Math of Big Data, Summer 2018

Prof: Gu

Name: **Forest Kobayashi**

HW #: **1**

Day: **Mon.** **Tue.** **Wed.** **Thu.** **Fri.**

Date: **05/15/2018**

No.	**Points**	**Acknowledgments**
1		Tim Player, Jacky Lee
2		
Total		

This Assignment is (check one):

☒ On Time ☐ Late, without deduction ☐ Late, with deduction

Comments: Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

The starter code for problem 2 part c and d can be found under the Resource tab on course website.

Note: You need to create a Github account for submission of the coding part of the homework. Please create a repository on Github to hold all your code and include your Github account username as part of the answer to problem 2.

Problem 1. (Linear Transformation)

Let $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ be a random vector. Show that expectation is linear:

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x} + \mathbf{b}] = A\mathbb{E}[\mathbf{x}] + \mathbf{b}.$$

Also show that

$$\operatorname{cov}[\mathbf{y}] = \operatorname{cov}[A\mathbf{x} + \mathbf{b}] = A \operatorname{cov}[\mathbf{x}]A^\top = A\mathbf{\Sigma}A^\top.$$

Solution:

(a) First, we prove some small lemmas.

Lemma 1.1. *Let X be a continuous random variable, and a be a scalar. Then $\mathbb{E}[aX] = a\mathbb{E}[X]$.*

Proof. Let X admit a density function $f(x)$. Then

$$\begin{aligned}\mathbb{E}[aX] &= \int_{-\infty}^{\infty} axf(x)\,\mathrm{d}x \\ &= a\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \\ &= a\mathbb{E}[X]\end{aligned}$$

■

Lemma 1.2. *Let X be a random variable, and let a be a scalar. Then $\mathbb{E}[X + a] = \mathbb{E}[X] + \mathbb{E}[a] = \mathbb{E}[X] + a$.*

Proof. Let X admit a density function $f(x)$. Then over $\mathbb{R}$, $f(x)$ must integrate to 1. Hence

$$\begin{aligned}\int_{-\infty}^{\infty} (x+a)f(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} xf(x) + af(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} af(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x + a\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= \mathbb{E}[X] + a \cdot 1 \\ &= \mathbb{E}[X] + a\end{aligned}$$

■

Lemma 1.3. *Let X and Y be continuous random variables. Then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

Proof. Let $X + Y$ have a density function $f(x, y)$. Then

$$\begin{aligned}\mathbb{E}[X + Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

■

Now we do the main problem. By the lemmas, we have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x} + \mathbf{b}]$$

$$= \mathbb{E}[A\mathbf{x}] + \mathbb{E}[\mathbf{b}]$$
$$= \mathbb{E}[A\mathbf{x}] + \mathbf{b}$$

It remains to show that $\mathbb{E}[A\mathbf{x}] = A\mathbb{E}[\mathbf{x}]$. Suppose $\mathbf{x}$ is an n-dimensional random vector in a space X:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

where each of the $x_i \in \mathbb{R}$. By definition, the expectation value of $\mathbf{x}$ is given by

$$\mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}[x_0] \\ \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_n] \end{bmatrix}$$

Let $A \in M_{m\times n}(\mathbb{R})$, with rows $\mathbf{a}_0$, $\mathbf{a}_1$, …, $\mathbf{a}_m$. Then $A\mathbb{E}[\mathbf{x}]$ is given by

$$A\mathbb{E}[\mathbf{x}] = \begin{bmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,n} \\ a_{1,0} & a_{1,1} & \cdots & a_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,0} & a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} \mathbb{E}[x_0] \\ \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_n] \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{i=0}^{n} a_{0,i}\mathbb{E}[x_i] \\ \sum_{i=0}^{n} a_{1,i}\mathbb{E}[x_i] \\ \vdots \\ \sum_{i=0}^{n} a_{n,i}\mathbb{E}[x_i] \end{bmatrix}$$

Now, we manipulate this expression to obtain $\mathbb{E}[A\mathbf{x}]$. Because the expected value of a sum is the sum of expected values (see lemmas), we have

$$\begin{bmatrix} \sum_{i=0}^{n} a_{0,i}\mathbb{E}[x_i] \\ \sum_{i=0}^{n} a_{1,i}\mathbb{E}[x_i] \\ \vdots \\ \sum_{i=0}^{n} a_{n,i}\mathbb{E}[x_i] \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\sum_{i=0}^{n} a_{0,i}x_i\right] \\ \mathbb{E}\left[\sum_{i=0}^{n} a_{1,i}x_i\right] \\ \vdots \\ \mathbb{E}\left[\sum_{i=0}^{n} a_{n,i}x_i\right] \end{bmatrix}$$

and now because the expected value of a vector is a vector of expected values,

$$\begin{bmatrix} \mathbb{E}\left[\sum_{i=0}^{n} a_{0,i}x_i\right] \\ \mathbb{E}\left[\sum_{i=0}^{n} a_{1,i}x_i\right] \\ \vdots \\ \mathbb{E}\left[\sum_{i=0}^{n} a_{n,i}x_i\right] \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} \sum_{i=0}^{n} a_{0,i}x_i \\ \sum_{i=0}^{n} a_{1,i}x_i \\ \vdots \\ \sum_{i=0}^{n} a_{m,i}x_i \end{bmatrix}\right]$$
$$= \mathbb{E}[A\mathbf{x}]$$

hence $A\mathbb{E}[\mathbf{x}] = \mathbb{E}[A\mathbf{x}]$, and so putting all the pieces together,

$$\boxed{\mathbb{E}[A\mathbf{x} + \mathbf{b}] = A\mathbb{E}[\mathbf{x}] + \mathbf{b}}$$

as desired. ■

(b) We have the following corollary to part (a):
Corollary 1.3.1. *Let A be a matrix, and let $\mathbf{x}$ be a random vector. Then $\mathbb{E}\left[\mathbf{x}A^\top\right] = \mathbb{E}[\mathbf{x}]A^\top$.*

Proof. Note that

$$\begin{aligned} \mathbb{E}\left[\mathbf{x}A^\top\right] &= \mathbb{E}\left[\left(\left(\mathbf{x}A^\top\right)^\top\right)^\top\right] \\ &= \mathbb{E}\left[\left(A\mathbf{x}^\top\right)^\top\right] \\ &= \left(\mathbb{E}\left[A\mathbf{x}^\top\right]\right)^\top \\ &= \left(A\mathbb{E}\left[\mathbf{x}^\top\right]\right)^\top \\ &= \left(\mathbb{E}\left[\mathbf{x}^\top\right]\right)^\top A^\top \\ &= \mathbb{E}[\mathbf{x}]A^\top \end{aligned}$$

■

Now, we proceed to the main proof. The covariance matrix $\operatorname{cov}[\mathbf{y}]$ is defined by

$$\operatorname{cov}[\mathbf{y}] = \begin{bmatrix} \mathbb{E}\Big[(y_0 - \mathbb{E}[y_0])(y_0 - \mathbb{E}[y_0])\Big] & \mathbb{E}\Big[(y_0 - \mathbb{E}[y_0])(y_1 - \mathbb{E}[y_1])\Big] & \cdots & \mathbb{E}\Big[(y_0 - \mathbb{E}[y_0])(y_n - \mathbb{E}[y_n])\Big] \\ \mathbb{E}\Big[(y_1 - \mathbb{E}[y_1])(y_0 - \mathbb{E}[y_0])\Big] & \mathbb{E}\Big[(y_1 - \mathbb{E}[y_1])(y_1 - \mathbb{E}[y_1])\Big] & \cdots & \mathbb{E}\Big[(y_1 - \mathbb{E}[y_1])(y_n - \mathbb{E}[y_n])\Big] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\Big[(y_n - \mathbb{E}[y_n])(y_0 - \mathbb{E}[y_0])\Big] & \mathbb{E}\Big[(y_n - \mathbb{E}[y_n])(y_1 - \mathbb{E}[y_1])\Big] & \cdots & \mathbb{E}\Big[(y_n - \mathbb{E}[y_n])(y_n - \mathbb{E}[y_n])\Big] \end{bmatrix}$$
$$= \begin{bmatrix} \operatorname{cov}(y_0, y_0) & \operatorname{cov}(y_0, y_1) & \cdots & \operatorname{cov}(y_0, y_n) \\ \operatorname{cov}(y_1, y_0) & \operatorname{cov}(y_1, y_1) & \cdots & \operatorname{cov}(y_1, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(y_n, y_0) & \operatorname{cov}(y_n, y_1) & \cdots & \operatorname{cov}(y_n, y_n) \end{bmatrix}$$

it is a property of covariance that for all random variables X and Y, with scalars a and b, $\operatorname{cov}(X + a, Y + b) = \operatorname{cov}(X, Y)$. Hence, since each $y_i = (A\mathbf{x})_i + b_i$, then $\forall i, j \in \{0, 1, \ldots, n\}$,

$$\operatorname{cov}[y_i, y_j] = \operatorname{cov}[(A\mathbf{x})_i + b_i, (A\mathbf{x})_j + b_j]$$

$$= \text{cov}\left[(A\mathbf{x})_i, (A\mathbf{x})_j\right]$$

thus, examining the matrix above, we see $\text{cov}\left[A\mathbf{x} + \mathbf{b}\right] = \text{cov}\left[A\mathbf{x}\right]$. It remains to show $\text{cov}\left[A\mathbf{x}\right] = A\,\text{cov}\left[\mathbf{x}\right]A^\top$.

Note that we can reexpress the covariance matrix by

$$\text{cov}\left[y\right] = \mathbb{E}\left[\begin{bmatrix} (y_0 - \mathbb{E}[y_0])(y_0 - \mathbb{E}[y_0]) & (y_0 - \mathbb{E}[y_0])(y_1 - \mathbb{E}[y_1]) & \cdots & (y_0 - \mathbb{E}[y_0])(y_n - \mathbb{E}[y_n]) \\ (y_1 - \mathbb{E}[y_1])(y_0 - \mathbb{E}[y_0]) & (y_1 - \mathbb{E}[y_1])(y_1 - \mathbb{E}[y_1]) & \cdots & (y_1 - \mathbb{E}[y_1])(y_n - \mathbb{E}[y_n]) \\ \vdots & \vdots & \ddots & \vdots \\ (y_n - \mathbb{E}[y_n])(y_0 - \mathbb{E}[y_0]) & (y_n - \mathbb{E}[y_n])(y_1 - \mathbb{E}[y_1]) & \cdots & (y_n - \mathbb{E}[y_n])(y_n - \mathbb{E}[y_n]) \end{bmatrix}\right]$$

$$= \mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top\right]$$

$$= \mathbb{E}\left[(A\mathbf{x} - \mathbb{E}[A\mathbf{x}])(A\mathbf{x} - \mathbb{E}[A\mathbf{x}])^\top\right]$$

by part (a),

$$= \mathbb{E}\left[(A\mathbf{x} - A\mathbb{E}[\mathbf{x}])(A\mathbf{x} - A\mathbb{E}[\mathbf{x}])^\top\right]$$

$$= \mathbb{E}\left[(A\mathbf{x} - A\mathbb{E}[\mathbf{x}])(\mathbf{x}^\top A^\top - \mathbb{E}[\mathbf{x}]^\top A^\top)\right]$$

$$= \mathbb{E}\left[A(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x}^\top - \mathbb{E}[\mathbf{x}]^\top)A^\top\right]$$

$$= \mathbb{E}\left[A(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top A^\top\right]$$

$$= A\mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top\right]A^\top$$

$$= A\,\text{cov}\left[\mathbf{x}\right]A^\top = A\mathbf{\Sigma}A^\top$$

hence, putting it all together, we have

$$\boxed{\text{cov}\left[A\mathbf{x} + \mathbf{b}\right] = A\,\text{cov}\left[\mathbf{x}\right]A^\top}$$

■

Problem 2.

Given the dataset $\mathcal{D} = \{(x, y)\} = \{(0, 1), (2, 3), (3, 6), (4, 8)\}$

(a) Find the least squares estimate $y = \boldsymbol{\theta}^\top \mathbf{x}$ by hand using Cramer's Rule.

(b) Use the normal equations to find the same solution and verify it is the same as part (a).

(c) Plot the data and the optimal linear fit you found.

(d) Find randomly generate 100 points near the line with white Gaussian noise and then compute the least squares estimate (using a computer). Verify that this new line is close to the original and plot the new dataset, the old line, and the new line.

Solution:

My github username is `redpanda1234`.

(a) By Cramer's rule, we have

$$\begin{aligned}
m &= \frac{n\displaystyle\sum_{i=1}^{n}[x_i y_i] - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\displaystyle\sum_{i=1}^{n}\left[x_i^2\right] - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{4(0+6+18+32) - (0+2+3+4)(1+3+6+8)}{4(0+4+9+16) - (0+2+3+4)^2} \\
&= \frac{4(56) - (9\cdot 8)}{4(29) - 9^2} \\
&= \frac{224 - 162}{116 - 81} \\
&= \boxed{\frac{62}{35}}
\end{aligned}$$

and

$$\begin{aligned}
b &= \frac{\left(\displaystyle\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\displaystyle\sum_{i=1}^{n}\left[x_i^2\right] - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{(0+4+9+16)(1+3+6+8) - (0+2+3+4)(0+6+18+32)}{35} \\
&= \frac{29\cdot 18 - 9\cdot 56}{35} \\
&= \frac{522 - 504}{35} \\
&= \boxed{\frac{18}{35}}
\end{aligned}$$

(b) By the normal equation,

$$X^\top \mathbf{y} = X^\top X \boldsymbol{\theta}$$

we have

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 6 \\ 8 \end{bmatrix}$$

thus

$$X^\top X = \begin{bmatrix} 4 & 9 \\ 9 & 29 \end{bmatrix}$$

and

$$X^\top \mathbf{y} = \begin{bmatrix} 18 \\ 56 \end{bmatrix}$$

hence

$$\begin{bmatrix} 18 \\ 56 \end{bmatrix} = \begin{bmatrix} 4 & 9 \\ 9 & 29 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

so we solve the following:

$$\left[\begin{array}{cc|c} 4 & 9 & 18 \\ 9 & 29 & 56 \end{array}\right]$$
$$\left[\begin{array}{cc|c} 4 & 9 & 18 \\ 1 & 11 & 20 \end{array}\right]$$
$$\left[\begin{array}{cc|c} 0 & -35 & -62 \\ 1 & 11 & 20 \end{array}\right]$$
$$\left[\begin{array}{cc|c} 0 & 1 & \frac{62}{35} \\ 1 & 11 & 20 \end{array}\right]$$
$$\left[\begin{array}{cc|c} 0 & 1 & \frac{62}{35} \\ 1 & 0 & \frac{18}{35} \end{array}\right]$$

thus

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} \frac{62}{35} \\ \frac{18}{35} \end{bmatrix}$$

as in part (a).

(c) See plot

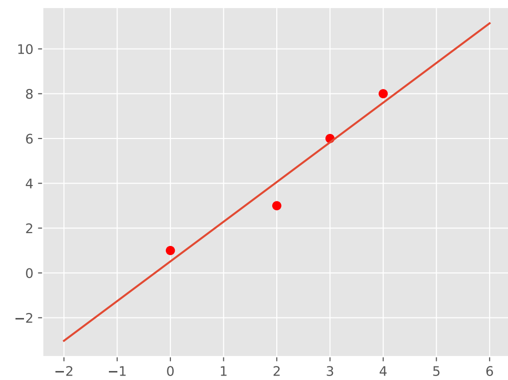

Figure 1: Optimal linear fit for $\mathcal{D}$

(d) See plot below. The values for m and b were calculated to be

$$m = 1.77377696 \qquad b = 0.40970185$$

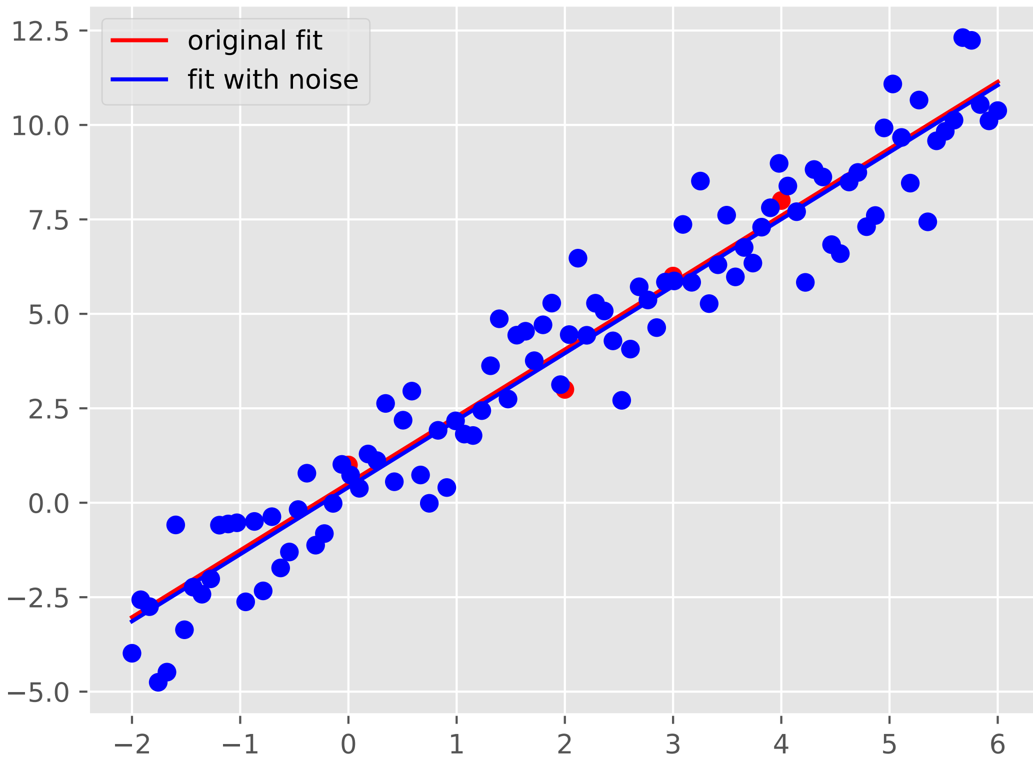

Figure 2: Optimal linear fit for data with noise