

Lecture Notes of Multivariate Statistics

Luo Luo

School of Data Science, Fudan University

April 2, 2022

1 Review of Linear Algebra

Theorem 1.1 (QR Factorization). *Prove the following results for Gram-Schmidt orthogonalization*

1. $r_{jj} \neq 0$ for all $i = 1, \dots, n$
2. $\|\mathbf{q}_i\|_2 = 1$ for all $i = 1, \dots, n$
3. $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $i = 1, \dots, n$ and $j < i$.

Proof. **Part 1:** Since each \mathbf{q}_i is a linear combination of $\{\mathbf{a}_1, \dots, \mathbf{a}_i\}$, the entry r_{jj} is zero means

$$r_{jj} = \left\| \mathbf{a}_n - \sum_{i=1}^{n-1} r_{in} \mathbf{q}_i \right\|_2 = 0,$$

then \mathbf{a}_n must be a linear combination of $\{\mathbf{a}_1, \dots, \mathbf{a}_{n-1}\}$, which validate the full rank assumption on \mathbf{A} .

Part 2: Just use the expression of r_{jj} .

Part 3: Recall that $r_{ij} = \mathbf{q}_i^\top \mathbf{a}_j$ for any $i \neq j$. We can verify

$$\mathbf{q}_1^\top \mathbf{q}_2 = \frac{\mathbf{q}_1^\top (\mathbf{a}_2 - r_{12} \mathbf{q}_1)}{r_{22}} = \frac{\mathbf{q}_1^\top (\mathbf{a}_2 - (\mathbf{q}_1^\top \mathbf{a}_2) \mathbf{q}_1)}{r_{22}} = \frac{\mathbf{q}_1^\top \mathbf{a}_2 - (\mathbf{q}_1^\top \mathbf{a}_2) \mathbf{q}_1^\top \mathbf{q}_1}{r_{22}} = 0$$

Suppose for $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $i = 1, \dots, n' - 1$ and $j < i$. Then for all $k = 1, 2, \dots, n' - 1$, we have

$$\mathbf{q}_k^\top \mathbf{q}_{n'} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - \sum_{i=1}^{n'-1} r_{in'} \mathbf{q}_i^\top \mathbf{q}_k}{r_{n'n'}} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - r_{kn'} \mathbf{q}_k^\top \mathbf{q}_k}{r_{n'n'}} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - r_{kn'}}{r_{n'n'}} = 0$$

Then we prove the result by induction. □

Theorem 1.2. *Prove $\|\mathbf{A}\|_2 = \sigma_1$.*

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be full SVD of \mathbf{A} . Then

$$\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\|_2 = \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\|_2$$

Then let $\mathbf{y} = \mathbf{V}^\top \mathbf{x}$. Since \mathbf{V} is orthogonal matrix, we have $\|\mathbf{y}\|_2 = \|\mathbf{V}^\top \mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1$. Hence,

$$\sup_{\|\mathbf{x}\|_2=1} \|\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\|_2 = \sup_{\|\mathbf{y}\|_2=1} \|\mathbf{\Sigma}\mathbf{y}\|_2 = \sup_{\|\mathbf{y}\|_2=1} \sqrt{\sum_{i=1}^r (\sigma_i y_i)^2} \leq \sigma_1.$$

We attain the maximum by taking $\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the corresponding \mathbf{x} is $\mathbf{V} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ □

Theorem 1.3 (Cholesky Factorization). *The symmetric positive-definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the decomposition of the form*

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top$$

where $\mathbf{L} \in \mathbb{R}^{n \times n}$ is a lower triangular matrix with real and positive diagonal entries.

Proof. For $n = 1$, it is trivial. Suppose it holds for $n - 1$, then any $\tilde{\mathbf{A}} \in \mathbb{R}^{(n-1) \times (n-1)}$ can be written as

$$\tilde{\mathbf{A}} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top$$

where $\tilde{\mathbf{L}} \in \mathbb{R}^{(n-1) \times (n-1)}$ is a lower triangular matrix with real and positive diagonal entries. Consider the case of n such that

$$\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{A}} & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad \text{where } \mathbf{a} \in \mathbb{R}^{n-1}, \quad \alpha \in \mathbb{R}.$$

Let

$$\mathbf{L}_1 = \begin{bmatrix} \tilde{\mathbf{L}}^{-1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

We have

$$\mathbf{L}_1^{-1} \mathbf{A} \mathbf{L}_1^{-\top} = \begin{bmatrix} \tilde{\mathbf{L}}^{-1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{L}}^{-\top} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{b} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \triangleq \mathbf{B} \in \mathbb{R}^{n \times n} \quad \text{where } \mathbf{b} \in \tilde{\mathbf{L}}^{-1} \mathbf{a} \in \mathbb{R}^{n-1}.$$

Let

$$\mathbf{L}_2 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{b}^\top & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Then

$$\mathbf{L}_2^{-1} \mathbf{B} \mathbf{L}_2^{-\top} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{b}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{b} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \alpha - \mathbf{b}^\top \mathbf{b} \end{bmatrix}.$$

Since \mathbf{A} is positive-definite, we have

$$\alpha - \mathbf{b}^\top \mathbf{b} = \alpha - \mathbf{a}^\top \tilde{\mathbf{L}}^{-\top} \tilde{\mathbf{L}}^{-1} \mathbf{a} = \alpha - \mathbf{a}^\top \tilde{\mathbf{L}}^{-\top} \tilde{\mathbf{L}}^{-1} \mathbf{a} = \alpha - \mathbf{a}^\top \tilde{\mathbf{A}}^{-1} \mathbf{a} > 0.$$

Let $\alpha - \mathbf{b}^\top \mathbf{b} = \lambda^2$, where $\lambda > 0$. Hence, we have

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \alpha - \mathbf{b}^\top \mathbf{b} \end{bmatrix} = \mathbf{L}_3 \mathbf{L}_3^\top, \quad \text{where } \mathbf{L}_3 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda \end{bmatrix}$$

which means $\mathbf{A} = \mathbf{L}\mathbf{L}^\top \in \mathbb{R}^{n \times n}$ where $\mathbf{L} = \mathbf{L}_1 \mathbf{L}_2 \mathbf{L}_3 \in \mathbb{R}^{n \times n}$ is a lower triangular matrix with real and positive diagonal entries. \square

Theorem 1.4. *Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution of minimization problem*

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

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

Proof. The Hessian of $f(\mathbf{x})$ is $\mathbf{A}^\top \mathbf{A} \succeq \mathbf{0}$, which means $f(\mathbf{x})$ is convex. Let $\mathbf{A} = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^\top$ be the condense SVD, where r is the rank of \mathbf{A} . Since $\nabla f(\mathbf{x}) = \mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b}$, we only needs to solve the linear system

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

We denote the solution of $\mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b} = \mathbf{0}$ be

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

We can verify that $\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \mathbf{y}$ is the solution of the linear system because

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

Hence, we have $\mathcal{X}_1 \subseteq \mathcal{X}$, where $\mathcal{X}_1 = \{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \mathbf{y}, \mathbf{y} \in \mathbb{R}^n\}$.

We also have

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

Hence, we have $\mathcal{X} = \{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top) \mathbf{x}\} \subseteq \mathcal{X}_1$. In conclusion, we have $\mathcal{X} = \mathcal{X}_1$. \square

2 The Multivariate Normal Distributions

Statistical Independence If $F(x, y) = F(x)G(y)$, we have

$$\begin{aligned} f(x, y) &= \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x)G(y)}{\partial x \partial y} \\ &= \frac{dF(x)}{dx} \frac{dG(y)}{dy} \\ &= f(x)g(y). \end{aligned}$$

If $f(x, y) = f(x)g(y)$, we have

$$\begin{aligned} F(x, y) &= \int_{-\infty}^y \int_{-\infty}^x f(u, v) du dv = \int_{-\infty}^y \int_{-\infty}^x f(u)g(v) du dv \\ &= \int_{-\infty}^y \int_{-\infty}^x f(u, v) du dv = \int_{-\infty}^x f(u) du \int_{-\infty}^y g(v) dv \\ &= F(x)G(y). \end{aligned}$$

Uncorrelated does not means independent Let $X \sim U(-1, 1)$ and

$$Y = \begin{cases} X, & X > 0 \\ -X, & X \leq 0 \end{cases}$$

Show X and Y are uncorrelated but they are NOT independent.

Conditional Distributions Let $y_1 = y$, $y_2 = y + \Delta y$. Then for a continuous density, the mean value theorem implies

$$\int_y^{y+\Delta y} g(v) dv = g(y^*)\Delta y,$$

where $y \leq y^* \leq y + \Delta y$. We also have

$$\int_y^{y+\Delta y} f(u, v) dv = f(u, y^*(u))\Delta y,$$

where $y \leq y^*(u) \leq y + \Delta y$. Connecting above results to

$$\Pr\{x_1 \leq X \leq x_2 \mid y_1 \leq Y \leq y_2\} = \frac{\int_{x_1}^{x_2} \int_{y_1}^{y_2} f(u, v) dv du}{\int_{y_1}^{y_2} g(v) dv}$$

with $y_1 = y$ and $y_2 = y + \Delta y$, we have

$$\begin{aligned} & \Pr\{x_1 \leq X \leq x_2 \mid y \leq Y \leq y + \Delta y\} \\ &= \frac{\int_{x_1}^{x_2} \int_y^{y+\Delta y} f(u, v) dv du}{\int_y^{y+\Delta y} g(v) dv} \\ &= \frac{\int_{x_1}^{x_2} f(u, y^*(u))\Delta y du}{g(y^*)\Delta y} \\ &= \int_{x_1}^{x_2} \frac{f(u, y^*(u))}{g(y^*)} du. \end{aligned} \tag{1}$$

For y such that $g(y) > 0$, we define $\Pr\{x_1 \leq X \leq x_2 \mid Y = y\}$, the probability that X lies between x_1 and x_2 , given that Y is y , as the limit of (1) as $\Delta y \rightarrow 0$. Thus

$$\Pr\{x_1 \leq X \leq x_2 \mid Y = y\} = \int_{x_1}^{x_2} \frac{f(u, y)}{g(y)} du = \int_{x_1}^{x_2} f(u \mid y) du. \tag{2}$$

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 $\mathbf{u} : \mathbb{R}^p \rightarrow \mathbb{R}^p$ such that

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

Assume the transformation \mathbf{u} from the x -space to the y -space is one-to-one, then the inverse transformation is \mathbf{u}^{-1} such that

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

Let the random variables Y_1, \dots, Y_p be defined by

$$Y_i = u_i(X_1, \dots, X_p), \quad i = 1, \dots, p,$$

then we have

$$\int_{\mathbf{u}(\Omega)} g(\mathbf{y}) d\mathbf{y} = \int_{\Omega} g(\mathbf{u}(\mathbf{x})) \text{abs}(|\mathbf{J}(\mathbf{x})|) d\mathbf{x}, \quad (3)$$

and

$$f(\mathbf{x}) = g(\mathbf{u}(\mathbf{x})) \text{abs}(|\mathbf{J}(\mathbf{x})|), \quad (4)$$

where the Jacobin matrix is

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

A roughly proof for above results:

- If $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathcal{S} \subset \mathbb{R}^p$ is a measurable set, then $m(\mathbf{A}\mathcal{S}) = |\det(\mathbf{A})|m(\mathcal{S})$. 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 $\{\Omega_i\}_i$. Then

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

- Consider notation Ω such that

$$\int_{\Omega} = \int_{x_1}^{x'_1} \cdots \int_{x_p}^{x'_p}$$

where $x_1 \leq x'_1, x_2 \leq x'_2, \dots, x_p \leq x'_p$. 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 (3). On the other hand, we have

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

Since it holds for any Ω , then

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

Lemma 2.1. *If \mathbf{Z} is an $m \times n$ random matrix, \mathbf{D} is an $l \times m$ real matrix, \mathbf{E} is an $n \times q$ real matrix, and \mathbf{F} is an $l \times q$ real matrix, then*

$$\mathbb{E}[\mathbf{DZE} + \mathbf{F}] = \mathbf{D}\mathbb{E}[\mathbf{Z}]\mathbf{E} + \mathbf{F}.$$

Proof. The element in the i -th row and j -th column of $\mathbb{E}[\mathbf{DZE} + \mathbf{F}]$ is

$$\mathbb{E} \left[\sum_{h,g} d_{ih} z_{hg} e_{gj} + f_{ij} \right] = \sum_{h,g} d_{ih} \mathbb{E}[z_{hg}] e_{gj} + f_{ij}$$

which is the element in the i -th row and j -th column of $\mathbf{D}\mathbb{E}[\mathbf{Z}]\mathbf{E} + \mathbf{F}$. □

Lemma 2.2. *If $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{f} \in \mathbb{R}^l$, where \mathbf{D} is an $l \times m$ real matrix, $\mathbf{x} \in \mathbb{R}^m$ is a random vector, then*

$$\mathbb{E}[\mathbf{y}] = \mathbf{D}\mathbb{E}[\mathbf{x}] + \mathbf{f} \quad \text{and} \quad \text{Cov}[\mathbf{y}] = \mathbf{D}\text{Cov}[\mathbf{x}]\mathbf{D}^\top.$$

Proof. We have

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

□

The Density Function of Multivariate Normal Distribution Let the spectral decomposition of \mathbf{A} be $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, then we take $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}^{-1/2}$ and it satisfies $\mathbf{C}^\top \mathbf{A} \mathbf{C} = \mathbf{I}$ and \mathbf{C} is non-singular. Define $\mathbf{y} = \mathbf{C}^{-1}(\mathbf{x} - \mathbf{b})$, then

$$\begin{aligned} K^{-1} &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp \left(-\frac{1}{2}(\mathbf{x} - \mathbf{b})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{b}) \right) dx_1 \dots dx_p \\ &= \frac{1}{\det(\mathbf{C}^{-1})} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp \left(-\frac{1}{2}\mathbf{y}^\top \mathbf{y} \right) dy_1 \dots dy_p \\ &= \det(\mathbf{C}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp \left(-\frac{1}{2} \sum_{i=1}^n y_i^2 \right) dy_1 \dots dy_p \\ &= \det(\mathbf{A}^{\frac{1}{2}}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp \left(-\frac{1}{2} y_p^2 \right) \cdots \exp \left(-\frac{1}{2} y_1^2 \right) dy_1 \dots dy_p \\ &= \det(\mathbf{A}^{\frac{1}{2}}) (2\pi)^{\frac{p}{2}}. \end{aligned}$$

The relation $\mathbf{y} = \mathbf{C}^{-1}(\mathbf{x} - \mathbf{b})$ means $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b}$ and $\mathbb{E}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}] + \mathbf{b}$. The transformation implies the density function of \mathbf{y} is

$$g(\mathbf{y}) = \det(\mathbf{C}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K \exp \left(-\frac{1}{2}(\mathbf{C}\mathbf{y} + \mathbf{b} - \mathbf{b})^\top \mathbf{A}(\mathbf{C}\mathbf{y} + \mathbf{b} - \mathbf{b}) \right) dy_1 \dots dy_p$$

$$\begin{aligned}
&= \det(\mathbf{C}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K \exp\left(-\frac{1}{2} \mathbf{y}^\top \mathbf{C}^\top \mathbf{A} \mathbf{C} \mathbf{y}\right) dy_1 \dots dy_p \\
&= K \det(\mathbf{C}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} \mathbf{y}^\top \mathbf{y}\right) dy_1 \dots dy_p \\
&= \frac{\det(\mathbf{C})}{\sqrt{(2\pi)^p \det(\mathbf{A})}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} \sum_{i=1}^p y_i^2\right) dy_1 \dots dy_p \\
&= \frac{1}{(2\pi)^{p/2}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} \sum_{i=1}^p y_i^2\right) dy_1 \dots dy_p.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[y_i] &= \frac{1}{(2\pi)^{p/2}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} y_i \exp\left(-\frac{1}{2} \sum_{j=1}^p y_j^2\right) dy_1 \dots dy_p \\
&= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_i \exp\left(-\frac{1}{2} y_i^2\right) dy_i \right) \prod_{j=1, j \neq i}^p \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} y_j^2\right) dy_j \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_i \exp\left(-\frac{1}{2} y_i^2\right) dy_i = 0.
\end{aligned}$$

Thus $\mathbb{E}[\mathbf{y}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}] + \mathbf{b} = \boldsymbol{\mu}$ implies $\mathbf{b} = \boldsymbol{\mu}$.

The relation $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b}$ means $\text{Cov}[\mathbf{x}] = \mathbf{C}\text{Cov}[\mathbf{y}]\mathbf{C}^\top = \mathbf{C}\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{C}^\top$. For each $i \neq j$, we have

$$\begin{aligned}
&\mathbb{E}[y_i y_j] \\
&= \frac{1}{(2\pi)^{p/2}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} y_i y_j \exp\left(-\frac{1}{2} \sum_{h=1}^p y_h^2\right) dy_1 \dots dy_p \\
&= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_i \exp\left(-\frac{1}{2} y_i^2\right) dy_i \right) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_j \exp\left(-\frac{1}{2} y_j^2\right) dy_j \right) \prod_{j=1, j \neq i, j}^p \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} y_h^2\right) dy_h \\
&= 0
\end{aligned}$$

We also have

$$\begin{aligned}
&\mathbb{E}[y_i^2] \\
&= \frac{1}{(2\pi)^{p/2}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} y_i^2 \exp\left(-\frac{1}{2} \sum_{h=1}^p y_h^2\right) dy_1 \dots dy_p \\
&= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_i^2 \exp\left(-\frac{1}{2} y_i^2\right) dy_i \right) \prod_{j=1, j \neq i}^p \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} y_h^2\right) dy_h = 1.
\end{aligned}$$

Hence, it holds that

$$\mathbb{E}[(y_i - \mathbb{E}[y_i])(y_j - \mathbb{E}[y_j])] = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

which implies $\boldsymbol{\Sigma} = \text{Cov}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{C}^\top = \mathbf{C}\mathbf{C}^\top$. Since $\mathbf{C}^\top \mathbf{A} \mathbf{C} = \mathbf{I}$, we obtain $\mathbf{A}^{-1} = \mathbf{C}\mathbf{C}^\top$ and $\boldsymbol{\Sigma} = \mathbf{A}^{-1} \succ \mathbf{0}$.

Theorem 2.1. Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Sigma} \succ \mathbf{0}$. Then

$$\mathbf{y} = \mathbf{C}\mathbf{x}$$

is distributed according to $\mathcal{N}_p(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$ for non-singular $\mathbf{C} \in \mathbb{R}^{p \times p}$.

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

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

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 $\mathbf{u}(\mathbf{x}) = \mathbf{C}\mathbf{x}$, $\mathbf{u}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}\mathbf{y}$ and $\mathbf{J}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}$. Hence, we have

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

where we use the fact

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

□

Theorem 2.2. If $\mathbf{x} = [x_1, \dots, x_p]^\top$ have a joint normal distribution. Let

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

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

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

such that

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

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

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

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

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

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

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

Thus the marginal distribution of $\mathbf{x}^{(1)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(1)}, \Sigma_{11})$ and the marginal distribution of $\mathbf{x}^{(2)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \Sigma_{22})$. We have prove two variables are independent.

Necessity (independent \implies uncorrelated): Let $1 \leq i \leq q$ and $q+1 \leq j \leq p$. The Independence means

$$\begin{aligned} \sigma_{ij} &= \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) f(x_1, \dots, x_p) dx_1 \dots dx_p \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) f(x_1, \dots, x_q) f(x_{q+1}, \dots, x_p) dx_1 \dots dx_p \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_i - \mu_i) f(x_1, \dots, x_q) dx_1 \dots dx_q \cdot \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_j - \mu_j) f(x_{q+1}, \dots, x_p) dx_{q+1} \dots dx_p \\ &= 0. \end{aligned}$$

□

Theorem 2.3. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\Sigma \succ \mathbf{0}$, 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 Σ , respectively.

Proof. We shall make a non-singular linear transformation \mathbf{B} to subvectors

$$\begin{aligned} \mathbf{y}^{(1)} &= \mathbf{x}^{(1)} + \mathbf{B}\mathbf{x}^{(2)} \\ \mathbf{y}^{(2)} &= \mathbf{x}^{(2)} \end{aligned}$$

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

$$\mathbf{0} = \mathbb{E} \left[(\mathbf{y}^{(1)} - \mathbb{E}[\mathbf{y}^{(1)}]) (\mathbf{y}^{(2)} - \mathbb{E}[\mathbf{y}^{(2)}])^\top \right]$$

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

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 has a normal distribution with

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

Since the transform is non-singular, we have

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

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

Theorem 2.4. 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 $\text{Cov}[\mathbf{z}] = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top$. 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 non-singular $\mathbf{T} \succ \mathbf{0}$. We can write

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

where $\mathbf{D}\mathbf{A} \in \mathbb{R}^{q \times r}$. If the rank of $\mathbf{D}\mathbf{A}$ is r , the formal definition of a normal distribution that includes the singular distribution implies \mathbf{z} follows normal distribution.

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

$$\mathbf{E} = \text{Cov}[\mathbf{z}] = \mathbf{D}\mathbf{A}\text{Cov}[\mathbf{y}]\mathbf{A}^\top\mathbf{D}^\top = \mathbf{D}\mathbf{A}\mathbf{T}\mathbf{A}^\top\mathbf{D}^\top \in \mathbb{R}^{r \times r}$$

is rank of s . There is a non-singular matrix

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

with $\mathbf{F}_1 \in \mathbb{R}^{s \times r}$ and $\mathbf{F}_2 \in \mathbb{R}^{(r-s) \times r}$ such that

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

Thus $(\mathbf{F}_1\mathbf{D}\mathbf{A})\mathbf{T}(\mathbf{F}_1\mathbf{D}\mathbf{A})^\top = \mathbf{I}_s$ means $\mathbf{F}_1\mathbf{D}\mathbf{A}$ is of rank s and the non-singularity of \mathbf{T} means $\mathbf{F}_2\mathbf{D}\mathbf{A} = \mathbf{0}$. Hence, we have

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

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

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

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

Theorem 2.5. For $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and every vector $\boldsymbol{\alpha} \in \mathbb{R}^{(p-q)}$, we have

$$\text{Var}[x_i^{(11.2)}] \leq \text{Var}[x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}],$$

for $i = 1, \dots, q$, where $x_i^{(11.2)}$ and x_i are the i -th entry of $\mathbf{x}^{(11.2)}$ and the i -th entry of \mathbf{x} respectively.

Proof. We denote

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_{(1)}^\top \\ \vdots \\ \boldsymbol{\beta}_{(q)}^\top \end{bmatrix}.$$

Since $\mathbf{x}^{(11.2)}$ is uncorrelated with $\mathbf{x}^{(2)}$ and

$$\mathbb{E}[\mathbf{x}^{(11.2)}] = \mathbb{E}[\mathbf{x}^{(1)} - (\boldsymbol{\mu}^{(1)} + \mathbf{B}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}))] = \mathbb{E}[\mathbf{x}^{(1)}] - \boldsymbol{\mu}^{(1)} + \mathbf{B}(\mathbb{E}[\mathbf{x}^{(2)}] - \boldsymbol{\mu}^{(2)}) = \mathbf{0},$$

we have

$$\begin{aligned} & \text{Var}[x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}] \\ &= \mathbb{E}[x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)} - \mathbb{E}[x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}]]^2 \\ &= \mathbb{E}[x_i - \mu_i - \boldsymbol{\alpha}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 \\ &= \mathbb{E}[x_i^{(11.2)} + \boldsymbol{\beta}_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}) - \boldsymbol{\alpha}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 \\ &= \mathbb{E}[x_i^{(11.2)} - \mathbb{E}[x_i^{(11.2)}] + (\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha})^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 \\ &= \text{Var}[x_i^{(11.2)}]^2 + \mathbb{E}[(x_i^{(11.2)} - \mathbb{E}[x_i^{(11.2)}])(\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha})^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})] + \mathbb{E}[(\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha})^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 \\ &= \text{Var}[x_i^{(11.2)}]^2 + (\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha})^\top \mathbb{E}[(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top] (\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha}) \\ &= \text{Var}[x_i^{(11.2)}]^2 + (\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha})^\top \text{Cov}(\mathbf{x}^{(2)}) (\boldsymbol{\beta}_{(i)} - \boldsymbol{\alpha}) \\ &\geq \text{Var}[x_i^{(11.2)}]^2, \end{aligned}$$

where the quadratic form attains its minimum of 0 at $\boldsymbol{\beta}_{(i)} = \boldsymbol{\alpha}$. \square

Remark 2.1. Observe that

$$\mathbb{E}[x_i] = \mu_i + \boldsymbol{\alpha}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})$$

Hence, the second equality in the proof means $\mu_i + \boldsymbol{\beta}_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})$ is the best linear predictor of x_i in the sense that of all functions of $\mathbf{x}^{(2)}$ of the form $\boldsymbol{\alpha}^\top \mathbf{x}^{(2)} + c$, the mean squared error of the above is a minimum.

Theorem 2.6. Under the setting of Theorem 2.5, we have

$$\text{Corr}\left(x_i, \beta_{(i)}^\top \mathbf{x}^{(2)}\right) \geq \text{Corr}\left(x_i, \alpha^\top \mathbf{x}^{(2)}\right).$$

Proof. Since the correlation between two variables is unchanged when either or both is multiplied by a positive constant, we can assume that

$$\mathbb{E}\left[\alpha^\top \mathbf{x}^{(2)}\right]^2 = \mathbb{E}\left[\beta_{(i)}^\top \mathbf{x}^{(2)}\right]^2.$$

Using Theorem 2.5, we have

$$\begin{aligned} \text{Var}[x_i^{(11.2)}] &\leq \text{Var}[x_i - \alpha^\top \mathbf{x}^{(2)}] \\ \iff \mathbb{E}[x_i - \mu_i - \beta_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 &\leq \mathbb{E}[x_i - \mu_i - \alpha^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]^2 \\ \iff \text{Var}[x_i] - \mathbb{E}[(x_i - \mu_i)\beta_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})] + \text{Var}[\beta_{(i)}^\top \mathbf{x}^{(2)}] \\ &\leq \text{Var}[x_i] - \mathbb{E}[(x_i - \mu_i)\alpha^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})] + \text{Var}[\alpha^\top \mathbf{x}^{(2)}] \\ \iff \frac{\mathbb{E}[(x_i - \mu_i)\alpha^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]}{\sqrt{\text{Var}[x_i]}\sqrt{\text{Var}[\alpha^\top \mathbf{x}^{(2)}]}} &\leq \frac{\mathbb{E}[(x_i - \mu_i)\beta_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})]}{\sqrt{\text{Var}[x_i]}\sqrt{\text{Var}[\beta_{(i)}^\top \mathbf{x}^{(2)}]}} \\ \iff \frac{\text{Cov}[x_i, \alpha^\top \mathbf{x}^{(2)}]}{\sqrt{\text{Var}[x_i]}\sqrt{\text{Var}[\alpha^\top \mathbf{x}^{(2)}]}} &\leq \frac{\mathbb{E}[x_i, \beta_{(i)}^\top \mathbf{x}^{(2)}]}{\sqrt{\text{Var}[x_i]}\sqrt{\text{Var}[\beta_{(i)}^\top \mathbf{x}^{(2)}]}} \end{aligned}$$

□

Theorem 2.7. Let $\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}$. If $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are independent and $g(\mathbf{x}) = g^{(1)}(\mathbf{x}^{(1)})g^{(2)}(\mathbf{x}^{(2)})$, its characteristic function is

$$\mathbb{E}[g(\mathbf{x})] = \mathbb{E}[g^{(1)}(\mathbf{x}^{(1)})]\mathbb{E}[g^{(2)}(\mathbf{x}^{(2)})].$$

Proof. Let $f(\mathbf{x}) = f^{(1)}(\mathbf{x}^{(1)})f^{(2)}(\mathbf{x}^{(2)})$ be the density of \mathbf{x} . If $g(x)$ is real-valued, we have

$$\begin{aligned} &\mathbb{E}[g(\mathbf{x})] \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(\mathbf{x})f(\mathbf{x}) \, dx_1 \dots dx_p \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g^{(1)}(\mathbf{x}^{(1)})g^{(2)}(\mathbf{x}^{(2)})f^{(1)}(\mathbf{x}^{(1)})f^{(2)}(\mathbf{x}^{(2)}) \, dx_1 \dots dx_p \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g^{(1)}(\mathbf{x}^{(1)})f^{(1)}(\mathbf{x}^{(1)}) \, dx_1 \dots dx_q \cdot \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g^{(2)}(\mathbf{x}^{(2)})f^{(2)}(\mathbf{x}^{(2)}) \, dx_{q+1} \dots dx_p \\ &= \mathbb{E}[g^{(1)}(\mathbf{x}^{(1)})]\mathbb{E}[g^{(2)}(\mathbf{x}^{(2)})]. \end{aligned}$$

If $g(x)$ is complex-valued, then we have

$$\begin{aligned} &g(\mathbf{x}) \\ &= [g_1^{(1)}(\mathbf{x}^{(1)}) + i g_2^{(1)}(\mathbf{x}^{(1)})][g_1^{(2)}(\mathbf{x}^{(2)}) + i g_2^{(2)}(\mathbf{x}^{(2)})] \\ &= g_1^{(1)}(\mathbf{x}^{(1)})g_1^{(2)}(\mathbf{x}^{(2)}) - g_2^{(1)}(\mathbf{x}^{(1)})g_2^{(2)}(\mathbf{x}^{(2)}) + i [g_1^{(1)}(\mathbf{x}^{(1)})g_2^{(2)}(\mathbf{x}^{(2)}) + g_2^{(1)}(\mathbf{x}^{(1)})g_1^{(2)}(\mathbf{x}^{(2)})] \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[g(\mathbf{x})] \\ &= \mathbb{E}[g_1^{(1)}(\mathbf{x}^{(1)})g_1^{(2)}(\mathbf{x}^{(2)})] - \mathbb{E}[g_2^{(1)}(\mathbf{x}^{(1)})g_2^{(2)}(\mathbf{x}^{(2)})] + i \mathbb{E}[g_1^{(1)}(\mathbf{x}^{(1)})g_2^{(2)}(\mathbf{x}^{(2)}) + g_2^{(1)}(\mathbf{x}^{(1)})g_1^{(2)}(\mathbf{x}^{(2)})] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}[g_1^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g_1^{(2)}(\mathbf{x}^{(2)})] - \mathbb{E}[g_2^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g_2^{(2)}(\mathbf{x}^{(2)})] \\
&\quad + i \mathbb{E}[g_1^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g_2^{(2)}(\mathbf{x}^{(2)})] + i \mathbb{E}[g_2^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g_1^{(2)}(\mathbf{x}^{(2)})] \\
&= \left[\mathbb{E}[g_1^{(1)}(\mathbf{x}^{(1)})] + i \mathbb{E}[g_2^{(1)}(\mathbf{x}^{(1)})] \right] \left[\mathbb{E}[g_1^{(2)}(\mathbf{x}^{(2)})] + i \mathbb{E}[g_2^{(2)}(\mathbf{x}^{(2)})] \right] \\
&= \mathbb{E}[g^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g^{(2)}(\mathbf{x}^{(2)})].
\end{aligned}$$

□

Theorem 2.8. *The characteristic function of \mathbf{x} distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is*

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

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

$$\begin{aligned}
\phi_0(\mathbf{t}) &= \mathbb{E} [\exp(i \mathbf{t}^\top \mathbf{y})] \\
&= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{\exp(i \mathbf{t}^\top \mathbf{y})}{(2\pi)^{p/2}} \exp \left(-\frac{1}{2} \mathbf{y}^\top \mathbf{y} \right) dy_1 \dots dy_p \\
&= \prod_{j=1}^p \left(\int_{-\infty}^{+\infty} \frac{\exp(i t_j y_j)}{(2\pi)^{p/2}} \exp \left(-\frac{1}{2} y_j^2 \right) dy_j \right) \\
&= \prod_{j=1}^p \left(\int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{p/2}} \exp \left(-\frac{1}{2} (y_j - i t_j)^2 - \frac{1}{2} t_j^2 \right) dy_j \right) \\
&= \prod_{j=1}^p \left(\exp \left(-\frac{1}{2} t_j^2 \right) \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{p/2}} \exp \left(-\frac{1}{2} z_j^2 \right) dz_j \right) \\
&= \prod_{j=1}^p \left(\exp \left(-\frac{1}{2} t_j^2 \right) \right) = \exp \left(-\frac{1}{2} \mathbf{t}^\top \mathbf{t} \right).
\end{aligned}$$

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

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

□

Remark 2.2. *Denote the characteristic function of $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\phi_{\mathbf{x}}(\mathbf{t}) = \exp(i \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t})$. For $\mathbf{z} = \mathbf{D}\mathbf{x}$, the characteristic function of \mathbf{z} is*

$$\phi_{\mathbf{z}}(\mathbf{t}) = \mathbb{E} [\exp(i \mathbf{t}^\top \mathbf{z})] = \mathbb{E} [\exp(i \mathbf{t}^\top \mathbf{D}\mathbf{x})] = \mathbb{E} [\exp(i (\mathbf{D}^\top \mathbf{t})^\top \mathbf{x})] = \exp \left(i \mathbf{t}^\top (\mathbf{D}\boldsymbol{\mu}) - \frac{1}{2} \mathbf{t}^\top (\mathbf{D}^\top \boldsymbol{\Sigma} \mathbf{D}) \mathbf{t} \right)$$

which implies $\mathbf{z} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}^\top \boldsymbol{\Sigma} \mathbf{D})$ and we prove Theorem 2.4.

Theorem 2.9. *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 $\text{Cov}[\mathbf{y}] = \boldsymbol{\Sigma}$. 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

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

Set $t = 1$, then we have

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

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

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

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

3 Estimation of the Mean Vector and the Covariance

Theorem 3.1. *If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p < N$, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are*

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

respectively.

Proof. The logarithm of the likelihood function is

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

We have

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

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

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

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

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

then

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

We can verify $l(\boldsymbol{\Psi})$ is concave on the domain of symmetric positive definite matrices, which means the maximum is taken by $\frac{\partial f(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = \mathbf{0}$, 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 3.1. *If $\mathbf{D} \in \mathbb{R}^{p \times p}$ is positive definite, the maximum of*

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

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

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

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

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

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

$$\begin{aligned} g(\mathbf{H}) &= -N \ln \det(\mathbf{D}) + N \ln \det(\mathbf{H}) - \text{tr}(\mathbf{H}) \\ &= -N \ln \det(\mathbf{D}) + N \ln (\det(\mathbf{T}))^2 - \text{tr}(\mathbf{T} \mathbf{T}^\top) \end{aligned}$$

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

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

$$\mathbf{G} = \frac{1}{N} \mathbf{D}.$$

□

Theorem 3.2. 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

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

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^* .

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

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

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. □

Corollary 3.1. If $\mathbf{x}_1, \dots, \mathbf{x}_N$ 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

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

□

Theorem 3.3. Suppose $\mathbf{x}_1, \dots, \mathbf{x}_N$ 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 $\boldsymbol{\nu} = \sum_{\beta=1}^N c_{\alpha\beta} \boldsymbol{\mu}_\beta$ for $\alpha = 1, \dots, N$ and $\mathbf{y}_1, \dots, \mathbf{y}_N$ are independent.

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

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

The covariance matrix between \mathbf{y}_α and \mathbf{y}_γ is

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

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

$$\text{Cov}[\mathbf{y}_\alpha, \mathbf{y}_\gamma] = \sum_{\beta=1}^N c_{\alpha\beta} c_{\gamma\beta} \boldsymbol{\Sigma} = \delta_{\alpha\gamma} \boldsymbol{\Sigma}.$$

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

Lemma 3.2. *If*

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

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

Proof. Let

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

We have

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

\square

Remark 3.1. We can also write $\mathbf{y}_\alpha = \mathbf{X}^\top \mathbf{c}_\alpha$ and $\mathbf{Y} = \mathbf{C}\mathbf{X}$ by defining \mathbf{Y} like \mathbf{X} .

Theorem 3.4. 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

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

By Lemma 3.2, we have

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

Lemma 3.2 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 $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \mathbf{A}$. 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 3.3 implies

$$\mathbb{E}[\bar{\mathbf{x}}] = \mathbb{E} \left[\frac{1}{\sqrt{N}} \sum_{\beta=1}^N b_{N\beta} \mathbf{x}_\beta \right] = \frac{1}{\sqrt{N}} \sum_{\beta=1}^N \frac{1}{\sqrt{N}} \boldsymbol{\mu} = \boldsymbol{\mu}, \quad \text{and} \quad \text{Cov}[\bar{\mathbf{x}}] = \frac{1}{N} \text{Cov} \left[\sum_{\beta=1}^N b_{N\beta} \mathbf{x}_\beta \right] = \frac{1}{N} \boldsymbol{\Sigma}.$$

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

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

and Theorem 3.3 implies $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. □

Theorem 3.5. Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be p -dimensional random vector and they are independent. Denote

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

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

$$\mathbb{E}[\hat{\boldsymbol{\Sigma}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

Proof. We have

$$\boldsymbol{\Sigma} = \text{Cov}[\mathbf{x}_\alpha] = \mathbb{E}[(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{x}_\alpha\mathbf{x}_\alpha^\top - \mathbf{x}_\alpha\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{x}_\alpha^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top] = \mathbb{E}[\mathbf{x}_\alpha\mathbf{x}_\alpha^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

and

$$\frac{1}{n}\boldsymbol{\Sigma} = \text{Cov}[\bar{\mathbf{x}}] = \text{Cov}[(\bar{\mathbf{x}} - \mathbb{E}[\bar{\mathbf{x}}])(\bar{\mathbf{x}} - \mathbb{E}[\bar{\mathbf{x}}])^\top] = \text{Cov}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Hence, we obtain

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

□

Theorem 3.6. Using the notation of Theorem 3.1, if $N > p$, the probability is 1 of drawing a sample so that

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

is positive definite.

Proof. The proof of Theorem 3.1 shows that $\mathbf{A} = \tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}}$ where

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

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

$$\Pr(\text{rank}(\tilde{\mathbf{Z}}) = p) = 1.$$

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

$$\begin{aligned} & \Pr(\mathbf{z}_1, \dots, \mathbf{z}_p \text{ are linearly dependent}) \\ & \leq \sum_{i=1}^p \Pr(\mathbf{z}_i \in \text{span}\{\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_i, \dots, \mathbf{z}_p\}) \\ & = p \Pr(\mathbf{z}_1 \in \text{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}) \\ & = p \mathbb{E}[\Pr(\mathbf{z}_1 \in \text{span}\{\mathbf{z}_2, \mathbf{z}_3, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p)] \\ & \leq p \mathbb{E}[\Pr(\text{there exists non-zero } \boldsymbol{\alpha} \in \mathbb{R}^p \text{ such that } \boldsymbol{\alpha}^\top \mathbf{z}_1 = 0 \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p)] \\ & = p \mathbb{E}[0] = 0 \end{aligned}$$

The second equality is obtained as follows

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

The second inequality is due to

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

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

Theorem 3.7. If $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independent observations from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

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

where

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

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

$$\prod_{\alpha=1}^M n(\mathbf{x}_\alpha \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

where the last step is due to

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

Hence, the density is a function of $\mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \{\bar{\mathbf{x}}, \mathbf{S}\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \mathbf{\Sigma}\}$. If $\boldsymbol{\mu}$ is given, it is a function of $\mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \sum_{\alpha=1}^N (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top$ and $\boldsymbol{\theta} = \mathbf{\Sigma}$. If $\mathbf{\Sigma}$ is given, it is a function of $\mathbf{t}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \bar{\mathbf{x}}$ (since \mathbf{S} can be viewed a function of \mathbf{t} for given) and $\boldsymbol{\theta} = \boldsymbol{\mu}$. \square