Lecture Notes of Multivariate Statistics

Weizhong Zhang

School of Data Science, Fudan University

March 12, 2023

1 Review of Linear Algebra

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

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

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

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

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

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

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

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

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

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

Then we prove the result by induction. □

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

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

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

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

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

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

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

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

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

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

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

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

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

Let

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

We have

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

Let

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

Then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We also have

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

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

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

2 The Multivariate Normal Distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

where the Jacobin matrix is

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

A roughly proof for above results:

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

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

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

- Consider notation Ω such that

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

where $x_1 \le x_1', x_2 \le x_2', \ldots, x_p \le x_p'$. Then the notation $\mathbf{u}(\Omega)$ in the integral should consider the order

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

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

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

Since it holds for any Ω, then

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

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

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

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

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

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

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

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

Proof. We have

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

□

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

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

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

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

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

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

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

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

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

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

We also have

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

where the last step is due to

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where we use the fact

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

□

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

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

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

Proof. Let

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

such that

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

Since the transform is non-singular, we have

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

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

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

Conditional Distribution by Schur Complement Recall that

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

which directly means the inverse of covariance of Normal distribution.

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

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

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

Proof. It is easy to verify $\mathbb{E}[\mathbf{z}] = \mathbf{D}\boldsymbol{\mu}$ and $\operatorname{Cov}[\mathbf{z}] = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top$. 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

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

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

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

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

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

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