n} \beta_i v_i$ is in the null-space:

$$Az = (\sum_{i=1}^{r} \sigma_i u_i v_i^T)(\sum_{i=r+1}^{n} \beta_i v_i) = 0.$$

Fundamental Subspaces

Theorem (Fundamental subspaces)

1. *The singular vectors $u_1, u_2, \ldots, u_r$ are an orthonormal basis in $\mathcal{R}(A)$ and*

$$\mathrm{rank}(A) = \dim(\mathcal{R}(A)) = r.$$

2. *The singular vectors $v_{r+1}, v_{r+2}, \ldots, v_n$ are an orthonormal basis in $\mathcal{N}(A)$ and*

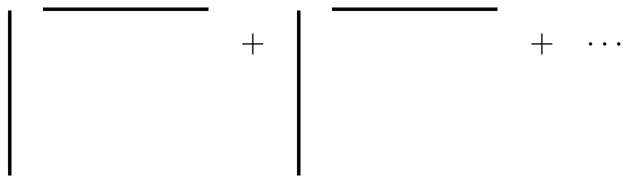
$$\dim(\mathcal{N}(A)) = n - r.$$

3. *The singular vectors $v_1, v_2, \ldots, v_r$ are an orthonormal basis in $\mathcal{R}(A^T)$.*
4. *The singular vectors $u_{r+1}, u_{r+2}, \ldots, u_m$ are an orthonormal basis in $\mathcal{N}(A^T)$.*

SVD matrix expansion

$$A = U\Sigma V^T$$

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T =$$



SVD of a matrix with full column rank I

```
A =  1    1
     1    2
     1    3
     1    4

>> [U,S,V]=svd(A)

U = 0.2195   -0.8073    0.0236    0.5472
    0.3833   -0.3912   -0.4393   -0.7120
    0.5472    0.0249    0.8079   -0.2176
    0.7110    0.4410   -0.3921    0.3824

S = 5.7794         0        V = 0.3220  -0.9467
         0    0.7738            0.9467   0.3220
         0         0
         0         0
```


Thin SVD

```
>> [U,S,V]=svd(A,0)

U = 0.2195   -0.8073
    0.3833   -0.3912
    0.5472    0.0249
    0.7110    0.4410

S = 5.7794         0
         0    0.7738

V = 0.3220   -0.9467
    0.9467    0.3220
```


Rank deficient matrix I

```
>> A(:,3)=A(:,1)+0.5*A(:,2)

A = 1.0000    1.0000    1.5000
    1.0000    2.0000    2.0000
    1.0000    3.0000    2.5000
    1.0000    4.0000    3.0000

>> [U,S,V]=svd(A,0)

U = 0.2612   -0.7948   -0.5000
    0.4032   -0.3708    0.8333
    0.5451    0.0533   -0.1667
    0.6871    0.4774   -0.1667


S = 7.3944         0         0
         0    0.9072         0
         0         0         0
```

Rank deficient matrix II

```
V = 0.2565   -0.6998    0.6667
    0.7372    0.5877    0.3333
    0.6251   -0.4060   -0.6667
```

SVD is rank-revealing!

Null Space

The third column of V is a basis vector in $N(A)$:

```
>> A*V(:,3)

ans =
    1.0e-15 *
          0
    -0.2220
    -0.2220
          0
```


Historical notes

SVD has many different names:

- First derivation of the SVD by Eugenio Beltrami (1873)
- Full proof by Camille Jordan (1874)
- James Joseph Sylvester (1889), independently discovers SVD
- Erhard Schmidt (1907), first to derive an optimal, low-rank approximation of a larger problem
- Hermann Weyl (1912) - determination of the rank in the presence of errors
- Eckart-Young decomposition and optimality properties of SVD (1936), psychometrics
- Numerically efficient algorithms to compute the SVD - works by Gene Golub 1970 (Golub-Kahan)

Best approximation / Eckart-Young Property I

Theorem

Assume that the matrix $A \in \mathbb{R}^{m\times n}$ has rank r and choose k, such that $r > k$. The Frobenius norm matrix approximation problem

$$\min_{\mathrm{rank}(Z)=k} \|A - Z\|_F$$

has the solution

$$Z = A_k = U_k \Sigma_k V_k^T,$$

where $U_k = (u_1, \ldots, u_k)$, $V_k = (v_1, \ldots, v_k)$, and $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$.

Recall: $\|A\|_F = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$

Best approximation / Eckart-Young Property II

Proof:

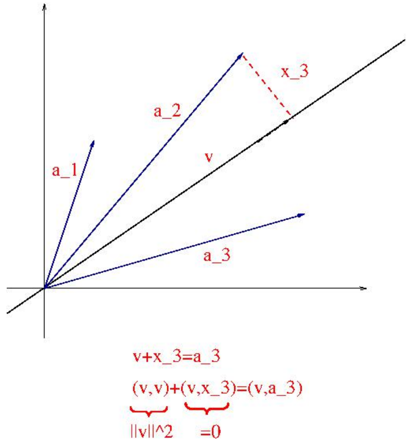
(1) Observe: if $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^*$, then $\|A - A_k\|_2 = \sigma_{k+1}$.

(2) Observe: Consider the subspace, spanned by the first $k+1$ singular vectors of A, W. Then, $\|Aw\|_2 \geq \sigma_{k+1}\|w\|_2, w \in W$.

(3) Assume that there exists a matrix B of rank k, such that $\|A - B\|_2 < \sigma_{k+1}$. Then, there exists a subspace $\widehat{W}$ of size $n-k$, such that $Bw = 0, w \in \widehat{W}$.

$\|Aw\|_2 = \|(A-B)w\|_2 \leq \|A-B\|_2\|w\|_2 \leq \sigma_{k+1}\|w\|_2$. From dimension arguments $W \cap \widehat{W} \neq \emptyset$.

Singular vectors, another view

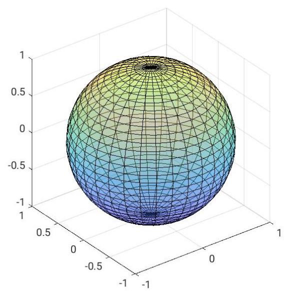
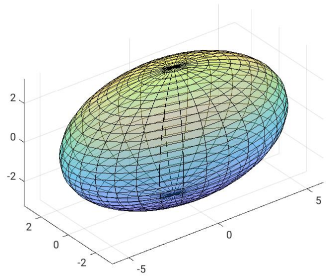


Consider the rows of $A(m,n)$ as points in an n-dimensional space and find the best linear fit through the origin.

$$\mathbf{v}_1 = arg \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|_2^2, \ \sigma_1 = \|A\mathbf{v}_1\|_2$$

$$\mathbf{v}_2 = arg \max_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_1} \|A\mathbf{v}\|_2^2$$

SVD, geometric view

$$A = U\Sigma V^* \quad AV = U\Sigma$$

Solving Least Squares problems by SVD

$$A\mathbf{x} = \mathbf{b}, A(m,n)$$
$$A = U\Sigma V$$
$$U\Sigma V\mathbf{x} = \mathbf{b} \ \rightarrow \mathbf{x} = V(\Sigma^{-1}(U^T\mathbf{b}))$$

Least Squares by SVD I

```
A =   1     1             b = 7.9700
      1     2                10.2000
      1     3                14.2000
      1     4                16.0000
      1     5                21.2000

>> [U1,S,V]=svd(A,0)

U1 =0.1600   -0.7579
    0.2853   -0.4675
    0.4106   -0.1772
    0.5359    0.1131
    0.6612    0.4035
```


Least Squares by SVD II

```
S = 7.6912    0        V = 0.2669   -0.9637
     0    0.9194           0.9637    0.2669

   >> x=V*(S\(U1'*b))

   x = 4.2360
       3.2260
```


Linear dependence – SVD

Theorem

Let the singular values of A satisfy

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Then the rank of A is equal to r.

Rank $=$ the number of linearly independent columns of A.

Linear dependence I

```
A=[1 1; 1 2; 1 3; 1 4]
singval=svd(A)

% Third col=linear combination of first two
A1=[A A(:,1)+0.5*A(:,2)]
singval1=svd(A1)
```

Linear dependence II

Result:

```
A =  1     1          singval = 5.7794
     1     2                     0.7738
     1     3
     1     4

A1 = 1.0000    1.0000    1.5000
     1.0000    2.0000    2.0000
     1.0000    3.0000    2.5000
     1.0000    4.0000    3.0000

singval1 = 7.3944
           0.9072
                0
```


Almost linear dependence I

```
A2=[A A(:,1)+0.5*A(:,2)+0.0001*randn(4,1)]
singval2=svd(A2)

--------------------------------------

A2 = 1.0000    1.0000    1.4999
     1.0000    2.0000    2.0001
     1.0000    3.0000    2.5000
     1.0000    4.0000    3.0001

singval2 = 7.3944
           0.9072
           0.0001
```


Computing the SVD in a numerically efficient way

Computing the SVD

1. Transform A to bidiagonal form by unitary transformations

$$Q_L A Q_R = B = \begin{bmatrix} * & * & & \\ & * & * & \\ & & \ddots & \ddots \\ & & & * \end{bmatrix}$$

2. Diagonalize B by two orthogonal transformations

$$\widetilde{Q}_L B \widetilde{Q}_R = \widetilde{Q}_L Q_L A Q_R \widetilde{Q}_R = \Sigma$$

The cost for the bidiagonalization is $4mn^2 - 4/3n^3$.
The cost for SVD: $4m^2n + 8mn^2 + 9n^3$.

Before defining a pseudoinverse: the inverse of a nonsingular matrix

Nothing easier:
If A is a square nonsingular matrix, then A^{-1} is a matrix of the same size as A, such that

$$A^{-1}A = AA^{-1} = I.$$

Properties:

I1 $(A^{-1})^{-1} = A$

I2 $(A^T)^{-1} = (A^{-1})^T$

I3 $(A^*)^{-1} = (^{-1})^*$

I4 $(AB)^{-1} = B^{-1}A^{-1}$

I5 If $A\mathbf{v} = \lambda\mathbf{v}$ and $A^{-1}\mathbf{w} = \mu\mathbf{w}$ then $\mu = 1/\lambda$.

A definition of a generalized inverse

Any matrix, satisfying

$$AXA = A.$$

Example: Solvability of a linear system $A\mathbf{x} = \mathbf{b}$.
Let $\mathbf{b}$ be in the range of A, i.e., there exist a vector $\mathbf{h}$, such that $\mathbf{b} = A\mathbf{h}$.
If X is a generalized inverse of A, then $\mathbf{x} = X\mathbf{b}$.
If $AXA = A$, then $A\mathbf{x} = AX\mathbf{b} = AXA\mathbf{h} = A\mathbf{h} = \mathbf{b}$

Generalized / Pseudo- inverses

- The Moore-Penrose pseudoinverse
- The Drazin inverse
- Weighted generalized inverses, group inverses
- The Bott-Duffin inverse (for constrained problems)

Moore-Penrose pseudoinverse I

The Moore-Penrose pseudoinverse A^+ is defined for any matrix and is unique.
Moreover, it brings notational and conceptual clarity to the study of solutions to arbitrary systems of linear equations and linear Least Squares problems.

Moore-Penrose pseudoinverse II

Consider $A \in \mathbb{R}_r^{m,n}$. The subscript r denotes the rank of A.

Theorem (Penrose, 1956)

Let $A \in \mathbb{R}_r^{m,n}$. Then $G = A^+$ if and only if

P1 $AGA = A$

P2 $GAG = G$

P3 $(AG)^* = AG$

P4 $(GA)^* = GA$

Furthermore, A^+ always exists and is unique.

The theorem is not constructive but gives criteria that can be checked.

Moore-Penrose pseudoinverse III

Example:
Let $A \in \mathbb{R}_r^{m,n}$.
Then, from the SVD decomposition of $A = U\Sigma V^T$ we find
$A^+ = V\Sigma^+ U^T$, where $\Sigma^+ = \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

Moore-Penrose pseudoinverse IV

Properties:

- $A^+ = (A^T A)^+ A^T = A^T (AA^T)^+$
- $(A^T)^+ = (A^+)^T$
- $(A^+)^+ = A$
- $(A^T A)^+ = A^+ (A^T)^+ = (A^T)^+ A^+$
- $\mathcal{R}(A^+) = \mathcal{R}(A^T) = \mathcal{R}(A^+ A) = \mathcal{R}(A^T A)$
- $\mathcal{N}(A)^+ = \mathcal{N}(AA^+) = \mathcal{N}((AA^T)^+) = \mathcal{N}(AA^T) = \mathcal{N}(A^T)$

Moore-Penrose pseudoinverse V

For linear systems $A\mathbf{x} = \mathbf{b}$ with non-unique solutions (such as under-determined systems), the pseudoinverse may be used to construct the solution of minimum Euclidean norm $\|\mathbf{x}\|_2$ among all solutions.
If $A\mathbf{x} = \mathbf{b}$ is consistent, the vector $\mathbf{x} = A^+\mathbf{b}$ is a solution, and satisfies $\|\mathbf{x}\|_2 \leq \|\mathbf{z}\|_2$ for all possible solutions $\mathbf{z}$.

Uniqueness of the Moor-Penrose pseudoinverse I

Let $A \in \mathbb{R}_r^{m,n}$. Assume that there are two matrices that satisfy the conditions:

$$\begin{array}{ll} AA^+A = A & ABA = A \\ A^+AA^+ = A^+ & BAB = B \\ (AA^+)^* = AA^+ & (AB)^* = AB \\ (A^+A)^* = A^+A & (BA)^* = BA \end{array}$$

Let $M_1 = AB - AA^+ = A(B - A^+)$. By the hypothesis, M_1 is self-adjoint (since it is the difference of two self-adjoint matrices) and

$$\begin{aligned} (M_1)^2 &= (AB - AA^+)A(B - A^+) \\ &= (\underbrace{ABA}_{A} - \underbrace{AA^+A}_{A})(B - A^+) = (A - A)(B - A^+)A = 0. \end{aligned}$$

Uniqueness of the Moor-Penrose pseudoinverse II

Since M_1 is self-adjoint, the fact that $M_1^2 = 0$ implies that $M_1 = 0$:

since for all x one has $\|M_1x\|^2 = (M1x, M1x) = (x, (M1)^2x) = 0$, implying $M_1 = 0$. This showed that $AB = AA^+$.

Following the same steps we can prove that $BA = A^+A$ (consider the self-adjoint matrix $M2 := BAA + A$ and proceed as above). Thus, $A^+ = A^+AA^+ = A^+(AA^+) = A^+AB = (A^+A)B = BAB = B$, thus A^+ is unique.

The Drazin Inverse

Defined for a square matrix.
Let A be a square matrix. The index k of A is the least nonnegative integer k such that $rank(A^{k+1}) = rank(A^k)$.
The Drazin inverse of A is the unique matrix A^D which satisfies

$$A^{k+1}A^D = A^k, \quad A^DAA^D = A^D, \quad AA^D = A^DA.$$

If A is invertible with inverse A^{-1}, then $A^D = A^{-1}$.
Example: Solving systems with a singular matrix by CG.
I. Ipsen, C. Meyer, The idea behind Krylov methods, *The American Mathematical Monthly*, 105 (1998)

Theoretical result

The following statements are equivalent:

- $Ax = b$ has a Krylov solution.
- $b \in R(A^i)$, where i is the index of the zero eigenvalue of A (the index i of an eigenvalue is the maximum size of a block, containing the eivenvalue in the Jordan canonical form).
- A^Db is a solution of $Ax = b$ and it is unique.

Computing the pseudoinverse from SVD

$$A = U\Sigma V^T \;\rightarrow\; A^{\dagger} = V\Sigma^{\dagger}U^T,$$

where $A = U \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T$ and $\Sigma^{\dagger} = \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix}$.

Bott-Duffin inverse

Constrained generalized inverse of a square matrix: We want to solve $Ax = b$, $A(n, n)$, where x should belong to a certain subspace L of R^n.
Denote P_L to be the orthogonal projection on L. Then the constrained problem $Ax = b,\ x \in L$ has a solution if

$$AP_L x = b$$

is solvable.
The generalized Bott-Duffin inverse is defined as

$$A^{(+)} = P_L(AP_L + P_{L^{\perp}})^{-1}$$

if the inverse on the right exists.

Least square problems

Given $A(m, n)$ with full column rank, $b(n, 1)$, consistent with A. We want to solve

$$Ax = b$$

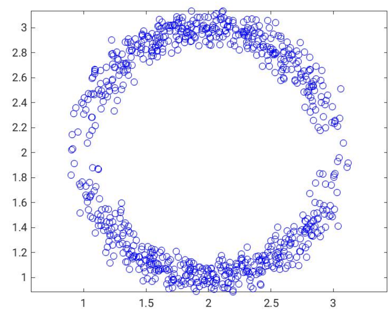
in the Least Squares sense, thus, $x = (A^TA)^{-1}A^Tb$.

We do not want to form A^TA because
– it is usually badly conditioned
– it is in general full even if A is sparse.

A^TA is symmetric positive definite and we have a method for such systems.

An example

Task: find a circle which best fits the points x_i, y_i, lying in a place, as shown in the figure



Thus, seek the best fir circle with radius R and center with coordinates a ad b. The task reduces to minimizing the algebraic distance

$$d(a, b, R) = \sum_{i=1}^{n}((x_i - a)^2 + (y_i - b)^2 - R^2)^2 = \|r\|^2$$

An example, cont

$$d(a,b,R) = \sum_{i=1}^{n}((x_i - a)^2 + (y_i - b)^2 - R^2)^2 = \|r\|^2$$

Here r is a residual vector and is nonlinear in a, b and R. However, we notice that

$$\begin{aligned} r_i &= R^2 - a^2 - b^2 + 2ax_i + 2by_i \\ &= \begin{bmatrix} 2x_i & 2y_i & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ R^2 - a^2 - b^2 \end{bmatrix} - (x_i + y_i)^2 \end{aligned}$$

Thus, the residual is linear wrt $\mathbf{z} = (a, b, R^2 - a^2 - b^2)$.

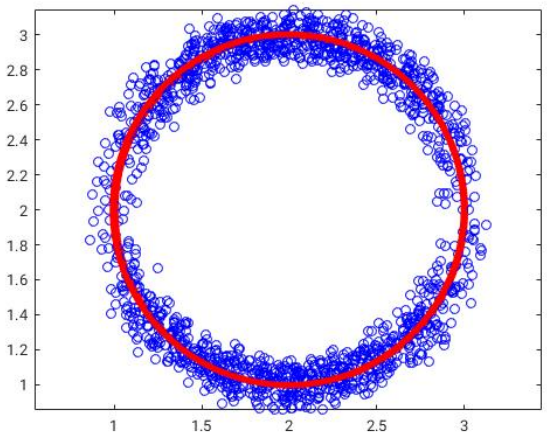
An example, cont.

Formulate as LS problem:

$$A = \begin{bmatrix} 2x_1 & 2y_1 & 1 \\ \vdots & \vdots & \vdots \\ 2x_n & 2y_n & 1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} x_1^{+} y_1^2 \\ \vdots \\ x_n^2 + y_n^2 \end{bmatrix} \quad d(a,b,R) = \|A\mathbf{z} - \mathbf{b}\|^2$$

```
A=[2*x 2*y ones(n,1)]:
b=x.^2+y.^2;
z=A\z;      <---- Solving LS is a linear algebra pro
a=z(1);
b=z(2);
R=sqrt(z(3)+a^2+b^2);
t=linspace(0,2*pi,100);
plot(x,y,'o',a+R*cos(t),b+R*sin(t),'r');
```


An example, cont.



Solving LS via QR and SVD

Via SVD:

$$A\mathbf{x} = \mathbf{b}, A(m,n)$$
$$A = U\Sigma V$$
$$U\Sigma V\mathbf{x} = \mathbf{b} \;\rightarrow \mathbf{x} = V(\Sigma^{-1}(U^T\mathbf{b}))$$

Via QR: $A = QR$, $\quad QR\mathbf{x} = \mathbf{b}, R\mathbf{x} = Q^T\mathbf{b}$

CGLS - Conjugate Gradient for Least Square problems

CGLS: solve the normal equation for $A \in R^{n \times m}$

History:
CG has appeared in a paper by Hestenes and Stiefel (1952). In that paper and in a followup paper by Stiefel (1952), a version of CG for solving the normal equation has peen presented. First result for using a preconditioned CG for solving Least Square problems appears in a paper by Lächli (1959).

CGLS

Recall the definition of a Krylov subspace, based on a vector $\boldsymbol{v} \in R^n$ and a matrix $B \in R^{n \times n}$,

$$\mathcal{K}_k(B, \boldsymbol{v}) = span\{\boldsymbol{v}, B\boldsymbol{v}, B^2\boldsymbol{v}, \cdots, B^{k-1}\boldsymbol{v}\}.$$

Let A be rectangular and denote $A^\dagger$ be its pseudoinverse. Denote $\widehat{\boldsymbol{x}} = A^\dagger \boldsymbol{b}$ - the pseudoinverse solution and the corresponding residual $\widehat{\boldsymbol{r}} = A\widehat{\boldsymbol{x}}$. Then, in the CG framework, $\widehat{\boldsymbol{x}}^k$ minimizes the following error functional:

$$E_\mu(\widehat{\boldsymbol{x}}^k) = (\widehat{\boldsymbol{x}} - \boldsymbol{x}^k)^T (A^T A)^\mu (\widehat{\boldsymbol{x}} + \boldsymbol{x}^k)$$

where $\widehat{\boldsymbol{x}}^k = (x)^0 + \mathcal{K}_k(A^T A, (s)^0),\ \boldsymbol{s}^0 = A^T(\boldsymbol{b} - A\boldsymbol{x}^0)$.

CGLS I

$$E_\mu(\boldsymbol{x}^k) = (\widehat{\boldsymbol{x}} - \boldsymbol{x}^k)^T (A^T A)^\mu (\widehat{\boldsymbol{x}} + \boldsymbol{x}^k)$$

Values of μ of practical interest:

- $\mu = 0$ minimizes $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^k\|_2^2$
- $\mu = 1$ minimizes $\|\widehat{\boldsymbol{r}} - \boldsymbol{r}^k\|_2^2 = \|\widehat{\boldsymbol{r}}\|_2^2 - \|\boldsymbol{r}^k\|_2^2$
  (due to the orthogonality relation $\widehat{\boldsymbol{r}} \perp \widehat{\boldsymbol{r}} - \boldsymbol{r}^k$)
- $\mu = 2$ minimizes $\|A^T(\widehat{\boldsymbol{r}} - \boldsymbol{r}^k)\|_2^2$

$\mu = 0$ - feasible only for consistent systems.
$\mu = 1$ - CGLS

CGLS II

Properties of CGSL:

- $E_\mu(\boldsymbol{x}^k)$ decreases monotonically.
- For $\mu = 1, 2$, $E_\nu(\boldsymbol{x}^k)$ decreases monotonically for all $\nu \le \mu$.
- for $\mu = 1$ also $\boldsymbol{r}^k$ decreases monotonically.
- The rate of convergence is estimated as follows:

$$E_\mu(\boldsymbol{x}^k) < 2\left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right)^k E_\mu(\boldsymbol{x}^0),$$

  where $\varkappa = \varkappa(A^TA)$.
- For $\mu = 1$, both $\|\widehat{\boldsymbol{r}} - \boldsymbol{r}^k\|$ and $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^k\|$ decrease monotonically, however $\|A^T\boldsymbol{r}^k\|$ does oscillate (not due to roundoff errors).

Algorithm CGLS

Unpreconditioned CG	Unreconditioned CGLS
x = x0	x = x0,
r = b -A*x	r = b-A*x; „
delta0 = (r,r)	g = s = A^T*r
g = r	delta0 = (s,s)
Repeat: h = A*g	Repeat: h = A*s
tau = delta0/(g,h)	tau = delta0/(h,h)
x = x + tau*g	x = x + tau*s
r = r + tau*h	r = r - tau*h
delta1 = (r,r)	s = A^T*r
if delta1 <= eps, stop	delta1 = (s,s)
beta = delta1/delta0	if delta1 <= eps, stop
g = r + beta*g	beta = delta1/delta0
	g = s + beta*g

CGSL I

Note: $x, g \in R^n$, $r, h \in R^m$, $(A \in R^{n\times m})$

With $\boldsymbol{s} = A^T(\boldsymbol{b} - A\boldsymbol{x})$, by construction, $\boldsymbol{x}$ minimizes

$$\boldsymbol{s}(A^TA)^{-1}\boldsymbol{s}$$

over the space $\mathcal{K}_k(A^TA, A^T\boldsymbol{b})$.
Thus, $\boldsymbol{s}^k \in T_k,\ T_k = \{A^T(\boldsymbol{b} - A\boldsymbol{x}) \,|\, \boldsymbol{x} \in \mathcal{K}_k(A^TA, A^T\boldsymbol{b})\}$ and any vector from T_k can be expressed as

$$\boldsymbol{s}^k = (I - A^TA\mathcal{P}_{k-1}(A^TA))A^T\boldsymbol{b} = \mathcal{R}_k(A^TA)A^T\boldsymbol{b},$$

where $\mathcal{P}_{k-1}$ is a polynomial of degree $k-1$ and $\mathcal{R}_k$ is a residual polynomial of degree less than or equal k and is normalized at zero, thus $\mathcal{R}_k(0) = 1$.

CGSL II

$$\|\boldsymbol{s}^k\|_{(A^TA)^{-1}} = \min_{\mathcal{R}\in\Pi_k} \|\mathcal{R}_k(A^TA)A^T\boldsymbol{b}^k\|_{(A^TA)^{-1}}$$

Consider the singular value decomposition of A, $A = U\Sigma V$.
Then

$$\boldsymbol{b} = \sum_{i=1}^m b_i\boldsymbol{u}_i, \quad A^T\boldsymbol{b} = \sum_{i=1}^n b_i\sigma_i\boldsymbol{v}_i$$

and

$$\|\boldsymbol{s}^k\|_{(A^TA)^{-1}} \min_{\mathcal{R}\in\Pi_k} \sum_{i=1}^n b_i^2\mathcal{R}_k^2(\sigma_i).$$

CGSL III

$$\|\mathbf{s}^k\|_{(A^TA)^{-1}} \min_{\mathcal{R}\in\Pi_k} \sum_{i=1}^{n} b_i^2 \mathcal{R}_k^2(\sigma_i).$$

Any polynomial from Π_k will give an upper bound. For the choise

$$\mathcal{R}_n(\sigma^2) = \left(1 - \frac{\sigma^2}{\sigma_1^2}\right)\left(1 - \frac{\sigma^2}{\sigma_2^2}\right)\cdots\left(1 - \frac{\sigma^2}{\sigma_n^2}\right)$$

we get $\|\mathbf{s}_n\|_{(A^TA)^{-1}} = 0$, which shows the final termination property of CGLS.
If A has only q distinkt singular values, then CGLS will converge in at most q iterations.

Algorithm: Preconditioned CGLS

A good preconditioner for CGLS: the distinkt singular values of the preconditioned matrix should be very few!
The normal equations for the preconditioned problem in factored form:

$$C^{-T}A^T(AC^{-1}\mathbf{y} - \mathbf{b}) = C^{-T}A^T(A\mathbf{x} - \mathbf{b}) = 0.$$

The convergence now depends on the condition number $\varkappa(AC^{-1})$.

Algorithm: Preconditioned CGLS

Unpreconditioned CGLS	Preconditioned CGLS
x = x0,	x = x0,
r = b-A*x;	r = b-A*x;
g = s = A^T*r	g = s = C^{-1} A^T*r
delta0 = (s,s)	delta0 = (s,s)
Repeat: h = A*s	Repeat: t=C^{-1}s; h = A*s
tau = delta0/(h,h)	tau = delta0/(h,h)
x = x + tau*s	x = x + tau*t
r = r - tau*h	r = r - tau*h
s = A^T*r	s = $C^{-1}A^T$*r
delta1 = (s,s)	delta1 = (s,s)
if delta1 <= eps, stop	if delta1 <= eps, stop
beta = delta1/delta0	beta = delta1/delta0
g = s + beta*g	g = s + beta*g

Demo