Best Approximation

Yinan Huang

Nov 1, 2020

Contents

1 Introduction

Best approximation (or least square error estimation) has tremendous applications in solving the linear problem which requires us to minimize the error. If we translate these problems into a projection problem in vector space language, then the best approximation is given by the orthogonal projection onto the solution space. We are going to introduct the best approximation problem in vector space and its dual problem, then apply them on linear approximation problems.

2 Best approximation problem in vector space

2.1 Best approximation problem

The best approximation problem is defined as following.

Definition 2.1. *We call the following a best approximation problem. Given a bunch of linearly independent vectors $\{\boldsymbol{w_i}\}$ and our goal vector $\boldsymbol{v}$, we are asked to find a vector $\hat{\boldsymbol{v}} \in span\{\boldsymbol{w_i}\}$ such that the error $\|\hat{v} - \boldsymbol{v}\|^2$ is minimized.*

According to projection theorem, the best vector $\hat{v} = \sum_i s_i \boldsymbol{w_i}$ should be the orthogonal projection of $\boldsymbol{v}$, which suggests the **normal equation** holds

$$\langle \hat{\boldsymbol{v}} - \boldsymbol{v} | \boldsymbol{w_i} \rangle = 0 \Rightarrow \sum_i s_i \langle \boldsymbol{w_i} | \boldsymbol{w_j} \rangle = \langle \boldsymbol{v} | \boldsymbol{w_j} \rangle . \tag{1}$$

We can define the **Grammian** as $G_{i,j} \equiv \langle \boldsymbol{w_j} | \boldsymbol{w_i} \rangle$ and $t_j \equiv \langle \boldsymbol{v} | \boldsymbol{w_j} \rangle$, then the normal equation can be written in matrix form

$$Gs = t. \tag{2}$$

Note that $G = A^T A$, where

$$A = \begin{pmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & ... \end{pmatrix} \tag{3}$$

Thus G is at least postive-semidefinite. As long as $\{\boldsymbol{w_i}\}$ are linearly independent, then $A\boldsymbol{v} \neq 0$ for any $\boldsymbol{v} \in V$, which implies $\|A\boldsymbol{v}\|^2 = \boldsymbol{v}^T G \boldsymbol{v} \geq 0$, and thus G is positive-definite and invertible as a result.

Theorem 2.1. *For best approximation problem, the best estimate $\hat{\boldsymbol{v}}$ is given by*

$$\hat{\boldsymbol{v}} = \sum_i (G^{-1}t)_i \boldsymbol{w}_i, \tag{4}$$

with

$$G_{i,j} \equiv \langle \boldsymbol{w_j} | \boldsymbol{w_i} \rangle , \quad t_j \equiv \langle \boldsymbol{v} | \boldsymbol{w}_j \rangle . \tag{5}$$

Proof 2.1. *The proof is simple from what we discussed previously. The key is to use projection theorem and G is invertible if $\{\boldsymbol{w_j}\}$ are linearly independent.*

2.2 Minimum norm problem

Minimum norm problem is called the **dual problem** of best approximation problem.

Definition 2.2. *We call the following a minimum norm problem. Given a bunch of linearly independent vectors $\{\boldsymbol{w_j}\}$ and inner product $\langle \boldsymbol{v} | \boldsymbol{w_j} \rangle = t_j$, we are asked to find $\boldsymbol{v} \in V$ such that $\langle \boldsymbol{v} | \boldsymbol{w_j} \rangle = t_j$ with the minimum norm $\|\boldsymbol{v}\|^2$.*

Let $U = \text{span}\{\boldsymbol{w_j}\}$, then $\boldsymbol{v} = \sum_i s_i \boldsymbol{w_j} + \boldsymbol{u}^\perp$, where $\boldsymbol{u}^\perp \in U^\perp$. We see that $\boldsymbol{u}^\perp$ has no contribution to the inner product, and thus should be zero if we want to minimize the norm of $\boldsymbol{v}$. So we can write down the equation that is very similar to normal equation

$$\langle \boldsymbol{v} | \boldsymbol{w_j} \rangle = \sum_i s_i \langle \boldsymbol{w_i} | \boldsymbol{w_j} \rangle = t_j. \tag{6}$$

By defining gramming we can also rewrite it in matrix form:

$$Gs = t. \tag{7}$$

The close relation between best approximation and minimum norm problem can be explained in a naive way. Consider the minimum norm problem is to find a vector lying in the hypter-plane $\langle \boldsymbol{v} | \boldsymbol{w}_j \rangle = t_j$ that is the closest to original point (has the shortest norm). We can always choose some vector $\boldsymbol{a}$ and move the whole space $V + \boldsymbol{a}$ such that hyper-plane after displacement $HP + \boldsymbol{a} = \{\boldsymbol{v} + \boldsymbol{a} | \langle \boldsymbol{v} | \boldsymbol{w}_j \rangle = t_j\}$ includes the original point, and the problem is equivalent to find $\boldsymbol{v}$ in this new hypter-plane such that it is closest to $\boldsymbol{a}$. We can instantly see that the closest vector to $\boldsymbol{a}$ is simply the orthogonal projection of $\boldsymbol{a}$ onto the new hyper-plane, so the minimum norm problem is translated into a best approximation problem.

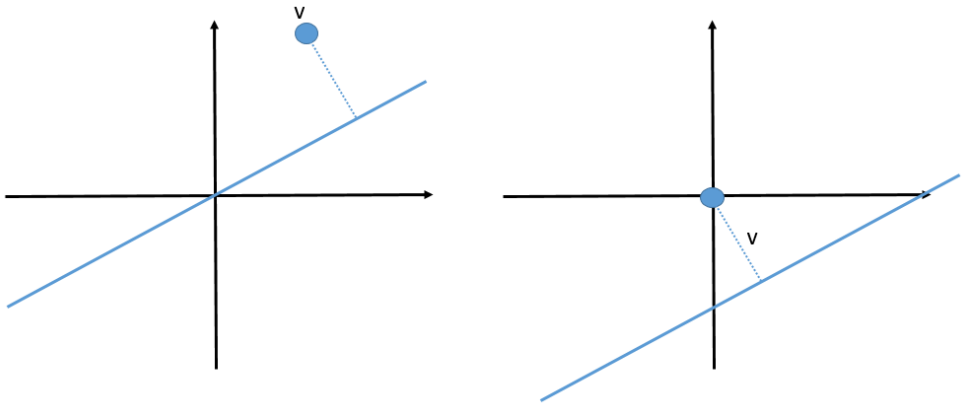

Figure 1: Left: projection $\boldsymbol{v}$ onto hyper-plane; right: find the closest vector on hyper-plane to original point.

3 Solving $Ax = b$

Now we are going to use what we knew to "solve" matrix equation $Ax = b$. To be specific, let $A \in F^{n,m}$, $x \in F^m$ and $b \in F^n$. There are several circumstances.

(1) **Determined equation: unique solution**

If we want a unique solution, we require: $b \in R(A)$ and column vectors of A are linearly independent. Then we can solve a unique solution x.

(2) **Overdetermined equation: No solution**

If $b \notin R(A)$, then we simply do not have solution. We say this is overdetermined, because if $n > m$, then it is very likely that $b \notin R(A)$ and thus we do not have solution due to too many equations but too less unknown variables. In the case that column vectors of A are linearly independent and $b \notin R(A)$, we can use best approximation method to get a "best" solution. Note that though best approximation can be applied in case that solution exists, we simply get a random solution depending on the linearly independent vectors we choose.

(3) **Underdetermined equation: multiple solutions**

If $b \in R(A)$ and column vectors of A are linearly dependent, then we have multiple solutions. If we want to find the solution with the minimum norm, then we get a minimum norm problem, by letting $t = b$, $\boldsymbol{v} = x$ and rows of A being $\boldsymbol{w_j}$. Note that minimum norm method is only applied when solution exists, or it is meanless to find a solution with minimized norm.

Conclusion: To solve or approximately solve $Ax = b$,

- If $b \in R(A)$ and column vectors of A are linearly independent, then we can find a unique solution. Actually in this case we can always omit linear dependent rows and get an invertible square matrix, and apply the inverse of the matrix to get the unique solution. Of course best approximation and minimum norm method can give the right answer as long as we omit the linear dependent columns/rows.
- If $b \in R(A)$ and column vectors of A are linearly dependent, then we delete the linear dependent rows (they represent the same equations) and then apply minimum norm method to find the "best" solution with the minimized norm. Though we can use best approximation, but it will give random solution instead of the minimum-norm solution.
- If $b \notin R(A)$ and column vectors of A are linearly independent, then apply best approximation to get the "best" solution with the minimized error.
- If $b \notin R(A)$ and column vectors of A are linearly dependent, then we omit the linear dependent columns and go back to best approximation problem.

4 Four fundamental subspaces

Four fundamental subspaces of a linear transform $A \in L(V, W)$, are range space of A, null space of $N(A)$, range space of $R(A^\dagger)$ and null space of $N(A^\dagger)$. Have a picture of these four fundamental subspaces is the key to understand linear transform.

Theorem 4.1 (Four fundamental subspaces)**.** *Let $A \in L(V, W)$ be a linear transform. Then $R(A)$, $N(A)$, $R(A^\dagger)$ and $N(A^\dagger)$ are called **four fundamental subspaces** with the following properties:*

(1) $R(A^\dagger) = N(A)^\perp$

(2) $R(A) = N(A^\dagger)^\perp$

(3) $\dim\{R(A)\} = \dim\{R(A^\dagger)\}$

Proof 4.1. *To prove (1), let $\boldsymbol{v} \in N(A)$, then for any $A^\dagger \boldsymbol{w} \in R(A^\dagger)$, we have $\langle A^\dagger \boldsymbol{w} | v \rangle = \langle \boldsymbol{w} | A\boldsymbol{v} \rangle = 0$, so $v \in R(A^\dagger)$, and thus $R(A^\dagger) = N(A)^\perp$.*

To prove (2), we simply use the similar argument in (1) but substitute A with $A^\dagger$.

To prove (3), we use the fundamental theorem $\dim\{R(A)\} + \dim\{N(A)\} = \dim\{V\}$*, so we get* $\dim\{R(V)\} = \dim\{R(A^\dagger)\}$.

The theorem basically says we can decompose V and W into two orthogonal complements: $V = R(A^\dagger) \oplus N(A)$ and $W = R(A) \oplus N(A^\dagger)$. Using this picture we can easily understand the inverse of the linear transform exists only if $N(A) = N(A^\dagger) = \{\mathbf{0}\}$. We are also going to see that Singular values decomposition (SVD) is simply finding basis of these four fundamental subspaces and factorize A using these basis. And the pseudo-inverse of transform is also based on these four fundamental subspaces.

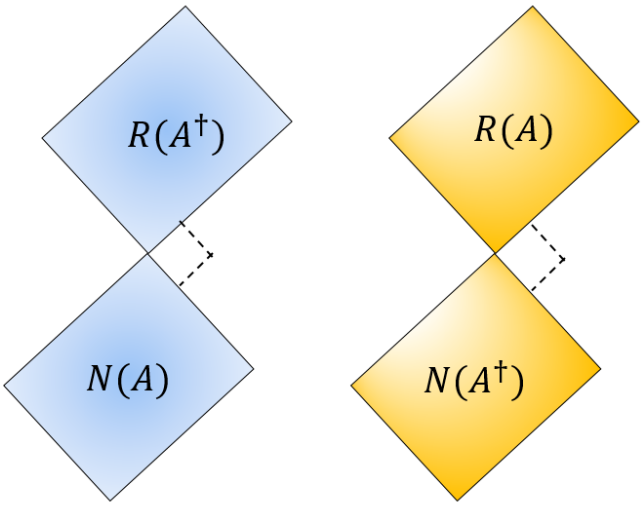

Figure 2: Four fundamental subspaces

5 SVD and pseudo-inverse

SVD is kind of a generalization of eigenvectors decomposition: for any linear map $A \in L(V, U)$, though we cannot talk about its eigenvectors in general (only operators have eigenvectors), we can always find orthonormal basis $\{\boldsymbol{v_i}\}$ of V and orthonormal basis $\{\boldsymbol{w_i}\}$ of W such that A is diagonal in these two basis:

$$A\boldsymbol{v_i} = \sigma_i \boldsymbol{w_i}, \tag{8}$$

where σ_i are called singular values. Then we can say A in terms of SVD (singular values decomposition) is

$$A = \sum_i \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^\dagger, \tag{9}$$

where $\boldsymbol{v_i}^\dagger$ is defined as the linear functional $\boldsymbol{v_i}^\dagger(\boldsymbol{v}) = \langle \boldsymbol{v} | \boldsymbol{v_i} \rangle$.

5.1 SVD construction

To construct SVD, the key is to first construct orthogonal basis $\{\boldsymbol{v_i}\}$ and $\{\boldsymbol{w_i}\}$. Let us consider eigenvectors of $A^\dagger A$:

$$A^\dagger A \boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v}_i. \tag{10}$$

Note that $A^\dagger A$ is a positive-semidefinite operator, so we can know that its eigenvectors $\{\boldsymbol{v_i}\}$ are orthonormal basis of V. We can classify $\{\boldsymbol{v_i}\}$ into two classes: one with positive σ_i^2 and others with $\sigma_i^2 = 0$. We are going to see that actually $\text{span}\{\boldsymbol{v_i} | \sigma_i^2 = 0\} = N(A)$ and $\text{span}\{\boldsymbol{v_i} | \sigma_i^2 > 0\} = R(A^\dagger)$.

Lemma 5.1. *Let $A \in L(V, W)$ and $A^\dagger A \boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v_i}$. Then span$\{\boldsymbol{v_i} | \sigma_i^2 = 0\} = N(A)$ and span$\{\boldsymbol{v_i} | \sigma_i^2 > 0\} = R(A^\dagger)$.*

Proof 5.1. *We first prove $\{\boldsymbol{v_i} | \sigma_i^2 = 0\} = N(A)$. Let $A^\dagger A \boldsymbol{v_i} = 0$, and if we assmue $A\boldsymbol{v_i} \neq 0$, then $A\boldsymbol{v_i} \in R(A)$, and thus $A\{v_i\} \notin N(A^\dagger)$ (from fact that $R(A) \oplus N(A^\dagger) = W$). So $A^\dagger A \boldsymbol{v}_i \notin 0$ since $A\boldsymbol{v}_i$ is not in $N(A^\dagger)$, which is contradictory to our assumption. So $A\boldsymbol{v_i} = 0$ and $\boldsymbol{v_i}$ with $\sigma_i^2 = 0$ is in $N(A)$. Let $\boldsymbol{v_i} \in N(A)$, then $A^\dagger A \boldsymbol{v_i} = 0$ by definition. Therefore span$\{\boldsymbol{v_i} | \sigma^2 = 0\} = N(A)$. Also note that span$\{\boldsymbol{v_i}\} = W$, so span$\{\boldsymbol{v_i} | \sigma^2 > 0\} = R(A^\dagger)$.*

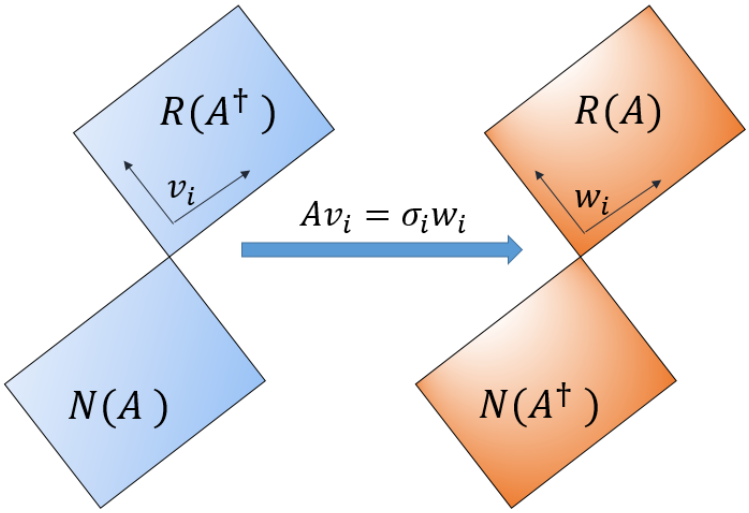

Figure 3: SVD: find two orthonormal basis of V and W such that $A\boldsymbol{v_i} = \sigma_i \boldsymbol{w}_i$, namely A is diagonal in these basis.

So now we see that eigenbasis of $A^\dagger A$ naturally form basis of $R(A^\dagger)$ and $N(A)$. Now the next step to construct the basis of space W.

Lemma 5.2. *Let $\boldsymbol{w_i} \equiv \frac{1}{\sigma_i} A\boldsymbol{v}_i$ for $\sigma_i^2 > 0$. Then $\{\boldsymbol{w_i}\}$ are orthogonal basis of $R(A)$. And it turns out $\boldsymbol{w_i}$ are eigenbasis of $AA^\dagger$, with $AA^\dagger \boldsymbol{w_i} = \sigma_i^2 \boldsymbol{w_i}$.*

Proof 5.2. *Note that*

$$\langle \boldsymbol{w_i} | \boldsymbol{w_j} \rangle = \frac{1}{\sigma_i \sigma_j} \langle A\boldsymbol{v_i} | A\boldsymbol{v_j} \rangle = \frac{1}{\sigma_i \sigma_j} \left\langle \boldsymbol{v_i} \middle| A^\dagger A \boldsymbol{v_j} \right\rangle = \delta_{i,j}.$$

So $\{\boldsymbol{w_i}\}$ are indeed orthonormal basis. Since for every $\boldsymbol{w} \in R(A)$, we can find a $\boldsymbol{v} \in R(A^\dagger)$ such that $A\boldsymbol{v} = \boldsymbol{w}$. Thus $\{\boldsymbol{w_i}\}$ are basis of $R(A)$.

Also note that

$$AA^\dagger \boldsymbol{w_i} = \frac{1}{\sigma_i} AA^\dagger A\boldsymbol{v_i} = \sigma_i A\boldsymbol{v_i} = \sigma_i^2 \boldsymbol{w_i}.$$

So simply from eigenbasis $\boldsymbol{v_i}$ of $A^\dagger A$ with eigenvalues $\sigma^2 > 0$, we can immediately construct k (dimension of $R(A^\dagger)$ or $R(A)$, or rank of A) orthonormal basis of $R(A)$. The other eigenvectors of $AA^\dagger$ will be basis of $N(A^\dagger)$, which we do not concern.

Up to now we have successfully construct orthonormal basis of $R(A^\dagger)$ and $R(A)$. Then we can find our linear transform is diagonal in these two basis by definition: $A\boldsymbol{v_i} = \sigma_i \boldsymbol{w_i}$ for $\sigma_i > 0$. For any vector $\boldsymbol{v} \in V$, we then have

$$A\boldsymbol{v} = \sum_i \langle \boldsymbol{v} | \boldsymbol{v_i} \rangle A\boldsymbol{v_i} = \sum_i \langle \boldsymbol{v} | \boldsymbol{v_i} \rangle \sigma_i \boldsymbol{w_i} = \sum_{i|\sigma_i > 0} \langle \boldsymbol{v} | \boldsymbol{v_i} \rangle \sigma_i \boldsymbol{w_i}. \tag{11}$$

Of course you can either sum all i or simply i such that $\sigma_i > 0$. Both of them are the same, since term $\sigma_i = 0$ does not contribute. This gives SVD decomposition of A.

Theorem 5.1. *Let $A \in L(V, W)$. Let $A^\dagger A\boldsymbol{v_i} = \sigma_i^2 \boldsymbol{v_i}$ and $AA^\dagger \boldsymbol{w}_i = \sigma_i^2 \boldsymbol{w_i}$. We can factorize A by so-called **Compact SVD**:*

$$A = \sum_{i|\sigma_i > 0} \sigma_i \boldsymbol{w}_i \boldsymbol{v}_i^\dagger, \tag{12}$$

or in matrix form

$$A = \sum_{i|\sigma_i>0} \sigma_i \boldsymbol{w_i} \boldsymbol{v}_i^T = W\Sigma V^T. \tag{13}$$

*Also we can do the **Full SVD**:*

$$A \sum_i \sigma_i \boldsymbol{w}_i \boldsymbol{v}_i^\dagger, \tag{14}$$

or in matrix form

$$A = \sum_i \sigma_i \boldsymbol{w_i} \boldsymbol{v}_i^T = \begin{pmatrix} W & W_0 \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V & V_0 \end{pmatrix}^T. \tag{15}$$

Proof 5.3. *It is automatically proved based on our lemma.*

To summarize, we first look at eigenbasis of $A^\dagger A$ and $AA^\dagger$, and use these basis turns out to "diagonalize" A in a wide sense. Essentially it is because eigenbasis of $A^\dagger A$ and $AA^\dagger$ with positive eigenvalues span $R(A^\dagger)$ and $R(A)$ respectively.

5.2 Peusdo-inverse

Having a picture of SVD, it is natural to extend concepts of inverse to an arbitrary linear transform. We knew that a linear transform $T \in L(V, W)$ is invertible if and only if $N(T) = \{0\}$ and $\dim\{V\} = \dim\{W\}$. For a general T with $\dim\{V\} \neq \dim\{W\}$, namely $M(T)$ being a non-square matrix, we cannot find its inverse.

From the picture of SVD, though inverse cannot be find since $N(T)$ and $N(T^\dagger)$ are not equal to $\{0\}$ in general, we can do our "best" to approximate to its inverse. A linear transform is characterize by $T\boldsymbol{v_i} = \sigma_i \boldsymbol{w_i}$, and we define the pseudo inverse of T as

$$T^{-1}\boldsymbol{w_i} = \frac{1}{\sigma_i}\boldsymbol{v_i}. \tag{16}$$

Note that we cannot perfectly find an inverse since: (1) There are some $\sigma_i = 0$, so we lose this part of information when applying T; (2) Not every vector $\boldsymbol{w} \in W$ can be expressed in $\{\boldsymbol{w_i}\}$. Our best choice is

$$T^{-1}\boldsymbol{w_i} = 0 \tag{17}$$

for those $\boldsymbol{w_i}$ in $N(T^\dagger)$.

Definition 5.1. *Let $T \in L(V, W)$ and $T = \sum_{i||\sigma_i>0} \sigma_i \boldsymbol{w_i} \boldsymbol{v_i}^\dagger$ be SVD of T. Then the **pseudo inverse** of T is defined by*

$$T^{-1} \equiv \sum_{i|\sigma_i>0} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{w}_i^\dagger. \tag{18}$$

And if $N(T) = N(T^\dagger) = \{0\}$, namely T is invertible, then T^{-1} is exactly the inverse of T.

5.3 $Ax = b$ Revisit

We are going to see either best approximation or minimum norm method is kind of pseudo inverse. Consider the full SVD:

$$A = \sum_i \sigma_i \boldsymbol{w_i} \boldsymbol{v}_i^T = \begin{pmatrix} W_1 & W_0 \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1 & V_0 \end{pmatrix}^T = W \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T. \tag{19}$$

Then the pseudo inverse can be represented as

$$A^{-1} = \sum_i \frac{1}{\sigma_i} \boldsymbol{v_i} \boldsymbol{w}_i^T = \begin{pmatrix} V_1 & V_0 \end{pmatrix} \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} W_1 & W_0 \end{pmatrix}^T = V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} W^T. \tag{20}$$

In best approximation problem, in order to apply normal equation, we usually require that A is full rank but $\dim\{W\} > \dim\{V\}$, so in this case A has the following form:

$$A_{\text{best approx}} = W \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T \tag{21}$$

Then if we look at the solution of normal equation

$$\begin{aligned} A^{-1}_{\text{best approx}} = (A^T A)^{-1} A^T &= \left[V \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}^T W^T W \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T \right]^{-1} V \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}^T W^T \\ &= V \Sigma^{-2} \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}^T W^T = V \begin{pmatrix} \Sigma^{-1} & 0 \end{pmatrix} W^T = A^{-1}, \end{aligned} \tag{22}$$

we find it has the exactly the same form as pseudo inverse. So best approx is simply a special form of pseudo inverse. We can do the same for minimum norm method, which usually has the following form:

$$A_{\text{min norm}} = W \begin{pmatrix} \Sigma & 0 \end{pmatrix} V^T \tag{23}$$

If we use normal equation, we will find $A^T A$ is not invertible, instead we use

$$A^{-1}_{\text{min norm}} = A^T (A A^T)^{-1} = V A_{\text{best approx}} = W \begin{pmatrix} \Sigma^{-}1 \\ 0 \end{pmatrix} W^T, \tag{24}$$

which is also the same as pseudo inverse. In general, the pseudo inverse is more universal than best approximation and minimum norm method.

6 Conclusion

We start with the best approximation and minimum norm problem, and deduce the normal equation. By introducing SVD, we find that both of best approximation and minimum norm method are special form of pseudo inverse.