Computational data mining

Fatemeh Mansoori

Ax = b

- 1) A is square and invertible and its condition number σ_1/σ_n is not large
  - The elimination will succeed.
  - We have $PA = LU$ or $A = LU$
- 2) m > n = r : There are too many equations Ax = b to expect a solution
  - If the columns of *A* are independent (invertible A^TA) and not too ill-conditioned then we solved the normal equation $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$
  - Vector b is probably not in column space of A, and Ax = b is impossible.
  - $A\widehat{\boldsymbol{x}}$ is the projection of b onto column space of A.
- 3) m < n. Ax = b have many solutions.
  - A has non zero null space.
  - Want to choose best x
  - Possible choices are x+ and x1
  - x+=A^+b. A^+ gives the minimum l2 norm solution with nullspace component = zero
  - x1 = minimum l1 norm solution. (this solution is often sparse: many zero components)

$Ax = b$

- 4) column of A may be in bad condition or near singular
  - A^TA will have very large inverse
  - Orthogonize the columns by Gram-Schmidt algorithm
- 5) column of A may be in bad condition or near singular
  - Gram-Schmidt may fail.
  - A different approach is to add a penalty term
- 6) A is too big
  - Best solution is random sampling of column

Least Squares problem

- Ax = b
  - Have distinct solution : If A is square and rank r = n and b is on the column space of A
  - unsolvable linear equations :
    - A is singular
    - b is not on the column space of A
    - Need to produce the best solution $\widehat{x}$
    - Least square method choose $\widehat{x}$ to make $||b - A\widehat{x}||^2$ as small as possible
    - Minimizing means that its derivation are zero

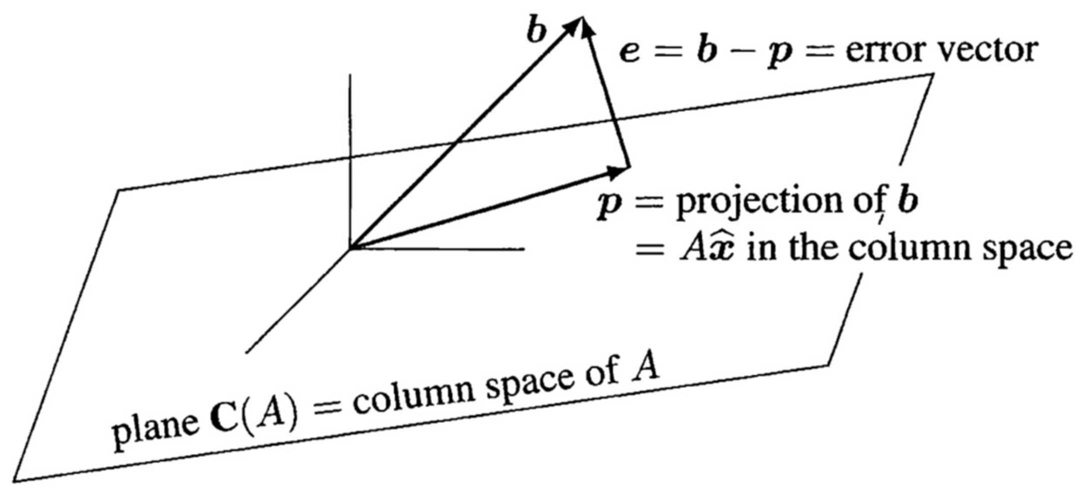
Least square problem

- Suppose you have m measurement mi
- The measurements are noisy.
- You want to find c + dx = b where the best line you can fit to the data
- The solution is to try to minimize the ||b-Ax||. Where :
  - m_is are in the matrix A . A= $\begin{bmatrix} 1 & m_1 \\ 1 & m_2 \\ \vdots & \vdots \\ 1 & m_n \end{bmatrix}$
  - x= $\begin{bmatrix} c \\ d \end{bmatrix}$

Least Squares problem

- b is not in the column space of *A*
  - *Ax* = *b* has no solution
  - project *b* onto the column space of *A*
  - $A^{\mathrm{T}} A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}.$
  - To invert $A^T A$ we need to know that *A* has independent columns
  - Show that e is peroendecular

  to all vector Ax in the column space



Least Squares problem

- A now has independent columns : $r = n$. That makes $A^T A$ positive definite and invertible
- We could check that our $\widehat{\boldsymbol{x}}$ is the same vector $\boldsymbol{x}^+ = A^+\boldsymbol{b}$ that came from the pseudoinverse
- There are no other $\widehat{\boldsymbol{x}}$ because the rank is assumed to be $r = n$
  - nullspace of A only contains the zero vector

When is $A^T A$ Invertible?

- $A^T A$ is invertible exactly when A has independent columns
- Always A and $A^T A$ have the same nullspace
  - Why?

Pseudoinverse of *A*

- If *A* is invertible then A^+ is A^{-1}
- If *A* is *m* by *n* then A^+ is *n* by *m*
- When *A* multiplies a vector *x* in its row space, this produces *Ax* in the column space
  - Those two spaces have equal dimension *r* (the rank).
  - Restricted to these spaces *A* is always invertible and A^+ inverts *A.*
  - $A^+Ax = x$ exactly when *x* is in the row space
  - $AA^+b = b$ when *b* is in the column space

- The nullspace of A^+ is the nullspace of A^T
  - It contains the vectors y in R^m with $A^Ty = 0$
  - vectors y are perpendicular to every Ax in the column space
  - For these y we accept $x^+ = A^+y = 0$ as the best solution to the unsolvable equation $Ax = y$

"pseudoinverse" when *A* has no inverse

Rule 1 If A has independent columns, then $\boldsymbol{A}^{+} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$ and so $A^{+}A = I$.

Rule 2 If A has independent rows, then $\boldsymbol{A}^{+} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$ and so $AA^{+} = I$.

Rule 3 A diagonal matrix Σ is inverted where possible—otherwise Σ^{+} has zeros

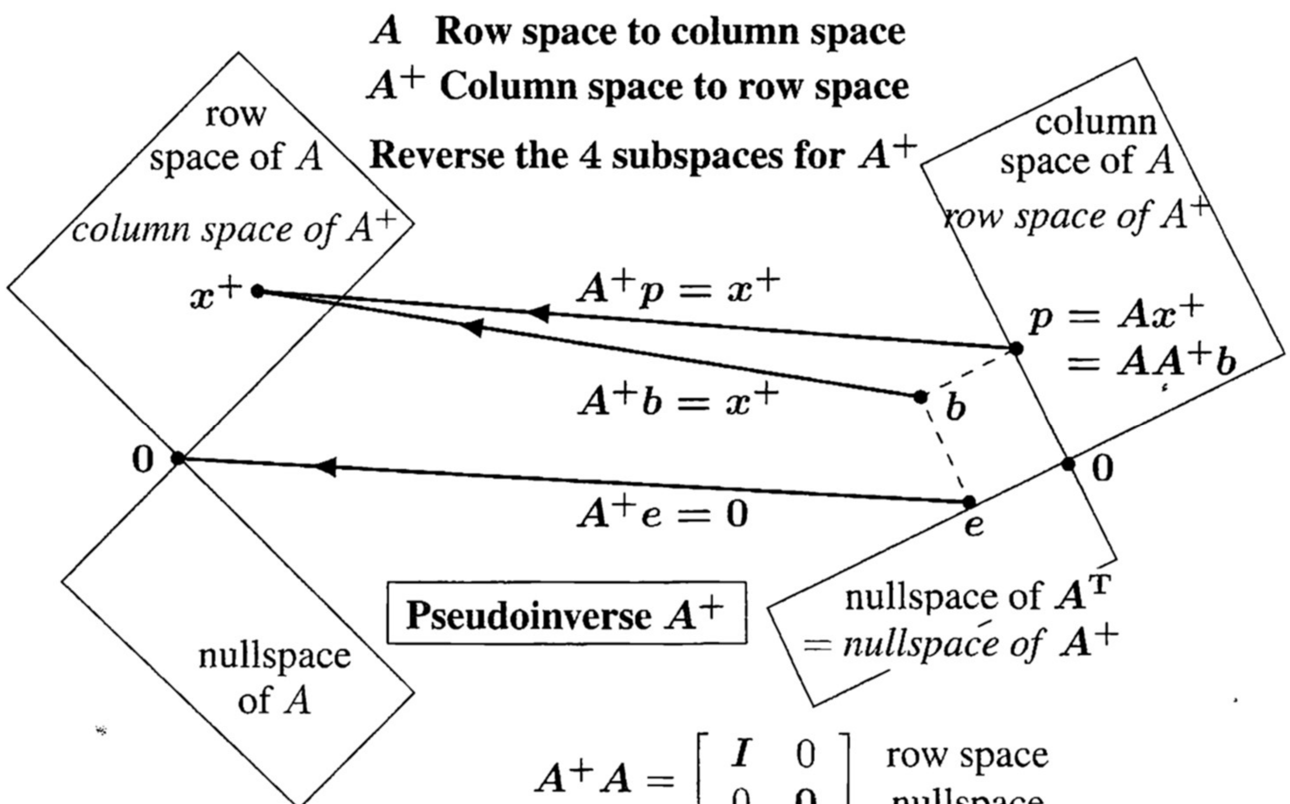
$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma_1} & 0 & 0 & 0 \\ 0 & \boldsymbol{\sigma_2} & 0 & 0 \\ 0 & 0 & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \boldsymbol{\Sigma}^{+} = \begin{bmatrix} \mathbf{1}/\boldsymbol{\sigma_1} & 0 & 0 \\ 0 & \mathbf{1}/\boldsymbol{\sigma_2} & 0 \\ 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 \end{bmatrix}$$

On the four subspaces

$$\boldsymbol{\Sigma}^{+}\boldsymbol{\Sigma} = \boldsymbol{I} \quad \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{+} = \boldsymbol{I}$$

$$\boldsymbol{\Sigma}^{+}\boldsymbol{\Sigma} = \mathbf{0} \quad \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{+} = \mathbf{0}$$

All matrices $\quad$ $\boxed{\textbf{The pseudoinverse of } A = U\Sigma V^{\mathrm{T}} \text{ is } A^{+} = V\Sigma^{+}U^{\mathrm{T}}.}$



$$A^+A = \begin{array}{c} \\ \end{array}\begin{bmatrix} I & 0 \\ 0 & \mathbf{0} \end{bmatrix} \begin{array}{l} \text{row space} \\ \text{nullspace} \end{array}$$

- pseudoinverse *A+* solves the least squares equation $A^TAx = A^Tb$ in one step

$A^+\boldsymbol{b} = V\Sigma^+U^{\mathrm{T}}\boldsymbol{b}$ is best possible

When r < n

- $x^+ = A^+b$ is the minimum norm least squares solution
- When A has independent columns and rank $r = n$
  - X^+ is the only least square solution
- if there are nonzero vectors x in the nullspace of A ($r < n$) they can be added to $x+$
  - The error $||b- A(x^+ +x)||$ is not affected when $Ax = 0$ but the length $|| x^+ +x||$ *will grow*

- Example: find the shortest square solution to $\begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}$

Gram schmidt

- The column of A are still to be independence
- Third approach to find the x is orthogonalzing the columns of A
  - Why : ? The operation count is doubled compared to the $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$,
  - Because : orthogonal vectors provide numerical stability
  - Stability is become import when A^TA is nearly singular.
  - The condition number of A^TA is its norm $||A^TA||$ times $||(A^TA)^{-1}||$
  - When σ_1^2/σ_n^2 is large, it is wise to orthogonalize the columns of A in advance
  -

singular matrix or ill conditioned

- Near singular matrix or ill conditioned
  - Matrix which has its determinate close to zero and whose inverse is unreliable
- The extend of ill-conditioned matrix id defined by its condition number
- Matrix A is ill-conditioned if it is invertible but can become non-invertible(singular) if some of its entries are changed ever so slightly
  - Condition number of A is a measure of how ill-conditioned A is
  - The bigger the condition number is the more ill-conditioned A is
  - Well-conditioned matrices have condition number close to 1

- Solving linear systems whose coefficient matrices are ill-conditioned is tricky
  - A small change in the data (right hand side vector), can lead to radically different answer.
- Example

$$A = \begin{bmatrix} 4.5 & 3.1 \\ 1.6 & 1.1 \end{bmatrix}, \quad b = \begin{bmatrix} 19.249 \\ 6.843 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 19.25 \\ 6.84 \end{bmatrix}.$$

```python
A = np.array([[4.5,3.1],
              [1.6,1.1]])

b1 = np.array([19.249, 6.843]).reshape(2,1)
b2 = np.array([19.25, 6.84]).reshape(2,1)

x1 = np.matmul(np.linalg.inv(A), b1)
print(x1)

x2 = np.matmul(np.linalg.inv(A), b2)
print(x2)


u,s,v = np.linalg.svd(A)
condition = s[0]/s[1]
print(condition)
```

```
[[3.94]
 [0.49]]
[[2.9]
 [2. ]]
3362.999702646099
```

Gram-Schmidt

Independent columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ lead to orthonormal $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$.

$\boldsymbol{q}_1 = \boldsymbol{a}_1 / ||\boldsymbol{a}_1||$.

Gram-Schmidt step **Orthogonalize** $\boldsymbol{A}_2 = \boldsymbol{a}_2 - (\boldsymbol{a}_2^{\mathrm{T}} \boldsymbol{q}_1)\, \boldsymbol{q}_1$
Normalize $\boldsymbol{q}_2 = \boldsymbol{A}_2 / ||\boldsymbol{A}_2||$

$\boldsymbol{A}_3 = \boldsymbol{a}_3 - (\boldsymbol{a}_3^{\mathrm{T}} \boldsymbol{q}_1)\, \boldsymbol{q}_1 - (\boldsymbol{a}_3^{\mathrm{T}} \boldsymbol{q}_2)\, \boldsymbol{q}_2$ **Normalize** $\boldsymbol{q}_3 = \dfrac{\boldsymbol{A}_3}{||\boldsymbol{A}_3||}$

$$\boldsymbol{a}_1 = ||\boldsymbol{a}_1||\, \boldsymbol{q}_1$$
$$\boldsymbol{a}_2 = (\boldsymbol{a}_2^{\mathrm{T}} \boldsymbol{q}_1)\, \boldsymbol{q}_1 + ||\boldsymbol{A}_2||\, \boldsymbol{q}_2$$
$$\boldsymbol{a}_3 = (\boldsymbol{a}_3^{\mathrm{T}} \boldsymbol{q}_1)\, \boldsymbol{q}_1 + (\boldsymbol{a}_3^{\mathrm{T}} \boldsymbol{q}_2)\, \boldsymbol{q}_2 + ||\boldsymbol{A}_3||\, \boldsymbol{q}_3$$

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}$$

$$\boxed{\textbf{Gram-Schmidt produces orthonormal } q\textbf{'s from independent } a\textbf{'s. Then } A = QR.}$$

$\widehat{x} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b$ is $(R^{\mathrm{T}}R)^{-1}R^{\mathrm{T}}Q^{\mathrm{T}}b$. This is exactly $\widehat{x} = R^{-1}Q^{\mathrm{T}}b$.

```python
Q,R = np.linalg.qr(A)
print(Q)
print(R)
qb1 = np.matmul(Q,b1)
qb2 = np.matmul(Q,b2)
print("----------")
print(qb1)
print(qb2)
print("----------")
print(np.matmul(np.linalg.inv(R),np.array([[-20.42],[-0.001]])))
print(np.matmul(np.linalg.inv(R),np.array([[-20.42],[0]])))
```

```
[[3.94]
 [0.49]]
[[2.9]
 [2. ]]
3362.999702646099
-0.009999999999999957
[[-0.94221469 -0.33500967]
 [-0.33500967  0.94221469]]
[[-4.77598157e+00 -3.28937617e+00]
 [ 0.00000000e+00 -2.09381042e-03]]
----------
[[-2.04291617e+01]
 [-1.02596711e-03]]
[[-2.04290989e+01]
 [-4.18762085e-03]]
----------
[[3.94662327]
 [0.47759816]]
[[4.27556088]
 [0.        ]]
```

What is the advantage of Gram Schmidt in solving Least square problem when A is ill conditioned ?

What is the condition number of Q

Gram-Schmidt with Column Pivoting

- straightforward description of Gram-Schmidt worked with the columns of A in their original order a1, a2, a3, ...
- This could be dangerous!
  - Then roundoff error could wipe us out
- We need column exchanges to pick the largest remaining column. Change the order of columns as we go.

Old Accept column $\boldsymbol{a}_j$ as next. Subtract its components in the directions $\boldsymbol{q}_1$ to $\boldsymbol{q}_{j-1}$

New When $\boldsymbol{q}_{j-1}$ is found, subtract the $\boldsymbol{q}_{j-1}$ component from **all remaining columns**

Algorithm for Gram Schmidt with Column Pivoting

$$\boldsymbol{i} \;= \operatorname{argmax} ||A_{j-1}(:,\ \ell)||$$ finds the largest column not yet chosen for the basis

$$\boldsymbol{q}_j = A_{j-1}(:,\ i)/||A_{j-1}(:,\ i)||$$ normalizes that column to give the new unit vector $\boldsymbol{q}_j$

$$Q_j = \begin{bmatrix} Q_{j-1} & \boldsymbol{q}_j \end{bmatrix}$$ updates Q_{j-1} with the new orthogonal unit vector $\boldsymbol{q}_j$

$$\boldsymbol{r}_j = \boldsymbol{q}_j^{\mathrm{T}} A_{j-1}$$ finds the row of inner products of $\boldsymbol{q}_j$ with remaining columns of $\boldsymbol{A}$

$$R_j = \begin{bmatrix} R_{j-1} \\ \boldsymbol{r}_j \end{bmatrix}$$ updates R_{j-1} with the new row of inner products

$$A_j = A_{j-1} - \boldsymbol{q}_j \boldsymbol{r}_j$$ subtracts the new rank-one piece from each column to give A_j

Least Squares with a Penalty Term

- If A has dependent columns and Ax = 0 has nonzero solutions, then A^TA cannot be invertible
  - This is where we need A^+
- A gentle approach will "regularize" least squares

Penalty term	Minimize $\|\|Ax-b\|\|^2+\delta^2\,\|\|x\|\|^2$	Solve $(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}+\boldsymbol{\delta}^{2}\boldsymbol{I})\,\widehat{\boldsymbol{x}}=A^{\mathrm{T}}\boldsymbol{b}$

- This fourth approach to least squares is called ridge regression
- The x in the above equation approaches shortest solution $\boldsymbol{x}^+ = A^+\boldsymbol{b}$

- The difficulty of computing A+ is to know if a singular value is zero or very small

The diagonal entry in Σ^+ is zero or extremely large

$$\textbf{From 0 to } \mathbf{2^{10}} \quad \begin{bmatrix} 2 & 0 \\ 0 & \mathbf{0} \end{bmatrix}^+ = \begin{bmatrix} 1/2 & 0 \\ 0 & \mathbf{0} \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} 2 & 0 \\ 0 & \mathbf{2^{-10}} \end{bmatrix}^+ = \begin{bmatrix} 1/2 & 0 \\ 0 & \mathbf{2^{10}} \end{bmatrix}$$

Pseudoinverse A^+ is the Limit of $(A^TA + \delta^2 I)^{-1}A^T$

- Suppose A is 1 by 1 matrix *(just a single number σ)*

For $\delta > 0$ $\quad (A^TA + \delta^2 I)^{-1}A^T = \left[\dfrac{\sigma}{\sigma^2 + \delta^2}\right]$ is 1 by 1

limit is zero if $\sigma = 0$. The limit is $\dfrac{1}{\sigma}$ if $\sigma \neq 0$.

Pseudoinverse A^+ is the Limit of $(A^TA + \delta^2 I)^{-1}A^T$

- Now suppose a diagonal matrix Σ.
- seeing the 1 by 1 case at every position along the main diagonal

Σ has positive entries σ_1 to σ_r and otherwise all zeros

$(\Sigma^T\Sigma + \delta^2 I)^{-1}\Sigma^T$ has positive diagonal entries $\dfrac{\sigma_i}{\sigma_i^2 + \delta^2}$ and otherwise all zeros.

Positive numbers approach $\frac{1}{\sigma_i}$. Zeros stay zero. When $\delta \to 0$ the limit is again Σ^+.

prove that the limit is A^+ for *every matrix* A.

$$\boxed{\lim_{\delta \to 0} \boldsymbol{V}\left[(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma} + \delta^2 \boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{\mathrm{T}}\right]\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{A}^{+}.}$$

Question

- Is it true that $(AB)^+ = B^+A^+$.
- $A = \begin{bmatrix} 1 & 0 \end{bmatrix}$ **and** $B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$(A^{\mathrm{T}})^+ = (A^+)^{\mathrm{T}}$$

$$(A^{\mathrm{T}}A)^+ = A^+(A^{\mathrm{T}})^+.$$

If C has full column rank and R has full row rank then $(CR)^+ = R^+C^+$ is true.

- By using above equation the pseudoinverse of A could be computed with out using SVD

- Which matrices have $A^+ = A$?