

Advanced Methods Homework 1

Bohao Tang

September 11, 2017

1 Vector spaces and inner products

1. *Proof.* Denote $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k\}$ be a basis of W . And we can expand them to a basis of \mathbb{R}^n as $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k, \mathbf{a}_{k+1}, \dots, \mathbf{a}_n\}$. Then we do Gram-Schmidt orthogonalization to the basis and get $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$. It's easy to see that $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k\}$ become an orthogonal basis of W (they are orthogonal, hence linear independent, and k is the dimension of W). Now we assert that $V = \text{span}\{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ will be W^\perp .

First, for every vector \mathbf{x} in W^\perp , it has a representation $\mathbf{x} = \sum_{i=1}^n \omega_i \mathbf{b}_i$ for some $\omega_i \in \mathbb{R}^n$. Since $\mathbf{x} \in W^\perp$, we have $\langle \mathbf{x}, \mathbf{b}_j \rangle = 0$, for all $j = 1, 2, \dots, k$. Therefore $\omega_j = 0$ for all $1 \leq j \leq k$, which means $\mathbf{x} \in V$ and $W^\perp \subset V$.

In the other side, for every vector $\mathbf{y} = \sum_{t=k+1}^n \gamma_t \mathbf{b}_t \in V$, it's quite direct to see that $\langle \mathbf{y}, \mathbf{b}_j \rangle = 0$, for all $j = 1, 2, \dots, k$, hence $\langle \mathbf{y}, \mathbf{w} \rangle = 0$, for all $\mathbf{w} \in W$. Therefore $\mathbf{y} \in W^\perp$ and $W^\perp \subset V$.

Now we get $W^\perp = V$, hence $\dim(W^\perp) = n - k$. Also we can see that W^\perp is unique. If not, we can merge the two different W^\perp and get a higher dimensional subspace Z that is orthogonal to W . Then $Z \oplus W \subset \mathbb{R}^n$ but $\dim(Z \oplus W) \geq n + 1 > n$, a contradiction. \square

2. (a) *Proof.* According to Cauchy inequality, $(\sum u_i v_i)^2 \leq \sum u_i^2 \sum v_i^2$. It's equivalent here that $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$. \square

(b) *Proof.*

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v})'(\mathbf{u} + \mathbf{v}) + (\mathbf{u} - \mathbf{v})'(\mathbf{u} - \mathbf{v}) \quad (1)$$

$$= \mathbf{u}'\mathbf{u} + 2\mathbf{u}'\mathbf{v} + \mathbf{v}'\mathbf{v} + \mathbf{u}'\mathbf{u} - 2\mathbf{u}'\mathbf{v} + \mathbf{v}'\mathbf{v} \quad (2)$$

$$= 2\mathbf{u}'\mathbf{u} + 2\mathbf{v}'\mathbf{v} = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 \quad (3)$$

\square

3. In the definition of projection in Lecture 2, $\mathbf{y} - \Pi(\mathbf{y}|\mathbf{x})$ should be orthogonal to \mathbf{x} . So there is nothing need to be proved. Maybe we should changed the definition here that (suppose $\mathbf{x} \neq \mathbf{0}$, or there is no meaningful projection)

$$\Pi(\mathbf{y}|\mathbf{x}) \triangleq \underset{\mathbf{u} \in \text{span}\{\mathbf{x}\}}{\text{argmin}} \|\mathbf{y} - \mathbf{u}\|^2$$

Proof. (This proof contains the part to assert that projection above is well defined) Suppose $\mathbf{u} = b\mathbf{x}$, then $\|\mathbf{y} - \mathbf{u}\|^2 = \mathbf{y}'\mathbf{y} - 2b\mathbf{x}'\mathbf{y} + b^2\mathbf{x}'\mathbf{x}$. It's a quadratic function

with highest coefficient $\mathbf{x}'\mathbf{x} > 0$, therefore have a unique minimizer. Differentiate it with b , let it be zero and we get the normal equation $\mathbf{x}'\mathbf{y} = b\mathbf{x}'\mathbf{x}$, which means (since $\Pi(\mathbf{y}|\mathbf{x}) = b\mathbf{x}$) : $\langle \Pi(\mathbf{y}|\mathbf{x}), \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ and therefore $\mathbf{y} - \Pi(\mathbf{y}|\mathbf{x}) \perp \mathbf{x}$. \square

2 Regression

1. Slope $\hat{\beta}_{yx}$ of regressing \mathbf{y} on \mathbf{x} is $\hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$, where $\hat{\rho}_{xy}$ is the sample correlation coefficient and $\hat{\sigma}_y, \hat{\sigma}_x$ are the sample standard deviation of y and x . So, slope $\hat{\beta}_{xy}$ of \mathbf{x} on \mathbf{y} is $\hat{\rho}_{yx} \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$. We have that $\hat{\rho}_{xy} = \hat{\rho}_{yx}$, so $\hat{\beta}_{xy}\hat{\beta}_{yx} = \hat{\rho}_{xy}^2$.
2. *Proof.* In the setting of mean only regression of \mathbf{y} , we have $\hat{\mu} = \bar{\mathbf{y}} = \sum_{i=1}^n y_i/n$. And the residual is $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}\mathbf{J}_n$, therefore sum of residual is $\sum r_i = \sum_1^n y_i - n \cdot \sum_1^n y_i/n = 0$. \square
3. *Proof.* In the setting of regression through origin of \mathbf{y} on \mathbf{x} , we have $\hat{\beta} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$. Therefore residual is $\mathbf{r} = \mathbf{y} - \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \mathbf{x}$, therefore $\langle \mathbf{r}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle - \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \cdot \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Rightarrow \mathbf{r} \perp \mathbf{x}$. \square
4. Notation as above, when $\mathbf{y} \perp \mathbf{x}$, we find that the residual $\mathbf{r} = \mathbf{y}$, which of course need not sum to zero.
5. *Proof.* In this setting of regressing \mathbf{y} on \mathbf{x} , we get the normal equation that:

$$\mathbf{J}_n' \mathbf{y} = \hat{\beta}_0 \mathbf{J}_n' \mathbf{J}_n + \hat{\beta}_1 \mathbf{J}_n' \mathbf{x} \quad (4)$$

$$\mathbf{x}' \mathbf{y} = \hat{\beta}_0 \mathbf{x}' \mathbf{J}_n + \hat{\beta}_1 \mathbf{x}' \mathbf{x} \quad (5)$$

also the residual $\mathbf{r} = \mathbf{y} - \hat{\beta}_0 \mathbf{J}_n - \hat{\beta}_1 \mathbf{x}$. Then:

$$\langle \mathbf{r}, \mathbf{J}_n \rangle = \mathbf{J}_n' \mathbf{y} - \hat{\beta}_0 \mathbf{J}_n' \mathbf{J}_n - \hat{\beta}_1 \mathbf{J}_n' \mathbf{x} = 0$$

and

$$\langle \mathbf{r}, \mathbf{x} \rangle = \mathbf{x}' \mathbf{y} - \hat{\beta}_0 \mathbf{x}' \mathbf{J}_n - \hat{\beta}_1 \mathbf{x}' \mathbf{x} = 0$$

So we get the residual $\mathbf{r} \perp \mathbf{J}_n$ and $\mathbf{r} \perp \mathbf{x}$. \square

3 Least squares

1. *Proof.* Given $\mathbf{H}^2 = \mathbf{H}$, then $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I}^2 - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}$. \square
2. After calculating, I think the method mentioned to estimate β_2 in this question is:
 - First, regressing \mathbf{y} on \mathbf{x}_1 through origin and get the residual \mathbf{r}_1 .
 - Second, regressing \mathbf{x}_2 on \mathbf{x}_1 through origin and get the residual \mathbf{r}_2 .
 - Finally, regressing \mathbf{r}_1 on \mathbf{r}_2 through origin and get the estimation $\hat{\beta}_2$.

Proof. To minimize $\|\mathbf{y} - \beta_1 \mathbf{x}_1 - \beta_2 \mathbf{x}_2\|^2$, we can first suppose β_2 be a constant and minimize through β_1 (it will be a function of β_2), and then minimize through β_2 to get the estimation. So the goal is firstly to find the $\text{argmin}_{\beta_1} \|(\mathbf{y} - \beta_2 \mathbf{x}_2) - \beta_1 \mathbf{x}_1\|^2$. This is simply a regression through origin and we can get that $\beta_1^{\min} = \frac{\langle \mathbf{y} - \beta_2 \mathbf{x}_2, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle}$.

Therefore, the second step is to find the argmin of:

$$\|\mathbf{y} - \beta_2 \mathbf{x}_2 - \frac{\langle \mathbf{y} - \beta_2 \mathbf{x}_2, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle} \mathbf{x}_1\|^2 = \|(\mathbf{y} - \frac{\langle \mathbf{y}, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle} \mathbf{x}_1) - \beta_2 (\mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle} \mathbf{x}_1)\|^2$$

Also, we can find that:

$$\mathbf{r}_1 = \mathbf{y} - \frac{\langle \mathbf{y}, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle} \mathbf{x}_1 \quad (6)$$

$$\mathbf{r}_2 = \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{x}_1 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle} \mathbf{x}_1 \quad (7)$$

Therefore, $\hat{\beta}_2$ comes from minimizing $\|\mathbf{r}_1 - \beta_2 \mathbf{r}_2\|^2$, which is just doing regression through origin of \mathbf{r}_1 on \mathbf{r}_2 . Then we finish the proof that the procedure described in the question is valid. \square

3. *Proof.* Suppose that \mathbf{X} is of size $m \times n$, where $m \leq n$ (or we replace \mathbf{X} by \mathbf{X}' , and we just need to prove the same thing). Then consider the singular value decomposition of \mathbf{X} is $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}$, where \mathbf{U}, \mathbf{V} are orthogonal matrices and $\mathbf{S} = (\text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_r, 0, \dots, 0\}, \mathbf{0}) : \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_r > 0$. Here $r = \text{rank}(\mathbf{X})$. Now consider the svd of $\mathbf{X}', \mathbf{X}'\mathbf{X}, \mathbf{X}\mathbf{X}'$, we have:

$$\mathbf{X}' = \mathbf{V}'\mathbf{S}'\mathbf{U}' \quad (8)$$

$$\mathbf{X}'\mathbf{X} = \mathbf{V}' \begin{pmatrix} \text{diag}\{\lambda_1^2, \lambda_2^2, \dots, \lambda_r^2, 0, \dots, 0\} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V} \quad (9)$$

$$\mathbf{X}\mathbf{X}' = \mathbf{U} \text{diag}\{\lambda_1^2, \lambda_2^2, \dots, \lambda_r^2, 0, \dots, 0\} \mathbf{U}' \quad (10)$$

Therefore we can see that the three matrices above all have and only have r nonzero singular values, which means they are all of rank r . Then we get $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}') = \text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}\mathbf{X}')$. \square

4. If the design matrix is orthogonal, then we have:

$$\mathbf{J} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\beta}$$

And then we get the normal equation as:

$$\frac{\partial \mathbf{J}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = 0$$

So we have $\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$.

4 Computing and analysis

Code is shown below and we simulate the data ourselves and test the L_1 norm of the differences of $\hat{\beta}$, fitted_value and residuals from two results.

1.

```
mylm <- function(y,X) {  
  D = cbind(1,X)  
  beta = solve(t(D) %*% D) %*% t(D) %*% y  
  fitted = D %*% beta  
  residuals = y - fitted  
  result = list(beta = beta,  
                fitted = fitted,  
                residuals = residuals)  
  return(result)  
}
```

2.

```
test.X = cbind(sample(1:100),sample(1:100),sample(1:100))  
beta = c(5,-1,4,2)  
test.y = cbind(1,test.X) %*% beta + rnorm(100)  
model = lm(test.y~test.X)  
mymodel = mylm(test.y,test.X)  
c(sum(abs(model$coefficients - mymodel$beta)),  
  sum(abs(model$fitted.values - mymodel$fitted)),  
  sum(abs(model$residuals - mymodel$residuals))  
)
```

```
## [1] 1.096900e-13 2.499334e-11 2.488536e-11
```