Lecture Notes of Multivariate Statistical Analysis

Luo Luo

School of Data Science, Fudan University

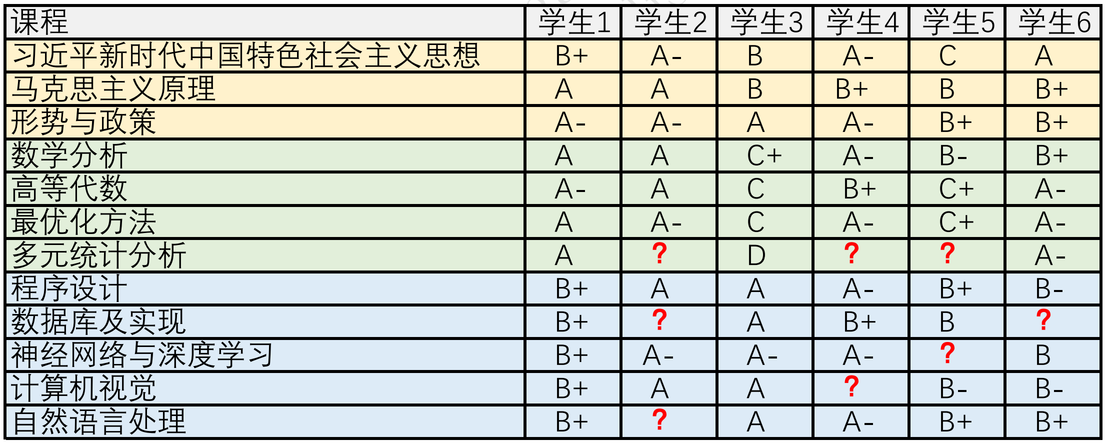
August 13, 2025

1 Review of Linear Algebra and Optimization

ganized by Luo Luo.
istribute.

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. Dimensionality reduction (Do you want to join my group? Are you good at math/programming?)
3. Prediction (Can a get A? Can I receive a Phd/master offer?)
4. Clustering (Course category. Which Phd/master advisor should 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 using $\mathbf{A}^\top$ 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

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

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

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

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

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

Recall the decomposition

$$\begin{aligned} \mathbf{M} = & \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix} \\ = & \begin{bmatrix} \mathbf{I} & \mathbf{BD}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}-\mathbf{BD}^{-1}\mathbf{B}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B}^\top & \mathbf{I} \end{bmatrix} \\ = & \mathbf{G}^\top \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \mathbf{G}, \end{aligned}$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B}^\top & \mathbf{I} \end{bmatrix}.$$

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

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

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

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

Part II: The remain leaves for homework. □

This document is organized by Luo Luo.
Please do not distribute.

Regularized Matrix Approximation For positive semi-definite $\mathbf{A} \in \mathbb{R}^{n\times n}$, we can verify

$$\big(\mathbf{U}_k(\mathbf{\Sigma}_k - \hat{\delta}\mathbf{I}_k)^{1/2}, \hat{\delta}\big) = \operatorname*{arg\,min}_{\text{rank}(\mathbf{B})\leq k, \delta\in\mathbb{R}} \big\|\mathbf{A} - (\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d)\big\|_F,$$

where

$$\hat{\delta} = \frac{1}{n-k}\sum_{i=k+1}^{n}\sigma_i.$$

For any unitary invariant norm $\|\cdot\|$, we have[1]

$$\|\mathbf{X} - \mathbf{Y}\| \geq \|\mathbf{\Sigma}(\mathbf{X}) - \mathbf{\Sigma}(\mathbf{Y})\|.$$

Applying this result, we have

$$\begin{aligned}
&\big\|\mathbf{A} - (\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d)\big\|_F \\
\geq&\big\|\mathbf{\Sigma}(\mathbf{A}) - \mathbf{\Sigma}(\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d)\big\|_F \\
=&\sum_{i=1}^{n}(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d))^2 \\
\geq&\sum_{i=k+1}^{n}(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d))^2 \\
=&\sum_{i=k+1}^{n}(\sigma_i(\mathbf{A}) - \delta)^2 \qquad \text{//because } \mathbf{B} \text{ is low-rank} \\
\geq&\sum_{i=k+1}^{n}(\sigma_i(\mathbf{A}) - \hat{\delta})^2.
\end{aligned}$$

Note that all above equalities occur by taking $\mathbf{B} = \mathbf{U}_k(\mathbf{\Sigma}_k - \hat{\delta}\mathbf{I}_k)^{1/2}$ and $\delta = \hat{\delta}$.

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

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

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 (Section 3.1.3 of Boyd and Vandenberghe [5])**.** *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 $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.

This document is organized by Luo Luo.
Please do not distribute.

[1] See Problem 18 in Section 3.5 (page 215) of Horn and Johnson [8] and Appendix B of Luo et al. [11].

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 $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Theorem 1.4 (Theorem 2.1.4 of Nesterov [14])**.** *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, then it is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ holds for any $\mathbf{x} \in \mathbb{R}^d$.*

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

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

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

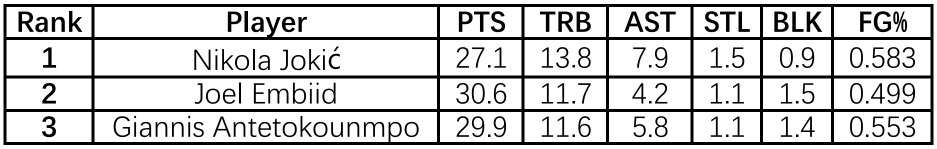
$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\mathbf{x}^* - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top\mathbf{A}\left(\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}\right) - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\dagger - \mathbf{I})\mathbf{b}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

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

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

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

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 $\mathcal{X} = \mathcal{X}_1$. □



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 2: MVP ranking of NBA season 2021-2022.

This document is organized by Luo Luo.
Please do not distribute.

Let

$$\mathbf{A} = \begin{bmatrix} 27.1 & 13.8 & 7.9 & 1.5 & 0.9 & 0.583 \\ 30.6 & 11.7 & 4.2 & 1.1 & 1.5 & 0.499 \\ 29.9 & 11.6 & 5.8 & 1.1 & 1.4 & 0.553 \end{bmatrix} \in \mathbb{R}^{3\times 6} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^3.$$

We want to find $\mathbf{x} \in \mathbb{R}^6$ to predict the MVP ranking 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

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

is a solution. The feature of Luka Dončić is $\mathbf{a} = [28.4, 9.1, 8.7, 1.2, 0.6, 0.634]^\top$. We achieve $\mathbf{a}^\top \mathbf{x} = 7.1260$. In real world, his ranking is 5. Note that using linear regression to predict ranking is not a good idea, because ranking is not continuous variable.

2 Random Vectors and Matrices

This document is organized by Luo Luo. Please do not distribute.

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*

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

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

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

Let $\mathbf{Y} = \mathbf{X}\mathbf{B}$, then

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

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

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

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

Theorem 2.2. *Let* $\mathbf{x} = \big[x_1, \dots, x_p\big]^\top$ *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 with $\mathbb{E}[x] = \mu$, we have*

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

Theorem 2.3. *Let $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{f}$, 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 $\mathrm{Cov}[\mathbf{y}] = \mathbf{D}\mathrm{Cov}[\mathbf{x}]\mathbf{D}^\top$.*

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.*

This document is organized by Luo Luo.
Please do not distribute.

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

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

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 $\mathbf{u}^{-1} = [u_1^{-1}(\mathbf{y}), \dots, u_p^{-1}(\mathbf{y})]^\top$ with

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

Let the density of $\mathbf{y} = [y_1, \dots, y_p]^\top$ 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

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

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

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

This document is organized by Luo Luo.
Please do not distribute.

- Consider notation Ω such that
$$\int_{\Omega} = \int_{x_1}^{x_1'} \cdots \int_{x_p}^{x_p'}$$
where $x_1 \leq x_1', x_2 \leq x_2', \dots, x_p \leq x_p'$. 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).

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

Since it holds for any Ω, then

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

Theorem 2.4. *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be p-dimensional random vector. 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

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

This document is organized by Luo Luo.
Please do not distribute.

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

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

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

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

Hence, we obtain

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

□

Matrix Presentation: Define $\mathbf{1}_N = [1, \dots, 1] \in \mathbb{R}^N$ and let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_p \end{bmatrix} \in \mathbb{R}^{N\times p} \in \mathbb{R}^{N\times p},$$

then we can write

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}^\top\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)\mathbf{X}.$$

Consider that

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha = \frac{1}{N}\mathbf{X}^\top\mathbf{1}_N \quad \text{and} \quad \begin{bmatrix} \bar{\mathbf{x}}^\top \\ \vdots \\ \bar{\mathbf{x}}^\top \end{bmatrix} = \mathbf{1}_N\bar{\mathbf{x}}^\top = \mathbf{1}_N\left(\frac{1}{N}\mathbf{X}^\top\mathbf{1}_N\right)^\top = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{X}.$$

then

$$\begin{aligned}
(N-1)\mathbf{S} =& \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = \begin{bmatrix}\mathbf{x}_1 - \bar{\mathbf{x}} & \dots & \mathbf{x}_N - \bar{\mathbf{x}}\end{bmatrix}\begin{bmatrix}(\mathbf{x}_1 - \bar{\mathbf{x}})^\top \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^\top\end{bmatrix} \\
=& \left(\mathbf{X} - \mathbf{1}_N\bar{\mathbf{x}}^\top\right)^\top\left(\mathbf{X} - \mathbf{1}_N\bar{\mathbf{x}}^\top\right) \\
=& \left(\mathbf{X} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{X}\right)^\top\left(\mathbf{X} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{X}\right) \\
=& \mathbf{X}^\top\left(\mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)^\top\left(\mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)\mathbf{X} \\
=& \mathbf{X}^\top\left(\mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)\mathbf{X},
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

where the last step is because of

$$\begin{aligned}&\left(\mathbf{I}-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)^\top\left(\mathbf{I}-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)=\left(\mathbf{I}-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)^2\\ =&\mathbf{I}-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^\top+\frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{1}_N\mathbf{1}_N^\top\\ =&\mathbf{I}-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^\top+\frac{1}{N^2}\mathbf{1}_N(\mathbf{1}_N^\top\mathbf{1}_N)\mathbf{1}_N^\top\\ =&\mathbf{I}-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^\top+\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\\ =&\mathbf{I}-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top.\end{aligned}$$

Geometrical Interpretation: We denote $\bar{\mathbf{x}}=\begin{bmatrix}\bar{x}_1 & \dots & \bar{x}_p\end{bmatrix}^\top$ and $\mathbf{d}_i=\mathbf{y}_i-\bar{x}_i\mathbf{1}_N\in\mathbb{R}^N$.

1. The projection of $\mathbf{y}_i$ onto the equal angular vector $\mathbf{1}_N$ is the vector $\bar{x}_i\mathbf{1}_N$. Let the projection be $\alpha\mathbf{1}_N$, that satisfies

$$0=(\mathbf{y}_i-\alpha\mathbf{1}_N)^\top\mathbf{1}_N=\mathbf{y}_i^\top\mathbf{1}_N-\alpha\mathbf{1}_N^\top\mathbf{1}_N=\sum_{j=1}^N x_{ji}-N\alpha=N\bar{x}_i-N\alpha.$$

   Hence, we achieve $\alpha=\bar{x}_i$ and the projection is $\bar{x}_i\mathbf{1}_N$ with length $\sqrt{n}|\bar{x}_i|$. This means the magnitude of $\bar{x}_i$ is related to the projection of $\mathbf{y}_i$ on $\mathbf{1}_N$.

2. We have

$$\begin{bmatrix}(\mathbf{x}_1-\bar{\mathbf{x}})^\top\\ \vdots\\ (\mathbf{x}_N-\bar{\mathbf{x}})^\top\end{bmatrix}=\begin{bmatrix}\mathbf{d}_1 & \dots & \mathbf{d}_p\end{bmatrix}\qquad\text{and}\qquad (N-1)\mathbf{S}=\begin{bmatrix}\mathbf{d}_1^\top\\ \vdots\\ \mathbf{d}_p^\top\end{bmatrix}\begin{bmatrix}\mathbf{d}_1 & \dots & \mathbf{d}_p\end{bmatrix},$$

   which means $s_{ij}=(\mathbf{d}_i^\top\mathbf{d}_j)/(N-1)$.

3. We have $r_{ij}=\dfrac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}=\dfrac{\mathbf{d}_i^\top\mathbf{d}_j}{\|\mathbf{d}_i\|_2\|\mathbf{d}_j\|_2}$.

Generalized Variance: We first introduce the following theorem.

Theorem 2.5 (Theorem 7.5.1 of Anderson [1]). *Define* $\mathbf{V}=[\mathbf{v}_1,\dots,\mathbf{v}_p]\in\mathbb{R}^{N\times p}$ *and let*

$$\mathrm{Vol}(\mathbf{v}_1,\dots,\mathbf{v}_p)$$

be the p-dimensional volume of the parallelotope with $\mathbf{v}_1,\dots,\mathbf{v}_p\in\mathbb{R}^N$ *as principal edges* ($N\geq p$)*, then*

$$\big(\mathrm{Vol}(\mathbf{v}_1,\dots,\mathbf{v}_p)\big)^2=\det(\mathbf{V}^\top\mathbf{V}).$$

Applying this theorem by taking $\mathbf{v}_i=\mathbf{d}_i$, we achieve

$$\mathbf{S}=\frac{1}{N-1}\mathbf{V}^\top\mathbf{V}\qquad\text{and}\qquad\big(\mathrm{Vol}(\mathbf{d}_1,\dots,\mathbf{d}_p)\big)^2=(N-1)^p\det(\mathbf{S}).$$

Remark 2.3. *For $p=2$, the p-dimensional volume is area. We allow the area be non-zero even the points lie in 3-dimensional space.*

Some observations:

This document is organized by Luo Luo.
Please do not distribute.

1. The volume will increase when the length of any $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i \mathbf{1}_N$ is increased.
2. The volume will increase if the residual vectors of fixed length are moved until they are at right angles to one another.
3. The volume will be small if just one of the s_{ii} is small or one of the deviation vectors lies nearly in the (hyper) plane formed by the others, or both.

Remark 2.4. *We can also see the view of hyperellipsoid. For unit ball*

$$\mathcal{B}^p = \{\mathbf{z} \in \mathbb{R}^p : \|\mathbf{z}\|_2 = 1\},$$

we have[2]

$$\mathrm{Vol}(\mathcal{B}^p) = \frac{2\pi^{p/2}}{p\Gamma(p/2)}.$$

Let SVD of $\mathbf{S}$ *be* $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$*. Since translate and rotation do not change the volume, we have*

$$\begin{aligned} V =& \mathrm{Vol}\left(\{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2\}\right) \\ =& \mathrm{Vol}\left(\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^\top \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top \mathbf{x} = c^2\}\right) \\ =& \mathrm{Vol}\left(\{\mathbf{y} \in \mathbb{R}^p : \mathbf{y}^\top \boldsymbol{\Lambda}^{-1} \mathbf{y} = c^2\}\right) \quad //\mathbf{y} = \mathbf{U}^\top \mathbf{x} \\ =& \mathrm{Vol}\left(\{\mathbf{y} \in \mathbb{R}^p : \mathbf{y}^\top (c^{-2}\boldsymbol{\Lambda}^{-1}) \mathbf{y} = 1\}\right). \end{aligned}$$

Let $y_i = c\sqrt{\lambda_i} z_i$*, then* $\mathbf{z} = [z_1, \dots, z_p]^\top$ *satisfies* $\mathbf{z}^\top \mathbf{z} = 1$ *and we have*

$$\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{z} = 1\} = \mathrm{Vol}(\mathcal{B}^p) = \frac{2\pi^{p/2}}{p\Gamma(p/2)}.$$

This implies

$$V = \mathrm{Vol}(\mathcal{B}^p) \left(\prod_{i=1}^p c\sqrt{\lambda_i} \right) = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \cdot c^p (\det(\mathbf{S}))^{1/2}.$$

Remark 2.5. *We can show that if* $N \leq p$*, then* $\det(\mathbf{S}) = 0$*. Note that*

$$\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^\top \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^\top \end{bmatrix} \in \mathbb{R}^{N \times p}$$

and

$$(N-1)\mathbf{S} = \sum_{\alpha=1}^N (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = \left(\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top\right)^\top \left(\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top\right),$$

which means $\mathrm{rank}(\mathbf{S}) \leq \mathrm{rank}(\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top) \leq N$*.*

- *If* $N < p$*, then* $\mathrm{rank}(\mathbf{S}) \leq N < p$*.*
- *If* $N = p$*, note that the sum of rows of* $\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top$ *is zero, which means this square matrix is not full rank, that is* $\mathrm{rank}(\mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^\top) < p$*. Hence, the matrix* $\mathbf{S}$ *is also not full rank.*

This document is organized by Luo Luo.
Please do not distribute.

[2] https://en.wikipedia.org/wiki/Volume_of_an_n-ball

3 Multivariate Normal Distribution

Theorem 3.1. *Suppose the p-dimensional random vector* $\mathbf{x}$ *has the density function*

$$f(\mathbf{x}) = K \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b})\right),$$

where $K \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^p$ *and* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *is symmetric positive definite. Then we have*

$$K = \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}}, \qquad \mathbf{b} = \boldsymbol{\mu} \qquad \textit{and} \qquad \mathbf{A} = \mathbf{\Sigma}^{-1}.$$

Proof. Let the spectral decomposition of $\mathbf{A} \in \mathbb{R}^{p\times p}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, then we take $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}^{-1/2}$ and it satisfies

$$\mathbf{C}^\top\mathbf{A}\mathbf{C} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^\top(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top)\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{I}$$

and $\mathbf{C}$ is non-singular. Define $\mathbf{y} = \mathbf{C}^{-1}(\mathbf{x}-\mathbf{b})$. We consider K, $\mathbf{b}$ and $\mathbf{A}$, respectively.

Part I: We have $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b}$ and

$$\begin{aligned}(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b}) =& (\mathbf{C}\mathbf{y}+\mathbf{b}-\mathbf{b})^\top \mathbf{A}(\mathbf{C}\mathbf{y}+\mathbf{b}-\mathbf{b}) \\ =& \mathbf{y}\mathbf{C}^\top\mathbf{A}\mathbf{C}\mathbf{y} = \mathbf{y}\mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{y} = \mathbf{y}^\top\mathbf{y}.\end{aligned}$$

Recall that for $y \sim \mathcal{N}(0,1)$, we have

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \mathrm{d}y = 1, \qquad \int_{-\infty}^{+\infty} y \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \mathrm{d}y = 0,$$
$$\text{and} \qquad \int_{-\infty}^{+\infty} y^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \mathrm{d}y = 1.$$

Thus, we have

$$\begin{aligned}1 =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} K\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top\mathbf{A}(\mathbf{x}-\mathbf{b})\right)\mathrm{d}x_1\dots\mathrm{d}x_p \\ =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} K\det(\mathbf{C})\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\right)\mathrm{d}y_1\dots\mathrm{d}y_p \\ =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} K\det(\mathbf{C})\exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right)\mathrm{d}y_1\dots\mathrm{d}y_p \\ =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} K\det(\mathbf{A}^{-\frac{1}{2}})\exp\left(-\frac{1}{2}y_p^2\right)\dots\exp\left(-\frac{1}{2}y_1^2\right)\mathrm{d}y_1\dots\mathrm{d}y_p \\ =& K\det(\mathbf{A}^{-\frac{1}{2}})(2\pi)^{\frac{p}{2}},\end{aligned}$$

which means

$$K = \frac{1}{\det(\mathbf{A}^{-\frac{1}{2}})(2\pi)^{\frac{p}{2}}} = \frac{\det(\mathbf{A}^{1/2})}{\sqrt{(2\pi)^p}}.$$

Part II: Directly consider the expectation and variance of $\mathbf{x}$ is not easy, so we first consider the ones of $\mathbf{y}$. The relation $\mathbf{y} = \mathbf{C}^{-1}(\mathbf{x}-\mathbf{b})$ means

$$\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b} \qquad \text{and} \qquad \mathbb{E}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}] + \mathbf{b}.$$

This document is organized by Luo Luo.
Please do not distribute.

The transformation implies the density function of $\mathbf{y}$ is

$$\begin{aligned}g(\mathbf{y}) =&K\det(\mathbf{C})\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\right) = \frac{\det(\mathbf{A}^{1/2})\det(\mathbf{C})}{\sqrt{(2\pi)^p}}\exp\left(-\frac{1}{2}\sum_{j=1}^p y_j^2\right)\\ =&\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\sum_{j=1}^p y_j^2\right) = \prod_{j=1}^p \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_j^2\right),\end{aligned}$$

where we use the fact

$$\begin{aligned}&\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top = (\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top)^{-1} = (\mathbf{C}\mathbf{C}^\top)^{-1}\\ \Longrightarrow&\det(\mathbf{A}) = \det((\mathbf{C}\mathbf{C}^\top)^{-1})\\ \Longrightarrow&\det(\mathbf{A})\det((\mathbf{C}\mathbf{C}^\top)) = 1\\ \Longrightarrow&\det(\mathbf{A}^{1/2})\det(\mathbf{C}) = 1\end{aligned}$$

Then for each $i = 1, \dots, p$, we have

$$\begin{aligned}\mathbb{E}[y_i] =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} y_i \prod_{j=1}^p \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_j^2\right)\,\mathrm{d}y_1\dots\,\mathrm{d}y_p\\ =& \left(\int_{-\infty}^{+\infty} y_i\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_i^2\right)\,\mathrm{d}y_i\right)\prod_{j=1,i\neq j}^p\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_j^2\right)\,\mathrm{d}y_j = 0,\end{aligned}$$

where the last step is because of

$$\int_{-\infty}^{+\infty} y_i\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_i^2\right)\,\mathrm{d}y_i = 0.$$

Thus, we have $\mathbb{E}[\mathbf{y}] = \mathbf{0}$ and

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

Part III: The relation $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b}$ means

$$\mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathrm{Cov}[\mathbf{y}]\mathbf{C}^\top.$$

Now we need to check the (i, j)-th element of $\mathrm{Cov}[\mathbf{y}]$, that is

$$\mathbb{E}[(y_i - \mathbb{E}[y_i])(y_j - \mathbb{E}[y_j])] = \mathbb{E}[y_iy_j],$$

where we use the fact $\mathbb{E}[\mathbf{y}] = \mathbf{0}$. For each $i \neq j$, we have

$$\begin{aligned}&\mathbb{E}[y_iy_j]\\ =&\frac{1}{(2\pi)^{p/2}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} y_iy_j\exp\left(-\frac{1}{2}\sum_{h=1}^p y_h^2\right)\,\mathrm{d}y_1\dots\mathrm{d}y_p\\ =&\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y_i\exp\left(-\frac{1}{2}y_i^2\right)\,\mathrm{d}y_i\right)\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y_j\exp\left(-\frac{1}{2}y_j^2\right)\mathrm{d}y_j\right)\\ &\cdot\prod_{j=1,h\neq i,j}^p\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}y_h^2\right)\,\mathrm{d}y_h\\ =&0.\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

We also have

$$
\begin{aligned}
&\mathbb{E}[y_i^2] \\
=&\frac{1}{(2\pi)^{p/2}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} y_i^2 \exp\left(-\frac{1}{2}\sum_{h=1}^{p} y_h^2\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\
=&\left(\int_{-\infty}^{+\infty} y_i^2\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_i^2\right)\,\mathrm{d}y_i\right)\prod_{j=1,h\neq i}^{p}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_h^2\right)\,\mathrm{d}y_h = 1,
\end{aligned}
$$

where the last step is due to

$$
\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}y_h^2\right)\,\mathrm{d}y_h
$$

corresponds to the pdf of $y_h \sim \mathcal{N}(0,1)$ and

$$
\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}y_h^2\exp\left(-\frac{1}{2}y_h^2\right)\,\mathrm{d}y_i
$$

corresponds to the variance of $y_i \sim \mathcal{N}(0,1)$. Hence, it holds that

$$
\mathbb{E}[(y_i-\mathbb{E}[y_i])(y_j-\mathbb{E}[y_j])] = \begin{cases}0, & i\neq j,\\ 1, & i=j.\end{cases} \implies \quad \mathrm{Cov}[\mathbf{y}] = \mathbf{I},
$$

which implies $\mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathrm{Cov}[\mathbf{y}]\mathbf{C}^\top = \mathbf{C}\mathbf{C}^\top$.
Since it holds

$$
\mathbf{C}^\top\mathbf{A}\mathbf{C} = (\mathbf{U}\mathbf{\Lambda}^{-1/2})^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}(\mathbf{U}\mathbf{\Lambda}^{-1/2}) = (\mathbf{\Lambda}^{-1/2})^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{I},
$$

we obtain

$$
(\mathbf{C}^\top\mathbf{A}\mathbf{C})^{-1} = \mathbf{I} \implies \mathbf{C}^{-1}\mathbf{A}^{-1}(\mathbf{C}^\top)^{-1} = \mathbf{I} \implies \mathbf{A}^{-1} = \mathbf{C}\mathbf{C}^\top.
$$

Hence, $\mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathbf{C}^\top = \mathbf{A}^{-1}$. □

This document is organized by Luo Luo.
Please do not distribute.

Bivariate Normal Distribution: Consider the bivariate normal distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where

$$
\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \text{and} \qquad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.
$$

We have

$$
\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix},
$$

which can be verifies as follows

$$
\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{22}\sigma_{11}-\sigma_{12}^2 & \sigma_{22}\sigma_{12}-\sigma_{12}\sigma_{22} \\ -\sigma_{12}\sigma_{11}+\sigma_{11}\sigma_{12} & -\sigma_{12}^2+\sigma_{11}\sigma_{22} \end{bmatrix} = (\sigma_{11}\sigma_{22}-\sigma_{12}^2)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
$$

Let $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, then we have

$$
\begin{aligned}
\det(\mathbf{\Sigma}) &= \sigma_{11}\sigma_{22}-\sigma_{12}^2 = \sigma_{11}\sigma_{22}\left(1-\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right) = \sigma_{11}\sigma_{22}\left(1-\rho^2\right), \\
\mathbf{\Sigma}^{-1} &= \frac{1}{\sigma_{11}\sigma_{22}\left(1-\rho^2\right)}\begin{bmatrix} \sigma_{22} & -\rho\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{bmatrix},
\end{aligned}
$$

and

$$\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
=&\frac{\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_{22} & -\rho\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}}{\sigma_{11}\sigma_{22}\,(1-\rho^2)} \\
=&\frac{\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_{22}(x_1-\mu_1)-\rho\sqrt{\sigma_{11}\sigma_{22}}\,(x_2-\mu_2) \\ -\rho\sqrt{\sigma_{11}\sigma_{22}}\,(x_1-\mu_1)+\sigma_{11}(x_2-\mu_2) \end{bmatrix}}{\sigma_{11}\sigma_{22}\,(1-\rho^2)} \\
=&\frac{\sigma_{22}(x_1-\mu_1)^2 - 2\rho\sqrt{\sigma_{11}\sigma_{22}}\,(x_1-\mu_1)(x_2-\mu_2) + \sigma_{11}(x_2-\mu_2)^2}{\sigma_{11}\sigma_{22}\,(1-\rho^2)} \\
=&\frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}}\right).
\end{aligned}$$

Hence, the density function is

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\,(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}}\right)\right).$$

If $\rho = 0$, then

$$\begin{aligned}
f(x_1,x_2) =& \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_{11}} - \frac{(x_2-\mu_2)^2}{2\sigma_{22}}\right) \\
=& \underbrace{\frac{1}{\sqrt{2\pi\sigma_{11}}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_{11}}\right)}_{f_1(x_1)} \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_{22}}\right)}_{f_2(x_2)},
\end{aligned}$$

which means x_1 and x_2 are independent since $f_1(\cdot)$ and $f_2(\cdot)$ are the density of x_1 and x_2, respectively.

Example 3.1. *Consider bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} > 0$. The eigenvalue λ of the covariance matrix $\boldsymbol{\Sigma}$ satisfies $\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0$, that is*

$$0 = \det\left(\begin{bmatrix} \sigma_{11}-\lambda & \sigma_{12} \\ \sigma_{12} & \sigma_{11}-\lambda \end{bmatrix}\right) = (\sigma_{11}-\lambda)^2 - \sigma_{12}^2 = (\sigma_{11}-\lambda+\sigma_{12})(\sigma_{11}-\lambda-\sigma_{12}).$$

This implies $\lambda_1 = \sigma_{11} + \sigma_{12}$ and $\lambda_2 = \sigma_{11} - \sigma_{12}$. For $\lambda_1 = \sigma_{11} + \sigma_{12}$, we have

$$\begin{aligned}
\boldsymbol{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \iff& \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = (\sigma_{11}+\sigma_{12}) \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} \\
\iff& \begin{bmatrix} \sigma_{11}u_{11} + \sigma_{12}u_{12} \\ \sigma_{12}u_{11} + \sigma_{11}u_{12} \end{bmatrix} = \begin{bmatrix} (\sigma_{11}+\sigma_{12})u_{11} \\ (\sigma_{11}+\sigma_{12})u_{12} \end{bmatrix} \\
\iff& u_1 = u_2.
\end{aligned}$$

By normalizing on $\mathbf{u}_1$, we obtain $\mathbf{u}_1 = \begin{bmatrix} 1/\sqrt{2}, 1/\sqrt{2} \end{bmatrix}^\top$. Similarly, we have $\mathbf{u}_2 = \begin{bmatrix} -1/\sqrt{2}, 1/\sqrt{2} \end{bmatrix}^\top$. Hence, the vertices of the ellipsoid are

$$\pm c\sqrt{\sigma_{11}+\sigma_{12}} \begin{bmatrix} \dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} \end{bmatrix} \qquad \textit{and} \qquad \pm\, c\sqrt{\sigma_{11}-\sigma_{12}} \begin{bmatrix} -\dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} \end{bmatrix}.$$

Remark 3.1. *We consider the case of $\boldsymbol{\mu} = \mathbf{0}$.*

This document is organized by Luo Lao. Please do not distribute.

1. *For $\mathbf{\Sigma} = \mathbf{I}$, it is a ball with radius c.*
2. *For $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$, it is an ellipsoid with vertices $\pm[c\sqrt{\lambda_1}, \dots, 0]^\top, \dots, \pm[0, \dots, c\sqrt{\lambda_p}\,]^\top$.*
3. *For $\mathbf{\Sigma}$ with SVD of the form $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, it is the rotation of the second case. Consider the ellipsoid*

   $$\{\mathbf{y} : \mathbf{y}^\top \mathbf{\Lambda}^{-1}\mathbf{y} = c^2\}.$$

   The ellipsoid

   $$\{\mathbf{x} : \mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x} = c^2\}$$

   corresponds to the transform $\mathbf{x} = \mathbf{U}\mathbf{y}$. Hence, we have

   $$\pm\mathbf{U}[c\sqrt{\lambda_1}, \dots, 0]^\top = \pm c\sqrt{\lambda_1}\mathbf{u}_1 \quad , \dots , \quad \pm\mathbf{U}[0, \dots, c\sqrt{\lambda_p}\,]^\top = \pm c\sqrt{\lambda_p}\mathbf{u}_p.$$

In above example, we have

$$\mathbf{U} = \begin{bmatrix} \dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \\ \sin\theta & \cos\theta \end{bmatrix}$$

with $\theta = \pi/4$, which is the rotation matrix ($\pi/4$ counterclockwise)

Theorem 3.2 (Linear Transformation)**.** *Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ and $\mathbf{\Sigma} \succ \mathbf{0}$. Then $\mathbf{y} = \mathbf{C}\mathbf{x}$ is distributed according to $\mathcal{N}_p(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top)$ for non-singular $\mathbf{C} \in \mathbb{R}^{p\times p}$.*

Proof. Let $f(x)$ be the density of $\mathbf{x}$ such that

$$f(\mathbf{x}) = n(\boldsymbol{\mu} \mid \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

and $g(\mathbf{y})$ be the density function of $\mathbf{y}$. The relation $\mathbf{x} = \mathbf{C}^{-1}\mathbf{y}$ implies

$$g(\mathbf{y}) = f(\mathbf{u}^{-1}(\mathbf{y}))|\det(\mathbf{J}^{-1}(\mathbf{y}))|$$

with

$$\mathbf{u}(\mathbf{x}) = \mathbf{C}\mathbf{x}, \quad \mathbf{u}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}\mathbf{y} \quad \text{and} \quad \mathbf{J}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}.$$

Hence, we have

$$\begin{aligned}
&g(\mathbf{y}) \\
=&f(\mathbf{C}^{-1}\mathbf{y})|\det(\mathbf{C}^{-1})| \\
=&\frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{C}^{-1}\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{C}^{-1}\mathbf{y}-\boldsymbol{\mu})\right)|\det(\mathbf{C}^{-1})| \\
=&\frac{|\det(\mathbf{C}^{-1})|}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu})^\top \mathbf{C}^{-\top}\mathbf{\Sigma}^{-1}\mathbf{C}^{-1}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu})\right) \\
=&\frac{1}{\sqrt{(2\pi)^p \det(\mathbf{C}\mathbf{\Sigma}^{-1}\mathbf{C}^\top)}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu})^\top \left(\mathbf{C}\mathbf{\Sigma}^{-1}\mathbf{C}^\top\right)^{-1}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu})\right) \\
=&n(\mathbf{C}\mu \mid \mathbf{C}\mathbf{\Sigma}^{-1}\mathbf{C}^\top),
\end{aligned}$$

where we use the fact

$$\frac{|\det(\mathbf{C}^{-1})|}{\sqrt{\det(\mathbf{\Sigma})}} = \frac{1}{\sqrt{|\det(\mathbf{C})|^2\det(\mathbf{\Sigma})}} = \frac{1}{\sqrt{|\det(\mathbf{C})|\det(\mathbf{\Sigma})|\det(\mathbf{C}^\top)|}} = \frac{1}{\sqrt{|\det(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top)|}}.$$

□

This document is organized by Luo Luo.
Please do not distribute.

Theorem 3.3 (Independence)**.** *If $\mathbf{x} = [x_1, \ldots, x_p]^\top$ have a joint normal distribution. Let*

1. $\mathbf{x}^{(1)} = [x_1, \ldots, x_q]^\top$,
2. $\mathbf{x}^{(2)} = [x_{q+1}, \ldots, x_p]^\top$.

for $q < p$. A necessary and sufficient condition for $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ to be independent is that each covariance of a variable from $\mathbf{x}^{(1)}$ and a variable from $\mathbf{x}^{(2)}$ is 0.

Proof. Let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \text{where} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

such that

- $\boldsymbol{\mu}^{(1)} = \mathbb{E}\left[\mathbf{x}^{(1)}\right]$,
- $\boldsymbol{\mu}^{(2)} = \mathbb{E}\left[\mathbf{x}^{(2)}\right]$,
- $\boldsymbol{\Sigma}_{11} = \mathbb{E}\left[\left(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}\right)\left(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}\right)^\top\right]$,
- $\boldsymbol{\Sigma}_{22} = \mathbb{E}\left[\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)^\top\right]$,
- $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top = \mathbb{E}\left[\left(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}\right)\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)^\top\right]$.

Sufficiency (uncorrelated $\Longrightarrow$ independent): The random vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are uncorrelated means

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix}.$$

The quadratic form of $n(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\begin{aligned}
&(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\
=&\begin{bmatrix} (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})^\top & (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)} \\ \mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)} \end{bmatrix} \\
=&\begin{bmatrix} (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})^\top & (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}) \\ \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}) \end{bmatrix} \\
=&(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}) + (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})
\end{aligned}$$

and we have $\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11}) \det(\boldsymbol{\Sigma}_{22})$.
Then we have

$$\begin{aligned}
&n(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \\
=&\frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \\
=&\frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11})}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})\right) \\
&\cdot \frac{1}{\sqrt{(2\pi)^{p-q} \det(\boldsymbol{\Sigma}_{22})}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})\right) \\
=&n(\boldsymbol{\mu}^{(1)} \mid \boldsymbol{\Sigma}_{11}) n(\boldsymbol{\mu}^{(2)} \mid \boldsymbol{\Sigma}_{22}),
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

which proves $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent.

Additionally, this result means the marginal distribution of variable $\mathbf{x}^{(1)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11})$ since

$$\begin{aligned}
f_1(\mathbf{x}^{(1)}) =& \int_{\mathbf{x}^{(2)} \in \mathbb{R}^{(p-q)}} n(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \, \mathrm{d}\mathbf{x}^{(2)} \\
=& \int_{\mathbf{x}^{(2)} \in \mathbb{R}^{(p-q)}} n(\boldsymbol{\mu}^{(1)} \mid \boldsymbol{\Sigma}_{11}) n(\boldsymbol{\mu}^{(2)} \mid \boldsymbol{\Sigma}_{22}) \, \mathrm{d}\mathbf{x}^{(2)} \\
=& n(\boldsymbol{\mu}^{(1)} \mid \boldsymbol{\Sigma}_{11}) \int_{\mathbf{x}^{(2)} \in \mathbb{R}^{(p-q)}} n(\boldsymbol{\mu}^{(2)} \mid \boldsymbol{\Sigma}_{22}) \, \mathrm{d}\mathbf{x}^{(2)} \\
=& n(\boldsymbol{\mu}^{(1)} \mid \boldsymbol{\Sigma}_{11}).
\end{aligned}$$

Similarly, the marginal distribution of $\mathbf{x}^{(2)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$.

Necessity (independent $\Longrightarrow$ uncorrelated): This is trivial.

□

Remark 3.2. *Note that normally distributed and uncorrelated assumptions do not imply the variables are independent. Suppose the random variable x has standard normal distribution such that $x \sim \mathcal{N}(0, 1)$. Let w has the Rademacher distribution such that*

$$\Pr(w = 1) = \Pr(w = -1) = \frac{1}{2},$$

and assume w is independent of x. Let $y = wx$, then we can show that y is a normal distributed variable; x and y are uncorrelated; x and y are not independent.

1. *Let $\Phi(\cdot)$ be the cumulative distribution function (CFD) of standard normal distribution such that*

   $$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}x^2\right).$$

   Then we obtain the CFD of y as follows

   $$\begin{aligned}
   &\Pr(y \le z) = \Pr(wx \le z) = \Pr(x \le z) \Pr(w = 1) + \Pr(-x \le z) \Pr(w = -1) \\
   =& \frac{1}{2} \Pr(x \le z) + \frac{1}{2} \Pr(-x \le z) = \frac{1}{2}\Phi(z) + \frac{1}{2}\Phi(z) = \Phi(z),
   \end{aligned}$$

   where we use the fact $\Pr(x \le z) = \Pr(-x \le z) = \Phi(z)$ since both x and $-x$ has standard normal distribution. This means $y \sim \mathcal{N}(0, 1)$.

2. *We have*

   $$\begin{aligned}
   \mathrm{Cov}[x, y] =& \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[xy] - 0 \cdot \mathbb{E}[y] = \mathbb{E}[xy] \\
   =& \mathbb{E}[wx^2] = \mathbb{E}[w]\mathbb{E}[x^2] = 0 \cdot \mathbb{E}[x^2] = 0,
   \end{aligned}$$

   which means x and y are uncorrelated.

3. *We have shown both $x \sim \mathcal{N}(0, 1)$ and $y \sim \mathcal{N}(0, 1)$. If they are independent, then $[x, y]^\top \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I})$. However, the definitions means $|x| = |y|$, which leads to $[x, y]^\top \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I})$ does not hold.*

Corollary 3.1. *We use the notation in Theorem 3.3 such that*

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

It shows that if $\mathbf{x}^{(1)}$ is uncorrelated with $\mathbf{x}^{(2)}$, the marginal distribution of $\mathbf{x}^{(1)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11})$ and the marginal distribution of $\mathbf{x}^{(2)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$.

This document is organized by Luo Luo.
Please do not distribute.

Proof. We have already shown $f_1(\mathbf{x}^{(1)}) = n(\boldsymbol{\mu}^{(1)} \mid \boldsymbol{\Sigma}_{11})$ in the proof of Theorem 3.3. □

Theorem 3.4. *If $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$, the marginal distribution of any set of components of $\mathbf{x}$ is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.*

Proof. We shall make a non-singular linear transformation $\mathbf{B} \in \mathbb{R}^{p \times p}$ to subvectors

$$\mathbf{y}^{(1)} = \mathbf{x}^{(1)} + \mathbf{B}\mathbf{x}^{(2)} \qquad \text{and} \qquad \mathbf{y}^{(2)} = \mathbf{x}^{(2)},$$

leading to the components of $\mathbf{y}^{(1)}$ are uncorrelated with the ones of $\mathbf{y}^{(2)}$. The matrix $\mathbf{B}$ should satisfy

$$\begin{aligned}
\mathbf{0} =& \mathbb{E}\left[\big(\mathbf{y}^{(1)} - \mathbb{E}[\mathbf{y}^{(1)}]\big)\big(\mathbf{y}^{(2)} - \mathbb{E}[\mathbf{y}^{(2)}]\big)^\top\right] \\
=& \mathbb{E}\left[\big(\mathbf{x}^{(1)} + \mathbf{B}\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(1)} + \mathbf{B}\mathbf{x}^{(2)}]\big)\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)^\top\right] \\
=& \mathbb{E}\left[\big(\mathbf{x}^{(1)} - \mathbb{E}[\mathbf{x}^{(1)}] + \mathbf{B}\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)\big)\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)^\top\right] \\
=& \mathbb{E}\left[\big(\mathbf{x}^{(1)} - \mathbb{E}[\mathbf{x}^{(1)}]\big)\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)^\top\right] + \mathbf{B} \cdot \mathbb{E}\left[\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)\big(\mathbf{x}^{(2)} - \mathbb{E}[\mathbf{x}^{(2)}]\big)^\top\right] \\
=& \boldsymbol{\Sigma}_{12} + \mathbf{B}\boldsymbol{\Sigma}_{22}.
\end{aligned}$$

Thus $\mathbf{B} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ and $\mathbf{y}^{(1)} = \mathbf{x}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{(2)}$. The vector

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{x}$$

is a non-singular transform of $\mathbf{x}$, and therefore Theorem 3.2 means $\mathbf{y}$ has a normal distribution with

$$\mathbb{E}\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

and

$$\begin{aligned}
\operatorname{Cov}\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} &= \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}
\end{aligned}$$

Thus, we have shown $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)} = \mathbf{x}^{(2)}$ are independent, which implies the marginal distribution of $\mathbf{x}^{(2)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$ by using Corollary 3.1. Because the numbering of the components of $\mathbf{x}$ is arbitrary, we have proved this theorem. □

This document is organized by Luo Luo.
Please do not distribute.

Theorem 3.5. *Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} \succ \mathbf{0}$. Then the q-dimensional random vector*

$$\mathbf{z} = \mathbf{D}\mathbf{x}$$

is distributed according to $\mathcal{N}_q(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$ for $\mathbf{D} \in \mathbb{R}^{q\times p}$ of rank $q \leq p$.

Proof. For any transformation $\mathbf{z} = \mathbf{D}\mathbf{x}$, for $\mathbf{D} \in \mathbb{R}^{q\times p}$ and p-dimensional random vector $\mathbf{x}$, we have

$$\mathbb{E}[\mathbf{z}] = \mathbf{D}\boldsymbol{\mu} \qquad \text{and} \qquad \mathrm{Cov}[\mathbf{z}] = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top.$$

For $q = p$, the result is the same as Theorem 3.2 by taking $\mathbf{C} = \mathbf{D}$. For $q < p$, we can find a matrix $\mathbf{E} \in \mathbb{R}^{(p-q)\times p}$ such that

$$\begin{bmatrix}\mathbf{z}\\ \mathbf{w}\end{bmatrix} = \begin{bmatrix}\mathbf{D}\\ \mathbf{E}\end{bmatrix}\mathbf{x}, \quad \text{where } \begin{bmatrix}\mathbf{D}\\ \mathbf{E}\end{bmatrix} \in \mathbb{R}^{p\times p} \text{ is non-singular and } \mathbf{w} \text{ is a } (p-q)\text{-dimensional random vector.}$$

Applying Theorem 3.2 by taking

$$\mathbf{C} = \begin{bmatrix}\mathbf{D}\\ \mathbf{E}\end{bmatrix}$$

indicates $\mathbf{z}$ and $\mathbf{w}$ have a joint normal distribution such that

$$\begin{bmatrix}\mathbf{z}\\ \mathbf{w}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{D}\boldsymbol{\mu}\\ \mathbf{E}\boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top & \mathbf{D}\boldsymbol{\Sigma}\mathbf{E}^\top\\ \mathbf{E}\boldsymbol{\Sigma}\mathbf{D}^\top & \mathbf{E}\boldsymbol{\Sigma}\mathbf{E}^\top\end{bmatrix}\right),$$

and its marginal normal distribution of $\mathbf{z}$ is $\mathcal{N}_q(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$. □

Singular normal distribution We show why the singular normal distribution has no density function. Suppose the mass of p-dimensional singular normal distribution concentrated on a given lower dimensional set $\mathcal{S}$ with $\dim(\mathcal{S}) < p$, then the measure of $\mathcal{S}$ is zero. Hence, we have $f(\mathbf{x}) = 0$ almost everywhere, which means

$$\int_{\mathbb{R}^p} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0.$$

It contradicts to the property of density function such that

$$\int_{\mathbb{R}^p} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1.$$

Remark 3.3. *There is no density if the measure is defined on $\mathbb{R}^p$. However, we can define a restriction of Lebesgue measure to the $\mathrm{rank}(\boldsymbol{\Sigma})$-dimensional affine subspace of $\mathbb{R}^p$, i.e. $\{\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{v} : \mathbf{v} \in \mathbb{R}^p\}$. With respect to this measure, the distribution has the density*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det^\dagger(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^\dagger(\mathbf{x}-\boldsymbol{\mu})\right),$$

where $k = \mathrm{rank}(\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}^\dagger$ is the pseudo inverse of $\boldsymbol{\Sigma}$ and $\det^\dagger(\boldsymbol{\Sigma})$ is the pseudo-determinant of $\boldsymbol{\Sigma}$ which is defined as the product of the non-zero singular values of $\boldsymbol{\Sigma}$.

Remark 3.4. *In general, suppose that $\mathbf{y} \sim \mathcal{N}_q(\boldsymbol{\nu}, \mathbf{T})$, $\boldsymbol{\lambda} \in \mathbb{R}^p$, and $\mathbf{A} \in \mathbb{R}^{p\times q}$ with $\mathbf{T} \succeq \mathbf{0}$ and $p > q$; then we say that*

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\lambda}$$

has a singular (degenerate) normal distribution in $\mathbb{R}^p$ space. Let $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}]$, then

$$\boldsymbol{\mu} = \mathbf{A}\mathbb{E}[\mathbf{y}] + \boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\nu} + \boldsymbol{\lambda} \quad \textit{and} \quad \boldsymbol{\Sigma} = \mathbf{A}\mathrm{Cov}[\mathbf{y}]\mathbf{A}^\top = \mathbf{A}\mathbf{T}\mathbf{A}^\top.$$

The matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ must be singular, since $\mathrm{rank}(\boldsymbol{\Sigma}) \leq \mathrm{rank}(\mathbf{A}) \leq q < p$.

This document is organized by Luo Luo.
Please do not distribute.

Theorem 3.6. *Let* $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then*

$$\mathbf{z} = \mathbf{D}\mathbf{x}$$

is distributed according to $\mathcal{N}_q(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$ *for any* $\mathbf{D} \in \mathbb{R}^{q\times p}$.

Proof. It is easy to verify $\mathbb{E}[\mathbf{z}] = \mathbf{D}\boldsymbol{\mu}$ and $\mathrm{Cov}[\mathbf{z}] = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top$. Hence, we only need to show $\mathbf{z}$ follows normal distribution.

Since $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it can be presented as

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\lambda}$$

where $\mathbf{A} \in \mathbb{R}^{p\times r}$, r is the rank of $\boldsymbol{\Sigma}$ and $\mathbf{y} \sim \mathcal{N}_r(\boldsymbol{\nu}, \mathbf{T})$ with $\boldsymbol{\nu} \in \mathbb{R}^r$ and non-singular $\mathbf{T} \in \mathbb{R}^{r\times r}$. The relationship $\mathbf{z} = \mathbf{D}\mathbf{x}$ means we can write

$$\mathbf{z} = \mathbf{D}\mathbf{A}\mathbf{y} + \mathbf{D}\boldsymbol{\lambda}, \tag{3}$$

where $\mathbf{D}\mathbf{A} \in \mathbb{R}^{q\times r}$. We consider the folllowing two cases.

1. If the rank of $\mathbf{D}\mathbf{A}$ is r, the formal definition of a normal distribution that conclude the singular distribution implies $\mathbf{z}$ follows normal distribution.

2. If the rank of $\mathbf{D}\mathbf{A}$ is less than r, say s, then

$$\mathbf{E} = \mathrm{Cov}[\mathbf{z}] = \mathbf{D}\mathbf{A}\mathrm{Cov}[\mathbf{y}]\mathbf{A}^\top\mathbf{D}^\top = \mathbf{D}\mathbf{A}\mathbf{T}\mathbf{A}^\top\mathbf{D}^\top \in \mathbb{R}^{q\times q} \tag{4}$$

   is rank of s. We desire $\mathbf{z}$ has the form of

$$\mathbf{z} = \mathbf{G}_1\mathbf{u}_1 + \tilde{\boldsymbol{\lambda}},$$

   with $\mathbf{G}_1 \in \mathbb{R}^{q\times s}$, $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^s$ and $\mathbf{u}_1$ has s-dimensional non-singular normal distribution.

   Note that there exists a non-singular matrix $\mathbf{F} \in \mathbb{R}^{q\times q}$ such that (apply SVD on $\mathbf{E}$ to find such $\mathbf{F}$)

$$\mathbf{F}\mathbf{E}\mathbf{F}^\top = \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{q\times q}. \tag{5}$$

   We partition

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{bmatrix} \in \mathbb{R}^{q\times q} \tag{6}$$

   with $\mathbf{F}_1 \in \mathbb{R}^{s\times q}$ and $\mathbf{F}_2 \in \mathbb{R}^{(q-s)\times q}$, then

$$\begin{aligned} \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \overset{(5)}{=} \mathbf{F}\mathbf{E}\mathbf{F}^\top \overset{(6)}{=} \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{bmatrix} \mathbf{E} \begin{bmatrix} \mathbf{F}_1^\top & \mathbf{F}_2^\top \end{bmatrix} &= \begin{bmatrix} \mathbf{F}_1\mathbf{E}\mathbf{F}_1^\top & \mathbf{F}_1\mathbf{E}\mathbf{F}_2^\top \\ \mathbf{F}_2\mathbf{E}\mathbf{F}_1^\top & \mathbf{F}_2\mathbf{E}\mathbf{F}_2^\top \end{bmatrix} \\ &\overset{(4)}{=} \begin{bmatrix} (\mathbf{F}_1\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_1\mathbf{D}\mathbf{A})^\top & (\mathbf{F}_1\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_2\mathbf{D}\mathbf{A})^\top \\ (\mathbf{F}_2\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_1\mathbf{D}\mathbf{A})^\top & (\mathbf{F}_2\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_2\mathbf{D}\mathbf{A})^\top \end{bmatrix}. \end{aligned}$$

   Thus, the fact $(\mathbf{F}_1\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_1\mathbf{D}\mathbf{A})^\top = \mathbf{I}_s$ means $\mathrm{rank}(\mathbf{F}_1\mathbf{D}\mathbf{A}) = s$, then matrix $\mathbf{T}$ is positive-definite means

$$\mathbf{F}_2\mathbf{D}\mathbf{A} = \mathbf{0} \tag{7}$$

   since $(\mathbf{F}_2\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_2\mathbf{D}\mathbf{A})^\top = \mathbf{0}$.

   Hence, we have

$$\mathbf{F}\mathbf{z} \overset{(3)}{=} \mathbf{F}(\mathbf{D}\mathbf{A}\mathbf{y} + \mathbf{D}\boldsymbol{\lambda}) \overset{(6)}{=} \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{bmatrix} \mathbf{D}\mathbf{A}\mathbf{y} + \mathbf{F}\mathbf{D}\boldsymbol{\lambda} = \begin{bmatrix} \mathbf{F}_1\mathbf{D}\mathbf{A}\mathbf{y} \\ \mathbf{F}_2\mathbf{D}\mathbf{A}\mathbf{y} \end{bmatrix} + \mathbf{F}\mathbf{D}\boldsymbol{\lambda} \overset{(7)}{=} \begin{bmatrix} \mathbf{F}_1\mathbf{D}\mathbf{A}\mathbf{y} \\ \mathbf{0} \end{bmatrix} + \mathbf{F}\mathbf{D}\boldsymbol{\lambda}.$$

This document is organized by Luo Luo.
Please don't distribute.

Let $\mathbf{u}_1 = \mathbf{F}_1\mathbf{D}\mathbf{A}\mathbf{y} \in \mathbb{R}^s$. Since random vector y has non-singular normal distribution and the rank of matrix $\mathbf{F}_1\mathbf{D}\mathbf{A} \in \mathbb{R}^{s\times r}$ is $s \leq r$, we conclude $\mathbf{u}_1$ has a non-singular normal distribution. Let $\mathbf{F}^{-1} = [\mathbf{G}_1, \mathbf{G}_2]$, where $\mathbf{G}_1 \in \mathbb{R}^{q\times s}$ and $\mathbf{G}_2 \in \mathbb{R}^{q\times(q-s)}$. Then we have

$$\mathbf{z} = \mathbf{F}^{-1}\left(\begin{bmatrix}\mathbf{u}_1\\ \mathbf{0}\end{bmatrix} + \mathbf{F}\mathbf{D}\boldsymbol{\lambda}\right) = [\mathbf{G}_1, \mathbf{G}_2]\begin{bmatrix}\mathbf{u}_1\\ \mathbf{0}\end{bmatrix} + \mathbf{D}\boldsymbol{\lambda} = \mathbf{G}_1\mathbf{u}_1 + \mathbf{D}\boldsymbol{\lambda}$$

which is of the form of the formal definition of normal distribution.

□

Theorem 3.7. *Let $\mathbf{U}$ be a $d\times k$ random matrix ($k \leq d$) and each of its entry is independent distributed according to $\mathcal{N}(0,1)$, then it holds that*

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d.$$

Remark 3.5. *In fact, we have* $\operatorname{rank}(\mathbf{U}) = k$ *with probability 1, which will be proved in later sections.*

1. *For $k = d$, we have $\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top = \mathbf{I}$.*

2. *For $k = 1$, it is $\mathbb{E}\left[\mathbf{u}(\mathbf{u}^\top\mathbf{u})^{-1}\mathbf{u}^\top\right] = (1/d)\mathbf{I}_d$, where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $\mathbf{y} = \mathbf{u}/\left\|\mathbf{u}\right\|_2$, then*

$$\mathbf{y} = [y_1, \dots, y_d]^\top \sim \operatorname{Unif}(\mathcal{S}^{k-1}), \quad \textit{where } \operatorname{Unif}(\mathcal{S}^{k-1}) \textit{ is the uniformly sphere distribution.}$$

   Note that we have $\mathbf{Q}\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for any unitary $\mathbf{Q} \in \mathbb{R}^{d\times d}$, which means $\mathbf{u}$ is rotation invariant. Then we conclude

$$\mathbf{Q}\mathbf{y} = \frac{\mathbf{Q}\mathbf{u}}{\left\|\mathbf{Q}\mathbf{u}\right\|_2} \sim \frac{\mathbf{u}}{\left\|\mathbf{u}\right\|_2} = \mathbf{y}.$$

   Let $\mathbf{Q} = [\mathbf{e}_1, \dots, -\mathbf{e}_i, \dots, \mathbf{e}_d]$ and $\hat{\mathbf{y}} = \mathbf{Q}\mathbf{y}$, then we have

$$\mathbf{y} = [y_1, \dots, y_i, \dots, y_d]^\top \quad \textit{and} \quad \hat{\mathbf{y}} = [y_1, \dots, -y_i, \dots, y_d]^\top$$

   has the same distribution. Hence, we have

$$\mathbb{E}[y_i] = -\mathbb{E}[y_i] \qquad \textit{and} \qquad \mathbb{E}[y_iy_j] = \mathbb{E}[(-y_i)y_j] \textit{ for any } i \neq j.$$

   This implies $\mathbb{E}[y_i] = 0$ and $\mathbb{E}[y_iy_j] = 0$ for any $i \neq j$. On the other hand, the variables $y_1, \dots, y_d$ are symmetric, which means

$$\mathbb{E}[y_1^2] = \cdots = \mathbb{E}[y_d^2].$$

   Combing with the fact

$$\sum_{i=1}^d y_i^2 = \sum_{i=1}^d \frac{y_i^2}{\sum_{i=1}^d y_i^2} = 1,$$

   we achieve $\mathbb{E}[y_1^2] = \cdots = \mathbb{E}[y_d^2] = 1/d$. Since we have showed $\mathbb{E}[\mathbf{y}] = \mathbf{0}$, we also have

$$\operatorname{Cov}[\mathbf{y}] = \mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top\right] = \mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \frac{1}{d}\mathbf{I}.$$

3. *The vector $\mathbf{y}$ in the case of $k = 1$ can be generalized to*

$$\mathbf{Y} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1/2}$$

   which is uniformly distributed on the Stiefel manifold $\mathcal{V}_{d,k} = \{\mathbf{A} \in \mathbb{R}^{d\times k} : \mathbf{A}^\top\mathbf{A} = \mathbf{I}_k\}$. You can read Section 2 of Chikuse [6] for more interesting results about statistics on Stiefel manifold. We can use the result of Theorem 3.7 to show the explicit convergence rates of block quasi-Newton methods [10], which builds a bridge for the theories between standard Newton method and conventional quasi-Newton method.

This document is organized by Luo Luo.
Please do not distribute.

Before prove Theorem 3.7, we first provide the following lemma.

Lemma 3.1. *Assume* $\mathbf{P} \in \mathbb{R}^{d\times r}$ *is column orthonormal* ($r \leq d$) *and* $\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ *is a* d*-dimensional multivariate normal distributed vector. Then we have*

$$\mathbb{E}\left[\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}}\right] = \frac{1}{r}\mathbf{P}\mathbf{P}^\top.$$

Proof. The distribution $\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ implies there exists a r-dimensional multivariate normal distributed vector $\mathbf{w} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ such that $\mathbf{v} = \mathbf{P}\mathbf{w}$. Thus we have

$$\mathbb{E}\left[\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}}\right] = \mathbb{E}\left[\frac{(\mathbf{P}\mathbf{w})(\mathbf{P}\mathbf{w})^\top}{(\mathbf{P}\mathbf{w})^\top(\mathbf{P}\mathbf{w})}\right] = \mathbb{E}\left[\frac{\mathbf{P}\mathbf{w}\mathbf{w}^\top\mathbf{P}^\top}{\mathbf{w}^\top\mathbf{w}}\right] = \mathbf{P}\mathbb{E}\left[\frac{\mathbf{w}\mathbf{w}^\top}{\mathbf{w}^\top\mathbf{w}}\right]\mathbf{P}^\top = \frac{1}{r}\mathbf{P}\mathbf{P}^\top,$$

where the last step is because of $\mathbf{w}/\left\|\mathbf{w}\right\|_2$ is uniform distributed on k-dimensional unit sphere and its covariance matrix is $r^{-1}\mathbf{I}_k$. □

Now we prove Theorem 3.7.

Proof. We prove this result by induction on k. The induction base $k = 1$ have been verified by previous remark. Now we assume

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d$$

holds for any $\mathbf{U} \in \mathbb{R}^{d\times k}$ that each of its entries are independently distributed according to $\mathcal{N}(0, 1)$. We define the random matrix

$$\hat{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix} \in \mathbb{R}^{d\times(k+1)},$$

where $\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ is independent distributed to $\mathbf{U}$. Then we have

$$\begin{aligned}
&\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top \\
=&\begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix}\left(\begin{bmatrix}\mathbf{U}^\top \\ \mathbf{v}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{U}^\top \\ \mathbf{v}^\top\end{bmatrix} \\
=&\mathbf{A} + \frac{(\mathbf{I}_d - \mathbf{A})\mathbf{v}\mathbf{v}^\top(\mathbf{I}_d - \mathbf{A})}{\mathbf{v}^\top(\mathbf{I}_d - \mathbf{A})\mathbf{v}},
\end{aligned}$$

where $\mathbf{A} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$. Since the rank of projection matrix $\mathbf{I}_d - \mathbf{A}$ is $d - k$, we have $\mathbf{I}_d - \mathbf{A} = \mathbf{Q}\mathbf{Q}^\top$ for some column orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{d\times(d-k)}$. Thus, we achieve

$$\begin{aligned}
&\mathbb{E}[\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})\hat{\mathbf{U}}^\top] \\
=&\mathbb{E}[\mathbf{A}] + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{v}}\left[\left.\frac{(\mathbf{I}_d - \mathbf{A})\mathbf{v}\mathbf{v}^\top(\mathbf{I}_d - \mathbf{A})}{\mathbf{v}^\top(\mathbf{I}_d - \mathbf{A})\mathbf{v}}\,\right|\,\mathbf{U}\right]\right] \\
=&\frac{k}{d}\mathbf{I}_d + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{v}}\left[\left.\frac{(\mathbf{Q}\mathbf{Q}^\top\mathbf{v})(\mathbf{v}^\top\mathbf{Q}\mathbf{Q}^\top)}{(\mathbf{v}^\top\mathbf{Q}\mathbf{Q}^\top)(\mathbf{Q}\mathbf{Q}^\top\mathbf{v})}\,\right|\,\mathbf{U}\right]\right] \\
=&\frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{Q}\mathbf{Q}^\top] \\
=&\frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{I}_d - \mathbf{A}] \\
=&\frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\left(\mathbf{I}_d - \frac{k}{d}\mathbf{I}_d\right) \\
=&\frac{k+1}{d}\mathbf{I}_d,
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

which completes the induction. In above derivation, the second equality is due to Lemma 3.1 with $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ and $r = d - k$ and the fact $\mathbf{Q}\mathbf{Q}^\top\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}\mathbf{Q}^\top)$ for given column orthonormal $\mathbf{Q}$; the third equality comes from the inductive hypothesis. □

Theorem 3.8 (Conditional Distribution)**.** *Let* $\mathbf{x}$ *be distributed according to* $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} \succ \mathbf{0}$. *We partition*

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \quad \textit{with} \quad \mathbf{x}^{(1)} \in \mathbb{R}^q \quad \textit{and} \quad \mathbf{x}^{(2)} \in \mathbb{R}^{p-q}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \textit{ with } \boldsymbol{\mu}^{(1)} \in \mathbb{R}^q \quad \textit{and} \quad \boldsymbol{\mu}^{(2)} \in \mathbb{R}^{p-q},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad \textit{with} \quad \boldsymbol{\Sigma}_{11} \in \mathbb{R}^{q\times q}, \quad \boldsymbol{\Sigma}_{12} \in \mathbb{R}^{q\times(p-q)}, \quad \boldsymbol{\Sigma}_{21} \in \mathbb{R}^{(p-q)\times q} \quad \textit{and} \quad \boldsymbol{\Sigma}_{22} \in \mathbb{R}^{(p-q)\times(p-q)}.$$

with Then the conditional density of $\mathbf{x}^{(1)}$ *given that* $\mathbf{x}^{(2)}$ *is*

$$f(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}) = \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{f(\mathbf{x}^{(2)})} = \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)^\top \boldsymbol{\Sigma}_{11.2}^{-1}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)\right),$$

where $\mathbf{x}_{11.2} = \mathbf{x}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{(2)}$, $\boldsymbol{\mu}_{11.2} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}$ *and* $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Proof. Recall the transformation

$$\mathbf{y}^{(1)} = \mathbf{x}^{(1)} + \mathbf{B}\mathbf{x}^{(2)} = \mathbf{x}_{11.2} \qquad \text{and} \qquad \mathbf{y}^{(2)} = \mathbf{x}^{(2)} \qquad \text{with} \qquad \mathbf{B} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1},$$

which is used to prove the marginal distribution of multivariate normal distribution.
We have proved $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are independent and has normal distribution with

$$\mathbb{E}\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \text{ and } \operatorname{Cov}\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}. \end{bmatrix}$$

Hence, the joint density of $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ is

$$\begin{aligned} g(\mathbf{y}) =& n(\mathbf{y}^{(1)} \mid \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \cdot n(\mathbf{y}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}) \\ =& n(\mathbf{y}^{(1)} \mid \boldsymbol{\mu}_{11.2}, \boldsymbol{\Sigma}_{11.2}) \cdot n(\mathbf{y}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}), \end{aligned}$$

where

$$n(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

Note that

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \qquad \text{and} \qquad \det\left(\begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\right) = 1.$$

Therefore, we have

$$\begin{aligned} & f(\mathbf{x}) = f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \\ =& n(\mathbf{y}^{(1)} \mid \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}_{11.2}, \boldsymbol{\Sigma}_{11.2}) \cdot n(\mathbf{y}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}) \\ =& n(\mathbf{x}_{11.2} \mid \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}_{11.2}, \boldsymbol{\Sigma}_{11.2}) \cdot n(\mathbf{x}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}) \\ =& \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)^\top \boldsymbol{\Sigma}_{11.2}^{-1}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)\right) \cdot f(\mathbf{x}^{(2)}). \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

This implies the conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\begin{aligned}f(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}) =& \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{f(\mathbf{x}^{(2)})} \\ =&\frac{1}{\sqrt{(2\pi)^q \det(\mathbf{\Sigma}_{11.2})}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)^\top \mathbf{\Sigma}_{11.2}^{-1}\left(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}\right)\right).\end{aligned}$$

□

Remark 3.6. *This theorem indicates the conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is*

$$\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}), \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\right),$$

since we have

$$\begin{aligned}&\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2} \\ =&\mathbf{x}^{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{x}^{(2)} - \left(\boldsymbol{\mu}^{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}\right) \\ =&\mathbf{x}^{(1)} - \left(\boldsymbol{\mu}^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})\right).\end{aligned}$$

Remark 3.7. *Note that $\mathbf{\Sigma}_{11.2}$ has the form of Schur complement and*

$$\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \preceq \mathbf{\Sigma}_{11}.$$

This means $\mathbf{\Sigma}_{11.2}$ is potentially "smaller" than $\mathbf{\Sigma}_{11}$ of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are heavily correlated. If $\mathbf{\Sigma}_{22}$ is large, the matrix $\mathbf{\Sigma}_{11.2}$ will be close to $\mathbf{\Sigma}_{11}$ since we have fixed the components with "large" covariance.

Theorem 3.9. *Consider the twice differentiable function $f(\mathbf{x}, \mathbf{y})$ which is strongly convex in $\mathbf{y}$. The problem*

$$\min_{\mathbf{x}\in\mathbb{R}^q, \mathbf{y}\in\mathbb{R}^{p-q}} f(\mathbf{x}, \mathbf{y})$$

can be formulated by

$$\min_{\mathbf{x}\in\mathbb{R}^q}\left\{P(\mathbf{x}) \triangleq \min_{\mathbf{y}\in\mathbb{R}^{p-q}} f(\mathbf{x}, \mathbf{y})\right\}.$$

Then we have

$$\nabla^2 P(\mathbf{x}) = \nabla^2_{xx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla^2_{xy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))(\nabla^2_{yy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})))^{-1}\nabla^2_{yx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})).$$

Proof. The implicit function theorem means $\mathbf{y}^*(\cdot)$ is differentiable. The optimality of $\mathbf{y}^*(\cdot)$ means

$$\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) = \mathbf{0}. \tag{8}$$

Taking total derivative on equation (8) achieves

$$\nabla^2_{yx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \nabla^2_{yy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\nabla \mathbf{y}^*(\mathbf{x}) = \mathbf{0}. \tag{9}$$

The Danskin's theorem [3] says

$$\nabla P(\mathbf{x}) = \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})). \tag{10}$$

Taking total derivative on equation (10) achieves

$$\nabla^2 P(\mathbf{x}) = \nabla^2_{xx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \nabla^2_{xy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\nabla \mathbf{y}^*(\mathbf{x}). \tag{11}$$

Combining equations (9) and (11), we have

$$\nabla^2 P(\mathbf{x}) = \nabla^2_{xx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla^2_{xy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big(\nabla^2_{yy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big)^{-1}\nabla^2_{yx} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})),$$

where the non-singularity of $\nabla^2_{yy} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is due to the strong convexity. □

Remark 3.8. *The property of Hessian shown in Theorem 3.9 is very similar to the counterpart covariance of normal distribution. Informally speaking, by fixing partial ones, the second-order information of the remains has the form of Schur complement. Theorem 3.9 is also useful in the analysis of second-order stationarity in nonconvex minimax optimization [12].*

This document is organized by Luo Luo.
Please do not distribute.

Characteristic Function: The characteristic function determines the density function uniquely (if exists).

Theorem 3.10. *If the p-dimensional random vector $\mathbf{x}$ has the density $f(\mathbf{x})$ and the characteristic function $\phi(\mathbf{t})$, then*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^p}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-\mathrm{i}\,\mathbf{t}^\top\mathbf{x})\,\phi(\mathbf{t})\,\mathrm{d}t_1\ldots\mathrm{d}t_p.$$

Proof. We have

$$\begin{aligned}
&\frac{1}{(2\pi)^p}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-\mathrm{i}\,\mathbf{t}^\top\mathbf{x})\,\phi(\mathbf{t})\,\mathrm{d}t_1\ldots\mathrm{d}t_p\\
=&\frac{1}{(2\pi)^p}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-\mathrm{i}\,\mathbf{t}^\top\mathbf{x})\left(\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{y})f(\mathbf{y})\,\mathrm{d}y_1\ldots\mathrm{d}y_p\right)\mathrm{d}t_1\ldots\mathrm{d}t_p\\
=&\frac{1}{(2\pi)^p}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-\mathrm{i}\,\mathbf{t}^\top(\mathbf{x}-\mathbf{y}))f(\mathbf{y})\,\mathrm{d}y_1\ldots\mathrm{d}y_p\,\mathrm{d}t_1\ldots\mathrm{d}t_p\\
=&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\mathbf{y})\prod_{j=1}^{p}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-\mathrm{i}\,t_j(x_j-y_j))\,\mathrm{d}t_i\right)\mathrm{d}y_1\ldots\mathrm{d}y_p\\
=&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(y_1,y_2,\ldots,y_p)\prod_{j=1}^{p}\delta(y_j-x_j)\,\mathrm{d}y_1\ldots\mathrm{d}y_p\\
=&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,y_2,\ldots,y_p)\prod_{j=2}^{p}\delta(y_j-x_j)\,\mathrm{d}y_2\ldots\mathrm{d}y_p\\
=&\cdots=\ldots\\
=&f(x_1,x_2,\ldots,x_p).
\end{aligned}$$

□

Remark 3.9. *The Dirac delta function $\delta(x)$ can be loosely thought of as a function on the real line which is zero everywhere except at the origin, where it is infinite,*

$$\delta(x) = \begin{cases} +\infty, & x = 0,\\ 0, & x \neq 0\end{cases}$$

and which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty}\delta(x)\,\mathrm{d}x = 1.$$

The integral of the time-delayed Dirac delta is

$$\int_{-\infty}^{\infty}f(t)\,\delta(t-T)\,\mathrm{d}t = f(T).$$

for any f that is is finite almost everywhere.

If the random variable does not have a density, the characteristic function uniquely defines the probability of any continuity interval.

Theorem 3.11 (Section 10.7 of Cramér [7]). *Let $\{F_j(\mathbf{x})\}$ be a sequence of cdfs, and let $\{\phi_j(\mathbf{t})\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(\mathbf{x})$ to converge to a cdf $F(\mathbf{x})$ is that, for every $\mathbf{t}$, $\phi_j(\mathbf{t})$ converges to a limit $\phi(\mathbf{t})$ that is continuous at $\mathbf{t}=0$. When this condition is satisfied, the limit $\phi(\mathbf{t})$ is identical with the characteristic function of the limiting distribution $F(\mathbf{x})$.*

This document is organized by Luo Luo.
Please do not distribute.

Theorem 3.12. *The characteristic function of* $\mathbf{x}$ *distributed according to* $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is*

$$\phi(\mathbf{t}) = \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}\right).$$

for every $\mathbf{t} \in \mathbb{R}^p$.

Proof. For standard normal distribution $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$, we have

$$\begin{aligned}\phi_0(\mathbf{t}) =& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top\mathbf{y}\right)\right] \\ =& \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{\exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{y})}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\ =& \prod_{j=1}^{p}\left(\int_{-\infty}^{+\infty}\frac{\exp(\mathrm{i}\,t_jy_j)}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_j^2\right)\,\mathrm{d}y_j\right)\end{aligned} \tag{12}$$

Now we want to calculate

$$\begin{aligned}I(t) \triangleq& \int_{-\infty}^{+\infty}\exp(\mathrm{i}\,ty)\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y \\ =& \int_{-\infty}^{+\infty}\cos(ty)\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y + \int_{-\infty}^{+\infty}\mathrm{i}\,\sin(ty)\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y \\ =& \int_{-\infty}^{+\infty}\cos(ty)\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y.\end{aligned}$$

Note that

$$\begin{aligned}I'(t) =& \int_{-\infty}^{+\infty} -y\sin(ty)\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y \\ =& \int_{-\infty}^{+\infty}\sin(ty)\,\mathrm{d}\left(\exp\left(-\frac{1}{2}y^2\right)\right) \\ =& \sin(ty)\exp\left(-\frac{1}{2}y^2\right)\Bigg|_{-\infty}^{\infty} - \int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}\sin(ty) \\ =& -\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}y^2\right)\cdot t\cos(ty)\,\mathrm{d}y = -tI(t).\end{aligned}$$

The solution of differential equation $I'(t) = -tI(t)$ has the form of

$$I(t) = c\exp\left(-\frac{1}{2}t^2\right).$$

We also have

$$I(0) = \int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y = \sqrt{2\pi},$$

which means

$$c = \sqrt{2\pi}.$$

Hence, we have

$$I(t) = \sqrt{2\pi}\exp\left(-\frac{1}{2}t^2\right).$$

This document is organized by Luo Luo.
Please do not distribute.

Combining this result with equation (12), we have

$$\phi_0(\mathbf{t}) = \prod_{j=1}^{p} \left(\int_{-\infty}^{+\infty} \frac{\exp(\mathrm{i}\, t_j y_j)}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right) \mathrm{d}y_j\right) = \prod_{j=1}^{p} \frac{I(t_j)}{\sqrt{2\pi}} = \prod_{j=1}^{p} \exp\left(-\frac{1}{2}t_j^2\right) = \exp\left(-\frac{1}{2}\mathbf{t}^\top\mathbf{t}\right).$$

For the general case of $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can write $\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu}$ such that $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$. Then we have

$$\begin{aligned}
\phi(\mathbf{t}) =& \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{x})\right] \\
=& \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{t}^\top(\mathbf{A}\mathbf{y} + \boldsymbol{\mu}))\right] \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}\right)\, \mathbb{E}\left[\exp(\mathrm{i}\,(\mathbf{A}^\top\mathbf{t})^\top\mathbf{y})\right] \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}\right)\, \phi_0\left(\mathbf{A}^\top\mathbf{t}\right) \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}\right)\, \exp\left(-\frac{1}{2}\mathbf{t}^\top\mathbf{A}\mathbf{A}^\top\mathbf{t}\right) \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}\right).
\end{aligned}$$

□

Remark 3.10. *Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with characteristic function $\phi_{\mathbf{x}}(\mathbf{t}) = \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}\right)$. For $\mathbf{z} = \mathbf{C}\mathbf{x}$, the characteristic function of $\mathbf{z}$ is*

$$\begin{aligned}
\phi_{\mathbf{z}}(\mathbf{t}) =& \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{z})\right] \\
=& \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{C}\mathbf{x})\right] \\
=& \mathbb{E}\left[\exp(\mathrm{i}\,(\mathbf{C}^\top\mathbf{t})^\top\mathbf{x})\right] \\
=& \phi_{\mathbf{x}}(\mathbf{C}^\top\mathbf{t}) \\
=& \exp\left(\mathrm{i}\,(\mathbf{C}^\top\mathbf{t})^\top\boldsymbol{\mu} - \frac{1}{2}(\mathbf{C}^\top\mathbf{t})^\top\boldsymbol{\Sigma}(\mathbf{C}^\top\mathbf{t})\right) \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top(\mathbf{C}\boldsymbol{\mu}) - \frac{1}{2}\mathbf{t}^\top(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)\mathbf{t}\right)
\end{aligned}$$

which implies $\mathbf{z} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$ and we prove Theorem 3.6.

Theorem 3.13. *If every linear combination of the components of a random vector $\mathbf{y}$ is normally distributed, then $\mathbf{y}$ is normally distributed.*

Proof. Let $\mathbf{y}$ is a random vector with $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathrm{Cov}[\mathbf{y}] = \boldsymbol{\Sigma}$. Suppose the univariate random variable $\mathbf{u}^\top\mathbf{y}$ (linear combination of $\mathbf{y}$) is normal distributed for any $\mathbf{u} \in \mathbb{R}^p$. The characteristic function of $\mathbf{u}^\top\mathbf{y}$ is

$$\begin{aligned}
\phi_{\mathbf{u}^\top\mathbf{y}}(t) =& \mathbb{E}\left[\exp(\mathrm{i}\, t\mathbf{u}^\top\mathbf{y})\right] \\
=& \exp\left(\mathrm{i}\, t\mathbb{E}[\mathbf{u}^\top\mathbf{y}] - \frac{1}{2}t^2\mathrm{Cov}[\mathbf{u}^\top\mathbf{y}]\right) \\
=& \exp\left(\mathrm{i}\, t\mathbf{u}^\top\boldsymbol{\mu} - \frac{1}{2}t^2\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}\right).
\end{aligned}$$

Set $t = 1$, then we have

$$\phi_{\mathbf{u}^\top\mathbf{y}}(1) = \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{u}^\top\mathbf{y})\right] = \exp\left(\mathrm{i}\,\mathbf{u}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}\right)$$

which implies the characteristic function of $\mathbf{y}$ is

$$\phi_{\mathbf{y}}(\mathbf{u}) = \exp\left(\mathrm{i}\,\mathbf{u}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}\right),$$

that is, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

This document is organized by Luo Luo. Please do not distribute.

Theorem 3.14. *We let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $\mathbf{z} = \mathbf{x} + \mathbf{y}$. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are independent, then we have $\mathbf{z} \sim \mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.*

Proof. Let $\phi_{\mathbf{x}}$, $\phi_{\mathbf{y}}$ and $\phi_{\mathbf{z}}$ be the characteristic functions of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. Then we have

$$\begin{aligned}
\phi_{\mathbf{z}}(\mathbf{t}) =& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top(\mathbf{x}+\mathbf{y})\right)\right] \\
=& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top\mathbf{x}\right)\exp\left(\mathrm{i}\,\mathbf{t}^\top\mathbf{y}\right)\right] \\
=& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top\mathbf{x}\right)\right]\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top\mathbf{y}\right)\right] \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}_1 - \frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}_1\mathbf{t}\right)\exp\left(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}_2 - \frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}_2\mathbf{t}\right) \\
=& \exp\left(\mathrm{i}\,\mathbf{t}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) - \frac{1}{2}\mathbf{t}^\top(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\mathbf{t}\right),
\end{aligned}$$

which is the characteristic function of $\mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. □

Theorem 3.15. *If the n-th moment of random variable x, denoted by $\mathbb{E}[x^n]$, exists and is finite, then its characteristic function is n times continuously differentiable and*

$$\mathbb{E}[x^n] = \frac{1}{\mathrm{i}^n}\frac{\mathrm{d}^n\phi(t)}{\mathrm{d}t^n}\bigg|_{t=0}$$

Proof. We prove this result as

$$\begin{aligned}
\frac{\mathrm{d}^n\phi(t)}{\mathrm{d}t^n} =& \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}\left[\exp(\mathrm{i}\,tx)\right] \\
=& \mathbb{E}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}\exp(\mathrm{i}\,tx)\right] \\
=& \mathbb{E}\left[(\mathrm{i}\,x)^n\exp(\mathrm{i}\,tx)\right] \\
=& \mathrm{i}^n\,\mathbb{E}\left[x^n\exp(\mathrm{i}\,tx)\right].
\end{aligned}$$

□

4 Estimation of the Mean Vector and the Covariance

Theorem 4.1. *If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $N > p$, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are*

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

respectively.

Proof. The likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{pN/2}(\det(\boldsymbol{\Sigma}))^{N/2}}\exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right).$$

The logarithm of the likelihood function is

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{pN}{2}\ln 2\pi - \frac{N}{2}\ln\left(\det(\boldsymbol{\Sigma})\right) - \frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}).$$

This document is organized by Luo Luo.
Please do not distribute.

We first focus on $\boldsymbol{\mu}$. It holds that

$$
\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\boldsymbol{\mu})\\
=&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})\\
=&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})+\sum_{\alpha=1}^{N}(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})+\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})+\sum_{\alpha=1}^{N}(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\\
=&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})+\sum_{\alpha=1}^{N}(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\\
\geq&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}}),
\end{aligned}
$$

where the equality holds when $\boldsymbol{\mu}=\bar{\mathbf{x}}$. Hence, the estimator of means should be $\hat{\boldsymbol{\mu}}=\bar{\mathbf{x}}$.

Now, we only need to study how to maximize

$$
-\frac{pN}{2}\ln 2\pi-\frac{N}{2}\ln\left(\det(\boldsymbol{\Sigma})\right)-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}}).
$$

We let $\boldsymbol{\Psi}=\boldsymbol{\Sigma}^{-1}$ and

$$
\begin{aligned}
l(\boldsymbol{\Psi})=&-\frac{pN}{2}\ln 2\pi-\frac{N}{2}\ln\left(\det(\boldsymbol{\Psi}^{-1})\right)-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Psi}(\mathbf{x}_\alpha-\bar{\mathbf{x}})\\
=&-\frac{pN}{2}\ln 2\pi+\frac{N}{2}\ln\left(\det(\boldsymbol{\Psi})\right)-\frac{1}{2}\sum_{\alpha=1}^{N}\operatorname{tr}\left((\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Psi}(\mathbf{x}_\alpha-\bar{\mathbf{x}})\right)\\
=&-\frac{pN}{2}\ln 2\pi+\frac{N}{2}\ln\left(\det(\boldsymbol{\Psi})\right)-\frac{1}{2}\sum_{\alpha=1}^{N}\operatorname{tr}\left((\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Psi}\right),
\end{aligned}
$$

then

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\Psi})}{\partial\boldsymbol{\Psi}}=&\frac{\partial}{\partial\boldsymbol{\Psi}}\left(-\frac{PN}{2}\ln 2\pi+\frac{N}{2}\ln\left(\det(\boldsymbol{\Psi})\right)-\frac{1}{2}\sum_{\alpha=1}^{N}\operatorname{tr}\left((\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Psi}\right)\right)\\
=&\frac{N}{2}\boldsymbol{\Psi}^{-1}-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top.
\end{aligned}
$$

We can verify $l(\boldsymbol{\Psi})$ is concave on the domain of symmetric positive definite matrices, which means the maximum is taken by $\dfrac{\partial f(\boldsymbol{\Psi})}{\partial\boldsymbol{\Psi}}=\mathbf{0}$, that is,

$$
\boldsymbol{\Sigma}=\boldsymbol{\Psi}^{-1}=\frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top.
$$

□

Lemma 4.1. *Given function $f:\mathbb{R}^d\to\mathbb{R}$ with convex domain $\operatorname{dom} f$, if for any $\mathbf{u}\in\operatorname{dom} f$ and $\mathbf{v}\in\mathbb{R}^d$, the function*

$$
g(t)=f(\mathbf{u}+t\mathbf{v})
$$

is convex in t on $\operatorname{dom} g=\{t:\mathbf{u}+t\mathbf{v}\in\operatorname{dom} f\}$, then f is convex on $\operatorname{dom} f$.

This document is organized by Luo Luo.
Please do not distribute.

Proof. The convexity of $g(t)$ means we have

$$\begin{aligned} & g(\lambda t_1 + (1-\lambda)t_2) \le \lambda g(t_1) + (1-\lambda) g(t_2) \\ \Longleftrightarrow & f(\mathbf{u} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{v}) \le \lambda f(\mathbf{u} + t_1\mathbf{v}) + (1-\lambda) f(\mathbf{u} + t_2\mathbf{v}) \end{aligned}$$

for any $\lambda \in [0,1]$ and $t_1, t_2 \in \operatorname{dom} g$. For any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, taking $\mathbf{u} \in \operatorname{dom} f$ and $\mathbf{v} \in \mathbb{R}^d$ such that

$$\begin{cases} \mathbf{u} + t_1\mathbf{v} = \mathbf{x}, \\ \mathbf{u} + t_2\mathbf{v} = \mathbf{y}, \end{cases} \tag{13}$$

leads to

$$\lambda\mathbf{x} + (1-\lambda)\mathbf{y} = \lambda(\mathbf{u} + t_1\mathbf{v}) + (1-\lambda)(\mathbf{u} + t_2\mathbf{v}) = \mathbf{u} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{v}$$

and

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = f(\mathbf{u} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{v}) \le \lambda f(\mathbf{u} + t_1\mathbf{v}) + (1-\lambda) f(\mathbf{u} + t_2\mathbf{v}) = \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}),$$

which is the definition of convexity on f. The condition (13) can be satisfied for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ by taking

$$\mathbf{u} = \frac{t_1\mathbf{y} - t_2\mathbf{x}}{t_1 - t_2} \qquad \text{and} \qquad \mathbf{v} = \frac{\mathbf{x} - \mathbf{y}}{t_1 - t_2}$$

with if $t_1 \neq t_2$. The case of $t_1 = t_2$ corresponds to $\mathbf{x} = \mathbf{y}$ which leads to the condition of convexity holds trivial. □

Remark 4.1. *We can show* $\operatorname{dom} g$ *is convex. Consider fixed* $\mathbf{u} \in \operatorname{dom} f$ *and* $\mathbf{v} \in \mathbb{R}^d$. *Since* $\operatorname{dom} f$ *is convex, the set* $\mathcal{C} = -\mathbf{u} + \operatorname{dom} f$ *is also convex. For any* $t_1, t_2 \in \operatorname{dom} g$, *we have*

$$\mathbf{u} + t_1\mathbf{v}, \mathbf{u} + t_2\mathbf{v} \in \operatorname{dom} f \quad \text{and} \quad t_1\mathbf{v}, t_2\mathbf{v} \in \mathcal{C}.$$

The convexity of $\mathcal{C}$ *means for any* $\lambda \in [0,1]$ *holds*

$$(\lambda t_1 + (1-\lambda)t_2)\mathbf{v} = \lambda(t_1\mathbf{v}) + (1-\lambda)(t_2\mathbf{v}) \in \mathcal{C}.$$

This implies $\lambda t_1 + (1-\lambda)t_2 \in \operatorname{dom} g$. *Hence, the set* $\operatorname{dom} g$ *is convex.*

Theorem 4.2. *The function* $h : \mathbb{S}^p_{++} \to \mathbb{R}$ *such that*

$$h(\mathbf{X}) = -\log\det(\mathbf{X})$$

is convex, where $\mathbb{S}^p_{++} = \{\mathbf{X} \in \mathbb{R}^{p \times p} : \mathbf{X} \succ \mathbf{0}\}$.

Proof. We denote $\mathbb{S}^p = \{\mathbf{Z} \in \mathbb{R}^{p \times p} : \mathbf{Z} = \mathbf{Z}^\top\}$. Lemma 4.1 indicates we only needs to show for any $\mathbf{X} \in \mathbb{S}^p_{++}$ and $\mathbf{Z} \in \mathbb{S}^p$, the function

$$g(t) = -\log\det(\mathbf{X} + t\mathbf{Z})$$

is convex on $\operatorname{dom} g = \{t : \mathbf{X} + t\mathbf{Z} \in \mathbb{S}^p_{++}\} = \{t : \mathbf{X} + t\mathbf{Z} \succ \mathbf{0}\}$. We have

$$\begin{aligned} g(t) &= -\ln\det(\mathbf{X} + t\mathbf{Z}) \\ &= -\ln\det(\mathbf{X}^{1/2}(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Z}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}) \\ &= -\ln\left(\det(\mathbf{X}^{1/2})\det(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Z}\mathbf{X}^{-1/2})\det(\mathbf{X}^{1/2})\right) \\ &= -\ln\left(\det(\mathbf{X})\det(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Z}\mathbf{X}^{-1/2})\right) \\ &= -\ln\det(\mathbf{X}) - \ln\det(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Z}\mathbf{X}^{-1/2}) \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
&= - \ln\det(\mathbf{X}) - \ln\prod_{i=1}^{p}(1+t\lambda_i) \\
&= - \ln\det(\mathbf{X}) - \sum_{i=1}^{p}\ln(1+t\lambda_i),
\end{aligned}
$$

where λ_i is the i-th eigenvalue of $\mathbf{X}^{-1/2}\mathbf{Z}\mathbf{X}^{-1/2}$. Therefore, we have

$$
g'(t) = -\sum_{i=1}^{p}\frac{t}{1+t\lambda_i} \quad \text{and} \quad g''(t) = \sum_{i=1}^{p}\frac{1}{(1+t\lambda_i)^2} > 0,
$$

which means g is convex. □

We can also achieve the solution by the following lemma with $\mathbf{G} = \mathbf{\Sigma}$ and $\mathbf{D} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$.

Lemma 4.2. *If $\mathbf{D} \in \mathbb{R}^{p\times p}$ is positive definite, the maximum of*

$$
f(\mathbf{G}) = -N\ln\det(\mathbf{G}) - \operatorname{tr}(\mathbf{G}^{-1}\mathbf{D})
$$

with respect to positive definite matrices $\mathbf{G}$ exists, occurs at $\mathbf{G} = \frac{1}{N}\mathbf{D}$.

Proof. Let $\mathbf{D} = \mathbf{E}\mathbf{E}^\top$ and $\mathbf{E}^\top\mathbf{G}^{-1}\mathbf{E} = \mathbf{H}$. Then we have $\mathbf{G} = \mathbf{E}\mathbf{H}^{-1}\mathbf{E}^\top$,

$$
\det(\mathbf{G}) = \det(\mathbf{E})\det(\mathbf{H}^{-1})\det(\mathbf{E}^\top) = \det(\mathbf{E}\mathbf{E}^\top)\det(\mathbf{H}^{-1}) = \frac{\det(\mathbf{D})}{\det(\mathbf{H})}
$$

and

$$
\operatorname{tr}(\mathbf{G}^{-1}\mathbf{D}) = \operatorname{tr}(\mathbf{G}^{-1}\mathbf{E}\mathbf{E}^\top) = \operatorname{tr}(\mathbf{E}^\top\mathbf{G}^{-1}\mathbf{E}) = \operatorname{tr}(\mathbf{H}).
$$

Then the function to be maximized (with respect to positive definite $\mathbf{H}$) is

$$
g(\mathbf{H}) = -N\ln\det(\mathbf{D}) + N\ln\det(\mathbf{H}) - \operatorname{tr}(\mathbf{H}).
$$

Let $\mathbf{H} = \mathbf{T}\mathbf{T}^\top$ here $\mathbf{T}$ is lower triangular. Then the maximum of

$$
\begin{aligned}
g(\mathbf{H}) &= - N\ln\det(\mathbf{D}) + N\ln\det(\mathbf{H}) - \operatorname{tr}(\mathbf{H}) \\
&= - N\ln\det(\mathbf{D}) + N\ln(\det(\mathbf{T}))^2 - \operatorname{tr}(\mathbf{T}\mathbf{T}^\top) \\
&= - N\ln\det(\mathbf{D}) + N\ln\left(\prod_{i=1}^{p}t_{ii}^2\right) - \sum_{i\geq j}t_{ij}^2 \\
&= - N\ln\det(\mathbf{D}) + \sum_{i=1}^{p}\left(N\ln(t_{ii}^2) - t_{ii}^2\right) - \sum_{i>j}t_{ij}^2
\end{aligned}
$$

occurs at $t_{ii}^2 = N$ and $t_{ij} = 0$ for $i \neq j$; that is $\mathbf{H} = N\mathbf{I}$. Then

$$
\mathbf{G} = \mathbf{E}\mathbf{H}^{-1}\mathbf{E}^\top = \frac{1}{N}\mathbf{E}\mathbf{E}^\top = \frac{1}{N}\mathbf{D}.
$$

□

Theorem 4.3. *Let $f(\theta)$ be a real-valued function defined on a set $\mathcal{S}$ and let ϕ be a single-valued function, with a single-valued inverse, on $\mathcal{S}$ to a set $\mathcal{S}^*$. Let*

$$
g(\theta^*) = f\left(\phi^{-1}(\theta^*)\right).
$$

Then if $f(\theta)$ attains a maximum at $\theta = \theta_0$, then $g(\theta^*)$ attains a maximum at $\theta^* = \theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique, so is the maximum of $g(\theta^*)$ at θ_0^*.

This document is organized by Luo Luo.
Please do not distribute.

Proof. By hypothesis $f(\theta_0) \geq f(\theta)$ for all $\theta \in \mathcal{S}$. Then for any $\theta^* \in \mathcal{S}^*$, we have

$$g(\theta^*) = f\left(\phi^{-1}(\theta^*)\right) = f(\theta) \leq f(\theta_0) = g(\phi(\theta_0)) = g(\theta_0^*).$$

Thus $g(\theta^*)$ attains a maximum at $\theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique, there is strict inequality above for $\theta \neq \theta_0$, and the maximum of $g(\theta^*)$ is unique. □

Theorem 4.4. *If $\boldsymbol{\phi} : \mathcal{S} \to \mathcal{S}^*$ is not one-to-one, we let*

$$\boldsymbol{\phi}^{-1}(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} : \boldsymbol{\theta}^* = \boldsymbol{\phi}(\boldsymbol{\theta})\}.$$

and the induced likelihood function

$$g(\boldsymbol{\theta}^*) = \sup\{f\,(\boldsymbol{\theta}) : \boldsymbol{\theta}^* = \boldsymbol{\phi}(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \mathcal{S}\}.$$

If $\hat{\boldsymbol{\theta}}$ maximize $f(\cdot)$, then $\hat{\boldsymbol{\theta}}^* \triangleq \phi(\hat{\boldsymbol{\theta}})$ also maximize $g(\cdot)$.

Proof. The definition of $g(\cdot)$ means

$$\sup_{\boldsymbol{\theta}^* \in \mathcal{S}^*} g(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}^* \in \mathcal{S}^*} \sup_{\boldsymbol{\theta}^* = \boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathcal{S}} f(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta} \in \mathcal{S}} f(\boldsymbol{\theta}).$$

Since $\hat{\boldsymbol{\theta}}$ maximizes $f(\cdot)$ and $\hat{\boldsymbol{\theta}}^* \triangleq \boldsymbol{\phi}(\hat{\boldsymbol{\theta}})$, we have

$$f(\hat{\boldsymbol{\theta}}) = \sup\{f(\boldsymbol{\theta}) : \hat{\boldsymbol{\theta}}^* = \boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathcal{S}\} = g(\hat{\boldsymbol{\theta}}^*),$$

where the last step is based on the definition of g. Since $\hat{\boldsymbol{\theta}}$ maximizes $f(\cdot)$, we also have

$$g(\hat{\boldsymbol{\theta}}^*) = f(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in \mathcal{S}} f(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}^* \in \mathcal{S}^*} g(\boldsymbol{\theta}^*)$$

by connecting above two results, which implies $\hat{\boldsymbol{\theta}}^*$ maximizes $g(\boldsymbol{\theta}^*)$. □

Corollary 4.1. *If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ constitutes a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\rho_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)$. Then the maximum likelihood estimator of ρ_{ij} is*

$$\hat{\rho}_{ij} = \frac{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^N (x_{j\alpha} - \bar{x}_j)^2}}$$

Proof. The set of parameters $\mu_i = \mu_i$, $\sigma_i^2 = \sigma_{ii}$ and $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ is a one-to-one transform of the set of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then the estimator of ρ is

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}} = \frac{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^N (x_{j\alpha} - \bar{x}_j)^2}}.$$

□

Remark 4.2. *The estimator $\hat{\rho}_{ij}$ does not correspond to estimators of mean and covariance uniquely. Note that any estimators $c\hat{\boldsymbol{\Sigma}}$ with $c > 0$ leads to above $\hat{\rho}_{ij}$.*

Theorem 4.5. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent, where $\mathbf{x}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma})$. Let $\mathbf{C} \in \mathbb{R}^{N \times N}$ be an orthogonal matrix, then*

$$\mathbf{y}_\alpha = \sum_{\beta=1}^N c_{\alpha\beta}\mathbf{x}_\beta \sim \mathcal{N}_p(\boldsymbol{\nu}_\alpha, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\nu}_\alpha = \sum_{\beta=1}^N c_{\alpha\beta}\boldsymbol{\mu}_\beta$ for $\alpha = 1, \ldots, N$ and $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent.

This document is organized by Luo Luo.
Please do not distribute.

Proof. The set of vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ have a joint normal distribution, because the entire set of components is a set of linear combinations of the components of $\mathbf{x}_1, \ldots, \mathbf{x}_N$, which have a joint normal distribution. The expected value of $\mathbf{y}_\alpha$ is

$$\mathbb{E}[\mathbf{y}_\alpha] = \mathbb{E}\left[\sum_{\beta=1}^{N} c_{\alpha\beta}\mathbf{x}_\beta\right] = \sum_{\beta=1}^{N} c_{\alpha\beta}\mathbb{E}\left[\mathbf{x}_\beta\right] = \sum_{\beta=1}^{N} c_{\alpha\beta}\boldsymbol{\mu}_\beta.$$

The covariance matrix between $\mathbf{y}_\alpha$ and $\mathbf{y}_\gamma$ is

$$\begin{aligned}
\mathrm{Cov}[\mathbf{y}_\alpha, \mathbf{y}_\gamma] =& \mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu}_\alpha)(\mathbf{y}_\gamma - \boldsymbol{\nu}_\gamma)^\top] \\
=& \mathbb{E}\left[\left(\sum_{\beta=1}^{N} c_{\alpha\beta}(\mathbf{x}_\beta - \boldsymbol{\mu}_\beta)\right)\left(\sum_{\xi=1}^{N} c_{\gamma\xi}(\mathbf{x}_\xi - \boldsymbol{\mu}_\xi)^\top\right)\right] \\
=& \sum_{\beta=1}^{N}\sum_{\xi=1}^{N} c_{\alpha\beta}c_{\gamma\xi}\mathbb{E}\left[(\mathbf{x}_\beta - \boldsymbol{\mu}_\beta)(\mathbf{x}_\xi - \boldsymbol{\mu}_\xi)^\top\right] \\
=& \sum_{\beta=1}^{N}\sum_{\xi=1}^{N} c_{\alpha\beta}c_{\gamma\xi}\delta_{\beta\xi}\boldsymbol{\Sigma} = \sum_{\beta=1}^{N} c_{\alpha\beta}c_{\gamma\beta}\boldsymbol{\Sigma},
\end{aligned}$$

where

$$\delta_{\beta\xi} = \begin{cases} 1, & \text{if } \beta = \xi, \\ 0, & \text{if } \beta \neq \xi. \end{cases}$$

If $\alpha = \gamma$, we have $\sum_{\beta=1}^{N} c_{\alpha\beta}c_{\gamma\beta} = \sum_{\beta=1}^{N} c_{\alpha\beta}c_{\alpha\beta} = 1$; otherwise, we have $\sum_{\beta=1}^{N} c_{\alpha\beta}c_{\gamma\beta} = 0$. Hence, we have

$$\mathrm{Cov}[\mathbf{y}_\alpha, \mathbf{y}_\gamma] = \sum_{\beta=1}^{N} c_{\alpha\beta}c_{\gamma\beta}\boldsymbol{\Sigma} = \begin{cases} \boldsymbol{\Sigma}, & \text{if } \alpha = \gamma, \\ \mathbf{0}, & \text{if } \alpha \neq \gamma. \end{cases}$$

The set of vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ have a joint normal distribution, we have proved $\mathrm{Cov}[\mathbf{y}_\alpha] = \boldsymbol{\Sigma}$ for $\alpha = 1, \ldots, N$ and $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent. □

Lemma 4.3. *If*

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix} = \begin{bmatrix} c_1^\top \\ c_2^\top \\ \vdots \\ c_p^\top \end{bmatrix} \in \mathbb{R}^{p\times p}$$

is orthogonal, then $\sum_{\alpha=1}^{N} \mathbf{x}_\alpha\mathbf{x}_\alpha^\top = \sum_{\beta=1}^{N} \mathbf{y}_\alpha\mathbf{y}_\alpha^\top$ *where* $\mathbf{y}_\alpha = \sum_{\beta=1}^{N} c_{\alpha\beta}\mathbf{x}_\alpha$ *for* $\alpha = 1, \ldots, N$.

Proof. Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_p^\top \end{bmatrix} \in \mathbb{R}^{p\times p}.$$

This document is organized by Luo Luo.
Please do not distribute.

We have

$$\sum_{\alpha=1}^{N} \mathbf{y}_\alpha \mathbf{y}_\alpha^\top = \sum_{\beta=1}^{N} \mathbf{X}^\top \mathbf{c}_\alpha \mathbf{c}_\alpha^\top \mathbf{X} = \mathbf{X}^\top \left(\sum_{\beta=1}^{N} \mathbf{c}_\alpha \mathbf{c}_\alpha^\top \right) \mathbf{X} = \mathbf{X}^\top \left(\mathbf{C}^\top \mathbf{C} \right) \mathbf{X} = \mathbf{X}^\top \mathbf{X} = \sum_{\beta=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top.$$

□

Theorem 4.6. *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be independent, each distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the mean of the sample*

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{\alpha=1}^{N} \mathbf{x}_\alpha$$

is distributed according to $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma})$ and independent of

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

Additionally, we have $N\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$, where $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $\alpha = 1, \dots, N$, and $\mathbf{z}_1, \dots, \mathbf{z}_{N-1}$ are independent.

Proof. There exists an orthogonal matrix $\mathbf{B} \in \mathbb{R}^{p \times p}$ such that

$$\mathbf{B} = \begin{bmatrix} \times & \times & \dots & \times \\ \times & \times & \dots & \times \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \dots & \frac{1}{\sqrt{N}} \end{bmatrix}$$

Let $\mathbf{A} = N\hat{\boldsymbol{\Sigma}}$ and let $\mathbf{z}_\alpha = \sum_{\beta=1}^{N} b_{\alpha\beta} \mathbf{x}_\beta$, then

$$\mathbf{z}_N = \sum_{\beta=1}^{N} b_{N\beta} \mathbf{x}_\beta = \sum_{\beta=1}^{N} \frac{\mathbf{x}_\beta}{\sqrt{N}} = \frac{N\bar{\mathbf{x}}}{\sqrt{N}} = \sqrt{N}\bar{\mathbf{x}}.$$

By Lemma 4.3 with $\mathbf{C} = \mathbf{B}$ and $\mathbf{y}_\alpha = \mathbf{z}_\alpha$, we have

$$\begin{aligned}
\mathbf{A} &= \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \bar{\mathbf{x}}^\top - \sum_{\alpha=1}^{N} \bar{\mathbf{x}} \mathbf{x}_\alpha^\top + \sum_{\alpha=1}^{N} \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top - \mathbf{z}_N \mathbf{z}_N^\top \\
&= \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Lemma 4.3 also states $\mathbf{z}_N$ is independent of $\mathbf{z}_1, \dots, \mathbf{z}_{N-1}$, then the mean vector $\bar{\mathbf{x}} = \frac{1}{\sqrt{N}}\mathbf{z}_N$ is independent of $\mathbf{A}$ and $\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{A}$. Since

$$\bar{\mathbf{x}} = \frac{1}{\sqrt{N}}\mathbf{z}_n = \frac{1}{\sqrt{N}}\sum_{\beta=1}^{N} b_{N\beta}\mathbf{x}_\beta,$$

Theorem 4.5 implies ($b_{N\beta} = 1/\sqrt{N}$)

$$\mathbb{E}[\bar{\mathbf{x}}] = \mathbb{E}\left[\frac{1}{\sqrt{N}}\sum_{\beta=1}^{N} b_{N\beta}\mathbf{x}_\beta\right] = \frac{1}{\sqrt{N}}\sum_{\beta=1}^{N} b_{N\beta}\boldsymbol{\mu} = \frac{1}{\sqrt{N}}\sum_{\beta=1}^{N} \frac{\boldsymbol{\mu}}{\sqrt{N}} = \boldsymbol{\mu} \tag{14}$$

and

$$\mathrm{Cov}[\bar{\mathbf{x}}] = \frac{1}{N}\mathrm{Cov}\left[\sum_{\beta=1}^{N} b_{N\beta}\mathbf{x}_\beta\right] = \frac{1}{N}\mathbf{\Sigma}. \tag{15}$$

Hence, we have $\bar{\mathbf{x}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{N}\mathbf{\Sigma}\right)$. For $\alpha = 1, \dots, N-1$, we also have

$$\mathbb{E}[\mathbf{z}_\alpha] = \mathbb{E}\left[\sum_{\beta=1}^{N} b_{\alpha\beta}\mathbf{x}_\beta\right] = \sum_{\beta=1}^{N} b_{\alpha\beta}\mathbb{E}\left[\mathbf{x}_\beta\right] = \sum_{\beta=1}^{N} b_{\alpha\beta}\boldsymbol{\mu} = \sum_{\beta=1}^{N} b_{\alpha\beta}b_{N\beta}\sqrt{N}\boldsymbol{\mu} = \mathbf{0}.$$

Theorem 4.5 also implies $\mathbf{z}_1, \dots, \mathbf{z}_N$ are independent with normal distribution and $\mathrm{Cov}[\mathbf{z}_\alpha] = \mathbf{\Sigma}$. □

Theorem 4.7. *Using the notation of Theorem 4.1, if $N > p$, the probability is 1 of drawing a sample so that*

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

is positive definite.

Proof. The proof of Theorem 4.1 shows that $\mathbf{A} = N\hat{\mathbf{\Sigma}} = \mathbf{Z}^\top\mathbf{Z}$ where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_{N-1}^\top \end{bmatrix} \in \mathbb{R}^{(N-1)\times p},$$

which means $\mathrm{rank}\big(\hat{\mathbf{\Sigma}}\big) = \mathrm{rank}(\mathbf{A}) = \mathrm{rank}\big(\mathbf{Z}\big)$. Then the probability is 1 of $\hat{\mathbf{\Sigma}} \succ \mathbf{0}$ is equivalent to

$$\Pr\big(\mathrm{rank}\big(\mathbf{Z}\big) = p\big) = 1.$$

Since appending rows at the end of $\mathbf{Z}$ will not increase its rank, we only needs to consider the case of $N = p + 1$ ($N - 1 = p$ and $\mathbf{Z} \in \mathbb{R}^{p\times p}$). We have

$$\begin{aligned}
&\Pr(\mathbf{z}_1, \dots, \mathbf{z}_p \text{ are linearly dependent})\\
=&\Pr\left(\bigcup_{i=1}^{p}\mathbf{z}_i \in \mathrm{span}\{\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_i, \dots, \mathbf{z}_p\}\right)\\
\leq&\sum_{i=1}^{p}\Pr\big(\mathbf{z}_i \in \mathrm{span}\{\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_i, \dots, \mathbf{z}_p\}\big)\\
=&p\Pr\big(\mathbf{z}_1 \in \mathrm{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}\big)
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
&= p\,\mathbb{E}\left[\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \mathbf{z}_3, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)\right] \\
&\leq p\mathbb{E}\left[\Pr\left(\text{there exists non-zero } \boldsymbol{\gamma} \in \mathbb{R}^p \text{ such that } \boldsymbol{\gamma}^\top \mathbf{z}_1 = \mathbf{0} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)\right] \\
&= p\mathbb{E}[0] = 0.
\end{aligned}
$$

The third equality is obtained as follows

$$
\begin{aligned}
&\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}\right) \\
=&\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right) f\left(\mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right) \mathrm{d}\boldsymbol{\alpha}_2 \dots \mathrm{d}\boldsymbol{\alpha}_p \\
=&\mathbb{E}\left[\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)\right].
\end{aligned}
$$

The second inequality is due to

$$
\begin{aligned}
&\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \mathbf{z}_3, \dots, \mathbf{z}_p\} \\
\Longrightarrow&\text{there exists } \boldsymbol{\beta} \in \mathbb{R}^{p-1} \text{ such that } \mathbf{z}_1 = [\mathbf{z}_2, \dots, \mathbf{z}_p]\boldsymbol{\beta} \\
\Longrightarrow&\text{there exists } \boldsymbol{\beta} \in \mathbb{R}^{p-1} \text{ and non-zero } \boldsymbol{\gamma} \in \mathbb{R}^p \text{ such that } \boldsymbol{\gamma}^\top \mathbf{z}_1 = \boldsymbol{\gamma}^\top [\mathbf{z}_2, \dots, \mathbf{z}_p]\boldsymbol{\beta} = 0 \\
&\left(\text{the columns of } [\mathbf{z}_2, \dots, \mathbf{z}_p]^\top = \begin{bmatrix} \mathbf{z}_2^\top \\ \vdots \\ \mathbf{z}_p^\top \end{bmatrix} \in \mathbb{R}^{(p-1)\times p} \text{ are linearly dependent, which means} \right.\\
&\left.\quad \text{there exists } \boldsymbol{\gamma} \neq \mathbf{0} \text{ such that } [\mathbf{z}_2, \dots, \mathbf{z}_p]^\top \boldsymbol{\gamma} = \mathbf{0}\right).
\end{aligned}
$$

The fourth equality is due to $\boldsymbol{\gamma}^\top \mathbf{z}_1 \sim \mathcal{N}(0, \boldsymbol{\gamma}^\top \boldsymbol{\Sigma} \boldsymbol{\gamma})$ and $\boldsymbol{\gamma}^\top \boldsymbol{\Sigma} \boldsymbol{\gamma} > 0$ for any nonzero $\boldsymbol{\gamma}$ since $\boldsymbol{\Sigma} \succ \mathbf{0}$. □

This document is organized by Luo Luo.
Please do not distribute.

Unbiasedness and Sufficiency: Consider the result of MLE for normal distribution:

1. We have

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

and

$$\mathbb{E}\big[\hat{\boldsymbol{\Sigma}}\big] = \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top\right] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

For general case, it also holds by the relations (see Theorem 2.4)

$$\boldsymbol{\Sigma} = \operatorname{Cov}[\mathbf{x}_\alpha] = \mathbb{E}\left[\mathbf{x}_\alpha \mathbf{x}_\alpha^\top\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \qquad \text{and} \qquad \frac{1}{N}\boldsymbol{\Sigma} = \operatorname{Cov}[\bar{\mathbf{x}}] = \mathbb{E}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

We also have

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

2. The sample covariance

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$$

   is an unbiased estimator of $\mathbf{\Sigma}$.

Theorem 4.8. *A statistic $\mathbf{t}(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$ if and only if the density $f(\mathbf{y};\boldsymbol{\theta})$ can be factored as*

$$f(\mathbf{y};\boldsymbol{\theta}) = g(\mathbf{t}(\mathbf{y});\boldsymbol{\theta})h(\mathbf{y}),$$

where $g(\mathbf{t}(\mathbf{y});\boldsymbol{\theta})$ and $h(\mathbf{y})$ are nonnegative and $h(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$.

Proof. The proof of this result for general case is not easy. Please refer to Bahadur [2]. □

Theorem 4.9. *If $\mathbf{x}_1,\dots,\mathbf{x}_N$ are independent observations from $\mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$, then*

1. *$\bar{\mathbf{x}}$ and $\mathbf{S}$ are sufficient for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$;*
2. *if $\boldsymbol{\mu}$ is given, $\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top$ is sufficient for $\mathbf{\Sigma}$;*
3. *if $\mathbf{\Sigma}$ is given, $\bar{\mathbf{x}}$ is sufficient for $\boldsymbol{\mu}$;*

where

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

Proof. The density of $\mathbf{x}_1,\dots,\mathbf{x}_N$ is

$$\begin{aligned}
&\prod_{\alpha=1}^{M} n(\mathbf{x}_\alpha \mid \boldsymbol{\mu},\mathbf{\Sigma})\\
=&(2\pi)^{-\frac{pN}{2}}\left(\det(\mathbf{\Sigma})\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right)\\
=&(2\pi)^{-\frac{pN}{2}}\left(\det(\mathbf{\Sigma})\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\right)\right) &(16)\\
=&(2\pi)^{-\frac{pN}{2}}\left(\det(\mathbf{\Sigma})\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\left(N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + (N-1)\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right)\right)\right) &(17)
\end{aligned}$$

where the last step is due to

$$\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\\
=&\sum_{\alpha=1}^{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + \sum_{\alpha=1}^{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\\
&+\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\\
=&N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + (N-1)\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right).
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

1. Let

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}, \qquad \mathbf{y} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \qquad \text{and} \qquad \mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \{\bar{\mathbf{x}}, \mathbf{S}\},$$

   then the expression (17) has the form of

$$g(\bar{\mathbf{x}}, \mathbf{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N), \boldsymbol{\theta}) h(\mathbf{x}_1, \dots, \mathbf{x}_N),$$

   with $h(\mathbf{x}_1, \dots, \mathbf{x}_N) = 1$.

2. We fix $\boldsymbol{\mu}$ and let

$$\boldsymbol{\theta} = \{\boldsymbol{\Sigma}\}, \qquad \mathbf{y} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \qquad \text{and} \qquad \mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top,$$

   then the expression (16) has the form of

$$g\left(\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top, \boldsymbol{\Sigma}\right) = g(\mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N), \boldsymbol{\theta}) h(\mathbf{x}_1, \dots, \mathbf{x}_N),$$

   with $h(\mathbf{x}_1, \dots, \mathbf{x}_N) = 1$.

3. We fix $\boldsymbol{\Sigma}$ and let

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}\}, \qquad \mathbf{y} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \qquad \text{and} \qquad \mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \bar{\mathbf{x}},$$

   then the expression (17) has the form of

$$g(\boldsymbol{\mu}, \bar{\mathbf{x}}) \exp\left(-\frac{(N-1)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}{2}\right) = g(\boldsymbol{\mu}, \bar{\mathbf{x}}) h(\mathbf{x}_1, \dots, \mathbf{x}_N),$$

   with $h(\mathbf{x}_1, \dots, \mathbf{x}_N) = \exp\left(-\frac{(N-1)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}{2}\right)$.

4. If we fix $\boldsymbol{\mu}$, the statistics $\mathbf{S}$ is not sufficient for $\boldsymbol{\Sigma}$. Consider that $\mathbf{S}$ and $\bar{\mathbf{x}}$ are independent and the information of the expression (17) cannot be only included by $\boldsymbol{\Sigma}$.

□

This document is organized by Luo Luo.
Please do not distribute.

Completeness: A family of distributions of $\mathbf{t}$ indexed by $\boldsymbol{\theta}$ is complete if for every real-valued function $g(\mathbf{t})$, we have

$$\mathbb{E}[g(\mathbf{t})] \equiv 0$$

identically in $\boldsymbol{\theta}$ implies $g(\mathbf{t}) = 0$ except for a set of $\mathbf{y}$ of probability 0 for every $\boldsymbol{\theta}$.

1. We can think $\mathbb{E}[g(\mathbf{t})] \equiv 0$ as function g does not provide information related to parameter $\boldsymbol{\theta}$.

2. The condition "$\mathbb{E}[g(\mathbf{t})] \equiv 0$ implies $g(\mathbf{t}) = 0$ except for a set of $\mathbf{t}$ of probability 0" says g does not provide information related to parameter $\boldsymbol{\theta}$ implies $g(\mathbf{t}) = 0$ almost everywhere.

3. If $g(\mathbf{t}) = 0$ does not hold almost everywhere, then g do provide information related to parameter $\boldsymbol{\theta}$.

4. All information of the variable $\mathbf{t}$ (almost everywhere) are useful to inference parameter $\boldsymbol{\theta}$.

5. There is no useless information in $\mathbf{t}$.

Theorem 4.10 (Anderson [1, Theorem 3.4.2])**.** *The sufficient set of statistics $\{\bar{\mathbf{x}}, \mathbf{S}\}$ is complete for $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ when the sample is drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Sufficiency and Cramer-Rao Inequality We first give some lemmas. We denote the density with parameter $\boldsymbol{\theta}$ by $f(\mathbf{x}, \boldsymbol{\theta})$ and

$$\mathbf{s} = \frac{\partial \ln g(\mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where g is the density on N samples and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and each $\mathbf{x}_\alpha$ has density $f(\mathbf{x}_\alpha, \boldsymbol{\theta})$. In the following analysis, we assume everything is well-defined.

Lemma 4.4. *We have* $\mathbb{E}[\mathbf{s}] = \mathbf{0}$*.*

Proof. We have

$$\begin{aligned}
\mathbb{E}[s_j] &= \int g(\mathbf{X}, \boldsymbol{\theta}) \frac{\partial \ln g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_j} \,\mathrm{d}\mathbf{X} \\
&= \int g(\mathbf{X}, \boldsymbol{\theta}) \frac{1}{g(\mathbf{X}, \boldsymbol{\theta})} \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_j} \,\mathrm{d}\mathbf{X} \\
&= \int \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_j} \,\mathrm{d}\mathbf{X} \\
&= \frac{\partial}{\partial \theta_j} \int g(\mathbf{X}, \boldsymbol{\theta}) \,\mathrm{d}\mathbf{X} \\
&= \frac{\partial}{\partial \theta_j} 1 = 0.
\end{aligned}$$

□

Remark 4.3. *Similarly, we also have*

$$\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

Lemma 4.5. *For unbiased estimator* $\mathbf{t}$ *of* $\boldsymbol{\theta}$ *(i.e.,* $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$*) based on* $\mathbf{X}$*, we have* $\mathrm{Cov}[\mathbf{t}, \mathbf{s}] = \mathbf{I}$.

Proof. We have

$$\begin{aligned}
&\mathrm{Cov}[t_j s_k] = \mathbb{E}\big[(t_j - \mathbb{E}[t_j])(s_k - \mathbb{E}[s_k])\big] \\
&= \int (t_j - \theta_j) \frac{\partial \ln g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} g(\mathbf{X}, \boldsymbol{\theta}) \,\mathrm{d}\mathbf{X} \\
&= \int (t_j - \theta_j) \frac{1}{g(\mathbf{X}, \boldsymbol{\theta})} \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} g(\mathbf{X}, \boldsymbol{\theta}) \,\mathrm{d}\mathbf{X} \\
&= \int (t_j - \theta_j) \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} \,\mathrm{d}\mathbf{X} \\
&= -\int g(\mathbf{X}, \boldsymbol{\theta}) \frac{\partial (t_j - \theta_j)}{\partial \theta_k} \,\mathrm{d}\mathbf{X} = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}
\end{aligned}$$

For the last line, we use the integrate by part

$$\begin{aligned}
&\int (t_j - \theta_j) \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} \,\mathrm{d}\theta_k \\
=& \int (t_j - \theta_j) \,\mathrm{d}g(\mathbf{X}, \boldsymbol{\theta}) \\
=& (t_j - \theta_j) g(\mathbf{X}, \boldsymbol{\theta}) - \int g(\mathbf{X}, \boldsymbol{\theta}) \,\mathrm{d}(t_j - \theta_j) \\
=& (t_j - \theta_j) g(\mathbf{X}, \boldsymbol{\theta}) - \int g(\mathbf{X}, \boldsymbol{\theta}) \frac{\partial (t_j - \theta_j)}{\partial \theta_k} \,\mathrm{d}\theta_k.
\end{aligned} \tag{18}$$

This document is organized by Luo Luo.
Please do not distribute.

Since we have

$$\mathbb{E}[t_j] = \theta_j \quad \Longleftrightarrow \quad 0 = \mathbb{E}[t_j - \theta_j] = \int (t_j - \theta_j) g(\mathbf{X}, \boldsymbol{\theta}) \, \mathrm{d}\mathbf{X},$$

taking integral w.r.t $\mathbf{X}$ on equation (18) leads to

$$\iint (t_j - \theta_j) \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} \, \mathrm{d}\theta_k \, \mathrm{d}\mathbf{X} = - \iint g(\mathbf{X}, \boldsymbol{\theta}) \frac{\partial (t_j - \theta_j)}{\partial \theta_k} \, \mathrm{d}\theta_k \, \mathrm{d}\mathbf{X}.$$

Taking derivative w.r.t θ_k on above equation means

$$\int (t_j - \theta_j) \frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_k} \, \mathrm{d}\mathbf{X} = - \int g(\mathbf{X}, \boldsymbol{\theta}) \frac{\partial (t_j - \theta_j)}{\partial \theta_k} \, \mathrm{d}\mathbf{X}.$$

□

Theorem 4.11. *Following notations and assumptions in this part, we have*

$$N\mathbb{E}\left[(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^\top\right] \succeq \left(\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right]\right)^{-1}.$$

Proof. For any constant nonzero vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, consider the correlation of $\mathbf{a}^\top \mathbf{t}$ and $\mathbf{b}^\top \mathbf{s}$, we have

$$1 \geq \frac{\mathrm{Cov}[\mathbf{a}^\top \mathbf{t}, \mathbf{b}^\top \mathbf{s}]}{\sqrt{\mathrm{Var}[\mathbf{a}^\top \mathbf{t}]\mathrm{Var}[\mathbf{b}^\top \mathbf{s}]}} = \frac{\mathbf{a}^\top \mathrm{Cov}[\mathbf{t}, \mathbf{s}] \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathrm{Cov}[\mathbf{t}] \mathbf{a}} \sqrt{\mathbf{b}^\top \mathrm{Cov}[\mathbf{s}] \mathbf{b}}} = \frac{\mathbf{a}^\top \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathrm{Cov}[\mathbf{t}] \mathbf{a}} \sqrt{\mathbf{b}^\top \mathrm{Cov}[\mathbf{s}] \mathbf{b}}}$$

Let $\mathbf{b} = (\mathrm{Cov}[\mathbf{s}])^{-1}\mathbf{a}$, we have

$$1 \geq \frac{\mathbf{a}^\top (\mathrm{Cov}[\mathbf{s}])^{-1} \mathbf{a}}{\sqrt{\mathbf{a}^\top \mathrm{Cov}[\mathbf{t}] \mathbf{a}} \sqrt{\mathbf{a}^\top (\mathrm{Cov}[\mathbf{s}])^{-1} \mathbf{a}}},$$

which means

$$\mathbf{a}^\top \mathrm{Var}[\mathbf{t}] \mathbf{a} \geq \mathbf{a}^\top \left(\mathrm{Var}[\mathbf{s}]\right)^{-1} \mathbf{a}$$

for any $\mathbf{a}$, i.e., $\mathrm{Var}[\mathbf{t}] \succeq \left(\mathrm{Var}[\mathbf{s}]\right)^{-1}$. Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^\top\right] \\
=&\mathrm{Var}[\mathbf{t}] \succeq \left(\mathrm{Var}[\mathbf{s}]\right)^{-1} \\
=&\left(\mathrm{Var}\left[\frac{\partial \ln g(\mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right)^{-1} \\
=&\left(\mathrm{Var}\left[\frac{\partial \ln \prod_{\alpha=1}^N f(\mathbf{x}_\alpha, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right)^{-1} \\
=&\left(\sum_{\alpha=1}^N \mathrm{Var}\left[\frac{\partial \ln f(\mathbf{x}_\alpha, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right)^{-1} \\
=&\left(N\mathrm{Var}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right)^{-1} \\
=&\frac{1}{N}\left(\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right]\right)^{-1},
\end{aligned}$$

where the last step use the statement in Remark 4.3.

□

This document is organized by Luo Luo. Please do not distribute.

Remark 4.4. *For sample $\mathbf{x}_1, \dots, \mathbf{x}_N$ from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we let $\mathbf{t} = \bar{\mathbf{x}}$, $\boldsymbol{\theta} = \boldsymbol{\mu}$ and*

$$f(\mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{x}, \boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

Then we have

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

We also have

$$\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x}),$$

which leads to

$$\begin{aligned} &\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right] \\ =&\mathbb{E}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\right] \\ =&\mathrm{Cov}\left[\boldsymbol{\Sigma}^{-1}\mathbf{x}\right] \\ =&\boldsymbol{\Sigma}^{-1}\mathrm{Cov}\left[\mathbf{x}\right]\boldsymbol{\Sigma}^{-1} \\ =&\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Hence, this case leads to the matrices in Theorem 4.11 are equal, which means $\bar{\mathbf{x}}$ is efficient.

Remark 4.5. *If we take $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, then the unbiased estimator $\{\bar{\mathbf{x}}, \mathbf{S}\}$ has efficiency $(N - 1/N)^{p(p+1)/2}$. The proof of this result is extremely complicated, please see the provided by TA in this link.*

Theorem 4.12. *Under the regularity condition, we have*

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))^\top\right] = -\mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))\right].$$

Proof. Recall that

$$\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta})) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x}, \boldsymbol{\theta})},$$

then we have

$$\begin{aligned} &\nabla^2_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta})) \\ =&\mathbf{J}_{\boldsymbol{\theta}}\left(\frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x}, \boldsymbol{\theta})}\right) \\ =&\frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) f(\mathbf{x}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})^\top}{(f(\mathbf{x}, \boldsymbol{\theta}))^2} \\ =&\frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x}, \boldsymbol{\theta})} - \nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))^\top. \end{aligned}$$

Since

$$\begin{aligned} \mathbb{E}\left[\frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x}, \boldsymbol{\theta})}\right] &= \int \frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x}, \boldsymbol{\theta})} f(\mathbf{x}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{x} \\ = \int \nabla^2_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{x} &= \nabla^2_{\boldsymbol{\theta}}\left(\int f(\mathbf{x}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{x}\right) = \nabla^2_{\boldsymbol{\theta}} 1 = 0, \end{aligned}$$

we have

$$\mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}}(\ln f(\mathbf{x}, \boldsymbol{\theta}))\right]$$

This document is organized by Luo Luo. Please do not distribute.

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta})}{f(\mathbf{x},\boldsymbol{\theta})}\right] - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x},\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x},\boldsymbol{\theta}))^\top\right] \\
&= -\,\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x},\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}(\ln f(\mathbf{x},\boldsymbol{\theta}))^\top\right].
\end{aligned}
$$

□

Remark 4.6. *In some statistical learning model, we let $f(\mathbf{x},\boldsymbol{\theta})$ be the probability/density corresponds to random sample $\mathbf{x}$ (including feature and label) with parameter $\boldsymbol{\theta}$. Maximizing the expected likelihood associate leads to the formulation*

$$
\max_{\boldsymbol{\theta}\in\mathbb{R}^p} L(\boldsymbol{\theta}) \triangleq \mathbb{E}[l(\mathbf{x},\boldsymbol{\theta})],
$$

where $l(\mathbf{x},\boldsymbol{\theta}) \triangleq \ln f(\mathbf{x},\boldsymbol{\theta})$. Theorem 4.12 means

$$
\nabla^2 L(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}}(\ln f(\mathbf{x},\boldsymbol{\theta}))\right] = -\mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\mathbf{x},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} l(\mathbf{x},\boldsymbol{\theta})^\top].
$$

For given dataset $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$, we typically approximate

$$
L(\boldsymbol{\theta}) = \mathbb{E}[l(\mathbf{x},\boldsymbol{\theta})] \approx \frac{1}{n}\sum_{\alpha=1}^n l(\mathbf{x}_\alpha,\boldsymbol{\theta})
$$

that leads to

$$
\min_{\boldsymbol{\theta}\in\mathbb{R}^p} F(\boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{\alpha=1}^n -l(\mathbf{x}_\alpha,\boldsymbol{\theta}), \quad \textit{such that } F(\boldsymbol{\theta}) \approx -L(\boldsymbol{\theta}).
$$

We can also introduce

$$
\nabla^2 F(\boldsymbol{\theta}) \approx -L(\boldsymbol{\theta}) = \mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\mathbf{x},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} l(\mathbf{x},\boldsymbol{\theta})^\top] \approx \frac{1}{n}\sum_{\alpha=1}^n \nabla_{\boldsymbol{\theta}} l(\mathbf{x}_\alpha,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} l(\mathbf{x}_\alpha,\boldsymbol{\theta})^\top
$$

to achieve approximate Newton iteration

$$
\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \left(\frac{1}{n}\sum_{\alpha=1}^n \nabla_{\boldsymbol{\theta}} l(\mathbf{x}_\alpha,\boldsymbol{\theta}^t)\nabla_{\boldsymbol{\theta}} l(\mathbf{x}_\alpha,\boldsymbol{\theta}^t)^\top\right)^{-1} \nabla F(\boldsymbol{\theta}^t).
$$

This document is organized by Luo Luo.
Please do not distribute.

Consistency and Asymptotic Normality Let sequences of random variables $\{x_n\}$ and $\{y_n\}$ satisfies

$$\operatorname*{plim}_{n\to+\infty} x_n = a \qquad \text{and} \qquad \operatorname*{plim}_{n\to+\infty} y_n = b,$$

then we have

$$\lim_{n\to+\infty} \Pr\left(|x_n - a| < \frac{\epsilon}{2}\right) = 1 \qquad \text{and} \qquad \lim_{n\to+\infty} \Pr\left(|y_n - b| < \frac{\epsilon}{2}\right) = 1$$

for any $\epsilon > 0$. If $|x_n + y_n - (a+b)| \geq \epsilon$, it must lead to $|x_n - a| \geq \epsilon/2$ or $|y_n - b| \geq \epsilon/2$. Then we have

$$\begin{aligned}
&\lim_{n\to+\infty} \Pr(|x_n + y_n - (a+b)| \geq \epsilon) \\
\leq &\lim_{n\to+\infty} \Pr\left(|x_n - a| \geq \frac{\epsilon}{2} \cup |y_n - b| \geq \frac{\epsilon}{2}\right) \\
\leq &\lim_{n\to+\infty} \Pr\left(|x_n - a| \geq \frac{\epsilon}{2}\right) + \Pr\left(|y_n - b| \geq \frac{\epsilon}{2}\right) = 0,
\end{aligned}$$

which means

$$\operatorname*{plim}_{n\to+\infty} (x_n + y_n) = a + b.$$

The other properties of plim is also similar to lim.

Theorem 4.13. *For sample $\mathbf{x}_1, \mathbf{x}_2 \ldots$ are independently and identically distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the estimators*

$$\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha \qquad \textit{and} \qquad \mathbf{S}_N = \frac{1}{N-1}\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}}_N)(\mathbf{x}_\alpha - \bar{\mathbf{x}}_N)^\top$$

are consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.

Proof. The weak law of large numbers (also called Khinchin's law) states that the sample average converges in probability towards the expected value. This directly means $\bar{\mathbf{x}}_N$ is a consistent estimator of $\boldsymbol{\mu}$.

We can write

$$\begin{aligned}
\mathbf{S} =& \frac{1}{N-1}\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \\
=& \frac{1}{N-1}\sum_{\alpha=1}^{N} \big((\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top + (\boldsymbol{\mu} - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top + (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^\top + (\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^\top\big) \\
=& \frac{1}{N-1}\sum_{\alpha=1}^{N} \big((\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top + (\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^\top\big) \\
=& \frac{N}{N-1} \cdot \frac{1}{N}\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top + \frac{N}{N-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^\top
\end{aligned}$$

and its entry satisfies

$$s_{ij} = \frac{N}{N-1} \cdot \frac{1}{N}\sum_{\alpha=1}^{N} (x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j) + \frac{N}{N-1}(\mu_i - \bar{x}_i)(\mu_j - \bar{x}_j).$$

Taking the limit in probability, we obtain

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} s_{ij} =& \operatorname*{plim}_{N\to\infty} \frac{N}{N-1} \cdot \frac{1}{N}\sum_{\alpha=1}^{N} (x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j)^\top + \operatorname*{plim}_{N\to\infty} \frac{N}{N-1}(\mu_i - \bar{x}_i)(\mu_j - \bar{x}_j)^\top \\
=& \mathbb{E}\big[(x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j)^\top\big] = \sigma_{ij}.
\end{aligned}$$

□

This document is organized by Luo Luo.
Please do not distribute.

Theorem 4.14. *Let p-component vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots$ be i.i.d with means $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance matrices $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$. Then the limiting distribution of*

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})$$

as $n \to +\infty$ is $\mathcal{N}(\mathbf{0}, \mathbf{T})$.

Proof. Let

$$\phi_n(\mathbf{t}, u) = \mathbb{E}\left[\exp\left(\mathrm{i}\, u \mathbf{t}^\top \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})\right)\right], \tag{19}$$

where $u \in \mathbb{R}$ and $\mathbf{t} \in \mathbb{R}^p$. For fixed $\mathbf{t}$, the function $\phi_n(\mathbf{t}, u)$ can be viewed as the characteristic function (with respect to u) of the random variable

$$z_n = \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{t}^\top \mathbf{y}_\alpha - \mathbf{t}^\top \mathbb{E}[\mathbf{y}_\alpha]).$$

Note that

$$\mathbb{E}[\mathbf{t}^\top \mathbf{y}_\alpha] = \mathbf{t}^\top \mathbb{E}[\mathbf{y}_\alpha] = \mathbf{t}^\top \boldsymbol{\nu}$$

and

$$\mathrm{Cov}[\mathbf{t}^\top \mathbf{y}_\alpha] = \mathbf{t}^\top \mathrm{Cov}[\mathbf{y}_\alpha] \mathbf{t} = \mathbf{t}^\top \mathbf{T} \mathbf{t}.$$

By the univariate central limit theorem, the limiting distribution of z_n is $\mathcal{N}(0, \mathbf{t}^\top \mathbf{T} \mathbf{t})$. Therefore, we have

$$\lim_{n \to \infty} \phi_n(\mathbf{t}, u) = \exp\left(-\frac{1}{2} u^2 \mathbf{t}^\top \mathbf{T} \mathbf{t}\right),$$

for any $u \in \mathbb{R}$ and $\mathbf{t} \in \mathbb{R}^p$. Let $u = 1$, then we obtain

$$\begin{aligned}
&\lim_{n \to \infty} \phi_n(\mathbf{t}, 1) \\
\overset{(19)}{=}\ & \lim_{n \to \infty} \mathbb{E}\left[\exp\left(\mathrm{i}\, \mathbf{t}^\top \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})\right)\right] \quad // \text{ characteristic function of } \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu}) \\
=\ & \exp\left(-\frac{1}{2} \mathbf{t}^\top \mathbf{T} \mathbf{t}\right)
\end{aligned}$$

for any $\mathbf{t} \in \mathbb{R}^p$. Since $\exp\left(-\mathbf{t}^\top \mathbf{T} \mathbf{t}/2\right)$ is continuous at $\mathbf{t} = \mathbf{0}$, the convergence is uniform in some neighborhood of $\mathbf{t} = \mathbf{0}$. The Applying Theorem 3.11 finish this proof. □

This document is organized by Luo Luo.
Please do not distribute.

Bayesian Statistics In Bayesian statistics, we regard the parameters as a random variable with prior distribution. We first revisit linear regression. Given dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ are the feature and the corresponding label of the i-th data. We suppose

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, N$ and given $\sigma > 0$. We regard $\boldsymbol{\beta}$ as parameter and form the likelihood on random variables $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$. Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times p} \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N.$$

We have

$$\mathbf{X}\boldsymbol{\beta} - \mathbf{y} = \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} \sim \mathcal{N}_N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right).$$

Maximizing the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{-\left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right\|_2^2}{2\sigma^2}\right)$$

can be view as solving the following optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right\|_2^2.$$

Suppose $\mathbf{X}^\top \mathbf{X}$ is non-singular, then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] = \mathbb{E}[\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}] = \boldsymbol{\beta}.$$

Since $\mathrm{Cov}[\mathbf{y}] = \mathrm{Cov}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}$, we have

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \mathrm{Cov}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathrm{Cov}[\mathbf{y}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

and

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\big(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\big).$$

We let $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, then

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} =& \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
=& \mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
=& (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} \\
=& (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)(\mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\epsilon}) \\
=& \mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\epsilon}) \\
=& \mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\epsilon} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon} \\
=& -(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \boldsymbol{\epsilon}
\end{aligned}$$

and

$$\mathbb{E}[\hat{\boldsymbol{\epsilon}}] = -\mathbb{E}[(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \boldsymbol{\epsilon}] = \mathbf{0}.$$

Now we can show $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are uncorrelated, since

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}] =& (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta} \\
=& (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}) \\
=& (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
=& -(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

and

$$\begin{aligned}\mathrm{Cov}\big[\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\epsilon}}\big] =&\mathbb{E}\big[(\hat{\boldsymbol{\beta}}-\mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\epsilon}}-\mathbb{E}[\hat{\boldsymbol{\epsilon}}])^\top\big]\\ =&\mathbb{E}\big[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\big]\\ =&\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\\ =&\sigma^2\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top-(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\right)\\ =&\mathbf{0}.\end{aligned}$$

Now we additionally suppose the parameter has a prior distribution of $\boldsymbol{\beta}\sim\mathcal{N}_p(\mathbf{0},\tau^2\mathbf{I})$. Then the posterior joint density given $\boldsymbol{\epsilon}$ is

$$\begin{aligned}f(\boldsymbol{\beta}\,|\,\boldsymbol{\epsilon}) =&\frac{f(\boldsymbol{\epsilon}\,|\,\boldsymbol{\beta})f(\boldsymbol{\beta})}{f(\boldsymbol{\epsilon})}\\ \propto&f(\boldsymbol{\epsilon}\,|\,\boldsymbol{\beta})f(\boldsymbol{\beta})\\ =&\frac{1}{(2\pi\sigma^2)^N}\exp\left(\frac{-\left\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\right\|_2^2}{2\sigma^2}\right)\cdot\frac{1}{(2\pi\tau^2)^{N/2}}\exp\left(\frac{-\boldsymbol{\beta}^2}{2\tau^2}\right),\end{aligned}$$

which also follows normal distribution. Maximizing above ones corresponds to the optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\frac{1}{2}\left\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\right\|_2^2+\frac{\sigma^2}{2\tau^2}\left\|\boldsymbol{\beta}\right\|_2^2,$$

which is the model of ridge regression. Let $\lambda=\sigma^2/\tau^2>0$, then the solution is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\lambda =&(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\\ =&(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}).\end{aligned}$$

Remark 4.7. *We can ignore the prior distribution of $\boldsymbol{\beta}$ and only consider the form of $\hat{\boldsymbol{\beta}}_\lambda$, then we have*

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] =&\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})]\\ =&\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}]+\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}]\\ =&(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}[\hat{\boldsymbol{\beta}}_\lambda] =&\mathbb{E}[(\hat{\boldsymbol{\beta}}_\lambda-\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda])(\hat{\boldsymbol{\beta}}_\lambda-\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda])^\top]\\ =&\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}]\\ =&\sigma^2(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\\ \preceq&\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}=\mathrm{Cov}[\hat{\boldsymbol{\beta}}].\end{aligned}$$

The last line come from SVD of $\mathbf{X}^\top\mathbf{X}=\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^\top$, which leads to

$$\begin{aligned}&(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\\ =&\mathbf{U}(\boldsymbol{\Gamma}+\lambda\mathbf{I})^{-1}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^\top\mathbf{U}(\boldsymbol{\Gamma}+\lambda\mathbf{I})^{-1}\mathbf{U}^\top\\ =&\mathbf{U}\,\mathrm{diag}\{\gamma_i/(\gamma_i+\lambda)^2\}_{i=1}^p\mathbf{U}^\top\end{aligned}$$

and

$$(\mathbf{X}^\top\mathbf{X})^{-1}=\mathbf{U}\,\mathrm{diag}\{1/\gamma_i\}_{i=1}^p\mathbf{U}^\top.$$

This document is organized by Luo Luo.
Please do not distribute.

Theorem 4.15. *If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independently distributed, each $\mathbf{x}_\alpha$ according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is normal with mean*

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}$$

and covariance matrix

$$\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

Proof. Applying Bayes rule, we have

$$\begin{aligned}
f(\boldsymbol{\mu} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) =& \frac{f(\mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\mu})}{f(\mathbf{x}_1, \ldots, \mathbf{x}_N)} = \frac{f(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \boldsymbol{\mu}) f(\boldsymbol{\mu})}{f(\mathbf{x}_1, \ldots, \mathbf{x}_N)} \\
\propto& f(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \boldsymbol{\mu}) f(\boldsymbol{\mu}) \\
=& \frac{1}{(2\pi)^{pN/2}(\det(\boldsymbol{\Sigma}))^{N/2}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right) \\
&\cdot \frac{1}{(2\pi)^{p/2}(\det(\boldsymbol{\Sigma}))^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\nu})^\top \boldsymbol{\Phi}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})\right).
\end{aligned}$$

Note that we only require focusing on the term related to $\boldsymbol{\mu}$. Consider the term in exponential, we have

$$\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}) \\
=& \sum_{\alpha=1}^{N}\left(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_\alpha\right) \\
=& N\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2N\bar{\mathbf{x}}_\alpha^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + N\mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_\alpha \\
=& \boldsymbol{\mu}^\top\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} - 2\bar{\mathbf{x}}^\top\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} + N\sum_{\alpha=1}^{N}\mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_\alpha
\end{aligned}$$

and

$$(\boldsymbol{\mu} - \boldsymbol{\nu})^\top \boldsymbol{\Phi}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}) = \boldsymbol{\mu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\mu} + \boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\nu}.$$

Summing over above results, we have

$$\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\nu})^\top \boldsymbol{\Phi}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}) \\
=& \boldsymbol{\mu}^\top\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} - 2\bar{\mathbf{x}}^\top\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} + N\sum_{\alpha=1}^{N}\mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_\alpha + \boldsymbol{\mu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\mu} + \boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\nu} \\
=& \boldsymbol{\mu}^\top\left(\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \boldsymbol{\Phi}^{-1}\right)\boldsymbol{\mu} - 2\left(\bar{\mathbf{x}}^\top\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\right)\boldsymbol{\mu} + N\sum_{\alpha=1}^{N}\mathbf{x}_\alpha^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_\alpha + \boldsymbol{\nu}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\nu}.
\end{aligned}$$

This implies the posterior of $\boldsymbol{\mu}$ has normal distribution with covariance (check the quadratic term of $\boldsymbol{\mu}$)

$$\left(\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \boldsymbol{\Phi}^{-1}\right)^{-1} = \boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

This document is organized by Luo Luo.
Please do not distribute.

We can verify this equality as follows

$$\begin{aligned}
\mathbf{I} =& \left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\right)\left(\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \boldsymbol{\Phi}^{-1}\right) \\
=& \left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\right)\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\right)\boldsymbol{\Phi}^{-1} \\
=& \boldsymbol{\Phi}\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} + \mathbf{I} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} \\
\Longleftrightarrow \mathbf{0} =& \left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} - \left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} - \left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} \\
\Longleftrightarrow \mathbf{0} =& \left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right) - \boldsymbol{\Phi} - \frac{1}{N}\boldsymbol{\Sigma},
\end{aligned}$$

where the last step multiplies $\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}$ on left and $\frac{1}{N}\boldsymbol{\Sigma}$ on right. We can find the mean by checking the linear term of $\boldsymbol{\mu}$, that is

$$\begin{aligned}
& \left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\right)\left(\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \boldsymbol{\Phi}^{-1}\boldsymbol{\nu}\right) \\
=& \left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\right)\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \left(\mathbf{I} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\right)\boldsymbol{\nu} \\
=& \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}.
\end{aligned}$$

□

We can also provide another proof of this theorem.

Proof. Since $\bar{\mathbf{x}}$ is sufficient for $\boldsymbol{\mu}$, we need only consider $\bar{\mathbf{x}}$, which has the distribution of $\boldsymbol{\mu} + \mathbf{w}$, where

$$\mathbf{w} \,|\, \boldsymbol{\mu} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{N}\boldsymbol{\Sigma}\right).$$

Consider that

$$\bar{\mathbf{x}} \,|\, \boldsymbol{\mu} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma}\right), \qquad \mathbf{w} \,|\, \boldsymbol{\mu} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{N}\boldsymbol{\Sigma}\right)$$

and

$$\begin{aligned}
f(\mathbf{w}, \boldsymbol{\mu}) =& f(\mathbf{w} \,|\, \boldsymbol{\mu}) f(\boldsymbol{\mu}) \\
=& n\left(\mathbf{w} \,\middle|\, \mathbf{0}, \frac{1}{N}\boldsymbol{\Sigma}\right) n\left(\boldsymbol{\mu} \,|\, \boldsymbol{\nu}, \boldsymbol{\Phi}\right),
\end{aligned}$$

which means $\mathbf{w}$ and $\boldsymbol{\mu}$ are independent. Then we have

$$\bar{\mathbf{x}} = \begin{bmatrix}\mathbf{I} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu} \\ \mathbf{w}\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\boldsymbol{\mu} \\ \mathbf{w}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\nu} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\boldsymbol{\Sigma}\end{bmatrix}\right).$$

Since we have $\bar{\mathbf{x}} = \boldsymbol{\mu} + \mathbf{w}$, then it holds

$$\begin{bmatrix}\boldsymbol{\mu} \\ \bar{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu} \\ \mathbf{w}\end{bmatrix},$$

This document is organized by Luo Luo.
Please do not distribute.

which implies

$$\begin{bmatrix}\boldsymbol{\mu}\\ \bar{\mathbf{x}}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\nu}\\ \boldsymbol{\nu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\Phi}\\ \boldsymbol{\Phi} & \boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\end{bmatrix}\right).$$

Consider the conditional distribution of $\boldsymbol{\mu}$ given $\bar{\mathbf{x}}$, we obtain the mean and covariance given $\bar{\mathbf{x}}$ are

$$\begin{aligned}&\boldsymbol{\nu}+\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}(\bar{\mathbf{x}}-\boldsymbol{\nu})\\ =&\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}}+\left(\mathbf{I}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\right)\boldsymbol{\nu}\\ =&\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}}+\frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}\end{aligned}$$

and

$$\boldsymbol{\Phi}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

□

Remark 4.8. *Let*

$$\mathbf{x}=\begin{bmatrix}\mathbf{x}^{(1)}\\ \mathbf{x}^{(2)}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right).$$

The conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\mathbf{x}^{(1)}\mid\mathbf{x}^{(2)}\sim\mathcal{N}\left(\boldsymbol{\mu}^{(1)}+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)}-\boldsymbol{\mu}^{(2)}), \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\right)$$

The last step in above proof corresponds to setting $\mathbf{x}^{(1)}=\boldsymbol{\mu}$ and $\mathbf{x}^{(2)}=\bar{\mathbf{x}}$.

Remark 4.9. *We have the following intuitions on this theorem:*

1. *If $\boldsymbol{\Phi}$ is small, i.e., close to $\mathbf{0}$, we have*

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}}+\frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}\approx\boldsymbol{\nu}.$$

   and (the second term is high-order w.r.t $\boldsymbol{\Phi}$)

$$\boldsymbol{\Phi}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}\approx\boldsymbol{\Phi}.$$

   This means the prior information is very strong, which heavily affects the posterior estimation.

2. *If $\boldsymbol{\Phi}$ is large (compared with $\boldsymbol{\Sigma}$ and $\mathbf{0}$) we have*

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}}+\frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}\approx\bar{\mathbf{x}}$$

   and

$$\boldsymbol{\Phi}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}=\left(\left(\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}+\boldsymbol{\Phi}^{-1}\right)^{-1}\approx\frac{1}{N}\boldsymbol{\Sigma}.$$

   This means the prior information is not strong, which weakly affects the posterior estimation.

Remark 4.10. *In Bayesian probability theory, if the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate distributions.*

Remark 4.11. *You can find more examples of conjugate prior on Wikipedia. A famous application of Bayesian model is latent Dirichlet allocation [4].*

This document is organized by Luo Luo.
Please do not distribute.

James-Stein Estimator: We show some intuition of James-Stein estimator

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

for each $\mathbf{x}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}, N\mathbf{I})$. In the view of above theorem, we let $\boldsymbol{\Sigma} = N\mathbf{I}$ and $\boldsymbol{\Phi} = \tau^2\mathbf{I}$. That is

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\nu}, \tau^2\mathbf{I}). \tag{20}$$

The variable $\bar{\mathbf{x}}$ has marginal distribution $\mathcal{N}\left(\boldsymbol{\nu}, \boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)$, then

$$\bar{\mathbf{x}} - \boldsymbol{\nu} \sim \mathcal{N}_p\left(\mathbf{0}, (\tau^2+1)\mathbf{I}\right), \qquad \frac{\bar{\mathbf{x}} - \boldsymbol{\nu}}{\sqrt{\tau^2+1}} \sim \mathcal{N}_p\left(\mathbf{0}, \mathbf{I}\right) \qquad \text{and} \qquad \frac{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}{\tau^2+1} \sim \chi^2(p).$$

This implies

$$\frac{\tau^2+1}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2} \sim \chi^{-2}(p)$$

and

$$\mathbb{E}\left[\frac{\tau^2+1}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right] = \frac{1}{p-2} \qquad \Longrightarrow \qquad \mathbb{E}\left[\frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right] = \frac{1}{\tau^2+1}.$$

The posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \dots, \mathbf{x}_N$ has mean

$$\begin{aligned}
&\boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu} \\
=&\frac{\tau^2}{\tau^2+1}\bar{\mathbf{x}} + \frac{1}{\tau^2+1}\boldsymbol{\nu} \\
=&\left(1 - \frac{1}{\tau^2+1}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu} \\
=&\left(1 - \mathbb{E}\left[\frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right]\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.
\end{aligned}$$

Interestingly, taking the expectation of the factor before term $\bar{\mathbf{x}} - \boldsymbol{\nu}$ in James-Stein estimator lead to the posterior distribution of $\boldsymbol{\mu}$ in Bayesian model.

It is amazing that biased estimator leads to smaller mean-squared error in expectation, even if there is no assumption of prior distribution.

Lemma 4.6. *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed to $\mathcal{N}_p(\boldsymbol{\mu}, N\mathbf{I})$, we have*

$$\mathbb{E}\left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2\right] = \sum_{\alpha=1}^{p} \mathrm{Var}(\bar{x}_\alpha) = p.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2\right] =&\mathbb{E}\left[\mathrm{tr}\big((\bar{\mathbf{x}} - \boldsymbol{\mu})^\top(\bar{\mathbf{x}} - \boldsymbol{\mu})\big)\right] \\
=&\mathbb{E}\left[\mathrm{tr}\big((\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\big)\right] \\
=&\mathrm{tr}\big(\mathbb{E}[(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top]\big) \\
=&\mathrm{tr}\big(\mathbb{E}[\mathrm{Cov}(\bar{\mathbf{x}})]\big) \\
=&\mathrm{tr}\big(\mathbf{I}\big) = p.
\end{aligned}$$

□

This document is organized by Luo Luo.
Please do not distribute

Lemma 4.7. *If $f(x)$ is a function such that*

$$f(b) - f(a) = \int_a^b f'(x)\,\mathrm{d}x$$

for all $a < b$ and if

$$\int_{-\infty}^{+\infty} |f'(x)| \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x < +\infty,$$

then

$$\int_{-\infty}^{+\infty} f(x)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x = \int_{-\infty}^{+\infty} f'(x) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x. \tag{21}$$

Proof. Since

$$\int_{-\infty}^{+\infty} (x-\theta) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)\,\mathrm{d}x = 0, \tag{22}$$

the LHS of (21) can be written as (subtract equation (22) times $f(\theta)$)

$$\begin{aligned}
&\int_{-\infty}^{+\infty} (f(x) - f(\theta))(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x \\
=&\int_{\theta}^{+\infty} (f(x) - f(\theta))(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x \\
&+ \int_{-\infty}^{\theta} (f(x) - f(\theta))(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x \\
=&\int_{\theta}^{+\infty} \int_{\theta}^{x} f'(y)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}y\,\mathrm{d}x \\
&- \int_{-\infty}^{\theta} \int_{x}^{\theta} f'(y)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}y\,\mathrm{d}x \\
=&\int_{\theta}^{+\infty} \int_{y}^{+\infty} f'(y)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x\mathrm{d}y \\
&- \int_{-\infty}^{\theta} \int_{-\infty}^{y} f'(y)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x\,\mathrm{d}y \\
=&\int_{\theta}^{+\infty} f'(y) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)\,\mathrm{d}y - \int_{-\infty}^{\theta} f'(y) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)\,\mathrm{d}y \\
=&\int_{-\infty}^{+\infty} f'(y) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)\,\mathrm{d}y \\
=&\int_{-\infty}^{+\infty} f'(x) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x,
\end{aligned}$$

where we use

$$\begin{aligned}
&\int (x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}x \\
=&\frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}(x-\theta)^2\right)\,\mathrm{d}\left(\frac{1}{2}(x-\theta)^2\right) \\
=&\frac{-1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

and

$$\begin{aligned}
&\lim_{x\to+\infty}\frac{-1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\theta)^2\right)\\
=&\lim_{x\to-\infty}\frac{-1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\theta)^2\right)\\
=&0.
\end{aligned}$$

□

Theorem 4.16. *Under the setting of Lemma 4.6, we let*

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1-\frac{p-2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}}-\boldsymbol{\nu})+\boldsymbol{\nu}$$

and $p > 3$. Then we have

$$\mathbb{E}\left[\|\mathbf{m}(\bar{\mathbf{x}})-\boldsymbol{\mu}\|_2^2\right] < \mathbb{E}\left[\|\bar{\mathbf{x}}-\boldsymbol{\mu}\|_2^2\right].$$

Proof. We have

$$\begin{aligned}
\Delta R(\boldsymbol{\mu}) =&\mathbb{E}\left[\|\bar{\mathbf{x}}-\boldsymbol{\mu}\|_2^2-\|\mathbf{m}(\bar{\mathbf{x}})-\boldsymbol{\mu}\|_2^2\right]\\
=&\mathbb{E}\left[\|\bar{\mathbf{x}}-\boldsymbol{\mu}\|_2^2-\left\|\left(1-\frac{p-2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}}-\boldsymbol{\nu})+\boldsymbol{\nu}-\boldsymbol{\mu}\right\|_2^2\right]\\
=&\mathbb{E}\left[\sum_{i=1}^p(\bar{x}_i-\mu_i)^2-\sum_{i=1}^p\left(\left(1-\frac{p-2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right)(\bar{x}_i-\nu_i)+\nu_i-\mu_i\right)^2\right]\\
=&\mathbb{E}\left[\sum_{i=1}^p(\bar{x}_i-\mu_i)^2-\sum_{i=1}^p\left(\bar{x}_i-\mu_i-\frac{p-2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}(\bar{x}_i-\nu_i)\right)^2\right]\\
=&\mathbb{E}\left[\frac{2(p-2)}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\sum_{i=1}^p(\bar{x}_i-\nu_i)(\bar{x}_i-\mu_i)-\sum_{i=1}^p\frac{(p-2)^2(\bar{x}_i-\nu_i)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^4}\right]\\
=&\mathbb{E}\left[2(p-2)\sum_{i=1}^p\frac{\bar{x}_i-\nu_i}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\cdot(\bar{x}_i-\mu_i)-\frac{(p-2)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right].
\end{aligned}$$

Using Lemma 4.7 with $\theta=\mu_i$, we have

$$f(\bar{x}_i)=\frac{\bar{x}_i-\nu_i}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\qquad\text{and}\qquad f'(\bar{x}_i)=\frac{1}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}-\frac{2(\bar{x}_i-\nu_i)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^4}.$$

Hence, we obtain

$$\begin{aligned}
\Delta R(\boldsymbol{\mu}) =&\mathbb{E}\left[2(p-2)\sum_{i=1}^p\left(\frac{1}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}-\frac{2(\bar{x}_i-\nu_i)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^4}\right)-\frac{(p-2)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right]\\
=&\mathbb{E}\left[\frac{2p(p-2)}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}-\frac{4(p-2)}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}-\frac{(p-2)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right]\\
=&\mathbb{E}\left[\frac{(p-2)^2}{\|\bar{\mathbf{x}}-\boldsymbol{\nu}\|_2^2}\right]>0.
\end{aligned}$$

□

This document is organized by Luo Luo.
Please do not distribute.

Remark 4.12. *We consider the bias and variance decomposition*

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\boldsymbol{\mu}\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]+\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu}\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]\right\|_2^2+2\mathbb{E}[(\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})])^\top(\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu})]+\mathbb{E}\left\|\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu}\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]\right\|_2^2+2\mathbb{E}[(\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]])^\top(\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu})+\mathbb{E}\left\|\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu}\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]\right\|_2^2+\left\|\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]-\boldsymbol{\mu}\right\|_2^2.
\end{aligned}$$

Unbiased estimator may leads to larger variance. The requirement of unbiasedness can be regarded as the constrained problem

$$\min_{\mathbb{E}[\mathbf{m}(\bar{\mathbf{x}})]=\boldsymbol{\mu}}\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\boldsymbol{\mu}\right\|_2^2,$$

while without the requirement of unbiasedness can be regarded as

$$\min_{\mathbf{m}(\bar{\mathbf{x}})\in\mathbb{R}^p}\mathbb{E}\left\|\mathbf{m}(\bar{\mathbf{x}})-\boldsymbol{\mu}\right\|_2^2.$$

Lemma 4.8. *Suppose that $\mathbf{z}\sim\mathcal{N}(\boldsymbol{\zeta},\mathbf{I})$, then*

$$\mathbb{E}\left\|g^+\big(\|\mathbf{z}\|_2\big)\mathbf{z}-\boldsymbol{\zeta}\right\|_2^2\leq\mathbb{E}\left\|g\big(\|\mathbf{z}\|_2\big)\mathbf{z}-\boldsymbol{\zeta}\right\|_2^2,$$

where

$$g^+(u)=\begin{cases}g(u), & \textit{if } g(u)\geq 0\\ 0, & \textit{otherwise}\end{cases}$$

for any function $g(u)$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left\|g\big(\|\mathbf{z}\|_2\big)\mathbf{z}-\boldsymbol{\zeta}\right\|_2^2-\mathbb{E}\left\|g^+\big(\|\mathbf{z}\|_2\big)\mathbf{z}-\boldsymbol{\zeta}\right\|_2^2\\
=&\mathbb{E}\left[\big(g(\|\mathbf{z}\|_2)\big)^2\|\mathbf{z}\|_2^2\right]-\mathbb{E}\left[\big(g^+(\|\mathbf{z}\|_2)\big)^2\|\mathbf{z}\|^2\right]+2\mathbb{E}\left[\boldsymbol{\zeta}^\top\mathbf{z}\left(g^+\big(\|\mathbf{z}\|_2\big)-g\big(\|\mathbf{z}\|_2\big)\right)\right]\\
\geq&2\mathbb{E}\left[\boldsymbol{\zeta}^\top\mathbf{z}\left(g^+\big(\|\mathbf{z}\|_2\big)-g\big(\|\mathbf{z}\|_2\big)\right)\right],
\end{aligned}$$

where the last step is due to $|g^+(\cdot)|\leq|g(\cdot)|$.

Let $\mathbf{P}$ be the orthogonal matrix such that $\mathbf{P}\mathbf{P}^\top=\mathbf{I}$ and

$$\mathbf{P}=\left[\frac{\boldsymbol{\zeta}}{\|\boldsymbol{\zeta}\|_2},\times,\dots,\times\right],$$

which means

$$\mathbf{P}^\top\boldsymbol{\zeta}=[\|\boldsymbol{\zeta}\|_2,0,\dots,0]^\top.$$

Let $\mathbf{y}=\mathbf{P}^\top\mathbf{z}$, then we have $\boldsymbol{\zeta}^\top\mathbf{z}=\boldsymbol{\zeta}^\top\mathbf{P}\mathbf{y}=(\mathbf{P}^\top\boldsymbol{\zeta})^\top\mathbf{y}=\|\boldsymbol{\zeta}\|_2\,y_1$ and $\mathbf{y}\sim\mathcal{N}(\mathbf{P}^\top\boldsymbol{\zeta},\mathbf{I})$. Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\zeta}^\top\mathbf{z}\left(g^+\big(\|\mathbf{z}\|_2\big)-g\big(\|\mathbf{z}\|_2\big)\right)\right]\\
=&\mathbb{E}\left[\|\boldsymbol{\zeta}\|_2\,y_1\left(g^+\big(\|\mathbf{y}\|_2\big)-g\big(\|\mathbf{y}\|_2\big)\right)\right]\\
=&\|\boldsymbol{\zeta}\|_2\int_{-\infty}^{+\infty}y_1\left(g^+\big(\|\mathbf{y}\|_2\big)-g\big(\|\mathbf{y}\|_2\big)\right)\frac{1}{(2\pi)^{\frac{p}{2}}}\exp\left(-\frac{1}{2}\left(\sum_{i=1}^p y_i^2-2y_1\|\boldsymbol{\zeta}\|_2+\|\boldsymbol{\zeta}\|_2^2\right)\right)\mathrm{d}\mathbf{y}\\
=&\frac{\|\boldsymbol{\zeta}\|_2\exp\left(-\frac{1}{2}\|\boldsymbol{\zeta}\|_2^2\right)}{(2\pi)^{\frac{p}{2}}}\int_{-\infty}^{+\infty}y_1\left(g^+\big(\|\mathbf{y}\|_2\big)-g\big(\|\mathbf{y}\|_2\big)\right)\exp\left(-\frac{1}{2}\sum_{i=1}^p y_i^2\right)\exp(y_1\|\boldsymbol{\zeta}\|_2)\,\mathrm{d}\mathbf{y}
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
=&\frac{\|\boldsymbol{\zeta}\|_2 \exp\left(-\frac{1}{2}\|\boldsymbol{\zeta}\|_2^2\right)}{(2\pi)^{\frac{p}{2}}} \\
&\cdot \int_{-\infty}^{+\infty} \cdots \int_{0}^{+\infty} y_1 \left(g^{+}\left(\|\mathbf{y}\|_2\right) - g\left(\|\mathbf{y}\|_2\right)\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{p} y_i^2\right) \left(\exp(y_1\|\boldsymbol{\zeta}\|_2) - \exp(-y_1\|\boldsymbol{\zeta}\|_2)\right) \mathrm{d}y_1 \ldots \mathrm{d}y_p \\
\geq& 0
\end{aligned}
$$

where the last step use $\exp(z) - \exp(-z) \geq 0$ for all $z \geq 0$. □

Theorem 4.17. *Let*

$$
\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu} \qquad \textit{and} \qquad \tilde{\mathbf{m}}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)^{+}(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},
$$

where $\bar{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$. Then we have $\mathbb{E}\|\tilde{\mathbf{m}}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \leq \mathbb{E}\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2$.

Proof. Use Lemma 4.8 by taking $g(u) = 1 - (p-2)/u$, $\mathbf{z} = \bar{\mathbf{x}} - \boldsymbol{\nu}$ and $\boldsymbol{\zeta} = \boldsymbol{\mu} - \boldsymbol{\nu}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)^{+}(\bar{\mathbf{x}} - \boldsymbol{\nu}) - (\boldsymbol{\mu} - \boldsymbol{\nu})\right] \\
\leq &\mathbb{E}\left[\left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) - (\boldsymbol{\mu} - \boldsymbol{\nu})\right],
\end{aligned}
$$

that is $\mathbb{E}\|\tilde{\mathbf{m}}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \leq \mathbb{E}\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2$. □

5 Hypothesis Testing

Directly apply hypothesis testing for p-dimensional case with significance level α:

1. For $\alpha = 0.05$ and $p = 1$, we have confidence level $1 - \alpha = 0.95$.
2. For $\alpha = 0.05$ and $p = 100$, we have confidence level $(1-\alpha)^p \approx 0.006$.
3. For $\alpha \approx 0.0005$ and $p = 100$, we have confidence level $(1-\alpha)^p > 0.95$.

We desire $\|\bar{\mathbf{x}} - \boldsymbol{\mu}_0\|$ (distance w.r.t some norm) being small, rather than each entry of $\bar{\mathbf{x}} - \boldsymbol{\mu}_0$ is small.

Theorem 5.1. *The density of $y \sim \chi^2(n)$ is*

$$
f(y;n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} \exp\left(-\dfrac{y}{2}\right), & y > 0, \\ 0, & \textit{otherwise}, \end{cases}
$$

where

$$
\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)\,\mathrm{d}t.
$$

Proof. We first provide the following results:

This document is organized by Luo Luo.
Please do not distribute.

1. We have $\Gamma\left(1/2\right) = \sqrt{\pi}$, because

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) =& \int_0^\infty t^{-1/2}\exp(-t)\,\mathrm{d}t \\ =& \int_0^\infty \left(\frac{1}{2}x^2\right)^{-1/2} \exp\left(-\frac{1}{2}x^2\right)\,\mathrm{d}\left(\frac{1}{2}x^2\right) \\ =& \int_0^\infty \frac{\sqrt{2}}{x}\exp\left(-\frac{1}{2}x^2\right)x\,\mathrm{d}x \\ =& \sqrt{2}\int_0^\infty \exp\left(-\frac{1}{2}x^2\right)\,\mathrm{d}x \\ =& 2\sqrt{\pi}\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\,\mathrm{d}x \\ =& \sqrt{\pi}.\end{aligned}$$

2. For $y_1 = x^2$ with $x \sim \mathcal{N}(0,1)$, we can show that the density function of y_1 is

$$\frac{1}{\sqrt{2\pi y_1}}\exp\left(-\frac{1}{2}y_1\right).$$

We define the non-negative random variable $\hat{x}$ whose density function is

$$\begin{cases}\dfrac{2}{\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\hat{x}^2\right), & \hat{x} \geq 0, \\ 0 & \hat{x} < 0.\end{cases} \tag{23}$$

Then the transform $\hat{x} = \sqrt{y_1}$ is one to one and the density of y_1 is

$$\frac{2}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_1\right)\frac{\mathrm{d}\sqrt{y_1}}{\mathrm{d}y_1} = \frac{1}{\sqrt{2\pi y_1}}\exp\left(-\frac{1}{2}y_1\right).$$

3. For beta function

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t,$$

we have

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Consider that

$$\begin{aligned}&\Gamma(\alpha)\Gamma(\beta) \\ =& \int_0^\infty x^{\alpha-1}\exp(-x)\,\mathrm{d}x\int_0^\infty y^{\beta-1}\exp(-y)\,\mathrm{d}y \\ =& \int_0^\infty\int_0^\infty x^{\alpha-1}y^{\beta-1}\exp(-(x+y))\,\mathrm{d}y\,\mathrm{d}x.\end{aligned}$$

Using the substitution $x = uv$ and $y = u(1-v)$, then the Jacobian matrix of the transformation is

$$\mathbf{J} = \begin{bmatrix}\frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}\end{bmatrix} = \begin{bmatrix} v & u \\ 1-v & -u\end{bmatrix}$$

This document is organized by Luo Luo.
Please do not distribute.

and $\det(\mathbf{J}) = -u$. Since $u = x + y$ and $v = x/(x + y)$, we have that the limits of integration for u are 0 to ∞ and the limits of integration for v are 0 to 1. Thus

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) =& \int_0^\infty \int_0^\infty x^{\alpha-1} y^{\beta-1} \exp(-(x+y)) \,\mathrm{d}y \,\mathrm{d}x \\
=& \int_0^1 \int_0^\infty (uv)^{\alpha-1} (u(1-v))^{\beta-1} \exp(-(uv + u(1-v)))|-u| \,\mathrm{d}u \,\mathrm{d}v \\
=& \int_0^1 \int_0^\infty u^{\alpha+\beta-1} v^{\alpha-1} (1-v)^{\beta-1} \exp(-u) \,\mathrm{d}u \,\mathrm{d}v \\
=& \int_0^1 v^{\alpha-1}(1-v)^{\beta-1} \,\mathrm{d}v \int_0^\infty u^{\alpha+\beta-1} \exp(-u) \,\mathrm{d}u \\
=& B(\alpha,\beta)\Gamma(\alpha+\beta).
\end{aligned}$$

4. If

$$F(z) = \int_{a(z)}^{b(z)} f(y,z) \,\mathrm{d}y,$$

then

$$F'(z) = \int_{a(z)}^{b(z)} \frac{\partial f(y,z)}{\partial z} \,\mathrm{d}x + f(b(z),z)b'(z) - f(a(z),z)a'(z).$$

We prove the density of Chi-square distribution by induction. For $n = 1$ and $y > 0$, we have

$$f(y;1) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{1}{2}y\right) = \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} y^{\frac{1}{2}-1} \exp\left(-\frac{y}{2}\right),$$

which corresponds to equation (23) by using $\Gamma(1/2) = \sqrt{\pi}$.
Suppose the statement holds for $n - 1$, that is

$$f(y; n-1) = \begin{cases} \dfrac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} y^{\frac{n-1}{2}-1} \exp\left(-\dfrac{y}{2}\right), & y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We consider $y_n = y_{n-1} + x_n^2$ such that $y_{n-1} \sim \chi^2(n-1)$ and $x_n \sim \mathcal{N}(0,1)$ are independent. Let F_1 be the corresponding cdf of $f(y;1)$. Then the cfd of y_n is

$$\begin{aligned}
&\Pr(y_n \le z) \\
=& \int_0^z \int_0^{z-y} f_{n-1}(y) f_1(x) \,\mathrm{d}x \,\mathrm{d}y \\
=& \int_0^z (F_1(z-y) - F_1(0)) f_{n-1}(y) \,\mathrm{d}x \,\mathrm{d}y \\
=& \int_0^z F_1(z-y) f_{n-1}(y) \,\mathrm{d}y
\end{aligned}$$

and the pdf of y_n is (let $y = tz$)

$$\begin{aligned}
&\int_0^z \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} (z-y)^{\frac{1}{2}-1} \exp\left(-\frac{z-y}{2}\right) \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right) \,\mathrm{d}y \\
=& \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \int_0^z (z-y)^{\frac{1}{2}-1} y^{\frac{n-1}{2}-1} \exp\left(-\frac{z}{2}\right) \,\mathrm{d}y
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
&=\frac{\exp\left(-\frac{z}{2}\right) z^{\frac{n-1}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}\int_0^1 (1-t)^{\frac{1}{2}-1}t^{\frac{n-1}{2}-1}\,\mathrm{d}t \\
&=\frac{\exp\left(-\frac{z}{2}\right) z^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} B\left(\frac{n-1}{2},\frac{1}{2}\right) \\
&=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1}\exp\left(-\frac{z}{2}\right).
\end{aligned}
$$

□

Theorem 5.2. *For $y \sim \chi^2_{n,\lambda}$, we have $\mathbb{E}[y] = n+\lambda$ and $\mathrm{Var}[y] = 2n+4\lambda$.*

Proof. We can write

$$
y = \sum_{i=1}^n x_i^2,
$$

where $x_i \sim \mathcal{N}(\mu_i, 1)$ and $x_1,\dots,x_n$ are independent. Then, we have

$$
\begin{aligned}
\mathbb{E}[y] =& \mathbb{E}\left[\sum_{i=1}^n x_i^2\right] = \sum_{i=1}^n \mathbb{E}\left[(x_i-\mu_i)^2 + 2\mu_i x_i - \mu_i^2\right] \\
=& \sum_{i=1}^n \left(\mathrm{Var}\left[x_i^2\right] + 2\mu_i^2 - \mu_i^2\right) = n+\lambda.
\end{aligned}
$$

Consider that $\phi(t) = \exp\big(\mathrm{i}\,\mu_i t - t^2/2\big)$, we have

$$
\mathbb{E}[x_i^2] = \frac{1}{\mathrm{i}^2}\frac{\mathrm{d}^2\phi(t)}{\mathrm{d}t^2}\bigg|_{t=0} = (-1)\cdot(t^2 - 2\mathrm{i}\,\mu_i t - \mu_i^2 - 1)\exp\left(\mathrm{i}\,\mu_i t - \frac{1}{2}t^2\right)\bigg|_{t=0} = \mu_i^2 + 1
$$

and

$$
\mathbb{E}[x_i^4] = \frac{1}{\mathrm{i}^4}\frac{\mathrm{d}^4\phi(t)}{\mathrm{d}t^4}\bigg|_{t=0} = (t^4 - 4\mathrm{i}\,\mu_i t^3 - (6+6\mu_i^2)t^2 + \mu_i^4 + 6\mu_i^2 + 3)\exp\left(\mathrm{i}\,\mu_i t - \frac{1}{2}t^2\right)\bigg|_{t=0} = \mu_i^4 + 6\mu_i^2 + 3.
$$

This implies

$$
\begin{aligned}
\mathrm{Var}[y] =& \mathrm{Var}\left[\sum_{i=1}^n x_i^2\right] = \sum_{i=1}^n \mathrm{Var}\left[x_i^2\right] = \sum_{i=1}^n \left(\mathbb{E}\left[x_i^4\right] - \left(\mathbb{E}[x_i^2]\right)^2\right) \\
=& \sum_{i=1}^n \mathbb{E}\left[\mu_i^4 + 6\mu_i^2 + 3 - (\mu_i^2+1)^2\right] \\
=& \sum_{i=1}^n \mathbb{E}\left[4\mu_i^2 + 2\right] = 4\lambda + 2n.
\end{aligned}
$$

□

Remark 5.1. *If the n-th moment of random variable x, denoted by $\mathbb{E}[x^n]$, exists and is finite, then its characteristic function is n times continuously differentiable and*

$$
\mathbb{E}[x^n] = \frac{1}{\mathrm{i}^n}\frac{\mathrm{d}^n\phi(t)}{\mathrm{d}t^n}\bigg|_{t=0},
$$

which is because of

$$
\frac{\mathrm{d}^n\phi(t)}{\mathrm{d}t^n} = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}\left[\exp(\mathrm{i}\,tx)\right] = \mathbb{E}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}\exp(\mathrm{i}\,tx)\right] = \mathbb{E}\left[(\mathrm{i}\,x)^n\exp(\mathrm{i}\,tx)\right] = \mathrm{i}^n\,\mathbb{E}\left[x^n\exp(\mathrm{i}\,tx)\right].
$$

This document is organized by Luo Luo.
Please do not distribute.

Theorem 5.3. *If the n-component random vector* $\mathbf{y}$ *is distributed according to* $\mathcal{N}_n(\boldsymbol{\nu}, \mathbf{T})$ *with* $\mathbf{T} \succ \mathbf{0}$*, then*

$$\mathbf{y}^\top \mathbf{T}^{-1}\mathbf{y} \sim \chi^2_{n,\lambda},$$

where

$$\lambda = \boldsymbol{\nu}^\top \mathbf{T}^{-1}\boldsymbol{\nu}.$$

If $\boldsymbol{\nu} = \mathbf{0}$*, the distribution is the central* χ^2_n*-distribution.*

Proof. Let SVD of $\mathbf{T}$ be

$$\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

Define

$$\mathbf{C} = \mathbf{D}^{-1/2}\mathbf{U}^\top \qquad \text{such that} \qquad \mathbf{T}^{-1} = \mathbf{C}^\top\mathbf{C}.$$

Let

$$\mathbf{z} = \mathbf{C}\mathbf{y},$$

then we have

$$\begin{aligned}
&\mathbf{y}^\top \mathbf{T}^{-1}\mathbf{y} \\
=&\mathbf{z}^\top(\mathbf{C}^{-1})^\top \mathbf{T}^{-1}\mathbf{C}^{-1}\mathbf{z} \\
=&\mathbf{z}^\top\mathbf{z} = \sum_{i=1}^n z_i^2.
\end{aligned}$$

We also have

$$\mathbf{C}\boldsymbol{\nu} = \mathbf{D}^{-1/2}\mathbf{U}^\top\boldsymbol{\nu} \qquad \text{and} \qquad \mathbf{C}\mathbf{T}\mathbf{C}^\top = (\mathbf{D}^{-1/2}\mathbf{U}^\top)(\mathbf{U}\mathbf{D}\mathbf{U}^\top)(\mathbf{U}\mathbf{D}^{-1/2}) = \mathbf{I},$$

which means

$$\mathbf{z} \sim \mathcal{N}(\mathbf{D}^{-1/2}\mathbf{U}^\top\boldsymbol{\nu}, \mathbf{I}).$$

Therefore, the variance $z_1, \dots, z_n$ are independent normal with unit variance and means $\mu_1, \dots, \mu_n$ such that

$$\sum_{i=1}^n \mu_i^2 = \left(\boldsymbol{\nu}^\top\mathbf{U}\mathbf{D}^{-1/2}\right)\left(\mathbf{D}^{-1/2}\mathbf{U}^\top\boldsymbol{\nu}\right) = \boldsymbol{\nu}^\top\mathbf{T}^{-1}\boldsymbol{\nu},$$

which means $\mathbf{y}^\top\mathbf{T}^{-1}\mathbf{y} = \mathbf{z}^\top\mathbf{z} \sim \chi^2_{n,\lambda}$, $\square$

Theorem 5.4. *The probability density function (pdf) for the noncentral Chi-squared distribution is*

$$f(v; p, \lambda) = \begin{cases} \displaystyle\sum_{k=0}^{\infty} \frac{(\lambda/2)^k \exp\left(-(\lambda/2)\right)}{k!} \cdot \frac{1}{2^{\frac{p}{2}+k}\Gamma\left(\frac{p}{2}+k\right)} v^{\frac{p}{2}+k-1} \exp\left(-\frac{v}{2}\right) & v > 0, \\ 0, & v \le 0. \end{cases}$$

where $B(\alpha, \beta) = \displaystyle\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$.

To proof Theorem 5.4, we first present some lemmas.

Lemma 5.1. *For* $a, b \in \mathbb{C}$ *such that* $\mathrm{Re}(a) > 0$*, we have*

$$\int_{-\infty}^{+\infty} \exp(-a(x-b)^2)\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}.$$

Remark 5.2. *Note that the terms*

$$\frac{(\lambda/2)^k \exp\left(-(\lambda/2)\right)}{k!} \qquad \textit{and} \qquad \frac{1}{2^{\frac{p+2k}{2}}\Gamma\left(\frac{p}{2}+k\right)} y^{\frac{p}{2}+k-1}\exp\left(-\frac{v}{2}\right)$$

are the probabilistic mass function of Poisson distribution with parameter $\lambda/2$ *and the probabilistic density function of central* χ^2*-distribution with degree of freedom* $p + 2k$.

This document is organized by Luo Luo.
Please do not distribute.

Hypothesis Testing for the Mean: If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with known $\mathbf{\Sigma}$, then we have

$$\bar{\mathbf{x}} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{N}\mathbf{\Sigma}\right), \qquad \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}) \qquad \text{and} \qquad \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \mathcal{N}_p(\boldsymbol{\mu} - \boldsymbol{\mu}_0, \mathbf{\Sigma}).$$

1. Using Theorem 5.3 with $\mathbf{y} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ and $\mathbf{T} = \mathbf{\Sigma}$ achieves $N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \chi_p^2$.
2. Using Theorem 5.3 with $\mathbf{y} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ and $\mathbf{T} = \mathbf{\Sigma}$ achieves $N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \chi_{p,\lambda}^2$ with $\lambda = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$.

Let $\chi_p^2(\alpha)$ be the number such that

$$\Pr\left\{N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) > \chi_p^2(\alpha)\right\} = \alpha.$$

It says that the probability is α that the weighted distance $\sqrt{(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})}$ is greater than $\sqrt{\chi_p^2(\alpha)/N}$.

Remark 5.3. *For $\mathbf{\Sigma} = \mathbf{I}$, it corresponds to Euclidean distance. We can rotation the coordinate to achieve diagonal covariance matrix, then it reduce the coordinate with large variance.*

To test the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is a specified vector, we use as our rejection region

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \chi_p^2(\alpha),$$

where $N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \chi_p^2$ if we suppose $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. If above inequality is satisfied, we reject the null hypothesis.

Without assuming $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, we regard

$$\mathcal{E} = \{\mathbf{x} \in \mathbb{R}^p : N(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\} \quad \text{and} \quad \mathcal{E}_0 = \{\mathbf{x} \in \mathbb{R}^p : N(\mathbf{x} - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) \leq \chi_p^2(\alpha)\}$$

as two ellipsoids in space of $\mathbf{x}$, centered at $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$ respectively. For small α, the sample mean $\bar{\mathbf{x}}$ lies in $\mathcal{E}$ with high probability $1 - \alpha$.

1. It can be seen intuitively that the probability is greater than α of rejecting the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ if $\boldsymbol{\mu}$ is very different from $\boldsymbol{\mu}_0$. Note that when $\boldsymbol{\mu}$ is far away from $\boldsymbol{\mu}_0$, the density of $\bar{\mathbf{x}}$ will be concentrated at a point near the edge or outside of $\mathcal{E}_0$.
2. If $\boldsymbol{\mu}_0$ is close $\boldsymbol{\mu}$, then ellipsoid $\mathcal{E}$ is also close to $\mathcal{E}_0$.

Remark 5.4. *The power function of a hypothesis test is the probability of rejecting the null space. In this case, it is achieved by noncentral χ^2-distribution.*

This document is organized by Luo Luo. Please do not distribute.

Two-Sample Problems: Suppose there are two samples $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{N_1}^{(1)}$ from $\mathcal{N}\big(\boldsymbol{\mu}^{(1)}, \mathbf{\Sigma}\big)$ and $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{N_2}^{(2)}$ from $\mathcal{N}\big(\boldsymbol{\mu}^{(2)}, \mathbf{\Sigma}\big)$, where $\mathbf{\Sigma}$ is known. We test the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$. Note that the sample means

$$\bar{\mathbf{x}}^{(1)} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} \mathbf{x}_\alpha^{(1)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)}, \frac{1}{N_1}\mathbf{\Sigma}\right) \qquad \text{and} \qquad \bar{\mathbf{x}}^{(2)} = \frac{1}{N_2}\sum_{\alpha=1}^{N_2} \mathbf{x}_\alpha^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(2)}, \frac{1}{N_2}\mathbf{\Sigma}\right).$$

are independent. Then we have

$$\begin{bmatrix}\bar{\mathbf{x}}^{(1)} \\ \bar{\mathbf{x}}^{(2)}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)}\end{bmatrix}, \begin{bmatrix}\frac{1}{N_1}\mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \frac{1}{N_2}\mathbf{\Sigma}\end{bmatrix}\right), \qquad \mathbf{y} = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} = \begin{bmatrix}\mathbf{I} & -\mathbf{I}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{x}}^{(1)} \\ \bar{\mathbf{x}}^{(2)}\end{bmatrix},$$

and

$$\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\nu}, \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\mathbf{\Sigma}\right) \qquad \text{where} \qquad \boldsymbol{\nu} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}.$$

Thus, we reduce the problem to testing mean $\boldsymbol{\nu} = \mathbf{0}$ for $\mathbf{y}$. We achieve the confidence region

$$\left\{\boldsymbol{\nu}^* \in \mathbb{R}^p : \frac{N_1 N_2}{N_1 + N_2}(\mathbf{y} - \boldsymbol{\nu}^*)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\nu}^*) \leq \chi_p^2(\alpha)\right\}.$$

and critical region

$$\frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) > \chi_p^2(\alpha).$$

Future Course: What about both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is unknown? Let $x_1, \dots, x_N$ be independently and identically drawn from the distribution $\mathcal{N}(\mu, \sigma^2)$, then the random variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$

has student t-distribution with $N - 1$ degrees of freedom, where

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^N x_\alpha \qquad \text{and} \qquad s^2 = \frac{1}{N-1}\sum_{\alpha=1}^N (x_\alpha - \bar{x})^2.$$

The multivariate analog of t^2 is

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

where

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^N \mathbf{x}_\alpha \qquad \text{and} \qquad \mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^N (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

If $\mathbf{x}_1, \dots, \mathbf{x}_N$ be independently and identically drawn from the distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The distribution of

$$\frac{T^2}{N-1} \cdot \frac{N-p}{p}$$

is noncentral F with p and $N - p$ degrees of freedom and noncentrality parameter $N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$.

Remark 5.5. *If y_1 is a noncentral χ^2-random variable with noncentrality parameter λ and d_1 degrees of freedom, and y_2 is a central χ^2-random variable with d_2 degrees of freedom that is independent of y_1, then*

$$x = \frac{y_1/d_1}{y_2/d_2}$$

is a noncentral F-distributed random variable with degrees of freedom d_1, d_2 and noncentral parameter λ.

This document is organized by Luo Luo.
Please do not distribute.

6 Sample Correlation Coefficients

Lemma 6.1. *Let $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{Q} \in \mathbb{R}^{p \times p}$ be a rank-r projection matrix, then*

$$\mathbf{x}^\top \mathbf{Q} \mathbf{x} \sim \chi_r^2.$$

Proof. The projection matrix $\mathbf{Q}$ has SVD as follows

$$\mathbf{Q} = \mathbf{U}\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{U}^\top, \qquad \text{where } \mathbf{U} \in \mathbb{R}^{p \times p} \text{ is orthogonal.}$$

Let $\mathbf{y} = \mathbf{U}^\top \mathbf{x}$, then we have $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and

$$\begin{aligned}\mathbf{x}^\top \mathbf{Q} \mathbf{x} =& \mathbf{x}^\top \mathbf{U} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^\top \mathbf{x} \\ =& \mathbf{y}^\top \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{y} = \sum_{\alpha=1}^{r} y_i \sim \chi_r^2.\end{aligned}$$

□

Lemma 6.2. *If $\mathbf{y}_1, \dots, \mathbf{y}_N$ are independently distributed, if*

$$\mathbf{y}_\alpha = \begin{bmatrix} \mathbf{y}_\alpha^{(1)} \\ \mathbf{y}_\alpha^{(2)} \end{bmatrix}$$

has the density $f(\mathbf{y}_\alpha)$ and if the conditional density of $\mathbf{y}_\alpha^{(2)}$ given $\mathbf{y}_\alpha^{(1)}$ is $f\big(\mathbf{y}_\alpha^{(2)} \mid \mathbf{y}_\alpha^{(1)}\big)$ for $\alpha = 1, \dots, n$. Then in the conditional distribution of $\mathbf{y}_1^{(2)}, \dots, \mathbf{y}_N^{(2)}$ given $\mathbf{y}_1^{(1)}, \dots, \mathbf{y}_N^{(1)}$, the random vectors $\mathbf{y}_1^{(1)}, \dots, \mathbf{y}_N^{(1)}$ are independent and the density of $\mathbf{y}_\alpha^{(2)}$ is $\prod_{\alpha=1}^N f\big(\mathbf{y}_\alpha^{(2)} \mid \mathbf{y}_\alpha^{(1)}\big)$.

Proof. The marginal density of $\mathbf{y}_1^{(1)}, \dots, \mathbf{y}_N^{(1)}$ is

$$\prod_{\alpha=1}^{N} f_1\big(\mathbf{y}_\alpha^{(1)}\big)$$

where $f_1\big(\mathbf{y}_\alpha^{(1)}\big)$ is the marginal density of $\mathbf{y}_\alpha^{(1)}$. The conditional density of $\mathbf{y}_1^{(2)}, \dots, \mathbf{y}_N^{(2)}$ given $\mathbf{y}_1^{(1)}, \dots, \mathbf{y}_N^{(1)}$ is

$$\frac{\prod_{\alpha=1}^N f\big(\mathbf{y}_\alpha\big)}{\prod_{\alpha=1}^N f_1\big(\mathbf{y}_\alpha^{(1)}\big)} = \prod_{\alpha=1}^{N} \frac{f\big(\mathbf{y}_\alpha^{(1)}, \mathbf{y}_\alpha^{(2)}\big)}{f_1\big(\mathbf{y}_\alpha^{(1)}\big)} = \prod_{\alpha=1}^{N} f\big(\mathbf{y}_\alpha^{(2)} \mid \mathbf{y}_\alpha^{(1)}\big).$$

□

This document is organized by Luo Luo. Please do not redistribute.

The distribution of sample correlation: We consider two-dimensional case. Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be observation from

$$\mathcal{N}_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right).$$

We denote

$$\mathbf{x}_\alpha = \begin{bmatrix} x_{1\alpha} \\ x_{2\alpha} \end{bmatrix}, \quad \bar{x}_i = \frac{1}{N}\sum_{\alpha=1}^{N} x_{i\alpha}, \quad \mathbf{A} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \quad \text{and} \quad a_{ij} = \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j).$$

We have shown that $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top = \begin{bmatrix} \sum_{\alpha=1}^{n} z_{1\alpha}^2 & \sum_{\alpha=1}^{n} z_{1\alpha} z_{2\alpha} \\ \sum_{\alpha=1}^{n} z_{1\alpha} z_{2\alpha} & \sum_{\alpha=1}^{n} z_{2\alpha}^2 \end{bmatrix} \quad \text{where } \mathbf{z}_\alpha = \begin{bmatrix} z_{1\alpha} \\ z_{2\alpha} \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right).$$

$n = N - 1$ and $\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent. We also denote the $2 \times n$ random matrix

$$[\mathbf{z}_1, \dots, \mathbf{z}_n] = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \end{bmatrix}, \quad \text{where } \mathbf{v}_1 = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1n} \end{bmatrix} \in \mathbb{R}^n \text{ and } \mathbf{v}_2 = \begin{bmatrix} z_{21} \\ \vdots \\ z_{2n} \end{bmatrix} \in \mathbb{R}^n.$$

We shall consider the sample correlation coefficient

$$r = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} = \frac{\mathbf{v}_1^\top \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2}$$

We also denote

$$a_{11.2} = a_{11} - \frac{a_{12}^2}{a_{22}} = \mathbf{v}_1^\top \mathbf{v}_1 - \frac{(\mathbf{v}_1^\top \mathbf{v}_2)^2}{\mathbf{v}_2^\top \mathbf{v}_2}.$$

Lemma 6.3. *Based on above notations, we have*

(a) $\dfrac{a_{11}}{\sigma_{11}} \sim \chi_n^2$ *and* $\dfrac{a_{22}}{\sigma_{22}} \sim \chi_n^2$*;*

(b) $a_{12} \mid a_{22} \sim \mathcal{N}\left(\sigma_{22}^{-1}\sigma_{12}a_{22}, \sigma_{11.2}a_{22}\right)$*;*

(c) $\dfrac{a_{11.2}}{\sigma_{11.2}} \sim \chi_{n-1}^2$ *is independent on* a_{12} *and* a_{22}*.*

Proof. **Part (a):** It is easy to show that

$$\frac{a_{11}}{\sigma_{11}} = \sum_{\alpha=1}^{n}\left(\frac{z_{1\alpha}}{\sqrt{\sigma_{11}}}\right)^2 \sim \chi_n^2 \qquad \text{and} \qquad \frac{a_{22}}{\sigma_{22}} = \sum_{\alpha=1}^{n}\left(\frac{z_{2\alpha}}{\sqrt{\sigma_{22}}}\right)^2 \sim \chi_n^2.$$

Part (b): Recall that for

$$\mathbf{x} = \begin{bmatrix}\mathbf{x}^{(1)} \\ \mathbf{x}^{(2)}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right),$$

the conditional density $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

For any $\mathbf{z}_\alpha$, we have

$$z_{1\alpha} \mid z_{2\alpha} \sim \mathcal{N}(\sigma_{22}^{-1}\sigma_{12}z_{2\alpha}, \sigma_{11.2}),$$

where $\sigma_{11.2} = \sigma_{11} - \sigma_{12}^2\sigma_{22}^{-1}$. Since variables $z_{11}, \dots, z_{1n}$ are independent, Lemma 6.2 means

$$\mathbf{v}_1 \mid \mathbf{v}_2 \sim \mathcal{N}_n\left(\sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2, \sigma_{11.2}\mathbf{I}_n\right).$$

Consider that $a_{22} = \mathbf{v}_2^\top \mathbf{v}_2$, we have

$$a_{12} = \mathbf{v}_2^\top \mathbf{v}_1 \mid \mathbf{v}_2 \sim \mathcal{N}_n\left(\sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2^\top\mathbf{v}_2, \sigma_{11.2}\mathbf{v}_2^\top\mathbf{v}_2\right) \quad \Longrightarrow \quad a_{12} \mid a_{22} \sim \mathcal{N}\left(\sigma_{22}^{-1}\sigma_{12}a_{22}, \sigma_{11.2}a_{22}\right)$$

Part (c): The distribution of $\mathbf{v}_1 \mid \mathbf{v}_2$ means

$$\mathbf{w} = \mathbf{v}_1 - \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2 \mid \mathbf{v}_2 \sim \mathcal{N}_n\left(\mathbf{0}, \sigma_{11.2}\mathbf{I}_n\right) \qquad \Longrightarrow \qquad \frac{\mathbf{w}}{\sqrt{\sigma_{11.2}}} \,\bigg|\, \mathbf{v}_2 \sim \mathcal{N}_n\left(\mathbf{0}, \mathbf{I}_n\right).$$

We can write

$$\begin{aligned} a_{11.2} =& a_{11} - \frac{a_{12}^2}{a_{22}} = \mathbf{v}_1^\top\mathbf{v}_1 - \frac{(\mathbf{v}_1^\top\mathbf{v}_2)^2}{\mathbf{v}_2^\top\mathbf{v}_2} \\ =& \mathbf{v}_1^\top\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{v}_1 = \mathbf{w}^\top\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{w}, \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

where the last step is because of

$$\begin{aligned}\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{w} =& \left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\left(\mathbf{v}_1 - \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2\right) \\ =& \mathbf{v}_1 - \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2 - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\left(\mathbf{v}_1 - \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2\right) \\ =& \mathbf{v}_1 - \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2 - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\mathbf{v}_1 + \sigma_{22}^{-1}\sigma_{12}\frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\mathbf{v}_2 \\ =& \left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{v}_1\end{aligned}$$

and

$$\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right) = \mathbf{I} - \frac{2\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2} + \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\cdot\frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2} = \mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}.$$

Applying Lemma 6.1, we have

$$\frac{a_{11.2}}{\sigma_{11.2}} = \frac{\mathbf{w}}{\sqrt{\sigma_{11.2}}}\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\frac{\mathbf{w}^\top}{\sqrt{\sigma_{11.2}}} \qquad \Longrightarrow \qquad \frac{a_{11.2}}{\sigma_{11.2}}\,\Big|\,\mathbf{v}_2 \sim \chi^2_{n-1}.$$

Note that the conditional distribution of $a_{11.2}$ given that $\mathbf{v}_2$ does not depend on $\mathbf{v}_2$. This implies the variable $a_{11.2}$ is independent on $\mathbf{v}_2$ (it also independent on a_{22}) and

$$\frac{a_{11.2}}{\sigma_{11.2}} \sim \chi^2_{n-1}.$$

Finally, we show that $a_{11.2}$ is independent on a_{12}. Note that

$$\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{v}_2 = \mathbf{0},$$

then variables

$$\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{w} \qquad \text{and} \qquad \mathbf{v}_2^\top\mathbf{w}$$

are independent for given $\mathbf{v}_2$. Thus, the random variables

$$a_{12} = \mathbf{v}_2^\top\mathbf{v}_1 = \mathbf{v}_2^\top(\mathbf{w} + \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2) = \mathbf{v}_2^\top\mathbf{w} + \sigma_{22}^{-1}\sigma_{12}\mathbf{v}_2^\top\mathbf{v}_2$$

and

$$a_{11.2} = \mathbf{w}^\top\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{w} = \left(\mathbf{w}\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\right)^\top\left(\mathbf{I}_n - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2}\right)\mathbf{w}$$

are independent for given $\mathbf{v}_2$, i.e,

$$f(a_{11.2}, a_{12}\,|\,\mathbf{v}_2) = f(a_{11.2}\,|\,\mathbf{v}_2)f(a_{12}\,|\,\mathbf{v}_2).$$

The independence on $a_{11.2}$ and a_{12} can be shown as follows

$$\begin{aligned}f(a_{11.2}, a_{12}) =& \int f(a_{11.2}, a_{12}, \mathbf{v}_2)\,\mathrm{d}\mathbf{v}_2 \\ =& \int f(a_{11.2}, a_{12}\,|\,\mathbf{v}_2)f(\mathbf{v}_2)\,\mathrm{d}\mathbf{v}_2 \\ =& \int f(a_{11.2}\,|\,\mathbf{v}_2)f(a_{12}\,|\,\mathbf{v}_2)f(\mathbf{v}_2)\,\mathrm{d}\mathbf{v}_2\end{aligned}$$

This document is organized by Luo Luo
Please do not distribute

$$
\begin{aligned}
&= \int f(a_{11.2}) f(a_{12} \,|\, \mathbf{v}_2) f(\mathbf{v}_2) \, \mathrm{d}\mathbf{v}_2 \\
&= f(a_{11.2}) \int f(a_{12} \,|\, \mathbf{v}_2) f(\mathbf{v}_2) \, \mathrm{d}\mathbf{v}_2 \\
&= f(a_{11.2}) \int f(a_{12}, \mathbf{v}_2) \, \mathrm{d}\mathbf{v}_2 \\
&= f(a_{11.2}) f(a_{12}).
\end{aligned}
$$

□

Consider that $a_{12} \mid a_{22} \sim \mathcal{N}\left(\sigma_{22}^{-1}\sigma_{12}a_{22}, \sigma_{11.2}a_{22}\right)$, we have

$$
x = \frac{a_{12} - \sigma_{22}^{-1}\sigma_{12}a_{22}}{\sqrt{\sigma_{11.2}a_{22}}} \,\Big|\, a_{22} \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad x = \frac{a_{12} - \sigma_{22}^{-1}\sigma_{12}a_{22}}{\sqrt{\sigma_{11.2}a_{22}}} \sim \mathcal{N}(0,1).
$$

We also have $y = a_{11.2}/\sigma_{11.2} \sim \chi^2_{n-1}$, which is independent on a_{12} and a_{22}. This means x is independent on y. Substituting the expression of x and y into above one, we have

$$
z = \frac{x}{\sqrt{y/(n-1)}} = \frac{\dfrac{a_{12} - \sigma_{22}^{-1}\sigma_{12}a_{22}}{\sqrt{\sigma_{11.2}a_{22}}}}{\sqrt{\dfrac{a_{11.2}}{(n-1)\sigma_{11.2}}}}
$$

Recall that $r = a_{12}/\sqrt{a_{11}a_{22}}$. If $\sigma_{12} = 0$, we have

$$
z = \frac{x}{\sqrt{y/(n-1)}} = \frac{\dfrac{a_{12}}{\sqrt{a_{22}}}}{\sqrt{\dfrac{a_{11} - a_{12}^2/a_{22}}{n-1}}} = \frac{\sqrt{n-1}\,\dfrac{a_{12}}{\sqrt{a_{11}a_{22}}}}{\sqrt{1 - \dfrac{a_{12}^2}{a_{11}a_{22}}}} = \frac{\sqrt{n-1}\,r}{\sqrt{1-r^2}} \sim t_{n-1}
$$

and

$$
z^2 = \frac{(n-1)r^2}{1-r^2}.
$$

Recall that z has density

$$
\frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})} \left(1 + \frac{z^2}{n-1}\right)^{-\frac{n}{2}}.
$$

and

$$
\frac{\mathrm{d}z}{\mathrm{d}r} = \frac{\sqrt{n-1}}{(1-r^2)^{3/2}},
$$

then the density of r is

$$
\begin{aligned}
&\frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})} \left(1 + \frac{r^2}{1-r^2}\right)^{-\frac{n}{2}} \cdot \frac{\sqrt{n-1}}{(1-r^2)^{3/2}} \\
=&\frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{n-1}{2})} \cdot \left(1-r^2\right)^{\frac{n-3}{2}} = \frac{\Gamma(\frac{N-1}{2})}{\sqrt{\pi}\Gamma(\frac{N-2}{2})} \cdot \left(1-r^2\right)^{\frac{N-4}{2}}.
\end{aligned}
$$

This document is organized by Luo Luo.
Please do not distribute.

Remark 6.1. *Let* $\mathbf{u}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|_2$ *and* $\mathbf{u}_2 = \mathbf{v}_2 / \|\mathbf{v}_2\|_2$*, we have*

$$r = \frac{a_{12}}{\sqrt{a_{11}a_{22}}} = \frac{\mathbf{v}_1^\top \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2} = \mathbf{u}_1^\top \mathbf{u}_2.$$

Recall that $\mathbf{u}_1 \sim \mathrm{Unif}(\mathcal{S}^{n-1})$ *and* $\mathbf{u}_2 \sim \mathrm{Unif}(\mathcal{S}^{n-1})$ *are independent since* $\sigma_{12} = 0$ *leads to* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *are independent. If* $n \to +\infty$*, the student t-distribution tends to standard normal distribution, then* $|r|$ *tends to zero. In other words, two independent uniformly distributed random vectors on high dimensional unit sphere are almost orthogonal.*

Remark 6.2. *Applying Lemma 6.3, we have*

$$\begin{aligned} f(\mathbf{A}) =& f(a_{11}, a_{12}, a_{22}) \\ =& f(a_{11.2}, a_{12}, a_{22}) \\ =& f(a_{11.2}) f(a_{12}, a_{22}) \\ =& f(a_{11.2}) f(a_{12} \,|\, a_{22}) f(a_{22}) \end{aligned}$$

where the second line is because of the transform from

$$[a_{11}, a_{12}, a_{22}]^\top \qquad \text{to} \qquad [a_{11.2}, a_{12}, a_{22}]^\top = [a_{11} - a_{12}^2 a_{22}, a_{12}, a_{22}]^\top$$

has Jacobian

$$\begin{bmatrix} \dfrac{\partial(a_{11} - a_{12}^2 a_{22})}{\partial a_{11}} & \dfrac{\partial(a_{11} - a_{12}^2 a_{22})}{\partial a_{12}} & \dfrac{\partial(a_{11} - a_{12}^2 a_{22})}{\partial a_{22}} \\ \dfrac{\partial a_{12}}{\partial a_{11}} & \dfrac{\partial a_{12}}{\partial a_{12}} & \dfrac{\partial a_{12}}{\partial a_{22}} \\ \dfrac{\partial a_{22}}{\partial a_{11}} & \dfrac{\partial a_{22}}{\partial a_{12}} & \dfrac{\partial a_{22}}{\partial a_{22}} \end{bmatrix} = \begin{bmatrix} 1 & \times & \times \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the third line is because of $a_{11.2}$ *is independent on* a_{12} *and* a_{22}*. In the case of* $\sigma_{12} = 0$*, we have*

$$\frac{a_{11.2}}{\sigma_{11.2}} \sim \chi^2_{n-1}, \qquad a_{12} \mid a_{22} \sim \mathcal{N}\left(\sigma_{22}^{-1}\sigma_{12}a_{22}, \sigma_{11.2}a_{22}\right) \qquad \text{and} \qquad \frac{a_{22}}{\sigma_{22}} \sim \chi^2_n,$$

the density of $\mathbf{A}$ *is*

$$\frac{(\det(\mathbf{A}))^{-\frac{n-3}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{A})\right)}{2^n \sqrt{\pi}\,(\det(\mathbf{\Sigma}))^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}.$$

This document is organized by Luo Luo.
Please do not distribute.

Remark 6.3. *In the case of* $\rho > 0$*, the derivation for the density of* r *is somewhat complicated, please see Section 4.2.2 of Anderson [1].*

Theorem 6.1. *If* x *and* y *are independently distributed,* x *having the distribution* $\mathcal{N}(0, 1)$ *and* y *having the* χ^2*-distribution with* m *degrees of freedom, then* $z = x/\sqrt{y/m}$ *(has t-distribution with* m *degrees of freedom) has the density*

$$\frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\,\Gamma(\frac{m}{2})} \left(1 + \frac{z^2}{m}\right)^{-\frac{m+1}{2}}.$$

Proof. We present the brief ideas:

1. Write down the joint density of x and y.
2. Use the fact $x = z\sqrt{y/m}$ to write the joint density of y and z.
3. Integrate out y on the joint density of y and z.

□

Theorem 6.2. *Consider the likelihood ratio test of the hypothesis that $\rho = \rho_0$ based on a sample $\mathbf{x}_1, \dots, \mathbf{x}_N$ from the bivariate normal distribution*

$$\mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right).$$

Define the set

$$\Omega = \big\{(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) : \boldsymbol{\mu} \in \mathbb{R}^2, \sigma_1 > 0, \sigma_2 > 0, \boldsymbol{\Sigma} \succ \mathbf{0}\big\}$$

and its subset

$$\omega = \big\{(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) : \boldsymbol{\mu} \in \mathbb{R}^2, \sigma_1 > 0, \sigma_2 > 0, \boldsymbol{\Sigma} \succ \mathbf{0}, \rho = \rho_0\big\}, \quad \textit{where} \quad \boldsymbol{\Sigma} = \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}.$$

The likelihood ratio criterion is

$$\frac{\sup_{\boldsymbol{\theta}\in\omega} L(\mathbf{x}, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Omega} L(\mathbf{x}, \boldsymbol{\theta})} = \left(\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right)^{\frac{N}{2}},$$

where

$$r = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}}, \quad \mathbf{A} = \begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \quad \textit{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha.$$

Proof. The maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ on Ω are

$$\boldsymbol{\Sigma}_\Omega = \frac{1}{N}\mathbf{A} \qquad \text{and} \qquad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha,$$

where

$$\mathbf{A} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = \begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix}.$$

The corresponding likelihood function value is

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Omega} L(\mathbf{x}, \boldsymbol{\theta}) = & \max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ = & \prod_{\alpha=1}^{N}\frac{1}{\sqrt{(2\pi)^p\det(\boldsymbol{\Sigma}_\Omega)}}\exp\left(-\frac{1}{2}(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\boldsymbol{\Sigma}_\Omega^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\right) \\ = & (2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\boldsymbol{\Sigma}_\Omega^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\right).\end{aligned}$$

Note that

$$\begin{aligned}& \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\boldsymbol{\Sigma}_\Omega^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}}) \\ = & \sum_{\alpha=1}^{N}\operatorname{tr}\left((\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\boldsymbol{\Sigma}_\Omega^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\right)\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
&=\sum_{\alpha=1}^{N} \operatorname{tr}\left(\boldsymbol{\Sigma}_\Omega^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\right)\\
&=\operatorname{tr}\left(\boldsymbol{\Sigma}_\Omega^{-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\right)\\
&=\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{A}\right)^{-1}\mathbf{A}\right)=\operatorname{tr}\left(N\mathbf{I}_p\right)=Np,
\end{aligned}
$$

then the likelihood maximized in Ω is

$$
\sup_{\boldsymbol{\theta}\in\Omega} L(\mathbf{x},\boldsymbol{\theta}) = \max_{\boldsymbol{\mu}\in\mathbb{R}^p,\boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right)
$$

We take $p=2$ in this proof, then

$$
\det(\boldsymbol{\Sigma}_\Omega)=\det\left(\frac{1}{N}\mathbf{A}\right)=\frac{1}{N^p}\det(\mathbf{A})=\frac{a_{11}a_{22}-a_{12}a_{21}}{N^2},
$$

which implies

$$
\sup_{\boldsymbol{\theta}\in\Omega} L(\mathbf{x},\boldsymbol{\theta})=\frac{N^N\exp\left(-N\right)}{(2\pi)^N\left(a_{11}a_{22}-a_{12}a_{21}\right)^{\frac{N}{2}}}=\frac{N^N\exp\left(-N\right)}{(2\pi)^N(1-r^2)^{\frac{N}{2}}a_{11}^{\frac{N}{2}}a_{22}^{\frac{N}{2}}},\quad \text{where}\quad r=\frac{a_{12}}{\sqrt{a_{11}a_{22}}}.
$$

Let $\sigma^2=\sigma_1\sigma_2$ and $\tau=\sigma_1/\sigma_2$. Under the null hypothesis $\rho=\rho_0$, we have

$$
\det(\boldsymbol{\Sigma})=\sigma_1^2\sigma_2^2-\sigma_1^2\sigma_2^2\rho_0^2=\sigma^4(1-\rho_0^2),\quad \boldsymbol{\Sigma}^{-1}=\frac{1}{1-\rho_0^2}\begin{bmatrix}\dfrac{1}{\sigma_1^2} & -\dfrac{\rho_0}{\sigma_1\sigma_2}\\ -\dfrac{\rho_0}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2}\end{bmatrix}
$$

and

$$
\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})=\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}\right)\\
&=\frac{1}{1-\rho_0^2}\operatorname{tr}\left(\begin{bmatrix}\dfrac{1}{\sigma_1^2} & -\dfrac{\rho_0}{\sigma_1\sigma_2}\\ -\dfrac{\rho_0}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2}\end{bmatrix}\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\right)\\
&=\frac{1}{1-\rho_0^2}\left(\frac{a_{11}}{\sigma_1^2}-\frac{2\rho_0 a_{12}}{\sigma_1\sigma_2}+\frac{a_{22}}{\sigma_2^2}\right)\\
&=\frac{1}{(1-\rho_0^2)\sigma^2}\left(\frac{a_{11}}{\tau}-2\rho_0 a_{12}+\tau a_{22}\right).
\end{aligned}
$$

Then the likelihood function under the null hypothesis ($\rho=\rho_0$) is

$$
\frac{1}{(2\pi)^N(1-\rho_0^2)^{\frac{N}{2}}(\sigma^2)^N}\exp\left(-\frac{a_{11}/\tau-2\rho_0a_{12}+\tau a_{22}}{2\sigma^2(1-\rho_0^2)}\right). \tag{24}
$$

The maximum of (24) with respect to τ occurs at

$$
\hat{\tau}=\sqrt{a_{11}/a_{22}},
$$

then the concentrated likelihood is

$$
\frac{1}{(2\pi)^N(1-\rho_0^2)^{\frac{N}{2}}(\sigma^2)^N}\exp\left(-\frac{\sqrt{a_{11}}\sqrt{a_{22}}\,(1-\rho_0 r)}{\sigma^2(1-\rho_0^2)}\right). \tag{25}
$$

This document is organized by Luo Luo.
Please do not distribute.

The maximum of (25) occurs at

$$\hat{\sigma}^2 = \frac{\sqrt{a_{11}}\sqrt{a_{22}}(1-\rho_0 r)}{N(1-\rho_0^2)},$$

which is because of $f(x) = \exp(-b/x)/x^N$ leads to

$$f'(x) = \frac{\exp(-\frac{b}{x})\cdot\frac{b}{x^2}\cdot x^N - \exp(-\frac{b}{x})\cdot Nx^{N-1}}{x^{2N}} = \frac{\exp(-\frac{b}{x})x^{N-2}(b-Nx)}{x^{2N}}.$$

Hence, taking $x = b/N$ results maximum value

$$f\left(\frac{b}{N}\right) = \frac{N^N\exp(-N)}{b^N}.$$

We obtain $\hat{\sigma}^2$ by setting

$$b = \frac{\sqrt{a_{11}}\sqrt{a_{22}}\,(1-\rho_0 r)}{1-\rho_0^2} \qquad \text{and} \qquad x = \sigma^2,$$

leading to

$$\sup_{\boldsymbol{\theta}\in\omega} L(\mathbf{x},\boldsymbol{\theta}) = \frac{N^N\exp(-N)\,(1-\rho_0^2)^{\frac{N}{2}}}{(2\pi)^N(1-\rho_0 r)^N a_{11}^{\frac{N}{2}} a_{22}^{\frac{N}{2}}}$$

The likelihood ratio criterion is, therefore,

$$\frac{\sup_{\boldsymbol{\theta}\in\omega} L(\mathbf{x},\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Omega} L(\mathbf{x},\boldsymbol{\theta})} = \left(\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right)^{\frac{N}{2}}.$$

□

Remark 6.4. *The likelihood ratio test is*

$$\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2} \leq c$$

where c is chosen by the prescribed significance level. The critical region can be written equivalently as

$$(\rho_0^2 c - \rho_0^2 + 1)r^2 - 2\rho_0 cr + c - 1 + \rho_0^2 \geq 0,$$

that is,

$$r > \frac{\rho_0 c + (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c - \rho_0^2 + 1} = r_1 \qquad \textit{and} \qquad r < \frac{\rho_0 c - (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c - \rho_0^2 + 1} = r_2.$$

Thus the likelihood ratio test of $H : \rho = \rho_0$ against alternatives $\rho \neq \rho_0$ has a rejection region of the form $r > r_1$ and $r < r_2$ (not chosen so that the probability of each inequality is $\alpha/2$ when H is true).

Lemma 6.4 (homework, see Section 2.6.2 of Anderson [1]). *For random vector*

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \textit{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{ji} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}.$$

Then $\mathbb{E}[(x_i-\mu_i)(x_j-\mu_j)(x_k-\mu_k)] = 0$ and $\mathbb{E}[(x_i-\mu_i)(x_j-\mu_j)(x_k-\mu_k)(x_l-\mu_l)] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

This document is organized by Luo Luo.
Please do not distribute.

Theorem 6.3. *Let*

$$\mathbf{A}(n) = \sum_{\alpha=1}^{N} \big(\mathbf{x}_\alpha - \bar{\mathbf{x}}_N\big)\big(\mathbf{x}_\alpha - \bar{\mathbf{x}}_N\big)^\top,$$

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ *are independently distributed according to* $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $n = N - 1$. *Then the limiting distribution of*

$$\mathbf{B}(n) = \frac{1}{\sqrt{n}}\big(\mathbf{A}(n) - n\boldsymbol{\Sigma}\big)$$

is normal with mean $\mathbf{0}$ *and covariance* $\mathbb{E}\big[b_{ij}(n)b_{kl}(n)\big] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

Proof. We have

$$\mathbf{A}(n) = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top,$$

where $\mathbf{z}_1, \dots, \mathbf{z}_n$ are distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We arrange the elements of

$$\mathbf{z}_\alpha \mathbf{z}_\alpha^\top = \begin{bmatrix} z_{1\alpha}^2 & z_{1\alpha}z_{2\alpha} & \cdots & z_{1\alpha}z_{p\alpha} \\ z_{2\alpha}z_{1\alpha} & z_{2\alpha}^2 & \cdots & z_{2\alpha}z_{p\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p\alpha}z_{1\alpha} & z_{p\alpha}z_{2\alpha} & \cdots & z_{p\alpha}^2 \end{bmatrix}$$

in a p^2-dimensional random vector such as

$$\mathbf{y}_\alpha = \begin{bmatrix} z_{1\alpha}^2 \\ z_{1\alpha}z_{2\alpha} \\ \vdots \\ z_{2\alpha}^2 \\ \vdots \\ z_{p\alpha}^2 \end{bmatrix}, \qquad \text{where we write } \mathbf{z}_\alpha = \begin{bmatrix} z_{1\alpha} \\ \vdots \\ z_{p\alpha} \end{bmatrix}.$$

The second moments of $\mathbf{y}_\alpha$ can be deduced from the forth moments of $\mathbf{z}_\alpha$ by using Lemma 6.4, that is,

$$\mathbb{E}[z_{i\alpha}z_{j\alpha}] = \sigma_{ij}, \qquad \mathbb{E}[z_{i\alpha}z_{j\alpha}z_{k\alpha}z_{l\alpha}] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},$$

and (covariance of the entries of $\mathbf{y}_\alpha$)

$$\begin{aligned} & \mathbb{E}[(z_{i\alpha}z_{j\alpha} - \mathbb{E}[z_{i\alpha}z_{j\alpha}])(z_{k\alpha}z_{l\alpha} - \mathbb{E}[z_{k\alpha}z_{l\alpha}])] \\ =& \mathbb{E}[(z_{i\alpha}z_{j\alpha} - \sigma_{ij})(z_{k\alpha}z_{l\alpha} - \sigma_{kl})] \\ =& \mathbb{E}[z_{i\alpha}z_{j\alpha}z_{k\alpha}z_{l\alpha}] - \sigma_{ij}\mathbb{E}[z_{k\alpha}z_{l\alpha}] - \sigma_{kl}\mathbb{E}[z_{i\alpha}z_{j\alpha}] + \sigma_{ij}\sigma_{kl} \\ =& \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} - \sigma_{ij}\sigma_{kl} - \sigma_{kl}\sigma_{ij} + \sigma_{ij}\sigma_{kl} \\ =& \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \end{aligned} \tag{26}$$

Arranging the elements of $\mathbf{A}(n)$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{w}(n) = \begin{bmatrix} a_{11}(n) \\ a_{12}(n) \\ \vdots \\ a_{22}(n) \\ \vdots \\ a_{pp}(n) \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\nu} = \begin{bmatrix} \sigma_{11} \\ \sigma_{12} \\ \vdots \\ \sigma_{22} \\ \vdots \\ \sigma_{pp} \end{bmatrix},$$

This document is organized by Luo Luo.
Please do not distribute.

we obtain

$$\frac{1}{\sqrt{n}}\big(\mathbf{w}(n) - n\boldsymbol{\nu}\big) = \frac{1}{\sqrt{n}}\sum_{\alpha=1}^{n}\big(\mathbf{y}_\alpha - \boldsymbol{\nu}\big).$$

Since $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance of $\mathbf{y}_\alpha$ satisfies (26), the multivariate central limit theorem implies the desired result. □

Theorem 6.4 ([15, Section 3.3])**.** *Let $\{\mathbf{u}(n)\}$ be a sequence of m-component random vectors and $\mathbf{b}$ a fixed vector such that*

$$\lim_{n\to\infty}\sqrt{n}(\mathbf{u}(n) - \mathbf{b}) \sim \mathcal{N}(\mathbf{0}, \mathbf{T}).$$

Let $\mathbf{f}(\mathbf{u})$ be a vector-valued function of $\mathbf{u}$ such that each component $f_j(\mathbf{u})$ has a nonzero differential at $\mathbf{u} = \mathbf{b}$, and define $\mathbf{\Phi}_\mathbf{b}$ with its (i,j)-th component being

$$\frac{\partial f_j(\mathbf{u})}{\partial u_i}\Big|_{\mathbf{u}=\mathbf{b}}.$$

Then $\sqrt{n}(\mathbf{f}(\mathbf{u}(n)) - f(\mathbf{b}))$ has the limiting distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Phi}_\mathbf{b}^\top\mathbf{T}\mathbf{\Phi}_\mathbf{b})$.

We can write

$$r(n) = \frac{a_{ij}(n)}{\sqrt{a_{ii}(n)}\sqrt{a_{jj}(n)}} = \frac{c_{ij}(n)}{\sqrt{c_{ii}(n)}\sqrt{c_{jj}(n)}},$$

with

$$c_{ii}(n) = \frac{a_{ii}(n)}{\sigma_{ii}}, \qquad c_{ij}(n) = \frac{a_{ij}(n)}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \quad \text{and} \quad c_{jj}(n) = \frac{a_{ii}(n)}{\sigma_{jj}}.$$

We can study $r(n)$ by $\mathbf{C}(n)$ and $\mathbf{z}_\alpha^*$ which is defined as

$$c_{ij}(n) = \sum_{\alpha=1}^{n}\begin{bmatrix} z_{i\alpha}^* \\ z_{j\alpha}^* \end{bmatrix}\begin{bmatrix} z_{i\alpha}^* & z_{j\alpha}^* \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} z_{i\alpha}^* \\ z_{j\alpha}^* \end{bmatrix} = \begin{bmatrix} \dfrac{z_{i\alpha}}{\sqrt{\sigma_{ii}}} \\ \dfrac{z_{j\alpha}}{\sqrt{\sigma_{jj}}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \quad \text{and} \quad \rho = \frac{\sigma_{ij}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{jj}}}.$$

We apply Theorem 6.3 with

$$\mathbf{A}(n) = \mathbf{C}(n) \qquad \text{and} \qquad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sigma_{jj} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

and

$$\mathbf{u}(n) = \frac{1}{n}\begin{bmatrix} c_{ii}(n) \\ c_{jj}(n) \\ c_{ij}(n) \end{bmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ \rho \end{bmatrix}$$

then the vector

$$\sqrt{n}(\mathbf{u}(n) - \mathbf{b}) = \frac{1}{\sqrt{n}}\left(\begin{bmatrix} c_{ii}(n) \\ c_{jj}(n) \\ c_{ij}(n) \end{bmatrix} - n\mathbf{b}\right)$$

has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{bmatrix} \sigma_{ii}\sigma_{ii} + \sigma_{ii}\sigma_{ii} & \sigma_{ij}\sigma_{ij} + \sigma_{ij}\sigma_{ij} & \sigma_{ii}\sigma_{ij} + \sigma_{ij}\sigma_{ii} \\ \sigma_{ij}\sigma_{ij} + \sigma_{ij}\sigma_{ij} & \sigma_{jj}\sigma_{jj} + \sigma_{jj}\sigma_{jj} & \sigma_{ij}\sigma_{jj} + \sigma_{jj}\sigma_{ij} \\ \sigma_{ii}\sigma_{ij} + \sigma_{ij}\sigma_{ii} & \sigma_{ij}\sigma_{jj} + \sigma_{jj}\sigma_{ij} & \sigma_{ii}\sigma_{jj} + \sigma_{ij}\sigma_{ij} \end{bmatrix} = \begin{bmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{bmatrix}.$$

This document is organized by Luo Luo.
Please do not distribute.

Applying Theorem 6.4 with $r = f(\mathbf{u}) = u_3 u_1^{-\frac{1}{2}} u_2^{-\frac{1}{2}}$, we have $f(\mathbf{b}) = \rho$ and

$$\mathbf{\Phi}_{\mathbf{b}} = \begin{bmatrix} \dfrac{\partial r}{\partial u_1}\Big|_{\mathbf{u}=\mathbf{b}} \\ \dfrac{\partial r}{\partial u_2}\Big|_{\mathbf{u}=\mathbf{b}} \\ \dfrac{\partial r}{\partial u_3}\Big|_{\mathbf{u}=\mathbf{b}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} u_3 u_1^{-\frac{3}{2}} u_2^{-\frac{1}{2}}\Big|_{\mathbf{u}=\mathbf{b}} \\ -\frac{1}{2} u_3 u_1^{-\frac{1}{2}} u_2^{-\frac{3}{2}}\Big|_{\mathbf{u}=\mathbf{b}} \\ u_1^{-\frac{1}{2}} u_2^{-\frac{1}{2}}\Big|_{\mathbf{u}=\mathbf{b}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{bmatrix}.$$

Thus, the covariance of the limiting distribution of $\sqrt{n}(r(n) - \rho)$ is

$$\begin{bmatrix} -\frac{1}{2}\rho & -\frac{1}{2}\rho & 1 \end{bmatrix} \begin{bmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{bmatrix} = (1-\rho^2)^2$$

and we have $\lim\limits_{n\to\infty} \dfrac{\sqrt{n}(r(n) - \rho)}{1-\rho^2} \sim \mathcal{N}(0,1)$.

Remark 6.5. *In the case of $\rho = \sigma_{12} = 0$, recall that we have shown that*

$$\frac{\sqrt{n-1}\,r(n)}{\sqrt{1-(r(n))^2}} \sim t_{n-1}.$$

Intuitively, the sample correlation $r(n)$ converges to $\rho = 0$ and t_{n-1} converges to $\mathcal{N}(0,1)$ when $n \to \infty$, which matches our asymptotic result.

7 The Wishart Distribution

In this section, we always suppose the Wishart distribution is non-singular and focus on the case of Wishart random matrix is positive-definite.

Lemma 7.1. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{C} \in \mathbb{R}^{q\times p}$, then*

$$\mathbf{C}\mathbf{A}\mathbf{C}^\top \sim \mathcal{W}_p(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top, n)$$

Proof. We can write

$$\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top,$$

where $\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent, each with the distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{y}_\alpha = \mathbf{C}\mathbf{z}_\alpha$, then we have

$$\mathbf{y}_\alpha \sim \mathcal{N}_p(\mathbf{0}, \mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top) \qquad \text{and} \qquad \mathbf{C}\mathbf{A}\mathbf{C}^\top = \sum_{\alpha=1}^{n} \mathbf{C}\mathbf{z}_\alpha \mathbf{z}_\alpha^\top \mathbf{C}^\top = \sum_{\alpha=1}^{n} \mathbf{y}_\alpha \mathbf{y}_\alpha^\top \sim \mathcal{W}_p(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top, n).$$

□

Lemma 7.2. *Let*

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times p}$$

where $\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent, each with the distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. For projection matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ with rank-r, we have

$$\mathbf{Z}^\top \mathbf{Q}\mathbf{Z} \sim \mathcal{W}_p(\mathbf{\Sigma}, r).$$

This document is organized by Luo Luo.
Please do not distribute.

Proof. We denote SVD of $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top, \qquad \text{where } \mathbf{V} \in \mathbb{R}^{p\times p} \text{ is orthogonal and } \mathbf{D} \in \mathbb{R}^{p\times p} \text{ is diagonal.}$$

Let

$$\mathbf{y}_\alpha = \boldsymbol{\Sigma}^{-1/2}\mathbf{z}_\alpha, \qquad \text{where} \quad \boldsymbol{\Sigma}^{-1/2} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^\top \quad (\text{we also denote } \boldsymbol{\Sigma}^{1/2} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^\top).$$

Then we have

$$\mathbf{y}_\alpha \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}).$$

The projection matrix $\mathbf{Q}$ has SVD as follows

$$\mathbf{Q} = \mathbf{U}\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{U}^\top, \qquad \text{where } \mathbf{U} \in \mathbb{R}^{n\times n} \text{ is orthogonal.}$$

Let

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1^\top\boldsymbol{\Sigma}^{-1/2} \\ \vdots \\ \mathbf{z}_n^\top\boldsymbol{\Sigma}^{-1/2} \end{bmatrix} = \mathbf{Z}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n\times p} \qquad \text{and} \qquad \mathbf{W} = \mathbf{U}^\top\mathbf{Y} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times p},$$

then we have

$$\mathbf{Y}^\top\mathbf{Q}\mathbf{Y} = \mathbf{Y}^\top\mathbf{U}\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{U}^\top\mathbf{Y} = \mathbf{W}^\top\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{W} = \sum_{\alpha=1}^{r}\mathbf{w}_\alpha\mathbf{w}_\alpha^\top.$$

Since each entry of $\mathbf{Y}$ has distribution $\mathcal{N}(0,1)$ independently and $\mathbf{U}$ is orthogonal, each entry of $\mathbf{W}$ also has distribution $\mathcal{N}(0,1)$ independently. This implies each $\mathbf{w}_\alpha$ has distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$ independently and

$$\mathbf{Y}^\top\mathbf{Q}\mathbf{Y} \sim \mathcal{W}_p(\mathbf{I}, r).$$

Note that $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Sigma}^{-1/2}$ implies

$$\mathbf{Z} = \mathbf{Y}\boldsymbol{\Sigma}^{1/2}$$

then we have

$$\mathbf{Z}^\top\mathbf{Q}\mathbf{Z} = \boldsymbol{\Sigma}^{1/2}\mathbf{Y}\mathbf{Q}\mathbf{Y}\boldsymbol{\Sigma}^{1/2}$$

and Lemma 7.1 leads to

$$\mathbf{Z}^\top\mathbf{Q}\mathbf{Z} \sim \mathcal{W}_p(\mathbf{I}, \boldsymbol{\Sigma}).$$

□

Before consider the density of Wishart distribution, we review some results for determinant of Jacobian in matrix transformation.

Lemma 7.3. *Let* $\mathbf{X} \in \mathbb{R}^{p\times q}$ *and let* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *be a constant matrix. Then Jacobian of the transform* $\mathbf{Y} = \mathbf{A}\mathbf{X}$ *(from* $\mathbb{R}^{p\times q}$ *to* $\mathbb{R}^{p\times q}$*) has the determinant* $(\det(\mathbf{A}))^p$.

Proof. We write the transform $\mathbf{Y} = \mathbf{A}\mathbf{X}$ as

$$\begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_q \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{x}_1 & \dots & \mathbf{A}\mathbf{x}_q \end{bmatrix},$$

which means

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_q \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{x}_1 \\ \vdots \\ \mathbf{A}\mathbf{x}_q \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} \end{bmatrix}\begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_q \end{bmatrix} \quad \Longrightarrow \quad \det\left(\begin{bmatrix} \mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} \end{bmatrix}\right) = (\det(\mathbf{A}))^q.$$

□

This document is organized by Luo Luo.
Please do not distribute.

Remark 7.1. *Given constant matrix* $\mathbf{B} \in \mathbb{R}^{q\times q}$ *and Jacobian of transform* $\mathbf{Z} = \mathbf{XB}$ *(from* $\mathbb{R}^{p\times q}$ *to* $\mathbb{R}^{q\times q}$*) has determinant* $(\det(\mathbf{B}))^p$*, since we have*

$$\begin{bmatrix} \mathbf{z}_{(1)}^\top \\ \vdots \\ \mathbf{z}_{(p)}^\top \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{(1)}^\top \mathbf{B} \\ \vdots \\ \mathbf{x}_{(p)}^\top \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{(1)}^\top \\ \vdots \\ \mathbf{x}_{(p)}^\top \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{B} \end{bmatrix} \quad \Longrightarrow \quad \begin{bmatrix} \mathbf{z}_{(1)} \\ \vdots \\ \mathbf{z}_{(p)} \end{bmatrix} \begin{bmatrix} \mathbf{B}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{B}^\top \end{bmatrix} \begin{bmatrix} \mathbf{x}_{(1)} \\ \vdots \\ \mathbf{x}_{(p)} \end{bmatrix}.$$

Following these notations, the Jacobian of transform $\mathbf{W} = \mathbf{AXB}$ *has determinant* $(\det(\mathbf{A}))^q(\det(\mathbf{B}))^p$*.*

Lemma 7.4. *Let* $\mathbf{X} \in \mathbb{R}^{p\times p}$ *be a symmetric matrix of* $p(p+1)/2$ *functionally independent real elements and let* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *be a non-singular constant matrix. Then Jacobian of transform* $\mathbf{Y} = \mathbf{AXA}^\top$ *has the determinant* $(\det(\mathbf{A}))^{p+1}$*.*

Proof. The non-singular $\mathbf{A}$ has be written as the product of elementary matrices, i.e.,

$$\mathbf{A} = \mathbf{E}_k \cdot \dots \cdot \mathbf{E}_1 \qquad \Longrightarrow \qquad \mathbf{Y} = \mathbf{E}_k \dots \mathbf{E}_1 \mathbf{X} \mathbf{E}_1^\top \dots \mathbf{E}_k^\top,$$

where $\mathbf{E}_k \dots \mathbf{E}_1$ and $\mathbf{E}_1^\top \dots \mathbf{E}_k^\top$ correspond to row transform and column transform respectively.
We consider the transform $\mathbf{Y}_i = \mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top$ as follows:

1. If $\mathbf{E}_i$ corresponds to row multiplication, we can write

$$\mathbf{E}_i = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & 1 & 0 & 0 & \cdots \\ \cdots & 0 & c_i & 0 & \cdots \\ \cdots & 0 & 0 & 1 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \text{and} \quad \mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & \times & c_i \cdot x_{\alpha-1,\alpha} & \times & \cdots \\ \cdots & c_i \cdot x_{\alpha,\alpha-1} & c_i^2 \cdot x_{\alpha,\alpha} & c_i x_{\alpha,\alpha+1} & \cdots \\ \cdots & \times & c_i \cdot x_{\alpha+1,\alpha} & \times & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

   Since the matrix $\mathbf{A}$ is symmetric, we only require considering the lower triangle (or upper triangle) of $\mathbf{A}$. There are $p-1$ off-diagonal entries is multiplied by c_i and one diagonal entry is multiplied by c_i^2. Hence, the Jacobian of this transform has determinant c_i^{p+1}.

2. If $\mathbf{E}_i$ corresponds to row addition, we consider the example of add the 1st row (column) to the 2nd row (column), that is

$$\mathbf{E}_i = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \text{and} \quad \mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top = \begin{bmatrix} x_{11} & x_{11} + x_{12} & x_{11} & \dots \\ x_{11} + x_{12} & x_{11} + 2x_{12} + x_{22} & x_{13} + x_{23} & \dots \\ x_{13} & x_{13} + x_{23} & x_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

   We focus on the changed entries in lower triangle of $\mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top$. Each of these entries depends on the corresponding entry of $\mathbf{X}_i$ whose position is on the more left column or the same column but on upper position. We arrange the lower triangle of $\mathbf{X}_i$ to a $p(p+1)/2$-dimensional vector by stack its column and transform it into the corresponding $p(p+1)/2$-dimensional vector of lower triangle of $\mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top$, then the Jacobian is a $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$ lower triangle matrix and each entry of its diagonal is one. Hence, the Jacobian of $\mathbf{E}_i \mathbf{X}_i \mathbf{E}_i^\top$ for symmetric $\mathbf{X}_i$ is 1.

Above observation indicates the Jacobian of $\mathbf{Y} = \mathbf{E}_k \dots \mathbf{E}_1 \mathbf{X} \mathbf{E}_1^\top \dots \mathbf{E}_k^\top$ has determinant

$$\prod_{i=1}^k c_i^{p+1} = \prod_{i=1}^k (\det(\mathbf{E}_i))^{p+1} = \left(\det\left(\prod_{i=1}^k \mathbf{E}_i\right)\right)^{p+1} = (\det(\mathbf{A}))^{p+1},$$

where we define $c_i = 1$ if $\mathbf{E}_i$ corresponds to addition matrix. □

This document is organized by Luo Luo. Please do not distribute.

Remark 7.2. *We typically include three types of elementary matrices, which correspond to three types of row operations (respectively, column operations), that is (a) the row switching, (b) row multiplication and (c) row addition (with nonzero weight). In fact, we only require row multiplication and row addition (with weight one):*

1. *For row addition (with nonzero weight), we have*

$$\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}c\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}c\mathbf{a}_1\\ \mathbf{a}_2+c\mathbf{a}_1\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2+c\mathbf{a}_1\end{bmatrix}.$$

2. *For row switching, we have*

$$\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2+\mathbf{a}_1\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}\mathbf{a}_1+(-1)\cdot(\mathbf{a}_2+\mathbf{a}_1)\\ \mathbf{a}_2+\mathbf{a}_1\end{bmatrix} = \begin{bmatrix}-\mathbf{a}_2\\ \mathbf{a}_2+\mathbf{a}_1\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}-\mathbf{a}_2\\ \mathbf{a}_1\end{bmatrix} \quad\longrightarrow\quad \begin{bmatrix}\mathbf{a}_2\\ \mathbf{a}_1\end{bmatrix}.$$

Lemma 7.5. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and partition $\mathbf{A}$ and $\mathbf{\Sigma}$ into q and $p-q$ rows and columns as*

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix} \qquad \textit{and} \qquad \mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix},$$

then we have

(a) $\mathbf{A}_{11} \sim \mathcal{W}_q(\mathbf{\Sigma}_{11}, n)$ *and* $\mathbf{A}_{22} \sim \mathcal{W}_{p-q}(\mathbf{\Sigma}_{22}, n)$*;*

(b) if $q = 1$, then

$$\mathbf{a}_{21} \mid \mathbf{A}_{22} \sim \mathcal{N}_{p-q}(\mathbf{A}_{22}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11.2}^2\mathbf{A}_{22})$$

where $\sigma_{11.2}^2 = \sigma_{11} - \boldsymbol{\sigma}_{21}^\top\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$*;*

(c) if $n > p-q$, then

$$\mathbf{A}_{11.2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \sim \mathcal{W}_q(\mathbf{\Sigma}_{11.2}, n-p+q)$$

is independent on $\mathbf{A}_{22}$ and $\mathbf{A}_{21}$, where $\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.

Proof. **Part (a):** Apply Lemma 7.1 with $\mathbf{C} = \begin{bmatrix}\mathbf{I}_q & \mathbf{0}\end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix}\mathbf{0} & \mathbf{I}_{p-q}\end{bmatrix}$ respectively.

Part (b): We partition the vector and matrix into q and $p-q$ columns (rows) as follows

$$\mathbf{z}_\alpha = \begin{bmatrix}\mathbf{z}_{1\alpha}\\ \mathbf{z}_{2\alpha}\end{bmatrix} \text{ and } \mathbf{Z} = \begin{bmatrix}\mathbf{z}_1^\top\\ \vdots\\ \mathbf{z}_n^\top\end{bmatrix} = \begin{bmatrix}\mathbf{z}_{11}^\top & \mathbf{z}_{21}^\top\\ \vdots & \vdots\\ \mathbf{z}_{1n}^\top & \mathbf{z}_{2n}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{bmatrix} \in \mathbb{R}^{n\times p}, \quad \text{where } \mathbf{Z}_1 = \begin{bmatrix}\mathbf{z}_{11}^\top\\ \vdots\\ \mathbf{z}_{1n}^\top\end{bmatrix} \text{ and } \mathbf{Z}_2 = \begin{bmatrix}\mathbf{z}_{21}^\top\\ \vdots\\ \mathbf{z}_{2n}^\top\end{bmatrix}$$

Then we can write

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix} = \mathbf{Z}^\top\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1^\top\\ \mathbf{Z}_2^\top\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_1^\top\mathbf{Z}_1 & \mathbf{Z}_1^\top\mathbf{Z}_2\\ \mathbf{Z}_2^\top\mathbf{Z}_1 & \mathbf{Z}_2^\top\mathbf{Z}_2\end{bmatrix}$$

Consider the conditional distribution of $\mathbf{z}_{1\alpha}$ for given $\mathbf{z}_{2\alpha}$, we have

$$\mathbf{z}_{1\alpha} \mid \mathbf{z}_{2\alpha} \sim \mathcal{N}_q\left(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{z}_{2\alpha}, \mathbf{\Sigma}_{11.2}\right).$$

In this statement, we take $q = 1$. We can write

$$\mathbf{\Sigma} = \begin{bmatrix}a_{11} & \boldsymbol{\sigma}_{21}^\top\\ \boldsymbol{\sigma}_{21} & \mathbf{A}_{22}\end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix}a_{11} & \mathbf{a}_{21}^\top\\ \mathbf{a}_{21} & \mathbf{A}_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{z}_1^\top\mathbf{z}_1 & \mathbf{z}_1^\top\mathbf{Z}_2\\ \mathbf{Z}_2^\top\mathbf{z}_1 & \mathbf{Z}_2^\top\mathbf{Z}_2\end{bmatrix} \quad \text{and} \quad z_{1\alpha} \mid \mathbf{z}_{2\alpha} \sim \mathcal{N}_q\left(\boldsymbol{\sigma}_{21}^\top\mathbf{\Sigma}_{22}^{-1}\mathbf{z}_{2\alpha}, \sigma_{11.2}\right),$$

This document is organized by Lu Lu.
Please do not distribute.

then

$$\boldsymbol{\sigma}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_{2\alpha} = \mathbf{z}_{2\alpha}^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \in \mathbb{R} \qquad \text{and} \qquad \begin{bmatrix} \boldsymbol{\sigma}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_{21} \\ \vdots \\ \boldsymbol{\sigma}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_{2n} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \\ \vdots \\ \mathbf{z}_{2n}^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \end{bmatrix} = \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \in \mathbb{R}^n.$$

Hence, we have

$$\mathbf{z}_1 = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1n} \end{bmatrix} \quad \text{and} \quad \mathbf{z}_1 \mid \mathbf{Z}_2 \sim \mathcal{N}_n(\mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11.2}\mathbf{I}) \quad \Longrightarrow \quad \mathbf{Z}_2^\top \mathbf{z}_1 \mid \mathbf{Z}_2 \sim \mathcal{N}_{p-q}(\mathbf{Z}_2^\top \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11.2}\mathbf{Z}_2^\top \mathbf{Z}_2),$$

that is $\mathbf{a}_{21} \,|\, \mathbf{A}_{22} \sim \mathcal{N}_{p-q}(\mathbf{A}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11.2}\mathbf{A}_{22})$.

Part (c): The partition on $\mathbf{Z}$ means

$$\begin{aligned} \mathbf{A}_{11.2} =& \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \\ =& \mathbf{Z}_1^\top \mathbf{Z}_1 - \mathbf{Z}_1^\top \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top \mathbf{Z}_1 \\ =& \mathbf{Z}_1^\top\big(\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top\big)\mathbf{Z}_1 \end{aligned}$$

We condition on $\mathbf{Z}_2$, then matrix

$$\mathbf{Q} = \mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top \in \mathbb{R}^{n\times n}$$

is a constant projection matrix with rank

$$n - \operatorname{rank}(\mathbf{Z}_2) = n - (p-q).$$

Recall that

$$\mathbf{z}_{1\alpha} \mid \mathbf{z}_{2\alpha} \sim \mathcal{N}_q\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_{2\alpha}, \boldsymbol{\Sigma}_{11.2}\right) \qquad \Longrightarrow \qquad \mathbf{w}_\alpha = \mathbf{z}_{1\alpha} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_{2\alpha} \mid \mathbf{z}_{2\alpha} \sim \mathcal{N}_q\left(\mathbf{0}, \boldsymbol{\Sigma}_{11.2}\right).$$

This implies

$$\sum_{\alpha=1}^{n} \mathbf{w}_\alpha \mathbf{w}_\alpha^\top \mid \mathbf{Z}_2 \sim \mathcal{W}_q(\boldsymbol{\Sigma}_{11.2}, n) \qquad \Longrightarrow \qquad \mathbf{W}^\top \mathbf{W} \mid \mathbf{Z}_2 \sim \mathcal{W}_q(\boldsymbol{\Sigma}_{11.2}, n),$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_n^\top \end{bmatrix} = \mathbf{Z}_1 - \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{n\times q}.$$

and Lemma 7.2 means

$$\mathbf{W}^\top \mathbf{Q}\mathbf{W} \mid \mathbf{Z}_2 \sim \mathcal{W}_q(\boldsymbol{\Sigma}_{11.2}, n-p+q) \qquad \Longrightarrow \qquad \mathbf{W}^\top \mathbf{Q}\mathbf{W} \sim \mathcal{W}_q(\boldsymbol{\Sigma}_{11.2}, n-p+q).$$

We also have

$$\begin{aligned} \mathbf{Q}\mathbf{W} =& (\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top)(\mathbf{Z}_1 - \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \\ =& \mathbf{Z}_1 - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top \mathbf{Z}_1 - \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} + \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top \mathbf{Z}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ =& \mathbf{Z}_1 - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top \mathbf{Z}_1 \\ =& \big(\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top\big)\mathbf{Z}_1 = \mathbf{Q}\mathbf{Z}_1. \end{aligned}$$

Since $\mathbf{Q}$ is a projection matrix, we have

$$\mathbf{A}_{11.2} = \mathbf{Z}_1^\top \mathbf{Q}\mathbf{Z}_1 = \mathbf{Z}_1^\top \mathbf{Q}\mathbf{W} = \mathbf{W}^\top \mathbf{Q}\mathbf{W} \sim \mathcal{W}_q(\boldsymbol{\Sigma}_{11.2}, n-p+q),$$

This document is organized by Luo Luo.
Please do not distribute.

which does not depend on $\mathbf{A}_{22}$. Hence, the random matrix $\mathbf{A}_{11.2}$ is independent on $\mathbf{A}_{22}$.

Finally, we prove that $\mathbf{A}_{11.2}$ is independent on $\mathbf{A}_{21}$. We first show they are independent if we condition on $\mathbf{Z}_2$ (or condition on $\mathbf{A}_{22}$). Note that each $\mathbf{w}_\alpha$ has distribution $\mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}_{11.2})$ independently and

$$\mathbf{Q}\mathbf{Z}_2 = (\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2\mathbf{Z}_2^\top)^{-1}\mathbf{Z}_2^\top)\mathbf{Z}_2 = \mathbf{0}.$$

For given $\mathbf{Z}_2$, then the random matrices

$$\mathbf{Q}\mathbf{W} \qquad \text{and} \qquad \mathbf{Z}_2^\top \mathbf{W}$$

are independent (homework). This implies

$$\mathbf{A}_{11.2} = \mathbf{W}^\top\mathbf{Q}\mathbf{W} = (\mathbf{W}\mathbf{Q})^\top\mathbf{Q}\mathbf{W} \quad \text{and} \quad \mathbf{A}_{21} = \mathbf{Z}_2^\top\mathbf{Z}_1 = \mathbf{Z}_2^\top(\mathbf{W} + \mathbf{Z}_2\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}) = \mathbf{Z}_2^\top\mathbf{W} + \mathbf{Z}_2^\top\mathbf{Z}_2\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

are independent for given $\mathbf{Z}_2$. Similar to bivariate case, we have

$$\begin{aligned}
& f(\mathbf{A}_{11.2}, \mathbf{A}_{12}) \\
=& \int f(\mathbf{A}_{11.2}, \mathbf{A}_{21} \,|\, \mathbf{Z}_2) f(\mathbf{Z}_2)\, \mathrm{d}\mathbf{Z}_2 \\
=& \int f(\mathbf{A}_{11.2} \,|\, \mathbf{Z}_2) f(\mathbf{A}_{21} \,|\, \mathbf{Z}_2) f(\mathbf{Z}_2)\, \mathrm{d}\mathbf{Z}_2 \\
=& \int f(\mathbf{A}_{11.2}) f(\mathbf{A}_{21} \,|\, \mathbf{Z}_2) f(\mathbf{Z}_2)\, \mathrm{d}\mathbf{Z}_2 \\
=& f(\mathbf{A}_{11.2}) \int f(\mathbf{A}_{21} \,|\, \mathbf{Z}_2) f(\mathbf{Z}_2)\, \mathrm{d}\mathbf{Z}_2 \\
=& f(\mathbf{A}_{11.2}) f(\mathbf{A}_{21}).
\end{aligned}$$

□

Remark 7.3. *For $q > 1$, we can generalized the statement (b) to*

$$\mathrm{vec}(\mathbf{A}_{21}^\top) \,|\, \mathbf{A}_{22} \sim \mathcal{N}_{q(p-q)}\big(\mathrm{vec}(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{A}_{22}), \mathbf{A}_{22} \otimes \mathbf{\Sigma}_{11.2}\big),$$

where the notation $\otimes$ is the Kronecker product (see Wikipedia).

Lemma 7.6. *The function density of $\mathbf{A} \sim \mathcal{W}_p(\mathbf{I}_p, n)$ is*

$$w_p(\mathbf{A} \,|\, \mathbf{I}_p, n) = \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{A})\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

Proof. We prove this lemma by induction. For $p = 1$, the variable $\mathbf{A} = a$ is a scalar and has χ_n^2-distribution with density

$$w_1(a \,|\, 1, n) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} a^{\frac{n}{2}-1} \exp\left(-\frac{a}{2}\right).$$

We suppose the density of $\mathbf{B} \sim \mathcal{W}_{p-1}(\mathbf{I}_{p-1}, n)$ is

$$w_{p-1}(\mathbf{B} \,|\, \mathbf{I}_{p-1}, n) = \frac{(\det(\mathbf{B}))^{\frac{n-p}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{B})\right)}{2^{\frac{n(p-1)}{2}} \pi^{\frac{(p-1)(p-2)}{4}} \prod_{i=1}^{p-1} \Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

For $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$, we partition $\mathbf{A}$ into 1 and $p-1$ rows and columns as follows

$$\mathbf{A} = \begin{bmatrix} a_{11} & \mathbf{a}_{21}^\top \\ \mathbf{a}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

Applying Lemma 7.5 with $\mathbf{\Sigma} = \mathbf{I}_p$ and $q = 1$, we have

This document is organized by Luo Luo.
Please do not distribute.

1. $\mathbf{A}_{22} \sim \mathcal{W}_{p-1}(\mathbf{\Sigma}_{22}, n) \implies \mathbf{A}_{22} \sim \mathcal{W}_{p-1}(\mathbf{I}_{p-1}, n)$;
2. $\mathbf{a}_{21} \,|\, \mathbf{A}_{22} \sim \mathcal{N}_{p-1}(\mathbf{0}, \mathbf{A}_{22})$;
3. $a_{11.2} = a_{11} - \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21} \sim \chi^2_{n-p+1}$ is independent on $\mathbf{A}_{22}$ and $\mathbf{a}_{21}$.

Note that matrix $\mathbf{A}$ is symmetric, then the density of $\mathbf{A}$ can be written as

$$\begin{aligned} f(\mathbf{A}) =& f(a_{11}, \mathbf{a}_{21}, \mathbf{A}_{22}) \\ =& f(a_{11.2}, \mathbf{a}_{21}, \mathbf{A}_{22}) \\ =& f(a_{11.2}) f(\mathbf{a}_{21}, \mathbf{A}_{22}) \\ =& f(a_{11.2}) f(\mathbf{a}_{21} \,|\, \mathbf{A}_{22}) f(\mathbf{A}_{22}), \end{aligned}$$

where the second line is because of the Jacobian of transform from $(a_{11}, \mathbf{a}_{21}, \mathbf{A}_{22})$ to $(a_{11.2}, \mathbf{a}_{21}, \mathbf{A}_{22})$ is an upper triangle matrix of the form (consider $a_{11.2} = a_{11} - \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21}$)

$$\begin{bmatrix} 1 & \times & \dots & \times \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Applying the induction base, we have

$$\begin{aligned} w_p(\mathbf{A} \,|\, \mathbf{I}_p, n) =& \frac{1}{2^{\frac{n-p+1}{2}} \Gamma\left(\frac{n-p+1}{2}\right)} a_{11.2}^{\frac{n-p+1}{2}-1} \exp\left(-\frac{a_{11.2}}{2}\right) \\ & \cdot \frac{1}{(2\pi)^{\frac{p-1}{2}} (\det(\mathbf{A}_{22}))^{\frac{1}{2}}} \exp\left(-\frac{-\mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21}}{2}\right) \\ & \cdot \frac{1}{2^{\frac{n(p-1)}{2}} \pi^{\frac{(p-1)(p-2)}{4}} \prod_{i=1}^{p-1} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot (\det(\mathbf{A}_{22}))^{\frac{n-p}{2}} \exp\left(-\frac{1}{2} \mathrm{tr}\,(\mathbf{A}_{22})\right). \end{aligned}$$

The terms related to exponential function hold

$$\begin{aligned} & a_{11.2} + \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21} + \mathrm{tr}(\mathbf{A}_{22}) \\ =& \mathrm{tr}(a_{11.2} + \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21} + \mathbf{A}_{22}) \\ =& \mathrm{tr}(a_{11} - \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21} + \mathbf{a}_{21}^\top \mathbf{A}_{22}^{-1} \mathbf{a}_{21} + \mathbf{A}_{22}) \\ =& \mathrm{tr}(\mathbf{A}). \end{aligned}$$

The terms related to determinant function hold (the last step use the property of Schur complement)

$$\begin{aligned} & a_{11.2}^{\frac{n-p+1}{2}-1} \cdot \frac{1}{(\det(\mathbf{A}_{22}))^{\frac{1}{2}}} \cdot (\det(\mathbf{A}_{22}))^{\frac{n-p}{2}} \\ =& a_{11.2}^{\frac{n-p-1}{2}} \cdot (\det(\mathbf{A}_{22}))^{\frac{n-p-1}{2}} \\ =& (\det(a_{11.2}) \det(\mathbf{A}_{22}))^{\frac{n-p-1}{2}} \\ =& ((\det(\mathbf{A}))^{\frac{n-p-1}{2}}. \end{aligned}$$

The terms related to the constants hold

$$\begin{aligned} w_p(\mathbf{A} \,|\, \mathbf{I}_p, n) =& \frac{1}{2^{\frac{n-p+1}{2}} \Gamma\left(\frac{n-p+1}{2}\right)} \cdot \frac{1}{(2\pi)^{\frac{p-1}{2}}} \cdot \frac{1}{2^{\frac{n(p-1)}{2}} \pi^{\frac{(p-1)(p-2)}{4}} \prod_{i=1}^{p-1} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \\ =& \frac{1}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}. \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Combing all above results, we achieve

$$w_p(\mathbf{A}\,|\,\mathbf{I}_p, n) = \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{A}\right)\right)}{2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

□

Theorem 7.1. *The density function of* $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ *is*

$$w_p(\mathbf{A}\,|\,\mathbf{\Sigma}, n) = \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}\left(\det(\mathbf{\Sigma})\right)^{\frac{n}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

for positive definite $\mathbf{A}$*.*

Proof. Let $\mathbf{B} = \mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}$, then Lemma 7.1 indicates $\mathbf{B} \sim \mathcal{W}_p(\mathbf{I}_p, n)$. Lemma 7.4 implies the Jacobian of transform from $\mathbf{A}$ to $\mathbf{B}$ has determinant

$$(\det(\mathbf{\Sigma}^{-1/2}))^{p+1} = (\det(\mathbf{\Sigma}))^{-\frac{p+1}{2}}.$$

Hence, we have

$$\begin{aligned}
& w_p(\mathbf{A}\,|\,\mathbf{\Sigma}, n) \\
=& w_p(\mathbf{B}\,|\,\mathbf{I}_p, n) \cdot (\det(\mathbf{\Sigma}))^{-\frac{p+1}{2}} \\
=& w_p(\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\,|\,\mathbf{I}_p, n) \cdot (\det(\mathbf{\Sigma}))^{-\frac{p+1}{2}} \\
=& \frac{\left(\det(\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2})\right)^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\right)\right)}{2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot (\det(\mathbf{\Sigma}))^{-\frac{p+1}{2}} \\
=& \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}\left(\det(\mathbf{\Sigma})\right)^{\frac{n}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.
\end{aligned}$$

□

Remark 7.4. *We define the multivariate gamma function as*

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\Big(t - \frac{1}{2}(i-1)\Big).$$

Then the density function of Wishart distribution can be written as

$$w_p(\mathbf{A}\,|\,\mathbf{\Sigma}, n) = \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}\Gamma_p\left(\frac{n}{2}\right)\left(\det(\mathbf{\Sigma})\right)^{\frac{n}{2}}}.$$

Corollary 7.1. *Let* $\mathbf{x}_1, \dots, \mathbf{x}_N$ *be independently distributed, each according to* $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$*, where* $N > p$*. Then*

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \sim \mathcal{W}_p\left(\frac{1}{N-1}\mathbf{\Sigma}, N-1\right).$$

Proof. The matrix $\mathbf{S}$ has the distribution of

$$\mathbf{S} = \sum_{\alpha=1}^{n}\frac{\mathbf{z}_\alpha}{\sqrt{n}}\left(\frac{\mathbf{z}_\alpha}{\sqrt{n}}\right)^\top,$$

where each $\mathbf{z}_1/\sqrt{n}, \dots, \mathbf{z}_n/\sqrt{n}$ are independently distributed, each according to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}/n)$, and $n = N-1$. We prove this result by the definition of Wishart distribution. □

This document is organized by Luo Luo.
Please do not distribute.

T^2-statistic and F-distribution: We consider the distribution of T^2-statistic

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

where

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha, \qquad \mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$$

and $\mathbf{x}_1, \dots, \mathbf{x}_N$, are independently distributed to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Lemma 7.7. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{I}_p, n)$ with $n \geq p$, then we have*

$$\mathbf{t}^\top \mathbf{A}\mathbf{t} \sim \chi^2_n \qquad \textit{and} \qquad \frac{1}{\mathbf{t}^\top \mathbf{A}^{-1}\mathbf{t}} \sim \chi^2_{n-p+1}$$

for any constant vector $\mathbf{t} \in \mathbb{R}^p$ with $\|\mathbf{t}\|_2 = 1$.

This document is organized by Luo Luo.
Please do not distribute.

Proof. The result of $\mathbf{t}^\top \mathbf{A}\mathbf{t} \sim \chi^2_n$ can be achieved by Lemma 7.1. We focus on the second result. For given $\mathbf{B} \in \mathbb{R}^{p\times p}$, we partition $\mathbf{B}$ and $\mathbf{B}^{-1}$ into q and $p-q$ rows and column as

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{B}^{-1} = \begin{bmatrix} \mathbf{B}_{11.2}^{-1} & \times \\ \times & \times \end{bmatrix}, \qquad \text{where } \mathbf{B}_{11.2} = \mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}.$$

In the view of Lemma 7.5 (c) with $q = 1$. We desire to construct some constant $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{B} \sim \mathcal{W}_p(\mathbf{I}_p, n)$ such that

$$\mathbf{t}^\top \mathbf{A}^{-1}\mathbf{t} = \mathbf{u}^\top \mathbf{B}^{-1}\mathbf{u} = \begin{bmatrix} 1 & \mathbf{0}^\top \end{bmatrix} \begin{bmatrix} b_{11.2}^{-1} & \times \\ \times & \times \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} = b_{11.2}^{-1} \qquad \text{and} \qquad b_{11.2} = \chi^2_{n-p+1}.$$

Specifically, we let

$$\mathbf{Q} = \begin{bmatrix} \mathbf{t}^\top \\ \mathbf{V}^\top \end{bmatrix} \in \mathbb{R}^{p\times p} \qquad \text{and} \qquad \mathbf{B} = \mathbf{Q}\mathbf{A}\mathbf{Q}^\top \quad \text{with} \quad \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_p.$$

Then we have

$$\mathbf{Q}\mathbf{t} = \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \qquad \text{and} \qquad \mathbf{t}^\top \mathbf{A}^{-1}\mathbf{t} = \mathbf{t}^\top\mathbf{Q}^\top(\mathbf{Q}\mathbf{A}\mathbf{Q}^\top)^{-1}\mathbf{Q}\mathbf{t} = \begin{bmatrix} 1 & \mathbf{0}^\top \end{bmatrix} \mathbf{B}^{-1} \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} = b_{11.2}^{-1}.$$

Since we construct $\mathbf{Q} \in \mathbb{R}^{p\times p}$ as an orthogonal matrix, Lemma 7.1 indicates

$$\mathbf{B} = \mathbf{Q}\mathbf{A}\mathbf{Q}^\top \sim \mathcal{W}_p(\mathbf{I}_p, n).$$

Lemma 7.5 (c) with $q = 1$, we have

$$b_{11.2} \sim \mathcal{W}_1(1, n-p+1) \qquad \Longrightarrow \qquad \frac{1}{\mathbf{t}^\top \mathbf{A}^{-1}\mathbf{t}} = b_{11.2}^{-1} \sim \chi^2_{n-p+1}.$$

□

Remark 7.5. *Suppose $\mathbf{B}_{12} = \mathbf{B}_{21}^\top$ and*

$$\begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-q} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{12}^\top & \mathbf{B}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Y}^\top & \mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{11}\mathbf{X} + \mathbf{B}_{12}\mathbf{Y}^\top & \mathbf{B}_{11}\mathbf{Y} + \mathbf{B}_{12}\mathbf{Z} \\ \mathbf{B}_{12}^\top\mathbf{X} + \mathbf{B}_{22}\mathbf{Y}^\top & \mathbf{B}_{12}^\top\mathbf{Y} + \mathbf{B}_{22}\mathbf{Z}. \end{bmatrix}$$

The equations

$$\begin{cases} \mathbf{B}_{11}\mathbf{X} + \mathbf{B}_{12}\mathbf{Y}^\top = \mathbf{I}_q \\ \mathbf{B}_{12}^\top\mathbf{X} + \mathbf{B}_{22}\mathbf{Y}^\top = \mathbf{0} \end{cases} \Longrightarrow \begin{cases} \mathbf{B}_{11}\mathbf{X} + \mathbf{B}_{12}\mathbf{Y}^\top = \mathbf{I}_q \\ \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{12}^\top\mathbf{X} + \mathbf{B}_{12}\mathbf{Y}^\top = \mathbf{0} \end{cases} \Longrightarrow \mathbf{X} = (\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{12}^\top)^{-1}.$$

Theorem 7.2. *Let $\mathbf{A} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$ and $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ be independent with $n \geq p$, then*

$$\frac{n-p+1}{p} \cdot \mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y} \sim F_{p,n-p+1}.$$

Proof. Let $\mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{-1/2}$, then Lemma 7.1 and Theorem 5.3 indicate

$$\mathbf{B} \sim \mathcal{W}_p(\mathbf{I}_p, n) \qquad \text{and} \qquad \mathbf{w}^\top\mathbf{w} = \mathbf{y}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} \sim \chi^2_p.$$

For given $\mathbf{w} \in \mathbb{R}^p$, applying Lemma 7.7 in the view of $\mathbf{t} = \mathbf{w}/\left\|\mathbf{w}\right\|_2$ achieves

$$\frac{1}{(\mathbf{w}/\left\|\mathbf{w}\right\|_2)^\top \mathbf{B}^{-1}(\mathbf{w}/\left\|\mathbf{w}\right\|_2)} = \frac{\mathbf{w}^\top\mathbf{w}}{\mathbf{w}^\top\mathbf{B}^{-1}\mathbf{w}} \,\Big|\, \mathbf{w} \sim \chi^2_{n-p+1},$$

which does not depend on $\mathbf{w}$. Hence, we have

$$\frac{\mathbf{w}^\top\mathbf{w}}{\mathbf{w}^\top\mathbf{B}^{-1}\mathbf{w}} \sim \chi^2_{n-p+1},$$

which is independent on $\mathbf{w}$ (also independent on $\mathbf{w}^\top\mathbf{w}$). Hence, the definition of F-distribution means

$$\frac{\mathbf{w}^\top\mathbf{w}/p}{(\mathbf{w}^\top\mathbf{w}/(\mathbf{w}^\top\mathbf{B}^{-1}\mathbf{w}))/(n-p+1)} \sim F_{p,n-p+1}.$$

We have finished the proof because of the fact

$$\frac{\mathbf{w}^\top\mathbf{w}/p}{(\mathbf{w}^\top\mathbf{w}/(\mathbf{w}^\top\mathbf{B}^{-1}\mathbf{w}))/(n-p+1)} = \frac{n-p+1}{p} \cdot \mathbf{w}^\top\mathbf{B}^{-1}\mathbf{w} = \frac{n-p+1}{p} \cdot \mathbf{y}\mathbf{A}^{-1}\mathbf{y}.$$

□

Corollary 7.2. *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be independently distributed to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $N > p$. Define*

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}), \quad \textit{where} \quad \mathbf{x} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha \quad \textit{and} \quad \mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top,$$

then we have

$$\frac{N-p}{(N-1)p} \cdot T^2 \sim F_{p,n-p+1}$$

Proof. We can write

$$\begin{aligned} T^2 =& N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\ =& N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \left(\frac{1}{N-1} \cdot \mathbf{A}\right)^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \\ =& (N-1)\big(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})\big)^\top \mathbf{A}^{-1}\big(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})\big) \end{aligned}$$

We let $n = N - 1$ and recall that that

$$\mathbf{A} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n) \qquad \text{and} \qquad \mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \bar{\mathbf{x}} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma}\right) \implies \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Since Theorem 4.6 says $\mathbf{A}$ and $\bar{\mathbf{x}}$ are independent, we can apply Theorem 7.2 to achieve

$$\frac{n-p+1}{p} \cdot \frac{T^2}{N-1} \sim F_{p,n-p+1} \qquad \implies \qquad \frac{N-p}{(N-1)p} \cdot T^2 \sim F_{p,n-p+1}.$$

□

This document is organized by Luo Luo.
Please do not distribute.

Differential and Jacobian We define $\mathbf{f}:\mathbb{R}^p\to\mathbb{R}^q$ such that

$$\mathbf{f}(\mathbf{x})=\begin{bmatrix}f_1(\mathbf{x})\\ \vdots\\ f_q(\mathbf{x})\end{bmatrix}$$

for given $\mathbf{x}\in\mathbb{R}^p$. Then we have

$$\mathrm{d}\mathbf{f}(\mathbf{x})=\begin{bmatrix}\mathrm{d}f_1(\mathbf{x})\\ \vdots\\ \mathrm{d}f_q(\mathbf{x})\end{bmatrix}=\begin{bmatrix}\displaystyle\sum_{k=1}^{p}\frac{\partial f_1(\mathbf{x})}{\partial x_k}\,\mathrm{d}x_k\\ \vdots\\ \displaystyle\sum_{k=1}^{p}\frac{\partial f_q(\mathbf{x})}{\partial x_k}\,\mathrm{d}x_k\end{bmatrix}=\begin{bmatrix}\dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_p}\\ \vdots & \ddots & \ddots\\ \dfrac{\partial f_q(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_q(\mathbf{x})}{\partial x_p}\end{bmatrix}\begin{bmatrix}\mathrm{d}x_1\\ \vdots\\ \mathrm{d}x_p\end{bmatrix}=\mathbf{J}(\mathbf{f}(\mathbf{x}))\,\mathrm{d}\mathbf{x},$$

where $\mathbf{J}(\mathbf{f}(\mathbf{x}))\in\mathbb{R}^{q\times p}$ is the Jacobian from $\mathbf{x}$ to $\mathbf{f}(\mathbf{x})$, also the Jacobian from $\mathrm{d}\mathbf{x}$ to $\mathrm{d}\mathbf{f}(\mathbf{x})$. Let $\mathbf{y}=\mathbf{f}(\mathbf{x})$, then $\mathbf{J}(\mathbf{f}(\mathbf{x}))$ is the Jacobian from $\mathbf{x}$ to $\mathbf{y}$, also the Jacobian from $\mathrm{d}\mathbf{x}$ to $\mathrm{d}\mathbf{y}$.

The differential of matrix variates $\mathbf{X}\in\mathbb{R}^{p\times q}$ and $\mathbf{Y}\in\mathbb{R}^{q\times r}$ hold that

$$\mathrm{d}(\mathbf{X}\mathbf{Y})=(\mathrm{d}\mathbf{X})\mathbf{Y}+\mathbf{X}\mathrm{d}\mathbf{Y},$$

since

$$\begin{aligned}&(\mathrm{d}(\mathbf{X}\mathbf{Y}))_{ij}=\mathrm{d}(\mathbf{X}\mathbf{Y})_{ij}\\ =&\mathrm{d}\sum_k x_{ik}y_{kj}=\sum_k\mathrm{d}(x_{ik}y_{kj})\\ =&\sum_k\big(x_{ik}\mathrm{d}y_{kj}+(\mathrm{d}x_{ik})y_{kj}\big)\\ =&\big(\mathbf{X}\mathrm{d}\mathbf{Y}+(\mathrm{d}\mathbf{X})\mathbf{Y}\big)_{ij}.\end{aligned}$$

Let $\mathbf{F}:\mathbb{R}^{p\times q}\to\mathbb{R}^{m\times n}$ such that

$$\mathbf{Y}=\mathbf{F}(\mathbf{X}),$$

then we constructed the Jacobian from $\mathbf{X}\in\mathbb{R}^{p\times q}$ to $\mathbf{Y}\in\mathbb{R}^{m\times n}$ by

$$\mathrm{d}(\mathrm{vec}(\mathbf{Y}))=\mathbf{J}(\mathbf{f}(\mathrm{vec}(\mathbf{X})))\,\mathrm{d}(\mathrm{vec}(\mathbf{X})),$$

where $\mathbf{f}:\mathbb{R}^{pq}\to\mathbb{R}^{mn}$ and $\mathbf{J}(\mathbf{f}(\mathrm{vec}(\mathbf{X})))\in\mathbb{R}^{mn\times pq}$.

Lemma 7.8. *Let $\mathbf{X}\in\mathbb{R}^{p\times p}$ be a symmetric non-singular matrix. Then Jacobian of transform $\mathbf{Y}=\mathbf{X}^{-1}$ has the determinant*

$$(-1)^{p(p+1)/2}(\det(\mathbf{X}))^{-(p+1)}.$$

Proof. Taking total differential on $\mathbf{X}\mathbf{Y}=\mathbf{I}$, we achieve

$$(\mathrm{d}\mathbf{X})\mathbf{Y}+\mathbf{X}\mathrm{d}\mathbf{Y}=\mathbf{0}\quad\Longrightarrow\quad-\mathrm{d}\mathbf{Y}=\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1}.$$

Then Lemma 7.4 means the Jacobian from $\mathrm{d}\mathbf{X}$ to $-\mathrm{d}\mathbf{Y}$ (also the Jacobian from $\mathbf{X}$ to $-\mathbf{Y}$) is

$$(\det(\mathbf{X}^{-1}))^{p+1}=(\det(\mathbf{X}))^{-(p+1)}.$$

Since $\mathbf{Y}$ is symmetric, the Jacobian from $-\mathbf{Y}$ to $\mathbf{Y}$ is $(-1)^{p(p+1)/2}$. Then the Jacobian from $\mathbf{X}$ to $\mathbf{Y}$ is

$$(-1)^{p(p+1)/2}(\det(\mathbf{X}))^{-(p+1)}.$$

□

This document is organized by Luo Luo.
Please do not distribute.

Theorem 7.3. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$, then the random matrix $\mathbf{B} = \mathbf{A}^{-1}$ has inverted (inverse) Wishart distribution, written as*

$$\mathbf{B} \sim \mathcal{W}_p^{-1}(\mathbf{\Psi}, n) \qquad \textit{or} \qquad \mathbf{B} \sim \mathcal{IW}_p(\mathbf{\Psi}, n).$$

The density of $\mathbf{B}$ is

$$w_p^{-1}(\mathbf{B} \,|\, \mathbf{\Psi}, n) = \frac{(\det(\mathbf{\Psi}))^{\frac{n}{2}} (\det(\mathbf{B}))^{-\frac{n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{B}^{-1}\right)\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

Proof. The Jacobian of the transform from $\mathbf{B}$ to $\mathbf{A} = \mathbf{B}^{-1}$ has determinant

$$(-1)^{p(p+1)/2} (\det(\mathbf{B}))^{-(p+1)}.$$

Recall that the density function of $\mathbf{A}$ is

$$w_p(\mathbf{A} \,|\, \mathbf{\Sigma}, n) = \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)},$$

then the density function of $\mathbf{B}$ is

$$\begin{aligned} &\frac{\left(\det(\mathbf{B}^{-1})\right)^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^{-1}\right)\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot (\det(\mathbf{B}))^{-(p+1)} \\ =& \frac{(\det(\mathbf{\Psi}))^{\frac{n}{2}} (\det(\mathbf{B}))^{-\frac{n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{B}^{-1}\right)\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}. \end{aligned}$$

□

Remark 7.6. *Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ and $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$. We partition $\mathbf{x}$, $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ into q and $p-q$ rows (column) as*

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \qquad \textit{and} \qquad \mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_{11} & \mathbf{\Psi}_{12} \\ \mathbf{\Psi}_{21} & \mathbf{\Psi}_{22} \end{bmatrix}.$$

Recall that $\mathrm{Cov}[\mathbf{x}^{(1)} \,|\, \mathbf{x}^{(2)}] = \mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ *and Remark 7.5 says*

$$\mathbf{\Psi}_{11} = \mathbf{\Sigma}_{11.2}^{-1} = (\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21})^{-1}.$$

The entries of $\mathbf{x}^{(1)}$ are conditional mutually independent for given $\mathbf{x}^{(2)}$ is equivalent to $\mathbf{\Sigma}_{11.2}$ is diagonal. That is, the sub-matrix $\mathbf{\Psi}_{11} = \mathbf{\Sigma}_{11.2}^{-1}$ is also diagonal. In other words, the variables x_i and x_j are conditionally independent for given all other coordinates of $\mathbf{x}$ is equivalent to $\psi_{ij} = 0$.

Theorem 7.4. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{\Sigma}$ has the a prior distribution $\mathcal{W}_p^{-1}(\mathbf{\Psi}, m)$, then the conditional distribution of $\mathbf{\Sigma}$ given $\mathbf{A}$ is the inverted Wishart distribution $\mathcal{W}_p^{-1}(\mathbf{A} + \mathbf{\Psi}, n+m)$.*

Proof. We target to calculate

$$f(\mathbf{\Sigma} \,|\, \mathbf{A}) = \frac{f(\mathbf{A}, \mathbf{\Sigma})}{f(\mathbf{A})} = \frac{f(\mathbf{A} \,|\, \mathbf{\Sigma}) f(\mathbf{\Sigma})}{f(\mathbf{A})} \propto f(\mathbf{A} \,|\, \mathbf{\Sigma}) f(\mathbf{\Sigma}).$$

We can calculate the value of

$$f(\mathbf{A}) = \int f(\mathbf{A} \,|\, \mathbf{\Sigma}) f(\mathbf{\Sigma}) \, \mathrm{d}\mathbf{\Sigma}$$

This document is organized by Luo Luo.
Please do not distribute.

by leveraging the density of inverted Wishart distribution, but it is unnecessary since we only require consider the terms related to $\mathbf{\Sigma}$. Following Bayes rule, we have

$$\begin{aligned}
f(\mathbf{\Sigma}\,|\,\mathbf{A}) &\propto f(\mathbf{A}\,|\,\mathbf{\Sigma})f(\mathbf{\Sigma}) \\
&= \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}(\det(\mathbf{\Sigma}))^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}}(\det(\mathbf{\Sigma}))^{-\frac{m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)} \\
&\propto (\det(\mathbf{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\mathbf{\Psi})\mathbf{\Sigma}^{-1}\right)\right) \\
&\propto w^{-1}(\mathbf{\Sigma}\mid\mathbf{A}+\mathbf{\Psi}, n+m).
\end{aligned}$$

Recall that $f(\mathbf{\Sigma}\,|\,\mathbf{A})$ is a density, then the conditional distribution of $\mathbf{\Sigma}$ for given $\mathbf{A}$ must follow

$$\mathcal{W}^{-1}(\mathbf{A}+\mathbf{\Psi}, n+m).$$

□

Remark 7.7. *We can also calculate the posterior density brutally. Consider that*

$$\begin{aligned}
&f(\mathbf{A},\mathbf{\Sigma}) = f(\mathbf{A}\,|\,\mathbf{\Sigma})f(\mathbf{\Sigma}) \\
=&\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}(\det(\mathbf{\Sigma}))^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}}(\det(\mathbf{\Sigma}))^{-\frac{m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)} \\
=&\frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}(\det(\mathbf{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\mathbf{\Psi})\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)}.
\end{aligned} \tag{27}$$

The marginal density $f(\mathbf{A})$ is the integral of (27) over the set of $\mathbf{\Sigma}$ positive definite.

Note that the brutal method does not require calculating the integral brutally. By leveraging the density function of inverted Wishart distribution, we have

$$\begin{aligned}
1 =& \int w^{-1}(\mathbf{\Sigma}\mid\mathbf{A}+\mathbf{\Psi}, n+m)\,\mathrm{d}\mathbf{\Sigma} \\
=&\frac{(\det(\mathbf{A}+\mathbf{\Psi}))^{\frac{n+m}{2}}(\det(\mathbf{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\mathbf{\Psi})\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n+m}{2}\right)},
\end{aligned}$$

which leads to

$$\begin{aligned}
f(\mathbf{A}) =& \int f(\mathbf{A},\mathbf{\Sigma})\,\mathrm{d}\mathbf{\Sigma} \\
=&\frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)}\int\frac{(\det(\mathbf{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\mathbf{\Psi})\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}}\mathrm{d}\mathbf{\Sigma} \\
=&\frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)}\cdot\Gamma_p\left(\frac{n+m}{2}\right)(\det(\mathbf{A}+\mathbf{\Psi}))^{-\frac{n+m}{2}}.
\end{aligned}$$

Then we have

$$\begin{aligned}
f(\mathbf{\Sigma}\mid\mathbf{A}) =& \frac{f(\mathbf{\Sigma},\mathbf{A})}{f(\mathbf{A})} \\
=&\frac{(\det(\mathbf{A}+\mathbf{\Psi}))^{\frac{n+m}{2}}(\det(\mathbf{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\mathbf{\Psi})\mathbf{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n+m}{2}\right)} \\
=&w^{-1}(\mathbf{\Sigma}\mid\mathbf{A}+\mathbf{\Psi}, n+m).
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Theorem 7.5 (Anderson [1, Theorem 7.7.3])**.** *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be observations from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have prior densities*

$$n\left(\boldsymbol{\mu} \;\middle|\; \boldsymbol{\nu}, \frac{\boldsymbol{\Sigma}}{K}\right) \qquad \textit{and} \qquad w^{-1}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}, m)$$

respectively, where $n = N - 1$. Then the posterior density of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given

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

is

$$n\left(\boldsymbol{\mu} \;\middle|\; \frac{N\bar{\mathbf{x}} + K\boldsymbol{\nu}}{N+K}, \frac{\boldsymbol{\Sigma}}{N+K}\right) \cdot w^{-1}\left(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi} + n\mathbf{S} + \frac{NK(\bar{\mathbf{x}} - \boldsymbol{\nu})(\bar{\mathbf{x}} - \boldsymbol{\nu})^\top}{N+K}, N + m\right).$$

Characteristic Function of Wishart Distribution We first provide a fundamental proof for the density of χ^2-distribution and a lemma for matrix transform.

Lemma 7.9. *The characteristic function of χ^2-distribution with the degree of freedom n is*

$$\phi(t) = (1 - 2\mathrm{i}t)^{-\frac{n}{2}}.$$

Proof. Let $x \sim \chi^2_n$, then its density is

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right).$$

We have (using the density of χ^2-distribution with the degree of freedom $2k + n$)

$$\begin{aligned}
\phi(t) =& \mathbb{E}\left[\exp(\mathrm{i}tx)\right] \\
=& \int_0^{+\infty} \exp(\mathrm{i}tx) \cdot \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} \left(\sum_{k=0}^{\infty} \frac{(\mathrm{i}tx)^k}{k!}\right) x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \int_0^{+\infty} x^{\frac{n+2k}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \cdot 2^{\frac{n+2k}{2}}\Gamma\left(\frac{n+2k}{2}\right) \int_0^{+\infty} \frac{1}{2^{\frac{n+2k}{2}}\Gamma\left(\frac{n+2k}{2}\right)} x^{\frac{n+2k}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \cdot 2^{\frac{n+2k}{2}}\Gamma\left(\frac{n+2k}{2}\right) \\
=& 1 + \sum_{k=1}^{\infty} \frac{(2\mathrm{i}t)^k}{k!} \cdot \prod_{j=0}^{k-1}\left(j + \frac{n}{2}\right) = (1 - 2\mathrm{i}t)^{-\frac{n}{2}}.
\end{aligned}$$

For the last step, we consider Taylor expansion of $f(x) = (1 - x)^{-\frac{n}{2}}$ at 0, that is

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)x^k}{k!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} \prod_{j=0}^{k-1}\left(j + \frac{n}{2}\right).$$

We take $x = 2\mathrm{i}t$ to achieve the desired result. □

This document is organized by Luo Luo.
Please do not distribute.

Lemma 7.10. *Given positive definite matrix* $\mathbf{X}$ *and positive semidefinite matrix* $\mathbf{Y}$*, there exists a non-singular matrix* $\mathbf{F}$ *such that* $\mathbf{F}^\top\mathbf{Y}\mathbf{F} = \mathbf{D}$ *and* $\mathbf{F}^\top\mathbf{X}\mathbf{F} = \mathbf{I}$*, where* $\mathbf{D}$ *is diagonal.*

Proof. Let the spectral decomposition of $\mathbf{X}$ be $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{U}_\mathbf{X}^\top$ and let $\mathbf{E} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-\frac{1}{2}}$, then we have

$$\mathbf{E}^\top\mathbf{X}\mathbf{E} = \mathbf{\Sigma}_\mathbf{X}^{-\frac{1}{2}}\mathbf{U}_\mathbf{X}^\top(\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{U}_\mathbf{X}^\top)\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-\frac{1}{2}} = \mathbf{I}.$$

Let $\mathbf{Z} = \mathbf{E}^\top\mathbf{Y}\mathbf{E}$ with the spectral decomposition $\mathbf{Z} = \mathbf{U}_\mathbf{Z}\mathbf{\Sigma}_\mathbf{Z}\mathbf{U}_\mathbf{Z}^\top$, then we have

$$\mathbf{\Sigma}_\mathbf{Z} = \mathbf{U}_\mathbf{Z}^\top\mathbf{Z}\mathbf{U}_\mathbf{Z} = \mathbf{U}_\mathbf{Z}^\top\mathbf{E}^\top\mathbf{Y}\mathbf{E}\mathbf{U}_\mathbf{Z}$$

and

$$\mathbf{U}_\mathbf{Z}^\top\mathbf{E}^\top\mathbf{X}\mathbf{E}\mathbf{U}_\mathbf{Z} = \mathbf{U}_\mathbf{Z}^\top(\mathbf{\Sigma}_\mathbf{X}^{-\frac{1}{2}}\mathbf{U}_\mathbf{X}^\top)(\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{U}_\mathbf{X}^\top)(\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-\frac{1}{2}})\mathbf{U}_\mathbf{Z} = \mathbf{I}$$

Letting $\mathbf{F} = \mathbf{E}\mathbf{U}_\mathbf{Z}$ and $\mathbf{D} = \mathbf{\Sigma}_\mathbf{Z}$ proves this lemma. □

Theorem 7.6. *If* $\mathbf{z}_1, \dots, \mathbf{z}_n$ *are independent, each with distribution* $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$*, then the characteristic function of* $a_{11}, \dots, a_{pp}$*,* $2a_{12}, \dots, 2a_{p-1,p}$*, where* a_{ij} *is the* (i,j)*-th element of*

$$\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha\mathbf{z}_\alpha^\top$$

is given by $\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right] = \left(\det\left(\mathbf{I} - 2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{n}{2}}$*, where* $\mathbf{\Theta} \in \mathbb{R}^{p\times p}$ *is symmetric.*

Proof. The characteristic function of $a_{11}, \dots, a_{pp}$, $2a_{12}, \dots, 2a_{p-1,p}$ is

$$\begin{aligned}
&\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathrm{tr}\left(\sum_{\alpha=1}^{n}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\mathbf{\Theta}\right)\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathrm{tr}\left(\sum_{\alpha=1}^{n}\mathbf{z}_\alpha^\top\mathbf{\Theta}\mathbf{z}_\alpha\right)\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\sum_{\alpha=1}^{n}\mathbf{z}_\alpha^\top\mathbf{\Theta}\mathbf{z}_\alpha\right)\right]\\
=&\prod_{\alpha=1}^{n}\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}_\alpha^\top\mathbf{\Theta}\mathbf{z}_\alpha\right)\right]\\
=&\left(\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]\right)^n,
\end{aligned}$$

where $\mathbf{z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Applying Lemma 7.10 with $\mathbf{Y} = \mathbf{\Theta}$ and $\mathbf{X} = \mathbf{\Sigma}^{-1}$ means there exists non-singular matrix $\mathbf{F}$ such that

$$\mathbf{F}^\top\mathbf{\Theta}\mathbf{F} = \mathbf{D} \qquad \text{and} \qquad \mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F} = \mathbf{I}$$

where $\mathbf{D} \in \mathbb{R}^{p\times p}$ is diagonal. If we set $\mathbf{z} = \mathbf{F}\mathbf{y}$, then we have

$$\mathbf{y} = \mathbf{F}^{-1}\mathbf{z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$$

and

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{y}^\top\mathbf{F}^\top\mathbf{\Theta}\mathbf{F}\mathbf{y}\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{y}^\top\mathbf{D}\mathbf{y}\right)\right]
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$
\begin{aligned}
&=\mathbb{E}\left[\prod_{j=1}^{p}\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right]\\
&=\prod_{j=1}^{p}\mathbb{E}\left[\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right].
\end{aligned}
$$

Note that the term of $\mathbb{E}\left[\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right]$ is the characteristic function of the χ^2-distribution with one degree of freedom, namely $(1-2\mathrm{i}d_{jj})^{-\frac{1}{2}}$. Thus, we have

$$
\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]=\prod_{j=1}^{p}(1-2\mathrm{i}d_{jj})^{-\frac{1}{2}}=(\det(\mathbf{I}-2\mathrm{i}\mathbf{D}))^{-\frac{1}{2}}.
$$

We also have

$$
\begin{aligned}
&\det(\mathbf{I}-2\mathrm{i}\mathbf{D})\\
=&\det\left(\mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F}-2\mathrm{i}\mathbf{F}^\top\mathbf{\Theta}\mathbf{F}\right)\\
=&\det\left(\mathbf{F}^\top\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)\mathbf{F}\right)\\
=&(\det(\mathbf{F}))^2\det\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)
\end{aligned}
$$

and $\mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F}=\mathbf{I}$ means $\det(\mathbf{F})=(\det(\mathbf{\Sigma}))^{\frac{1}{2}}$. Combing the above results, we obtain

$$
\det(\mathbf{I}-2\mathrm{i}\mathbf{D})=\det(\mathbf{\Sigma})\det\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)=\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)
$$

and

$$
\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right]=\left(\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]\right)^n=\left(\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{n}{2}}.
$$

□

Remark 7.8. *If $\mathbf{\Phi}\in\mathbb{R}^{p\times p}$ is not symmetric, we have*

$$
\mathrm{tr}(\mathbf{A}\mathbf{\Phi})=\mathrm{tr}(\mathbf{\Phi}^\top\mathbf{A}^\top)=\mathrm{tr}(\mathbf{A}^\top\mathbf{\Phi}^\top)=\mathrm{tr}(\mathbf{A}\mathbf{\Phi}^\top)
$$

since $\mathbf{A}\in\mathbb{R}^{p\times p}$ is symmetric. Hence, we have

$$
\mathrm{tr}(\mathbf{A}\mathbf{\Phi})=\frac{\mathrm{tr}(\mathbf{A}\mathbf{\Phi})+\mathrm{tr}(\mathbf{A}\mathbf{\Phi}^\top)}{2}=\mathrm{tr}\left(\mathbf{A}\cdot\frac{\mathbf{\Phi}+\mathbf{\Phi}^\top}{2}\right).
$$

Applying Theorem 7.6 with $\mathbf{\Theta}=(\mathbf{\Phi}+\mathbf{\Phi}^\top)/2$, we have

$$
\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Phi}))\right]=\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathrm{tr}\left(\mathbf{A}\cdot\frac{\mathbf{\Phi}+\mathbf{\Phi}^\top}{2}\right)\right)\right]=\left(\det\left(\mathbf{I}-\mathrm{i}(\mathbf{\Phi}+\mathbf{\Phi}^\top)\mathbf{\Sigma}\right)\right)^{-\frac{n}{2}}.
$$

8 More Matrix Variate Distribution

Theorem 8.1. *Let $\mathbf{A}\sim\mathcal{W}_p(\mathbf{I},n)$ and $\mathbf{B}\sim\mathcal{W}_p(\mathbf{\Sigma}^{-1},m)$ be independent, then*

$$
\mathbf{U}=\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2},
$$

has matrix F-distribution with n and m degrees of freedom. Its density function is

$$
f(\mathbf{U})=\frac{\Gamma_p\left(\frac{m+n}{2}\right)(\det(\mathbf{\Sigma}))^{-\frac{n}{2}}}{\Gamma_p(\frac{m}{2})\Gamma_p(\frac{n}{2})}\cdot(\det(\mathbf{U}))^{\frac{n-p-1}{2}}\left(\det(\mathbf{I}+\mathbf{U}\mathbf{\Sigma}^{-1})\right)^{-\frac{m+n}{2}}.
$$

This document is organized by Luo Luo. Please do not distribute.

Proof. We first consider the conditional distribution of $\mathbf{U}$ for given $\mathbf{C} = \mathbf{B}^{-1}$. Since we have

$$\mathbf{A} \sim \mathcal{W}_p(\mathbf{I}, n) \qquad \text{and} \qquad \mathbf{U} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \mathbf{C}^{1/2}\mathbf{A}\mathbf{C}^{1/2},$$

Lemma 7.4 indicates $\mathbf{U} \mid \mathbf{C} \sim \mathcal{W}_p(\mathbf{C}, n)$, and we have

$$f(\mathbf{U}\,|\,\mathbf{C}) = w(\mathbf{U}\,|\,\mathbf{C}, n) = \frac{(\det(\mathbf{U}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{U}\right)\right)}{2^{\frac{np}{2}}\Gamma_p(\frac{n}{2})(\det(\mathbf{C}))^{\frac{n}{2}}}.$$

The distribution $\mathbf{B} \sim \mathcal{W}_p(\mathbf{\Sigma}^{-1}, m)$ implies

$$\mathbf{C} = \mathbf{B}^{-1} \sim \mathcal{W}_p^{-1}(\mathbf{\Sigma}, m)$$

and we have

$$f(\mathbf{C}) = w^{-1}(\mathbf{C} \mid \mathbf{\Sigma}, m) = \frac{(\det(\mathbf{\Sigma}))^{\frac{m}{2}} (\det(\mathbf{C}))^{-\frac{m+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}\mathbf{C}^{-1}\right)\right)}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)}.$$

Hence, the density of $\mathbf{U}$ is

$$\begin{aligned}
f(\mathbf{U}) &= \int f(\mathbf{U}, \mathbf{C})\,\mathrm{d}\mathbf{C} = \int f(\mathbf{U}\,|\,\mathbf{C}) f(\mathbf{C})\,\mathrm{d}\mathbf{C} \\
&= \int \frac{(\det(\mathbf{\Sigma}))^{\frac{m}{2}} (\det(\mathbf{U}))^{\frac{n-p-1}{2}} (\det(\mathbf{C}))^{-\frac{m+n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{U}+\mathbf{\Sigma})\mathbf{C}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p(\frac{m}{2})\Gamma_p\left(\frac{n}{2}\right)}\,\mathrm{d}\mathbf{C}.
\end{aligned}$$

In the view of density function $w_p^{-1}(\mathbf{C}\,|\,\mathbf{U}+\mathbf{\Sigma}, m+n)$, we have

$$\begin{aligned}
1 &= \int w_p^{-1}(\mathbf{C}\,|\,\mathbf{U}+\mathbf{\Sigma}, m+n)\,\mathrm{d}\mathbf{C} \\
&= \int \frac{(\det(\mathbf{U}+\mathbf{\Sigma}))^{\frac{m+n}{2}} (\det(\mathbf{C}))^{-\frac{m+n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{U}+\mathbf{\Sigma})\mathbf{C}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{m+n}{2}\right)}\,\mathrm{d}\mathbf{C},
\end{aligned}$$

that is

$$\frac{\Gamma_p\left(\frac{m+n}{2}\right)}{(\det(\mathbf{U}+\mathbf{\Sigma}))^{\frac{m+n}{2}}} = \int \frac{(\det(\mathbf{C}))^{-\frac{m+n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{U}+\mathbf{\Sigma})\mathbf{C}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}}\,\mathrm{d}\mathbf{C}.$$

Hence, we have

$$\begin{aligned}
f(\mathbf{U}) =& \frac{(\det(\mathbf{\Sigma}))^{\frac{m}{2}} (\det(\mathbf{U}))^{\frac{n-p-1}{2}}}{\Gamma_p(\frac{m}{2})\Gamma_p(\frac{n}{2})} \int \frac{(\det(\mathbf{C}))^{-\frac{m+n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{U}+\mathbf{\Sigma})\mathbf{C}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}}\,\mathrm{d}\mathbf{C} \\
=& \frac{(\det(\mathbf{\Sigma}))^{\frac{m}{2}} (\det(\mathbf{U}))^{\frac{n-p-1}{2}}}{\Gamma_p(\frac{m}{2})\Gamma_p(\frac{n}{2})} \cdot \frac{\Gamma_p\left(\frac{m+n}{2}\right)}{(\det(\mathbf{U}+\mathbf{\Sigma}))^{\frac{m+n}{2}}} \\
=& \frac{\Gamma_p(\frac{m+n}{2})\,(\det(\mathbf{\Sigma}))^{-\frac{n}{2}}}{\Gamma_p(\frac{m}{2})\Gamma_p(\frac{n}{2})} \cdot (\det(\mathbf{U}))^{\frac{n-p-1}{2}} \left(\det(\mathbf{I}+\mathbf{U}\mathbf{\Sigma}^{-1})\right)^{-\frac{m+n}{2}},
\end{aligned}$$

where the last step is because of

$$\frac{\det(\mathbf{\Sigma})}{\det(\mathbf{U}+\mathbf{\Sigma})} = \frac{1}{\det(\mathbf{\Sigma}^{-1})\det(\mathbf{U}+\mathbf{\Sigma})} = \frac{1}{\det(\mathbf{\Sigma}^{-1}\mathbf{U}+\mathbf{I})}.$$

□

This document is organized by Luo Luo.
Please do not distribute.

Remark 8.1. *The original definition of matrix Beta distribution is based on the density of the form*

$$\int w(\mathbf{U}\,|\,\mathbf{C}, n)\cdot w^{-1}(\mathbf{C}\,|\,\mathbf{\Sigma}, m)\,\mathrm{d}\mathbf{C}.$$

Remark 8.2. *For $p = 1$ and $\mathbf{\Sigma} = n/m$, it reduce to the classical F-distribution in univariate case*

$$f(u) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(\frac{m}{2})\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{n}{2}} u^{\frac{n}{2}-1}\left(1+\frac{m}{n}\cdot u\right)^{-\frac{m+n}{2}} = \frac{1}{B\left(\frac{m}{2},\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{n}{2}} u^{\frac{n}{2}-1}\left(1+\frac{m}{n}\cdot u\right)^{-\frac{m+n}{2}}.$$

It is natural to define the multivariate Beta function as

$$B_p(a, b) = \frac{\Gamma_p(a)\Gamma_p(b)}{\Gamma_p(a+b)}.$$

Remark 8.3. *Let $\mathbf{B} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ be SVD of $\mathbf{B}$, we define $\mathbf{B}^{1/2} = \mathbf{Q}\mathbf{D}^{1/2}\mathbf{Q}^\top$. We can also define $\mathbf{B}^{1/2}$ be a $p \times p$ matrix that satisfies $\mathbf{B} = \mathbf{B}^{1/2}(\mathbf{B}^{1/2})^\top$.*

Remark 8.4. *The application of matrix F-distribution can be found in Mulder and Pericchi's paper [13].*

Theorem 8.2. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{B} \sim \mathcal{W}_p(\mathbf{\Sigma}, m)$ be independent, then*

$$\mathbf{W} = (\mathbf{A}+\mathbf{B})^{-1/2}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1/2}$$

has matrix Beta distribution with parameters $n/2$ and $m/2$. Its density function is

$$f(\mathbf{W}) = \frac{1}{B_p(\frac{n}{2},\frac{m}{2})}\cdot(\det(\mathbf{W}))^{\frac{n-p-1}{2}}(\det(\mathbf{I}-\mathbf{W}))^{\frac{m-p-1}{2}}$$

if $\mathbf{0} \prec \mathbf{W} \prec \mathbf{I}$ and 0 elsewhere.

Proof. Since random matrices $\mathbf{A}$ and $\mathbf{B}$ are independent, we can write their joint density as

$$f(\mathbf{A}, \mathbf{B}) = f(\mathbf{A})f(\mathbf{B}) = w_p(\mathbf{A}\,|\,\mathbf{\Sigma}, n)\cdot w_p(\mathbf{B}\,|\,\mathbf{\Sigma}, m).$$

We define

$$\mathbf{D} = \mathbf{A} + \mathbf{B}.$$

We shall show the joint density of $\mathbf{D}$ and $\mathbf{W}$. Consider the determinant of Jacobian for following transforms

$$\begin{pmatrix}\mathbf{W}\\ \mathbf{D}\end{pmatrix} \xrightarrow{(\det(\mathbf{D}))^{\frac{p+1}{2}}} \begin{pmatrix}\mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2}\\ \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\\ \mathbf{D}\end{pmatrix} \xrightarrow{1} \begin{pmatrix}\mathbf{A}\\ \mathbf{D}-\mathbf{A}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix}.$$

The definition of $\mathbf{W}$ and $\mathbf{D}$ implies

$$\mathbf{A} = \mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2} \qquad \text{and} \qquad \mathbf{B} = \mathbf{D}^{1/2}(\mathbf{I}-\mathbf{W})\mathbf{D}^{1/2}.$$

Hence, the joint density of $\mathbf{W}$ and $\mathbf{D}$ is

$$\begin{aligned}&w_p(\mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2}\,|\,\mathbf{\Sigma}, n)\cdot w_p(\mathbf{D}^{1/2}(\mathbf{I}-\mathbf{W})\mathbf{D}^{1/2}\,|\,\mathbf{\Sigma}, m)\cdot(\det(\mathbf{D}))^{\frac{p+1}{2}}\\ =&\frac{(\det(\mathbf{W}))^{\frac{n-p-1}{2}}(\det(\mathbf{I}-\mathbf{W}))^{\frac{m-p-1}{2}}}{2^{\frac{(m+n)p}{2}}\Gamma_p(\frac{n}{2})\Gamma_p(\frac{m}{2})(\det(\mathbf{\Sigma}))^{\frac{m+n}{2}}}\cdot(\det(\mathbf{D}))^{\frac{m+n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{D}\right)\right).\end{aligned}$$

In the view of $w_p(\mathbf{D}\,|\,\mathbf{\Sigma}, m+n)$, we have

$$1 = \int w_p(\mathbf{D}\,|\,\mathbf{\Sigma}, m+n)\,\mathrm{d}\mathbf{D} = \int \frac{(\det(\mathbf{D}))^{\frac{m+n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{D}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p(\frac{m+n}{2})(\det(\mathbf{\Sigma}))^{\frac{m+n}{2}}}\,\mathrm{d}\mathbf{D}.$$

This document is organized by Luo Luo.
Please do not distribute.

Hence, we have

$$2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{m+n}{2}\right)(\det(\mathbf{\Sigma}))^{\frac{m+n}{2}} = \int (\det(\mathbf{D}))^{\frac{m+n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{D}\right)\right) \mathrm{d}\mathbf{D}$$

and

$$\begin{aligned} f(\mathbf{W}) =& \int w_p(\mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2} \,|\, \mathbf{\Sigma}, n) \cdot w_p(\mathbf{D}^{1/2}(\mathbf{I}-\mathbf{W})\mathbf{D}^{1/2} \,|\, \mathbf{\Sigma}, m) \cdot (\det(\mathbf{D}))^{\frac{p+1}{2}} \,\mathrm{d}\mathbf{D} \\ =& \frac{1}{B_p\left(\frac{n}{2},\frac{m}{2}\right)} \cdot (\det(\mathbf{W}))^{\frac{n-p-1}{2}} \left(\det(\mathbf{I}-\mathbf{W})\right)^{\frac{m-p-1}{2}}. \end{aligned}$$

□

Remark 8.5. *For given independent random variables $a \sim \chi_n^2$ and $b \sim \chi_m^2$, we have*

$$w = \frac{a}{a+b} \sim \mathrm{Beta}\left(\frac{n}{2}, \frac{m}{2}\right)$$

with density function

$$f(w) = \frac{1}{B(\frac{m}{2},\frac{n}{2})} \cdot w^{\frac{n}{2}-1}(1-w)^{\frac{m}{2}-1}, \qquad \textit{where } B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t.$$

In Theorem 5.1, we also shown that

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

which matches the definition of multivariate Beta function in Remark 8.2!

This document is organized by Luo Luo.
Please do not distribute.

9 Revisiting Likelihood Ratio Criterion

Theorem 9.1. *If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ with $N > p$. Define the likelihood ratio criterion as*

$$\lambda = \frac{\max_{\mathbf{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}_0, \mathbf{\Sigma})}{\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \mathbf{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \mathbf{\Sigma})},$$

where

$$L(\boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^{-\frac{pN}{2}}(\det(\mathbf{\Sigma}))^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha-\boldsymbol{\mu})\right).$$

then we have

$$\lambda^{\frac{2}{N}} = \frac{1}{1+T^2/(N-1)},$$

where

$$T^2 = N(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^\top\mathbf{S}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0), \qquad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha \qquad \textit{and} \qquad \mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha-\bar{\mathbf{x}})(\mathbf{x}_\alpha-\bar{\mathbf{x}})^\top.$$

Proof. The maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}}_\Omega = \bar{\mathbf{x}} \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}}_\Omega = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

Following the proof of Theorem 6.2, we have

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right).$$

If we restrict $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, the likelihood function is maximized at

$$\boldsymbol{\Sigma}_\omega = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)^\top.$$

Similarly, we also have

$$\max_{\boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right).$$

Thus the likelihood ratio criterion is

$$\begin{aligned}
\lambda =& \frac{(2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right)}{(2\pi)^{-\frac{pN}{2}}\left(\det(\boldsymbol{\Sigma}_\omega)\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right)} = \frac{\left(\det(\boldsymbol{\Sigma}_\omega)\right)^{\frac{N}{2}}}{\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{\frac{N}{2}}} \\
=& \frac{\left(\det\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\right)\right)^{\frac{N}{2}}}{\left(\det\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)^\top\right)\right)^{\frac{N}{2}}} \\
=& \frac{\left(\det\left(\mathbf{A}\right)\right)^{\frac{N}{2}}}{\left(\det\left(\mathbf{A} + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\right)\right)^{\frac{N}{2}}},
\end{aligned}$$

where $\mathbf{A} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = (N-1)\mathbf{S}$ and the last step is because of

$$\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)^\top \\
=&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \\
=&\sum_{\alpha=1}^{N}\left((\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top + (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\right) \\
=&\mathbf{A} + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\lambda^{\frac{2}{N}} =& \frac{\det\left(\mathbf{A}\right)}{\det\left(\mathbf{A} + \left(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\right)\left(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\right)\right)} \\
=& \frac{\det\left(\mathbf{A}\right)}{\left(1 + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\mathbf{A}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\right)\det\left(\mathbf{A}\right)} \\
=& \frac{1}{1 + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\mathbf{A}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)} \\
=& \frac{1}{1 + T^2/(N-1)}
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

where

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = (N-1)N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{A}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

The second line uses the property of Schur complement to obtain

$$\det\left(\begin{bmatrix} \mathbf{A} & \mathbf{u} \\ -\mathbf{u}^\top & 1 \end{bmatrix}\right) = \det\left(\mathbf{A} + \mathbf{u}\mathbf{u}^\top\right) \cdot \det(1) = \det\left(\begin{bmatrix} 1 & -\mathbf{u}^\top \\ \mathbf{u} & \mathbf{A} \end{bmatrix}\right) = \left(1 + \mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}\right) \cdot \det(\mathbf{A}),$$

with $\mathbf{u} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$. Recall that The decomposition

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

means $\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$. Similarly, we also have $\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})$. □

Remark 9.1. *The condition $\lambda^{2/N} > c$ for some $c \in (0,1)$ is equivalent to*

$$\lambda^{\frac{2}{N}} = \frac{1}{1 + T^2/(N-1)} > c \qquad \Longleftrightarrow \qquad T^2 < \frac{(N-1)(1-c)}{c}.$$

Multivariate Analysis of Variance (MANOVA): We consider testing the equality of means with common covariance. Let $\mathbf{x}_\alpha^{(g)}$ be an observation from the g-th population $\mathcal{N}_p(\boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma})$ for $\alpha = 1, \dots, N_g$ and $g = 1, \dots, q$. We wish to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g.$$

The likelihood function is

$$L(\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(q)}, \boldsymbol{\Sigma}) = \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{pN_g}{2}}(\det(\boldsymbol{\Sigma}))^{\frac{N_g}{2}}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \boldsymbol{\mu}^{(g)}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_\alpha^{(g)} - \boldsymbol{\mu}^{(g)}\right)\right).$$

We denote

1. We let $\boldsymbol{\theta} = \{\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(q)}, \boldsymbol{\Sigma}\}$ be the parameters.
2. The set Ω is the space in which $\boldsymbol{\Sigma}$ is positive definite and each $\boldsymbol{\mu}^{(g)}$ is any p-dimensional vector.
3. The set ω is the space in which $\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}$ (p-dimensional vectors) and $\boldsymbol{\Sigma}$ is positive definite matrix.

We first consider the space Ω. The maximum likelihood estimators of $\boldsymbol{\mu}^{(g)}$ can be achieved by considering each population separately, that is

$$\hat{\boldsymbol{\mu}}_\Omega^{(g)} = \bar{\mathbf{x}}^{(g)} = \frac{1}{N_g}\sum_{\alpha=1}^{N_g} \mathbf{x}_\alpha^{(g)}.$$

Substitute $\{\hat{\boldsymbol{\mu}}_\Omega^{(g)}\}$ into L, we obtain

$$\begin{aligned} L(\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma}) =& \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{pN_g}{2}}(\det(\boldsymbol{\Sigma}))^{\frac{N_g}{2}}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)\right) \\ =& \frac{1}{(2\pi)^{\frac{pN}{2}}(\det(\boldsymbol{\Sigma}))^{\frac{N}{2}}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)\left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)^\top \boldsymbol{\Sigma}^{-1}\right)\right) \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$\propto \frac{1}{(\det(\mathbf{\Sigma}))^{\frac{N}{2}}} \exp\Bigg(-\frac{1}{2}\mathrm{tr}\Bigg(\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g} \big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)^\top \mathbf{\Sigma}^{-1}\Bigg)\Bigg),$$

where $N = N_1 + \cdots + N_g$. In the view of Theorem 6.2 (the term $\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)$ plays the role of $\mathbf{A}$ in Theorem 6.2), we have

$$\hat{\mathbf{\Sigma}}_\Omega = \frac{1}{N}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)^\top \quad \text{and} \quad \sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) = (2\pi)^{-\frac{pN}{2}}\big(\det(\hat{\mathbf{\Sigma}}_\Omega)\big)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right).$$

We then consider the space ω. The maximum likelihood estimators can be obtained by considering N identical independent normal observation, where $N = N_1 + \cdots + N_g$. Hence, we have

$$\hat{\boldsymbol{\mu}}_\omega^{(g)} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\mathbf{x}_\alpha^{(g)}, \qquad \hat{\mathbf{\Sigma}}_w = \frac{1}{N}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\big)^\top$$

and

$$\sup_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta}) = (2\pi)^{-\frac{pN}{2}}\big(\det(\hat{\mathbf{\Sigma}}_\omega)\big)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right).$$

Thus the likelihood ratio criterion is

$$\lambda = \frac{\sup_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})} = \frac{(2\pi)^{-\frac{pN}{2}}\big(\det(\hat{\mathbf{\Sigma}}_\omega)\big)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right)}{(2\pi)^{-\frac{pN}{2}}\big(\det(\hat{\mathbf{\Sigma}}_\Omega)\big)^{-\frac{N}{2}}\exp\left(-\frac{1}{2}pN\right)},$$

which means

$$\lambda = \frac{(\det(\hat{\mathbf{\Sigma}}_\Omega))^{\frac{N}{2}}}{(\det(\hat{\mathbf{\Sigma}}_\omega))^{\frac{N}{2}}}.$$

Now we consider the distribution of $\lambda^{\frac{2}{N}}$ under hypothesis $H_0 : \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(g)}$. Since each observation are independent, each

$$\mathbf{A}_g = \sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)^\top \quad \text{has distribution} \;\; \mathbf{A}_g \sim \mathcal{W}_p(\mathbf{\Sigma}, N_g - 1).$$

independently. We denote

$$\mathbf{A} = N\hat{\mathbf{\Sigma}}_\Omega = \sum_{g=1}^{q}\mathbf{A}_g = \sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\big)^\top,$$

which has distribution

$$\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, N - q).$$

We can write

$$\begin{aligned} N\hat{\mathbf{\Sigma}}_\omega =& \sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\big)^\top \\ =& \sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)} + \bar{\mathbf{x}}^{(g)} - \bar{\mathbf{x}}\big)\big(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)} + \bar{\mathbf{x}}^{(g)} - \bar{\mathbf{x}}\big)^\top \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

$$\begin{aligned}
=&\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\Big(\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)^\top+\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)\big(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}}\big)^\top+\big(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}}\big)\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)^\top\\
&\qquad+\big(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}}\big)\big(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}}\big)^\top\Big)\\
=&\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)\big(\mathbf{x}_\alpha^{(g)}-\bar{\mathbf{x}}^{(g)}\big)^\top+\sum_{g=1}^{q}N_g(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}})(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}})^\top\\
=&\mathbf{A}+\mathbf{B},
\end{aligned}$$

where

$$\mathbf{A}+\mathbf{B}=N\hat{\mathbf{\Sigma}}_\omega\sim\mathcal{W}_p(\mathbf{\Sigma},N-1)\qquad\text{and}\qquad\mathbf{B}=\sum_{g=1}^{q}N_g(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}})(\bar{\mathbf{x}}^{(g)}-\bar{\mathbf{x}})^\top.$$

We can show $\mathbf{B}$ is independent on $\mathbf{A}$ and $\mathbf{B}\sim\mathcal{W}_p(\mathbf{\Sigma},q-1)$ as follows.

Remark 9.2. *We denote*

$$\mathbf{X}=\begin{bmatrix}\mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_{N_1}^{(1)} & \mathbf{x}_1^{(2)} & \cdots & \mathbf{x}_{N_q}^{(q)}\end{bmatrix}^\top\in\mathbb{R}^{N\times p},$$

then we have

$$\mathbf{A}+\mathbf{B}=N\hat{\mathbf{\Sigma}}_\omega=\mathbf{X}^\top\mathbf{Q}_1\mathbf{X},\qquad\mathbf{A}=\mathbf{X}^\top\mathbf{Q}_2\mathbf{X}\qquad\textit{and}\qquad\mathbf{B}=\mathbf{X}^\top\mathbf{Q}_3\mathbf{X},$$

where

$$\mathbf{Q}_1=\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top,\qquad\mathbf{Q}_2=\begin{bmatrix}\mathbf{I}_{N_1}-\frac{1}{N_1}\mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{N_2}-\frac{1}{N_2}\mathbf{1}_{N_2}\mathbf{1}_{N_2}^\top & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{N_q}-\frac{1}{N_q}\mathbf{1}_{N_q}\mathbf{1}_{N_q}^\top\end{bmatrix}$$

$$\textit{and}\qquad\mathbf{Q}_3=\begin{bmatrix}\left(\frac{1}{N_1}-\frac{1}{N}\right)\mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top & -\frac{1}{N}\mathbf{1}_{N_1}\mathbf{1}_{N_2}^\top & \cdots & -\frac{1}{N}\mathbf{1}_{N_1}\mathbf{1}_{N_q}^\top\\ -\frac{1}{N}\mathbf{1}_{N_2}\mathbf{1}_{N_1}^\top & \left(\frac{1}{N_2}-\frac{1}{N}\right)\mathbf{1}_{N_2}\mathbf{1}_{N_2}^\top & \cdots & -\frac{1}{N}\mathbf{1}_{N_2}\mathbf{1}_{N_q}^\top\\ \vdots & \vdots & \ddots & \vdots\\ -\frac{1}{N}\mathbf{1}_{N_q}\mathbf{1}_{N_1}^\top & -\frac{1}{N}\mathbf{1}_{N_q}\mathbf{1}_{N_2}^\top & \cdots & \left(\frac{1}{N_q}-\frac{1}{N}\right)\mathbf{1}_{N_q}\mathbf{1}_{N_q}^\top\end{bmatrix}$$

are projection matrix with rank $N-1$, $N-q$ and $q-1$ respectively. We let $\mathbf{Z}=\mathbf{\Sigma}^{-1/2}\mathbf{X}$, then we have

$$\mathbf{A}=(\mathbf{Q}_2\mathbf{X})^\top\mathbf{Q}_2\mathbf{X}=(\mathbf{Q}_2\mathbf{\Sigma}^{1/2}\mathbf{Z})^\top\mathbf{Q}_2\mathbf{\Sigma}^{1/2}\mathbf{Z},\quad\mathbf{B}=(\mathbf{Q}_3\mathbf{X})^\top\mathbf{Q}_3\mathbf{X}=(\mathbf{Q}_3\mathbf{\Sigma}^{1/2}\mathbf{Z})^\top\mathbf{Q}_3\mathbf{\Sigma}^{1/2}\mathbf{Z},$$

and each column of $\mathbf{Z}$ has normal distribution with covariance $\mathbf{I}$. Therefore, the fact $\mathbf{Q}_2\mathbf{Q}_3=\mathbf{0}$ implies

$$(\mathbf{Q}_2\mathbf{\Sigma}^{1/2}\mathbf{Z})^\top\mathbf{Q}_3\mathbf{\Sigma}^{1/2}\mathbf{Z}=\mathbf{Z}^\top\mathbf{\Sigma}^{1/2}\mathbf{Q}_2\mathbf{Q}_3\mathbf{\Sigma}^{1/2}\mathbf{Z}=\mathbf{0},$$

which means $\mathbf{A}$ and $\mathbf{B}$ are independent.

Theorem 7.6 indicates the characteristic function of $a_{11},\dots,a_{pp}$, $2a_{12},\dots,2a_{p-1,p}$ is

$$\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right]=\left(\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{N-q}{2}},$$

and the characteristic function of $(a_{11}+b_{11}),\dots,(a_{pp}+b_{pp})$, $2(a_{12}+b_{12}),\dots,2(a_{p-1,p}+b_{p-1,p})$ is

$$\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}((\mathbf{A}+\mathbf{B})\mathbf{\Theta}))\right]=\left(\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{N-1}{2}},$$

This document is organized by Luo Luo.
Please do not distribute.

where $\mathbf{\Theta} \in \mathbb{R}^{p\times p}$ is symmetric. $\mathbf{A}$ and $\mathbf{B}$ are independent. Hence, we have

$$\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}((\mathbf{A}+\mathbf{B})\mathbf{\Theta}))\right] = \mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right]\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{B}\mathbf{\Theta}))\right],$$

which means the characteristic function of $b_{11}, \dots, b_{pp}$, $2b_{12}, \dots, 2b_{p-1,p}$ is

$$\begin{aligned}\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{B}\mathbf{\Theta}))\right] =& \frac{\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}+\mathbf{B})\mathbf{\Theta})\right]}{\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta})\right]} \\ =& \left(\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{q-1}{2}}.\end{aligned}$$

Hence, we achieve $\mathbf{B} \sim \mathcal{W}_p(\mathbf{\Sigma}, q-1)$.

For two independent Wishart distributed variables $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{B} \sim \mathcal{W}_p(\mathbf{\Sigma}, m)$ with $n \geq p$, the ratio $\det(\mathbf{A})/\det(\mathbf{A}+\mathbf{B})$ has Wilks' Lambda distribution with degrees of freedom n and m, written as

$$\frac{\det(\mathbf{A})}{\det(\mathbf{A}+\mathbf{B})} \sim \Lambda_{p,n,m}.$$

Theorem 9.2. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$ and $\mathbf{B} \sim \mathcal{W}_p(\mathbf{\Sigma}, m)$ be two independent Wishart distributed variables with $n \geq p$, then we can write*

$$\frac{\det(\mathbf{A})}{\det(\mathbf{A}+\mathbf{B})} = \prod_{i=1}^{p} u_i \sim \Lambda_{p,n,m},$$

where $u_1, \dots, u_p$ are independent distributed as

$$u_i \sim \mathrm{Beta}\left(\frac{n+1-i}{2}, \frac{m}{2}\right).$$

Theorem 9.3. *Let $\mathbf{A} \sim \mathcal{W}_p(\mathbf{\Sigma}, n)$, we have*

$$\det(\mathbf{A}) = \det(\mathbf{\Sigma})\prod_{i=1}^{p} v_i$$

where $v_1, \dots, v_p$ are independent distributed as $v_i \sim \chi^2_{n-i+1}$.

Proof. We let

$$\mathbf{C} = \mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2} \sim \mathcal{W}_p(\mathbf{I}, n),$$

then

$$\det(\mathbf{A}) = \det(\mathbf{\Sigma})\det(\mathbf{C}).$$

For $p = 1$, we have

$$v_1 = \mathbf{C} \sim \chi^2_n \qquad \text{and} \qquad \det(\mathbf{A}) = \det(\mathbf{\Sigma})v_1.$$

For $p \geq 2$, we suppose the $(p-1)\times(p-1)$ random matrix with distribution $\mathcal{W}_{p-1}(\mathbf{I}, n)$ can be written as $\prod_{i=1}^{p-1} v_i$, where $v_1, \dots, v_{p-1}$ are independent distributed as $v_i \sim \chi^2_{n-i+1}$. We partition $\mathbf{C}$ into 1 and $p-1$ rows and columns as

$$\mathbf{C} = \begin{bmatrix} c_{11} & \mathbf{c}_{12}^\top \\ \mathbf{c}_{21} & \mathbf{C}_{22}. \end{bmatrix}$$

We let

$$c_{11.2} = c_{11} - \mathbf{c}_{12}\mathbf{C}_{22}^{-1}\mathbf{c}_{12}^\top,$$

This document is organized by Luo Luo.
Please do not distribute.

then Lemma 7.5 means $c_{11.2} \sim \chi^2_{n-p+1}$ is independent on $\mathbf{C}_{22} \sim \mathcal{W}_{p-1}(\mathbf{I}, n)$. The induction hypothesis indicates we can write

$$\det(\mathbf{C}_{22}) = \prod_{i=1}^{p-1} v_i,$$

where $v_1, \ldots, v_{p-1}$ are independent distributed as $v_i \sim \chi^2_{n-i+1}$. Hence, we let $v_1 = c_{11.2}$ and achieve

$$\det(\mathbf{C}) = c_{11.2} \det(\mathbf{C}_{22}) = \prod_{i=1}^{p} v_i,$$

where $u_1, \ldots, u_p$ are independent distributed as $v_i \sim \chi^2_{n-i+1}$. □

Remark 9.3. *The proofs for the two above theorems provided by Anderson [1, Section 8.4] are very complicated. Can you follow the proof of Theorem 9.3 to prove Theorem 9.2?*

10 Multivariate Linear Regression

Given dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$ are the feature and the corresponding output of the i-th data. We suppose

$$\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\epsilon}_i \qquad \text{with} \quad \mathbf{B} \in \mathbb{R}^{p \times q} \quad \text{and} \quad \boldsymbol{\epsilon}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$$

for $i = 1, \ldots, N$, $\boldsymbol{\Sigma} \succ 0$ and $N > p$. We regard $\mathbf{B} \in \mathbb{R}^{p \times q}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ as parameters, then

$$\boldsymbol{\epsilon}_i = \mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}).$$

Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times p}, \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times q} \qquad \text{and} \qquad \mathbf{E} = \begin{bmatrix} \boldsymbol{\epsilon}_1^\top \\ \vdots \\ \boldsymbol{\epsilon}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times q}.$$

We also suppose $\mathbf{X}$ is full rank.

We construct the posterior likelihood function

$$\begin{aligned} L(\mathbf{B}, \boldsymbol{\Sigma}) &= \prod_{\alpha=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{B}^\top \mathbf{x}_\alpha - \mathbf{y}_\alpha)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{B}^\top \mathbf{x}_\alpha - \mathbf{y}_\alpha)\right) \\ &= \frac{1}{(2\pi)^{Np/2} (\det(\boldsymbol{\Sigma}))^{N/2}} \exp\left(-\frac{1}{2} \mathrm{tr}\left((\mathbf{X}\mathbf{B} - \mathbf{Y}) \boldsymbol{\Sigma}^{-1} (\mathbf{X}\mathbf{B} - \mathbf{Y})^\top\right)\right), \end{aligned}$$

where the last step can be verified by

$$\begin{aligned} &(\mathbf{X}\mathbf{B} - \mathbf{Y}) \boldsymbol{\Sigma}^{-1} (\mathbf{X}\mathbf{B} - \mathbf{Y})^\top = (\mathbf{X}\mathbf{B} - \mathbf{Y}) \boldsymbol{\Sigma}^{-1} (\mathbf{B}^\top \mathbf{X}^\top - \mathbf{Y}^\top) \\ =& \begin{bmatrix} (\mathbf{B}^\top \mathbf{x}_1 - \mathbf{y}_1)^\top \\ \vdots \\ (\mathbf{B}^\top \mathbf{x}_N - \mathbf{y}_N)^\top \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mathbf{B}^\top \mathbf{x}_1 - \mathbf{y}_1 & \cdots & \mathbf{B}^\top \mathbf{x}_N - \mathbf{y}_N \end{bmatrix}. \end{aligned}$$

We first consider minimizing

$$f(\mathbf{B}) = \mathrm{tr}\left((\mathbf{X}\mathbf{B} - \mathbf{Y}) \boldsymbol{\Sigma}^{-1} (\mathbf{B}^\top \mathbf{X}^\top - \mathbf{Y}^\top)\right)$$

with respect to $\mathbf{B}$. We provide some trick to compute the gradient of such matrix variate function.

This document is organized by Luo Luo.
Please do not distribute.

Remark 10.1. *Recall the relationship between differential and derivative/gradient as follows*

1. *For single value input function $f : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathrm{d}f(x) = f'(x)\,\mathrm{d}x.$$

2. *For vector input function $f : \mathbb{R}^p \to \mathbb{R}$, we have*

$$\mathrm{d}f(\mathbf{x}) = \sum_{i=1}^{p} \frac{\partial f(\mathbf{x})}{\partial x_i} \cdot \mathrm{d}x_i = \langle \nabla f(\mathbf{x}), \mathrm{d}\mathbf{x} \rangle,$$

where

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_p} \end{bmatrix} \in \mathbb{R}^p \qquad \textit{and} \qquad \mathrm{d}\mathbf{x} = \begin{bmatrix} \mathrm{d}x_1 \\ \vdots \\ \mathrm{d}x_p \end{bmatrix} \in \mathbb{R}^p.$$

3. *For matrix input function $f : \mathbb{R}^{p \times q} \to \mathbb{R}$, we have*

$$\begin{aligned} \mathrm{d}f(\mathbf{X}) =& \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\partial f(\mathbf{X})}{\partial x_{ij}} \cdot \mathrm{d}x_{ij} \\ =& \langle \nabla f(\mathbf{X}), \mathrm{d}\mathbf{X} \rangle \\ =& \mathrm{tr}\big(\nabla f(\mathbf{X})^\top \mathrm{d}\mathbf{X}\big), \end{aligned}$$

where

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{X})}{\partial x_{11}} & \dots & \dfrac{\partial f(\mathbf{X})}{\partial x_{1q}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f(\mathbf{X})}{\partial x_{p1}} & \dots & \dfrac{\partial f(\mathbf{X})}{\partial x_{pq}} \end{bmatrix} \in \mathbb{R}^{p \times q} \qquad \textit{and} \qquad \mathrm{d}\mathbf{X} = \begin{bmatrix} \mathrm{d}x_{11} & \dots & \mathrm{d}x_{1q} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{p1} & \dots & \mathrm{d}x_{pq} \end{bmatrix} \in \mathbb{R}^{p \times q}.$$

This implies if the differential $\mathrm{d}f(\mathbf{X})$ has the form of

$$\mathrm{d}f(\mathbf{X}) = \mathrm{tr}\big(\mathbf{A}\mathrm{d}\mathbf{X}\big),$$

then the gradient of $f(\mathbf{X})$ is $\mathbf{A}^\top$.

This document is organized by Luo Luo.
Please do not distribute.

We can write

$$\begin{aligned} f(\mathbf{B}) =& \mathrm{tr}\left((\mathbf{X}\mathbf{B} - \mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{X}^\top - \mathbf{Y}^\top)\right) \\ =& \mathrm{tr}\left(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\right) - 2\mathrm{tr}\left(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\right) + \mathrm{tr}\left(\mathbf{Y}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\right), \end{aligned} \tag{28}$$

then we write its differential as follows

$$\begin{aligned} \mathrm{d}f(\mathbf{B}) =& \mathrm{d}\mathrm{tr}\left(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\right) - 2\mathrm{d}\mathrm{tr}\left(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\right) + \mathrm{d}\mathrm{tr}\left(\mathbf{Y}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\right) \\ =& \mathrm{tr}\left(\mathrm{d}(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top)\right) - 2\mathrm{tr}\left(\mathrm{d}(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top)\right). \end{aligned} \tag{29}$$

For the first term, we have

$$\begin{aligned} & \mathrm{d}(\mathbf{X}\mathbf{B} \cdot \mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top) \\ =& \mathrm{d}(\mathbf{X}\mathbf{B}) \cdot \mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top + \mathbf{X}\mathbf{B} \cdot \mathrm{d}(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top) \end{aligned}$$

$$=\mathbf{X}(\mathrm{d}\mathbf{B})\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top+\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}(\mathrm{d}\mathbf{B}^\top)\mathbf{X}^\top,$$

which implies

$$\begin{aligned}
&\mathrm{tr}\left(\mathrm{d}(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top)\right)\\
=&\mathrm{tr}\left(\mathbf{X}(\mathrm{d}\mathbf{B})\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\right)+\mathrm{tr}\left(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}(\mathrm{d}\mathbf{B}^\top)\mathbf{X}^\top\right)\\
=&\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)+\mathrm{tr}\left((\mathrm{d}\mathbf{B}^\top)\mathbf{X}^\top\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\right)\\
=&\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)+\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)\\
=&2\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)
\end{aligned}\tag{30}$$

For the second term, we have

$$\begin{aligned}
&2\mathrm{tr}\left(\mathrm{d}(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{Y}^\top)\right)\\
=&2\mathrm{tr}\left(\mathbf{X}(\mathrm{d}\mathbf{B})\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\right)\\
=&2\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)
\end{aligned}\tag{31}$$

Substituting equations (30) and (31) into (29), we have

$$\mathrm{d}f(\mathbf{B})=2\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)-2\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{Y}^\top\mathbf{X}\mathrm{d}\mathbf{B}\right)=\mathrm{tr}\left(2\mathbf{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}-\mathbf{Y}^\top\mathbf{X})\mathrm{d}\mathbf{B}\right),$$

which means

$$\begin{aligned}
\nabla f(\mathbf{B})=&\big(2\mathbf{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}-\mathbf{Y}^\top\mathbf{X})\big)^\top\\
=&2\big(\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}-\mathbf{Y}^\top\mathbf{X}\big)^\top\mathbf{\Sigma}^{-1}\\
=&2\big(\mathbf{X}^\top\mathbf{X}\mathbf{B}-\mathbf{X}^\top\mathbf{Y}\big)\mathbf{\Sigma}^{-1}.
\end{aligned}$$

Hence, taking the gradient of $f(\cdot)$ be zero leads to

$$\hat{\mathbf{B}}=\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

We also require checking the convexity of $f(\cdot)$. We only needs to consider the quadratic term (the first term) in equation (28), that is

$$\begin{aligned}
g(\mathbf{B})=&\mathrm{tr}\big(\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\mathbf{X}^\top\big)\\
=&\mathrm{tr}\big(\mathbf{\Sigma}^{-1/2}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1/2}\big)\\
=&\mathrm{tr}\big(\mathbf{C}^\top\mathbf{X}^\top\mathbf{X}\mathbf{C}\big)\\
=&\sum_{i=1}^{q}\mathbf{c}_i^\top\mathbf{X}^\top\mathbf{X}\mathbf{c}_i\\
=&\mathrm{vec}(\mathbf{C})^\top(\mathbf{I}_q\otimes(\mathbf{X}^\top\mathbf{X}))\mathrm{vec}(\mathbf{C})\triangleq h(\mathbf{C}),
\end{aligned}\tag{32}$$

where

$$\mathbf{C}=\mathbf{B}\mathbf{\Sigma}^{-1/2}=\begin{bmatrix}\mathbf{c}_1 & \cdots & \mathbf{c}_q\end{bmatrix}\in\mathbb{R}^{p\times q},\quad \mathrm{vec}(\mathbf{C})=\begin{bmatrix}\mathbf{c}_1\\ \vdots\\ \mathbf{c}_q\end{bmatrix}\in\mathbb{R}^{pq}$$

and

$$\mathbf{I}_q\otimes(\mathbf{X}^\top\mathbf{X})=\begin{bmatrix}\mathbf{X}^\top\mathbf{X} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}^\top\mathbf{X}\end{bmatrix}\in\mathbb{R}^{pq\times pq}.$$

This document is organized by Luo Luo.
Please do not distribute.

Since the assumptions of $\mathbf{X} \in \mathbb{R}^{N\times p}$ is full rank and $N > p$ indicate $\mathbf{X}^\top\mathbf{X} \succ \mathbf{0}$, we have

$$\mathbf{I}_q \otimes (\mathbf{X}^\top\mathbf{X}) \succ \mathbf{0},$$

which means $h(\mathbf{C})$ is convex with respect to $\mathbf{C}$. The definition of $\mathbf{C}$ means

$$g(\mathbf{B}) = h(\mathbf{C}) = h(\mathbf{B}\mathbf{\Sigma}^{-1/2}).$$

is also convex, This is because of For any $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{p\times q}$ and $\alpha \in [0, 1]$, we have

$$\begin{aligned}
&g(\alpha\mathbf{B}_1 + (1-\alpha)\mathbf{B}_2) \\
=&h(\alpha\mathbf{B}_1\mathbf{\Sigma}^{-1/2} + (1-\alpha)\mathbf{B}_2\mathbf{\Sigma}^{-1/2}) \\
\leq&\alpha h(\mathbf{B}_1\mathbf{\Sigma}^{-1/2}) + (1-\alpha) f(\mathbf{B}_2\mathbf{\Sigma}^{-1/2}) \\
=&\alpha g(\mathbf{B}_1) + (1-\alpha) g(\mathbf{B}_2),
\end{aligned}$$

where the inequality is based on the convexity of $h(\cdot)$. Now we conclude $\hat{\mathbf{B}}$ is the minimizer.

Then we consider minimize $L(\mathbf{B}, \mathbf{\Sigma})$ with respect to $\mathbf{\Sigma}$. We also present the details if you are not familiar with the derivation of maximum likelihood estimator of multivariate normal distribution. We denote

$$\mathbf{G} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{Y}.$$

Taking $\mathbf{B} = \hat{\mathbf{B}}$, the term in exponential of likelihood function is

$$\begin{aligned}
&\mathrm{tr}\big((\mathbf{X}\hat{\mathbf{B}} - \mathbf{Y})\mathbf{\Sigma}^{-1}(\hat{\mathbf{B}}^\top\mathbf{X}^\top - \mathbf{Y}^\top)\big) \\
=&\mathrm{tr}\big((\hat{\mathbf{B}}^\top\mathbf{X}^\top - \mathbf{Y}^\top)(\mathbf{X}\hat{\mathbf{B}} - \mathbf{Y})\mathbf{\Sigma}^{-1}\big) \\
=&\mathrm{tr}\big(\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{Y}^\top)(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} - \mathbf{Y})\mathbf{\Sigma}^{-1}\big) \\
=&\mathrm{tr}\big(\mathbf{Y}^\top(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{I})(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{I})\mathbf{Y})\mathbf{\Sigma}^{-1}\big) \\
=&\mathrm{tr}\big(\mathbf{G}\mathbf{\Sigma}^{-1}\big).
\end{aligned}$$

Recall that the matrix $\mathbf{G}$ plays the role of $\mathbf{A}$ in maximum likelihood estimation of multivariate normal distribution. Hence, the maximum likelihood estimator of $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{Y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{Y}.$$

For details, the minimization problem with respect to $\mathbf{\Sigma}$ can be formulated by taking logarithm on L and omitting the constant, which leads to

$$\min_{\mathbf{\Sigma}\in\mathbb{R}^p_{++}} \frac{N}{2}\ln(\det(\mathbf{\Sigma})) + \frac{1}{2}\mathrm{tr}(\mathbf{G}\mathbf{\Sigma}^{-1}).$$

Take $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$, then it can be written as

$$\min_{\mathbf{\Psi}\in\mathbb{R}^p_{++}} g(\mathbf{\Psi}) = -\frac{N}{2}\ln(\det(\mathbf{\Psi})) + \frac{1}{2}\mathrm{tr}(\mathbf{G}\mathbf{\Psi}).$$

We have shown $g(\mathbf{\Psi})$ is convex in Theorem 4.2. Hence, taking

$$\nabla g(\mathbf{\Psi}) = -\frac{N}{2}\mathbf{\Psi}^{-1} + \frac{1}{2}\mathbf{G} = \mathbf{0}$$

leads to

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Psi}}^{-1} = \frac{1}{N}\mathbf{G} = \frac{1}{N}\mathbf{Y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{Y}.$$

This document is organized by Luo Luo.
Please do not distribute.

Remark 10.2. *We can ignore the statistical model and directly formulate multivariate linear regression as*

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \|\mathbf{X}\mathbf{B}-\mathbf{Y}\|_F^2 .$$

The statistical model corresponds to the optimization problem with weighted norm

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \|\mathbf{X}\mathbf{B}-\mathbf{Y}\|_{\mathbf{\Sigma}^{-1}}^2 ,$$

where

$$\|\mathbf{X}\mathbf{B}-\mathbf{Y}\|_{\mathbf{\Sigma}^{-1}}^2 = \left\|(\mathbf{X}\mathbf{B}-\mathbf{Y})\mathbf{\Sigma}^{-1/2}\right\|_F^2 = \operatorname{tr}\left((\mathbf{X}\mathbf{B}-\mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{X}^\top-\mathbf{Y}^\top)\right).$$

Theorem 10.1. *Following the notations is this section, we additionally write*

$$\mathbf{B} = \begin{bmatrix}\boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_q\end{bmatrix} \in \mathbb{R}^{q\times p} \qquad \text{and} \qquad \hat{\mathbf{B}} = \begin{bmatrix}\hat{\boldsymbol{\beta}}_1 & \cdots & \hat{\boldsymbol{\beta}}_q\end{bmatrix} \in \mathbb{R}^{q\times p}.$$

Then the joint distribution of $\hat{\boldsymbol{\beta}}_1,\dots,\hat{\boldsymbol{\beta}}_N$ is normal and we have

1. $\mathbb{E}[\hat{\boldsymbol{\beta}}_i] = \boldsymbol{\beta}_i$;
2. $\operatorname{Cov}[\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}_j] = \sigma_{ij}(\mathbf{X}^\top\mathbf{X})^{-1}$;
3. $\hat{\mathbf{\Sigma}} \sim \mathcal{W}_q\left(\frac{1}{N}\mathbf{\Sigma}, N-p\right)$.

Proof. We can write $\mathbf{Y} = \mathbf{X}\mathbf{B}+\mathbf{E}$ and

$$\hat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{B}+\mathbf{E}) = \mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}.$$

This means the random component of each $\hat{b}_i$ is linear combination of entries of $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} \mathcal{N}_q(\mathbf{0},\mathbf{\Sigma})$. Hence, each $\hat{b}_i$ has normal distribution, and the joint distribution of $\hat{\boldsymbol{\beta}}_1,\dots,\hat{\boldsymbol{\beta}}_N$ is also normal.
Part 1: Since $\mathbb{E}[\mathbf{E}] = \mathbf{0}$, we have $\mathbb{E}[\hat{\mathbf{B}}] = \mathbf{B}$, that is $\mathbb{E}[\hat{\boldsymbol{\beta}}_i] = \boldsymbol{\beta}_i$.
Part 2: We can write

$$\mathbf{E} = \begin{bmatrix}\boldsymbol{\epsilon}_{(1)} & \cdots & \boldsymbol{\epsilon}_{(q)}\end{bmatrix} = \begin{bmatrix}\epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1q}\\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2q}\\ \vdots & \vdots & \ddots & \vdots\\ \epsilon_{N1} & \epsilon_{N2} & \dots & \epsilon_{Nq}\end{bmatrix} \in \mathbb{R}^{N\times q} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}_{(i)},$$

then we have

$$\begin{aligned}\operatorname{Cov}[\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}_j] =& \mathbb{E}\left[(\hat{\boldsymbol{\beta}}_i - \mathbb{E}[\hat{\boldsymbol{\beta}}_i])(\hat{\boldsymbol{\beta}}_j - \mathbb{E}[\hat{\boldsymbol{\beta}}_j])^\top\right]\\ =& \mathbb{E}\left[(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top\right]\\ =& \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}_{(i)}\boldsymbol{\epsilon}_{(j)}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\right].\end{aligned}$$

Note that the rows of $\mathbf{E} \in \mathbb{R}^{N\times q}$ are mutually independent, hence we have

$$\mathbb{E}[\epsilon_{ki}\epsilon_{lj}] = \begin{cases}\sigma_{ij}, & \text{if } k = l,\\ 0, & \text{if } k \neq l,\end{cases} \qquad \text{and} \qquad \mathbb{E}\left[\boldsymbol{\epsilon}_{(i)}\boldsymbol{\epsilon}_{(j)}^\top\right] = \begin{bmatrix}\epsilon_{1i}\epsilon_{1j} & \epsilon_{1i}\epsilon_{2j} & \dots & \epsilon_{1i}\epsilon_{Nj}\\ \epsilon_{2i}\epsilon_{1j} & \epsilon_{2i}\epsilon_{2j} & \dots & \epsilon_{2i}\epsilon_{Nj}\\ \vdots & \vdots & \ddots & \vdots\\ \epsilon_{Ni}\epsilon_{1j} & \epsilon_{Ni}\epsilon_{2j} & \dots & \epsilon_{Ni}\epsilon_{Nj}\end{bmatrix} = \sigma_{ij}\mathbf{I}_N.$$

Thus, we achieve

$$\operatorname{Cov}[\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}_j] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\left[\boldsymbol{\epsilon}_{(i)}\boldsymbol{\epsilon}_{(j)}^\top\right]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma_{ij}(\mathbf{X}^\top\mathbf{X})^{-1}.$$

This document is organized by Luo Luo.
Please do not distribute.

Part 3: Consider that

$$\begin{aligned}\hat{\mathbf{\Sigma}} =& \frac{1}{N}\mathbf{Y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{Y} \\ =& \frac{1}{N}(\mathbf{X}\mathbf{B} + \mathbf{E})^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)(\mathbf{X}\mathbf{B} + \mathbf{E})\end{aligned}$$

and

$$(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{X}\mathbf{B} = \mathbf{X}\mathbf{B} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{B} = \mathbf{0},$$

then we have

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{E}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{E}.$$

Consider that each row of $\mathbf{E}$ (that is $\boldsymbol{\epsilon}_i$) has distribution $\mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma})$ independently and $\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a projection matrix with rank-$(N - q)$, then Lemma 7.2 means

$$N\hat{\mathbf{\Sigma}} \sim \mathcal{W}_q\left(\mathbf{\Sigma}, N - p\right) \qquad \text{and} \qquad \hat{\mathbf{\Sigma}} \sim \mathcal{W}_q\left(\frac{1}{N}\mathbf{\Sigma}, N - p\right).$$

□

Remark 10.3. *We define*

$$\operatorname{vec}(\mathbf{B}^\top) = \begin{bmatrix}\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_q\end{bmatrix} \qquad \textit{and} \qquad \operatorname{vec}(\hat{\mathbf{B}}^\top) = \begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_q\end{bmatrix},$$

then we can write

$$\operatorname{vec}(\hat{\mathbf{B}}^\top) \sim \mathcal{N}_{pq}(\operatorname{vec}(\mathbf{B}^\top), \mathbf{\Sigma} \otimes (\mathbf{X}^\top\mathbf{X})^{-1}),$$

where

$$\mathbf{\Sigma} \otimes (\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix}\sigma_{11}(\mathbf{X}^\top\mathbf{X})^{-1} & \dots & \sigma_{1p}(\mathbf{X}^\top\mathbf{X})^{-1} \\ \vdots & \ddots & \vdots \\ \sigma_{p1}(\mathbf{X}^\top\mathbf{X})^{-1} & \dots & \sigma_{pp}(\mathbf{X}^\top\mathbf{X})^{-1}\end{bmatrix}.$$

More specifically, the random vector

$$\begin{bmatrix}\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_N\end{bmatrix} \textit{ has covariance } \begin{bmatrix}\operatorname{Cov}[\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_1] & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2] & \cdots & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_N] \\ \operatorname{Cov}[\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_1] & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_2] & \cdots & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_N] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}[\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\beta}}_1] & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\beta}}_2] & \cdots & \operatorname{Cov}[\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\beta}}_N]\end{bmatrix} = \begin{bmatrix}\sigma_{11}(\mathbf{X}^\top\mathbf{X})^{-1} & \dots & \sigma_{1p}(\mathbf{X}^\top\mathbf{X})^{-1} \\ \vdots & \ddots & \vdots \\ \sigma_{p1}(\mathbf{X}^\top\mathbf{X})^{-1} & \dots & \sigma_{pp}(\mathbf{X}^\top\mathbf{X})^{-1}\end{bmatrix}.$$

Remark 10.4. *The distribution $\hat{\mathbf{\Sigma}} \sim \mathcal{W}_q\left(\frac{1}{N}\mathbf{\Sigma}, N - p\right)$ means*

$$\mathbb{E}\big[\hat{\mathbf{\Sigma}}\big] = \frac{N - p}{N}\mathbf{\Sigma},$$

which is biased. Recall that the maximum likelihood estimator of covariance matrix for multivariate normal distribution is also biased.

This document is organized by Luo Luo.
Please do not distribute.

Bayesian Multivariate Linear Regression: We can additionally suppose each b_{ij} independently follows

$$b_{ij} \sim \mathcal{N}(0, \tau^2),$$

then the likelihood function is

$$\begin{aligned} L(\mathbf{B}, \mathbf{\Sigma}) =& \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{B}^\top \mathbf{x}_i - \mathbf{y}_i)^\top \mathbf{\Sigma}^{-1} (\mathbf{B}^\top \mathbf{x}_i - \mathbf{y}_i)\right) \cdot \prod_{i=1}^{p}\prod_{j=1}^{q} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{b_{ij}^2}{2\tau^2}\right) \\ \propto& \frac{1}{(\det(\mathbf{\Sigma}))^{N/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{XB}-\mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{XB}-\mathbf{Y})^\top\right) - \frac{1}{2\tau^2}\left\|\mathbf{B}\right\|_F^2\right). \end{aligned}$$

Minimizing function $L(\mathbf{B}, \mathbf{\Sigma})$ with respect to $\mathbf{B}$, only needs to solve

$$f_\tau(\mathbf{B}) = \mathrm{tr}\left((\mathbf{XB}-\mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{XB}-\mathbf{Y})^\top\right) + \frac{1}{\tau^2}\left\|\mathbf{B}\right\|_F^2.$$

The analysis on $f(\mathbf{B})$ indicates $f_\tau(\mathbf{B})$ is convex and

$$\nabla f_\tau(\mathbf{B}) = 2\big(\mathbf{X}^\top\mathbf{XB} - \mathbf{X}^\top\mathbf{Y}\big)\mathbf{\Sigma}^{-1} + \frac{2}{\tau^2}\mathbf{B}.$$

Taking $\nabla f_\tau(\mathbf{B}) = \mathbf{0}$ leads to

$$\mathbf{0} = \big(\tau^2\mathbf{X}^\top\mathbf{XB} + \mathbf{B\Sigma} - \tau^2\mathbf{X}^\top\mathbf{Y}\big)\mathbf{\Sigma}^{-1},$$

which is a Sylvester equation (see details on Wikipedia). We have

$$\begin{aligned} & \tau^2\mathbf{X}^\top\mathbf{XB} + \mathbf{B\Sigma} = \tau^2\mathbf{X}^\top\mathbf{Y} \\ \iff\quad & (\mathbf{I}_q \otimes \tau^2\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma} \otimes \mathbf{I}_p)\mathrm{vec}(\mathbf{B}) = \mathrm{vec}(\tau^2\mathbf{X}^\top\mathbf{Y}), \end{aligned}$$

which means

$$\mathrm{vec}(\hat{\mathbf{B}}) = (\mathbf{I}_q \otimes \tau^2\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma} \otimes \mathbf{I}_p)^{-1}\mathrm{vec}(\tau^2\mathbf{X}^\top\mathbf{Y}).$$

We do not like this model since the solution depends on $\mathbf{\Sigma}$.

We prefer to consider the prior distribution as

$$\boldsymbol{\beta}_{(i)} \overset{\text{i.i.d}}{\sim} \mathcal{N}_q(\mathbf{0}, \tau^2\mathbf{\Sigma}),$$

where

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_{(1)}^\top \\ \vdots \\ \boldsymbol{\beta}_{(p)}^\top \end{bmatrix} \in \mathbb{R}^{p\times q},$$

that is

$$\begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \vdots \\ \boldsymbol{\beta}_{(p)} \end{bmatrix} \sim \mathcal{N}_{pq}\left(\mathbf{0}, \mathbf{I} \otimes \mathbf{\Sigma}\right),$$

where

$$\mathbf{I} \otimes (\tau^2\mathbf{\Sigma}) = \begin{bmatrix} \tau^2\mathbf{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \tau^2\mathbf{\Sigma} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \tau^2\mathbf{\Sigma} \end{bmatrix}.$$

This document is organized by Luo Luo.
Please do not distribute.

Then the likelihood function is

$$\begin{aligned}
&L(\mathbf{B},\mathbf{\Sigma})\\
=&\prod_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^p\det(\mathbf{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{B}^\top\mathbf{x}_i-\mathbf{y}_i)^\top\mathbf{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{x}_i-\mathbf{y}_i)\right)\\
&\cdot\prod_{j=1}^{p}\frac{1}{\sqrt{(2\pi)^q\det(\mathbf{\Sigma})}}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}_{(j)}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\beta}_{(j)}\right)\\
\propto&\frac{1}{(\det(\mathbf{\Sigma}))^{N/2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{XB}-\mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{XB}-\mathbf{Y})^\top\right)-\frac{1}{2\tau^2}\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top\right).
\end{aligned}$$

Minimizing function $L(\mathbf{B},\mathbf{\Sigma})$ with respect to $\mathbf{B}$, only needs to solve

$$f_\lambda(\mathbf{B})=\mathrm{tr}\left((\mathbf{XB}-\mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{XB}-\mathbf{Y})^\top\right)+\lambda\mathrm{tr}(\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}^\top),$$

where $\lambda=1/\tau^2$.

The analysis on $f(\mathbf{B})$ indicates $f_\tau(\mathbf{B})$ is convex and

$$\nabla f_\lambda(\mathbf{B})=2\big(\mathbf{X}^\top\mathbf{XB}-\mathbf{X}^\top\mathbf{Y}\big)\mathbf{\Sigma}^{-1}+2\lambda\mathbf{B}\mathbf{\Sigma}^{-1}.$$

Taking $\nabla f_\lambda(\mathbf{B})=\mathbf{0}$ leads to

$$\hat{\mathbf{B}}_\lambda=(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

We can also obtain

$$\begin{aligned}
N\hat{\mathbf{\Sigma}}_\lambda=&(\mathbf{X}\hat{\mathbf{B}}_\lambda-\mathbf{Y})^\top(\mathbf{X}\hat{\mathbf{B}}_\lambda-\mathbf{Y})+\lambda\hat{\mathbf{B}}_\lambda^\top\hat{\mathbf{B}}_\lambda\\
=&\hat{\mathbf{B}}_\lambda^\top\mathbf{X}^\top\mathbf{X}\hat{\mathbf{B}}_\lambda-\mathbf{Y}^\top\mathbf{X}\hat{\mathbf{B}}_\lambda-\hat{\mathbf{B}}_\lambda^\top\mathbf{X}^\top\mathbf{Y}+\mathbf{Y}^\top\mathbf{Y}+\lambda\hat{\mathbf{B}}_\lambda^\top\hat{\mathbf{B}}_\lambda\\
=&\hat{\mathbf{B}}_\lambda^\top(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})\hat{\mathbf{B}}_\lambda-2\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}+\mathbf{Y}^\top\mathbf{Y}\\
=&\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}-2\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}+\mathbf{Y}^\top\mathbf{Y}\\
=&\mathbf{Y}^\top(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top)\mathbf{Y}
\end{aligned}$$

by following the analysis of ordinary multivariate linear regression, that is

$$\hat{\mathbf{\Sigma}}_\lambda=\frac{1}{N}\mathbf{Y}^\top(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top)\mathbf{Y}.$$

Remark 10.5. *We can ignore the prior distribution on* $\mathbf{B}$ *and consider the form of* $\hat{\mathbf{B}}_\lambda$*, then*

$$\begin{aligned}
\mathbb{E}\big[\hat{\mathbf{B}}_\lambda\big]=&\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}]\\
=&\mathbb{E}[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{XB}+\mathbf{E})]\\
=&(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{XB},
\end{aligned}$$

which is biased.

Remark 10.6. *Let* $\hat{\boldsymbol{\beta}}_{\lambda,i}$ *be the i-th column of* $\hat{\mathbf{B}}_\lambda$*, then we have*

$$\begin{aligned}
&\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\lambda,i},\hat{\boldsymbol{\beta}}_{\lambda,j}]\\
=&\mathbb{E}\big[(\hat{\boldsymbol{\beta}}_{\lambda,i}-\mathbb{E}[\hat{\boldsymbol{\beta}}_{\lambda,i}])(\hat{\boldsymbol{\beta}}_{\lambda,j}-\mathbb{E}[\hat{\boldsymbol{\beta}}_{\lambda,j}])^\top\big]\\
=&\mathbb{E}\big[(\hat{\boldsymbol{\beta}}_{\lambda,i}-\boldsymbol{\beta}_{\lambda,i})(\hat{\boldsymbol{\beta}}_{\lambda,j}-\boldsymbol{\beta}_{\lambda,j})^\top\big]\\
=&\mathbb{E}\big[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}_{(i)}\boldsymbol{\epsilon}_{(j)}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\big].
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Recall that $\mathbb{E}\big[\boldsymbol{\epsilon}_{(i)}\boldsymbol{\epsilon}_{(j)}^\top\big] = \sigma_{ij}\mathbf{I}_N$*, then we have*

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\lambda,i}, \hat{\boldsymbol{\beta}}_{\lambda,j}] = \sigma_{ij}\mathbb{E}\big[(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\big].$$

Therefore, we achieve

$$\mathrm{Cov}[\mathrm{vec}(\mathbf{B}_\lambda^\top)] = \boldsymbol{\Sigma}\otimes(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}.$$

Since we have shown that $(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1} \preceq (\mathbf{X}^\top\mathbf{X})^{-1}$ *in Remark 4.7, we have*

$$\mathrm{Cov}[\mathrm{vec}(\mathbf{B}_\lambda^\top)] \preceq \mathrm{Cov}[\mathrm{vec}(\mathbf{B}^\top)].$$

Remark 10.7. *It is natural to involve inverted Wishart distribution as the prior distribution of* $\boldsymbol{\Sigma}$*. We can try to do it as an exercise.*

11 Principal Components Analysis

Let p-dimensional random vector $\mathbf{x}$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{u}\in\mathbb{R}^p$ such that $\|\mathbf{u}\|_2 = 1$ and maximizes the variance of $\mathbf{u}^\top\mathbf{x}$, then we call $\mathbf{u}^\top\mathbf{x}$ is the first principle component of $\mathbf{x}$.

Theorem 11.1. *Let* $\mathbf{x}$ *be a* p*-dimensional random vector with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}\succ\mathbf{0}$*. Let* $\mathbf{u}_1\in\mathbb{R}^p$ *with* $\|\mathbf{u}_1\|_2 = 1$ *and maximizing the variance of* $\mathbf{u}_1^\top\mathbf{x}$ *must satisfy*

$$(\boldsymbol{\Sigma}-\lambda_1\mathbf{I})\mathbf{u}_1 = \mathbf{0},$$

where λ_1 *is the largest root of*

$$\det(\boldsymbol{\Sigma}-\lambda\mathbf{I}) = 0.$$

Proof. Consider that $\mathrm{Var}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}$, then we have

$$\mathrm{Var}[\mathbf{u}_1^\top\mathbf{x}] = \mathbb{E}[\mathbf{u}_1^\top\mathbf{x}(\mathbf{u}_1^\top\mathbf{x})] = \mathbb{E}[\mathbf{u}_1^\top\mathbf{x}\mathbf{x}^\top\mathbf{u}_1] = \mathbf{u}_1^\top\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{u}_1 = \mathbf{u}_1^\top\boldsymbol{\Sigma}\mathbf{u}_1.$$

Hence, we can find $\mathbf{u}_1\in\mathbb{R}^p$ by solving the optimization problem

$$\max_{\|\mathbf{u}\|_2^2=1} \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}.$$

Define the Lagrangian function

$$L(\mathbf{u},\lambda) = \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u} - \lambda(\mathbf{u}^\top\mathbf{u}-1),$$

where λ is a Lagrangian multiplier. The optimal condition implies

$$\mathbf{0} = \frac{\partial L(\mathbf{u},\lambda)}{\partial\mathbf{u}} = 2\boldsymbol{\Sigma}\mathbf{u} - 2\lambda\mathbf{u},$$

that is $(\boldsymbol{\Sigma}-\lambda\mathbf{I})\mathbf{u} = \mathbf{0}$. The constraint $\|\mathbf{u}\|_2 = 1$ means $\boldsymbol{\Sigma}-\lambda\mathbf{I}$ is singular. Then λ must satisfy

$$\det(\boldsymbol{\Sigma}-\lambda\mathbf{I}) = 0.$$

This implies λ and $\mathbf{u}$ must be the eigenvalue of $\boldsymbol{\Sigma}$ and corresponding eigenvector respectively, then

$$\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u} = \mathbf{u}^\top(\lambda\mathbf{u}) = \lambda.$$

Hence, maximizing the objective corresponds to taking λ_1 be the largest eigenvalue of $\boldsymbol{\Sigma}$. □

Remark 11.1. *We call random variable* $y_1 = \mathbf{u}_1^\top\mathbf{x}$ *as the first principle component of* $\mathbf{x}$*.*

This document is organized by Luo Luo.
Please do not distribute.

Remark 11.2. *Theorem 11.1 means* $\mathbf{u}_1$ *is the eigenvector of* $\boldsymbol{\Sigma}$ *associate with the largest eigenvalue. The variance of* y_1 *can be written as*

$$\mathrm{Var}[y_1]=\mathrm{Var}[\mathbf{u}_1^\top\mathbf{x}]=\mathbf{u}_1^\top(\boldsymbol{\Sigma}\mathbf{u}_1)=\mathbf{u}_1^\top(\lambda_1\mathbf{u}_1)=\lambda_1\mathbf{u}_1^\top\mathbf{u}_1=\lambda_1,$$

which is the largest eigenvalue of $\boldsymbol{\Sigma}$*.*

Remark 11.3. *For the second principle components* $y_2=\mathbf{u}_2^\top\mathbf{x}$*, we determine* $\mathbf{u}_2$ *by maximizing the variance of* y_2 *under the constraints* $\|\mathbf{u}_2\|_2=1$ *and* y_2 *be uncorrelated with* y_1*. The uncorrelated condition implies*

$$0=\mathrm{Cov}[y_2,y_1]=\mathbb{E}[y_2y_1]=\mathbb{E}\big[\mathbf{u}_2^\top\mathbf{x}\,\mathbf{u}_1{}^\top\mathbf{x}\big]=\mathbb{E}\big[\mathbf{u}_2^\top\mathbf{x}\mathbf{x}^\top\mathbf{u}_1\big]=\mathbf{u}_2^\top\boldsymbol{\Sigma}\mathbf{u}_1=\mathbf{u}_2^\top(\lambda_1\mathbf{u}_1)=\lambda_1\mathbf{u}_2^\top\mathbf{u}_1,$$

which means

$$\mathbf{u}_2^\top\mathbf{u}_1=0.$$

Hence, we can find $\mathbf{u}_2$ *by solving the optimization problem*

$$\max_{\mathbf{u}\in\mathbb{R}^p}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u},\qquad s.t\quad \|\mathbf{u}\|_2^2=1\quad and\quad \mathbf{u}^\top\mathbf{u}_1=0.$$

Remark 11.4. *For the* k*-th principle component* $y_k=\mathbf{u}_k^\top\mathbf{x}$*, we determine* $\mathbf{u}_k$ *by maximizing the variance of* y_k *under the constraints* $\|\mathbf{u}_k\|_2=1$ *and* y_k *be uncorrelated with* $y_1,\dots,y_{k-1}$*. We can show* $\mathbf{u}_k\in\mathbb{R}^p$ *is the eigenvector corresponds to the* k*-th largest eigenvalue of* $\boldsymbol{\Sigma}\in\mathbb{R}^{p\times p}$ *by induction.*

Suppose we have achieved $\mathbf{u}_1,\dots,\mathbf{u}_{k-1}\in\mathbb{R}^p$ *and* $\lambda_1,\dots,\lambda_{k-1}\in\mathbb{R}$ *such that* λ_i *and* $\mathbf{u}_i$ *are the* i*-th largest eigenvalue and the corresponding eigenvector for* $i=1,\dots,k-1$*, that is*

$$\boldsymbol{\Sigma}\mathbf{u}_i=\lambda_i\mathbf{u}_i\qquad for\quad i=1,\dots,k-1.$$

Similarly to above analysis, the uncorrelated condition implies

$$0=\mathbb{E}[\mathbf{u}_k^\top\mathbf{x}\mathbf{x}^\top\mathbf{u}_i]=\mathbf{u}_k^\top\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{u}_i=\mathbf{u}_k^\top\boldsymbol{\Sigma}\mathbf{u}_i=\lambda\mathbf{u}_k^\top\mathbf{u}_i\qquad\Longrightarrow\qquad\mathbf{u}_k^\top\mathbf{u}_i=0.$$

for $i=1,\dots,k-1$*. Hence, we can find* $\mathbf{u}_k$ *by solving the optimization problem*

$$\max_{\mathbf{u}\in\mathbb{R}^p}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u},\qquad s.t\quad \|\mathbf{u}\|_2^2=1\quad and\quad \mathbf{u}^\top\mathbf{u}_1=\dots=\mathbf{u}^\top\mathbf{u}_{k-1}=0.$$

Define the Lagrangian function

$$L_k(\mathbf{u},\lambda,\boldsymbol{\nu})=\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}-\lambda(\mathbf{u}^\top\mathbf{u}-1)-2\sum_{i=1}^{k-1}\nu_i\mathbf{u}^\top\mathbf{u}_i$$

where λ*,* $\nu_1,\dots,\nu_{k-1}$ *are Lagrangian multipliers. The optimal condition implies*

$$\mathbf{0}=\frac{\partial L_k(\mathbf{u},\lambda)}{\partial\mathbf{u}}=2\boldsymbol{\Sigma}\mathbf{u}-2\lambda\mathbf{u}-2\sum_{i=1}^{k-1}\nu_i\mathbf{u}_i.$$

Multiplying on the left by $\mathbf{u}_j^\top$ *for* $j=1,\dots,k-1$*, we have*

$$\begin{aligned}\mathbf{0}=&2\mathbf{u}_j{}^\top\boldsymbol{\Sigma}\mathbf{u}-2\lambda\mathbf{u}_j^\top\mathbf{u}-2\sum_{i=1}^{k-1}\nu_i\mathbf{u}_j^\top\mathbf{u}_i\\=&2\lambda_j\mathbf{u}_j{}^\top\mathbf{u}-2\lambda\mathbf{u}_j{}^\top\mathbf{u}-2\nu_j\mathbf{u}_j^\top\mathbf{u}_j\\=&-2\nu_j\mathbf{u}_j^\top\mathbf{u}_j.\end{aligned}$$

Therefore we have $\nu_j=0$ *for* $j=1,\dots,k-1$ *and the optimal condition means*

$$(\boldsymbol{\Sigma}-\lambda\mathbf{I})\mathbf{u}=\mathbf{0}.$$

This implies λ *and* $\mathbf{u}$ *must be the eigenvalue of* $\boldsymbol{\Sigma}$ *and corresponding eigenvector respectively. The orthogonal constraint* $\mathbf{u}^\top\mathbf{u}_j=0$ *for* $j=1,\dots,k-1$ *means* λ *and* $\mathbf{u}$ *must be the* k*-largest eigenvalue of* $\boldsymbol{\Sigma}$ *and corresponding eigenvector.*

This document is organized by Luo Luo.
Please do not distribute.

The above analysis indicates

$$\mathbf{U}_k = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{bmatrix} \in \mathbb{R}^{p\times k} \qquad \text{and} \qquad \mathbf{\Lambda}_k = \begin{bmatrix}\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k\end{bmatrix} \in \mathbb{R}^{k\times k}$$

contains the top-k eigenvectors and eigenvalues pairs of $\mathbf{\Sigma}$, that is

$$\mathbf{\Sigma}\mathbf{U}_k = \mathbf{U}_k\mathbf{\Lambda} \qquad \text{with} \qquad \mathbf{U}_k^\top\mathbf{U}_k = \mathbf{I}.$$

Remark 11.5. *Taking $k = p$, we obtain*

$$\mathbf{y} = \begin{bmatrix}y_1 \\ \vdots \\ y_p\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1^\top\mathbf{x} \\ \vdots \\ \mathbf{u}_p^\top\mathbf{x}\end{bmatrix} = \mathbf{U}^\top\mathbf{x},$$

then we have

$$\mathrm{Cov}[\mathbf{y}] = \mathbb{E}[\mathbf{U}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}] = \mathbf{U}^\top\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{U} = \mathbf{U}^\top\mathbf{\Sigma}\mathbf{U} = \mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U} = \mathbf{\Lambda}.$$

We take $k \ll p$ for dimensionality reduction. We keep $\mathbf{U}_k \in \mathbb{R}^{p\times k}$ and transform $\mathbf{x} \in \mathbb{R}^p$ to $\mathbf{U}_k^\top\mathbf{x} \in \mathbb{R}^k$. Then we can recover the information of $\mathbf{x}$ by

$$\hat{\mathbf{x}} = \mathbf{U}_k(\mathbf{U}_k^\top\mathbf{x}) \in \mathbb{R}^p.$$

We partition $\mathbf{U}$ and $\mathbf{\Lambda}$ into

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_k & \mathbf{U}_\perp\end{bmatrix} \qquad \text{and} \qquad \mathbf{\Lambda} = \begin{bmatrix}\mathbf{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\perp\end{bmatrix},$$

then

$$\begin{aligned}\mathrm{Cov}[\hat{\mathbf{x}}] =& \mathrm{Cov}[\mathbf{U}_k\mathbf{U}_k^\top\mathbf{x}] = \mathbb{E}[\mathbf{U}_k\mathbf{U}_k^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_k\mathbf{U}_k^\top] = \mathbf{U}_k\mathbf{U}_k^\top\mathbf{\Sigma}\mathbf{U}_k\mathbf{U}_k^\top \\ =& \mathbf{U}_k\mathbf{U}_k^\top\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{U}_k^\top \\ \mathbf{U}_\perp^\top\end{bmatrix}\mathbf{U}_k\mathbf{U}_k^\top \\ =& \mathbf{U}_k\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\mathbf{U}_k^\top \\ =& \begin{bmatrix}\mathbf{U}_k & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{U}_k^\top \\ \mathbf{0}\end{bmatrix} \\ =& \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{U}_k^\top,\end{aligned}$$

which is the best rank-k approximation of $\mathbf{\Sigma}$.

This document is organized by Luo Luo.
Please do not distribute.

PCA for sample covariance: Given observation $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^p$, we construct sample covariance

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top, \qquad \text{where } \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha.$$

Let spectral decomposition of $\mathbf{S}$ be

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U},$$

where $\mathbf{U} \in \mathbb{R}^{p\times p}$ is orthogonal and $\mathbf{\Lambda} \in \mathbb{R}^{p\times p}$ is diagonal. We write

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N\times p},$$

which results the sample principle components

$$\mathbf{Y} = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^\top \mathbf{U}_k \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^\top \mathbf{U}_k \end{bmatrix} = \mathbf{H}\mathbf{X}\mathbf{U}_k \in \mathbb{R}^{N\times k}, \qquad \text{where} \qquad \mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \in \mathbb{R}^{N\times N}.$$

Note that the matrix $\mathbf{S}$ can be written as ($\mathbf{H}$ is a projection matrix, i.e., $\mathbf{H}^2 = \mathbf{H}$), then we have

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}^\top\mathbf{H}\mathbf{H}\mathbf{X} = \frac{1}{N-1}\mathbf{X}^\top\mathbf{H}\mathbf{X} \in \mathbb{R}^{p\times p}.$$

We can keep $\bar{\mathbf{x}} \in \mathbb{R}^p$, $\mathbf{U}_k \in \mathbb{R}^{p\times k}$ and $\mathbf{Y} \in \mathbb{R}^{N\times k}$ to estimate $\mathbf{X} \in \mathbb{R}^{N\times p}$ by

$$\hat{\mathbf{X}} = \mathbf{Y}\mathbf{U}_k^\top + \mathbf{1}_N\bar{\mathbf{x}}^\top \in \mathbb{R}^{N\times p}.$$

Lemma 11.1. *For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$, then the matrices $\mathbf{AB} \in \mathbb{R}^{m\times m}$ and $\mathbf{BA} \in \mathbb{R}^{n\times n}$ have the same nonzero eigenvalues.*

Proof. Let $\lambda \neq 0$ such that $\mathbf{ABv} = \lambda\mathbf{v}$ for some $\mathbf{v} \neq \mathbf{0}$, then we have

$$\mathbf{BA}(\mathbf{Bv}) = \mathbf{B}(\mathbf{ABv}) = \mathbf{B}(\lambda\mathbf{v}) = \lambda(\mathbf{Bv}),$$

which means λ is also an eigenvalue of $\mathbf{BA}$. □

Principle Coordinate Analysis: We consider the case of $p \geq N \geq k$. We define matrix

$$\mathbf{T} = \frac{1}{N-1}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \in \mathbb{R}^{N\times N}$$

with spectral decomposition

$$\mathbf{T} = \mathbf{V}\mathbf{\Gamma}\mathbf{V}^\top,$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is orthogonal and $\mathbf{\Gamma} \in \mathbb{R}^{N\times N}$ is diagonal. We denote SVD of $\mathbf{HX}/\sqrt{N-1} \in \mathbb{R}^{N\times p}$ as

$$\frac{1}{\sqrt{N-1}}\mathbf{HX} = \mathbf{P}\mathbf{D}\mathbf{Q}^\top,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ is orthogonal, $\mathbf{D} \in \mathbb{R}^{N\times N}$ is diagonal and $\mathbf{Q} \in \mathbb{R}^{p\times N}$ is column orthogonal. We have

$$\begin{aligned} &\mathbf{T} = \mathbf{PDQ}^\top(\mathbf{PDQ}^\top)^\top = \mathbf{PD}^2\mathbf{P}^\top \qquad \Longrightarrow \qquad \mathbf{V} = \mathbf{P} \in \mathbb{R}^{N\times N}, \quad \mathbf{\Gamma} = \mathbf{D}^2 \in \mathbb{R}^{N\times N}, \\ \text{and} \quad &\mathbf{S} = (\mathbf{PDQ}^\top)^\top\mathbf{PDQ}^\top = \mathbf{QD}^2\mathbf{Q}^\top \qquad \Longrightarrow \qquad \mathbf{U} = \mathbf{Q} \in \mathbb{R}^{p\times N}, \quad \mathbf{\Lambda} = \mathbf{D}^2 \in \mathbb{R}^{N\times N}. \end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Then the matrix $\mathbf{Y} \in \mathbb{R}^{N\times k}$ can be written as

$$\begin{aligned}\mathbf{Y} =& \frac{1}{\sqrt{N-1}}\mathbf{H}\mathbf{X}\mathbf{U}_k = \mathbf{P}\mathbf{D}\mathbf{Q}^\top\mathbf{Q}_k \\ =& \begin{bmatrix}\mathbf{P}_k & \mathbf{P}_{N-k}\end{bmatrix}\begin{bmatrix}\mathbf{D}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{N-k}\end{bmatrix}\begin{bmatrix}\mathbf{U}_k^\top \\ \mathbf{U}_{N-k}^\top\end{bmatrix}\mathbf{U}_k \\ =& \mathbf{P}_k\mathbf{D}_k = \mathbf{V}_k\mathbf{\Gamma}_k^{1/2} \in \mathbb{R}^{N\times k},\end{aligned}$$

which can be obtained by SVD of $\mathbf{T} \in \mathbb{R}^{N\times N}$ or $\mathbf{H}\mathbf{X} \in \mathbb{R}^{N\times p}$. It is unnecessary to construct $\mathbf{S} \in \mathbb{R}^{p\times p}$.

Remark 11.6. *We consider the case of $p \gg N$, which requires require the complexity $\mathcal{O}(Np)$ to obtain the matrix $\mathbf{H}\mathbf{X} \in \mathbb{R}^{N\times p}$ and the complexity $\mathcal{O}(N^2p)$ to obtain the matrix $\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \in \mathbb{R}^{N\times N}$. However, we require the complexity $\mathcal{O}(Np^2)$ to obtain the matrix $\mathbf{S}$.*

Kernel PCA: Note that the matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ corresponds to outer product the centralized data, that is

$$\begin{aligned}\mathbf{X}^\top\mathbf{H}\mathbf{H}\mathbf{X} =& \begin{bmatrix}\mathbf{x}_1\mathbf{H} & \cdots & \mathbf{x}_N\mathbf{H}\end{bmatrix}\begin{bmatrix}\mathbf{H}\mathbf{x}_1^\top \\ \vdots \\ \mathbf{H}\mathbf{x}_N^\top\end{bmatrix} \\ =& \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha\mathbf{H})(\mathbf{x}_\alpha\mathbf{H})^\top \in \mathbb{R}^{p\times p}.\end{aligned}$$

The matrix $\mathbf{T} \in \mathbb{R}^{N\times N}$ corresponds to the inner product, since

$$\begin{aligned}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} =& \mathbf{H}\begin{bmatrix}\mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{bmatrix}\mathbf{H} \\ =& \mathbf{H}\begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \dots & \mathbf{x}_1^\top\mathbf{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_N^\top\mathbf{x}_1 & \mathbf{x}_N^\top\mathbf{x}_2 & \dots & \mathbf{x}_N^\top\mathbf{x}_N\end{bmatrix}\mathbf{H} \in \mathbb{R}^{N\times N}.\end{aligned}$$

We can map the sample $\mathbf{x}_\alpha \in \mathcal{X} \subseteq \mathbb{R}^p$ to the feature space $\mathcal{H} \subseteq \mathbb{R}^d$, that is

$$\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H},$$

and define the corresponding kernel function (inner product)

$$K(\mathbf{x}, \mathbf{y}) \triangleq \boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\phi}(\mathbf{y}).$$

Then we replace the matrix $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{N\times N}$ with the kernel matrix

$$\begin{aligned}\mathbf{K} =& \begin{bmatrix}\boldsymbol{\phi}(\mathbf{x}_1)^\top \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\phi}(\mathbf{x}_1) & \cdots & \boldsymbol{\phi}(\mathbf{x}_N)\end{bmatrix} \\ =& \begin{bmatrix}\boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_2) & \dots & \boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_2) & \dots & \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_N)\end{bmatrix} \\ =& \begin{bmatrix}K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \dots & K(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_N, \mathbf{x}_1) & K(\mathbf{x}_N, \mathbf{x}_2) & \dots & K(\mathbf{x}_N, \mathbf{x}_N)\end{bmatrix} \in \mathbb{R}^{N\times N}.\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

We can replace $\mathbf{T} \in \mathbb{R}^{N\times N}$ with

$$\mathbf{T}_K = \frac{1}{N-1}\mathbf{HKH}$$

and achieve kernel PCA by spectral decomposition on $\mathbf{T}_K$. There are some examples of kernel functions:

1. We define the polynomial kernel as

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + c)^d$$

for some $c \in \mathbb{R}$ and $d \in \mathbb{N}$. The dimensionality of corresponding $\boldsymbol{\phi}(\cdot)$ is $\mathcal{O}(p^d)$.

2. We define the Gaussian kernel (radial basis function kernel) as

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\right)$$

for some $\sigma > 0$. The dimension of corresponding $\boldsymbol{\phi}(\cdot)$ is $+\infty$.

The expression of corresponding $\boldsymbol{\phi}(\cdot)$ of above kernels can be found by Please refer to Link-1 and Link-2 for the detailed expressions of corresponding $\boldsymbol{\phi}(\cdot)$ for above kernels.

Remark 11.7. *For extreme large N, we can approximate $\mathbf{K} \in \mathbb{R}^{N\times N}$ by*

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12}^\top \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{11} \\ \mathbf{K}_{21} \end{bmatrix} \mathbf{K}_{11}^\dagger \begin{bmatrix} \mathbf{K}_{11}^\top & \mathbf{K}_{12}^\top \end{bmatrix}$$

or sample some columns of $\mathbf{K}$ to construct the estimator [16].

Remark 11.8. *The theory of kernel is also useful to analyze neural network with large layer width [9].*

12 Probabilistic PCA

Let $\mathbf{y}_1, \dots, \mathbf{y}_N \in \mathbb{R}^p$ be N independent observations and we have

$$\mathbf{y}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha,$$

where

$$\mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}) \qquad \text{and} \qquad \boldsymbol{\epsilon}_\alpha \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I})$$

are independent for some $\sigma^2 > 0$ and $N > q$. We target to estimate parameters

$$\mathbf{W} \in \mathbb{R}^{p\times q}, \qquad \boldsymbol{\mu} \in \mathbb{R}^p \qquad \text{and} \qquad \sigma \in (0, +\infty)$$

by maximum likelihood estimation for given $\mathbf{y}_1, \dots, \mathbf{y}_N$, where $q < p$.

MLE for PPCA: The independence between $\mathbf{x}_\alpha$ and $\boldsymbol{\epsilon}_\alpha$ implies

$$\mathbf{y}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}).$$

We construct the likelihood function

$$L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \prod_{\alpha=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top (\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_\alpha - \boldsymbol{\mu})\right),$$

This document is organized by Luo Luo.
Please do not distribute.

then we have

$$\ln L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) \propto -\frac{N}{2}\ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) - \frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_\alpha - \boldsymbol{\mu}).$$

The positive definiteness of $\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ means the maximum likelihood estimator of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{y}_\alpha.$$

Substituting $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ into $\ln L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2)$, we obtain

$$\begin{aligned}
\ln L(\bar{\mathbf{y}}, \mathbf{W}, \sigma^2) \propto & -\frac{N}{2}\ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) - \frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_\alpha - \bar{\mathbf{y}}) \\
= & -\frac{N}{2}\ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) - \frac{1}{2}\mathrm{tr}\left(\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\right) \\
= & -\ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) - \mathrm{tr}\big(\hat{\boldsymbol{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big),
\end{aligned}$$

where

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top.$$

Now we focus on minimizing the following function

$$f(\mathbf{W}, \sigma^2) = \ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) + \mathrm{tr}\big(\hat{\boldsymbol{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big).$$

The gradient with respect to $\mathbf{W} \in \mathbb{R}^{p\times q}$ is

$$\begin{aligned}
\frac{\partial f(\mathbf{W}, \sigma^2)}{\partial \mathbf{W}} =& 2(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W} - 2(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\hat{\boldsymbol{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W} \\
=& 2\mathbf{C}^{-1}\mathbf{W} - 2\mathbf{C}^{-1}\hat{\boldsymbol{\Sigma}}\mathbf{C}^{-1}\mathbf{W},
\end{aligned}$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$. We present how to achieve the gradient in the following remarks.

Remark 12.1. *Let $h : \mathbb{R}^{p\times q} \to \mathbb{R}$ with composite structure as*

$$h(\mathbf{W}) = g(\mathbf{C}(\mathbf{W}))$$

such that $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ and $\mathbf{C} : \mathbb{R}^{p\times q} \to \mathbb{R}^{m\times n}$. We can construct the chain rule as follows

$$\begin{aligned}
\frac{\partial h(\mathbf{W})}{\partial w_{ij}} =& \frac{\partial g(\mathbf{C}(\mathbf{W}))}{\partial w_{ij}} \\
=& \sum_{k=1}^{m}\sum_{l=1}^{n}\frac{\partial g(\mathbf{C})}{\partial c_{kl}}\frac{\partial c_{kl}(\mathbf{W})}{\partial w_{ij}} \\
=& \sum_{k=1}^{m}\sum_{l=1}^{n}\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)_{kl}\left(\frac{\partial \mathbf{C}(\mathbf{W})}{\partial w_{ij}}\right)_{kl} \\
=& \mathrm{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^\top\left(\frac{\partial \mathbf{C}(\mathbf{W})}{\partial w_{ij}}\right)\right) \\
=& \frac{\mathrm{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^\top \partial\mathbf{C}(\mathbf{W})\right)}{\partial w_{ij}}.
\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Hence, we have

$$\frac{\partial h(\mathbf{W})}{\partial \mathbf{W}} = \frac{\operatorname{tr}\left(\left(\dfrac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^{\top} \partial \mathbf{C}(\mathbf{W})\right)}{\partial \mathbf{W}}.$$

Remark 12.2. *Note that we write ∂ before $\mathbf{C}(\mathbf{W})$ (rather than before trace), which means we take derivative on $\mathbf{W}$ by regarding $\partial g(\mathbf{C})/\partial \mathbf{C}$ is fixed.*

We fix some $\sigma^2 > 0$ and consider the differential with respect to $\mathbf{W} \in \mathbb{R}^{p \times q}$. We denote

$$\mathbf{C}(\mathbf{W}) = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}, \qquad g(\mathbf{C}) = \ln\det(\mathbf{C}), \qquad \textit{and} \qquad h(\mathbf{W}) = g(\mathbf{C}(\mathbf{W})) = \ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}).$$

For the term of logarithmic determinant, we have

$$\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}} = \mathbf{C}^{-1},$$

and the chain rule implies

$$\begin{aligned}\frac{\partial h(\mathbf{W})}{\partial \mathbf{W}} =& \frac{\operatorname{tr}\big(\mathbf{C}^{-1}\partial(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})\big)}{\partial \mathbf{W}} \\ =& 2\mathbf{C}^{-1}\mathbf{W} = 2(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W}.\end{aligned}$$

For the term of trace, we have

$$\mathrm{d}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) = \mathrm{d}(\mathbf{W}\mathbf{W}^\top) = (\mathrm{d}\mathbf{W}) \cdot \mathbf{W}^\top + \mathbf{W} \cdot \mathrm{d}\mathbf{W}^\top$$

and

$$\begin{aligned}\mathbf{0} =& \mathrm{d}\big(\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})\big) \\ =& \mathrm{d}\big(\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big) \cdot (\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) + \hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1} \cdot \mathrm{d}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) \\ =& \mathrm{d}\big(\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big)\mathbf{C} + \hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\big((\mathrm{d}\mathbf{W})\mathbf{W}^\top + \mathbf{W}\mathrm{d}\mathbf{W}^\top\big),\end{aligned}$$

which implies

$$\mathrm{d}\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1} = -\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\big((\mathrm{d}\mathbf{W})\mathbf{W}^\top + \mathbf{W}\mathrm{d}\mathbf{W}^\top\big)\mathbf{C}^{-1}.$$

Taking trace on above equations, we achieve

$$\begin{aligned}&\operatorname{tr}\big(\mathrm{d}\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big) \\ =& -\operatorname{tr}\big(\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}(\mathrm{d}\mathbf{W})\mathbf{W}^\top\mathbf{C}^{-1}\big) - \operatorname{tr}\big(\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W}(\mathrm{d}\mathbf{W}^\top)\mathbf{C}^{-1}\big) \\ =& -\operatorname{tr}\big(\mathbf{W}^\top\mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathrm{d}\mathbf{W}\big) - \operatorname{tr}\big((\mathrm{d}\mathbf{W}^\top)\mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W}\big) \\ =& -2\operatorname{tr}\big(\mathbf{W}^\top\mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathrm{d}\mathbf{W}\big).\end{aligned}$$

Therefore, we have

$$\frac{\partial \operatorname{tr}\big(\hat{\mathbf{\Sigma}}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\big)}{\partial \mathbf{W}} = -2\mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W}.$$

Come back to MLE, we want to find $\mathbf{W} \in \mathbb{R}^{p \times q}$ such that

$$\frac{\partial f(\mathbf{W}, \sigma^2)}{\partial \mathbf{W}} = 2\mathbf{C}^{-1}\mathbf{W} - 2\mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W} = \mathbf{0}, \qquad \text{where} \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}.$$

This document is organized by Luo Luo.
Please do not distribute.

Let $\mathbf{W} = \mathbf{Q}\mathbf{D}\mathbf{V}^\top$ be condense SVD of $\mathbf{W}$, where $\mathbf{Q} \in \mathbb{R}^{p\times q}$, $\mathbf{D} \in \mathbb{R}^{q\times q}$ and $\mathbf{V} \in \mathbb{R}^{q\times q}$. Denote $\mathbf{Q}_\perp \in \mathbb{R}^{(p-q)\times q}$ be the orthogonal complement of $\mathbf{Q}$, then we have

$$\begin{aligned}\mathbf{C} =& \mathbf{Q}\mathbf{D}^2\mathbf{Q}^\top + \sigma^2 \begin{bmatrix}\mathbf{Q} & \mathbf{Q}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{Q}^\top \\ \mathbf{Q}_\perp^\top\end{bmatrix} \\ =& \begin{bmatrix}\mathbf{Q} & \mathbf{Q}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{D}^2 + \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{Q}^\top \\ \mathbf{Q}_\perp^\top\end{bmatrix}\end{aligned}$$

and

$$\mathbf{C}^{-1} = \begin{bmatrix}\mathbf{Q} & \mathbf{Q}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{D} + \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{Q}^\top \\ \mathbf{Q}_\perp^\top\end{bmatrix} = \mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{Q}^\top + \sigma^{-2}\mathbf{Q}_\perp\mathbf{Q}_\perp^\top.$$

Taking the gradient be zero, we have

$$\begin{aligned}& \mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W} = \mathbf{C}^{-1}\mathbf{W} \\ \iff\ & \hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W} = \mathbf{Q}\mathbf{D}\mathbf{V}^\top.\end{aligned}$$

The decomposition of $\mathbf{C}$ and $\mathbf{C}^{-1}$ implies

$$\begin{aligned}& \hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W} \\ =& \hat{\mathbf{\Sigma}}\left(\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{Q}^\top + \sigma^{-2}\mathbf{Q}_\perp\mathbf{Q}_\perp^\top\right)\mathbf{Q}\mathbf{D}\mathbf{V}^\top \\ =& \hat{\mathbf{\Sigma}}\left(\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{Q}^\top\right)\mathbf{Q}\mathbf{D}\mathbf{V}^\top \\ =& \hat{\mathbf{\Sigma}}\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{D}\mathbf{V}^\top.\end{aligned}$$

Hence, we obtain

$$\begin{aligned}& \hat{\mathbf{\Sigma}}\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{D}\mathbf{V}^\top = \mathbf{Q}\mathbf{D}\mathbf{V}^\top \\ \iff& \hat{\mathbf{\Sigma}}\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1}\mathbf{D} = \mathbf{Q}\mathbf{D} \\ \iff& \hat{\mathbf{\Sigma}}\mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I})^{-1} = \mathbf{Q} \\ \iff& \hat{\mathbf{\Sigma}}\mathbf{Q} = \mathbf{Q}(\mathbf{D}^2 + \sigma^2\mathbf{I}),\end{aligned}$$

where the column orthogonal matrix $\mathbf{Q}$ and the diagonal matrix $\mathbf{D}^2 + \sigma^2\mathbf{I}$ correspond to the eigenvalue decomposition of $\hat{\mathbf{\Sigma}}$. Hence, all potential solution of $\mathbf{W}$ has the form of

$$\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \tag{33}$$

where $\mathbf{U}_q$ contains q eigenvectors of $\hat{\mathbf{\Sigma}}$, $\mathbf{\Lambda}$ is diagonal matrix with corresponding eigenvalues and $\mathbf{R}^{q\times q}$ is any orthogonal matrix. Using the expression (33), we obtain

$$\begin{aligned}\mathbf{C} =& \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} \\ =& \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{U}_q^\top + \sigma^2\mathbf{I} \\ =& \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{U}_q^\top + \sigma^2\left(\mathbf{U}_q\mathbf{U}_q^\top + \mathbf{U}_{p-q}\mathbf{U}_{p-q}^\top\right) \\ =& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{p-q}^\top\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\mathbf{C}^{-1}\hat{\mathbf{\Sigma}} =& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_q^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{p-q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{p-q}^\top\end{bmatrix} \\ =& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{\Lambda}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{p-q}^\top\end{bmatrix}.\end{aligned}$$

This document is organized by Luo Luo.
Please do not distribute.

Therefore, the function f can be written as

$$\begin{aligned} f =& \ln\det(\mathbf{C}) + \mathrm{tr}(\mathbf{C}^{-1}\hat{\mathbf{\Sigma}})) \\ =& \sum_{i=1}^{q} \ln\lambda_i + (p-q)\ln\sigma^2 + q + \frac{1}{\sigma^2}\sum_{j=q+1}^{p}\lambda_j. \end{aligned}$$

Minimizing f over σ^2 achieves

$$\sigma^2 = \frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j. \tag{34}$$

So we have

$$\begin{aligned} f =& \sum_{i=1}^{q}\ln\lambda_i + (p-q)\ln\left(\frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j\right) + p \\ =& -\sum_{j=q+1}^{p}\ln\lambda_i + \sum_{i=1}^{p}\ln\lambda_i + (p-q)\ln\left(\frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j\right) + p. \end{aligned}$$

Since $\sum_{i=1}^{p}\ln\lambda_i = \ln\det(\hat{\mathbf{\Sigma}})$ is fixed, we only need to select $\lambda_{q+1},\dots,\lambda_p$ to minimize

$$\ln\left(\frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j\right) - \frac{1}{p-q}\sum_{j=q+1}^{p}\ln\lambda_i.$$

Suppose that $\lambda_{q+1} = \max\{\lambda_{q+1},\dots,\lambda_p\}$, then we have

$$\lambda_{q+1} \geq \frac{\lambda_{q+2}+\cdots+\lambda_p}{p-1-q}.$$

We introduce the following function to determine λ_{q+1}:

$$g(x) = \ln\left(\frac{1}{p-q}\left(x + \sum_{j=q+2}^{p}\lambda_j\right)\right) - \frac{1}{p-q}\left(\ln x + \sum_{j=q+2}^{p}\ln\lambda_i\right).$$

Then we have

$$g'(x) = \frac{1}{x + \sum_{j=q+2}^{p-1}\lambda_j} - \frac{1}{(p-q)x} \geq 0$$

when (x corresponds to $\lambda_{q+1} = \max\{\lambda_{q+1},\dots,\lambda_p\}$)

$$x \geq \frac{\lambda_{q+2}+\cdots+\lambda_p}{p-1-q},$$

which implies $g(x)$ is increasing. Therefore, we should take $\lambda_{q+1},\dots,\lambda_p$ as the smallest $p-q$ eigenvalues. In the view of equations (33) and (34), we have

$$\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \qquad \text{and} \qquad \mathbf{W} = \frac{1}{p-q}\sum_{j=q+1}^{q}\lambda_j, \tag{35}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\dots,\lambda_q)$ such that λ_j is the j-th largest eigenvalue of $\hat{\mathbf{\Sigma}}$ and $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ is consist of the corresponding eigenvectors.

This document is organized by Luo Luo.
Please do not distribute.

Remark 12.3. *We have showed that the MLE estimators $\hat{\mathbf{W}}$ and $\hat{\sigma}^2$ also minimize the Frobenius norm error in Section 1, that is*

$$(\hat{\mathbf{W}}, \hat{\sigma}^2) = \mathop{\arg\min}_{\mathbf{W}\in\mathbb{R}^{p\times q}, \sigma^2\in\mathbb{R}} \left\|\hat{\boldsymbol{\Sigma}} - \left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\right\|_F.$$

In this view, PCA corresponds to the best low-rank approximation and PPCA leads to the best regularized approximation, which is firstly observed and formally presented by Zhang [18]. If we use other unitary invariant norm to measure the approximation error, PCA also leads to the best low-rank approximation, while PPCA not. The regularized approximation for spectral norm and its extension in streaming case was studied by Luo et al. [11].

Remark 12.4. *If we take $\sigma^2 \to 0$ in the solution, PPCA tends to ordinary PCA.*

Remark 12.5. *If we directly establish the model with $\sigma^2 = 0$ by assuming*

$$\mathbf{y}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} \qquad \textit{and} \qquad \mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$$

for parameters $\mathbf{W} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{\mu} \in \mathbb{R}^p$, then $\mathbf{y}_\alpha$ has singular normal distribution

$$\mathbf{y}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top).$$

We can construct the pseudo likelihood function on low-dimensional space $\{\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{v} : \mathbf{v} \in \mathbb{R}^p\}$, that is

$$\frac{1}{(2\pi)^{Nq/2}\left(\det^\dagger(\mathbf{W}\mathbf{W}^\top)\right)^{N/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top(\mathbf{W}\mathbf{W}^\top)^\dagger(\mathbf{x}_\alpha - \boldsymbol{\mu})\right).$$

Minimizing on $\boldsymbol{\mu}$ leads to the equation $(\mathbf{W}\mathbf{W}^\top)^\dagger(\bar{\mathbf{x}} - \boldsymbol{\mu}) = \mathbf{0}$, which lacks the uniqueness of the solution.

The EM Algorithm for PPCA For the model

$$\mathbf{y}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha, \quad \text{where} \quad \mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}) \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I})$$

are independent. We establish the EM algorithm as follows:

1. We consider $\{\mathbf{x}_\alpha\}_{\alpha=1}^N$ to be missing data and $\{\mathbf{x}_\alpha, \mathbf{y}_\alpha\}_{\alpha=1}^N$ to be the complete data.
2. The posterior of $\mathbf{x}_\alpha$ given $\mathbf{y}_\alpha$ is

$$\begin{aligned}
& p(\mathbf{x}_\alpha \,|\, \mathbf{y}_\alpha) \\
\propto & \, p(\mathbf{y}_\alpha \,|\, \mathbf{x}_\alpha)\, p(\mathbf{x}_\alpha) \\
= & \, n\big(\mathbf{y}_\alpha \,|\, \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu}, \sigma^2\mathbf{I}\big)\, n(\mathbf{x}_\alpha \,|\, \mathbf{0}, \mathbf{I}) \\
\propto & \exp\left(-\frac{\|\mathbf{y}_\alpha - \mathbf{W}\mathbf{x}_\alpha - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|\mathbf{x}_\alpha\|_2^2}{2}\right) \\
\propto & \exp\left(\frac{1}{2\sigma^2}\left(\mathbf{x}_\alpha^\top\mathbf{W}\mathbf{W}^\top\mathbf{x}_\alpha - 2(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{x} + \sigma^2\left\|\mathbf{x}_\alpha\right\|_2^2\right)\right) \\
= & \exp\left(\frac{1}{2\sigma^2}\left(\mathbf{x}_\alpha^\top\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\mathbf{x}_\alpha - 2(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\mathbf{x}_\alpha\right)\right).
\end{aligned}$$

   Hence, it is normal distribution such that

$$\mathbf{x}_\alpha \mid \mathbf{y}_\alpha \sim \mathcal{N}\left(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right), \tag{36}$$

   where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

This document is organized by Luo Luo.
Please do not distribute.

3. The joint density of $\{\mathbf{x}_\alpha, \mathbf{y}_\alpha\}_{\alpha=1}^N$ is

$$\prod_{\alpha=1}^N n\big(\mathbf{y}_\alpha \,|\, \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu}, \sigma^2\mathbf{I}\big)\, n(\mathbf{x}_\alpha \,|\, \mathbf{0}, \mathbf{I}).$$

In E-step, we take the expectation of the log-likelihood with respect to the distributions $p(\mathbf{x}_\alpha \,|\, \mathbf{y}_\alpha)$:

$$\begin{aligned} l_C =& \mathbb{E}\left[\ln\left(\prod_{\alpha=1}^N p(\mathbf{x}_\alpha \,|\, \mathbf{y}_\alpha)\right)\right] \\ =& -\sum_{\alpha=1}^N \left(\frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top - \frac{1}{2\sigma^2}\langle\mathbf{x}_\alpha\rangle^\top\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top \right. \\ & \qquad \left. + \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{W}^\top\mathbf{W}\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\right) + \frac{\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle}{2}\right) + C. \end{aligned}$$

where $\langle\mathbf{x}_\alpha\rangle = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu})$ and $\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle = \sigma^2\mathbf{M}^{-1} + \langle\mathbf{x}_\alpha\rangle\langle\mathbf{x}_\alpha\rangle^\top$.
In the M-step, the expectation l_C is maximized with respect to $\mathbf{W}$ and σ^2 giving new parameter

$$\begin{aligned} \mathbf{W}_+ =& \left(\sum_{\alpha=1}^N(\mathbf{y}_\alpha - \boldsymbol{\mu})\langle\mathbf{x}_\alpha\rangle^\top\right)\left(\sum_{\alpha=1}^N\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\right)^{-1} \\ =& \sum_{\alpha=1}^N\left((\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\right)\left(N\sigma^2\mathbf{M}^{-1} + \sum_{\alpha=1}^N\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\right)^{-1} \\ =& \big(N\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\left(N\sigma^2\mathbf{M}^{-1} + N\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\right)^{-1} \\ =& \hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)^{-1} \\ =& \hat{\boldsymbol{\Sigma}}\mathbf{W}\big(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\big)^{-1} \end{aligned}$$

and

$$\begin{aligned} \sigma_+^2 =& \frac{1}{Nd}\sum_{\alpha=1}^N\left(\|\mathbf{y}_\alpha - \boldsymbol{\mu}\|_2^2 - 2\langle\mathbf{x}_\alpha\rangle^\top\mathbf{W}_+^\top(\mathbf{y}_\alpha - \boldsymbol{\mu}) + \operatorname{tr}\left(\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\mathbf{W}_+^\top\mathbf{W}_+\right)\right) \\ =& \frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) - \sum_{\alpha=1}^N 2\operatorname{tr}\big((\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big) \\ & \quad + \sum_{\alpha=1}^N\operatorname{tr}\big((\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1})\mathbf{W}_+^\top\mathbf{W}_+\big)\Big) \\ =& \frac{1}{d}\big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) - 2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big) + \operatorname{tr}\big((\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1})\mathbf{W}_+^\top\mathbf{W}_+\big)\big) \\ =& \frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) - 2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big) + \operatorname{tr}\Big(\mathbf{W}_+\big(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\mathbf{W}_+^\top\big)\big) \\ =& \frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) - 2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big) + \operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\big(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\big)^{-1}\big(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\mathbf{W}_+^\top\big)\big) \\ =& \frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) - 2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big) + \operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\big)\big) \\ =& \frac{1}{d}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\right). \end{aligned}$$

Remark 12.6. *See Wu's 1983 paper for the convergence analysis of EM algorithms.*

This document is organized by Leo Luo.
Please do not distribute.

References

[1] Theodore W. Anderson. *An introduction to multivariate statistical analysis.* Wiley, 2003.

[2] Raghu R. Bahadur. Sufficiency and statistical decision functions. *The Annals of Mathematical Statistics*, pages 423–462, 1954.

[3] Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty.* PhD thesis, Massachusetts Institute of Technology, 1971.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[5] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[6] Yasuko Chikuse. *Statistics on special manifolds*, volume 1. Springer, 2003.

[7] Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

[8] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis.* Cambridge University Press, 1994.

[9] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8580–8589, 2018.

[10] Chengchang Liu, Cheng Chen, and Luo Luo. Symmetric rank-k methods. *arXiv preprint arXiv:2303.16188*, 2023.

[11] Luo Luo, Cheng Chen, Zhihua Zhang, Wu-Jun Li, and Tong Zhang. Robust frequent directions with application in online learning. *Journal of Machine Learning Research*, 20(1):1697–1737, 2019.

[12] Luo Luo, Yujun Li, and Cheng Chen. Finding second-order stationary points in nonconvex-strongly-concave minimax optimization. In *Conference on Neural Information Processing Systems*, pages 36667–36679, 2022.

[13] Joris Mulder and Luis Raúl Pericchi. The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, 13(4):1193–1214, 2018.

[14] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[15] Robert J. Serfling. *Approximation theorems of mathematical statistics.* John Wiley & Sons, 1980.

[16] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2000.

[17] C.F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

[18] Zhihua Zhang. The matrix ridge approximation: algorithms and applications. *Machine Learning*, 97: 227–258, 2014.

This document is organized by Luo Luo.
Please do not distribute.