

# Lecture Notes of Multivariate Statistics

## Lecture 01

Weizhong Zhang

School of Data Science, Fudan University

May 30, 2023

### 1 Review of Linear Algebra

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

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

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

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

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

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

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

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

Suppose for  $\mathbf{q}_i^\top \mathbf{q}_j = 0$  for all  $\mathbf{q}_i^\top \mathbf{q}_j = 0$  for all  $i = 1, \dots, n' - 1$  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

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

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

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

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

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

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

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

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

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

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

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

Let

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

We have

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

Let

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

Then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Theorem 1.6.** Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{b} \in \mathbb{R}^m$ , the solution of minimization problem

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

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

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

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

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

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

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

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

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

We also have

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

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

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

## 2 The Multivariate Normal Distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Transform of Variables** Let the density of  $X_1, \dots, X_p$  be  $f(x_1, \dots, x_p)$ . Consider the  $p$  real-valued functions  $\mathbf{u} : \mathbb{R}^p \rightarrow \mathbb{R}^p$  such that

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

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

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

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

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

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

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

and

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

where the Jacobin matrix is

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

A roughly proof for above results:

- If  $\mathbf{A} \in \mathbb{R}^{p \times p}$  and  $\mathcal{S} \subset \mathbb{R}^p$  is a measurable set, then  $m(\mathbf{A}\mathcal{S}) = |\det(\mathbf{A})| m(\mathcal{S})$ . Let  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$  where  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal and  $\mathbf{\Sigma}$  is diagonal with nonnegative entries. Multiplying by  $\mathbf{V}^\top$  doesn't change the measure of  $\mathcal{S}$ . Multiplying by  $\mathbf{\Sigma}$  scales along each axis, so the measure gets multiplied by  $|\det(\mathbf{\Sigma})| = |\det(\mathbf{A})|$ . Multiplying by  $\mathbf{U}$  doesn't change the measure.

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

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

- Consider notation  $\Omega$  such that

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

where  $x_1 \leq x'_1, x_2 \leq x'_2, \dots, x_p \leq x'_p$ . Then the notation  $\mathbf{u}(\Omega)$  in the integral should consider the order

$$\int_{\mathbf{u}(\Omega)} = \int_{\min\{u_1(x_1), u_1(x'_1)\}}^{\max\{u_1(x_1), u_1(x'_1)\}} \cdots \int_{\min\{u_p(x_p), u_p(x'_p)\}}^{\max\{u_p(x_p), u_p(x'_p)\}}$$

By using even tinier subsets  $\Omega_i$ , the approximation would be even better so we see by a limiting argument that we actually obtain (3). On the other hand, we have ( $f$  is density functions of  $\mathbf{x}$  on  $\Omega$ ;  $g$  is density function of  $\mathbf{y}$  on  $\mathbf{u}(\Omega)$ ;  $\mathbf{y} = \mathbf{u}(\mathbf{x})$  means  $\mathbf{x}$  and  $\mathbf{y} = \mathbf{u}(\mathbf{x})$  are one-to-one mapping).

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

Since it holds for any  $\Omega$ , then

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

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

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

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

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

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

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

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

*Proof.* We have

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

□

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

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

Directly consider the expectation and variance of  $\mathbf{x}$  is not easy, so we first consider the ones of  $\mathbf{y}$ . The relation  $\mathbf{y} = \mathbf{C}^{-1}(\mathbf{x} - \mathbf{b})$  means  $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b}$  and  $\mathbb{E}[\mathbf{x}] = \mathbf{C}\mathbb{E}[\mathbf{y}] + \mathbf{b}$ . The transformation implies the density function of  $\mathbf{y}$  is

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

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

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

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

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

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

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

We also have

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

where the last step is due to

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

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

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

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

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

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

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

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

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

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

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

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

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



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

where we use the fact

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

□

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

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

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

*Proof.* Let

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

such that

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

**Sufficiency (uncorrelated  $\implies$  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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

Since the transform is non-singular, we have

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

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

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

**Conditional Distribution by Schur Complement** Recall that

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

which directly means the inverse of covariance of Normal distribution.

**Theorem 2.4.** Let  $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then

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

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

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

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

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

where  $\mathbf{A} \in \mathbb{R}^{p \times r}$ ,  $r$  is the rank of  $\boldsymbol{\Sigma}$  and  $\mathbf{y} \sim \mathcal{N}_r(\boldsymbol{\nu}, \mathbf{T})$  with non-singular  $\mathbf{T} \succ \mathbf{0}$ . We can write

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

*Proof.* We denote

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

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

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

we have

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

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

**Remark 2.1.** Observe that

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

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

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

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

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

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

Using Theorem 2.5, we have

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

□

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

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

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

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

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

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

and

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

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

□

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

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

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

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

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

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

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

□

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

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

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

**Theorem 2.9.** *If every linear combination of the components of a random vector  $\mathbf{y}$  is normally distributed, then  $\mathbf{y}$  is normally distributed.*

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

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

Set  $t = 1$ , then we have

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

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

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

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

**Theorem 2.10.** *Let  $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ ,  $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$  and  $\mathbf{z} = \mathbf{x} + \mathbf{y}$ . Suppose that  $\mathbf{x}$  and  $\mathbf{y}$  are independent. Prove  $\mathbf{z} \sim \mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$ .*

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

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

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

### 3 Estimation of the Mean Vector and the Covariance

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

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

respectively.

*Proof.* The logarithm of the likelihood function is

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

We have

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

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

Now, we only need to study how to maximize

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

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

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

then

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

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

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

□

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

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

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



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

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

and

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

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

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

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

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

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

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

□

**Theorem 3.2.** Let  $f(\theta)$  be a real-valued function defined on a set  $\mathcal{S}$  and let  $\phi$  be a single-valued function, with a single-valued inverse, on  $\mathcal{S}$  to a set  $\mathcal{S}^*$ . Let

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

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

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

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

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

**Theorem 3.3.** If  $\phi : \mathcal{S} \rightarrow \mathcal{S}^*$  is not one-to-one, we let

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

and the induced likelihood function

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

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

*Proof.* The definition means

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

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

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

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

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

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

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

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

$\square$

**Theorem 3.4.** *Suppose  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are independent, where  $\mathbf{x}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma})$ . Let  $\mathbf{C} \in \mathbb{R}^{N \times N}$  be an orthogonal matrix, then*

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

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

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

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

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

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

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

where

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

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

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

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

**Lemma 3.2.** *If*

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

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

*Proof.* Let

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

We have

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

$\square$

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

**Theorem 3.5.** *Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be independent, each distributed according to  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ . Then the mean of the sample*

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

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

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

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

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

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

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

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

By Lemma 3.2, we have

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

Lemma 3.2 also states  $\mathbf{z}_N$  is independent of  $\mathbf{z}_1, \dots, \mathbf{z}_{N-1}$ , then the mean vector  $\bar{\mathbf{x}} = \frac{1}{\sqrt{N}} \mathbf{z}_N$  is independent of  $\mathbf{A}$  and  $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \mathbf{A}$ . Since  $\bar{\mathbf{x}} = \frac{1}{\sqrt{N}} \mathbf{z}_n = \frac{1}{\sqrt{N}} \sum_{\beta=1}^N b_{N\beta} \mathbf{x}_\beta$ , Theorem 3.4 implies

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

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

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

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

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

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

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

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

*Proof.* We have

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

and

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

Hence, we obtain

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

□

**Theorem 3.7.** Using the notation of Theorem 3.1, if  $N > p$ , the probability is 1 of drawing a sample so that

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

is positive definite.

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

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

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

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

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

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

The second equality is obtained as follows

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

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

**Theorem 3.8.** *If  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are independent observations from  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then*

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

where

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

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

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

where the last step is due to

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

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

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

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

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

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

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

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

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

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

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

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

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

**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  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .

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

*Proof.* We have

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

$\square$

**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  $\mathcal{C}[\mathbf{t}, \mathbf{s}] = \mathbf{I}$ .*



*Proof.* We have

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

where the last line holds since

$$\begin{aligned}
& \int (t_j - \theta_j) g(\mathbf{X}, \boldsymbol{\theta}) d\mathbf{X} \\
&= \int t_j g(\mathbf{X}, \boldsymbol{\theta}) d\mathbf{X} - \theta_j \int g(\mathbf{X}, \boldsymbol{\theta}) d\mathbf{X} \\
&= \mathbb{E} t_j - \theta_j \\
&= 0
\end{aligned}$$

and therefore

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

□

**Theorem 3.10.** *Under the regularity condition (everything is well-defined, integration and differentiation can be swapped), we have*

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

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

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

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

$$1 \geq \frac{\mathbf{a}^\top (\mathcal{C}[\mathbf{s}])^{-1} \mathbf{a}}{\sqrt{\mathbf{a}^\top \mathcal{C}[\mathbf{t}] \mathbf{a} \mathbf{a}^\top (\mathcal{C}[\mathbf{s}])^{-1} \mathbf{a}}}$$

which means

$$\mathbf{a}^\top \mathcal{C}[\mathbf{t}] \mathbf{a} \geq \mathbf{a}^\top (\mathcal{C}[\mathbf{s}])^{-1} \mathbf{a}$$

for any nonzero  $\mathbf{a}$ . Hence, we have

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

□

**Theorem 3.11.** Let  $p$ -component vectors  $\mathbf{y}_1, \mathbf{y}_2, \dots$  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  $n \rightarrow +\infty$  is  $\mathcal{N}(\mathbf{0}, \mathbf{T})$ .

*Proof.* Let

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

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

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

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

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

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.  $\square$

**Theorem 3.12.** If  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are independently distributed, each  $x_\alpha$  according to  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , and if  $\boldsymbol{\mu}$  has an a prior distribution  $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$ , 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  $\boldsymbol{\mu} + \mathbf{y}$ , where

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

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

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

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

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

then

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

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

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

□

**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

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

**Lemma 3.5.** If  $f(x)$  is a function such that

$$f(b) - f(a) = \int_a^b f'(x) \, dx$$

for all  $a < b$  and if

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

then

$$\int_{-\infty}^{+\infty} f(x)(x - \theta) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2}(x - \theta)^2 \right) \, dx = \int_{-\infty}^{+\infty} f'(x) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2}(x - \theta)^2 \right) \, dx. \quad (5)$$

*Proof.* Since  $(x - \theta) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2}(x - \theta)^2 \right)$  is odd function, the LHS of (5) can be written as

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

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

where we use

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

and

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

□

**Lemma 3.6.** *Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are independently distributed to  $\mathcal{N}_p(\boldsymbol{\mu}, N\mathbf{I})$ , we have*

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

*Proof.* We have

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

□

**Theorem 3.13.** *Under the setting of Lemma 3.6, we let*

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

*and  $p > 3$ . Then  $\mathbb{E} \left[ \|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \right] < \mathbb{E} \left[ \|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2 \right]$ .*

*Proof.* We have

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

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

Using Lemma 3.5 with  $\theta = \mu_i$ ,

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

Hence, we obtain

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

□

**Remark 3.4.** We consider the bias and variance decomposition

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

Unbiased estimator may leads to larger variance.

**Lemma 3.7.** Suppose that  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$ , then

$$\mathbb{E} \|g^+(\|\mathbf{x}\|_2) \mathbf{x} - \boldsymbol{\mu}\|_2^2 \leq \mathbb{E} \|g(\|\mathbf{x}\|_2) \mathbf{x} - \boldsymbol{\mu}\|_2^2,$$

where

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

for any function  $g(u)$ .

*Proof.* We have

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

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

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

which means

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

Let  $\mathbf{y} = \mathbf{P}^\top \mathbf{x}$ , then we have  $\boldsymbol{\mu}^\top \mathbf{x} = \boldsymbol{\mu}^\top \mathbf{P}\mathbf{y} = (\mathbf{P}^\top \boldsymbol{\mu})^\top \mathbf{y} = \|\boldsymbol{\mu}\|_2 y_1$  and

$$\begin{aligned}
& \mathbb{E} [\boldsymbol{\mu}^\top \mathbf{x} (g^+(\|\mathbf{x}\|_2) - g(\|\mathbf{x}\|_2))] \\
&= \mathbb{E} [\|\boldsymbol{\mu}\|_2 y_1 (g^+(\|\mathbf{y}\|_2) - g(\|\mathbf{y}\|_2))] \\
&= \|\boldsymbol{\mu}\|_2 \int_{-\infty}^{+\infty} y_1 (g^+(\|\mathbf{y}\|_2) - g(\|\mathbf{y}\|_2)) \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2} \left(\sum_{i=1}^p y_i^2 - 2y_1 \|\boldsymbol{\mu}\|_2 + \|\boldsymbol{\mu}\|_2^2\right)\right) \mathrm{d}\mathbf{y} \\
&= \frac{\|\boldsymbol{\mu}\|_2 \exp\left(-\frac{1}{2} \|\boldsymbol{\mu}\|_2^2\right)}{(2\pi)^{\frac{p}{2}}} \int_{-\infty}^{+\infty} y_1 (g^+(\|\mathbf{y}\|_2) - g(\|\mathbf{y}\|_2)) \exp\left(-\frac{1}{2} \sum_{i=1}^p y_i^2\right) \exp(y_1 \|\boldsymbol{\mu}\|_2) \mathrm{d}\mathbf{y} \\
&= \frac{\|\boldsymbol{\mu}\|_2 \exp\left(-\frac{1}{2} \|\boldsymbol{\mu}\|_2^2\right)}{(2\pi)^{\frac{p}{2}}} \\
&\quad \cdot \int_{-\infty}^{+\infty} \dots \int_0^{+\infty} y_1 (g^+(\|\mathbf{y}\|_2) - g(\|\mathbf{y}\|_2)) \exp\left(-\frac{1}{2} \sum_{i=1}^p y_i^2\right) (\exp(y_1 \|\boldsymbol{\mu}\|_2) - \exp(-y_1 \|\boldsymbol{\mu}\|_2)) \mathrm{d}y_1 \dots \mathrm{d}y_p,
\end{aligned}$$

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

**Theorem 3.14.** *Let*

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

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

*Proof.* Use Lemma 3.7 with  $g(u) = 1 - (p-2)/u$ ,  $\mathbf{x} = \bar{\mathbf{x}} - \boldsymbol{\nu}$  and replace  $\boldsymbol{\mu}$  by  $\boldsymbol{\mu} - \boldsymbol{\nu}$ .  $\square$

## 4 $T^2$ -Statistic

**Theorem 4.1.** *For  $y \sim \chi^2(n)$ , we have  $\mathbb{E}[y] = n$  and  $\text{Var}[y] = 2n$ .*

*Proof.* We can write

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

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

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

and

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

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

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

□

**Theorem 4.2.** *The density of  $y \sim \chi^2(n)$  is*

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

where

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

*Proof.* We first provide the following results:

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

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

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

$$\frac{2}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_1\right) \frac{d\sqrt{y_1}}{dy_1} = \frac{1}{\sqrt{2\pi y_1}} \exp\left(-\frac{1}{2}y_1\right).$$

3. For beta function

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

we have

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

Consider that

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

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  $\infty$  and the limits of integration for  $v$  are 0 to 1. Thus

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

4. If

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

then

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

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

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



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 cdf of  $y_n$  is

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

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

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

□

**Theorem 4.3.** *If the  $n$ -component vector  $\mathbf{y}$  is distributed according to  $\mathcal{N}(\boldsymbol{\nu}, \mathbf{T})$  with  $\mathbf{T} \succ \mathbf{0}$ , then*

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

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

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

Then we have

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

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  $z_1, \dots, z_n$  are independently normally distributed with means  $\lambda_1, \dots, \lambda_n$  respectively, and variances 1. By definition this is the noncentral  $\chi^2$ -distribution with noncentrality parameter  $\sum_{i=1}^n \lambda_i^2 = \boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu}$ . □

**Theorem 4.4.** *The probability density function (pdf) for the noncentral  $\chi^2$ -distribution is*

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

*Proof.*  $\chi_p^2(\tau^2)$  with  $\tau^2 = \sum_{i=1}^p \lambda_i^2$  can be constructed via  $\mathbf{y}^\top \mathbf{y}$  with  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\lambda}, \mathbf{I})$ .

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  $i = 1, \dots, p$ . 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  $\tau = \sqrt{\boldsymbol{\lambda}^\top \boldsymbol{\lambda}}$ . 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  $\chi^2$ -distribution with  $p - 1$  degrees of freedom, and  $z_1$  and  $w$  have as joint density

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

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):

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

The joint density of  $v$  and  $u = z_1/\sqrt{v}$  is ( $dz_1 = \sqrt{v} du$ )

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

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  $u = \sqrt{s}$  ( $du = \frac{\sqrt{s}}{2} ds$ ) to obtain

$$\begin{aligned} & \int_{-1}^1 (1 - u^2)^{\frac{p-3}{2}} u^{2\beta} du \\ &= 2 \int_0^1 (1 - u^2)^{\frac{p-3}{2}} u^{2\beta} du \\ &= \int_0^1 (1 - s)^{\frac{p-3}{2}} s^{\beta-\frac{1}{2}} ds \\ &= B\left(\frac{p-1}{2}, \beta + \frac{1}{2}\right) \\ &= \frac{\Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\beta + \frac{1}{2}\right)}{\Gamma\left(\frac{p}{2} + \beta\right)} \end{aligned}$$

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  $v > 0$ . □

**Theorem 4.5.** Define the likelihood ratio criterion as

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

where

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

then we have

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

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

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

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

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

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

Furthermore, we have

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

because of

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

Similarly, we also have

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

Thus the likelihood ratio criterion is

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

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

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

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

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

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 4.6.** *Suppose  $\mathbf{y}_1, \dots, \mathbf{y}_m$  are independent with  $\mathbf{y}_\alpha$  distributed according to  $\mathcal{N}(\Gamma \mathbf{w}_\alpha, \Phi)$ , 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  $\mathbf{u}_1, \dots, \mathbf{u}_{m-r}$  are independently distributed according to  $\mathcal{N}(\mathbf{0}, \Phi)$  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 4.7.** *Let  $T^2 = \mathbf{y}^\top \mathbf{S}^{-1} \mathbf{y}$ , where  $\mathbf{y}$  is distributed according to  $\mathcal{N}_p(\boldsymbol{\nu}, \boldsymbol{\Sigma})$  and  $n\mathbf{S}$  is independently distributed as  $\sum_{\alpha=1}^n \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$  with  $\mathbf{z}_1, \dots, \mathbf{z}_n$  independent, each with distribution  $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ . 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  $\boldsymbol{\nu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\nu}$ . If  $\boldsymbol{\nu} = \mathbf{0}$ , the distribution is central  $F$ .*

**Theorem 4.8.** Let  $T^2 = \mathbf{y}^\top \mathbf{S}^{-1} \mathbf{y}$ , where  $\mathbf{y}$  is distributed according to  $\mathcal{N}_p(\boldsymbol{\nu}, \boldsymbol{\Sigma})$  and  $n\mathbf{S}$  is independently distributed as  $\sum_{\alpha=1}^n \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$  with  $\mathbf{z}_1, \dots, \mathbf{z}_n$  independent, each with distribution  $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ . 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  $\boldsymbol{\nu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\nu}$ . If  $\boldsymbol{\nu} = \mathbf{0}$ , the distribution is central  $F$ .

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

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

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

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

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  $j = 2, \dots, p$ . Then

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

where  $b^{ij}$  is the  $(i, j)$ -th entry of  $\mathbf{B}^{-1}$ . Using Schur Complement, we have

$$\frac{1}{b^{11}} = b_{11} - \mathbf{b}_{(1)}^\top \mathbf{B}_{22}^{-1} \mathbf{b}_{(1)} \triangleq b_{11.2, \dots, p} \quad (6)$$

with

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

and

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

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

$$\mathbf{B} = \sum_{\alpha=1}^n \mathbf{Q} \mathbf{z}_\alpha^* (\mathbf{Q} \mathbf{z}_\alpha^*)^\top = \sum_{\alpha=1}^n \mathbf{v}_\alpha^* (\mathbf{v}_\alpha^*)^\top = \begin{bmatrix} \sum_{\alpha=1}^n (\mathbf{v}_{\alpha 1}^*)^2 & \sum_{\alpha=1}^n \mathbf{v}_{\alpha,1}^* (\mathbf{v}_{\alpha,2-p}^*)^\top \\ \sum_{\alpha=1}^n \mathbf{v}_{\alpha,1}^* (\mathbf{v}_{\alpha,2-p}^*) & \sum_{\alpha=1}^n (\mathbf{v}_{\alpha,2-p}^*) (\mathbf{v}_{\alpha,2-p}^*)^\top \end{bmatrix} = \begin{bmatrix} b_{11} & \mathbf{b}_{(1)}^\top \\ \mathbf{b}_{(1)} & \mathbf{B}_{22} \end{bmatrix},$$

where  $\mathbf{v}_\alpha = \mathbf{Q} \mathbf{z}_\alpha^*$  are independent, each with distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  since  $\mathbf{Q} \mathbf{D} \Sigma \mathbf{D}^\top \mathbf{Q}^\top = \mathbf{I}$ . We denote

$$\mathbf{G} = b_{(1)}^\top \mathbf{B}_{22}^{-1} = \sum_{\alpha=1}^m \mathbf{v}_{\alpha,1}^* (\mathbf{v}_{\alpha,2-p}^*)^\top \mathbf{B}_{22}^{-1}$$

By Theorem 4.6, the random variable

$$\begin{aligned} b_{11.2, \dots, p} &= b_{11} - \left( b_{(1)}^\top \mathbf{B}_{22}^{-1} \right) \mathbf{B}_{22} \mathbf{B}_{22}^{-1} b_{(1)} \\ &= \sum_{\alpha=1}^n (\mathbf{v}_{\alpha 1}^*)^2 - \mathbf{G} \mathbf{B}_{22}^{-1} \mathbf{G}^\top \end{aligned}$$

is conditionally distributed as

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

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

**Remark 4.2.** The equation (6) is based on the fact

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B} \mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1} \mathbf{C} & \mathbf{I} \end{bmatrix} \quad (7)$$

and

$$\begin{aligned} \mathbf{M}^{-1} &= \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \left( \begin{bmatrix} \mathbf{I} & \mathbf{B} \mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1} \mathbf{C} & \mathbf{I} \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1} \mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B} \mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} \mathbf{B} \mathbf{D}^{-1} \\ -\mathbf{D}^{-1} \mathbf{C} (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{C} (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} \mathbf{B} \mathbf{D}^{-1} \end{bmatrix}. \end{aligned}$$

**Theorem 4.9.** Let  $u$  be distributed according to the  $\chi^2$ -distribution with  $a$  degrees of freedom and  $w$  be distributed according to the  $\chi^2$ -distribution with  $b$  degrees of freedom. The density of  $v = u/(u + w)$ , when  $u$  and  $w$  are independent is

$$\frac{1}{B\left(\frac{a}{2}, \frac{b}{2}\right)} v^{\frac{a}{2}-1} (1-v)^{\frac{b}{2}-1}, \quad (8)$$

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

*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  $u$  and  $w$  is

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

and the joint density of  $v$  and  $z$  is

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

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

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

Hence,

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

□

**Remark 4.3.** Beta distribution is a conjugate prior the binomial random variable. The binomial random variable  $X$  with parameters  $n$  and  $\theta$  has the probability mass function

$$f(X = k | n, \theta) = C_n^k \theta^k (1-\theta)^{n-k}.$$

Let  $\theta$  follows Beta distribution (prior distribution) with parameters  $a$  and  $b$  whose density function is

$$g(\theta | a, b) = \frac{1}{B(a, b)} v^{a-1} (1-v)^{b-1}.$$

Then we can write the density for the posterior distribution of  $\theta$  by Bayes rule

$$P(\theta | X = k) = \frac{P(X = k | \theta) P(\theta)}{P(X = k)}$$

$$\begin{aligned}
&= \frac{C_n^k \theta^k (1-\theta)^{n-k} \cdot \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}}{P(X=k)} \\
&= \frac{C_n^k}{P(X=k)B(a,b)} \theta^{k+a-1} (1-\theta)^{n-k+b-1}.
\end{aligned}$$

Since  $C_n^k/(P(X=k)B(a,b))$  is independent on  $\theta$ , it follows Beta distribution with parameters  $k+a$  and  $n-k+b$  is density.

**Theorem 4.10.** Let  $x_1, x_2, \dots$  be a sequence of independently identically distributed random vectors with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . Let

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

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  $N \rightarrow \infty$  is the  $\chi^2$ -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 stochastically 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  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ . The theorem follows from Theorem 4.3.  $\square$

**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  $\lambda_1$  of  $\det(\mathbf{v}\mathbf{v}^\top - \lambda \mathbf{B}) = 0$  associate with vector  $\boldsymbol{\beta} \neq \mathbf{0}$  satisfying

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

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 4.4.** 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  $\det(N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top - \lambda \mathbf{A}) = 0$ .  $\square$

**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.  $\square$



**Theorem 4.11.** 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

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

and

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

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

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

and

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

On the other hand, we can write

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

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. □

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

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

then

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

*Proof.* The definitions mean

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

then

$$\begin{aligned} \text{Cov}(\mathbf{z}_1, \mathbf{z}_2) &= \begin{bmatrix} a_1 \mathbf{I} & a_2 \mathbf{I} & \dots & a_N \mathbf{I} \end{bmatrix} \text{Cov} \left( \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_N \end{bmatrix}, \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_N \end{bmatrix} \right) \begin{bmatrix} b_1 \mathbf{I} \\ b_2 \mathbf{I} \\ \vdots \\ b_N \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} a_1 \mathbf{I} & a_2 \mathbf{I} & \dots & a_N \mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_N \end{bmatrix} \begin{bmatrix} b_1 \mathbf{I} \\ b_2 \mathbf{I} \\ \vdots \\ b_N \mathbf{I} \end{bmatrix} \\ &= \sum_{\alpha=1}^N a_\alpha b_\alpha \Sigma_\alpha. \end{aligned}$$

□

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

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

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

$$\text{Cov}(\mathbf{y}_\alpha, \mathbf{y}_{\alpha'}) = \begin{cases} \Sigma_1 + \frac{N_1}{N_2} \Sigma_2, & \alpha = \alpha', \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

*Proof.* We have

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

We first consider the case of  $\alpha = \alpha'$ . The independence means the covariance matrix of  $[\mathbf{x}_\alpha^{(1)}; \mathbf{z}_\alpha]^\top$  has the form of

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

where

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

Hence, we only need to focus on the covariance matrix of

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

Lemma 4.4 means

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

which means  $\text{Cov}(\mathbf{y}_\alpha, \mathbf{y}_\alpha) = \mathbf{\Sigma}_1 + (N_1/N_2) \mathbf{\Sigma}_2$ .

Then we consider the case of  $\alpha \neq \alpha'$ . We have

$$\begin{aligned} &\mathbf{y}_\alpha - \mathbb{E}[\mathbf{y}_\alpha] \\ &= \mathbf{x}_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}} \mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1 N_2}} \sum_{\beta=1}^{N_1} \mathbf{x}_\beta^{(2)} - \frac{1}{N_2} \sum_{\gamma=1}^{N_2} \mathbf{x}_\gamma^{(2)} - (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \\ &= \mathbf{x}_\alpha^{(1)} - \boldsymbol{\mu}^{(1)} - \sqrt{\frac{N_1}{N_2}} (\mathbf{x}_\alpha^{(2)} - \boldsymbol{\mu}^{(2)}) + \frac{1}{\sqrt{N_1 N_2}} \sum_{\beta=1}^{N_1} (\mathbf{x}_\beta^{(2)} - \boldsymbol{\mu}^{(2)}) - \frac{1}{N_2} \sum_{\gamma=1}^{N_2} (\mathbf{x}_\gamma^{(2)} - \boldsymbol{\mu}^{(2)}) \\ &= \mathbf{x}_\alpha^{(1)} - \boldsymbol{\mu}^{(1)} - \sqrt{\frac{N_1}{N_2}} (\mathbf{x}_\alpha^{(2)} - \boldsymbol{\mu}^{(2)}) + \left( \frac{1}{\sqrt{N_1 N_2}} - \frac{1}{N_2} \right) \sum_{\beta=1}^{N_1} (\mathbf{x}_\beta^{(2)} - \boldsymbol{\mu}^{(2)}) - \frac{1}{N_2} \sum_{\gamma=N_1+1}^{N_2} (\mathbf{x}_\gamma^{(2)} - \boldsymbol{\mu}^{(2)}) \end{aligned}$$

and

$$\begin{aligned} &\mathbf{y}_{\alpha'} - \mathbb{E}[\mathbf{y}_{\alpha'}] \\ &= \mathbf{x}_{\alpha'}^{(1)} - \boldsymbol{\mu}^{(1)} - \sqrt{\frac{N_1}{N_2}} (\mathbf{x}_{\alpha'}^{(2)} - \boldsymbol{\mu}^{(2)}) + \left( \frac{1}{\sqrt{N_1 N_2}} - \frac{1}{N_2} \right) \sum_{\beta=1}^{N_1} (\mathbf{x}_\beta^{(2)} - \boldsymbol{\mu}^{(2)}) - \frac{1}{N_2} \sum_{\gamma=N_1+1}^{N_2} (\mathbf{x}_\gamma^{(2)} - \boldsymbol{\mu}^{(2)}). \end{aligned}$$

The independence means

$$\begin{aligned}
& \mathbb{E} \left[ (\mathbf{y}_\alpha - \mathbb{E}[\mathbf{y}_\alpha]) (\mathbf{y}_{\alpha'} - \mathbb{E}[\mathbf{y}_{\alpha'}])^\top \right] \\
&= -2\sqrt{\frac{N_1}{N_2}} \left( \frac{1}{\sqrt{N_1 N_2}} - \frac{1}{N_2} \right) \mathbf{\Sigma}_2 + \left( \frac{1}{\sqrt{N_1 N_2}} - \frac{1}{N_2} \right)^2 N_1 \mathbf{\Sigma}_2 + \frac{N_2 - N_1}{N_2^2} \mathbf{\Sigma}_2 \\
&= \left( -2 \left( \frac{1}{N_2} - \frac{\sqrt{N_1}}{N_2 \sqrt{N_2}} \right) + \left( \frac{1}{N_1 N_2} - \frac{2}{N_2 \sqrt{N_1 N_2}} + \frac{1}{N_2^2} \right) N_1 + \frac{1}{N_2} - \frac{N_1}{N_2^2} \right) \mathbf{\Sigma}_2 \\
&= \mathbf{0}.
\end{aligned}$$

□

## 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  $f(\mathbf{y}_\alpha^{(2)} | \mathbf{y}_\alpha^{(1)})$  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  $\mathbf{y}_1^{(2)}, \dots, \mathbf{y}_N^{(2)}$  are independent and the density of  $\mathbf{y}_\alpha^{(2)}$  is  $f(\mathbf{y}_\alpha^{(2)} | \mathbf{y}_\alpha^{(1)})$ .*

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

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

where  $f_1(\mathbf{y}_\alpha^{(1)})$  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

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

□

**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*

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

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

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

*Proof.* The conditional distribution of  $z_{2\alpha}$  given  $z_{1\alpha}$  is  $\mathcal{N}(\beta z_{1\alpha}, \sigma^2)$ . Let  $\mathbf{v}_i = [z_{i1}, \dots, z_{in}]^\top$  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  $z_{21}, \dots, z_{2n}$  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

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

Apply Theorem 3.4 with  $x_\alpha = z_{2\alpha}$  and  $y_\alpha = \sum_{\gamma=1}^n c_{\alpha\gamma} z_{2\gamma}$  for  $\alpha = 1, \dots, n$ , where the first row of orthogonal matrix  $\mathbf{C}$  is  $(1/c) \mathbf{v}_1^\top$ . Then  $y_1, \dots, y_n$  are independently normally distributed with variance  $\sigma^2$  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  $\alpha = 2, \dots, n$ , the random variable  $u/\sigma^2$  has a  $\chi^2$ -distribution with  $n-1$  degrees of freedom.  $\square$

**Theorem 5.2.** *If  $x$  and  $y$  are independently distributed,  $x$  having the distribution  $\mathcal{N}(0, 1)$  and  $y$  having the  $\chi^2$ -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

$$\begin{aligned} f_{t,y}(t, y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2 y}{2m}\right) \cdot \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} \exp\left(-\frac{y}{2}\right) \cdot \frac{dt \sqrt{y/m}}{dt} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2 y}{2m}\right) \cdot \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} \exp\left(-\frac{y}{2}\right) \cdot \left(\frac{y}{m}\right)^{\frac{1}{2}} \\ &= \frac{1}{2^{\frac{m+1}{2}} \sqrt{m\pi} \Gamma(\frac{m}{2})} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right) y\right) \cdot y^{\frac{m-1}{2}}. \end{aligned} \tag{9}$$

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

$$\begin{aligned}\Gamma(\alpha) &= \int_0^{+\infty} \tilde{t}^{\alpha-1} \exp(-\tilde{t}) d\tilde{t} \\ &= \int_0^{+\infty} \left(\frac{t^2}{2m} + \frac{1}{2}\right)^{\alpha-1} y^{\alpha-1} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right)y\right) \left(\frac{t^2}{2m} + \frac{1}{2}\right) dy \\ &= \left(\frac{t^2}{2m} + \frac{1}{2}\right)^{\alpha} \int_0^{+\infty} y^{\alpha-1} \exp\left(-\left(\frac{t^2}{2m} + \frac{1}{2}\right)y\right) dy\end{aligned}\tag{10}$$

where we use the substitution

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

Connecting (9) and (10) with  $\alpha = \frac{m+1}{2}$ , we have

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

□

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

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

The set  $\Omega$  consists of  $\mu_1, \mu_2, \sigma_1, \sigma_2$  and  $\rho$  such that

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

and the set  $\omega$  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

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

*Proof.* We have shown in the section of  $T^2$ -statistic that the likelihood maximized in  $\Omega$  is

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

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

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

Let  $\sigma^2 = \sigma_1\sigma_2$  and  $\tau = \sigma_1/\sigma_2$ . Under the null hypothesis ( $\rho = \rho_0$ ), we have

$$\det(\mathbf{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_1^2\sigma_1^2\rho_0^2 = \sigma^4(1 - \rho_0^2), \quad \mathbf{\Sigma}^{-1} = \frac{1}{1 - \rho_0^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho_0}{\sigma_1\sigma_2} \\ -\frac{\rho_0}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix}$$

and

$$\begin{aligned} & \sum_{\alpha=1}^N (\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top \mathbf{\Sigma}^{-1} (\mathbf{x}_\alpha - \bar{\mathbf{x}}) = \text{tr}(\mathbf{\Sigma}^{-1} \mathbf{A}) \\ &= \frac{1}{1 - \rho_0^2} \left( \frac{a_{11}}{\sigma_1^2} - \frac{2\rho_0 a_{12}}{\sigma_1\sigma_2} + \frac{a_{22}}{\sigma_2^2} \right) \\ &= \frac{1}{(1 - \rho_0^2)\sigma^2} \left( \frac{a_{11}}{\tau} - 2\rho_0 a_{12} + \tau a_{22} \right). \end{aligned}$$

Then the likelihood function under the null hypothesis ( $\rho = \rho_0$ ) is

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

The maximum of (11) with respect to  $\tau$  occurs at

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

then the concentrated likelihood is

$$\frac{1}{(2\pi)^N (1 - \rho_0^2)^{\frac{N}{2}} (\sigma^2)^N} \exp \left( -\frac{\sqrt{a_{11}} \sqrt{a_{22}} (1 - \rho_0 r)}{\sigma^2(1 - \rho_0^2)} \right). \quad (12)$$

The maximum of (12) occurs at

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

which is because of  $f(x) = \exp(-b/x)/x^N$  leads to

$$f'(x) = \frac{\exp(-\frac{b}{x}) \cdot \frac{b}{x^2} \cdot x^N - \exp(-\frac{b}{x}) \cdot Nx^{N-1}}{x^{2N}} = \frac{\exp(-\frac{b}{x}) x^{N-2} (b - Nx)}{x^{2N}}.$$

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

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

where

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

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

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

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

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

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

*Proof.* We have

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

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

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

and

$$\mathbb{E}[(z_{i\alpha}z_{j\alpha} - \sigma_{ij})(z_{k\alpha}z_{l\alpha} - \sigma_{kl})] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \quad (13)$$

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

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



we obtain

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

Since  $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\mu}$  and covariance of  $\mathbf{y}_\alpha$  satisfies (13), 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

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

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  $\mathbf{x}_1, \dots, \mathbf{x}_N$  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.  $\square$

## 6 The Wishart Distribution

**Theorem 6.1.** Let  $\mathbf{z}_1, \dots, \mathbf{z}_n$  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  $i = 1, \dots, p$ . Then the density of  $\mathbf{T}^*$  is

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

*Proof.* Let  $\mathbf{C}$  be the lower triangular matrix ( $c_{ij} = 0, i < j$ ) such that  $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$  and  $c_{ii} > 0$ . Define  $\mathbf{y}_\alpha = \mathbf{C}^{-1}\mathbf{z}_\alpha$  for  $\alpha = 1, \dots, n$ , 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  $t_{i1}, \dots, t_{ii-1}$  are independently distributed and  $t_{ij}$  is distributed according to  $\mathcal{N}(0, 1)$  for  $i > j$ ; and  $t_{ii}$  has the  $\chi^2$ -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  $dw/dt_{ii} = 2t_{ii}$ )

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

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  $\mathbf{T}^* = \mathbf{C}\mathbf{T}$  can be written as

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

Since the matrix of the transformation is triangular, its determinant is the product of the diagonal elements, namely,  $\prod_{i=1}^p c_{ii}$ . 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

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

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

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

□

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

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

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

*Proof.* Following the proof of Theorem 6.1, we only need 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

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

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

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

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 density of  $\mathbf{A}$  is

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

□

**Corollary 6.1.** Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be independently distributed, each according to  $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 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  $\mathcal{W}(\frac{1}{n}\boldsymbol{\Sigma}, n)$ .

*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  $\mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{\Sigma})$ . Theorem 6.2 implies this corollary.  $\square$

**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  $\mathbf{A} = \mathbf{U}_\mathbf{A} \boldsymbol{\Sigma}_\mathbf{A} \mathbf{U}_\mathbf{A}^\top$  and  $\mathbf{E} = \mathbf{U}_\mathbf{A} \boldsymbol{\Sigma}_\mathbf{A}^{-\frac{1}{2}}$ , 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  $\mathbf{B}^* = \mathbf{U}_{\mathbf{B}^*} \boldsymbol{\Sigma}_{\mathbf{B}^*} \mathbf{U}_{\mathbf{B}^*}^\top$ , then

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

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

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

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

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

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

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

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

For the last step, we consider Taylor expansion on  $f(x) = (1 - x)^{-\frac{n}{2}}$  at  $x = 0$ , that is

$$f(x) = \sum_{k=0}^{\infty} \frac{f'(0)x^k}{k!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} \prod_{j=0}^{k-1} \left(j + \frac{n}{2}\right).$$

We take  $x = 2it$ .  $\square$

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

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

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

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

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

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

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

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

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

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

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

We also have

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

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

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

and

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

□

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

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

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

*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  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ . 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  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{11})$ , and  $\mathbf{A}_{11}$  is distributed as

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

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

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

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

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

*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  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ . 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  $\boldsymbol{\Sigma}$  be portioned. Since  $\boldsymbol{\Sigma}_{ij} = \mathbf{0}$ , the sets  $\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_n^{(1)}, \dots, \mathbf{z}_1^{(q)}, \dots, \mathbf{z}_n^{(q)}$  are independent. Then  $\mathbf{A}_{11} = \sum_{\alpha=1}^n \mathbf{z}_\alpha^{(1)} (\mathbf{z}_\alpha^{(1)})^\top, \dots, \mathbf{A}_{qq} = \sum_{\alpha=1}^n \mathbf{z}_\alpha^{(q)} (\mathbf{z}_\alpha^{(q)})^\top$  are independent. The rest of the proof follows from Theorem 6.4. □

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

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

then the density of the sample correlation coefficients is given by

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

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

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

to

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

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

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

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

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

Combing all above results proves this theorem.  $\square$

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

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

$$\begin{aligned} f(\mathbf{A}, \Sigma) &= \frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}\mathbf{A})\right)}{2^{\frac{np}{2}} (\det(\Sigma))^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{(\det(\Psi))^{\frac{m}{2}} (\det(\Sigma))^{-\frac{m+p+1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})\right)}{2^{\frac{mp}{2}} \Gamma_p\left(\frac{m}{2}\right)} \\ &= \frac{(\det(\Psi))^{\frac{m}{2}} (\det(\Sigma))^{-\frac{n+m+p+1}{2}} (\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\text{tr}((\mathbf{A} + \Psi)\Sigma^{-1})\right)}{2^{\frac{(m+n)p}{2}} \Gamma_p\left(\frac{n}{2}\right) \Gamma_p\left(\frac{m}{2}\right)} \end{aligned} \quad (14)$$

for  $\mathbf{A}$  and  $\Sigma$  are positive definite. The marginal density of  $\mathbf{A}$  is the integral of (14) over the set of  $\Sigma$  positive definite. Since

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

we have

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

Then

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

□

## 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  $\text{tr}(\mathbf{F}\mathbf{A}\mathbf{F}^\top \mathbf{G}) > 0$  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

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

□

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

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



The maximum likelihood estimator of  $\Sigma$  is give by

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

*Proof.* The likelihood function is

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

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

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

and

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

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

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

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

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

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

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

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

Then we have

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

Lemma 7.1 means

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

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  $\mathbf{G} = \boldsymbol{\Sigma}$  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  $\boldsymbol{\Sigma}$  is  $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\mathbf{D}$ . □

**Remark 7.1.** *Let*

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

*We consider the least square problem.*

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

*Then, taking the gradient of  $f$  be zero means*

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

*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*

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

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

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

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

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

**Theorem 7.2.** The maximum likelihood estimator  $\mathbf{B}$  based on a set of  $N$  observations, the  $\alpha$ -th from  $\mathcal{N}(\mathbf{B}\mathbf{z}_\alpha, \Sigma)$ , 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  $\hat{\Sigma}$  multiplied by  $N$  is independently distributed according to  $\mathcal{W}(\Sigma, N - q)$ , 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

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

From Theorem 4.6, it follows that

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

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

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  $\beta_{ig}$ , 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

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

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

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

$$\begin{aligned}
&= \sigma_{ii} a^{gg} - \sigma_{ii} \sum_{h=1}^q \sum_{h'=1}^q a_{hh'} a^{hg} a^{h'g} \\
&= \sigma_{ii} a^{gg} - \sigma_{ii} a^{gg} = 0.
\end{aligned}$$

Thus

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

□

**Theorem 7.4.** *The likelihood ratio criterion*

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

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  $\alpha = 1, \dots, N$  and non-singular  $\mathbf{D}$ .

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

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

then

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

□

**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  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ , the likelihood ratio criterion for testing the hypothesis*

$$\Sigma = \sigma_0^2 \Psi_0$$

where  $\Psi_0$  is specified and  $\sigma^2$  is not specified, is

$$\frac{(\det(\mathbf{A}\Psi_0^{-1}))^{\frac{N}{2}}}{(\text{tr}(\mathbf{A}\Psi_0^{-1})/p)^{\frac{pN}{2}}}.$$

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

$$\mathbf{C}\Psi_0\mathbf{C}^\top = \mathbf{I}.$$

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

$$\text{tr}(\mathbf{A}^*) = \text{tr} \left( \sum_{\alpha=1}^N (\mathbf{x}_\alpha^* - \bar{\mathbf{x}}_\alpha^*) (\mathbf{x}_\alpha^* - \bar{\mathbf{x}}_\alpha^*)^\top \right) = \text{tr}(\mathbf{C}\mathbf{A}\mathbf{C}^\top) = \text{tr}(\mathbf{A}\mathbf{C}^\top\mathbf{C}) = \text{tr}(\mathbf{A}\Psi_0^{-1})$$

and

$$\det(\mathbf{A}^*) = \det(\mathbf{C}\mathbf{A}\mathbf{C}^\top) = \det(\mathbf{C})^2 \det(\mathbf{A}) = (\det(\Psi_0))^{-1} \det(\mathbf{A}) = \det(\mathbf{A}\Psi_0^{-1}).$$

Thus

$$\frac{(\det(\mathbf{A}^*))^{\frac{N}{2}}}{(\text{tr}(\mathbf{A}^*)/p)^{\frac{pN}{2}}} = \frac{(\det(\mathbf{A}\Psi_0^{-1}))^{\frac{N}{2}}}{(\text{tr}(\mathbf{A}\Psi_0^{-1})/p)^{\frac{pN}{2}}}.$$

□