Lecture_notes4

February 23, 2021

1 MATH310 - Lecture_notes4

1.1 Pseudo-inverse

,

Definition (the pseudoinverse of a matrix A with full rank)

The *pseudo-inverse* of a $m \times n$ matrix A with full rank $r = n$ (n is the number of columns in A) is denoted A^+ and defined as

$$\boxed{A^+ = (A^tA)^{-1}A^t.}$$

$A^+A = \underbrace{(A^tA)^{-1}A^t}_{A^+}A$

$= I$

Note that A^+ is $n \times m$ and has the property $A^+A = I_n$.

Exercise 1:
Use the "thin" SVD of $A = U\Sigma V^t$ to show that

$$\boxed{A^+ = V\Sigma^{-1}U^t.}$$

Exercise 2:
Show that the least squares solution of $A\mathbf{x} = \mathbf{b}$ is

$$\boxed{\hat{\mathbf{x}} = A^+\mathbf{b}}$$

when A has full rank.

1.2 The minimum norm solution of underdetermined systems

,

If $rank(A) = r < n$ we say that A is *rank-deficient*.

We say that linear systems $A\mathbf{x} = \mathbf{b}$ with rank-deficient coefficient matrices are *underdetermined*.

,

Definition (the pseudoinverse of any matrix)

Lets extend the definition of the pseudo-inverse to also include matrices $\boxed{A = U_r\Sigma_rV_r^t = \sum_{i=1}^r \sigma_i \mathbf{v}_i\mathbf{u}_i^t}$ of any rank r based on the result of exercise 1:

$$\boxed{A^+ = V_r \Sigma_r^{-1} U_r^t = \sum_{i=1}^{r} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^t.}$$

Underdetermined systems can still be solved, but there is no longer a unique solution $\hat{\mathbf{x}}$.

However

Theorem (Minimum norm solution)

If $rank(A) = r < n$, then there are infinitely many least squares solutions of $A\mathbf{x} = \mathbf{b}$. Among all the least squares solutions, the particular solution

$$\boxed{\mathbf{x}_0 = V_r(\Sigma_r^{-1} U_r^t \mathbf{b}) = A^+\mathbf{b}}$$

has the smallest possible norm.

Proof:
It is clear that $\mathbf{x}_0 \in span(V_r) = Col(A^t)$, i.e. the solution $\mathbf{x}_0$ is a linear combination of the right singular vectors $\mathbf{V}_r = A^t U_r \Sigma_r^{-1}$ that are all linear combinations of the rows in A. The assumption $r < n$ implies that the null-space $Nul(A) \neq \{\mathbf{0}\}$ $(\dim(Nul(A) = n - r)$.

Let $\hat{\mathbf{x}} \neq \mathbf{x}_0$ be any least squares solution, i.e. $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$, where $\hat{\mathbf{b}}$ is the projection of $\mathbf{b}$ onto $Col(A)$. Then $A\hat{\mathbf{x}} = \hat{\mathbf{b}} = A\mathbf{x}_0$ and therefore $A(\hat{\mathbf{x}} - \mathbf{x}_0) = \hat{\mathbf{b}} - \hat{\mathbf{b}} = \mathbf{0}$, i.e. $\hat{\mathbf{x}} - \mathbf{x}_0 \in Nul(A)$. The latter means that $(\hat{\mathbf{x}} - \mathbf{x}_0)$ is orthogonal both to the rows in A, as well as every linear combination of these rows, such as $\mathbf{x}_0$.

Therefore, by Pythagoras theorem we have

$$\boxed{\|\hat{\mathbf{x}}\|^2 = \|(\hat{\mathbf{x}} - \mathbf{x}_0) + \mathbf{x}_0\|^2 = \overset{>0}{\|\hat{\mathbf{x}} - \mathbf{x}_0\|^2} + \|\mathbf{x}_0\|^2 \geq \|\mathbf{x}_0\|^2.}$$

Finally we note that for any vector $\mathbf{n} \in Nul(A)$, $\hat{\mathbf{x}} = \mathbf{x}_0 + \mathbf{n}$ is a least squares solution because $A\hat{\mathbf{x}} = A(\mathbf{x}_0 + \mathbf{n}) = A\mathbf{x}_0 + A\mathbf{n} = \hat{\mathbf{b}} + \mathbf{0} = \hat{\mathbf{b}}$, and because $Nul(A) \neq \{\mathbf{0}\}$ there are infinitely many choices for $\mathbf{n}$. ■

`Exercise 3:`
Verify that $\mathbf{x}_0 = V_r \Sigma_r^{-1} U_r^t \mathbf{b}$ above really is a least sqares solution of $A\mathbf{x} = \mathbf{b}$ when $rank(A) = r < n$.

1.3 The condition number of a matrix and poorly conditioned systems

,

If $A = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^t = U_r \Sigma_r V_r^t$ has rank r, the associated matrix (operator) norm $\|A\|_2$ defined as the supremum of $\|A\mathbf{x}\|_2$ over all unit vectors $\mathbf{x} \in \mathbb{S}^{n-1} = \{\mathbf{x} | \ \mathbf{x}^t\mathbf{x} = 1\} \subseteq \mathbb{R}^n$.

Because the unit sphere $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ is a *compact set* (a set that is closed and bounded in the mathematical sense), and the function $f(\mathbf{x}) = \|A\mathbf{x}\|_2$ is continuous, there is a particular choice $\mathbf{x}_0 \in \mathbb{S}^{n-1}$ that produces the supremum, i.e.

$$\|A\|_2 = \|A\mathbf{x}_0\|_2.$$

Because any candidate unit vector $\mathbf{x} = \sum_{i=1}^{r} c_i \mathbf{v} = V_r \mathbf{c}$ must be a linear combination of the right singular vectors from the reduced SVD of A, we have

$$\|A\mathbf{x}\|_2^2 = \mathbf{x}^t A^t A \mathbf{x} = \mathbf{c}^t V_r^t A^t A V_r \mathbf{c} = \mathbf{c}^t V_r^t V_r \Sigma_r U_r^t U_r \Sigma_r V_r^t V_r \mathbf{c} = \mathbf{c}^t \Sigma_r^2 \mathbf{c} = \sum_{i=1}^{r} c_i^2 \sigma_i^2 \leq \sum_{i=1}^{r} c_i^2 \sigma_1^2 = \sigma_1^2 \sum_{i=1}^{r} c_i^2 = \sigma_1^2.$$

Consequently, by choosing $\mathbf{x}_0 = \mathbf{v}_1$ (the right singular vector associated with the largest singular value σ_1), we obtain the maximum value defining the matrix operator norm $\|A\|_2 = \sqrt{\|A\mathbf{x}_0\|_2^2} = \sigma_1$.

Hence, the matrix norms of A and A^+ are

$$\|A\|_2 = \sigma_1$$

$$\|A^+\|_2 = \sigma_r^{-1}$$

and the **condition number** of A can be expressed in terms of these norms as

$$\boxed{\kappa(A) = \|A\|_2 \|A^+\|_2 = \frac{\sigma_1}{\sigma_r}.}$$

The Julia-script "**Condition_number.jl**" demonstrates some unfavourable consequences for the solution of a system $A\mathbf{x} = \mathbf{b}$ when A has a *large condition number*. Such sysems are called *poorly conditioned*.

1.4 Rank k regularization

,

Definition (Truncated pseudo-inverse)
By omitting the terms in $\sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^t$ associated with some of the smaller singular values, we obtain a *truncated pseudoinverse*. By keeping the k first terms the associated truncated pseudo-inverse of rank k is

$$\boxed{A_k^+ = V_k \Sigma_k^{-1} U_k^t = \sum_{i=1}^{k} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^t.}$$

Definition (rank k regularized solution)

$$\boxed{\mathbf{x} = A_k^+ \mathbf{b}}$$

is called the rank *k regularized* solution of $A\mathbf{x} = \mathbf{b}$.

,

The rank *k regularized* solution of $A\mathbf{x} = \mathbf{b}$ (a.k.a. *reduced rank regression*) is closely related to the so-called *principal component regression (PCR)* based on the first k principal components of the coefficient matrix A (were we assume that all the columns are centred to have mean values equal to 0).

1.5 Principal component regression (PCR)

,

Recall that for principal component analysis (PCA) we assume that the data matrix $X \in \mathbb{R}^{m\times n}$ has rank r, and that each X-column is arranged according to a common ordering of observations of corresponding real-valued random vectors with mean zero.

The data matrix is always assumed to be centered, i.e. the mean of each X-column is equal to 0 (zero).

From the SVD of the data matrix $X = U_r \Sigma_r V_r^t$ we have:

- The right singular vectors $\mathbf{v}_i$ $(i = 1, \cdots, r)$ from the columns of V_r define *the principal components directions* (also known as *the loadings*) of X.
- The left singular vectors $\mathbf{u}_i$ $(i = 1, \cdots, r)$ from the columns of U_r define *the normalized principal components*, and the associated vectors $\mathbf{t}_i = \sigma_i \mathbf{u}_i = X\mathbf{v}_i$ are called the *principal component scores* of X.

The truncated rank k pseudo-inverse of X is $X_k^+ = V_k \Sigma_k^{-1} U_k^t = \sum_{i=1}^k \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^t$.

For a data matrix $X \in \mathbb{R}^{m\times n}$ and response vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$ the corresponding k-component **principal component regression (PCR) coefficients** are defined as the rank k solution

$$\boxed{\boxed{\hat{\beta}_k} = X_k^+ \mathbf{y} = X_k^+ \mathbf{y}_0.}$$

of the system $\boxed{X\beta = \mathbf{y}_0,}$ where $\mathbf{y}_0 = \mathbf{y} - \bar{y}$, $(\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i)$ is the mean centered version of $\mathbf{y}$.

To predict the response value $\hat{y}$ for a new datapoint (sample) $\mathbf{x}^t \in \mathbb{R}^n$ based on the k-component **PCR-model** we also include a constant term $\beta_{0,k}$ to calculate

$$\boxed{\hat{y} = \boxed{\beta_{0,k}} + \mathbf{x}^t \boxed{\hat{\beta}_k}}$$

for $\boxed{\beta_{0,k} = \bar{y} - \bar{\mathbf{x}}^t \hat{\beta}_k}$, where $\bar{\mathbf{x}}^t$ is the (row) vector of column means used for centering of the data matrix X and $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$. Note that for the particular choice $\mathbf{x} = \bar{\mathbf{x}}$ we obtain the prediction

$$\boxed{\hat{y} = \beta_{0,k} + \bar{\mathbf{x}}^t \hat{\beta}_k = \bar{y} - \bar{\mathbf{x}}^t \hat{\beta}_k + \bar{\mathbf{x}}^t \hat{\beta}_k = \bar{y},}$$

i.e. from the mean of the observed X-data we predict the mean of the observed $\mathbf{y}$-data.

`Exercise 4:`

Let $M_k = [\mathbf{u}_0 \ \mathbf{u}_1 \cdots \mathbf{u}_k] \in \mathbb{R}^{m\times(k+1)}$, where $\mathbf{u}_0 = \frac{1}{\sqrt{m}}\mathbf{1} = \begin{bmatrix} \frac{1}{\sqrt{m}} \\ \frac{1}{\sqrt{m}} \\ \vdots \\ \frac{1}{\sqrt{m}} \end{bmatrix}$ is the constant vector of norm 1 in

$\mathbb{R}^m$.

- a) Explain why $M_k^t M_k = I_{k+1}$.
- b) The projection mapping onto the column space $Col(M_k)$ is given by $H_k = M_k M_k^t$. Verify that the projection $\hat{\mathbf{y}} = H_k \mathbf{y}$ of $\mathbf{y}$ onto $Col(M_k)$ is identical to the fitted values for the X-data of the k-component PCR model:

$$\boxed{\hat{\mathbf{y}} = \mathbf{1}\beta_{0,k} + X_1 \hat{\beta}_k,}$$

where $X_1 = \mathbf{1}\bar{\mathbf{x}}^t + X$ is the uncentered version of the $m \times n$ data matrix X.