Multivariate Distributions and Moments

The term *Probability* is used in our everyday life. In an experiment of tossing a fair coin for example, the probability it lands on head is 0.5. What does that mean? One explanation known as the Bayesian interpretation, it represents the probability as a measure of uncertainty about something [Mur12]. In other words, it is related to our information regarding the considered experiment. Different concepts and mathematical explanations regarding probabilities are presented in this chapter.

Random Vectors

Let $X_1, \ldots, X_n, n \in \mathbb{N}$ be random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

$$X_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{R}), \quad i = 1, \ldots, p$$

where $\mathcal{R}$ is the Borel σ-algebra generated by the open sets of $\mathbb{R}$.

- The vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ is called a random vector.
- Analogously, the matrix $\mathbf{X} = (X_{ij})_{\substack{1 \le i \le p, \\ 1 \le j \le n}}$, composed of the random variables X_{ij} as its elements, is called a random matrix.
- The joint distribution of a random vector is uniquely described by its multivariate distribution function:

  $$F(x_1, \ldots, x_p) = \mathbb{P}(X_1 \le x_1, \ldots, X_p \le x_p), \quad \text{where } (x_1, \ldots, x_p)^T \in \mathbb{R}^p,$$

  and x_i is a realization of X_i, with $i = 1, \ldots, p$.
- A random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ is called absolutely continuous if there exists an integrable function $f(x_1, \ldots, x_n) \ge 0$ such that:

  $$F(x_1, \ldots, x_p) = \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(x_1, \ldots, x_p) \mathrm{d}x_1 \ldots \mathrm{d}x_p.$$

  where f is the probability density function (pdf) and F is the cumulative distribution function (cdf).

Example. (Multivariate normal distribution) The multivariate normal (or Gaussian) distribution of the random vector $\mathbf{X} \in \mathbb{R}^p$ has the following pdf:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

where $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, and the parameters: $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, where $\boldsymbol{\Sigma} \succ 0$. This pdf can be denoted by $\mathbf{X} = (X_1, \ldots, X_p)^T \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that $\boldsymbol{\Sigma}$ must have full rank. There exists an $n-$dimensional Gaussian random variable, if $\mathrm{rk}(\boldsymbol{\Sigma}) < p$, however it has no density function with respect to $p-$ dimensional Lebesgue measure λ^p.

Expectation and Covariance

Definition 3.1. Given a random variable $\mathbf{X} = (X_1, \ldots, X_p)^T$.

(a) The expectation (vector) of $\mathbf{X}$, $\mathbb{E}(\mathbf{X})$, is defined by:

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))^T.$$

(b) The covariance matrix of $\mathbf{X}$, $\mathrm{Cov}(\mathbf{X})$, is defined by:

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T\right).$$

The expectation vector $\mathbb{E}(\mathbf{X})$ is constructed component-wise of $\mathbb{E}(X_i)$, where $i = 1, \ldots, p$. Furthermore, the covariance matrix has the covariance value $\mathrm{Cov}(X_i, X_j)$ as its (i,j)-th element given by

$$(\mathrm{Cov}(\mathbf{X}))_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\left((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right).$$

Theorem 3.2. *Given the random vectors* $\mathbf{X} = (X_1, \ldots, X_p)^T$, *and* $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$, *the following statements hold:*

(a) $\mathbb{E}(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{b}$

(b) $\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Y})$

(c) $\mathrm{Cov}(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\mathrm{Cov}(\mathbf{X})\mathbf{A}^T$

(d) $\mathrm{Cov}(\mathbf{X} + \mathbf{Y}) = \mathrm{Cov}(\mathbf{X}) + \mathrm{Cov}(\mathbf{Y})$, *if* $\mathbf{X}$ *and* $\mathbf{Y}$ *are stochastically independent.*

(e) $\mathrm{Cov}(\mathbf{X}) \succeq 0$, *i.e., the covariance matrix is non-negative definite.*

Proof. Prove (a)-(d) as exercise. Regarding e) assume that $\mathbf{a} \in \mathbb{R}^p$ be a vector, then

$$\mathbf{a}^T \mathrm{Cov}(\mathbf{X})\mathbf{a} \overset{(c)}{=} \mathrm{Cov}(\mathbf{a}^T\mathbf{X}) = \mathrm{Var}(\mathbf{a}^T\mathbf{X}) \geq 0.$$

□

Show as an exercise that if $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \text{and} \quad \mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}.$$

Theorem 3.3 (Steiner's rule). *Given a random vector* $\mathbf{X} = (X_1, \ldots, X_p)^T$*, it holds that*

$$\mathbb{E}\left((\mathbf{X}-\mathbf{b})(\mathbf{X}-\mathbf{b})^T\right) = \mathrm{Cov}(\mathbf{X}) + (\mathbf{b}-\mathbb{E}(\mathbf{X}))(\mathbf{b}-\mathbb{E}(\mathbf{X}))^T, \qquad \forall\, \mathbf{b} \in \mathbb{R}^p.$$

Proof. Let $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$. Note that

$$\mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{b}-\boldsymbol{\mu})^T\right) = \mathbb{E}\left(\mathbf{X}-\boldsymbol{\mu}\right)\left(\mathbf{b}-\boldsymbol{\mu}\right)^T = 0,$$

and $\mathbb{E}(a) = a, \forall a \in \mathbb{R}^p$, then

$$\begin{aligned}
\mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{b})(\mathbf{X}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{b})^T\right) &= \mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right) + \mathbb{E}\left((\boldsymbol{\mu}-\mathbf{b})(\boldsymbol{\mu}-\mathbf{b})^T\right) \\
&\quad + \mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu})(\boldsymbol{\mu}-\mathbf{b})^T\right) + \mathbb{E}\left((\boldsymbol{\mu}-\mathbf{b})(\mathbf{X}-\boldsymbol{\mu})^T\right) \\
&= \mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right) + \mathbb{E}\left((\boldsymbol{\mu}-\mathbf{b})(\boldsymbol{\mu}-\mathbf{b})^T\right) \\
&= \mathrm{Cov}(\mathbf{X}) + (\mathbf{b}-\mathbb{E}(\mathbf{X}))(\mathbf{b}-\mathbb{E}(\mathbf{X}))^T,
\end{aligned}$$

where we used the linearity of expectation to show that

$$\mathbb{E}\left((\mathbf{X}-\boldsymbol{\mu})(\boldsymbol{\mu}-\mathbf{b})^T\right) = \left((\mathbb{E}(\mathbf{X})-\boldsymbol{\mu})(\boldsymbol{\mu}-\mathbf{b})^T\right) = 0.$$

□

Theorem 3.4. *Let* $\mathbf{X}$ *be a random vector with* $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ *and* $\mathrm{Cov}(\mathbf{X}) = \mathbf{V}$*. Then*

$$\mathbb{P}(\mathbf{X} \in \mathrm{Im}(\mathbf{V}) + \boldsymbol{\mu}) = 1.$$

Proof. Let $\ker(\mathbf{V}) = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{V}\mathbf{x} = 0\}$ be the kernel (or null space) of $\mathbf{V}$. Assume a basis as $\ker(\mathbf{V}) = \langle \mathbf{a}_1, \ldots, \mathbf{a}_r \rangle$. For $i = 1, \ldots, r$, it holds that

$$a_i^T \mathbf{V} a_i = \mathrm{Var}(a_i^T \mathbf{V}) = 0.$$

Hence, $\mathbf{a}_i^T \mathbf{X}$ should be almost surely equal to its expectation, namely, $\mathbb{E}(\mathbf{a}_i^T \mathbf{X}) = \mathbf{a}_i^T \boldsymbol{\mu}$. In other words,

$$\mathbb{P}(\mathbf{a}_i^T \mathbf{X} = \mathbf{a}_i^T \boldsymbol{\mu}) = 1, \text{ i.e., } \mathbb{P}(\mathbf{a}_i^T (\mathbf{X}-\boldsymbol{\mu}) = 0) = 1,$$

and

$$\mathbb{P}(\mathbf{X}-\boldsymbol{\mu} \in \mathbf{a}_i^{\perp}) = 1.$$

Given the fact that for an arbitrary random variable Z and two closed sets A and B, the following expression is valid

$$\mathbb{P}(Z \in A) = \mathbb{P}(Z \in B) = 1 \implies \mathbb{P}(Z \in A \cap B) = 1, \tag{3.1}$$

it holds that

$$\mathbb{P}((\mathbf{X}-\boldsymbol{\mu}) \in \mathbf{a}_1^{\perp} \cap \cdots \cap \mathbf{a}_r^{\perp}) = 1.$$

However, given that $\mathrm{Im}(\mathbf{V}) = \ker(\mathbf{V})^{\perp} =< \mathbf{a}_1, \ldots, \mathbf{a}_r >^{\perp} = \mathbf{a}_1^{\perp} \cap \cdots \cap \mathbf{a}_r^{\perp}$. Therefore,

$$\mathbb{P}((\mathbf{X}-\boldsymbol{\mu}) \in \mathrm{Im}(\mathbf{V})) = 1.$$

Prove (3.1) as an exercise. □

Conditional Distribution

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a random vector such that $\mathbf{X} = (\mathbf{Y}_1, \mathbf{Y}_2)^T$ where $\mathbf{Y}_1 = (X_1, \ldots, X_k)$ and $\mathbf{Y}_2 = (X_{k+1}, \ldots, X_p)$. Suppose that $\mathbf{X}$ is absolutely continuous with density $f_{\mathbf{X}}$. Then the conditional density of $\mathbf{Y}_1$ given $\mathbf{Y}_2 = \mathbf{y}_2$ is denoted by

$$f_{\mathbf{Y}_1|\mathbf{Y}_2}(\mathbf{y}_1 \mid \mathbf{y}_2) = \frac{f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)}{f_{\mathbf{Y}_2}(\mathbf{y}_2)},$$

where $\mathbf{y}_1 \in \mathbb{R}^k$ is a realization of $\mathbf{Y_1}$. Furthermore, it also holds that

$$\mathbb{P}(\mathbf{Y}_1 \in B \mid \mathbf{Y}_2 = \mathbf{y}_2) = \int_B f_{\mathbf{Y}_1|\mathbf{Y}_2}(\mathbf{y}_1 \mid \mathbf{y}_2)\mathrm{d}\mathbf{y}_1, \quad \forall B \in \mathcal{R}^k.$$

Theorem 3.5 ([Mur12, Theorem 4.3.1]). *Suppose that* $(\mathbf{Y}_1, \mathbf{Y}_2) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where*

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \quad \textit{and} \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}.$$

Then

(a) The distribution of $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *are given by* $\mathbf{Y}_1 \sim N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ *and* $\mathbf{Y}_2 \sim N_{p-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, *respectively.*

(b) The conditional density $f_{\mathbf{Y}_1|\mathbf{Y}_2}(\mathbf{y}_1 \mid \mathbf{y}_2)$ *is given by the multivariate normal distribution* $f_{\mathbf{Y}_1|\mathbf{Y}_2}(\mathbf{y}_1 \mid \mathbf{y}_2) \sim N_k(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$. *The parameters* $\boldsymbol{\mu}_{1|2}$ *and* $\boldsymbol{\Sigma}_{1|2}$ *are defined as*

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2)), \end{aligned}$$

and

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}.$$

Note that $\boldsymbol{\Sigma}_{1|2}$ is the Schur complement, introduced in the previous chapter.

Maximum Likelihood Estimation

Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ is a random sample from a pdf $f(x; \boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ is a parameter vector. The function $L(\mathbf{x}; \boldsymbol{\vartheta})$ is referred to as the likelihood function, and defined as

$$L(\mathbf{x}; \boldsymbol{\vartheta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\vartheta}). \tag{3.2}$$

Furthermore, the function $\ell(\mathbf{x};\boldsymbol{\vartheta})$ represents the log-likelihood function, and is defined as

$$\ell(\mathbf{x};\boldsymbol{\vartheta}) = \log L(\mathbf{x};\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\vartheta}). \tag{3.3}$$

For a given sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, one can notice that both functions in (3.2) and (3.3) depend on the parameters in $\boldsymbol{\vartheta}$. Therefore, $\boldsymbol{\vartheta}$ is needed to be determined such that it fits the data in $\mathbf{x}$ through $L(\mathbf{x};\boldsymbol{\vartheta})$, or equivalently $\ell(\mathbf{x};\boldsymbol{\vartheta})$. The estimate of $\boldsymbol{\vartheta}$, denoted by $\hat{\boldsymbol{\vartheta}}$, is obtained as

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} \ell(\mathbf{x};\boldsymbol{\vartheta})$$

and called the maximum likelihood estimate (MLE) of $\boldsymbol{\vartheta}$.

Theorem 3.6. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n, n \in \mathbb{N}$, be i.i.d samples obtained from the distribution $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by*

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \overline{\mathbf{x}}, \quad \textit{and} \quad \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T.$$

Proof. In order to prove this theorem, Steiner's Lemma is required at this point along with the following fundamentals of matrix differentiation. For arbitrary matrices $\mathbf{V}, \mathbf{A}$ and vector $\mathbf{y}$, the following statements, which to be proved as exercise, are considered afterwards for simplification purposes.

- $\frac{\partial}{\partial \mathbf{V}} \log\det \mathbf{V} = (\mathbf{V}^{-1})^T$, if $\mathbf{V}$ is invertible.
- $\frac{\partial}{\partial \mathbf{V}} \mathrm{tr}(\mathbf{VA}) = \mathbf{A}^T$.
- $\frac{\partial(\mathbf{y}^T\mathbf{A}\mathbf{y})}{\partial \mathbf{y}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{y}$.

Starting with the log-likelihood function

$$\ell(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \left(\log \frac{1}{(2\pi)^{p/2}} + \log \frac{1}{|\,\boldsymbol{\Sigma}|^{1/2}} - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right),$$

in order to be maximized, the additive constants can be dropped, hence the log-likelihood function is defined by

$$\ell^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{p}{2} \log |\,\boldsymbol{\Sigma}^{-1}\,| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \tag{3.4}$$

By utilizing the previously presented fundamentals, and choosing $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda}$, (3.4) can be

reformulated as

$$\begin{aligned}\ell^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \frac{n}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \frac{n}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left(\boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right) \\ &= \frac{n}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right). \end{aligned} \tag{3.5}$$

Based on Steiner's rule, the summation term in (3.5) can be rewritten as

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T &= \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^T \\ &= n\mathbf{S}_n + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^T, \end{aligned} \tag{3.6}$$

hence $\ell^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is given by

$$\begin{aligned}\ell^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \frac{n}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}(n\mathbf{S}_n + (\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^T)\right) \\ &= \frac{n}{2}\log|\boldsymbol{\Lambda}| - \frac{n}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{S}_n\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^T\right). \end{aligned} \tag{3.7}$$

Based on (3.6), note that $n\mathbf{S}_n \preceq \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$, with equality when $\overline{\mathbf{x}} = \boldsymbol{\mu}$. Thus, $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$, and (3.7) is formulated as

$$\ell^*(\boldsymbol{\Lambda}) = \frac{n}{2}\log|\,\boldsymbol{\Lambda}\,| - \frac{n}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{S}_n\right). \tag{3.8}$$

Moreover, in order to find $\hat{\boldsymbol{\Lambda}}$, the derivative of log-likelihood function in (3.8) with respect to $\boldsymbol{\Lambda}$ is calculated to find its zeros, as

$$\frac{\partial\ell^*(\boldsymbol{\Lambda})}{\partial\boldsymbol{\Lambda}} = 0 \implies \frac{n}{2}\boldsymbol{\Lambda}^{-1} - \frac{n}{2}\mathbf{S}_n = 0,$$

hence $\hat{\boldsymbol{\Lambda}}^{-1} = \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n$.
Another way to obtain similar results is by taking the partial derivative of $\ell^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{\mu}$ and subsequently with respect to $\boldsymbol{\Lambda}$ to find the function's zeros, as follows

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ell^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = -\frac{1}{2}(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T)(\overline{\mathbf{x}} - \boldsymbol{\mu}) = -\boldsymbol{\Lambda}(\overline{\mathbf{x}} - \boldsymbol{\mu}) = 0 \implies \hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}.$$

$$\frac{\partial}{\partial\boldsymbol{\Lambda}}\ell^*(\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}, \boldsymbol{\Lambda}) = \frac{n}{2}\boldsymbol{\Lambda}^{-1} - \frac{n}{2}\mathbf{S}_n - \frac{1}{2}(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^T = 0 \implies \hat{\boldsymbol{\Lambda}}^{-1} = \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n.$$

□