Suppose that \( \mathbf{X} \sim N_p(\mu, \Sigma) \) and that \( x_1, x_2, \ldots, x_N \) are the observed values in a random sample of size \( N \) from this distribution.

The likelihood function is

\[
L = \prod_{\alpha=1}^{N} n_p (x_\alpha | \mu, \Sigma) = \frac{1}{\left(\sqrt{\det 2\pi \Sigma}\right)^N} \exp \left[ -\frac{1}{2} \sum_{\alpha=1}^{N} (x_\alpha - \mu)' \Sigma^{-1} (x_\alpha - \mu) \right].
\]
The data values $x_1, x_2, \ldots, x_N$ are fixed; we view $L$ as a function of $\mu$ and $\Sigma$, and to distinguish between the population parameters and variables, we denote the latter $\mu^*$ and $\Sigma^*$.

We often work with

$$-2 \log L = N \log(\det 2\pi \Sigma^*) + \sum_{\alpha=1}^{N} (x_\alpha - \mu^*)' \Sigma^{-1} (x_\alpha - \mu^*)$$

Because $\log(\cdot)$ is monotone increasing, maximizing $L$ is equivalent to minimizing $-2 \log L$. 
The sample mean vector is

\[ \bar{x} = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \]

The matrix of sums of squares and cross products of deviations is

\[ A = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) (x_\alpha - \bar{x})' \]

with \((i,j)\) element

\[ \sum_{\alpha=1}^{N} (x_{i,\alpha} - \bar{x}_i) (x_{j,\alpha} - \bar{x}_j) \]
Note that for any \( \xi_1, \xi_2, \ldots, \xi_N \) and any positive definite \( \Sigma^* \),

\[
\sum_{\alpha=1}^{N} \xi'_\alpha \Sigma^{-1} \xi_\alpha = \sum_{\alpha=1}^{N} \text{trace} \left( \Sigma^{-1} \xi_\alpha \xi'_\alpha \right) = \text{trace} \left( \Sigma^{-1} \sum_{\alpha=1}^{N} \xi_\alpha \xi'_\alpha \right)
\]

Note also the matrix generalization of the scalar result:

\[
\sum_{\alpha=1}^{N} (x_\alpha - b) (x_\alpha - b)' = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) (x_\alpha - \bar{x})' + N (\bar{x} - b) (\bar{x} - b)'
\]
So \(-2 \log L\), as a function of \(\mu^*\) and \(\Sigma^*\), can be written

\[-2 \log L = N \log(\det 2\pi \Sigma^*) + \text{trace}\left(\Sigma^{*-1} A\right) + N (\bar{x} - \mu^*)' \Sigma^{*-1} (\bar{x} - \mu^*)\]

Since \(\mu^*\) appears only in the last term, \(-2 \log L\) is clearly minimized at \(\mu^* = \bar{x}\), regardless of the value of \(\Sigma\).

The maximum likelihood estimate of \(\mu\) is therefore

\[\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha,\]

whether \(\Sigma\) is known or unknown.
To find the $\Sigma^*$ that minimizes $-2 \log L$, use the result that for any positive definite $B$,

$$\text{trace}(CB) - N \log \det C$$

is minimized at $C^{-1} = \frac{1}{N} B$.

So for any $\mu^*$, $-2 \log L$ is minimized at

$$\Sigma^* = \frac{1}{N} \left[ A + N (\bar{x} - \mu^*) (\bar{x} - \mu^*)' \right]$$

In particular, when $\mu$ is unknown, and is estimated by $\hat{\mu} = \bar{x}$, the maximum likelihood estimate of $\Sigma$ is

$$\hat{\Sigma} = \frac{1}{N} A = \frac{1}{N} \sum_{\alpha=1}^{N} (x_{\alpha} - \bar{x}) (x_{\alpha} - \bar{x})'.$$
Distribution of the Sample Mean

If we write $\mathbf{Z}$ as the $(Np)$-dimensional random vector

$$\mathbf{Z} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

its density is

$$\frac{1}{\sqrt{\det 2\pi \Sigma_Z}} \exp \left[ -\frac{1}{2} (\mathbf{z} - \mu_Z)' \Sigma_Z^{-1} (\mathbf{z} - \mu_Z) \right]$$
where

\[ \mu_Z = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \]

and

\[ \Sigma_Z = \begin{pmatrix} \Sigma & 0 & \ldots & 0 \\ 0 & \Sigma & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma \end{pmatrix} \]

That is, \( Z \sim N_{np}(\mu_Z, \Sigma_Z) \).
Also $\bar{X} = CZ$, where

$$C_{p \times Np} = \frac{1}{N} (I \ I \ \ldots \ I)$$

So $\bar{X} \sim N_p (\mu_{\bar{X}}, \Sigma_{\bar{X}})$, where

$$\mu_{\bar{X}} = C \mu_Z = \mu,$$

$$\Sigma_{\bar{X}} = C \Sigma_Z C' = \frac{1}{N} \Sigma.$$

That is, $\bar{X} \sim N_p (\mu, \frac{1}{N} \Sigma)$. 
Joint Distribution of the Sample Mean and Covariance Matrix

- To find the joint distribution of $\hat{\mu}$ and $\hat{\Sigma}$, we use the following:
  - Suppose that $X_1, X_2, \ldots, X_N$ are independent, and $X_\alpha \sim N_p(\mu_\alpha, \Sigma)$.
  - If $C$ is an $N \times N$ orthogonal matrix, and we transform from $X_1, X_2, \ldots, X_N$ to $Y_1, Y_2, \ldots, Y_N$ using
    \[ Y_\alpha = \sum_{\beta=1}^{N} c_{\alpha, \beta} X_\beta, \quad \alpha = 1, 2, \ldots, N, \]
  then $Y_1, Y_2, \ldots, Y_N$ are independent, and $Y_\alpha \sim N_p(\nu_\alpha, \Sigma)$, where
    \[ \nu_\alpha = \sum_{\beta=1}^{N} c_{\alpha, \beta} \mu_\beta. \]

- The proof is similar to the previous argument.
In the case of a random sample from $N_p(\mu, \Sigma)$, the means are equal:

$\mu_1 = \mu_2 = \ldots = \mu_N$.

Also, we can construct $C$ so that its last row is

$$
\left( \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}} \right)
$$

Then $Y_N = \sqrt{N} \bar{X}$.

Also, because the other rows of $C$ are orthogonal to the last,

$$
\sum_{\beta=1}^{N} c_{\alpha,\beta} = 0, \quad \alpha < N,
$$

so $Y_\alpha \sim N_p(0, \Sigma)$ for $\alpha = 1, 2, \ldots, N - 1$. 
Again because of the orthogonality of $C$,

$$\sum_{\alpha=1}^{N} X_\alpha X'_\alpha = \sum_{\alpha=1}^{N} Y_\alpha Y'_\alpha$$

But $Y_N Y'_N = N \bar{X} \bar{X}'$, so

$$A = \sum_{\alpha=1}^{N} X_\alpha X'_\alpha - N \bar{X} \bar{X}' = \sum_{\alpha=1}^{N-1} Y_\alpha Y'_\alpha$$

Since $Y_N$ is independent of $Y_1, Y_2, \ldots, Y_{N-1}$, it follows that $\bar{X}$ is independent of $A$. 
Summary

- The maximum likelihood estimators $\hat{\mu}$ and $\hat{\Sigma}$ are independent.
- $\hat{\mu} = \bar{X} \sim N_p (\mu, \frac{1}{N} \Sigma)$.
- $\hat{\Sigma} = \frac{1}{N} A$ is distributed as
  \[ \frac{1}{N} \sum_{\alpha=1}^{N-1} Y_\alpha Y'_\alpha \]
  where $Y_1, Y_2, \ldots, Y_{N-1}$ are independent $N_p(0, \Sigma)$.
- The distribution of $A$ is the Wishart distribution $W_p(\Sigma, N - 1)$.
- Since trivially $E(A) = (N - 1) \Sigma$, $\hat{\Sigma}$ is a biased estimator of $\Sigma$, and we often use the sample covariance matrix
  \[ S = \frac{N}{N - 1} \hat{\Sigma} = \frac{1}{N - 1} \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \]
  which is the sample value of an unbiased estimator.