Tsinghua-Berkeley Shenzhen Institute
INFORMATION THEORY AND STATISTICAL LEARNING
Fall 2020

Homework 1

HANMO CHEN September 18, 2020

- **Acknowledgments:** For Problem 1.2(b), I refer to some answer on StackExchange `https://math.stackexchange.com/a/108303`
- **Collaborators:** I finish this homework all by myself.
- *I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.*

  Hanmo Chen

1.1. *Mathematical Expectation, Variance, and Covariance Matrix*

(a) i. $\mathbb{E}[x|y] = \mathbb{E}[\mathbb{E}[x|yz]|y]$

Proof.

$$\mathbb{E}[x|yz] = \sum_x x P_{X|YZ}(x|yz) \tag{1}$$

Denote $\mathbb{E}[x|yz]$ as $g(y,z)$

$$\begin{aligned} E[g(y,z)|y] &= \sum_z g(y,z) P_{Z|Y}(z|y) \\ &= \sum_z \left(\sum_x x P_{X|YZ}(x|yz) \right) P_{Z|Y}(z|y) \\ &= \sum_{x,z} x f_{X|YZ}(x|yz) P_{Z|Y}(z|y) \\ &= \sum_{x,z} x P_{X,Z|Y}(x,z|y) \\ &= \sum_x x P_{X|Y}(x|y) \\ &= \mathbb{E}[x|y] \end{aligned} \tag{2}$$

So we have $\mathbb{E}[x|y] = \mathbb{E}[\mathbb{E}[x|yz]|y]$ □

ii. $\mathbb{E}[xg(y)|y] = g(y)\mathbb{E}[x|y]$

Proof.

$$\begin{aligned}\mathbb{E}[xg(y)|y] &= \sum_x xg(y)P_{X|Y}(x|y) \\ &= g(y)\sum_x xP_{X|Y}(x|y) \quad (g(y) \text{ is constant with } y \text{ fixed}) \\ &= g(y)\mathbb{E}[x|y]\end{aligned} \tag{3}$$

□

iii. $\mathbb{E}[x\mathbb{E}[x|y]] = \mathbb{E}\left[(\mathbb{E}[x|y])^2\right]$

Proof. Denote $\mathbb{E}[x|y]$ as $g(y)$, using the conclusion from Equation [3], considering the following expression $\mathbb{E}[xg(y)|y]$

$$\mathbb{E}[xg(y)|y] = g(y)\mathbb{E}[x|y] = (\mathbb{E}[x|y])^2 \tag{4}$$

Considering the expectation of each side in the above equation,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[xg(y)|y]] &= \mathbb{E}[xg(y)] \quad (\text{Total Expectation Rule}) \\ &= \mathbb{E}[x\mathbb{E}(x|y)] \\ &= \mathbb{E}[(\mathbb{E}[x|y])^2]\end{aligned} \tag{5}$$

□

iv. $\mathrm{var}(x) = \mathbb{E}[\mathrm{var}(x|y)] + \mathrm{var}(\mathbb{E}[x|y])$

Proof. Because $\mathrm{var}(x|y) = \mathbb{E}[(x - \mathbb{E}[x|y])^2|y] = \mathbb{E}[x^2|y] - (\mathbb{E}[x|y])^2$,

$$\begin{aligned}\mathbb{E}[\mathrm{var}(x|y)] &= \mathbb{E}[\mathbb{E}[x^2|y] - (\mathbb{E}[x|y])^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[(\mathbb{E}[x|y])^2]\end{aligned} \tag{6}$$

Meanwhile,

$$\begin{aligned}\mathrm{var}(\mathbb{E}[x|y]) &= \mathbb{E}[(\mathbb{E}[x|y])^2] - (\mathbb{E}[\mathbb{E}[x|y]])^2 \\ &= \mathbb{E}[(\mathbb{E}[x|y])^2] - (\mathbb{E}[x])^2\end{aligned} \tag{7}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\mathrm{var}(x|y)] + \mathrm{var}(\mathbb{E}[x|y]) =& \mathbb{E}[x^2] - \mathbb{E}[(\mathbb{E}[x|y])^2] \\ &+ \mathbb{E}[(\mathbb{E}[x|y])^2] - (\mathbb{E}[x])^2 \\ =& \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \\ =& \mathrm{var}(x)\end{aligned} \tag{8}$$

□

(b) i. $\mathrm{cov}(\underline{x}) = \mathbb{E}[\mathrm{cov}(\underline{x}|y)] + \mathrm{cov}(\mathbb{E}[\underline{x}|y])$

Proof. Because $\mathrm{cov}(\underline{x}|y) = \mathbb{E}[\underline{x} \cdot \underline{x}^T|y] - \mathbb{E}[\underline{x}|y] \cdot (\mathbb{E}[\underline{x}|y])^T$,

$$\begin{aligned}\mathbb{E}[\mathrm{cov}(\underline{x}|y)] &= \mathbb{E}[\mathbb{E}[\underline{x} \cdot \underline{x}^T|y] - \mathbb{E}[\underline{x}|y] \cdot (\mathbb{E}[\underline{x}|y])^T] \\ &= \mathbb{E}[\underline{x} \cdot \underline{x}^T] - \mathbb{E}[\mathbb{E}[\underline{x}|y] \cdot (\mathbb{E}[\underline{x}|y])^T]]\end{aligned} \tag{9}$$

Meanwhile,

$$\begin{aligned}\operatorname{cov}(\mathbb{E}[\underline{x}|y]) &= \mathbb{E}[\mathbb{E}[\underline{x}|y]\cdot(\mathbb{E}[\underline{x}|y])^T] - \mathbb{E}[\mathbb{E}[x|y]]\cdot(\mathbb{E}[\mathbb{E}[x|y]])^T \\ &= \mathbb{E}[\mathbb{E}[\underline{x}|y]\cdot(\mathbb{E}[\underline{x}|y])^T] - \mathbb{E}[x]\cdot(\mathbb{E}[x])^T\end{aligned} \tag{10}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\operatorname{cov}(\underline{x}|y)] + \operatorname{cov}(\mathbb{E}[\underline{x}|y]) =& \mathbb{E}[\underline{x}\cdot\underline{x}^T] - \mathbb{E}[\mathbb{E}[\underline{x}|y]\cdot(\mathbb{E}[\underline{x}|y])^T]] \\ &+ \mathbb{E}[\mathbb{E}[\underline{x}|y]\cdot(\mathbb{E}[\underline{x}|y])^T] - \mathbb{E}[x]\cdot(\mathbb{E}[x])^T \\ =& \mathbb{E}[\underline{x}\cdot\underline{x}^T] - \mathbb{E}[x]\cdot(\mathbb{E}[x])^T \\ =& \operatorname{cov}(\underline{x})\end{aligned} \tag{11}$$

□

ii. *Proof.* Denote $\Sigma = \operatorname{cov}(\underline{x}) = \mathbb{E}[(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T], \underline{\mu} = \mathbb{E}[\underline{x}]$, then

$$\det(\Sigma) = 0 \Longleftrightarrow \exists \underline{b} \in \mathbb{R}^k \backslash \{\underline{0}\} \text{ such that } \Sigma\underline{b} = 0 \tag{12}$$

"$\Longrightarrow$":
Denote $\underline{\alpha} = \underline{x} - \underline{\mu}$,

$$\Sigma\underline{b} = \mathbb{E}[\underline{\alpha}\underline{\alpha}^T]\underline{b} = \mathbb{E}[\underline{\alpha}\underline{\alpha}^T\underline{b}] = 0 \tag{13}$$

Mutiplied by $\underline{b}^T$, and denote $\underline{\beta} = \underline{b}^T\underline{\alpha}$, we have

$$\mathbb{E}[\underline{b}^T\underline{\alpha}\underline{\alpha}^T\underline{b}] = \mathbb{E}[\underline{\beta}\underline{\beta}^T] = 0 \tag{14}$$

Suppose $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)^T$, considering the diagonal elements of $\underline{\beta}\underline{\beta}^T$, the i-th diagonal element is $\beta_i^2 \geqslant 0$, so $\mathbb{E}[\beta_i^2] \geqslant 0$, but $\mathbb{E}[\underline{\beta}\underline{\beta}^T] = 0$, which means $\beta_i = 0$ and $\underline{\beta} = \underline{0}$.

$$\underline{b}^T(\underline{x} - \underline{\mu}) = 0 \Longleftrightarrow \underline{b}^T\underline{x} = \underline{b}^T\underline{\mu} \tag{15}$$

So $\exists \underline{c} = \underline{b} \in \mathbb{R}^k \backslash \{\underline{0}\}$ such that $\underline{c}^T\underline{x} = \underline{b}^T\underline{\mu}$ is a constant.

"$\Longleftarrow$":
Because $\underline{c}^T\underline{x}$ is a constant,

$$\underline{c}^T\underline{x} = \mathbb{E}[\underline{c}^T\underline{x}] = \underline{c}^T\mathbb{E}[\underline{x}] = \underline{c}^T\underline{\mu} \tag{16}$$

That is, $\underline{c}^T(\underline{x} - \underline{\mu}) = 0$, so $\underline{c}^T(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T = 0$,

$$\mathbb{E}[\underline{c}^T(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] = \underline{c}^T\mathbb{E}[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] = c^T\Sigma = 0 \tag{17}$$

So $\det(\Sigma) = 0$ □

1.2. (a) *Proof.* To minimize
$L = \mathbb{E}[(y-ax-b)^2] = \mathbb{E}[y^2] + a^2\mathbb{E}[x^2] + b^2 - 2a\mathbb{E}[xy] + 2ab\mathbb{E}[x] - 2b\mathbb{E}[y]$,
let the partial derivative equals 0.

$$\begin{cases} \dfrac{\partial L}{\partial a} = 2a\mathbb{E}[x^2] - 2\mathbb{E}[xy] + 2b\mathbb{E}[x] = 0 \\ \dfrac{\partial L}{\partial b} = 2b + 2a\mathbb{E}[x] - 2\mathbb{E}[y] = 0 \end{cases} \tag{18}$$

Solve the equation, we have

$$a^* = \frac{\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]}{\mathbb{E}[x^2] - (\mathbb{E}[x])^2} = \frac{\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]}{\text{var}(x)} \tag{19}$$

When $\text{var}(x) = \text{var}(y)$ we have $\rho(x, y) = a^*$ $\square$

(b) *Proof.* x and y is independent means $P_{X,Y}(x, y) = P_X(x)P_Y(y)$, and $\rho(f(x), g(y)) = 0$ means $\mathbb{E}[f(x)]\mathbb{E}[g(y)] = \mathbb{E}[f(x)g(y)]$

"$\Longrightarrow$":

Suppose $P_{X,Y}(x, y) = P_X(x)P_Y(y)$, for $\forall f, g$

$$\begin{aligned} \mathbb{E}[f(x)g(y)] &= \sum_{x,y} f(x)g(y)P_{X,Y}(x, y) \\ &= \sum_{x,y} f(x)g(y)P_X(x)P_Y(y)dxdy \\ &= \sum_{x} f(x)P_X(x) \sum_{y} g(y)P_Y(y) \\ &= \mathbb{E}[f(x)]\mathbb{E}[g(y)] \end{aligned} \tag{20}$$

"$\Longleftarrow$":

Suppose $\mathbb{E}[f(x)]\mathbb{E}[g(y)] = \mathbb{E}[f(x)g(y)]$, for all $A, B \in \mathbb{R}$ Let $f(x)$ be the indicator random variable $f(x) = \mathbb{I}(x \in A)$, $g(y) = \mathbb{I}(y \in B)$

$$\begin{aligned} P(x \in A, y \in B) &= \mathbb{E}[\mathbb{I}(x \in A)\mathbb{I}(y \in B)] = \mathbb{E}[f(x)g(y)] \\ &= \mathbb{E}[f(x)]\mathbb{E}[g(y)] = \mathbb{E}[\mathbb{I}(x \in A)]\mathbb{E}[\mathbb{I}(y \in B)]] \\ &= P(x \in A)P(y \in B) \end{aligned} \tag{21}$$

So $P_{X,Y}(x, y) = P_X(x)P_Y(y)$ which means X, Y are independent. $\square$

1.3. Let $X \sim N(0, 1)$,the *pdf* of X is $f_X(x) = -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, denote the cumulative distribution function $\Phi(x) = P(X \leqslant x)$ and $\Phi'(x) = f_X(x)$.

Suppose $Y = exp(X)$, we try to find the cumulative distribution function $P(Y \leqslant y)$,

$$P(Y \leqslant y) = P(X \leqslant \ln y) = \Phi(\ln y) \tag{22}$$

Therefore, the *pdf* of Y is,

$$f_Y(y) = \frac{d\Phi(\ln y)}{dy} = f_X(\ln y)\frac{1}{y} = \frac{1}{\sqrt{2\pi}y}e^{-\frac{1}{2}(\ln y)^2}, y \geqslant 0 \tag{23}$$