Lecture Notes of Multivariate Statistics

Luo Luo

School of Data Science, Fudan University

June 16, 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 $\{\mathbf{a}_1, \cdots, \mathbf{a}_i\}$, 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 $\{\mathbf{a}_1, \cdots, \mathbf{a}_{n-1}\}$, 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

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

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 $\mathbf{V}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ □

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. □

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

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

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

Proof. 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

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

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

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

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

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

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 $y \le y^* \le y + \Delta y$. 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 $X_1, \ldots, X_p$ be $f(x_1, \ldots, x_p)$. Consider the p real-valued functions $\mathbf{u} : \mathbb{R}^p \to \mathbb{R}^p$ such that

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

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

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

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

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

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

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

A roughly proof for above results:

- 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 $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\boldsymbol{\Sigma}$ is diagonal with nonnegative entries. Multiplying by $\mathbf{V}^\top$ doesn't change the measure of $\mathcal{S}$. Multiplying by $\boldsymbol{\Sigma}$ scales along each axis, so the measure gets multiplied by $|\det(\boldsymbol{\Sigma})| = |\det(\mathbf{A})|$. Multiplying by $\mathbf{U}$ doesn't change the measure.
- We consider the probability of $\mathbf{x}$ in Ω and $\mathbf{y}$ in $\mathbf{u}(\Omega)$; and partition Ω into $\{\Omega_i\}_i$. 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

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

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*

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

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 $\mathbf{y} = \mathbf{Dx} + \mathbf{f} \in \mathbb{R}^l$, 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 $\mathbf{A} = \mathbf{U\Lambda U}^\top$, then we take $\mathbf{C} = \mathbf{U\Lambda}^{-1/2}$ 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

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

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

$$g(\mathbf{y}) = \det(\mathbf{C}) \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K \exp\left(-\frac{1}{2}(\mathbf{Cy} + \mathbf{b} - \mathbf{b})^\top\mathbf{A}(\mathbf{Cy} + \mathbf{b} - \mathbf{b})\right) \mathrm{d}y_1 \dots \mathrm{d}y_p$$

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

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

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

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

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

We also have

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

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

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

and $g(\mathbf{y})$ be the density function of $\mathbf{y}$. The relation $\mathbf{x} = \mathbf{C}^{-1}\mathbf{y}$ implies $g(\mathbf{y}) = f(\mathbf{u}^{-1}(\mathbf{y}))|\det(\mathbf{J}^{-1}(\mathbf{y}))|$ with $\mathbf{u}(\mathbf{x}) = \mathbf{C}\mathbf{x}$, $\mathbf{u}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}\mathbf{y}$ and $\mathbf{J}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}$. Hence, we have

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

where we use the fact

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

$\square$

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

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

such that

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

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

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

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

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

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

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

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

□

Theorem 2.3. *If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$, the marginal distribution of any set of components of $\mathbf{x}$ is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.*

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

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

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

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

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

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. □

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

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

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 $s \le r$, we conclude $\mathbf{u}_1$ has a non-singular normal distribution. Let $\mathbf{F}^{-1} = [\mathbf{G}_1, \mathbf{G}_2]$, where $\mathbf{G}_1 \in \mathbb{R}^{r\times s}$ 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. ☐

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 $i = 1, \dots, q$, 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

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

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

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

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

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

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

□

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

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*

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

respectively.

Proof. The logarithm of the likelihood function is

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

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

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

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

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. □

Corollary 3.1. *If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ constitutes a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\rho_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$. Then the maximum likelihood estimator of ρ_{ij} is*

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

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

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

□

Theorem 3.3. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent, where $\mathbf{x}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma})$. Let $\mathbf{C} \in \mathbb{R}^{N\times N}$ be an orthogonal matrix, then*

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

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

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

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

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

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

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*

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

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

Proof. Let

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

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 3.4. *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

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

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

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

and Theorem 3.3 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*

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

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

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

Hence, we obtain

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

□

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 $\mathbf{A} = \widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}$ 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 $\operatorname{rank}(\hat{\boldsymbol{\Sigma}}) = \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\widetilde{\mathbf{Z}})$. Then the probability is 1 of $\hat{\boldsymbol{\Sigma}} \succ \mathbf{0}$ is equivalent to

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

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

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

The second equality is obtained as follows

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

The second inequality is due to

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

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 $\boldsymbol{\Sigma} \succ \mathbf{0}$. □

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

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

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

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

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

where the last step is due to

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

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 3.8 (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 3.4. 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

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

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

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

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

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

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

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 $g(\bar{\mathbf{x}},n\mathbf{S})=0$ for 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* $\operatorname{Cov}[\mathbf{t},\mathbf{s}]=\mathbf{I}$.

Proof. We have

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

where the last line use the integrate by part

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

and $\mathbb{E}[t_j] = \theta_j$. □

Theorem 3.9. *Under the regularity condition (everything is well-defined, integration and differentiation can be swapped), we have*

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

where $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$ and $f(\mathbf{x}, \boldsymbol{\theta})$ is the density of the distribution with respect to the components of $\boldsymbol{\theta}$.

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

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

We let $\mathbf{b}$ which satisfies $\mathbf{b}^\top \mathrm{Var}[\mathbf{s}] \mathbf{b} = 1$, then

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

which implies $\mathbf{a}^\top \mathrm{Var}[\mathbf{t}] \mathbf{a} \geq \mathbf{a}^\top \left(\mathrm{Var}[\mathbf{s}] \right)^{-1} \mathbf{a}$ for any nonzero $\mathbf{a}$. Hence, we have

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

□

Theorem 3.10. *Let p-component vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots$ be i.i.d with means $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance matrices $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$. Then the limiting distribution of*

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

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

Proof. Let

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

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

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

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

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

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

$$\phi_n(\mathbf{t}, 1) = \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{t}^\top \frac{1}{\sqrt{n}}\sum_{\alpha=1}^{n}(\mathbf{y}_\alpha - \boldsymbol{\nu})\right)\right] = \exp\left(-\frac{1}{2}\mathbf{t}^\top\mathbf{T}\mathbf{t}\right)$$

for any $\mathbf{t} \in \mathbb{R}^p$. Since $\exp\left(-\frac{1}{2}\mathbf{t}^\top\mathbf{T}\mathbf{t}\right)$ is continuous at $\mathbf{t} = \mathbf{0}$, the convergence is uniform in some neighborhood of $\mathbf{t} = \mathbf{0}$. The theorem follows. □

Theorem 3.11. *If $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed, each x_α according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Psi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \dots, \mathbf{x}_N$ is normal with mean*

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

and covariance matrix

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

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

$$\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{N}\boldsymbol{\Sigma}\right)$$

and is independent of $\boldsymbol{\mu}$. Then we have

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

which implies $\bar{\mathbf{x}} \sim \mathcal{N}\left(\boldsymbol{\nu}, \boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)$. Since we have

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

then

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

Consider the conditional distribution of $\boldsymbol{\mu}$ given $\bar{\mathbf{x}}$, we obtain the desired result. □

Remark 3.3. *Let*

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

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

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

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

Theorem 3.12. *For* $y \sim \chi^2(n)$, *we have* $\mathbb{E}[y] = n$ *and* $\mathrm{Var}[y] = 2n$.

Proof. We can write

$$y = \sum_{i=1}^n x_i^2,$$

where $x_1, \ldots, x_n$ are independent standard normal variables. Then, we have

$$\mathbb{E}[y] = \mathbb{E}\left[\sum_{i=1}^n x_i^2 \right] = \sum_{i=1}^n \mathbb{E}\left[x_i^2\right] = \sum_{i=1}^n \mathrm{Var}\left[x_i^2\right] = n$$

and

$$\mathrm{Var}[y] = \mathrm{Var}\left[\sum_{i=1}^n x_i^2 \right] = \sum_{i=1}^n \mathrm{Var}\left[x_i^2\right] = \sum_{i=1}^n \mathbb{E}\left[x_i^4 - \left(\mathbb{E}[x_i^2]\right)^2 \right] = \sum_{i=1}^n \mathbb{E}\left[3 - 1\right] = 2n.$$

We use the fact $\mathbb{E}[x_i^4] = 3$ because of $\phi(t) = \exp\left(-\frac{1}{2}t^2\right)$ and

$$\mathbb{E}[x_i^4] = \frac{1}{\mathrm{i}^4} \frac{\mathrm{d}^4 \phi(t)}{\mathrm{d}t^4}\bigg|_{t=0} = (t^4 - 6t^2 + 3)\exp\left(-\frac{1}{2}t^2\right)\bigg|_{t=0} = 3.$$

$\square$

Theorem 3.13. *The density of* $y \sim \chi^2(n)$ *is*

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

where

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

Proof. We first provide the following results:

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

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

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

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

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

We define the positive random variable $\hat{x}$ whose density function is

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

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

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

3. For beta function

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

we have

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

Consider that

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

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

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

and $\det(\mathbf{J}) = -u$. Since $u = x+y$ and $v = x/(x+y)$, we have that the limits of integration for u are 0 to ∞ and the limits of integration for v are 0 to 1. Thus

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

4. If

$$F(z) = \int_{a(z)}^{b(z)} f(y,z)\,\mathrm{d}y,$$

then

$$F'(z) = \int_{a(z)}^{b(z)} \frac{\partial f(y,z)}{\partial z}\,\mathrm{d}x + f(b(z),z)b'(z) - f(a(z),z)a'(z).$$

We prove the density of Chi-square distribution by induction. For $n = 1$ and $y > 0$, we have

$$f(y;1) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{1}{2}y\right) = \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} y^{\frac{1}{2}-1} \exp\left(-\frac{y}{2}\right).$$

Suppose the statement holds for $n-1$, that is

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

We consider $y_n = y_{n-1} + x_n^2$ such that $y_{n-1} \sim \chi^2(n-1)$ and $x_n \sim \mathcal{N}(0,1)$ are independent. Let F_1 be the corresponding cdf of $f(y;1)$. Then the cfd of y_n is

$$\begin{aligned} &\Pr(y_n \leq z) \\ =&\int_0^z \int_0^{z-y} f_{n-1}(y) f_1(x)\,\mathrm{d}x\,\mathrm{d}y \\ =&\int_0^z (F_1(z-y) - F_1(0)) f_{n-1}(y)\,\mathrm{d}x\,\mathrm{d}y \\ =&\int_0^z F_1(z-y) f_{n-1}(y)\,\mathrm{d}y \end{aligned}$$

and the pdf of y_n is (let $y = tz$)

$$\begin{aligned} &\int_0^z \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}(z-y)^{\frac{1}{2}-1} \exp\left(-\frac{z-y}{2}\right) \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right)\,\mathrm{d}y \\ =&\frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \int_0^z (z-y)^{\frac{1}{2}-1} y^{\frac{n-1}{2}-1} \exp\left(-\frac{z}{2}\right)\,\mathrm{d}y \\ =&\frac{\exp\left(-\frac{z}{2}\right) z^{\frac{n-1}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \int_0^1 (1-t)^{\frac{1}{2}-1} t^{\frac{n-1}{2}-1}\,\mathrm{d}t \\ =&\frac{\exp\left(-\frac{z}{2}\right) z^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} B\left(\frac{n-1}{2}, \frac{1}{2}\right) \\ =&\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1} \exp\left(-\frac{z}{2}\right). \end{aligned}$$

□

Theorem 3.14. *If the n-component vector $\mathbf{y}$ is distributed according to $\mathcal{N}(\boldsymbol{\nu}, \mathbf{T})$ with $\mathbf{T} \succ \mathbf{0}$, then*

$$\mathbf{y}^\top \mathbf{T}^{-1} \mathbf{y} \sim \chi_n^2\left(\boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu}\right).$$

If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is the central χ^2-distribution.

Proof. Let $\mathbf{C}$ be a non-singular matrix such that $\mathbf{C}\mathbf{T}\mathbf{C}^\top = \mathbf{I}$. Define $\mathbf{z} = \mathbf{C}\mathbf{y}$, then $\mathbf{z}$ is normally distributed with mean

$$\mathbf{C}\mathbb{E}[\mathbf{y}] = \mathbf{C}\boldsymbol{\nu} \triangleq \boldsymbol{\lambda}$$

and covariance matrix

$$\mathbb{E}\left[(\mathbf{z}-\boldsymbol{\lambda})(\mathbf{z}-\boldsymbol{\lambda})^\top\right] = \mathbf{C}\mathbb{E}\left[(\mathbf{y}-\boldsymbol{\nu})(\mathbf{y}-\boldsymbol{\nu})^\top\right]\mathbf{C}^\top = \mathbf{C}\mathbf{T}\mathbf{C}^\top = \mathbf{I}.$$

Then we have

$$\mathbf{y}^\top\mathbf{T}^{-1}\mathbf{y} = \mathbf{z}^\top\mathbf{C}^{-\top}\mathbf{T}^{-1}\mathbf{C}^{-1}\mathbf{z} = \mathbf{z}^\top\left(\mathbf{C}\mathbf{T}\mathbf{C}^\top\right)^{-1}\mathbf{z} = \mathbf{z}^\top\mathbf{z},$$

which is the sum of squares of the components of $\mathbf{z}$. Similarly, we have $\boldsymbol{\nu}^\top\mathbf{T}^{-1}\boldsymbol{\nu} = \boldsymbol{\lambda}^\top\boldsymbol{\lambda}$. Thus, the random variable $\mathbf{y}^\top\mathbf{T}^{-1}\mathbf{y}$ is distributed as $\sum_{i=1}^n z_i^2$, where $z_1,\ldots,z_n$ are independently normally distributed with means $\lambda_1,\ldots,\lambda_n$ respectively, and variances 1. By definition this is the noncentral χ^2-distribution with noncentrality parameter $\sum_{i=1}^n \lambda_i^2 = \boldsymbol{\nu}^\top\mathbf{T}^{-1}\boldsymbol{\nu}$. $\square$

Theorem 3.15. *The probability density function (pdf) for the noncentral F-distribution is*

$$f(v;p,\tau^2) = \begin{cases} \dfrac{\exp\left(-\frac{1}{2}(\tau^2+v)\right)v^{\frac{p}{2}-1}}{2^{\frac{p}{2}}\sqrt{\pi}} \displaystyle\sum_{\beta=0}^{\infty} \frac{\tau^{2\beta}v^{\beta}\Gamma\left(\beta+\frac{1}{2}\right)}{(2\beta)!\,\Gamma\left(\frac{p}{2}+\beta\right)} & v>0, \\ 0, & \textit{otherwise.} \end{cases}$$

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

Proof. Let $\mathbf{Q}$ be $p\times p$ orthogonal matrix with elements of the first row being

$$q_{i1} = \frac{\lambda_i}{\sqrt{(\boldsymbol{\lambda})^\top\boldsymbol{\lambda}}}$$

for $i=1,\ldots,p$. Then $\mathbf{z} = \mathbf{Q}\mathbf{y}$ is distributed according to $\mathcal{N}(\boldsymbol{\tau},\mathbf{I})$, where

$$\boldsymbol{\tau} = \begin{bmatrix}\tau\\0\\\vdots\\0\end{bmatrix},$$

where $\tau = \boldsymbol{\lambda}^\top\boldsymbol{\lambda}$. Let $\mathbf{v} = \mathbf{y}^\top\mathbf{y} = \mathbf{z}^\top\mathbf{z} = \sum_{i=1}^p z_i^2$. Then $w = \sum_{i=2}^p z_i^2$ has a χ^2-distribution with $p-1$ degrees of freedom, and z_1 and w have as joint density

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(z_1-\tau)^2\right)\frac{1}{2^{\frac{p-1}{2}}\Gamma\left(\frac{p-1}{2}\right)}w^{\frac{p-1}{2}-1}\exp\left(-\frac{w}{2}\right)\\
=&C\exp\left(-\frac{1}{2}\left(\tau^2+z_1^2+w\right)\right)w^{\frac{p-3}{2}}\exp\left(\tau z\right)\\
=&C\exp\left(-\frac{1}{2}\left(\tau^2+z_1^2+w\right)\right)w^{\frac{p-3}{2}}\sum_{\alpha=0}^{\infty}\frac{\tau^\alpha z_1^\alpha}{\alpha!}
\end{aligned}$$

where $C^{-1} = 2^{\frac{p}{2}}\sqrt{\pi}\Gamma\left(\frac{p-1}{2}\right)$. The joint density of $v = w + z_1^2$ and z_1 is obtained by substituting $w = v - z_1^2$ (the Jacobian being 1):

$$C\exp\left(-\frac{1}{2}\left(\tau^2+v\right)\right)(v-z_1^2)^{\frac{p-3}{2}}\sum_{\alpha=0}^{\infty}\frac{\tau^\alpha z_1^\alpha}{\alpha!}.$$

The joint density of v and $u = z_1/\sqrt{v}$ is ($\mathrm{d}z_1 = \sqrt{v}\mathrm{d}u$)

$$C \exp\left(-\frac{1}{2}\left(\tau^2 + v\right)\right) v^{\frac{p-2}{2}} (1-u^2)^{\frac{p-3}{2}} \sum_{\alpha=0}^{\infty} \frac{\tau^\alpha v^{\frac{\alpha}{2}} u^\alpha}{\alpha!}.$$

The admissible range of z given v is $-\sqrt{v}$ to $\sqrt{v}$, and the admissible range of u is -1 to 1. When we integrate above joint density with respect to u term by term, the terms for a odd integrate to 0, since such a term is an odd function of u. In the other integrations we substitute $u = \sqrt{s}$ ($\mathrm{d}u = \frac{\sqrt{s}}{2}\mathrm{d}s$) to obtain

$$\begin{aligned}
&\int_{-1}^{1} (1-u^2)^{\frac{p-3}{2}} u^{2\beta}\,\mathrm{d}u \\
=&2\int_{0}^{1} (1-u^2)^{\frac{p-3}{2}} u^{2\beta}\,\mathrm{d}u \\
=&\int_{0}^{1} (1-s)^{\frac{p-3}{2}} s^{\beta-\frac{1}{2}}\,\mathrm{d}s \\
=&B\left(\frac{p-1}{2}, \beta+\frac{1}{2}\right) \\
=&\frac{\Gamma(\frac{p-1}{2})\Gamma(\beta+\frac{1}{2})}{\Gamma(\frac{p}{2}+\beta)}
\end{aligned}$$

by the usual properties of the beta and gamma functions. Thus the density of v is

$$\frac{1}{2^{\frac{p}{2}}\sqrt{\pi}} \exp\left(-\frac{1}{2}(\tau^2+v)\right) v^{\frac{p}{2}-1} \sum_{\beta=0}^{\infty} \frac{\tau^{2\beta} v^\beta \Gamma\left(\beta+\frac{1}{2}\right)}{(2\beta)!\,\Gamma\left(\frac{p}{2}+\beta\right)}$$

for $v > 0$. $\square$

4 T^2-Statistic

Theorem 4.1. *Define the likelihood ratio criterion as*

$$\lambda = \frac{\max_{\boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})},$$

where

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

then we have

$$\lambda^{\frac{2}{N}} = \frac{1}{1 + T^2/(N-1)},$$

where $T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$.

Proof. The maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

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

If we restrict $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, the likelihood function is maximized at

$$\hat{\boldsymbol{\Sigma}}_\omega = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)^\top.$$

Furthermore, we have

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}} \left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}pN\right)$$

because of

$$\begin{aligned}
&\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}_\Omega^{-1}(\mathbf{x}_\alpha - \bar{\boldsymbol{\mu}}) \\
=&\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_\Omega^{-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\boldsymbol{\mu}})(\mathbf{x}_\alpha - \bar{\boldsymbol{\mu}})^\top\right) \\
=&\mathrm{tr}\left(n\mathbf{I}_p\right) = np.
\end{aligned}$$

Similarly, we also have

$$\max_{\boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}} \left(\det(\boldsymbol{\Sigma}_\omega)\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}pN\right).$$

Thus the likelihood ratio criterion is

$$\begin{aligned}
\lambda =&\frac{(2\pi)^{-\frac{pN}{2}} \left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}pN\right)}{(2\pi)^{-\frac{pN}{2}} \left(\det(\boldsymbol{\Sigma}_\omega)\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}pN\right)} = \frac{\left(\det(\boldsymbol{\Sigma}_\omega)\right)^{\frac{N}{2}}}{\left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{\frac{N}{2}}} \\
=&\frac{\left(\det\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\right)\right)^{\frac{N}{2}}}{\left(\det\left(\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)^\top\right)\right)^{\frac{N}{2}}} = \frac{\left(\det\left(\mathbf{A}\right)\right)^{\frac{N}{2}}}{\left(\det\left(\mathbf{A} + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\right)\right)^{\frac{N}{2}}}
\end{aligned}$$

where $\mathbf{A} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = (N-1)\mathbf{S}$. Hence, we obtain

$$\begin{aligned}
\lambda^{\frac{2}{N}} =&\frac{\det\left(\mathbf{A}\right)}{\det\left(\mathbf{A} + \left(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\right)\left(\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\right)\right)} \\
=&\frac{1}{1 + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\mathbf{A}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)} \\
=&\frac{1}{1 + T^2/(N-1)}
\end{aligned}$$

where $T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = (N-1)N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top\mathbf{A}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ and we use the property of Schur complement to obtain

$$\det\left(\begin{bmatrix}\mathbf{A} & \mathbf{u} \\ -\mathbf{u}^\top & 1\end{bmatrix}\right) = \det\left(\mathbf{A} + \mathbf{u}\mathbf{u}^\top\right) = \det\left(\begin{bmatrix}1 & -\mathbf{u}^\top \\ \mathbf{u} & \mathbf{A}\end{bmatrix}\right) = \det(\mathbf{A})\left(1 + \mathbf{u}\mathbf{A}^{-1}\mathbf{u}^\top\right)$$

with $\mathbf{u} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$. Recall that The decomposition

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

means we have $\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$. □

Lemma 4.1. *For any $p \times p$ non-singular matrices $\mathbf{C}$ and $\mathbf{H}$ and any vector $\mathbf{k}$, we have*

$$\mathbf{k}^\top \mathbf{H}^{-1}\mathbf{k} = (\mathbf{C}\mathbf{k})^\top (\mathbf{C}\mathbf{H}\mathbf{C}^\top)^{-1}(\mathbf{C}\mathbf{k}).$$

Proof. We have $(\mathbf{C}\mathbf{k})^\top (\mathbf{C}\mathbf{H}\mathbf{C}^\top)^{-1}(\mathbf{C}\mathbf{k}) = \mathbf{k}^\top \mathbf{C}^\top (\mathbf{C}^\top)^{-1}(\mathbf{H})^{-1}\mathbf{C}^{-1}(\mathbf{C}\mathbf{k}) = \mathbf{k}^\top \mathbf{H}^{-1}\mathbf{k}$. □

Remark 4.1. *This lemma means*

$$T^{*2} = N(\bar{\mathbf{x}}^* - \mathbf{0})^\top (\mathbf{S}^*)^{-1}(\bar{\mathbf{x}}^* - \mathbf{0}) = N(\mathbf{C}\bar{\mathbf{x}} - \mathbf{0})^\top (\mathbf{C}\mathbf{S}\mathbf{C})^{-1}(\mathbf{C}\bar{\mathbf{x}}^* - \mathbf{0}) = N(\bar{\mathbf{x}} - \mathbf{0})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}}^* - \mathbf{0}) = T^2.$$

Theorem 4.2. *Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are independent with $\mathbf{y}_\alpha$ distributed according to $\mathcal{N}(\mathbf{\Gamma}\mathbf{w}_\alpha, \mathbf{\Phi})$, where $\mathbf{w}_\alpha$ is an r-component vector. Let $\mathbf{H} = \sum_{\alpha=1}^m \mathbf{w}_\alpha \mathbf{w}_\alpha^\top$ assumed non-singular, $\mathbf{G} = \sum_{\alpha=1}^m \mathbf{y}_\alpha \mathbf{w}_\alpha^\top \mathbf{H}^{-1}$ and*

$$\mathbf{C} = \sum_{\alpha=1}^{m} (\mathbf{y}_\alpha - \mathbf{G}\mathbf{w}_\alpha)(\mathbf{y}_\alpha - \mathbf{G}\mathbf{w}_\alpha)^\top = \sum_{\alpha=1}^{m} \mathbf{y}_\alpha \mathbf{y}_\alpha^\top - \mathbf{G}\mathbf{H}\mathbf{G}^\top.$$

Then $\mathbf{C}$ is distributed as

$$\sum_{\alpha=1}^{m-r} \mathbf{u}_\alpha \mathbf{u}_\alpha^\top$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_{m-r}$ are independently distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{\Phi})$ independently of $\mathbf{G}$.

Proof. Theorem 4.3.3 of "Theodore W. Anderson. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons Inc; 3rd Edition." □

Theorem 4.3. *Let $T^2 = \mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y}$, where $\mathbf{y}$ is distributed according to $\mathcal{N}_p(\boldsymbol{\nu}, \mathbf{\Sigma})$ and $n\mathbf{S}$ is independently distributed as $\sum_{\alpha=1}^n \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$ with $\mathbf{z}_1, \ldots, \mathbf{z}_n$ independent, each with distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Then the random variable*

$$\frac{T^2}{n} \cdot \frac{n-p+1}{p}$$

is distributed as a noncentral F-distribution with p and $n-p+1$ degrees of freedom and noncentrality parameter $\boldsymbol{\nu}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is central F.

Proof. Let $\mathbf{D}$ be a non-singular matrix such that $\mathbf{D}\mathbf{\Sigma}\mathbf{D}^\top = \mathbf{I}$, and define

$$\mathbf{y}^* = \mathbf{D}\mathbf{y}, \quad \mathbf{S}^* = \mathbf{D}\mathbf{S}\mathbf{D}^\top, \quad \boldsymbol{\nu}^* = \mathbf{D}\boldsymbol{\nu}.$$

Lemma 4.1 means

$$T^2 = (\mathbf{y}^*)^\top (\mathbf{S}^*)^{-1}\mathbf{y}^*$$

where $\mathbf{y}^*$ is distributed according to $\mathcal{N}(\boldsymbol{\nu}^*, \mathbf{I})$ and

$$n\mathbf{S}^* = \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha^* (\mathbf{z}_\alpha^*)^\top = \sum_{\alpha=1}^{N-1} \mathbf{D}\mathbf{z}_\alpha (\mathbf{D}\mathbf{z}_\alpha)^\top$$

with $\mathbf{z}_\alpha^* = \mathbf{D}\mathbf{z}_\alpha$ independent, each with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We also have

$$\boldsymbol{\nu}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\nu} = (\mathbf{D}\boldsymbol{\nu})^\top (\mathbf{D}\mathbf{\Sigma}\mathbf{D}^\top)^{-1}(\mathbf{D}\boldsymbol{\nu}^*) = (\boldsymbol{\nu}^*)^\top \boldsymbol{\nu}^*.$$

Let the first row of a $p \times p$ orthogonal matrix $\mathbf{Q}$ be defined by

$$q_{i1} = \frac{y_i^*}{\sqrt{(\mathbf{y}^*)^\top \mathbf{y}^*}}$$

for $i = 1, \ldots, p$. Since $\mathbf{Q}$ depends on $\mathbf{y}^*$, it is a random matrix. Now let

$$\mathbf{u} = \mathbf{Q}\mathbf{y}^* \qquad \text{and} \qquad \mathbf{B} = \mathbf{Q}(n\mathbf{S}^*)\mathbf{Q}^\top,$$

where $n = N - 1$. The definition of $\mathbf{Q}$ means

$$u_1 = \sum_{i=1}^{p} q_{1i} y_i^* = \frac{\sum_{i=1}^{p} (y_i^*)^2}{\sqrt{(\mathbf{y}^*)^\top \mathbf{y}^*}} = \sqrt{(\mathbf{y}^*)^\top \mathbf{y}^*}$$

and

$$u_j = \sum_{i=1}^{p} q_{ji} y_i^* = \sqrt{(\mathbf{y}^*)^\top \mathbf{y}^*} \sum_{i=1}^{p} q_{ji} q_{1i} = 0$$

for $j = 2, \ldots, p$. Then

$$\begin{aligned}
\frac{T^2}{n} =& (\mathbf{y}^*)^\top (\mathbf{S}^*)^{-1} \mathbf{y}^* = (\mathbf{Q}\mathbf{u})^\top (\mathbf{Q}^\top \mathbf{B} \mathbf{Q})^{-1} \mathbf{Q}^\top \mathbf{u} = \mathbf{u}^\top \mathbf{Q}^\top \mathbf{Q}^\top \mathbf{B}^{-1} \mathbf{Q} \mathbf{Q}^\top \mathbf{u} = \mathbf{u}^\top \mathbf{B}^{-1} \mathbf{u} \\
=& \begin{bmatrix} u_1 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} b^{11} & b^{12} & \ldots & b^{1p} \\ b^{21} & b^{22} & \ldots & b^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b^{p1} & b^{p2} & \ldots & b^{pp} \end{bmatrix} \begin{bmatrix} u_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = u_1^2 b^{11}
\end{aligned}$$

where b^{ij} the entries of $\mathbf{B}^{-1}$. Using Schur Complement, we have

$$\frac{1}{b^{11}} = b_{11} - \mathbf{b}_{(1)}^\top \mathbf{B}_{22}^{-1} \mathbf{b}_{(1)} \triangleq b_{11.2,\ldots,p}$$

where

$$\mathbf{B} = \begin{bmatrix} b_{11} & \mathbf{b}_{(1)}^\top \\ \mathbf{b}_{(1)} & \mathbf{B}_{22} \end{bmatrix}$$

and

$$\frac{T^2}{n} = \frac{u_1^2}{b_{11.2,\ldots,p}} = \frac{(\mathbf{y}^*)^\top \mathbf{y}^*}{b_{11.2,\ldots,p}}.$$

The conditional distribution of $\mathbf{B}$ given $\mathbf{Q}$ is that of

$$\mathbf{B} = \sum_{\alpha=1}^{n} \mathbf{Q}\mathbf{z}_\alpha^* (\mathbf{Q}\mathbf{z}_\alpha^*)^\top = \sum_{\alpha=1}^{n} \mathbf{v}_\alpha^* (\mathbf{v}_\alpha^*)^\top,$$

where $\mathbf{v}_\alpha = \mathbf{Q}\mathbf{z}_\alpha^*$ are independent, each with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ since $\mathbf{Q}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top \mathbf{Q}^\top = \mathbf{I}$. By Theorem 4.2, the random variable $b_{11.2,\ldots,p}$ is conditionally distributed as

$$\sum_{\alpha=1}^{n-(p-1)} w_\alpha^2$$

where conditionally the w_α^2 are independent, each with the distribution $\mathcal{N}(0,1)$; that is, $b_{11.2,\ldots,p}$ is conditionally distributed as χ^2 with $n-(p-1)$ degrees of freedom. Since the conditional distribution of $b_{11.2,\ldots,p}$ does not depend on $\mathbf{Q}$, it is unconditionally distributed as χ^2. The quantity $\mathbf{y}^* \mathbf{y}*$ has a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter $(\boldsymbol{\nu}^*)^\top \boldsymbol{\nu}^* = \boldsymbol{\nu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\nu}^\top$ Then T is distributed as the ratio of a noncentral χ^2 and an independent χ^2. $\square$

Theorem 4.4. *Let u be distributed according to the χ^2-distribution with a degrees of freedom and w be distributed according to the χ^2-distribution with b degrees of freedom. The density of $v = u/(u+w)$, when u and w are independent is*

$$\frac{1}{B\left(\frac{a}{2}, \frac{b}{2}\right)} v^{\frac{a}{2}-1}(1-v)^{\frac{b}{2}-1}, \tag{5}$$

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

Proof. Let

$$v = \frac{u}{u+w} \quad \text{and} \quad z = u + w.$$

Then $u = vz$, $w = (1-v)z$ and

$$\det(\mathbf{J}(v,z)) = \det\left(\begin{bmatrix} \frac{\partial u}{\partial v} & \frac{\partial u}{\partial z} \\ \frac{\partial w}{\partial v} & \frac{\partial w}{\partial z} \end{bmatrix}\right) = \det\left(\begin{bmatrix} z & v \\ -z & 1-v \end{bmatrix}\right) = z.$$

Since v and w are independent, the joint density of v and w is

$$f_{u,v}(u,w) = \frac{1}{2^{\frac{a}{2}}\Gamma\left(\frac{a}{2}\right)} u^{\frac{a}{2}-1} \exp\left(-\frac{u}{2}\right) \cdot \frac{1}{2^{\frac{b}{2}}\Gamma\left(\frac{b}{2}\right)} w^{\frac{b}{2}-1} \exp\left(-\frac{w}{2}\right)$$

and the joint density of v and z is

$$\begin{aligned} f_{v,z}(v,z) =& f_{u,v}(vz,(1-v)z)\det(\mathbf{J}(v,z)) \\ =& \frac{1}{2^{\frac{a}{2}}\Gamma\left(\frac{a}{2}\right)} (vz)^{\frac{a}{2}-1} \exp\left(-\frac{vz}{2}\right) \cdot \frac{1}{2^{\frac{b}{2}}\Gamma\left(\frac{b}{2}\right)} ((1-v)z)^{\frac{b}{2}-1} \exp\left(-\frac{(1-v)z}{2}\right) \cdot z \\ =& \frac{1}{2^{\frac{a+b}{2}}\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} v^{\frac{a}{2}-1} \cdot (1-v)^{\frac{b}{2}-1} z^{\frac{a+b}{2}-1} \exp\left(-\frac{z}{2}\right). \end{aligned}$$

Consider that the density of χ^2-distribution with $a+b$ degrees of freedom, we have

$$\int_{-\infty}^{\infty} \frac{1}{2^{\frac{a+b}{2}}\Gamma\left(\frac{a+b}{2}\right)} z^{\frac{a+b}{2}-1} \exp\left(-\frac{z}{2}\right)\,\mathrm{d}z = 1.$$

Hence,

$$\begin{aligned} f_z(z) =& \int_{-\infty}^{\infty} f_{v,z}(v,z)\,\mathrm{d}z \\ =& \frac{1}{2^{\frac{a+b}{2}}\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} v^{\frac{a}{2}-1}(1-v)^{\frac{b}{2}-1} \int_{-\infty}^{\infty} z^{\frac{a+b}{2}-1} \exp\left(-\frac{z}{2}\right)\,\mathrm{d}z \\ =& \frac{2^{\frac{a+b}{2}}\Gamma\left(\frac{a+b}{2}\right)}{2^{\frac{a+b}{2}}\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} v^{\frac{a}{2}-1}(1-v)^{\frac{b}{2}-1} \\ =& \frac{1}{B\left(\frac{a}{2}+\frac{b}{2}\right)} v^{\frac{a}{2}-1}(1-v)^{\frac{b}{2}-1}. \end{aligned}$$

□

Theorem 4.5. *Let $x_1, x_2, \ldots$ be a sequence of independently identically distributed random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let*

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

and

$$T_N^2 = N(\bar{\mathbf{x}}_N - \boldsymbol{\mu}_0)^\top \mathbf{S}_N^{-1}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}_0).$$

Then the limiting distribution of T_N^2 as $N \to \infty$ is the χ^2-distribution with p degrees of freedom if $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.

Proof. By the central limit theorem, the limiting distribution of $\sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu})$ is $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The sample covariance matrix converges sarcastically to $\boldsymbol{\Sigma}$. Then the limiting distribution of T^2 is the distribution of

$$\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

where $\mathbf{y}$ has the distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The theorem follows from Theorem 3.14. □

Lemma 4.2. *If $\mathbf{v}$ is a vector of p components and if $\mathbf{B}$ is a non-singular $p \times p$ matrix, then $\mathbf{v}^\top \mathbf{B}^{-1}\mathbf{v}$ is the nonzero root of*

$$\det(\mathbf{v}\mathbf{v}^\top - \lambda \mathbf{B}) = 0.$$

Proof. The non-zero root λ_1 of $\det(\mathbf{v}\mathbf{v}^\top - \lambda\mathbf{B}) = 0$ associate with vector $\boldsymbol{\beta} \neq \mathbf{0}$ satisfying

$$(\mathbf{v}\mathbf{v}^\top - \lambda_1 \mathbf{B})\boldsymbol{\beta} = \mathbf{0} \Longrightarrow \mathbf{v}\mathbf{v}^\top \boldsymbol{\beta} = \lambda_1 \mathbf{B}\boldsymbol{\beta} \Longrightarrow \left(\mathbf{v}^\top \mathbf{B}^{-1}\mathbf{v}\right)\mathbf{v}^\top \boldsymbol{\beta} = \lambda_1 \mathbf{v}^\top \boldsymbol{\beta}.$$

We can obtain that $\mathbf{v}^\top \boldsymbol{\beta} \neq 0$, otherwise $(\mathbf{v}\mathbf{v}^\top - \lambda_1\mathbf{B})\boldsymbol{\beta} = \mathbf{0}$ means $\mathbf{B}\boldsymbol{\beta} = \mathbf{0}$ which is impossible since $\mathbf{B}$ is non-singular. Hence $\lambda_1 = \mathbf{v}^\top \mathbf{B}^{-1}\mathbf{v}$.

Remark 4.2. *Using this lemma with $\mathbf{v} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ and $\mathbf{B} = \mathbf{A}$, we can prove $T^2/(N-1)$ is the non-zero root of* $\det\left(N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top - \lambda\mathbf{A}\right) = 0.$

□

Lemma 4.3. *For any positive definite matrix $\mathbf{S} \in \mathbb{R}^{p \times p}$ and $\mathbf{y}, \boldsymbol{\gamma} \in \mathbb{R}^p$, we have*

$$(\boldsymbol{\gamma}^\top \mathbf{y})^2 \leq (\boldsymbol{\gamma}^\top \mathbf{S}\boldsymbol{\gamma})(\mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y}).$$

Proof. For $\boldsymbol{\gamma} = \mathbf{0}$, the result is trivial. Otherwise, let

$$b = \frac{\boldsymbol{\gamma}^\top \mathbf{y}}{\boldsymbol{\gamma}^\top \mathbf{S}\boldsymbol{\gamma}}.$$

Then we have

$$\begin{aligned}
0 \leq& (\mathbf{y} - b\mathbf{S}\boldsymbol{\gamma})^\top \mathbf{S}^{-1}(\mathbf{y} - b\mathbf{S}\boldsymbol{\gamma}) \\
=& \mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y} - b\mathbf{y}^\top \mathbf{S}^{-1}\mathbf{S}\boldsymbol{\gamma} - b\boldsymbol{\gamma}^\top \mathbf{S}\mathbf{S}^{-1}\mathbf{y} - b^2\boldsymbol{\gamma}^\top \mathbf{S}\mathbf{S}^{-1}\mathbf{S}\boldsymbol{\gamma} \\
=& \mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y} - 2b\mathbf{y}^\top \boldsymbol{\gamma} + b^2\boldsymbol{\gamma}^\top \mathbf{S}\boldsymbol{\gamma} \\
=& \mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y} - \frac{(\boldsymbol{\gamma}^\top \mathbf{y})^2}{\boldsymbol{\gamma}^\top \mathbf{S}\boldsymbol{\gamma}},
\end{aligned}$$

which implies the desired result. □

Theorem 4.6. *Let $\{\mathbf{x}_\alpha^{(i)}\}$ for $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$ be samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \dots, q$, respectively and suppose*

$$\sum_{i=1}^q \beta_i \boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}.$$

where $\beta_1, \dots, \beta_q$ are given scalars and $\boldsymbol{\mu}$ is a given vector. Define the criterion

$$T^2 = c\left(\sum_{i=1}^q \beta_i \bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right)\mathbf{S}^{-1}\left(\sum_{i=1}^q \beta_i \bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right)^\top$$

where

$$\bar{\mathbf{x}}^{(i)} = \frac{1}{N_i}\sum_{\alpha=1}^{N_i}\mathbf{x}_\alpha^{(i)}, \qquad \frac{1}{c} = \sum_{i=1}^{q}\frac{\beta_i^2}{N_i}$$

and

$$\left(\sum_{i=1}^{q} N_i - q\right) S = \sum_{i=1}^{q}\sum_{\alpha=1}^{N_i}\left(\mathbf{x}_\alpha^{(i)} - \bar{\mathbf{x}}^{(i)}\right)\left(\mathbf{x}_\alpha^{(i)} - \bar{\mathbf{x}}^{(i)}\right)^\top.$$

Then this T^2 has the T^2-distribution with $\sum_{i=1}^{q} N_i - q$ degrees of freedom.

Proof. Since $\mathbf{x}_\alpha^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, we have

$$\bar{\mathbf{x}}^{(i)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \frac{1}{N_i}\boldsymbol{\Sigma}\right) \quad \Longrightarrow \quad \beta_i\left(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i\right) \sim \mathcal{N}\left(0, \frac{\beta_i^2}{N_i}\boldsymbol{\Sigma}\right).$$

and

$$\sum_{i=1}^{q}\beta_i\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu} = \sum_{i=1}^{q}\beta_i\left(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}\right) \sim \mathcal{N}\left(\mathbf{0}, \sum_{i=1}^{q}\frac{\beta_i^2}{N_i}\boldsymbol{\Sigma}\right) \Longrightarrow \sqrt{c}\left(\sum_{i=1}^{q}\beta_i\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right).$$

On the other hand, we can write

$$\sum_{i=1}^{q}\sum_{\alpha=1}^{N_i}\left(\mathbf{x}_\alpha^{(i)} - \bar{\mathbf{x}}^{(i)}\right)\left(\mathbf{x}_\alpha^{(i)} - \bar{\mathbf{x}}^{(i)}\right)^\top = \sum_{i=1}^{q}\sum_{\alpha=1}^{N_i-1}\mathbf{z}_\alpha^{(i)}(\mathbf{z}_\alpha^{(i)})^\top$$

where $\mathbf{z}_\alpha^{(i)}$ are independent and $\mathbf{z}_\alpha^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Hence,

$$T^2 = \sqrt{c}\left(\sum_{i=1}^{q}\beta_i\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right)\mathbf{S}^{-1}\left(\sqrt{c}\left(\sum_{i=1}^{q}\beta_i\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right)\right)^\top$$

has the T^2-distribution with $\sum_{i=1}^{q} N_i - q$ degrees of freedom. $\square$

Lemma 4.4. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be independent samples from $\mathcal{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ for $i = 1, \ldots, m$. Define*

$$\mathbf{z}_1 = \sum_{\alpha=1}^{N} a_\alpha\mathbf{x}_\alpha \quad \textit{and} \quad \mathbf{z}_2 = \sum_{\alpha=1}^{N} b_\alpha\mathbf{x}_\alpha,$$

then

$$\operatorname{Cov}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{\alpha=1}^{N} a_\alpha b_\alpha\boldsymbol{\Sigma}_\alpha.$$

Proof. The definitions mean

$$\mathbf{z}_1 = \begin{bmatrix} a_1\mathbf{I} & a_2\mathbf{I} & \ldots & a_N\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \ldots\\ \mathbf{x}_N\end{bmatrix} \quad \text{and} \quad \mathbf{z}_2 = \begin{bmatrix} b_1\mathbf{I} & b_2\mathbf{I} & \ldots & b_N\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \ldots\\ \mathbf{x}_N\end{bmatrix},$$

then

$$\operatorname{Cov}(\mathbf{z}_1, \mathbf{z}_2) = \begin{bmatrix} a_1\mathbf{I} & a_2\mathbf{I} & \ldots & a_N\mathbf{I}\end{bmatrix}\operatorname{Cov}\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \ldots\\ \mathbf{x}_N\end{bmatrix}, \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \ldots\\ \mathbf{x}_N\end{bmatrix}\right)\begin{bmatrix} b_1\mathbf{I}\\ b_2\mathbf{I}\\ \vdots\\ b_N\mathbf{I}\end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} a_1\mathbf{I} & a_2\mathbf{I} & \dots & a_N\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_N \end{bmatrix} \begin{bmatrix} b_1\mathbf{I} \\ b_2\mathbf{I} \\ \vdots \\ b_N\mathbf{I} \end{bmatrix} \\
&= \sum_{\alpha=1}^{N} a_\alpha b_\alpha \boldsymbol{\Sigma}_\alpha.
\end{aligned}
$$

□

Lemma 4.5. *Let* $\{\mathbf{x}_\alpha^{(i)}\}$ *for* $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$ *be independent samples from* $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_i)$ *for* $i = 1, 2$, *respectively. We suppose* $N_1 < N_2$ *and define*

$$
\mathbf{y}_\alpha = \mathbf{x}_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)} - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}\mathbf{x}_\gamma^{(2)},
$$

for $\alpha = 1, \dots, N_1$. *Then we have*

$$
\bar{\mathbf{y}} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1}\mathbf{y}_\alpha = \bar{\mathbf{x}}_\alpha^{(1)} - \bar{\mathbf{x}}_\alpha^{(2)}
$$

and

$$
\mathrm{Cov}(\mathbf{y}_\alpha, \mathbf{y}_{\alpha'}) = \begin{cases} \boldsymbol{\Sigma}_1 + \frac{N_1}{N_2}\boldsymbol{\Sigma}_2, & \alpha = \alpha', \\ \mathbf{0}, & \textit{otherwise.} \end{cases}
$$

Proof. We have

$$
\begin{aligned}
\bar{\mathbf{y}} =& \frac{1}{N_1}\sum_{\alpha=1}^{N_1}\mathbf{y}_\alpha \\
=& \frac{1}{N_1}\sum_{\alpha=1}^{N_1}\left(\mathbf{x}_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)} - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}\mathbf{x}_\gamma^{(2)}\right) \\
=& \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} + \frac{1}{N_1}\sum_{\alpha=1}^{N_1}\left(\sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)}\right) \\
=& \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} + \frac{1}{N_1}\sum_{\alpha=1}^{N_1}\sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)} \\
=& \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}.
\end{aligned}
$$

For the covariance matrix, we only show the case of $\alpha = \alpha'$ and leave the other case as homework. The independence means the matrix $\mathrm{Cov}(\mathbf{y}_\alpha, \mathbf{y}_\alpha)$ has the form of

$$
\begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \times \end{bmatrix},
$$

which means we only needs to focus on the covariance matrix of

$$
\begin{aligned}
\mathbf{z}_\alpha =& -\sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)} - \frac{1}{N_1}\sum_{\gamma=1}^{N_2}\mathbf{x}_\gamma^{(2)} \\
=& \sum_{\gamma=1}^{\alpha-1}\left(\frac{1}{N_1N_2} - \frac{1}{N_2}\right)\mathbf{x}_\gamma^{(2)} + \left(\frac{1}{N_1N_2} - \frac{1}{N_2} - \sqrt{\frac{N_1}{N_2}}\right)\mathbf{x}_\alpha^{(2)}
\end{aligned}
$$

$$+\sum_{\gamma=\alpha+1}^{N_1}\left(\frac{1}{N_1N_2}-\frac{1}{N_2}\right)\mathbf{x}_\gamma^{(2)}+\sum_{\gamma=N_1+1}^{N_2}\left(-\frac{1}{N_2}\right)\mathbf{x}_\gamma^{(2)}$$

Lemma 4.4 means

$$\begin{aligned}\operatorname{Cov}(\mathbf{z}_\alpha,\mathbf{z}_\alpha)=&\Bigg((\alpha-1)\left(\frac{1}{N_1N_2}-\frac{1}{N_2}\right)^2+\left(\frac{1}{N_1N_2}-\frac{1}{N_2}-\sqrt{\frac{N_1}{N_2}}\right)^2\\&+(N-\alpha)\left(\frac{1}{N_1N_2}-\frac{1}{N_2}\right)^2+(N_2-N_1)\sum_{\gamma=N_1+1}^{N_2}\left(-\frac{1}{N_2}\right)^2\Bigg)\boldsymbol{\Sigma}_2=\frac{N_1}{N_2}\boldsymbol{\Sigma}_2,\end{aligned}$$

which means $\operatorname{Cov}(\mathbf{y}_\alpha,\mathbf{y}_\alpha)=\boldsymbol{\Sigma}_1+\frac{N_1}{N_2}\boldsymbol{\Sigma}_2$. □

5 Sample Correlation Coefficients

Lemma 5.1. *If $\mathbf{y}_1,\dots,\mathbf{y}_N$ are independently distributed, if*

$$\mathbf{y}_\alpha=\begin{bmatrix}\mathbf{y}_\alpha^{(1)}\\ \mathbf{y}_\alpha^{(2)}\end{bmatrix}$$

has the density $f(\mathbf{y}_\alpha)$ and if the conditional density of $\mathbf{y}_\alpha^{(2)}$ given $\mathbf{y}_\alpha^{(1)}$ is $f\big(\mathbf{y}_\alpha^{(2)}\mid\mathbf{y}_\alpha^{(1)}\big)$ for $\alpha=1,\dots,n$. Then in the conditional distribution of $\mathbf{y}_1^{(2)},\dots,\mathbf{y}_N^{(2)}$ given $\mathbf{y}_1^{(1)},\dots,\mathbf{y}_N^{(1)}$, the random vectors $\mathbf{y}_1^{(1)},\dots,\mathbf{y}_N^{(1)}$ are independent and the density of $\mathbf{y}_\alpha^{(2)}$ is $f\big(\mathbf{y}_\alpha^{(2)}\mid\mathbf{y}_\alpha^{(1)}\big)$.

Proof. The marginal density of $\mathbf{y}_1^{(1)},\dots,\mathbf{y}_N^{(1)}$ is

$$\prod_{\alpha=1}^N f_1\big(\mathbf{y}_\alpha^{(1)}\big)$$

where $f_1\big(\mathbf{y}_\alpha^{(1)}\big)$ is the marginal density of $\mathbf{y}_\alpha^{(1)}$, and the conditional density of $\mathbf{y}_1^{(2)},\dots,\mathbf{y}_N^{(2)}$ given $\mathbf{y}_1^{(1)},\dots,\mathbf{y}_N^{(1)}$ is

$$\frac{\prod_{\alpha=1}^N f\big(\mathbf{y}_\alpha\big)}{\prod_{\alpha=1}^N f_1\big(\mathbf{y}_\alpha^{(1)}\big)}=\prod_{\alpha=1}^N\frac{f\big(\mathbf{y}_\alpha^{(1)},\mathbf{y}_\alpha^{(2)}\big)}{f_1\big(\mathbf{y}_\alpha^{(1)}\big)}=\prod_{\alpha=1}^N f\big(\mathbf{y}_\alpha^{(2)}\mid\mathbf{y}_\alpha^{(1)}\big).$$

□

Theorem 5.1. *If the pairs $(z_{11},z_{21}),\dots,(z_{1n},z_{2n})$ are independent and each pair are distributed according to*

$$\begin{bmatrix}z_{1\alpha}\\ z_{2\alpha}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_1^2&\sigma_1\sigma_2\rho\\ \sigma_1\sigma_2\rho&\sigma_2^2\end{bmatrix}\right),\quad\text{where }\alpha=1,\dots,n,$$

then given $z_{11},z_{12},\dots,z_{1n}$, the conditional distributions of

$$b=\frac{\sum_{\alpha=1}^n z_{2\alpha}z_{1\alpha}}{\sum_{i=1}^n z_{1\alpha}^2}\quad\text{and}\quad\frac{u}{\sigma^2}=\sum_{\alpha=1}^n\frac{(z_{2\alpha}-bz_{1\alpha})^2}{\sigma^2}$$

are $\mathcal{N}\left(\beta,\sigma^2/c^2\right)$ and χ^2-distribution with $n-1$ degrees of freedom, respectively; and b and u are independent, where

$$\beta=\frac{\rho\sigma_2}{\sigma_1},\quad\sigma^2=\sigma_2^2(1-\rho^2)\quad\text{and}\quad c^2=\sum_{i=1}^n z_{1\alpha}^2.$$

Proof. The conditional distribution of $z_{2\alpha}$ given $z_{1\alpha}$ is $\mathcal{N}(\beta z_{1\alpha}, \sigma^2)$. Let $\mathbf{v}_i = [z_{i1}, \ldots, z_{in}]^\top$ for $i = 1, 2$. Lemma 5.1 means the density of $\mathbf{v}_2$ given $\mathbf{v}_1$ is $\mathcal{N}(\beta \mathbf{v}_1, \sigma^2 \mathbf{I})$ since $z_{21}, \ldots, z_{2n}$ are independent. We also have

$$\mathbf{v}_1^\top (\mathbf{v}_2 - b\mathbf{v}_1) = \mathbf{v}_1^\top \left(\mathbf{v}_2 - \frac{\mathbf{v}_1^\top \mathbf{v}_2}{\mathbf{v}_1^\top \mathbf{v}_1} \mathbf{v}_1 \right) = 0$$

and

$$u = (\mathbf{v}_2 - b\mathbf{v}_1)^\top (\mathbf{v}_2 - b\mathbf{v}_1) = \mathbf{v}_2^\top \mathbf{v}_2 - 2b\mathbf{v}_1^\top \mathbf{v}_2 + b^2 \mathbf{v}_2^\top = \mathbf{v}_2^\top \mathbf{v}_2 - b^2 \mathbf{v}_1^\top \mathbf{v}_1.$$

Apply Theorem 3.3 with $x_\alpha = z_{2\alpha}$ and $y_\alpha = \sum_{\gamma=1}^n c_{\alpha\gamma} z_{2\gamma}$ for $\alpha = 1, \ldots, n$, where the first row of orthogonal matrix $\mathbf{C}$ is $(1/c)\mathbf{v}_1^\top$. Then $y_1, \ldots, y_n$ are independently normally distributed with variance σ^2 and means

$$\mathbb{E}[y_1] = \sum_{\gamma=1}^n c_{1\gamma} \mathbb{E}[z_{2\gamma}] = \sum_{\gamma=1}^n c_{1\gamma} \beta z_{1\gamma} = \beta c,$$

and

$$\mathbb{E}[y_\alpha] = \sum_{\gamma=1}^n c_{\alpha\gamma} \mathbb{E}[z_{2\gamma}] = \sum_{\gamma=1}^n c_{\alpha\gamma} \beta z_{1\gamma} = 0.$$

Thus, we have

$$b = \frac{\sum_{\alpha=1}^n z_{2\alpha} z_{1\alpha}}{\sum_{i=1}^n z_{1\alpha}^2} = \frac{\sum_{\alpha=1}^n c z_{2\alpha} c_{1\alpha}}{c^2} = \frac{y_1}{c} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{c^2}\right).$$

and

$$u = \sum_{\alpha=1}^n z_{2\alpha}^2 - b^2 \sum_{\alpha=1}^n z_{1\alpha}^2 = \sum_{\alpha=1}^n y_\alpha^2 - y_1^2 = \sum_{\alpha=2}^n y_\alpha^2,$$

which is independent of b. Since we have $y_\alpha \sim \mathcal{N}(0, \sigma^2)$ for $\alpha = 2, \ldots, n$, the random variable u/σ^2 has a χ^2-distribution with $n-1$ degrees of freedom. ◻

Theorem 5.2. *If x and y are independently distributed, x having the distribution $\mathcal{N}(0,1)$ and y having the χ^2-distribution with m degrees of freedom, then $t = x/\sqrt{y/m}$ (has t-distribution with m degrees of freedom) has the density*

$$\frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\,\Gamma(\frac{m}{2})} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}.$$

Proof. The joint density of x and y is

$$f_{x,y}(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} \exp\left(-\frac{y}{2}\right).$$

The definition of t means $x = t\sqrt{y/m}$, then the joint density of t and y is

$$\begin{aligned}
f_{t,y}(t,y) =& \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2 y}{2m}\right) \cdot \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} y^{\frac{m}{2}-1} \exp\left(-\frac{y}{2}\right) \cdot \frac{\mathrm{d}t\sqrt{y/m}}{\mathrm{d}t} \\
=& \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2 y}{2m}\right) \cdot \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} y^{\frac{m}{2}-1} \exp\left(-\frac{y}{2}\right) \cdot \left(\frac{y}{m}\right)^{\frac{1}{2}} \\
=& \frac{1}{2^{\frac{m+1}{2}}\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right) y\right) \cdot y^{\frac{m-1}{2}}.
\end{aligned} \tag{6}$$

The density of t can be obtained by integrating out y. Consider the expression of gamma function

$$
\begin{aligned}
\Gamma(\alpha) &= \int_0^{+\infty} \tilde{t}^{\alpha-1} \exp(-\tilde{t})\,\mathrm{d}\tilde{t} \\
&= \int_0^{+\infty} \left(\frac{t^2}{2m} + \frac{1}{2}\right)^{\alpha-1} y^{\alpha-1} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right) y\right) \left(\frac{t^2}{2m} + \frac{1}{2}\right) \mathrm{d}y \\
&= \left(\frac{t^2}{2m} + \frac{1}{2}\right)^{\alpha} \int_0^{+\infty} y^{\alpha-1} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right) y\right) \mathrm{d}y
\end{aligned} \tag{7}
$$

where we use the substitution

$$
\tilde{t} = \left(\frac{t^2}{2m} + \frac{1}{2}\right) y.
$$

Connecting (6) and (7) with $\alpha = \frac{m+1}{2}$, we have

$$
\begin{aligned}
f_t(t) &= \int_0^{+\infty} f_{t,y}(t,y)\,\mathrm{d}y \\
&= \frac{1}{2^{\frac{m}{2}}\sqrt{m\pi}\,\Gamma\left(\frac{m+1}{2}\right)} \int_0^{+\infty} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right) y\right) \cdot y^{\frac{m-1}{2}}\,\mathrm{d}y \\
&= \frac{1}{2^{\frac{m}{2}}\sqrt{m\pi}\,\Gamma\left(\frac{m+1}{2}\right)} \left(\frac{t^2}{2m} + \frac{1}{2}\right)^{-\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right) \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \left(\frac{t^2}{m} + 1\right)^{-\frac{m+1}{2}}.
\end{aligned}
$$

□

Theorem 5.3. *Let us consider the likelihood ratio test of the hypothesis that $\rho = \rho_0$ based on a sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from the bivariate normal distribution*

$$
\mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right).
$$

The set Ω consists of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ such that

$$
\sigma_1 > 0, \quad \sigma_2 > 0 \quad \textit{and} -1 < \rho < 1
$$

and the set ω is the subset for which $\rho = \rho_0$. The likelihood ratio criterion is

$$
\frac{\sup_\omega L(\mathbf{x}, \boldsymbol{\theta})}{\sup_\Omega L(\mathbf{x}, \boldsymbol{\theta})} = \left(\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right)^{\frac{N}{2}},
$$

where

$$
r = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}}, \quad \mathbf{A} = \begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \quad \textit{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha.
$$

Proof. We have shown in the proof of Theorem 4.1 that the likelihood maximized in Ω is

$$
\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}} \left(\det(\boldsymbol{\Sigma}_\Omega)\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}pN\right)
$$

where

$$
\boldsymbol{\Sigma}_\Omega = \frac{1}{N}\mathbf{A} \quad \text{with} \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha, \quad \mathbf{A} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top = \begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} \quad \text{and} \quad p = 2.
$$

Then we have

$$\det(\mathbf{\Sigma}_\Omega) = \frac{a_{11}a_{22} - a_{12}a_{21}}{N^2},$$

which implies

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \mathbf{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{N^N \exp\left(-N\right)}{(2\pi)^N \left(a_{11}a_{22} - a_{12}a_{21}\right)^{\frac{N}{2}}} = \frac{N^N \exp\left(-N\right)}{(2\pi)^N (1-r^2)^{\frac{N}{2}} a_{11}^{\frac{N}{2}} a_{22}^{\frac{N}{2}}}.$$

Let $\sigma^2 = \sigma_1\sigma_2$ and $\tau = \sigma_1/\sigma_2$, then the likelihood function under the null hypothesis ($\rho = \rho_0$) is

$$\frac{1}{(2\pi)^N (1-\rho_0^2)^{\frac{N}{2}} (\sigma^2)^N} \exp\left(-\frac{a_{11}/\tau + \tau/a_{22} - 2\rho_0 a_{12}}{2\sigma^2(1-\rho_0^2)}\right) \tag{8}$$

The maximum of (8) with respect to τ occurs at

$$\hat{\tau} = \sqrt{a_{11}/a_{22}},$$

then the concentrated likelihood is

$$\frac{1}{(2\pi)^N (1-\rho_0^2)^{\frac{N}{2}} (\sigma^2)^N} \exp\left(-\frac{\sqrt{a_{11}}\,\sqrt{a_{22}}\,(1-\rho_0 r)}{\sigma^2(1-\rho_0^2)}\right). \tag{9}$$

The maximum of (9) occurs at

$$\hat{\sigma}^2 = \frac{\sqrt{a_{11}}\,\sqrt{a_{22}}(1-\rho_0 r)}{N(1-\rho_0^2)}.$$

The likelihood ratio criterion is, therefore,

$$\frac{\sup_\omega L(\mathbf{x}, \boldsymbol{\theta})}{\sup_\Omega L(\mathbf{x}, \boldsymbol{\theta})} = \left(\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right)^{\frac{N}{2}}.$$

□

Lemma 5.2. *For random vector*

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \text{and} \qquad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}.$$

Then $\mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)] = 0$ *and* $\mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)(x_l - \mu_l)] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

Theorem 5.4. *Let*

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

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ *are independently distributed according to* $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ *and* $n = N - 1$. *Then the limiting distribution of*

$$\mathbf{B}(n) = \frac{1}{\sqrt{n}}\left(\mathbf{A}(n) - n\mathbf{\Sigma}\right)$$

is normal with mean $\mathbf{0}$ *and covariance* $\mathbb{E}\left[b_{ij}(n) b_{kl}(n)\right] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

Proof. We have

$$\mathbf{A}(n) = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top,$$

where $\mathbf{z}_1, \dots, \mathbf{z}_n$ are distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We arrange the elements of $\mathbf{z}_\alpha \mathbf{z}_\alpha^\top$ in a vector such as

$$\mathbf{y}_\alpha = \begin{bmatrix} z_{1\alpha}^2 \\ z_{1\alpha} z_{2\alpha} \\ \vdots \\ z_{2\alpha}^2 \\ \vdots \\ z_{p\alpha}^2 \end{bmatrix}.$$

The second moments of $\mathbf{y}_\alpha$ can be deduced from the forth moments of $\mathbf{z}_\alpha$ by using Lemma 5.2, that is,

$$\mathbb{E}[z_{i\alpha} z_{j\alpha}] = \sigma_{ij}, \qquad \mathbb{E}[z_{i\alpha} z_{j\alpha} z_{k\alpha} z_{l\alpha}] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},$$

and

$$\mathbb{E}[(z_{i\alpha} z_{j\alpha} - \sigma_{ij})(z_{k\alpha} z_{l\alpha} - \sigma_{kl})] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \tag{10}$$

Arranging the elements of $\mathbf{\Sigma}$ and $\mathbf{A}(n)$ as

$$\boldsymbol{\nu} = \begin{bmatrix} \sigma_{11} \\ \sigma_{12} \\ \vdots \\ \sigma_{22} \\ \vdots \\ \sigma_{pp} \end{bmatrix} \qquad \text{and} \qquad \mathbf{w}(n) = \begin{bmatrix} a_{11}(n) \\ a_{12}(n) \\ \vdots \\ a_{22}(n) \\ \vdots \\ a_{pp}(n) \end{bmatrix}$$

we obtain

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

Since $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\mu}$ and covariance of $\mathbf{y}_\alpha$ satisfies (10), the multivariate central limit theorem implies the desired result. $\square$

Remark 5.1. *In the analysis for the asymptotic distribution of sample correlation, we apply this theorem with*

$$\mathbf{A}(n) = \mathbf{C}(n) \quad \textit{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Then the covariance matrix of limiting distribution of the vector

$$\sqrt{n}(\mathbf{u}(n) - \mathbf{b}) = \frac{1}{\sqrt{n}} \left(\begin{bmatrix} c_{ii}(n) \\ c_{jj}(n) \\ c_{ij}(n) \end{bmatrix} - n\mathbf{b} \right)$$

is

$$\begin{bmatrix} \sigma_{11}\sigma_{11} + \sigma_{11}\sigma_{11} & \sigma_{12}\sigma_{12} + \sigma_{12}\sigma_{12} & \sigma_{11}\sigma_{12} + \sigma_{12}\sigma_{11} \\ \sigma_{12}\sigma_{12} + \sigma_{12}\sigma_{12} & \sigma_{22}\sigma_{22} + \sigma_{22}\sigma_{22} & \sigma_{21}\sigma_{22} + \sigma_{22}\sigma_{21} \\ \sigma_{11}\sigma_{12} + \sigma_{12}\sigma_{11} & \sigma_{21}\sigma_{22} + \sigma_{22}\sigma_{21} & \sigma_{11}\sigma_{22} + \sigma_{12}\sigma_{21} \end{bmatrix} = \begin{bmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{bmatrix}.$$

Theorem 5.5. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partition the variables as*

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

Define $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$,

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{\mathbf{x}}^{(1)} \\ \bar{\mathbf{x}}^{(2)} \end{bmatrix} = \frac{1}{N}\sum_{\alpha=1}^{N} \begin{bmatrix} \mathbf{x}_\alpha^{(1)} \\ \mathbf{x}_\alpha^{(2)} \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

Then the maximum likelihood estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\mathbf{B}$, $\boldsymbol{\Sigma}_{11.2}$ and $\boldsymbol{\Sigma}_{22}$ are

$$\hat{\boldsymbol{\mu}}^{(1)} = \bar{\mathbf{x}}^{(1)}, \quad \hat{\boldsymbol{\mu}}^{(2)} = \bar{\mathbf{x}}^{(2)}, \quad \hat{\mathbf{B}} = \mathbf{A}_{12}\mathbf{A}_{22}^{-1},$$

$$\hat{\boldsymbol{\Sigma}}_{11.2} = \frac{1}{N}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{22} = \frac{1}{N}\mathbf{A}_{22}.$$

Proof. The correspondence between $\boldsymbol{\Sigma}$ and $(\boldsymbol{\Sigma}_{11.2}, \mathbf{B}, \boldsymbol{\Sigma}_{22})$ is one-by-one since

$$\boldsymbol{\Sigma}_{12} = \mathbf{B}\boldsymbol{\Sigma}_{22} \quad \text{and} \quad \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{11.2} + \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{B}^\top,$$

which implies the desired result. □

6 The Wishart Distribution

Theorem 6.1. *Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be independently distributed, each according to $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $n \geq p$; let*

$$\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top = \mathbf{T}^*\mathbf{T}^{*\top},$$

where $t_{ij}^* = 0$ for $i < j$, and $t_{ii}^* > 0$ for $i = 1, \ldots, p$. Then the density of $\mathbf{T}^*$ is

$$\frac{\prod_{i=1}^{p} {t_{ii}^*}^{n-i} \exp\left(-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\mathbf{T}^*\mathbf{T}^{*\top}\big)\right)}{2^{\frac{p(n-2)}{2}}\pi^{\frac{p(p-1)}{4}}(\det(\boldsymbol{\Sigma}))^{\frac{n}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

Proof. Let $\mathbf{C}$ be the lower triangular matrix ($c_{ij} = 0$, $i < j$) such that $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$ and $c_{ii} > 0$. Define $\mathbf{y}_\alpha = \mathbf{C}^{-1}\mathbf{z}_\alpha$ for $\alpha = 1, \ldots, n$, which are be independently distributed, each according to $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$. We have $\mathbf{T}^*\mathbf{T}^{*\top} = \sum_{\alpha=1}^{n} \mathbf{C}\mathbf{y}_\alpha\mathbf{y}_\alpha^\top\mathbf{C}^\top = \mathbf{C}\mathbf{T}\mathbf{T}^\top\mathbf{C}^\top$. Let $\mathbf{T} = \mathbf{C}^{-1}\mathbf{T}^*$, then the matrix $\mathbf{T}$ is the lower triangular with $t_{ii} > 0$ and we have

$$\mathbf{T}\mathbf{T}^\top = \mathbf{C}^{-1}\mathbf{T}^*\mathbf{T}^{*\top}\mathbf{C}^{-1} = \sum_{\alpha=1}^{n} \mathbf{C}^{-1}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\mathbf{C}^{-1} = \sum_{\alpha=1}^{n} \mathbf{y}_\alpha\mathbf{y}_\alpha^\top.$$

The lemma in slides have shown that random variables $t_{i1}, \ldots, t_{ii-1}$ are independently distributed and t_{ij} is distributed according to $\mathcal{N}(0, 1)$ for $i > j$; and t_{ii} has the χ^2-distribution with $n - i + 1$ degrees of freedom. Hence, the density of $w = t_{ii}^2$ is

$$\frac{1}{2^{\frac{1}{2}(n+1-i)}\Gamma\left(\frac{1}{2}(n+1-i)\right)} w^{\frac{1}{2}(n+1-i)-1} \exp\left(-\frac{w}{2}\right)$$

and the density of $t_{ii} = \sqrt{w}$ is (using $\mathrm{d}w/\mathrm{d}t_{ii} = 2t_{ii}$)

$$\frac{1}{2^{\frac{1}{2}(n+1-i)}\Gamma\left(\frac{1}{2}(n+1-i)\right)} (t_{ii}^2)^{\frac{1}{2}(n+1-i)-1} \exp\left(-\frac{t_{ii}^2}{2}\right) \cdot (2t_{ii}) = \frac{1}{2^{\frac{n-i-1}{2}}\Gamma\left(\frac{1}{2}(n+1-i)\right)} t_{ii}^{n-i} \exp\left(-\frac{t_{ii}^2}{2}\right)$$

Then the joint density of t_{ij} for $j = 1, \dots, i$, $i = 1, \dots, p$ is

$$
\begin{aligned}
&\prod_{i=1}^{p}\prod_{j=1}^{i-1} \frac{\exp\left(-\frac{1}{2}t_{ij}^2\right)}{\sqrt{2\pi}} \cdot \prod_{i=1}^{p} \frac{t_{ii}^{n-i}\exp\left(-\frac{1}{2}t_{ii}^2\right)}{2^{\frac{n-i-1}{2}}\Gamma\left(\frac{1}{2}(n+1-i)\right)}\\
=&\prod_{i=1}^{p} \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{i-1} t_{ij}^2\right)}{(2\pi)^{\frac{i-1}{2}}} \cdot \prod_{i=1}^{p} \frac{t_{ii}^{n-i}\exp\left(-\frac{t_{ii}^2}{2}\right)}{2^{\frac{n-i-1}{2}}\Gamma\left(\frac{1}{2}(n+1-i)\right)}\\
=&\prod_{i=1}^{p} \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{i} t_{ij}^2\right)t_{ii}^{n-i}}{2^{\frac{n}{2}-1}\pi^{\frac{i-1}{2}}\Gamma\left(\frac{1}{2}(n+1-i)\right)}\\
=&\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{i} t_{ij}^2\right)\prod_{i=1}^{p} t_{ii}^{n-i}}{2^{\frac{p(n-2)}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.
\end{aligned}
$$

The Jacobian of the transformation from $\mathbf{T}$ to $\mathbf{T}^* = \mathbf{CT}$ can be written as

$$
\begin{bmatrix} t_{11}^* \\ t_{21}^* \\ t_{22}^* \\ \vdots \\ t_{p1}^* \\ \vdots \\ t_{pp}^* \end{bmatrix}
=
\begin{bmatrix}
c_{11} & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\times & c_{22} & 0 & \cdots & 0 & \cdots & 0 \\
\times & \times & c_{22} & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\
\times & \times & \times & \cdots & c_{pp} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\times & \times & \times & \dots & \times & \dots & c_{pp}
\end{bmatrix}
\begin{bmatrix} t_{11} \\ t_{21} \\ t_{22} \\ \vdots \\ t_{p1} \\ \vdots \\ t_{pp} \end{bmatrix}.
$$

Since the matrix of the transformation is triangular, its determinant is the product of the diagonal elements, namely, $\prod_{i=1}^{p} c_{ii}^i$. The Jacobian of the transformation from $\mathbf{T}$ to $\mathbf{T}^*$ is the reciprocal of the determinant. We also have $t_{ii} = t_{ii}^*/c_{ii}$, $\prod_{i=1}^{p} c_{ii}^2 = \det(\mathbf{C})\det(\mathbf{C}^\top) = \det(\mathbf{\Sigma})$ and

$$
\begin{aligned}
&\sum_{i=1}^{p}\sum_{j=1}^{i} t_{ij}^2 = \operatorname{tr}\left(\mathbf{T}\mathbf{T}^\top\right) = \operatorname{tr}\left(\mathbf{C}^{-1}\mathbf{T}^*{\mathbf{T}^*}^\top\mathbf{C}^{-\top}\right)\\
=&\operatorname{tr}\left(\mathbf{T}^*{\mathbf{T}^*}^\top\mathbf{C}^{-\top}\mathbf{C}^{-1}\right) = \operatorname{tr}\left(\mathbf{T}^*{\mathbf{T}^*}^\top\mathbf{\Sigma}^{-1}\right)
\end{aligned}
$$

Then the density of $\mathbf{T}^*$ is

$$
\begin{aligned}
&\frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{T}^*{\mathbf{T}^*}^\top\mathbf{\Sigma}^{-1}\right)\right)\prod_{i=1}^{p}(t_{ii}^*/c_{ii})^{n-i}}{2^{\frac{p(n-2)}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot \left(\prod_{i=1}^{p} c_{ii}^i\right)^{-1}\\
=&\frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{T}^*{\mathbf{T}^*}^\top\mathbf{\Sigma}^{-1}\right)\right)\prod_{i=1}^{p}{t_{ii}^*}^{n-i}}{2^{\frac{p(n-2)}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot \left(\prod_{i=1}^{p} c_{ii}\right)^{n}\\
=&\frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}^*{\mathbf{T}^*}^\top\right)\right)\prod_{i=1}^{p}{t_{ii}^*}^{n-i}}{2^{\frac{p(n-2)}{2}}\pi^{\frac{p(p-1)}{4}}(\det(\mathbf{\Sigma}))^{\frac{n}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}.
\end{aligned}
$$

□

Theorem 6.2. *Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independently distributed, each according to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $n \geq p$. Then the density of $\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha\mathbf{z}_\alpha^\top$ is*

$$
\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}(\det(\mathbf{\Sigma}))^{\frac{n}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}
$$

for $\mathbf{A}$ positive definite, and 0 otherwise.

Proof. Following the proof of Theorem 6.1, we only needs to consider the transformation from $\mathbf{T}^*$ to $\mathbf{A}$. The relation $\mathbf{A} = \mathbf{T}^*{\mathbf{T}^*}^\top$ means we can write

$$a_{hi} = \sum_{j=1}^{i} t^*_{hj} t^*_{ij} \quad \text{for } h \geq i.$$

Then we have

$$\frac{\partial a_{hi}}{\partial t^*_{kl}} = 0 \quad \text{for } k > i; \text{ or } k = h, l > i.$$

that is, $\partial a_{hi}/\partial t^*_{kl} = 0$ if k, l, is beyond h, i in the lexicographic ordering. The Jacobian matrix of the transformation from $\mathbf{A}$ to $\mathbf{T}^*$ is a lower triangular matrix with diagonal elements

$$\begin{aligned}
\frac{\partial a_{hh}}{\partial t^*_{hh}} &= 2t^*_{hh} \quad \text{for} \quad h = 1, \dots, p; \\
\frac{\partial a_{hi}}{\partial t^*_{hi}} &= t^*_{ii} \quad \text{for} \quad h > i;
\end{aligned}$$

The determinant of the Jacobian matrix is therefore

$$2^p \prod_{i=1}^{p} {t^*_{ii}}^{p+1-i}$$

The Jacobian of the transformation from $\mathbf{T}^*$ to $\mathbf{A}$ is the reciprocal. Hence, the desnity of $\mathbf{A}$ is

$$\begin{aligned}
&\frac{\prod_{i=1}^{p} {t^*_{ii}}^{n-i} \exp\left(-\frac{1}{2}\mathrm{tr}\big(\mathbf{\Sigma}^{-1}\mathbf{A}\big)\right)}{2^{\frac{p(n-2)}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot \left(2^p \prod_{i=1}^{p} {t^*_{ii}}^{p+1-i}\right)^{-1} \\
=&\frac{\prod_{i=1}^{p} {t^*_{ii}}^{n-p-1} \exp\left(-\frac{1}{2}\mathrm{tr}\big(\mathbf{\Sigma}^{-1}\mathbf{A}\big)\right)}{2^{\frac{pn}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \\
=&\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\big(\mathbf{\Sigma}^{-1}\mathbf{A}\big)\right)}{2^{\frac{pn}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}.
\end{aligned}$$

□

Corollary 6.1. *Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be independently distributed, each according to $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $N > p$. Then the distribution of $\mathbf{S} = \frac{1}{n}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$ is $\mathcal{W}\left(\frac{1}{n}\mathbf{\Sigma}, n\right)$.*

Proof. The matrix $\mathbf{S}$ has the distribution of

$$\mathbf{S} = \sum_{\alpha=1}^{n} \frac{\mathbf{z}_\alpha}{\sqrt{n}} \left(\frac{\mathbf{z}_\alpha}{\sqrt{n}}\right)^\top,$$

where each $\frac{\mathbf{z}_1}{\sqrt{n}}, \dots, \frac{\mathbf{z}_n}{\sqrt{n}}$ are independently distributed, each according to $\mathcal{N}(\mathbf{0}, \frac{1}{n}\mathbf{I})$. Theorem 6.2 implies this corollary. □

Lemma 6.1. *Given $\mathbf{B}$ positive semidefinite and $\mathbf{A}$ positive definite, there exists a non-singular matrix $\mathbf{F}$ such that $\mathbf{F}^\top\mathbf{B}\mathbf{F} = \mathbf{D}$ and $\mathbf{F}^\top\mathbf{A}\mathbf{F} = \mathbf{I}$, where $\mathbf{D}$ is diagonal.*

Proof. Let the spectral decomposition of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top$ and $\mathbf{E} = \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}^{-\frac{1}{2}}$, then $\mathbf{E}^\top\mathbf{A}\mathbf{E} = \mathbf{I}$. Let the spectral decomposition of $\mathbf{B}^* = \mathbf{E}^\top\mathbf{B}\mathbf{E}$ be $\mathbf{B}^* = \mathbf{U}_{\mathbf{B}^*}\mathbf{\Sigma}_{\mathbf{B}^*}\mathbf{U}_{\mathbf{B}^*}^\top$, then

$$\mathbf{\Sigma}_{\mathbf{B}^*} = \mathbf{U}_{\mathbf{B}^*}^\top\mathbf{B}^*\mathbf{U}_{\mathbf{B}^*} = \mathbf{U}_{\mathbf{B}^*}^\top\mathbf{E}^\top\mathbf{B}\mathbf{E}\mathbf{U}_{\mathbf{B}^*}.$$

Letting $\mathbf{F} = \mathbf{E}\mathbf{U}_{\mathbf{B}^*}$ and $\mathbf{D} = \mathbf{\Sigma}_{\mathbf{B}^*}$ proves this lemma. □

Lemma 6.2. *The characteristic function of chi-square distribution with the degree of freedom n is*

$$\phi(t) = (1 - 2\mathrm{i}t)^{-\frac{n}{2}}.$$

Proof. Let x be distributed according to χ^2-distribution with the degree of freedom n, then its density is

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right).$$

We have (using the density of χ^2-distribution with the degree of freedom $2k+n$)

$$\begin{aligned}
\phi(t) =& \mathbb{E}\left[\exp(\mathrm{i}tx)\right] \\
=& \int_0^{+\infty} \exp(\mathrm{i}tx) \cdot \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} \left(\sum_{k=0}^{\infty} \frac{(\mathrm{i}tx)^k}{k!}\right) x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \int_0^{+\infty} x^{k+\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \cdot 2^{k+\frac{n}{2}}\Gamma\left(k+\frac{n}{2}\right) \int_0^{+\infty} \frac{1}{2^{k+\frac{n}{2}}\Gamma\left(k+\frac{n}{2}\right)} x^{k+\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathrm{d}x \\
=& \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \sum_{k=0}^{\infty} \frac{(\mathrm{i}t)^k}{k!} \cdot 2^{k+\frac{n}{2}}\Gamma\left(k+\frac{n}{2}\right) \\
=& 1 + \sum_{k=1}^{\infty} \frac{(2\mathrm{i}t)^k}{k!} \cdot \prod_{j=0}^{k-1} \left(j + \frac{n}{2}\right) \\
=& (1 - 2\mathrm{i}t)^{-\frac{n}{2}}.
\end{aligned}$$

□

Theorem 6.3. *If $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent, each with distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, then the characteristic function of $a_{11}, \ldots, a_{pp}$, $2a_{12}, \ldots, 2a_{p-1,p}$, where a_{ij} is the (i,j)-th element of*

$$\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$$

is given by $\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A\Theta}))\right] = \left(\det\left(\mathbf{I} - 2\mathrm{i}\mathbf{\Theta\Sigma}\right)\right)^{-\frac{n}{2}}$, where $\mathbf{\Theta} \in \mathbb{R}^{p\times p}$ is symmetric.

Proof. The characteristic function of $a_{11}, \ldots, a_{pp}$, $2a_{12}, \ldots, 2a_{p-1,p}$ is

$$\begin{aligned}
&\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A\Theta}))\right] \\
=& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathrm{tr}\left(\sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top \mathbf{\Theta}\right)\right)\right] \\
=& \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathrm{tr}\left(\sum_{\alpha=1}^{n} \mathbf{z}_\alpha^\top \mathbf{\Theta} \mathbf{z}_\alpha\right)\right)\right] \\
=& \mathbb{E}\left[\exp\left(\mathrm{i}\sum_{\alpha=1}^{n} \mathbf{z}_\alpha^\top \mathbf{\Theta} \mathbf{z}_\alpha\right)\right] \\
=& \prod_{\alpha=1}^{n} \mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}_\alpha^\top \mathbf{\Theta} \mathbf{z}_\alpha\right)\right]
\end{aligned}$$

$$=\left(\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]\right)^n,$$

where $\mathbf{z}\sim\mathcal{N}_p(\mathbf{0},\mathbf{\Sigma})$. Lemma 6.1 means there exists non-singular matrix $\mathbf{F}$ such that

$$\mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F}=\mathbf{I}\quad\text{and}\quad\mathbf{F}^\top\mathbf{\Theta}\mathbf{F}=\mathbf{D},$$

where $\mathbf{D}\in\mathbb{R}^{p\times p}$ is diagonal. If we set $\mathbf{z}=\mathbf{F}\mathbf{y}$, then

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{y}^\top\mathbf{F}^\top\mathbf{\Theta}\mathbf{F}\mathbf{y}\right)\right]\\
=&\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{y}^\top\mathbf{D}\mathbf{y}\right)\right]\\
=&\mathbb{E}\left[\prod_{j=1}^p\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right]\\
=&\prod_{j=1}^p\mathbb{E}\left[\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right].
\end{aligned}$$

Note that the term of $\mathbb{E}\left[\exp\left(\mathrm{i}\,d_{jj}y_j^2\right)\right]$ is the characteristic function of the χ^2-distribution with one degree of freedom, namely $(1-2\mathrm{i}d_{jj})^{-\frac{1}{2}}$. Thus, we have

$$\mathbb{E}\left[\exp\left(\mathrm{i}\,\mathbf{z}^\top\mathbf{\Theta}\mathbf{z}\right)\right]=\prod_{j=1}^p(1-2\mathrm{i}d_{jj})^{-\frac{1}{2}}=(\det(\mathbf{I}-2\mathrm{i}\mathbf{D}))^{-\frac{1}{2}}.$$

We also have

$$\begin{aligned}
&\det(\mathbf{I}-2\mathrm{i}\mathbf{D})\\
=&\det\left(\mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F}-2\mathrm{i}\mathbf{F}^\top\mathbf{\Theta}\mathbf{F}\right)\\
=&\det\left(\mathbf{F}^\top\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)\mathbf{F}\right)\\
=&(\det(\mathbf{F}))^2\det\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)
\end{aligned}$$

and $\mathbf{F}^\top\mathbf{\Sigma}^{-1}\mathbf{F}=\mathbf{I}$ means $\det(\mathbf{F})=(\det(\mathbf{\Sigma}))^{\frac{1}{2}}$. Combing the above results, we obtain

$$\det(\mathbf{I}-2\mathrm{i}\mathbf{D})=\det(\mathbf{\Sigma})\det\left(\mathbf{\Sigma}^{-1}-2\mathrm{i}\mathbf{\Theta}\right)=\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)$$

and

$$\mathbb{E}\left[\exp(\mathrm{i}\,\mathrm{tr}(\mathbf{A}\mathbf{\Theta}))\right]=\left(\det\left(\mathbf{I}-2\mathrm{i}\mathbf{\Theta}\mathbf{\Sigma}\right)\right)^{-\frac{n}{2}}.$$

□

Theorem 6.4. *Let $\mathbf{A}$ and $\mathbf{\Sigma}$ be partitioned into q and $p-q$ rows and columns,*

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11}&\mathbf{A}_{12}\\\mathbf{A}_{21}&\mathbf{A}_{22}\end{bmatrix},\qquad\mathbf{\Sigma}=\begin{bmatrix}\mathbf{\Sigma}_{11}&\mathbf{\Sigma}_{12}\\\mathbf{\Sigma}_{21}&\mathbf{\Sigma}_{22}\end{bmatrix}$$

If $\mathbf{A}$ is distributed according to $\mathcal{W}(\mathbf{\Sigma},n)$, then $\mathbf{A}_{11}$ is distributed according to $\mathcal{W}(\mathbf{\Sigma}_{11},n)$.

Proof. The assumption means $\mathbf{A}$ is distributed as $\mathbf{A}=\sum_{\alpha=1}^n\mathbf{z}_\alpha\mathbf{z}_\alpha^\top$, where the $\mathbf{z}_\alpha$ are independent, each with the distribution $\mathcal{N}(\mathbf{0},\mathbf{\Sigma})$. Partition $\mathbf{z}_\alpha$ into subvectors of q and $p-q$ components such that

$$\mathbf{z}_\alpha=\begin{bmatrix}\mathbf{z}_\alpha^{(1)}\\\mathbf{z}_\alpha^{(2)}\end{bmatrix}.$$

Then $\mathbf{z}_1^{(1)},\dots,\mathbf{z}_\alpha^{(n)}$ are independent, each with the distribution $\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_{11})$, and $\mathbf{A}_{11}$ is distributed as

$$\sum_{\alpha=1}^n\mathbf{z}_\alpha^{(1)}\left(\mathbf{z}_\alpha^{(1)}\right)^\top,$$

which has the distribution $\mathcal{W}(\mathbf{\Sigma}_{11},n)$. □

Theorem 6.5. *Let* $\mathbf{A}$ *and* $\mathbf{\Sigma}$ *be partitioned into* $p_1, \ldots, p_q$ *rows and columns with* $p = p_1, \ldots, p_q$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1q} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{q1} & \cdots & \mathbf{A}_{qq} \end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \cdots & \mathbf{\Sigma}_{1q} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{q1} & \cdots & \mathbf{\Sigma}_{qq} \end{bmatrix}$$

If $\mathbf{\Sigma} = \mathbf{0}$ *for* $i \neq j$ *and if* $\mathbf{A} \sim \mathcal{W}(\mathbf{\Sigma}, n)$, *then* $\mathbf{A}_{11}, \ldots, \mathbf{A}_{qq}$ *are independently distributed and* $\mathbf{A}_{jj} \sim \mathcal{W}(\mathbf{\Sigma}_{jj}, n)$ *for* $j = 1, \ldots, q$.

Proof. The assumption means $\mathbf{A}$ is distributed as $\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$, where the $\mathbf{z}_\alpha$ are independent, each with the distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Partition $\mathbf{z}_\alpha$ into subvectors

$$\mathbf{z}_\alpha = \begin{bmatrix} \mathbf{z}_\alpha^{(1)} \\ \vdots \\ \mathbf{z}_\alpha^{(q)} \end{bmatrix}.$$

as $\mathbf{A}$ and $\mathbf{\Sigma}$ be portioned. Since $\mathbf{\Sigma}_{ij} = \mathbf{0}$, the sets $\mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_n^{(1)}, \ldots, \mathbf{z}_1^{(q)}, \ldots, \mathbf{z}_n^{(q)}$ are independent. Then $\mathbf{A}_{11} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha^{(1)} \left(\mathbf{z}_\alpha^{(1)}\right)^\top, \ldots, \mathbf{A}_{qq} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha^{(q)} \left(\mathbf{z}_\alpha^{(q)}\right)^\top$ are independent. The rest of the proof follows from Theorem 6.4. ☐

Theorem 6.6. *If* $\mathbf{x}_1, \ldots, \mathbf{x}_N$ *are independent, each with distribution* $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, *where*

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix}$$

then the density of the sample correlation coefficients is given by

$$\frac{\left(\Gamma\left(\frac{n}{2}\right)\right)^p \left(\det\left([r_{ij}]_{ij}\right)\right)^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)}.$$

where $n = N - 1$.

Proof. The density of $\mathbf{A}$ is

$$\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} \frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{np}{2}} \prod_{i=1}^{p} \sigma_{ii}^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)}.$$

We consider the transformation

1. $a_{ij} = \sqrt{a_{ii}}\,\sqrt{a_{jj}}\, r_{ij}$ for $i < j$,
2. $a_{ii} = a_{ii}$ otherwise,

which is from

$$\{r_{ij} : i < j, \quad i, j = 1, \ldots, p\} \cup \{a_{ii} : i = 1, \ldots, p\}$$

to

$$\{a_{ij} : i < j, \quad i, j = 1, \ldots, p\} \cup \{a_{ii} : i = 1, \ldots, p\}.$$

The determinant of Jacobian for this transformation is

$$\prod_{i=1}^{p} \prod_{j=1}^{i-1} \sqrt{a_{ii}}\,\sqrt{a_{jj}} = \prod_{i=1}^{p} a_{ii}^{\frac{p-1}{2}}.$$

The joint density of $\{r_{ij} : i < j, \quad i, j = 1, \dots, p\} \cup \{a_{ii} : i = 1, \dots, p\}$ is

$$
\begin{aligned}
&\frac{\left(\det\left([\sqrt{a_{ii}}\,\sqrt{a_{jj}}\,r_{ij}]_{ij}\right)\right)^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{np}{2}}\prod_{i=1}^{p}\sigma_{ii}^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \cdot \prod_{i=1}^{p} a_{ii}^{\frac{p-1}{2}} \\
=&\frac{\left(\prod_{i=1}^{p} a_{ii}\right)^{\frac{n-p+1}{2}}\left(\det\left([r_{ij}]_{ij}\right)\right)^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{np}{2}}\prod_{i=1}^{p}\sigma_{ii}^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \cdot \prod_{i=1}^{p} a_{ii}^{\frac{p-1}{2}} \\
=&\frac{\left(\det\left([r_{ij}]_{ij}\right)\right)^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)} \cdot \prod_{i=1}^{p}\frac{a_{ii}^{\frac{n}{2}-1}\exp\left(-\frac{a_{ii}}{2\sigma_{ii}}\right)}{2^{\frac{n}{2}}\sigma_{ii}^{\frac{n}{2}}},
\end{aligned}
$$

where $r_{ii} = 1$. Let $u_i = a_{ii}/(2\sigma_{ii})$, then

$$
\int_0^\infty \frac{a_{ii}^{\frac{n}{2}-1}\exp\left(-\frac{a_{ii}}{2\sigma_{ii}}\right)}{2^{\frac{n}{2}}\sigma_{ii}^{\frac{n}{2}}}\,\mathrm{d}a_{ii} = \int_0^\infty u_i^{\frac{n}{2}-1}\exp\left(-u_i\right)\,\mathrm{d}u_i = \Gamma\Big(\frac{n}{2}\Big).
$$

Combing all above results proves this theorem. □

Theorem 6.7. *If* $\mathbf{A}$ *has the distribution* $\mathcal{W}(\boldsymbol{\Sigma}, n)$ *and* $\boldsymbol{\Sigma}$ *has the a prior distribution* $\mathcal{W}^{-1}(\boldsymbol{\Psi}, m)$*, then the conditional distribution of* $\boldsymbol{\Sigma}$ *given* $\mathbf{A}$ *is the inverted Wishart distribution* $\mathcal{W}^{-1}(\mathbf{A} + \boldsymbol{\Psi}, n + m)$*.*

Proof. The joint density of $\mathbf{A}$ and $\boldsymbol{\Sigma}$,

$$
\begin{aligned}
f(\mathbf{A}, \boldsymbol{\Sigma}) =&\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}}(\det(\boldsymbol{\Sigma}))^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{(\det(\boldsymbol{\Psi}))^{\frac{m}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)\right)}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)} \\
=&\frac{(\det(\boldsymbol{\Psi}))^{\frac{m}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{n+m+p+1}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)}
\end{aligned}
\tag{11}
$$

for $\mathbf{A}$ and $\boldsymbol{\Sigma}$ are positive definite. The marginal density of $\mathbf{A}$ is the integral of (11) over the set of $\boldsymbol{\Sigma}$ positive definite. Since

$$
\begin{aligned}
1 =& \int w^{-1}(\boldsymbol{\Sigma} \mid \mathbf{A} + \boldsymbol{\Psi}, n + m)\,\mathrm{d}\boldsymbol{\Sigma} \\
=&\frac{(\det(\mathbf{A}+\boldsymbol{\Psi}))^{\frac{n+m}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n+m}{2}\right)},
\end{aligned}
$$

we have

$$
\begin{aligned}
f(\mathbf{A}) =& \int f(\mathbf{A}, \boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{\Sigma} \\
=&\frac{(\det(\boldsymbol{\Psi}))^{\frac{m}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)}\int\frac{(\det(\boldsymbol{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}}\mathrm{d}\boldsymbol{\Sigma} \\
=&\frac{(\det(\boldsymbol{\Psi}))^{\frac{m}{2}}(\det(\mathbf{A}))^{\frac{n-p-1}{2}}}{\Gamma_p\left(\frac{n}{2}\right)\Gamma_p\left(\frac{m}{2}\right)} \cdot \Gamma_p\left(\frac{n+m}{2}\right)(\det(\mathbf{A}+\boldsymbol{\Psi}))^{-\frac{n+m}{2}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
f(\boldsymbol{\Sigma} \mid \mathbf{A}) =& \frac{f(\boldsymbol{\Sigma}, \mathbf{A})}{f(\mathbf{A})} \\
=&\frac{(\det(\mathbf{A}+\boldsymbol{\Psi}))^{\frac{n+m}{2}}(\det(\boldsymbol{\Sigma}))^{-\frac{n+m+p+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{A}+\boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\right)\right)}{2^{\frac{(m+n)p}{2}}\Gamma_p\left(\frac{n+m}{2}\right)} \\
=&w^{-1}(\boldsymbol{\Sigma} \mid \mathbf{A} + \boldsymbol{\Psi}, n + m).
\end{aligned}
$$

□

7 Multivariate Linear Regression

Lemma 7.1. *If $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{G} \in \mathbb{R}^{p\times p}$ are positive definite, then $\operatorname{tr}\big(\mathbf{F}\mathbf{A}\mathbf{F}^\top\mathbf{G}\big) > 0$ for non-zero $\mathbf{F} \in \mathbb{R}^{p\times p}$.*

Proof. Let $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$ and $\mathbf{G} = \mathbf{K}\mathbf{K}^\top$, then

$$\begin{aligned}
&\operatorname{tr}\big(\mathbf{F}\mathbf{A}\mathbf{F}^\top\mathbf{G}\big)\\
=&\operatorname{tr}\big(\mathbf{F}\mathbf{H}\mathbf{H}^\top\mathbf{F}^\top\mathbf{K}\mathbf{K}^\top\big)\\
=&\operatorname{tr}\big(\mathbf{H}^\top\mathbf{F}^\top\mathbf{K}\mathbf{K}^\top\mathbf{F}\mathbf{H}\big)\\
=&\operatorname{tr}\big(\mathbf{H}^\top\mathbf{F}^\top\mathbf{G}\mathbf{F}\mathbf{H}\big) > 0.
\end{aligned}$$

□

Theorem 7.1. *If $\mathbf{x}_\alpha$ is an observation from $\mathcal{N}_q(\mathbf{B}\mathbf{z}_\alpha, \mathbf{\Sigma})$ for $\alpha = 1, \dots, N$, where $[\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{N\times q}$ of rank q is given, $\mathbf{\Sigma} \in \mathbb{R}^{q\times q}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$ and $N \geq p+q$, the maximum likelihood estimator of $\mathbf{B}$ is given by $\hat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{-1}$ where*

$$\mathbf{C} = \sum_{\alpha=1}^N \mathbf{x}_\alpha\mathbf{z}_\alpha^\top \qquad \text{and} \qquad \mathbf{A} = \sum_{\alpha=1}^N \mathbf{z}_\alpha\mathbf{z}_\alpha^\top.$$

The maximum likelihood estimator of $\mathbf{\Sigma}$ is give by

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

Proof. The likelihood function is

$$L = \frac{1}{(2\pi)^{\frac{N}{2}}(\det(\mathbf{\Sigma}))^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)\right)$$

Recall that in the maximum likelihood estimation for normal distribution, we use the fact

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

and

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

We shall do the similar thing for the exponential in L. We have

$$\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha) = \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top\right);$$

and for any $\mathbf{H} \in \mathbb{R}^{p\times q}$, it holds that

$$\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top$$

$$= \sum_{\alpha=1}^{N} \Big((\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)^\top + (\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)(\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top + (\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)^\top$$
$$+ (\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \Big).$$

We hope

$$\sum_{\alpha=1}^{N} (\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)^\top = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \mathbf{H}\mathbf{z}_\alpha)(\mathbf{H}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top = \mathbf{0}$$

Hence, we select $\mathbf{H} = \hat{\mathbf{H}}$ as follows

$$\begin{aligned}
& \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)(\hat{\mathbf{H}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top = \mathbf{0} \\
\Longleftarrow & \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)\mathbf{z}_\alpha^\top (\hat{\mathbf{H}} - \mathbf{B})^\top = \mathbf{0} \\
\Longleftarrow & \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)\mathbf{z}_\alpha^\top = \mathbf{0} \\
\Longleftarrow & \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{z}_\alpha^\top = \hat{\mathbf{H}} \sum_{\alpha=1}^{N} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top \\
\Longleftarrow & \hat{\mathbf{H}} = \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{z}_\alpha^\top \left(\sum_{\alpha=1}^{N} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top \right)^{-1}.
\end{aligned}$$

Then we have

$$\sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top = \sum_{\alpha=1}^{N} \Big((\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)^\top + (\hat{\mathbf{H}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\hat{\mathbf{H}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \Big).$$

Lemma 7.1 means

$$\begin{aligned}
& \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \right) \\
= & \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \sum_{\alpha=1}^{N} \Big((\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)^\top + (\hat{\mathbf{H}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\hat{\mathbf{H}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \Big) \right) \\
\geq & \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{H}}\mathbf{z}_\alpha)^\top \right),
\end{aligned}$$

where the equality holds by taking $\mathbf{B} = \hat{\mathbf{H}}$. Hence, the maximum likelihood estimator of $\mathbf{B}$ is given by $\hat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{-1}$. Using Lemma 3.1 with $\mathbf{G} = \mathbf{\Sigma}$ and

$$\mathbf{D} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)^\top,$$

we obtain the the maximum likelihood estimator of $\mathbf{\Sigma}$ is $\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{D}$. □

Remark 7.1. *Let*

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \quad \textit{and} \quad \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_N^\top \end{bmatrix}.$$

We consider the least square problem.

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} f(\mathbf{B}) \triangleq \frac{1}{2}\left\|\mathbf{B}\mathbf{Z}^\top - \mathbf{X}^\top\right\|_F^2,$$

Then, taking the gradient of f be zero means

$$\nabla f(\mathbf{B}) = \frac{\partial}{\partial \mathbf{B}}\mathrm{tr}\left(\frac{1}{2}\mathbf{B}\mathbf{Z}^\top\mathbf{Z}\mathbf{B}^\top - \mathbf{B}\mathbf{Z}^\top\mathbf{X} + \frac{1}{2}\mathbf{X}^\top\mathbf{X}\right) = \mathbf{B}\mathbf{Z}\mathbf{Z}^\top - \mathbf{X}^\top\mathbf{Z} = \mathbf{0}.$$

Hence, we have $\mathbf{B} = \mathbf{X}^\top\mathbf{Z}(\mathbf{Z}\mathbf{Z}^\top)^{-1} = \mathbf{C}\mathbf{A}^{-1} = \hat{\mathbf{B}}$.

Remark 7.2. *The proof means*

$$\begin{aligned}
&\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \\
=&\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)^\top + \sum_{\alpha=1}^N(\hat{\mathbf{B}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\hat{\mathbf{B}}\mathbf{z}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top \\
=&N\hat{\mathbf{\Sigma}} + (\hat{\mathbf{B}} - \mathbf{B})\left(\sum_{\alpha=1}^N \mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)(\hat{\mathbf{B}} - \mathbf{B})^\top \\
=&N\hat{\mathbf{\Sigma}} + (\hat{\mathbf{B}} - \mathbf{B})\mathbf{A}(\hat{\mathbf{B}} - \mathbf{B})^\top.
\end{aligned}$$

Hence, the joint density of $\mathbf{x}_1, \dots, \mathbf{x}_N$ can be written as

$$\begin{aligned}
&\frac{1}{(2\pi)^{\frac{N}{2}}(\det(\mathbf{\Sigma}))^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)\right) \\
=&\frac{1}{(2\pi)^{\frac{N}{2}}(\det(\mathbf{\Sigma}))^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\sum_{\alpha=1}^N(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \mathbf{B}\mathbf{z}_\alpha)^\top\right)\right) \\
=&\frac{1}{(2\pi)^{\frac{N}{2}}(\det(\mathbf{\Sigma}))^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\left(N\hat{\mathbf{\Sigma}} + (\hat{\mathbf{B}} - \mathbf{B})\mathbf{A}(\hat{\mathbf{B}} - \mathbf{B})^\top\right)\right)\right),
\end{aligned}$$

which implies $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Sigma}}$ form a sufficient set statistics for $\mathbf{B}$ and $\mathbf{\Sigma}$.

Theorem 7.2. *The maximum likelihood estimator $\mathbf{B}$ based on a set of N observations, the α-th from $\mathcal{N}(\mathbf{B}\mathbf{z}_\alpha, \mathbf{\Sigma})$, is normally distributed with mean $\mathbf{B}$, and the covariance matrix of the i-th and j-th rows of $\hat{\mathbf{B}}$ is $\sigma_{ij}\mathbf{A}^{-1}$, where $\mathbf{A} = \sum_{\alpha=1}^N \mathbf{z}_\alpha\mathbf{z}_\alpha^\top$. The maximum likelihood estimator $\hat{\mathbf{\Sigma}}$ multiplied by N is independently distributed according to $\mathcal{W}(\mathbf{\Sigma}, N-q)$, where q is the number of components of $\mathbf{z}_\alpha$.*

Proof. For the estimator $\hat{\mathbf{B}}$, we have

$$\mathbb{E}[\hat{\mathbf{B}}] = \mathbb{E}\left[\sum_{\alpha=1}^N \mathbf{x}_\alpha\mathbf{z}_\alpha^\top\mathbf{A}^{-1}\right] = \sum_{\alpha=1}^N \mathbf{B}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top\mathbf{A}^{-1} = \mathbf{B}\left(\sum_{\alpha=1}^N \mathbf{z}_\alpha\mathbf{z}_\alpha^\top\right)\mathbf{A}^{-1} = \mathbf{B}$$

and

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top\right] \\
=&\mathbf{A}^{-1}\mathbb{E}\left[\sum_{\alpha=1}^N\left(x_{i\alpha} - \mathbb{E}[x_{i\alpha}]\right)\mathbf{z}_\alpha\sum_{\gamma=1}^N\left(x_{j\gamma} - \mathbb{E}[x_{j\gamma}]\right)\mathbf{z}_\gamma^\top\right]\mathbf{A}^{-1} \\
=&\mathbf{A}^{-1}\sum_{\alpha=1}^N\sum_{\gamma=1}^N\mathbb{E}\left[\left(x_{i\alpha} - \mathbb{E}[x_{i\alpha}]\right)\left(x_{j\gamma} - \mathbb{E}[x_{j\gamma}]\right)\right]\mathbf{z}_\alpha\mathbf{z}_\gamma^\top\mathbf{A}^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{A}^{-1} \sum_{\alpha=1}^{N} \sum_{\gamma=1}^{N} \delta_{\alpha\gamma} \sigma_{ij} \mathbf{z}_\alpha \mathbf{z}_\gamma^\top \mathbf{A}^{-1} \\
&= \mathbf{A}^{-1} \sum_{\alpha=1}^{N} \sigma_{ij} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top \mathbf{A}^{-1} \\
&= \mathbf{A}^{-1} \big(\sigma_{ij} \mathbf{A} \mathbf{A}^{-1} \big) \\
&= \sigma_{ij} \mathbf{A}^{-1}.
\end{aligned}
$$

From Theorem 4.2, it follows that

$$
\begin{aligned}
N\hat{\mathbf{\Sigma}} &= \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)^\top \\
&= \sum_{\alpha=1}^{N} \Big(\mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \mathbf{x}_\alpha \mathbf{z}_\alpha^\top \hat{\mathbf{B}}^\top - \hat{\mathbf{B}}\mathbf{z}_\alpha \mathbf{x}_\alpha^\top + \hat{\mathbf{B}}\mathbf{z}_\alpha \mathbf{z}_\alpha^\top \hat{\mathbf{B}}^\top \Big) \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{z}_\alpha^\top \hat{\mathbf{B}}^\top - \sum_{\alpha=1}^{N} \hat{\mathbf{B}}\mathbf{z}_\alpha \mathbf{x}_\alpha^\top + \sum_{\alpha=1}^{N} \hat{\mathbf{B}}\mathbf{z}_\alpha \mathbf{z}_\alpha^\top \hat{\mathbf{B}}^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \hat{\mathbf{B}}\mathbf{A}\mathbf{B}^\top - \hat{\mathbf{B}}\mathbf{A}\mathbf{B}^\top + \hat{\mathbf{B}}\mathbf{A}\hat{\mathbf{B}}^\top \\
&= \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha^\top - \hat{\mathbf{B}}\mathbf{A}\hat{\mathbf{B}}^\top.
\end{aligned}
$$

is distributed according to $\mathcal{W}(\mathbf{\Sigma}, N-q)$. □

Theorem 7.3. *The least squares estimator $\hat{\mathbf{B}}$ is the best linear unbiased estimator of $\mathbf{B}$.*

Proof. Let

$$
\tilde{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} f_{j\alpha} x_{j\alpha}
$$

be arbitrary unbiased estimator of β_{ig}, which satisfied

$$
\sum_{\alpha=1}^{N} f_{j\alpha} z_{h\alpha} = \begin{cases} 1, & j = i, h = g, \\ 0, & \text{otherwise.} \end{cases}
$$

Let a^{hg} be the (h,g)-th element of $\mathbf{A}^{-1}$, then the least square estimator can be written as

$$
\hat{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{h=1}^{q} x_{i\alpha} z_{h\alpha} a^{hg},
$$

where $\mathbf{A} = \sum_{\alpha=1}^{N} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$. Then we have

$$
\begin{aligned}
& \mathbb{E}\big[(\tilde{\beta}_{ig} - \beta_{ig})^2\big] \\
=& \mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig} + (\tilde{\beta}_{ig} - \hat{\beta}_{ig}))^2\big] \\
=& \mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig})^2\big] + \mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig})(\tilde{\beta}_{ig} - \hat{\beta}_{ig})\big] + \mathbb{E}\big[(\tilde{\beta}_{ig} - \hat{\beta}_{ig})^2\big]
\end{aligned}
$$

Let $u_{i\alpha} = \mathbb{E}[x_{i\alpha}]$. Since both $\tilde{\beta}_{ig}$ and $\hat{\beta}_{ig}$ are unbiased estimator of β_{ig}, we have

$$
\tilde{\beta}_{ig} - \beta_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} f_{j\alpha} u_{j\alpha}, \quad \hat{\beta}_{ig} - \beta_{ig} = \sum_{\alpha=1}^{N} \sum_{h=1}^{q} u_{i\alpha} z_{h\alpha} a^{hg},
$$

and

$$\tilde{\beta}_{ig} - \hat{\beta}_{ig} = \sum_{\alpha=1}^{N}\sum_{j=1}^{p}\left(f_{j\alpha} - \delta_{ij}\sum_{h=1}^{q} z_{h\alpha}a^{hg}\right)u_{j\alpha},$$

where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. Then we have

$$\begin{aligned}
&\mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig})(\tilde{\beta}_{ig} - \hat{\beta}_{ig})\big] \\
=&\mathbb{E}\left[\sum_{\alpha=1}^{N}\sum_{\gamma=1}^{N}\sum_{h=1}^{q} z_{h\alpha}a^{hg}u_{i\alpha}\sum_{j=1}^{p}\left(f_{j\gamma} - \delta_{ij}\sum_{h'=1}^{q} z_{h'\gamma}a^{h'g}\right)u_{j\gamma}\right] \\
=&\sum_{\alpha=1}^{N}\sum_{h=1}^{q}\sum_{j=1}^{p} z_{h\alpha}a^{hg}\left(f_{j\alpha} - \delta_{ij}\sum_{h'=1}^{q} z_{h'\alpha}a^{h'g}\right)\sigma_{ij} \\
=&\sigma_{ii}a^{gg} - \sigma_{ii}\sum_{h=1}^{q}\sum_{h'=1}^{q} a_{hh'}a^{hg}a^{h'g} \\
=&\sigma_{ii}a^{gg} - \sigma_{ii}a^{gg} = 0.
\end{aligned}$$

Thus

$$\mathbb{E}\big[(\tilde{\beta}_{ig} - \beta_{ig})^2\big] \geq \mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig})^2\big] + \mathbb{E}\big[(\tilde{\beta}_{ig} - \hat{\beta}_{ig})^2\big] \geq \mathbb{E}\big[(\hat{\beta}_{ig} - \beta_{ig})^2\big].$$

□

Theorem 7.4. *The likelihood ratio criterion*

$$\lambda = \frac{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\Omega\big)\big)^{\frac{N}{2}}}{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\omega\big)\big)^{\frac{N}{2}}}.$$

for testing the null hypothesis $\mathbf{B}_1 = \mathbf{0}$ is invariant with respect to transformations $\mathbf{x}_\alpha^* = \mathbf{D}\mathbf{x}_\alpha$ for $\alpha = 1, \ldots, N$ and non-singular $\mathbf{D}$.

Proof. The estimators in terms of $\mathbf{x}_\alpha^*$ are

$$\begin{aligned}
\hat{\mathbf{B}}^* =&\mathbf{D}\mathbf{C}^{-1}\mathbf{A} = \mathbf{D}\hat{\mathbf{B}}, \\
\hat{\boldsymbol{\Sigma}}_\Omega^* =&\frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{D}\mathbf{x}_\alpha - \mathbf{D}\hat{\mathbf{B}}\mathbf{z}_\alpha)(\mathbf{D}\mathbf{x}_\alpha - \mathbf{D}\hat{\mathbf{B}}\mathbf{z}_\alpha)^\top = \mathbf{D}\hat{\boldsymbol{\Sigma}}_\Omega\mathbf{D}^\top, \\
\hat{\mathbf{B}}_{2\omega}^* =&\mathbf{D}\big(\mathbf{C}_2 - \mathbf{B}_1^*\mathbf{A}_{12}\big)\mathbf{A}_{22}^{-1} = \mathbf{D}\hat{\mathbf{B}}_{2\omega}, \\
\hat{\boldsymbol{\Sigma}}_\omega^* =&\frac{1}{N}\sum_{\alpha=1}^{N}\big(\mathbf{D}\mathbf{y}_\alpha - \mathbf{D}\hat{\mathbf{B}}_{2\omega}\mathbf{z}_\alpha^{(2)}\big)\big(\mathbf{D}\mathbf{y}_\alpha - \mathbf{D}\hat{\mathbf{B}}_{2\omega}\mathbf{z}_\alpha^{(2)}\big)^\top = \mathbf{D}\hat{\boldsymbol{\Sigma}}_\omega\mathbf{D}^\top,
\end{aligned}$$

then

$$\lambda^* = \frac{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\Omega^*\big)\big)^{\frac{N}{2}}}{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\omega^*\big)\big)^{\frac{N}{2}}} = \frac{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\Omega\big)\big)^{\frac{N}{2}}}{\big(\det\big(\hat{\boldsymbol{\Sigma}}_\omega\big)\big)^{\frac{N}{2}}}.$$

□

Theorem 7.5. *The statistic*

$$V_1 = \frac{\prod_{g=1}^{q}(\det(\mathbf{A}_g))^{\frac{n_g}{2}}}{(\det(\mathbf{A}))^{\frac{n}{2}}}.$$

is invariant with respect to linear transformation

$$\mathbf{x}^{*(g)} = \mathbf{C}\mathbf{x}^{(g)} + \boldsymbol{\nu}^{(g)}.$$

Proof. We have

$$V_1^* = \frac{\prod_{g=1}^q (\det(\mathbf{A}_g^*))^{\frac{n_g}{2}}}{(\det(\mathbf{A}^*))^{\frac{n}{2}}} = \frac{\prod_{g=1}^q (\det(\mathbf{C}\mathbf{A}_g\mathbf{C}^\top))^{\frac{n_g}{2}}}{(\det(\mathbf{C}\mathbf{A}\mathbf{C}^\top))^{\frac{n}{2}}} = \frac{\prod_{g=1}^q (\det(\mathbf{A}_g))^{\frac{n_g}{2}}}{(\det(\mathbf{A}))^{\frac{n}{2}}} = V_1.$$

□

Theorem 7.6. *Given a set of p-component observation vectors* $\mathbf{x}_1, \dots, \mathbf{x}_N$ *from* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, the likelihood ratio criterion for testing the hypothesis*

$$\boldsymbol{\Sigma} = \sigma_0^2 \Psi_0$$

where Ψ_0 *is specified and* σ^2 *is not specified, is*

$$\frac{(\det(\mathbf{A}\boldsymbol{\Psi}_0^{-1}))^{\frac{N}{2}}}{(\operatorname{tr}(\mathbf{A}\boldsymbol{\Psi}_0^{-1})/p)^{\frac{pN}{2}}}.$$

Proof. Let $\mathbf{C}$ be matrix such that

$$\mathbf{C}\boldsymbol{\Psi}_0\mathbf{C}^\top = \mathbf{I}.$$

and $\mathbf{x}_\alpha^* = \mathbf{C}\mathbf{x}$, $\boldsymbol{\mu}^* = \mathbf{C}\boldsymbol{\mu}$, $\boldsymbol{\Sigma}^* = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top$. Then we have

$$\operatorname{tr}(\mathbf{A}^*) = \operatorname{tr}\left(\sum_{\alpha=1}^N \left(\mathbf{x}_\alpha^* - \bar{\mathbf{x}}_\alpha^*\right)\left(\mathbf{x}_\alpha^* - \bar{\mathbf{x}}_\alpha^*\right)^\top\right) = \operatorname{tr}(\mathbf{C}\mathbf{A}\mathbf{C}^\top) = \operatorname{tr}(\mathbf{A}\mathbf{C}^\top\mathbf{C}) = \operatorname{tr}(\mathbf{A}\boldsymbol{\Psi}_0^{-1})$$

and

$$\det(\mathbf{A}^*) = \det(\mathbf{C}\mathbf{A}\mathbf{C}^\top) = \det(\mathbf{C}))^2 \det(\mathbf{A}) = (\det(\boldsymbol{\Psi}_0))^{-1}\det(\mathbf{A}) = \det(\mathbf{A}\boldsymbol{\Psi}_0^{-1}).$$

Thus

$$\frac{(\det(\mathbf{A}^*))^{\frac{N}{2}}}{(\operatorname{tr}(\mathbf{A}^*)/p)^{\frac{pN}{2}}} = \frac{(\det(\mathbf{A}\boldsymbol{\Psi}_0^{-1}))^{\frac{N}{2}}}{(\operatorname{tr}(\mathbf{A}\boldsymbol{\Psi}_0^{-1})/p)^{\frac{pN}{2}}}.$$

□

8 Principal Components

Theorem 8.1. *Let* $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ *be positive definite. A vector* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta}\|_2 = 1$ *maximizing* $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}$ *must satisfy*

$$(\boldsymbol{\Sigma} - \lambda_1\mathbf{I})\boldsymbol{\beta} = \mathbf{0},$$

where λ_1 *is the largest root of*

$$\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0.$$

Proof. Let

$$\phi(\boldsymbol{\beta}, \lambda) = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}^\top\boldsymbol{\beta} - 1),$$

where λ is a Lagrange multiplier. A vector $\boldsymbol{\beta}$ maximizing $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}$ must satisfy

$$\mathbf{0} = \frac{\partial\phi(\boldsymbol{\beta}, \lambda)}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta},$$

that is $(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{0}$. The constraint $\|\boldsymbol{\beta}\|_2 = 1$ means $\boldsymbol{\Sigma} - \lambda\mathbf{I}$ is singular. Then λ must satisfy

$$\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0.$$

We also have

$$\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} = \lambda,$$

which implies our result. □

Remark 8.1. *For the second principle components $\boldsymbol{\beta}$, we require*

$$0 = \mathbb{E}\big[\boldsymbol{\beta}^\top \mathbf{x} {\boldsymbol{\beta}^{(1)}}^\top \mathbf{x}\big] = \mathbb{E}\big[\boldsymbol{\beta}^\top \mathbf{x}\mathbf{x}^\top \boldsymbol{\beta}^{(1)}\big] = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}^{(1)}.$$

Let

$$\phi_2(\boldsymbol{\beta}, \lambda, \boldsymbol{\nu}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}^\top \boldsymbol{\beta} - 1) - 2\nu \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)}.$$

We require

$$\mathbf{0} = \frac{\partial \phi_2(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} - 2\nu\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)}.$$

Multiplying on the left by ${\boldsymbol{\beta}^{(1)}}^\top$, we have

$$\mathbf{0} = 2{\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda{\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\beta} - 2\nu{\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = -2\nu\lambda_1.$$

Therefore $\nu = 0$ and $\boldsymbol{\beta}$ must satisfy $(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\beta}^\top \boldsymbol{\beta}^{(1)} = 0$, where

$$\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0.$$

Hence, we should take λ by the second-largest root of $\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0$.

Remark 8.2. *For the $(r+1)$-th step, we let*

$$\phi_{r+1}(\boldsymbol{\beta}, \lambda, \boldsymbol{\nu}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}^\top \boldsymbol{\beta} - 1) - 2\sum_{i=1}^{r} \nu_i \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)}$$

and

$$\mathbf{0} = \frac{\partial \phi_{r+1}(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} - 2\sum_{i=1}^{r} \nu_i \boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)}.$$

Similarly, we have $v_j = 0$ and $(\boldsymbol{\Sigma} - \lambda_j\mathbf{I})\boldsymbol{\beta}^{(j)} = \mathbf{0}$ and λ_j is the root of $\det(\boldsymbol{\Sigma} - \lambda\mathbf{I}) = 0$

Remark 8.3. *For the stationary point on surfaces $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} = C$, we let*

$$\psi(\mathbf{x}, \lambda) = \mathbf{x}^\top \mathbf{x} - \lambda \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}.$$

Then

$$\mathbf{0} = \frac{\partial \psi(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = 2\mathbf{x} - 2\lambda\boldsymbol{\Sigma}^{-1}\mathbf{x},$$

that is $\boldsymbol{\Sigma}\mathbf{x} = \lambda\mathbf{x}$. Thus the vectors $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(p)}$ give the principal axis of the ellipsoid. The transformation $\mathbf{u} = \mathbf{B}^\top \mathbf{x}$ is a rotation of the coordinate axes so that the new axes are in the direction of the principal axes of the ellipsoid. In the new coordinates, the ellipsoid is

$$\mathbf{u}^\top \boldsymbol{\Lambda}^{-1}\mathbf{u} = C.$$

Theorem 8.2. *An orthogonal transformation $\mathbf{v} = \mathbf{C}\mathbf{x}$ of a random vector $\mathbf{x}$ with $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ leaves invariant the generalized variance and the sum of the variances of the components.*

Proof. Let $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}$. The generalized variance of $\mathbf{v}$ is

$$\det(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top) = \det(\mathbf{C})\det(\boldsymbol{\Sigma})\det(\mathbf{C}^\top) = \det(\boldsymbol{\Sigma}).$$

The sum of the variances of the components of $\mathbf{v}$ is

$$\sum_{i=1}^{p} \mathbb{E}[v_i^2] = \operatorname{tr}(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top) = \operatorname{tr}(\boldsymbol{\Sigma}\mathbf{C}^\top\mathbf{C}) = \operatorname{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \mathbb{E}[x_i^2].$$

□

Theorem 8.3. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be N observations from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has p different characteristic roots and $N > p$. Then maximum likelihood estimators of $\lambda_1, \ldots, \lambda_p$ and $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(p)}$ consists of the roots $\lambda_1 > \cdots > \lambda_p$ of*

$$\det(\hat{\boldsymbol{\Sigma}} - \lambda \mathbf{I}) = 0$$

and corresponding vectors $\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(p)}$ satisfying $\|\hat{\boldsymbol{\beta}}^{(i)}\|_2 = 1$ and

$$(\hat{\boldsymbol{\Sigma}} - \lambda_i \mathbf{I})\hat{\boldsymbol{\beta}}^{(i)} = \mathbf{0}$$

for $i = 1, \ldots, p$, where $\hat{\boldsymbol{\Sigma}}$ is the the maximum likelihood estimate of $\boldsymbol{\Sigma}$.

Proof. When the roots of $\det(\boldsymbol{\Sigma} - \lambda \mathbf{I})$ are different, each vector $\boldsymbol{\beta}^{(i)}$ uniquely defined except that it can be replaced by $-\boldsymbol{\beta}^{(i)}$. If we require that the first nonzero component of $-\boldsymbol{\beta}^{(i)}$ be positive, then $-\boldsymbol{\beta}^{(i)}$ is uniquely defined. Then the variables $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\mathbf{B}$ is a is a single-valued function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Hence, the set of maximum likelihood estimates of $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\mathbf{B}$ is the same function of $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ (restriction that the first nonzero component of $\boldsymbol{\beta}^{(i)}$ must be positive). ☐

Remark 8.4. *If $\boldsymbol{\Sigma}$ is non-singular, the probability is 1 that the roots of $\lambda_1, \ldots, \lambda_p$ are different. Please see Masashi Okamoto. "Distinctness of the eigenvalues of a quadratic form in a multivariate sample." The Annals of Statistics (1973): 763-765.*

Theorem 8.4. *Let $n\mathbf{S} \sim \mathcal{W}(\boldsymbol{\Sigma}, n)$ and $(\lambda_1, \boldsymbol{\beta}^{(1)})$, $(\lambda_p, \boldsymbol{\beta}^{(p)})$ be two distinct eigen-pairs of $\boldsymbol{\Sigma}$ with $\|\boldsymbol{\beta}^{(1)}\|_2 = \|\boldsymbol{\beta}^{(p)}\|_2 = 1$, then*

$$\frac{n{\boldsymbol{\beta}^{(1)}}^\top \mathbf{S}\boldsymbol{\beta}^{(1)}}{\lambda_1} \quad \textit{and} \quad \frac{n{\boldsymbol{\beta}^{(p)}}^\top \mathbf{S}\boldsymbol{\beta}^{(p)}}{\lambda_p}.$$

are independently distrusted as χ^2-distribution with n degrees of freedom.

Proof. We have

$$n\mathbf{S} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top,$$

where $\mathbf{z}_\alpha$ are independently distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then we have ${\boldsymbol{\beta}^{(1)}}^\top \mathbf{z}_\alpha \sim \mathcal{N}(0, \lambda_1)$, since ${\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = \lambda_1 {\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\beta}^{(1)} = \lambda_1$. Hence, it holds that

$$\frac{n{\boldsymbol{\beta}^{(1)}}^\top \mathbf{S}\boldsymbol{\beta}^{(1)}}{\lambda_1} = \sum_{\alpha=1}^{n} \frac{{\boldsymbol{\beta}^{(1)}}^\top \mathbf{z}_\alpha \mathbf{z}_\alpha^\top \boldsymbol{\beta}^{(1)}}{\lambda_1} = \sum_{\alpha=1}^{n} \left(\frac{{\boldsymbol{\beta}^{(1)}}^\top \mathbf{z}_\alpha}{\sqrt{\lambda_1}} \right)^2 \sim \chi_n^2.$$

are distrusted as χ^2-distribution with n degrees of freedom We also have the similar result for λ_p and $\boldsymbol{\beta}^{(p)}$.

Consider that ${\boldsymbol{\beta}^{(1)}}^\top \mathbf{z}_\alpha$ and ${\boldsymbol{\beta}^{(p)}}^\top \mathbf{z}_\alpha$ are normal distributed with zero mean and

$$\mathbb{E}\left[{\boldsymbol{\beta}^{(1)}}^\top \mathbf{z}_\alpha {\boldsymbol{\beta}^{(p)}}^\top \mathbf{z}_\alpha\right] = {\boldsymbol{\beta}^{(1)}}^\top \mathbb{E}\left[\mathbf{z}_\alpha \mathbf{z}_\alpha^\top\right] \boldsymbol{\beta}^{(p)} = {\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}^{(p)} = \lambda_p {\boldsymbol{\beta}^{(1)}}^\top \boldsymbol{\beta}^{(p)} = 0.$$

Hence, we have proved the desired independence. ☐

Remark 8.5. *Let l and u be two numbers such that*

$$1 - \epsilon = \Pr\left\{nl \le \chi_n^2\right\} \Pr\left\{\chi_n^2 \le nu\right\}.$$

Then we have

$$1 - \epsilon = \Pr\left\{ nl \le \frac{n{\boldsymbol{\beta}^{(1)}}^\top \mathbf{S}\boldsymbol{\beta}^{(1)}}{\lambda_1}, \frac{n{\boldsymbol{\beta}^{(p)}}^\top \mathbf{S}\boldsymbol{\beta}^{(p)}}{\lambda_p} \le nu \right\}$$

$$
\begin{aligned}
&= \Pr\left\{\lambda_1 \leq \frac{\boldsymbol{\beta}^{(1)\top}\mathbf{S}\boldsymbol{\beta}^{(1)}}{l}, \frac{\boldsymbol{\beta}^{(p)\top}\mathbf{S}\boldsymbol{\beta}^{(p)}}{u} \leq \lambda_p\right\} \\
&\leq \Pr\left\{\lambda_1 \leq \frac{\max_{\|\mathbf{b}\|_2=1}\mathbf{b}^\top\mathbf{S}\mathbf{b}}{l}, \frac{\min_{\|\mathbf{b}\|_2=1}\mathbf{b}^\top\mathbf{S}\mathbf{b}}{u} \leq \lambda_p\right\} \\
&= \Pr\left\{\lambda_1 \leq \frac{l_1}{l}, \frac{l_p}{u} \leq \lambda_p\right\} = \Pr\left\{\frac{l_p}{u} \leq \lambda_p \leq \lambda_1 \leq \frac{l_1}{l}\right\}.
\end{aligned}
$$

9 Canonical Correlations

We consider the problem

$$
\max_{\substack{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}=1 \\ \boldsymbol{\gamma}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}=1}} \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma},
$$

where

$$
\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix} \succ \mathbf{0}.
$$

Let

$$
\psi(\boldsymbol{\alpha},\boldsymbol{\gamma},\lambda,\mu) = \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma} - \frac{\lambda}{2}(\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}-1) - \frac{\mu}{2}(\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}-1).
$$

The vectors of derivatives set equal to zero are

$$
\begin{aligned}
\frac{\partial\psi(\boldsymbol{\alpha},\boldsymbol{\gamma},\lambda,\mu)}{\partial\boldsymbol{\alpha}} &= \boldsymbol{\Sigma}_{12}\boldsymbol{\gamma} - \lambda\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \mathbf{0}, \\
\frac{\partial\psi(\boldsymbol{\alpha},\boldsymbol{\gamma},\lambda,\mu)}{\partial\boldsymbol{\gamma}} &= \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\alpha} - \mu\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma} = \mathbf{0}.
\end{aligned}
$$

Multiplication of above ones on the left by $\boldsymbol{\alpha}^\top$ and $\boldsymbol{\gamma}^\top$ respectively, we have

$$
\begin{aligned}
\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma} - \lambda\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \mathbf{0}, \\
\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}_{12}^\top\boldsymbol{\alpha} - \mu\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma} = \mathbf{0}.
\end{aligned}
$$

The constraint means $\lambda = \mu = \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}$. Setting derivatives be zero also can be written as

$$
\begin{bmatrix}-\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\gamma}\end{bmatrix} = \mathbf{0}.
$$

The positive definiteness of $\boldsymbol{\Sigma}$ means $\boldsymbol{\alpha} \neq \mathbf{0}$ and $\boldsymbol{\gamma} \neq \mathbf{0}$, then

$$
\det\left(\begin{bmatrix}-\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22}\end{bmatrix}\right) = 0.
$$

Remark 9.1. *Let*

$$
\boldsymbol{\xi} = \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\gamma}\end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{bmatrix} \quad \textit{and} \quad \mathbf{B} = \begin{bmatrix}\mathbf{0} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \mathbf{0}\end{bmatrix}.
$$

We have the form of generalized eigenvalue decomposition

$$
\mathbf{B}\boldsymbol{\xi} = \lambda\mathbf{A}\boldsymbol{\xi} \quad \textit{and} \quad \det(\mathbf{B}-\lambda\mathbf{A}) = 0.
$$

If $\mathbf{B} = \mathbf{I}$, it is eigenvalue decomposition. For $\mathbf{A} \succ \mathbf{0}$, we have

$$
\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\xi} = \lambda\boldsymbol{\xi} \quad \textit{and} \quad \det(\mathbf{A}^{-1}\mathbf{B}-\lambda\mathbf{I}) = 0,
$$

which corresponds to eigenvalue decomposition on $\mathbf{A}^{-1}\mathbf{B}$.

Remark 9.2. *At $(r+1)$-th step, the uncorrelated conditions for $u = \boldsymbol{\alpha}^\top \mathbf{x}^{(1)}$ and $v = \boldsymbol{\gamma}^\top \mathbf{x}^{(2)}$ are*

$$0 = \mathbb{E}[uu_i] = \mathbb{E}\big[\boldsymbol{\alpha}^\top \mathbf{x}^{(1)} {\mathbf{x}^{(1)}}^\top \boldsymbol{\alpha}^{(i)}\big] = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(i)},$$
$$0 = \mathbb{E}[vv_i] = \mathbb{E}\big[\boldsymbol{\gamma}^\top \mathbf{x}^{(2)} {\mathbf{x}^{(2)}}^\top \boldsymbol{\gamma}^{(i)}\big] = \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(i)}.$$

for $i = 1, \ldots, r$. Then

$$\mathbb{E}[uv_i] = \mathbb{E}\big[\boldsymbol{\alpha}^\top \mathbf{x}^{(1)} {\mathbf{x}^{(2)}}^\top \boldsymbol{\gamma}^{(i)}\big] = \boldsymbol{\alpha}^\top \mathbb{E}\big[\mathbf{x}^{(1)} {\mathbf{x}^{(2)}}^\top\big] \boldsymbol{\gamma}^{(i)} = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}^{(i)} = \lambda \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(i)} = 0.$$
$$\mathbb{E}[vu_i] = \mathbb{E}\big[\boldsymbol{\gamma}^\top \mathbf{x}^{(2)} {\mathbf{x}^{(1)}}^\top \boldsymbol{\alpha}^{(i)}\big] = \boldsymbol{\gamma}^\top \mathbb{E}\big[\mathbf{x}^{(2)} {\mathbf{x}^{(1)}}^\top\big] \boldsymbol{\alpha}^{(i)} = \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha}^{(i)} = \lambda \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(i)} = 0.$$

We now maximize $\mathbb{E}[u_{r+1} v_{r+1}]$. Let

$$\psi_{r+1}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda, \mu) = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \frac{\lambda}{2}(\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - 1) - \frac{\mu}{2}(\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} - 1) - \sum_{i=1}^r \nu_i \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(i)} - \sum_{i=1}^r \theta_i \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(i)}.$$

The vectors of derivatives set equal to zero are

$$\frac{\partial \psi_{r+1}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda, \mu, \boldsymbol{\nu}, \boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \lambda \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - \sum_{i=1}^r \nu_i \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(i)} = \mathbf{0},$$
$$\frac{\partial \psi_{r+1}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda, \mu, \boldsymbol{\nu}, \boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\alpha} - \mu \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} - \sum_{i=1}^r \theta_i \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(i)} = \mathbf{0}.$$

Multiplication of above ones on the left by ${\boldsymbol{\alpha}^{(j)}}^\top$ and ${\boldsymbol{\gamma}^{(j)}}^\top$ for any $j \le r$ respectively gives

$$0 = {\boldsymbol{\alpha}^{(j)}}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \lambda {\boldsymbol{\alpha}^{(j)}}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - \sum_{i=1}^r \nu_i {\boldsymbol{\alpha}^{(j)}}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(i)} = -\nu_j {\boldsymbol{\alpha}^{(j)}}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}^{(j)},$$
$$0 = {\boldsymbol{\gamma}^{(j)}}^\top \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\alpha} - \mu {\boldsymbol{\gamma}^{(j)}}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} - \sum_{i=1}^r \theta_i {\boldsymbol{\gamma}^{(j)}}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(i)} = -\theta_j {\boldsymbol{\gamma}^{(j)}}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}^{(j)}.$$

Hence, we have $v_j = \theta_j = 0$. Then the condition of derivatives is

$$\begin{bmatrix} -\lambda \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix} = \mathbf{0}.$$

where λ satisfies

$$\det\left(\begin{bmatrix} -\lambda \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda \boldsymbol{\Sigma}_{22} \end{bmatrix}\right) = 0;$$

and α and γ satisfy

$$\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 1, \quad \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} = 1, \quad \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}^{(i)} = 0, \quad \textit{and} \quad \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha}^{(i)} = 0.$$

Theorem 9.1. *The canonical correlations are invariant with respect to transformations*

$$\begin{cases} \mathbf{x}^{*(1)} = \mathbf{C}_1 \mathbf{x}^{(1)}, \\ \mathbf{x}^{*(2)} = \mathbf{C}_2 \mathbf{x}^{(2)}, \end{cases}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are non-singular. Additionally, any function of $\boldsymbol{\Sigma}$ that is invariant (under any such transformation) is a function of the canonical correlations.

Proof. The canonical correlations of $\mathbf{x}^{*(1)}$ and $\mathbf{x}^{*(2)}$ are the roots of

$$\begin{aligned}0 =& \det\left(\begin{bmatrix}-\lambda\mathbf{C}_1\mathbf{\Sigma}_{11}\mathbf{C}_1 & \mathbf{C}_1\mathbf{\Sigma}_{12}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{\Sigma}_{21}\mathbf{C}_1^\top & -\lambda\mathbf{C}_2\mathbf{\Sigma}_{22}\mathbf{C}_2^\top\end{bmatrix}\right) \\ =& \det\left(\begin{bmatrix}\mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2\end{bmatrix}\right)\det\left(\begin{bmatrix}-\lambda\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & -\lambda\mathbf{\Sigma}_{22}\end{bmatrix}\right)\det\left(\begin{bmatrix}\mathbf{C}_1^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^\top\end{bmatrix}\right),\end{aligned}$$

which are equivalent to the canonical correlations of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$.

If $\mathbf{f}(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{12}, \mathbf{\Sigma}_{22})$ be a vector function such that $\mathbf{f}(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{12}, \mathbf{\Sigma}_{22}) = \mathbf{f}(\mathbf{C}_1\mathbf{\Sigma}_{11}\mathbf{C}_1^\top, \mathbf{C}_1\mathbf{\Sigma}_{12}\mathbf{C}_2^\top, \mathbf{C}_2\mathbf{\Sigma}_{22}\mathbf{C}_2^\top)$ for any non-singular $\mathbf{C}_1$ and $\mathbf{C}_2$. Let $\mathbf{C}_1 = \mathbf{A}^\top$ and $\mathbf{C}_2 = \mathbf{\Gamma}^\top$, then $\mathbf{f}(\mathbf{C}_1\mathbf{\Sigma}_{11}\mathbf{C}_1^\top, \mathbf{C}_1\mathbf{\Sigma}_{12}\mathbf{C}_2^\top, \mathbf{C}_2\mathbf{\Sigma}_{22}\mathbf{C}_2^\top) = f(\mathbf{I}, \mathrm{diag}(\mathbf{\Lambda}, \mathbf{0}), \mathbf{I})$. □

10 Factor Analysis and Probabilistic PCA

Theorem 10.1. *Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ be independent, then*

$$\mathbf{z} = \mathbf{x} + \mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2).$$

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

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

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

MLE for Probabilistic PCA The maximum likelihood estimators of $\boldsymbol{\mu}$ is $\bar{\mathbf{t}}$, which can be observed by following MLE for normal distribution. By omitting the constant term, we focus on minimizing

$$f = \ln\det(\mathbf{C}) + \mathrm{tr}(\mathbf{C}^{-1}\hat{\mathbf{\Sigma}})),$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$. The gradient of $\mathbf{W}$ is

$$\frac{\partial f}{\partial \mathbf{W}} = \mathbf{C}^{-1}\hat{\mathbf{\Sigma}}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}.$$

Let $\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^\top$ be condense SVD of $\mathbf{W}$, where $\mathbf{U} \in \mathbb{R}^{d\times q}$, $\mathbf{L} \in \mathbb{R}^{q\times q}$ and $\mathbf{V} \in \mathbb{R}^{q\times q}$. Denote $\mathbf{U}_1$ be the orthogonal complement of $\mathbf{U}$, then we have

$$\begin{aligned}\mathbf{C} =&\mathbf{U}\mathbf{L}^2\mathbf{U}^\top + \sigma^2\begin{bmatrix}\mathbf{U} & \mathbf{U}_1\end{bmatrix}\begin{bmatrix}\mathbf{U} \\ \mathbf{U}_1\end{bmatrix} \\ =&\begin{bmatrix}\mathbf{U}^\top & \mathbf{U}_1^\top\end{bmatrix}\begin{bmatrix}\mathbf{L}^2+\sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top \\ \mathbf{U}_1^\top\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\mathbf{C}^{-1} =&\begin{bmatrix}\mathbf{U}^\top & \mathbf{U}_1^\top\end{bmatrix}\begin{bmatrix}\mathbf{L}+\sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{U}^\top \\ \mathbf{U}_1^\top\end{bmatrix} \\ =&\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{U}^\top + \sigma^{-2}\mathbf{U}_1\mathbf{U}_1^\top.\end{aligned}$$

Taking the gradient be zero, we have

$$\mathbf{C}^{-1}\hat{\boldsymbol{\Sigma}}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W} = \mathbf{0} \Longleftrightarrow \hat{\boldsymbol{\Sigma}}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}.$$

The decomposing of $\mathbf{C}$ and $\mathbf{C}^{-1}$ implies

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}\mathbf{C}^{-1}\mathbf{W} =& \hat{\boldsymbol{\Sigma}}\left(\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{U}^\top + \sigma^{-2}\mathbf{U}_1\mathbf{U}_1^\top\right)\mathbf{U}\mathbf{L}\mathbf{V}^\top \\
=& \hat{\boldsymbol{\Sigma}}\left(\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{U}^\top\right)\mathbf{U}\mathbf{L}\mathbf{V}^\top \\
=& \hat{\boldsymbol{\Sigma}}\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{L}\mathbf{V}^\top.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
& \hat{\boldsymbol{\Sigma}}\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{L}\mathbf{V}^\top = \mathbf{U}\mathbf{L}\mathbf{V}^\top \\
\Longleftrightarrow & \hat{\boldsymbol{\Sigma}}\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1}\mathbf{L} = \mathbf{U}\mathbf{L} \\
\Longleftrightarrow & \hat{\boldsymbol{\Sigma}}\mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I})^{-1} = \mathbf{U} \\
\Longleftrightarrow & \hat{\boldsymbol{\Sigma}}\mathbf{U} = \mathbf{U}(\mathbf{L}^2+\sigma^2\mathbf{I}),
\end{aligned}$$

where $\mathbf{U}$ and the diagonal matrix $\mathbf{L}^2 + \sigma^2\mathbf{I}$ correspond to the eigenvalue decomposition of $\hat{\boldsymbol{\Sigma}}$. Hence, all potential solution of $\mathbf{W}$ has the form of

$$\mathbf{W} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{12}$$

where $\mathbf{U}_q$ contains q eigenvectors of $\hat{\boldsymbol{\Sigma}}$, $\boldsymbol{\Lambda}$ is diagonal matrix with corresponding eigenvalues and $\mathbf{R}^{q\times q}$ is any orthogonal matrix. Using the expression (12), we obtain

$$\begin{aligned}
\mathbf{C} =& \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} \\
=& \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{U}_q^\top + \sigma^2\mathbf{I} \\
=& \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{U}_q^\top + \sigma^2\left(\mathbf{U}_q\mathbf{U}_q^\top + \mathbf{U}_{d-q}\mathbf{U}_{d-q}^\top\right) \\
=& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{d-q}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{d-q}^\top\end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{C}^{-1}\hat{\boldsymbol{\Sigma}} =& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{d-q}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_q^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{d-q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{d-q}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{d-q}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{d-q}^\top\end{bmatrix} \\
=& \begin{bmatrix}\mathbf{U}_q & \mathbf{U}_{d-q}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\boldsymbol{\Lambda}_{d-q}\end{bmatrix}\begin{bmatrix}\mathbf{U}_q^\top \\ \mathbf{U}_{d-q}^\top\end{bmatrix}.
\end{aligned}$$

Therefore, the function f can be written as

$$f = \ln\det(\mathbf{C}) + \operatorname{tr}(\mathbf{C}^{-1}\hat{\boldsymbol{\Sigma}})) = (d-q)\ln\sigma^2 + \sum_{i=1}^{q}\ln\lambda_i + q + \frac{1}{\sigma^2}\sum_{j=q+1}^{d}\lambda_j.$$

Minimizing f over σ^2 achieves

$$\sigma^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j.$$

So we have

$$f =(d-q)\ln\left(\frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j\right) + \sum_{i=1}^{q}\ln\lambda_i + d$$

$$=(d-q)\ln\left(\frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j\right)-\sum_{j=q+1}^{d}\ln\lambda_i+\sum_{i=1}^{d}\ln\lambda_i+d.$$

Since $\sum_{i=1}^{d}\ln\lambda_i=\mathrm{tr}(\hat{\boldsymbol{\Sigma}})$ is fixed, we only need to select $\lambda_{q+1},\ldots,\lambda_d$ to minimize

$$\ln\left(\frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j\right)-\frac{1}{d-q}\sum_{j=q+1}^{d}\ln\lambda_i.$$

Suppose that $\lambda_{q+1}=\max_j\{\lambda_j\}_{j=q+1}^{d}$, then we have

$$\lambda_q\geq\frac{\lambda_{q+1}+\cdots+\lambda_d}{d-1-q}.$$

We introduce the following function to determine λ_q:

$$g(x)=\ln\left(\frac{1}{d-q}\left(x+\sum_{j=q+2}^{d}\lambda_j\right)\right)-\frac{1}{d-q}\left(\ln x+\sum_{j=q+2}^{d}\ln\lambda_i\right).$$

Then we have

$$g'(x)=\frac{1}{x+\sum_{j=q+1}^{d-1}\lambda_j}-\frac{1}{(d-q)x}\geq 0$$

when (x corresponds to $\lambda_{q+1}=\max_j\{\lambda_j\}_{j=q+1}^{d}$)

$$x\geq\frac{\lambda_{q+1}+\cdots+\lambda_d}{d-1-q},$$

which implies $g(x)$ is decreasing. Therefore, we should take $\lambda_{q+1},\ldots,\lambda_d$ as the smallest $d-q$ eigenvalues.

Optimality of MLE Estimator We show that the MLE estimator also minimize the Frobenius norm error

$$\left(\hat{\mathbf{W}},\hat{\sigma}^2\right)=\operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{R}^{d\times q},\sigma^2\in\mathbb{R}^+}\left\|\hat{\boldsymbol{\Sigma}}-\left(\mathbf{W}\mathbf{W}^\top+\sigma^2\mathbf{I}\right)\right\|_F.$$

The following lemma comes from page 215 of book "Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis, Vol. 2.* Cambridge University Press, 1991". We can find a proof in Appendix B of paper "Luo Luo, Cheng Chen, Zhihua Zhang, Wu-Jun Li, Tong Zhang. Robust Frequent Directions with Application in Online Learning. *Journal of Machine Learning Research*, 20(45):1-41, 2019."

Lemma 10.1. *Let* $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$ *and* $q=\min\{m,n\}$*. Define the diagonal matrix* $\mathbf{D}(\mathbf{A})$ *whose* (i,i)*-th element is the* i*-th singular value of* $\mathbf{A}$ *and the others are zero. We define* $\mathbf{D}(\mathbf{A})$*. Then we have*

$$\|\mathbf{A}-\mathbf{B}\|\geq\|\mathbf{D}(\mathbf{A})-\mathbf{D}(\mathbf{B})\|.$$

Based on above lemma, we have

$$\begin{aligned}&\left\|\hat{\boldsymbol{\Sigma}}-\left(\mathbf{W}\mathbf{W}^\top+\sigma^2\mathbf{I}\right)\right\|_F\\ \geq&\left\|\mathbf{D}(\hat{\boldsymbol{\Sigma}})-\mathbf{D}\left(\mathbf{W}\mathbf{W}^\top+\sigma^2\mathbf{I}\right)\right\|_F\\ =&\sum_{i=1}^{d}\left(\lambda_i-\lambda_i\left(\mathbf{W}\mathbf{W}^\top\right)-\sigma^2\right)^2\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{i=q+1}^{d} \left(\lambda_i - \lambda_i\big(\mathbf{W}\mathbf{W}^\top\big) - \sigma^2\right)^2 \\
&= \sum_{i=q+1}^{d} \left(\lambda_i - \sigma^2\right)^2 \\
&\geq \sum_{i=q+1}^{d} \left(\lambda_i - \hat{\sigma}^2\right)^2,
\end{aligned}
$$

where $\mathbf{W} = \hat{\mathbf{W}}$ and $\sigma^2 = \hat{\sigma}^2$ lead all equality hold.

The EM Algorithm for PPCA For the model

$$
\mathbf{t}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha,
$$

where $\mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I})$ are independent.

1. We consider $\{\mathbf{x}_\alpha\}_{\alpha=1}^N$ to be missing data and $\{\mathbf{x}_\alpha, \mathbf{t}_\alpha\}_{\alpha=1}^N$ to be the complete data.

2. The posterior of $\mathbf{x}$ given $\mathbf{t}$ is

$$
\begin{aligned}
& p(\mathbf{x} \,|\, \mathbf{t}) \\
\propto & p(\mathbf{t} \,|\, \mathbf{x}) p(\mathbf{x}) \\
= & n\big(\mathbf{t} \,|\, \mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\big)\, n(\mathbf{x} \,|\, \mathbf{0}, \mathbf{I}) \\
\propto & \exp\left(-\frac{\|\mathbf{t}_\alpha - \mathbf{W}\mathbf{x}_\alpha - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{x}_\alpha\|_2^2}{2}\right) \\
\propto & \exp\left(\frac{1}{2\sigma^2}\left(\mathbf{x}^\top\mathbf{W}\mathbf{W}^\top\mathbf{x} - 2(\mathbf{t} - \boldsymbol{\mu})^\top\mathbf{W}\mathbf{x} + \sigma^2\|\mathbf{x}\|_2^2\right)\right) \\
= & \exp\left(\frac{1}{2\sigma^2}\left(\mathbf{x}^\top\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\mathbf{x} - 2(\mathbf{t} - \boldsymbol{\mu})^\top\mathbf{W}\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\mathbf{x}\right)\right).
\end{aligned}
$$

   Hence, it is normal distribution such that

$$
\mathbf{x} \mid \mathbf{t} \sim \mathcal{N}\left(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right),
$$

   where $\mathbf{M} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

3. The joint density of $\{\mathbf{x}_\alpha, \mathbf{t}_\alpha\}_{\alpha=1}^N$ is

$$
\prod_{\alpha=1}^{N} n\big(\mathbf{t}_\alpha \,|\, \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu}, \sigma^2\mathbf{I}\big)\, n(\mathbf{x}_\alpha \,|\, \mathbf{0}, \mathbf{I}).
$$

In E-step, we take the expectation of the log-likelihood with respect to the distributions $p(\mathbf{x}_\alpha \,|\, \mathbf{t}_\alpha)$:

$$
\begin{aligned}
& l_C \\
= & \mathbb{E}\left[\ln\left(\prod_{\alpha=1}^{N} p(\mathbf{x}_\alpha \,|\, \mathbf{t}_\alpha)\right)\right] \\
= & -\sum_{\alpha=1}^{N}\left(\frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\mathbf{t}_\alpha - \boldsymbol{\mu})(\mathbf{t}_\alpha - \boldsymbol{\mu})^\top - \frac{1}{2\sigma^2}\langle\mathbf{x}_\alpha\rangle^\top\mathbf{W}^\top(\mathbf{t}_\alpha - \boldsymbol{\mu})^\top + \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{W}^\top\mathbf{W}\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\right) + \frac{\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle}{2}\right) + C.
\end{aligned}
$$

where $\langle\mathbf{x}_\alpha\rangle = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t}_\alpha - \boldsymbol{\mu})$ and $\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle = \sigma^2\mathbf{M}^{-1} + \langle\mathbf{x}_\alpha\rangle\langle\mathbf{x}_\alpha\rangle^\top$.

In the M-step, the expectation l_C is maximised with respect to $\mathbf{W}$ and σ^2 giving new parameter

$$
\begin{aligned}
\tilde{\mathbf{W}} &= \left(\sum_{\alpha=1}^{N}(\mathbf{t}_\alpha-\boldsymbol{\mu})\langle\mathbf{x}_\alpha\rangle^\top\right)\left(\sum_{\alpha=1}^{N}\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\right)^{-1}\\
&=\sum_{\alpha=1}^{N}\left((\mathbf{t}_\alpha-\boldsymbol{\mu})(\mathbf{t}_\alpha-\boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\right)\left(N\sigma^2\mathbf{M}^{-1}+\sum_{\alpha=1}^{N}\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t}_\alpha-\boldsymbol{\mu})(\mathbf{t}_\alpha-\boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\\
&=\left(N\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\right)\left(N\sigma^2\mathbf{M}^{-1}+N\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\\
&=\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\\
&=\hat{\boldsymbol{\Sigma}}\mathbf{W}\left(\sigma^2\mathbf{I}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\right)^{-1}
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{\sigma}^2 =&\frac{1}{Nd}\sum_{\alpha=1}^{N}\left(\|\mathbf{t}_\alpha-\boldsymbol{\mu}\|_2^2-2\langle\mathbf{x}_\alpha\rangle^\top\tilde{\mathbf{W}}^\top(\mathbf{t}_\alpha-\boldsymbol{\mu})+\operatorname{tr}\left(\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\top\rangle\tilde{\mathbf{W}}^\top\tilde{\mathbf{W}}\right)\right)\\
=&\frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}})-\sum_{\alpha=1}^{N}2\operatorname{tr}\big((\mathbf{t}_\alpha-\boldsymbol{\mu})(\mathbf{t}_\alpha-\boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)\\
&+\sum_{\alpha=1}^{N}\operatorname{tr}\Big(\big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t}_\alpha-\boldsymbol{\mu})(\mathbf{t}_\alpha-\boldsymbol{\mu})^\top\mathbf{W}\mathbf{M}^{-1}\big)\tilde{\mathbf{W}}^\top\tilde{\mathbf{W}}\big)\big)\\
=&\frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}})-2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)+\operatorname{tr}\Big(\big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\tilde{\mathbf{W}}^\top\tilde{\mathbf{W}}\big)\big)\\
=&\frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}})-2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)+\operatorname{tr}\Big(\tilde{\mathbf{W}}\big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\tilde{\mathbf{W}}^\top\big)\big)\\
=&\frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}})-2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)+\operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\big(\sigma^2\mathbf{I}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\big)^{-1}\big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\big)\tilde{\mathbf{W}}^\top\big)\big)\\
=&\frac{1}{d}\Big(\operatorname{tr}(\hat{\boldsymbol{\Sigma}})-2\operatorname{tr}\big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)+\operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\big)\big)\\
=&\frac{1}{d}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\right).
\end{aligned}
$$