Lecture Notes of Multivariate Statistics
Lecture 01

Weizhong Zhang
School of Data Science, Fudan University

March 28, 2023

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

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

$$r_{jj} = \left\| \mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij} \mathbf{q}_i \right\|_2 = 0,$$

then $\mathbf{a}_j$ must be a linear combination of $\{\mathbf{a}_1, \cdots, \mathbf{a}_{j-1}\}$, which validates 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 $i = 1, \ldots, n' - 1$ and $j < i$. Then for all $k = 1, 2, \ldots, n' - 1$, we have

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

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

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

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

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

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

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 $\mathbf{L} \in \mathbb{R}^{\times n}$ 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 $\widetilde{\mathbf{A}} \in \mathbb{R}^{(n-1)\times(n-1)}$ can be written as

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

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

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

Let

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

We have

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

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

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

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

Proof. Because the Hessian is continuous and positive definite at x^*, we can choose a radius $r > 0$ so that $\nabla^2 f(\mathbf{x})$ remains positive definite for all $\mathbf{x}$ in the open ball $\mathcal{D} = \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}^*\|_2 < r\}$. Taking any nonzero vector $\mathbf{p}$ with $\|\mathbf{p}\|_2 < r$, we have $\mathbf{x}^* + \mathbf{p} \in \mathcal{D}$ and so

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \mathbf{p}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} = f(\mathbf{x}^*) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p},$$

where $\mathbf{z} = \mathbf{x}^* + t\mathbf{p}$ for some $t \in (0, 1)$. Since $\mathbf{z} \in \mathcal{D}$, we have $\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} > 0$, and therefore $f(\mathbf{x}^* + \mathbf{p}) > f(\mathbf{x}^*)$, giving the result. □

Theorem 1.5. *Suppose $\mathbf{x}^*$ is a local minimizer of twice differentiable $f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}$.*

Proof. Suppose for contradiction that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Define the vector $p = -\nabla f(\mathbf{x}^*)$, which leads to that $\mathbf{p}^\top \nabla f(\mathbf{x}^*) < 0$. Because ∇f is continuous near $\mathbf{x}^*$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla f(\mathbf{x}^* + t\mathbf{p}) < 0,$$

for all for any $t \in [0, T]$. We have by Taylor's theorem that

$$f(\mathbf{x}^* + \bar{t}\mathbf{p}) = f(\mathbf{x}^*) + \bar{t}\mathbf{p}^\top \nabla f(x^* + t\mathbf{p}),$$

for some $t \in (0, \bar{t})$. Therefore, $f(x^* + \bar{t}\mathbf{p}) < f(x^*)$ for all $\bar{t} \in (0, T]$. We have found a direction leading away from x^* along which f decreases, so x^* is not a local minimizer, and we have $\nabla^2 f(\mathbf{x}) = \mathbf{0}$.

For contradiction, assume that $\nabla^2 f(\mathbf{x}^*)$ is not positive semidefinite. Then we can choose a vector $\mathbf{p}$ such that $\mathbf{p}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{p} < 0$. Because $\nabla^2 f(\mathbf{x})$ is continuous near $\mathbf{x}^*$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p} < 0$$

for all $t \in [0, T]$. By doing a Taylor series expansion around x^*, we have for all $\bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$ that

$$f(\mathbf{x}^* + \bar{t}\mathbf{p}) = f(\mathbf{x}^*) + \bar{t}\mathbf{p}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\bar{t}^2\mathbf{p}^\top \nabla^2(\mathbf{x}^* + t\mathbf{p})\bar{t}^2\mathbf{p} < f(\mathbf{x}^*).$$

We have found a direction from $\mathbf{x}^*$ along which f is decreasing, and so again, $\mathbf{x}^*$ is not a local minimizer. □

Theorem 1.6. *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 $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top$ 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

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

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

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

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

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

We also have

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

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

2 The Multivariate Normal Distributions

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

$$
\begin{aligned}
f(x, y) =& \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x)G(y)}{\partial x \partial y} \\
=& \frac{\mathrm{d}F(x)}{\mathrm{d}x} \frac{\mathrm{d}G(y)}{\mathrm{d}y} \\
=& f(x)g(y).
\end{aligned}
$$

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

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

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

$$
\int_{y}^{y+\Delta y} g(v) \, \mathrm{d}v = g(y^*) \Delta y,
$$

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

$$
\int_{y}^{y+\Delta y} f(u, v) \, \mathrm{d}v = f(u, y^*(u)) \Delta y,
$$

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

$$\Pr\{x_1 \le X \le x_2 \mid y_1 \le Y \le y_2\} = \frac{\int_{x_1}^{x_2}\int_{y_1}^{y_2} f(u,v)\,\mathrm{d}v\,\mathrm{d}u}{\int_{y_1}^{y_2} g(v)\mathrm{d}v}$$

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

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

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

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

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

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

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

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

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

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

and the density of $Y_1, \dots, Y_p$ be $g(\mathbf{y})$. Then we have

$$\int_{\mathbf{u}(\Omega)} g(\mathbf{y})\mathrm{d}\mathbf{y} = \int_{\Omega} g\left(\mathbf{u}(\mathbf{x})\right) \mathrm{abs}(|\mathbf{J}(\mathbf{x})|)\mathrm{d}\mathbf{x}, \tag{3}$$

and

$$f(\mathbf{x}) = g\left(\mathbf{u}(\mathbf{x})\right) \mathrm{abs}(|\mathbf{J}(\mathbf{x})|), \tag{4}$$

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

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

- Consider notation Ω such that

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

where $x_1 \le x_1', x_2 \le x_2', \ldots, x_p \le x_p'$. 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 (f is density functions of $\mathbf{x}$ on Ω; g is density function of $\mathbf{y}$ on $\mathbf{u}(\Omega)$; $\mathbf{y} = \mathbf{u}(\mathbf{x})$ means $\mathbf{x}$ and $\mathbf{y} = \mathbf{u}(\mathbf{x})$ are one-to-one mapping).

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

Since it holds for any Ω, then

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

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*

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

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

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

Proof. We have

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

□

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 $\mathbf{C}^\top\mathbf{AC}=\mathbf{I}$ and $\mathbf{C}$ is non-singular. Define $\mathbf{y}=\mathbf{C}^{-1}(\mathbf{x}-\mathbf{b})$, then

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

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 $\mathbf{x}=\mathbf{Cy}+\mathbf{b}$ and $\mathbb{E}[\mathbf{x}]=\mathbf{C}\mathbb{E}[\mathbf{y}]+\mathbf{b}$. The transformation implies the density function of $\mathbf{y}$ is

$$
\begin{aligned}
g(\mathbf{y})=&\det(\mathbf{C})K\exp\left(-\frac{1}{2}(\mathbf{Cy}+\mathbf{b}-\mathbf{b})^\top\mathbf{A}(\mathbf{Cy}+\mathbf{b}-\mathbf{b})\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\
=&\det(\mathbf{C})K\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{C}^\top\mathbf{ACy}\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\
=&K\det(\mathbf{C})\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\
=&\frac{\det(\mathbf{C})}{\sqrt{(2\pi)^p\det(\mathbf{A})}}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}y_i^2\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p \\
=&\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}y_i^2\right)\,\mathrm{d}y_1\ldots\mathrm{d}y_p.
\end{aligned}
$$

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

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

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

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 $\mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathrm{Cov}[\mathbf{y}]\mathbf{C}^\top = \mathbf{C}\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{C}^\top$. For each $i \neq j$, we have

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

We also have

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

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

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

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

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

which implies $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{C}^\top = \mathbf{C}\mathbf{C}^\top$. 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 $f(x)$ be the density of $\mathbf{x}$ such that

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

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

$$g(\mathbf{y})$$

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

where we use the fact

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

□

Theorem 2.2. *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

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

such that

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

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

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

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

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

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

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

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

$\square$

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

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

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

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

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

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

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. □

Singular Normal Distribution The mass is concentrated on a linear set $\mathcal{S}$. For any $x \notin \mathcal{S}$, there exists $\mathcal{B}(x, r)$ such that $r > 0$ and $\mathcal{B} \cap \mathcal{S} = \emptyset$. If the distribution of x has density function f, then $f(x) = 0$ holds for any $x \notin \mathcal{S}$. Since the measure of $\mathcal{S}$ is zero, we have $f(x) = 0$ almost everywhere, which means the integration of $f(x)$ on the whole space is 0.

Conditional Distribution by Schur Complement Recall that

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{B}\mathbf{D}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I}\end{bmatrix},$$

which directly means the inverse of covariance of Normal distribution.

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 $\mathrm{Cov}[\mathbf{z}] = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top$. 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 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

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

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

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

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

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

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

Theorem 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*

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

for $i = 1, \ldots, q$, 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

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

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

Remark 2.1. *Observe that*

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

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*

$$\operatorname{Corr}\left(x_i, \boldsymbol{\beta}_{(i)}^\top \mathbf{x}^{(2)}\right) \geq \operatorname{Corr}\left(x_i, \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}\right).$$

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

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

Using Theorem 2.5, we have

$$\begin{aligned}
&\operatorname{Var}\big[x_i^{(11.2)}\big] \leq \operatorname{Var}\big[x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}\big] \\
\Longleftrightarrow &\mathbb{E}\big[x_i - \mu_i - \boldsymbol{\beta}_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}))\big]^2 \leq \mathbb{E}\big[x_i - \mu_i - \boldsymbol{\alpha}^\top \big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big)\big]^2 \\
\Longleftrightarrow \quad &\operatorname{Var}[x_i] - \mathbb{E}\big[(x_i - \mu_i)\boldsymbol{\beta}_{(i)}^\top \big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big)\big] + \operatorname{Var}\big[\boldsymbol{\beta}_{(i)}^\top \mathbf{x}^{(2)}\big] \\
&\leq \operatorname{Var}[x_i] - \mathbb{E}\big[(x_i - \mu_i)\boldsymbol{\alpha}^\top \big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big)\big] + \operatorname{Var}\big[\boldsymbol{\alpha}^\top \mathbf{x}^{(2)}\big] \\
\Longleftrightarrow &\frac{\mathbb{E}\big[(x_i - \mu_i)\boldsymbol{\alpha}^\top \big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big)\big]}{\sqrt{\operatorname{Var}[x_i]}\sqrt{\operatorname{Var}\big[\boldsymbol{\alpha}^\top \mathbf{x}^{(2)})\big]}} \leq \frac{\mathbb{E}\big[(x_i - \mu_i)\boldsymbol{\beta}_{(i)}^\top \big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big)\big]}{\sqrt{\operatorname{Var}[x_i]}\sqrt{\operatorname{Var}\big[\boldsymbol{\beta}^\top \mathbf{x}^{(2)})\big]}} \\
\Longleftrightarrow &\frac{\operatorname{Cov}\big[x_i, \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}\big]}{\sqrt{\operatorname{Var}[x_i]}\sqrt{\operatorname{Var}\big[\boldsymbol{\alpha}^\top \mathbf{x}^{(2)})\big]}} \leq \frac{\mathbb{E}\big[x_i, \boldsymbol{\beta}_{(i)}^\top \mathbf{x}^{(2)}\big]}{\sqrt{\operatorname{Var}[x_i]}\sqrt{\operatorname{Var}\big[\boldsymbol{\beta}^\top \mathbf{x}^{(2)})\big]}}
\end{aligned}$$

□

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*

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

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

$$\begin{aligned}
&\mathbb{E}[g(\mathbf{x})] \\
=&\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(\mathbf{x})f(\mathbf{x})\,\mathrm{d}x_1 \dots \mathrm{d}x_p \\
=&\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g^{(1)}(\mathbf{x}^{(1)})g^{(2)}(\mathbf{x}^{(2)})f^{(1)}(\mathbf{x}^{(1)})f^{(2)}(\mathbf{x}^{(2)})\,\mathrm{d}x_1 \dots \mathrm{d}x_p \\
=&\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g^{(1)}(\mathbf{x}^{(1)})f^{(1)}(\mathbf{x}^{(1)})\,\mathrm{d}x_1 \dots \mathrm{d}x_q \cdot \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g^{(2)}(\mathbf{x}^{(2)})f^{(2)}(\mathbf{x}^{(2)})\,\mathrm{d}x_{q+1} \dots \mathrm{d}x_p \\
=&\mathbb{E}\big[g^{(1)}(\mathbf{x}^{(1)})\big]\mathbb{E}\big[g^{(2)}(\mathbf{x}^{(2)})\big].
\end{aligned}$$

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

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

and

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

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

□

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

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

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

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

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

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

□

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

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

which implies $\mathbf{z} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}^\top \mathbf{\Sigma}\mathbf{D})$ *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 $\mathrm{Cov}[\mathbf{y}] = \boldsymbol{\Sigma}$. 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

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

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

Theorem 2.10. *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. Prove* $\mathbf{z} \sim \mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

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

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

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

3 Estimation of the Mean Vector and the Covariance

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

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

respectively.

Proof. The logarithm of the likelihood function is

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

We have

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

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 $l(\Psi)$ 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 3.1. *If $\mathbf{D}\in\mathbb{R}^{p\times p}$ is positive definite, the maximum of*

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

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 $\mathbf{L}$ 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

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

□

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*

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

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^*.

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 3.3. *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

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

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

Proof. The definition means

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

The definition of $\hat{\boldsymbol{\theta}}^* = \boldsymbol{\phi}(\hat{\boldsymbol{\theta}})$ means

$$f(\hat{\boldsymbol{\theta}}) = \sup_{\hat{\boldsymbol{\theta}}^*=\boldsymbol{\phi}(\boldsymbol{\theta})} f(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}^*)$$

Since $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ maximize $f(\boldsymbol{\theta})$, we have

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

which implies $\hat{\boldsymbol{\theta}}^*$ maximize $g(\boldsymbol{\theta}^*)$. □

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 3.4. *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 $\boldsymbol{\nu}_\alpha = \sum_{\beta=1}^N c_{\alpha\beta}\boldsymbol{\mu}_\beta$ 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 $\mathbf{y}_\alpha$ 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

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

$$
\begin{aligned}
&= \sum_{\beta=1}^{N} \sum_{\xi=1}^{N} c_{\alpha\beta} c_{\gamma\xi} \delta_{\beta\xi} \mathbf{\Sigma} \\
&= \sum_{\beta=1}^{N} c_{\alpha\beta} c_{\gamma\beta} \mathbf{\Sigma},
\end{aligned}
$$

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

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

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

Lemma 3.2. *If*

$$
\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix} = \begin{bmatrix} c_1^\top \\ c_2^\top \\ \vdots \\ c_N^\top \end{bmatrix} \in \mathbb{R}^{N \times N}
$$

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

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

We have

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

□

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 3.5. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_N$ *be independent, each distributed according to* $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. *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 3.4 implies

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

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

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

and Theorem 3.4 implies $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. □

Theorem 3.6. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be p-dimensional random vector and they are independent. Denote*

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

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

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

Proof. We have

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

and

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

Hence, we obtain

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

□

Theorem 3.7. *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 $\mathbf{A} = \widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}$ where

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

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

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

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

$$
\begin{aligned}
&\Pr(\mathbf{z}_1, \dots, \mathbf{z}_p \text{ are linearly dependent}) \\
\leq &\sum_{i=1}^{p} \Pr\left(\mathbf{z}_i \in \operatorname{span}\{\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_i, \dots, \mathbf{z}_p\}\right) \\
= &p \Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}\right) \\
= &p\, \mathbb{E}\left[\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \mathbf{z}_3, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)\right] \\
= &p\mathbb{E}[0] = 0
\end{aligned}
$$

The second equality is obtained as follows

$$
\begin{aligned}
&\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}\right) \\
= &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\}, \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right) \mathrm{d}\boldsymbol{\alpha}_2 \dots \mathrm{d}\boldsymbol{\alpha}_p \\
= &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right) \Pr\left(\mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right) \mathrm{d}\boldsymbol{\alpha}_2 \dots \mathrm{d}\boldsymbol{\alpha}_p \\
= &\mathbb{E}\left[\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)\right] \\
= &0
\end{aligned}
$$

The last equality holds since $\Pr\left(\mathbf{z}_1 \in \operatorname{span}\{\mathbf{z}_2, \dots, \mathbf{z}_p\} \mid \mathbf{z}_2 = \boldsymbol{\alpha}_2, \dots, \mathbf{z}_p = \boldsymbol{\alpha}_p\right)$ is the probability of the event that $\mathbf{z}_1$ lies in a subspace with the dimension no higher than $p - 1$. □

Theorem 3.8. *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

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

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

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

where the last step is due to

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

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

Hence, the density is a function of $\mathbf{t}(\mathbf{x}_1,\dots,\mathbf{x}_N)=\{\bar{\mathbf{x}},\mathbf{S}\}$ and $\boldsymbol{\theta}=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$. 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 $\boldsymbol{\theta}=\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\theta}=\boldsymbol{\mu}$. □

Theorem 3.9 (Theorem 3.4.2, Page 84)**.** *The sufficient set of statistics $\bar{\mathbf{x}}$, $\mathbf{S}$ is complete for $\boldsymbol{\mu},\boldsymbol{\Sigma}$ when the sample is drawn from $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.*

Proof. We introduce $\mathbf{z}_1,\dots,\mathbf{z}_N$ by following the proof of Theorem 3.5. For any function $g(\bar{\mathbf{x}},n\mathbf{S})$, we have

$$
\begin{aligned}
&0\equiv\mathbb{E}[g(\bar{\mathbf{x}},n\mathbf{S})]\\
=&\int\cdots\int K\big(\det(\boldsymbol{\Sigma})\big)^{-\frac{N}{2}}g\left(\bar{\mathbf{x}},\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)\exp\left(-\frac{1}{2}\left(\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}_\alpha+N(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right)\right)\mathrm{d}\mathbf{z}_1\dots\mathrm{d}\mathbf{z}_{N-1}\,\mathrm{d}\bar{\mathbf{x}}.
\end{aligned}
$$

for any $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where $K=\sqrt{N}(2\pi)^{-\frac{1}{2}pN}$. Let $\boldsymbol{\Sigma}^{-1}=\mathbf{I}-2\boldsymbol{\Omega}$ such that symmetric $\boldsymbol{\Omega}$ and $\mathbf{I}-2\boldsymbol{\Omega}\succ 0$. Let $\boldsymbol{\mu}=(\mathbf{I}-2\boldsymbol{\Omega})^{-1}\mathbf{t}=\boldsymbol{\Sigma}\mathbf{t}$. Then, we have

$$
\begin{aligned}
&0\\
\equiv&\int\cdots\int K\big(\det(\boldsymbol{\Sigma})\big)^{-\frac{N}{2}}g\left(\bar{\mathbf{x}},\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)\\
&\exp\left(-\frac{1}{2}\left(\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}_\alpha+N\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}-2N\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}+N\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)\mathrm{d}\mathbf{z}_1\dots\mathrm{d}\mathbf{z}_{N-1}\,\mathrm{d}\bar{\mathbf{x}}\\
=&\int\cdots\int K\big(\det(\boldsymbol{\Sigma})\big)^{-\frac{N}{2}}g\left(\bar{\mathbf{x}},\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)\\
&\exp\left(-\frac{1}{2}\left(\sum_{\alpha=1}^{N-1}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)+N\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)-2N\bar{\mathbf{t}}^\top\bar{\mathbf{x}}+N\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}\right)\right)\mathrm{d}\mathbf{z}_1\dots\mathrm{d}\mathbf{z}_{N-1}\,\mathrm{d}\bar{\mathbf{x}}\\
=&\int\cdots\int K\big(\det(\mathbf{I}-2\boldsymbol{\Omega})\big)^{\frac{N}{2}}g\left(\bar{\mathbf{x}},\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)\\
&\exp\left(-\frac{1}{2}\left(\mathrm{tr}\left((\mathbf{I}-2\boldsymbol{\Omega})\left(\sum_{\alpha=1}^{N-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top+N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)\right)-2N\bar{\mathbf{t}}^\top\bar{\mathbf{x}}+N\mathbf{t}^\top(\mathbf{I}-2\boldsymbol{\Omega})^{-1}\mathbf{t}\right)\right)\mathrm{d}\mathbf{z}_1\dots\mathrm{d}\mathbf{z}_{N-1}\,\mathrm{d}\bar{\mathbf{x}}\\
=&\big(\det(\mathbf{I}-2\boldsymbol{\Omega})\big)^{\frac{N}{2}}\exp\left(-\frac{1}{2}N\mathbf{t}^\top(\mathbf{I}-2\boldsymbol{\Omega})^{-1}\mathbf{t}\right)\\
&\int\cdots\int g\left(\bar{\mathbf{x}},\mathbf{B}-N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)\exp\left(\mathrm{tr}(\boldsymbol{\Omega}\mathbf{B})+\mathbf{t}^\top(N\bar{\mathbf{x}})\right)n\left(\bar{\mathbf{x}}\mid\mathbf{0},\frac{1}{N}\mathbf{I}\right)\prod_{\alpha=1}^{N-1}n(\mathbf{z}_\alpha\mid\mathbf{0},\mathbf{I})\,\mathrm{d}\mathbf{z}_1\dots\mathrm{d}\mathbf{z}_{N-1}\,\mathrm{d}\bar{\mathbf{x}}\\
=&\big(\det(\mathbf{I}-2\boldsymbol{\Omega})\big)^{\frac{N}{2}}\exp\left(-\frac{1}{2}N\mathbf{t}^\top(\mathbf{I}-2\boldsymbol{\Omega})^{-1}\mathbf{t}\right)\\
&\int g\left(\bar{\mathbf{x}},\mathbf{B}-N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)\exp\left(\mathrm{tr}(\boldsymbol{\Omega}\mathbf{B})+\mathbf{t}^\top(N\bar{\mathbf{x}})\right)n\left(\bar{\mathbf{x}}\mid\mathbf{0},\frac{1}{N}\mathbf{I}\right)\,\mathrm{d}\bar{\mathbf{x}}\\
=&\big(\det(\mathbf{I}-2\boldsymbol{\Omega})\big)^{\frac{N}{2}}\exp\left(-\frac{1}{2}N\mathbf{t}^\top(\mathbf{I}-2\boldsymbol{\Omega})^{-1}\mathbf{t}\right)\mathbb{E}\left[g\left(\bar{\mathbf{x}},\mathbf{B}-N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right)\exp\left(\mathrm{tr}(\boldsymbol{\Omega}\mathbf{B})+\mathbf{t}^\top(N\bar{\mathbf{x}})\right)\right].
\end{aligned}
$$

where $\mathbf{B} = \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top + N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top$. Thus

$$\begin{aligned}0 \equiv& \mathbb{E}\left[g\left(\bar{\mathbf{x}}, \mathbf{B} - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right) \exp\left(\operatorname{tr}(\mathbf{\Omega B}) + \mathbf{t}^\top(N\bar{\mathbf{x}})\right)\right] \\ =& \iint g\left(\bar{\mathbf{x}}, \mathbf{B} - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right) \exp\left(\operatorname{tr}(\mathbf{\Omega B}) + \mathbf{t}^\top(N\bar{\mathbf{x}})\right) h(\bar{\mathbf{x}}, \mathbf{B}) \,\mathrm{d}\bar{\mathbf{x}} \,\mathrm{d}\mathbf{B}\end{aligned}$$

where $h(\bar{\mathbf{x}}, \mathbf{B})$ is the joint density of $\bar{\mathbf{x}}$ and $\mathbf{B}$. Consider that

$$\iint g\left(\bar{\mathbf{x}}, \mathbf{B} - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right) \exp\left(\operatorname{tr}(\mathbf{\Omega B}) + \mathbf{t}^\top(N\bar{\mathbf{x}})\right) h(\bar{\mathbf{x}}, \mathbf{B}) \,\mathrm{d}\bar{\mathbf{x}} \,\mathrm{d}\mathbf{B}$$

is the Laplace transform of $g\left(\bar{\mathbf{x}}, \mathbf{B} - N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\right) h(\bar{\mathbf{x}}, \mathbf{B})$. Then we have $g(\bar{\mathbf{x}}, n\mathbf{S})h(\bar{\mathbf{x}}, \mathbf{B}) = 0$ almost everywhere. Hence, we have

$$\begin{aligned}0 =& \iint |g(\bar{\mathbf{x}}, n\mathbf{S})h(\bar{\mathbf{x}}, \mathbf{B})| \,\mathrm{d}\bar{\mathbf{x}} \,\mathrm{d}\mathbf{B} \\ =& \iint |g(\bar{\mathbf{x}}, n\mathbf{S})| h(\bar{\mathbf{x}}, \mathbf{B})| \,\mathrm{d}\bar{\mathbf{x}} \,\mathrm{d}\mathbf{B} \\ =& \iint |g(\bar{\mathbf{x}}, n\mathbf{S})| \mathrm{d}m(\bar{\mathbf{x}}, \mathbf{B}).\end{aligned}$$

Hence, we have $g(\bar{\mathbf{x}}, n\mathbf{S}) = 0$ almost everywhere. □

Cramer-Rao Inequality We first give some lemmas. We denote the density of observation 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 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

Lemma 3.3. *We have* $\mathbb{E}[\mathbf{s}] = \mathbf{0}$.

Proof. We have

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

□

Remark 3.2. *Similarly, we also have*

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

Lemma 3.4. *For unbiased estimator* $\mathbf{t}$ *of* $\boldsymbol{\theta}$*, we have* $\operatorname{Cov}[\mathbf{t}, \mathbf{s}] = \mathbf{I}$.