Lecture Notes of Multivariate Statistical Analysis

Luo Luo

School of Data Science, Fudan University

September 7, 2023

1 Introduction and Review of Linear Algebra/Optimization

There are some applications in multivariate statistics:

1. Investigating of the dependency among variables (Should you take this course? Are you good at math?)
2. Hypotheses testing (Can I achieve grade A?)
3. Dimensionality reduction (Do you want to join my group? Are you good at math/programming?)
4. Prediction (Can I receive an Phd offer?)
5. Clustering (Course category. Which Phd/master advisor shoud I select?)

课程	学生1	学生2	学生3	学生4	学生5	学生6
习近平新时代中国特色社会主义思想	B+	A-	B	A-	C	A
马克思主义原理	A	A	B	B+	B	B+
形势与政策	A-	A-	A	A-	B+	B+
数学分析	A	A	C+	A-	B-	B+
高等代数	A-	A	C	B+	C+	A-
最优化方法	A	A-	C	A-	C+	A-
多元统计分析	A	?	D	?	?	A-
程序设计	B+	A	A	A-	B+	B-
数据库及实现	B+	?	A	B+	B	?
神经网络与深度学习	B+	A-	A-	A-	?	B
计算机视觉	B+	A	A	?	B-	B-
自然语言处理	B+	?	A	A-	B+	B+

Figure 1: Grading of some students.

Notation of transpose: I do not like use $\mathbf{A}'$ to present the transpose of $\mathbf{A}$.

1. In MATLAB, the notation $\mathbf{A}'$ presents the conjugate transpose. I recommend use $\mathbf{A}^\top$ to present the transpose and $\mathbf{A}^H$ to present the conjugate transpose.
2. The prime usually presents derivative.

Property of trace: For $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$ and $\mathbf{C} \in \mathbb{R}^{p\times m}$, we have

$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB}).$$

However, we cannot write

$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{ACB}),$$

since the product $\mathbf{CB}$ may be even undefined.

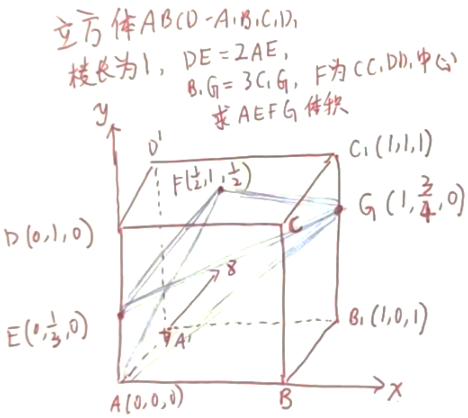
Inverse: For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{D} \in \mathbb{R}^{p\times n}$, we have

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

if $\mathbf{A}$ and $\mathbf{A}+\mathbf{BCD}$ are non-singular. Take $\mathbf{B} = \mathbf{u} \in \mathbb{R}^d$, $\mathbf{C} = 1 \in \mathbb{R}$ and $\mathbf{D} = \mathbf{u}^\top \in \mathbb{R}^{1\times d}$, then we have

$$(\mathbf{A}+\mathbf{uu}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{u}(1+\mathbf{uu}^\top)^{-1}\mathbf{u}^\top\mathbf{A}^{-1},$$

which takes $\mathcal{O}(d^2)$ flops for given $\mathbf{A}^{-1}$.



A.	$\dfrac{\sqrt{2}}{48}$
B.	$\dfrac{1}{36}$
C.	$\dfrac{\sqrt{3}}{60}$
D.	$\dfrac{\sqrt{5}-1}{40}$

$$V = \frac{1}{6}\det\left(\begin{bmatrix} 0 & 1/3 & 0 \\ 1/2 & 1 & 1/2 \\ 1 & 3/4 & 0 \end{bmatrix}\right)$$

Figure 2: Understanding the meaning of determinant.

Theorem 1.1 (Property of Schur Complement)**.** *We consider the symmetric matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}$$

with non-singular $\mathbf{D} \in \mathbb{R}^{q\times q}$ *and let* $\mathbf{S} = \mathbf{A} - \mathbf{BD}^{-1}\mathbf{B}^\top \in \mathbb{R}^{p\times p}$*, then*

1. $\mathbf{M} \succ \mathbf{0} \iff \mathbf{D} \succ \mathbf{0}$ *and* $\mathbf{S} \succ \mathbf{0}$.
2. *If* $\mathbf{D} \succ \mathbf{0}$*, then* $\mathbf{M} \succeq \mathbf{0} \iff \mathbf{S} \succeq \mathbf{0}$.

Proof. **Part I**: The condition $\mathbf{M} \succ \mathbf{0}$ means for any $\mathbf{x} = [0,\ldots,0,\mathbf{u}]^\top \in \mathbb{R}^{p+q}$ with nonzero $\mathbf{u} \in \mathbb{R}^q$, we have $\mathbf{x}^\top\mathbf{Mx} > 0$, which implies

$$\begin{bmatrix} \mathbf{0}^\top & \mathbf{u}^\top \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{u} \end{bmatrix} = \mathbf{u}^\top\mathbf{Du} > 0.$$

Recall the decomposition

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{BD}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}-\mathbf{BD}^{-1}\mathbf{B}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B}^\top & \mathbf{I} \end{bmatrix} = \mathbf{G}^\top \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \mathbf{G} \quad \text{where} \quad \mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B}^\top & \mathbf{I} \end{bmatrix}.$$

It is obviously that $\mathbf{G}$ is inevitable. For any nonzero $\mathbf{w} \in \mathbb{R}^{p+q}$, we have

$$\mathbf{G}^{-1}\mathbf{w} \neq \mathbf{0} \quad \Longrightarrow \quad \mathbf{w}^\top(\mathbf{G}^{-1})^\top\mathbf{MG}^{-1}\mathbf{w} > 0.$$

For $\mathbf{w} = [\mathbf{v}^\top, \mathbf{0}^\top]^\top$ with any $\mathbf{v} \in \mathbb{R}^p$, we have

$$\begin{bmatrix} \mathbf{v}^\top & \mathbf{0}^\top \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} > 0 \quad \Longrightarrow \quad \mathbf{v}^\top\mathbf{Sv} > 0.$$

Part II: Leave for homework. $\square$

Hessian and higher order expansion: Taylor's expansion of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a} \in \mathbb{R}^n$ is

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \text{higher order terms}.$$

For single variable case, we have

$$f(x) \approx f(a) + f'(a)(x - a) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^2 f'(a) + \frac{1}{6}(x - a)^3 f'(a).$$

For multivariate case, we have

$$\begin{aligned} f(\mathbf{x}) \approx f(\mathbf{a}) + &\nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \cdot (x_i - a_i)(x_j - a_j) \\ &+ \frac{1}{6} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial^3 f(\mathbf{x})}{\partial x_i \partial x_j \partial x_k} \cdot (x_i - a_i)(x_j - a_j)(x_k - a_k). \end{aligned}$$

Now we provide some results for optimization.

Theorem 1.2. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then it is convex if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Theorem 1.3. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, then $\mathbf{x}^*$ is the global minimizer of $f(\cdot)$ if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Proof. Suppose that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. For any $\mathbf{x} \in \mathbb{R}^d$, Theorem 1.2 means

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle = f(\mathbf{x}^*).$$

Suppose that $\mathbf{x}^*$ is the global minimizer of $f(\cdot)$. For any $\lambda > 0$ and $\mathbf{p} \in \mathbb{R}^d$, we have $f(\mathbf{x}^* + \lambda \mathbf{p}) \geq f(\mathbf{x}^*)$ that means

$$\frac{f(\mathbf{x}^* + \lambda \mathbf{p}) - f(\mathbf{x}^*)}{\lambda} \geq 0.$$

Taking $\lambda \to 0^+$, then

$$\lim_{\lambda \to 0^+} \frac{f(\mathbf{x}^* + \lambda \mathbf{p}) - f(\mathbf{x}^*)}{\lambda} = \langle \nabla f(\mathbf{x}^*), \mathbf{p} \rangle \geq 0.$$

Above results also holds for $\mathbf{p} \leq 0$. Hence, we conclude $\nabla f(\mathbf{x}^*) = \mathbf{0}$. $\square$

Remark 1.1. *Let $\mathbf{u}(\lambda) = \mathbf{x}^* + \lambda \mathbf{p}$ and $g(\lambda) = f(\mathbf{x}^* + \lambda \mathbf{p}) = f(\mathbf{u}(\lambda))$, then*

$$\lim_{\lambda \to 0^+} \frac{f(\mathbf{x}^* + \lambda \mathbf{p}) - f(\mathbf{x}^*)}{\lambda} = g'(\lambda) = \sum_{i=1}^{d} \frac{\partial f}{\partial u_i} \frac{\partial u_i}{\partial \lambda} \bigg|_{\lambda = 0} = \langle \nabla f(\mathbf{x}^*), \mathbf{p} \rangle.$$

Theorem 1.4. *Suppose $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$ and that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succ \mathbf{0}$. Then $\mathbf{x}^*$ is a strict local minimizer of $f(\cdot)$.*

Theorem 1.5. *Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution of minimization problem*

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \triangleq \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 .$$

is $\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$.

Proof. We can verify

$$\nabla f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A} \succeq \mathbf{0},$$

which means $f(\mathbf{x})$ is convex. Hence, we only needs to solve the linear system

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}.$$

If $\mathbf{A}^\top\mathbf{A}$ is full rank, we have

$$\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}.$$

Otherwise, we let $\mathbf{A} = \mathbf{U}_r\boldsymbol{\Sigma}_l r\mathbf{V}_r^\top$ be the condense SVD, where r is the rank of $\mathbf{A}$. We denote the solution of $\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}$ be

$$\mathcal{X} = \left\{\mathbf{x} : \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}\right\}$$

and denote

$$\mathcal{X}_1 = \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}, \quad \mathbf{y} \in \mathbb{R}^n\right\}.$$

We can verify that $\mathbf{x}^* = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y} \in \mathcal{X}_1$ satisfies $\mathbf{x}^* \in \mathcal{X}$ as follows

$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\mathbf{x}^* - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top\mathbf{A}\left(\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}\right) - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\dagger - \mathbf{I})\mathbf{b} + \mathbf{A}^\top\mathbf{A}\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{y} \\
=&\mathbf{V}_r\boldsymbol{\Sigma}_r\mathbf{U}_r^\top(\mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^\top\mathbf{V}_r\boldsymbol{\Sigma}_r^{-1}\mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r\boldsymbol{\Sigma}_r\mathbf{U}_r^\top\mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^\top\left(\mathbf{I} - \mathbf{V}_r\boldsymbol{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^\top\right)\mathbf{y} \\
=&\mathbf{V}_r\boldsymbol{\Sigma}_r\mathbf{U}_r^\top(\mathbf{U}_r\mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r\boldsymbol{\Sigma}_r^2\mathbf{V}_r^\top\left(\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top\right)\mathbf{y} \\
=&\mathbf{V}_r\boldsymbol{\Sigma}_r(\mathbf{U}_r^\top - \mathbf{U}_r^\top)\mathbf{b} + \mathbf{V}_r\boldsymbol{\Sigma}_r^2\left(\mathbf{V}_r^\top - \mathbf{V}_r^\top\right)\mathbf{y} \\
=&\mathbf{0}.
\end{aligned}$$

Hence, we have $\mathcal{X}_1 \subseteq \mathcal{X}$.

For any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\mathbf{V}_r\boldsymbol{\Sigma}_r^2\mathbf{V}_r^\top\mathbf{x} - \mathbf{V}_r\boldsymbol{\Sigma}_r\mathbf{U}_r^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\boldsymbol{\Sigma}_r^2\mathbf{V}_r^\top\mathbf{x} - \boldsymbol{\Sigma}_r\mathbf{U}_r^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\mathbf{V}_r^\top\mathbf{x} = \boldsymbol{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{b} \\
\Longleftrightarrow&\mathbf{V}_r\mathbf{V}_r^\top\mathbf{x} = \mathbf{V}_r\boldsymbol{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{b} \\
\Longleftrightarrow&\mathbf{x} - (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} \\
\Longleftrightarrow&\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}.
\end{aligned}$$

Then $\mathbf{x} \in \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x}\right\} \subseteq \mathcal{X}_1$, which means $\mathcal{X} \subseteq \mathcal{X}_1$. Hence, we have $\mathcal{X} = \mathcal{X}_1$. □

Remark 1.2. *We consider gradient descent method. Taking fixed stepsize $\eta > 0$, we have*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t) = \mathbf{x}_t - \eta(\mathbf{A}^\top\mathbf{A}\mathbf{x}_t - \mathbf{A}^\top\mathbf{b}),$$

which takes $\mathcal{O}(mn)$ flops for each iteration, while the closed form solution requires $\mathcal{O}(mn^2)$.

Rank	Player	PTS	TRB	AST	STL	BLK	FG%
1	Nikola Jokić	27.1	13.8	7.9	1.5	0.9	0.583
2	Joel Embiid	30.6	11.7	4.2	1.1	1.5	0.499
3	Giannis Antetokounmpo	29.9	11.6	5.8	1.1	1.4	0.553

Figure 3: MVP ranking of NBA season 2022-2023.

Let $\mathbf{A} = \begin{bmatrix} 27.1 & 13.8 & 7.9 & 1.5 & 0.9 & 0.583 \\ 30.6 & 11.7 & 4.2 & 1.1 & 1.5 & 0.499 \\ 29.9 & 11.6 & 5.8 & 1.1 & 1.4 & 0.553 \end{bmatrix} \in \mathbb{R}^{3\times 6}$ and $\mathbf{b} = [1,2,3]^\top \in \mathbb{R}^3$. We want to find $\mathbf{x} \in \mathbb{R}^6$ to predict the MVP rank for a player with statistic $\mathbf{a} \in \mathbb{R}^6$ by $\mathbf{a}^\top \mathbf{x}$. Note that $\operatorname{rank}(\mathbf{A}^\top \mathbf{A}) < 6$, then

$$\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b} = [0.3754, -1.0710, 0.7275, -0.1729, 0.1051, 0.0407].$$

The feature of Luka is $\mathbf{a} = [28.4, 9.1, 8.7, 1.2, 0.6, 0.634]$. We achieve $\mathbf{a}^\top \mathbf{x} = 7.1260$.

2 Random Vectors and Matrices

Theorem 2.1. *Let $\mathbf{X}$ and $\mathbf{Y}$ be random matrices off the same dimension, and let $\mathbf{A}$ and $\mathbf{B}$ be conformable matrices of constants. Then we have*

$$\mathbb{E}[\mathbf{X}+\mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}] \qquad \textit{and} \qquad \mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B}.$$

Proof. It follows the univariate properties of expectation $\mathbb{E}[x_1 + x_2] = \mathbb{E}[x_1] + \mathbb{E}[x_2]$ and $\mathbb{E}[c_1 x_1] = c_1 \mathbb{E}[x_1]$ for random variables x, y and constant c. It implies

$$\mathbb{E}[c_1 x_1 + \cdots + c_n x_n] = c_1 \mathbb{E}[x_1] + \cdots + c_n \mathbb{E}[x_n].$$

Let $\mathbf{Y} = \mathbf{XB}$, then

$$(\mathbb{E}[\mathbf{AY}])_{ij} = \mathbb{E}[(\mathbf{AY})_{ij}] = \mathbb{E}\left[\sum_k a_{ik} y_{kj}\right] = \sum_k a_{ik} \mathbb{E}[y_{kj}] = \sum_k a_{ik} (\mathbb{E}[\mathbf{Y}])_{kj},$$

which means $\mathbb{E}[\mathbf{AY}] = \mathbf{A}\mathbb{E}[\mathbf{Y}]$ (that is $\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[\mathbf{XB}]$). Similarly, we have

$$(\mathbb{E}[\mathbf{XB}])_{ij} = \mathbb{E}[(\mathbf{XB})_{ij}] = \mathbb{E}\left[\sum_k x_{ik} b_{kj}\right] = \sum_k \mathbb{E}[x_{ik}] b_{kj} = \sum_k (\mathbb{E}[\mathbf{X}])_{ik} b_{kj},$$

which means $\mathbb{E}[\mathbf{XB}] = \mathbb{E}[\mathbf{X}]\mathbf{B}$. Thus, we achieve $\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[\mathbf{XB}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B}$. ☐

Theorem 2.2. *Let $\mathbf{x} = \begin{bmatrix} x_1, \ldots, x_p \end{bmatrix}^\top$ be a random vector and we denote $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. Then we have*

$$\operatorname{Cov}[\mathbf{x}] = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Proof. We have

$$\begin{aligned} \operatorname{Cov}[\mathbf{x}] =& \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right] \\ =& \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top - \boldsymbol{\mu}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right] \\ =& \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \mathbb{E}\big[\boldsymbol{\mu}\mathbf{x}^\top\big] - \mathbb{E}\big[\mathbf{x}\boldsymbol{\mu}^\top\big] + \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^\top] \\ =& \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \boldsymbol{\mu}\mathbb{E}\big[\mathbf{x}^\top\big] - \mathbb{E}\left[\mathbf{x}\right]\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ =& \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ =& \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top, \end{aligned}$$

where the third and fourth lines use Theorem 2.1. ☐

Remark 2.1. *For single random variable x, we have*

$$\operatorname{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2.$$

Theorem 2.3. *Let $\mathbf{y} = \mathbf{Dx} + \mathbf{f}$, where $\mathbf{D}$ is an $n \times p$ constant matrix, $\mathbf{x}$ is a p-dimensional random vector and $\mathbf{f}$ is a n-dimensional constant vector, then*

$$\operatorname{Cov}[\mathbf{y}] = \mathbf{D}\operatorname{Cov}[\mathbf{x}]\mathbf{D}^\top.$$

Proof. We have

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{y})\\
=&\mathbb{E}\left[(\mathbf{y}-\mathbb{E}[\mathbf{y}])(\mathbf{y}-\mathbb{E}[\mathbf{y}])^\top\right]\\
=&\mathbb{E}\left[(\mathbf{D}\mathbf{x}+\mathbf{f}-\mathbb{E}[\mathbf{D}\mathbb{E}[\mathbf{x}]+\mathbf{f}])(\mathbf{D}\mathbf{x}+\mathbf{f}-\mathbb{E}[\mathbf{D}\mathbb{E}[\mathbf{x}]+\mathbf{f}])^\top\right]\\
=&\mathbb{E}[(\mathbf{D}\mathbf{x}-\mathbf{D}\mathbb{E}[\mathbf{x}])(\mathbf{D}\mathbf{x}-\mathbf{D}\mathbb{E}[\mathbf{x}])^\top]\\
=&\mathbb{E}[\mathbf{D}(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^\top\mathbf{D}^\top]\\
=&\mathbf{D}\mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^\top]\mathbf{D}^\top\\
=&\mathbf{D}\mathrm{Cov}[\mathbf{x}]\mathbf{D}^\top.
\end{aligned}$$

□

Example 2.1. *Let $\mathbf{x}=[x_1,x_2]^\top$ be a random vector with*

$$\mathbb{E}[\mathbf{x}]=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \qquad \textit{and} \qquad \mathrm{Cov}[\mathbf{x}]=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}.$$

Let $\mathbf{z}=[z_1,z_2]$ such that $z_1=x_1-x_2$ and $z_2=x_1+x_2$.

1. *Find the $\mathbb{E}[\mathbf{z}]$ and $\mathrm{Cov}[\mathbf{z}]$.*
2. *Find the condition that leads to z_1 and z_2 be uncorrelated.*

Solution: We can write

$$\mathbf{z}=\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\mathbf{x}=\mathbf{C}\mathbf{x}.$$

Then we have

$$\begin{aligned}
\mathbb{E}[\mathbf{z}]=&\mathbf{C}\mathbb{E}[\mathbf{x}]\\
=&\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\\
=&\begin{bmatrix}\mu_1-\mu_2\\ \mu_1+\mu_2\end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}[\mathbf{z}]=&\mathbf{C}\mathrm{Cov}[\mathbf{z}]\mathbf{C}^\top\\
=&\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}\\
=&\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\sigma_{11}-\sigma_{12} & \sigma_{11}+\sigma_{12}\\ \sigma_{21}-\sigma_{22} & \sigma_{21}+\sigma_{22}\end{bmatrix}\\
=&\begin{bmatrix}\sigma_{11}-2\sigma_{12}+\sigma_{22} & \sigma_{11}-\sigma_{12}\\ \sigma_{11}-\sigma_{22} & \sigma_{11}+2\sigma_{12}+\sigma_{22}.\end{bmatrix}
\end{aligned}$$

If $\sigma_{11}=\sigma_{22}$, then variables z_1 and z_2 are uncorrelated.

Remark 2.2. *The random vector with diagonal covariance matrix is easy to deal with. Note that the transform based on $\mathbf{C}$ does not loss any information since $\mathbf{C}$ is full rank.*

Transform of Variables Let the density of $x_1, \ldots, x_p$ be $f(x_1, \ldots, x_p)$. Consider the p real-valued functions $\mathbf{u} : \mathbb{R}^p \to \mathbb{R}^p$ such that $\mathbf{u} = [u_1(\mathbf{x}), \ldots, u_p(\mathbf{x})]^\top$ with

$$y_i = u_i(x_1, \ldots, x_p), \qquad i = 1, \ldots, p.$$

Assume the transformation $\mathbf{u}$ from the space of $\mathbf{x}$ to the space of $\mathbf{y}$ is one-to-one, then the inverse transformation is $\mathbf{u}^{-1}$ such that $\mathbf{u}^{-1} = [u_1^{-1}(\mathbf{y}), \ldots, u_p^{-1}(\mathbf{y})]^\top$ with

$$x_i = u_i^{-1}(y_1, \ldots, y_p), \qquad i = 1, \ldots, p.$$

Let the density of $\mathbf{y} = [y_1, \ldots, y_p]^\top$ be $g(\mathbf{y})$. Then we have

$$\int_{\mathbf{u}(\Omega)} g(\mathbf{y}) \mathrm{d}\mathbf{y} = \int_{\Omega} g\left(\mathbf{u}(\mathbf{x})\right) |\det(\mathbf{J}(\mathbf{x}))| \mathrm{d}\mathbf{x}, \tag{1}$$

and

$$f(\mathbf{x}) = g\left(\mathbf{u}(\mathbf{x})\right) \operatorname{abs}(|\mathbf{J}(\mathbf{x})|), \tag{2}$$

where the Jacobin matrix is

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} & \cdots & \dfrac{\partial u_1}{\partial x_p} \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} & \cdots & \dfrac{\partial u_2}{\partial x_p} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial u_p}{\partial x_1} & \dfrac{\partial u_p}{\partial x_2} & \cdots & \dfrac{\partial u_p}{\partial x_p} \end{bmatrix}.$$

A roughly proof for above results:

- Let $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathcal{S} \subset \mathbb{R}^p$ be a measurable set. We define

  $$\mathbf{A}\mathcal{S} = \{\mathbf{A}\mathbf{s} : \mathbf{s} \in \mathcal{S}\}.$$

  then we can show $m(\mathbf{A}\mathcal{S}) = |\det(\mathbf{A})| m(\mathcal{S})$. Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\boldsymbol{\Sigma}$ is diagonal with nonnegative entries. Multiplying by $\mathbf{V}^\top$ doesn't change the measure of $\mathcal{S}$. Multiplying by $\boldsymbol{\Sigma}$ scales along each axis, so the measure gets multiplied by $|\det(\boldsymbol{\Sigma})| = |\det(\mathbf{A})|$. Multiplying by $\mathbf{U}$ doesn't change the measure.

- We consider the probability of $\mathbf{x}$ in Ω and $\mathbf{y}$ in $\mathbf{u}(\Omega)$; and partition Ω into $\cup_i \Omega_i$. Then

  $$\begin{aligned} & \int_{\mathbf{u}(\Omega)} g(\mathbf{y}) \mathrm{d}\mathbf{y} \\ \approx & \sum_i g(\mathbf{u}(\mathbf{x}_i)) m\left(\mathbf{u}(\Omega_i)\right) \\ \approx & \sum_i g(\mathbf{u}(\mathbf{x}_i)) m(\mathbf{u}(\mathbf{x}_i) + \mathbf{J}(\mathbf{x}_i)(\Omega_i - \mathbf{x}_i)) \\ = & \sum_i g(\mathbf{u}(\mathbf{x}_i)) m(\mathbf{J}(\mathbf{x}_i)\Omega_i) \\ = & \sum_i g(\mathbf{u}(\mathbf{x}_i)) |\det(\mathbf{J}(\mathbf{x}_i))| m(\Omega_i) \\ \approx & \int_{\Omega} g(\mathbf{u}(\mathbf{x})) \, |\det(\mathbf{J}(\mathbf{x}))| \, \mathrm{d}\mathbf{x}. \end{aligned}$$

- Consider notation Ω such that

$$\int_{\Omega} = \int_{x_1}^{x_1'} \cdots \int_{x_p}^{x_p'}$$

where $x_1 \leq x_1', x_2 \leq x_2', \ldots, x_p \leq x_p'$. Then the notation $\mathbf{u}(\Omega)$ in the integral should consider the order

$$\int_{\mathbf{u}(\Omega)} = \int_{\min\{u_1(x_1),u_1(x_1')\}}^{\max\{u_1(x_1),u_1(x_1')\}} \cdots \int_{\min\{u_p(x_p),u_p(x_p')\}}^{\max\{u_p(x_p),u_p(x_p')\}}$$

By using even tinier subsets Ω_i, the approximation would be even better so we see by a limiting argument that we actually obtain (1). On the other hand, we have (f is density functions of $\mathbf{x}$ on Ω; g is density function of $\mathbf{y}$ on $\mathbf{u}(\Omega)$; $\mathbf{y} = \mathbf{u}(\mathbf{x})$ means $\mathbf{x}$ and $\mathbf{y} = \mathbf{u}(\mathbf{x})$ are one-to-one mapping).

$$\int_{\Omega} f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\mathbf{u}(\Omega)} g(\mathbf{y})\mathrm{d}\mathbf{y} = \int_{\Omega} g(\mathbf{u}(\mathbf{x}))\mathrm{abs}(|\mathbf{J}(\mathbf{x})|)\mathrm{d}\mathbf{x}.$$

Since it holds for any Ω, then

$$f(\mathbf{x}) = g(\mathbf{u}(\mathbf{x}))\mathrm{abs}(|\mathbf{J}(\mathbf{x})|).$$

Theorem 2.4. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be p-dimensional random vector and they are independent. Denote*

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha \quad \textit{and} \quad \hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

If $\mathbb{E}[\mathbf{x}_1] = \cdots = \mathbb{E}[\mathbf{x}_N] = \boldsymbol{\mu}$ and $\mathrm{Cov}[\mathbf{x}_1] = \cdots = \mathrm{Cov}[\mathbf{x}_N] = \mathbf{\Sigma}$, then we have

$$\mathbb{E}[\bar{\mathbf{x}}] = \boldsymbol{\mu}, \qquad \mathrm{Cov}[\bar{\mathbf{x}}] = \frac{1}{N}\mathbf{\Sigma}, \qquad \textit{and} \qquad \mathbb{E}\big[\hat{\mathbf{\Sigma}}\big] = \frac{N-1}{N}\mathbf{\Sigma}.$$

Proof. **Part I:** We have

$$\mathbb{E}[\bar{\mathbf{x}}] = \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha\right] = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbb{E}\left[\mathbf{x}_\alpha\right] = \boldsymbol{\mu}$$

and

$$\begin{aligned}
\mathrm{Cov}[\bar{\mathbf{x}}] =& \mathbb{E}\left[\left(\frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha - \boldsymbol{\mu}\right)\left(\frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha - \boldsymbol{\mu}\right)^\top\right] \\
=& \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right)\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right)^\top\right] \\
=& \frac{1}{N^2}\mathbb{E}\left[\sum_{\alpha=1}^{N}\sum_{\beta=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\beta - \boldsymbol{\mu})^\top\right].
\end{aligned}$$

Since random vectors $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$ are independent when $\alpha \neq \beta$, the covariance of $x_{\alpha i}$ and $x_{\alpha j}$ is zero, that is

$$\mathbb{E}[(x_{\alpha i} - \mu_i)(x_{\beta j} - \mu_j)] = 0,$$

which is just the (i,j)-th entry of $\mathbb{E}[(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\beta - \boldsymbol{\mu})^\top]$. Hence, we have

$$\mathrm{Cov}[\bar{\mathbf{x}}] = \frac{1}{N}\mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\right] = \frac{1}{N}\mathbf{\Sigma}.$$

Part II: Applying Theorem 2.2 on $\mathbf{x}_\alpha$, we have

$$\mathbf{\Sigma} = \mathrm{Cov}[\mathbf{x}_\alpha] = \mathbb{E}\left[\mathbf{x}_\alpha \mathbf{x}_\alpha^\top\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

Applying Part I and Theorem 2.2 on $\bar{\mathbf{x}}$, we have

$$\frac{1}{N}\mathbf{\Sigma} = \mathrm{Cov}[\bar{\mathbf{x}}] = \mathrm{Cov}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Hence, we obtain

$$\begin{aligned}
\mathbb{E}\big[\hat{\mathbf{\Sigma}}\big] =& \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\right] \\
=& \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N}\left(\mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \bar{\mathbf{x}}\mathbf{x}_\alpha^\top - \mathbf{x}_\alpha \bar{\mathbf{x}}^\top + \bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)\right] \\
=& \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right] \\
=& \mathbb{E}\left[\mathbf{x}_\alpha \mathbf{x}_\alpha^\top\right] - \mathbb{E}\left[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right] \\
=& \mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \left(\frac{1}{n}\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) \\
=& \frac{N-1}{N}\mathbf{\Sigma}.
\end{aligned}$$

□

Consider minimization problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x} - \mathbf{q}^\top \mathbf{x},$$

where $\mathbf{M} \in \mathbb{R}^{n\times n}$ is positive semi-definite and $\mathbf{q} \in \mathbb{R}^n$. We have

$$\begin{aligned} f(\mathbf{x}_{t+1}) \leq & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 \\ \leq & f(\mathbf{x}_t) - \left(\eta - \frac{L\eta^2}{2}\right)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \\ \leq & f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2, \end{aligned} \tag{3}$$

where $L = \lambda_1(\mathbf{M})$ and $\eta = 1/L$. There exists $\mathbf{x}^* \in \mathbb{R}^d$ such that $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Then we have

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^*) = & \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} - \left(\frac{1}{2}\mathbf{x}^{*\top}\mathbf{A}\mathbf{x}^* - \mathbf{b}^\top \mathbf{x}^*\right) \\ = & \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{x}^{*\top}\mathbf{A}\mathbf{x} - \left(\frac{1}{2}\mathbf{x}^{*\top}\mathbf{A}\mathbf{x}^* - \mathbf{x}^{*\top}\mathbf{A}\mathbf{x}^*\right) \\ = & \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{x}^{*\top}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^{*\top}\mathbf{A}\mathbf{x}^* \\ = & \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*) \end{aligned}$$

and

$$\left\|\nabla f(\mathbf{x})\right\|_2^2 = \left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2 = \left\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\right\|_2^2 = (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}^2 (\mathbf{x} - \mathbf{x}^*).$$

Taking $\mu = \lambda_k(\mathbf{A})$, where $\lambda_k(\mathbf{A})$ is the smallest nonzero eigenvalue of $\mathbf{A}$. Then it holds that $\mu\mathbf{A} \preceq \mathbf{A}^2$ and

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\left\|\nabla f(\mathbf{x})\right\|_2^2. \tag{4}$$

Combining (3) and (4), we have

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq f(\mathbf{x}_t) - \frac{\eta}{2}\cdot 2\mu(f(\mathbf{x}_t) - f(\mathbf{x}^*)),$$

which implies

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_t) - \eta\mu(f(\mathbf{x}_t) - f(\mathbf{x}^*)) - f(\mathbf{x}^*) = \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^*)).$$