

# BIOS 767 Homework 2

Mingwei Fei

## 1 Problem 1- Derive fundamental properties of the multivariate normal distribution

Consider the  $n \times 1$  random vector  $Y$  satisfying  $Y \sim MVN(\mu, \Sigma)$ ,  $\Sigma$  is positive definite matrix.

- (a) Let  $Z = aY + c$ , where  $a$  is an  $r \times n$  matrix of constants having rank  $r$  ( $r \leq n$ ) and  $c$  is a  $r \times 1$  vector of constants. Show using the definition of moment generating function that  $Z \sim Normal(a\mu + c, a\Sigma a')$ .

As  $Y \sim MVN(\mu, \Sigma)$ ,  $\Sigma$  is positive definite matrix.

Then we can decompose  $\Sigma = BB^T$  for some nonsingular matrix  $B$  since  $\Sigma$  is positive definite.

$$X = B^{-1}(Y - \mu), \quad Y = \mu + BX$$

So we have  $X_1, \dots, X_n$  are independent standard normal,  $X = (X_1, \dots, X_n)^T \sim MVN(0, I_n)$ .

$$\begin{aligned} M_Y(t) &= E[e^{t^T Y}] = E[e^{t^T(\mu + BX)}] = e^{t^T \mu} E[e^{l^T X}], \quad l^T = t^T B \\ &= e^{t^T \mu} E[e^{\sum_{i=1}^n l_i Y_i}], \quad l = (l_1, \dots, l_n) \\ &= e^{t^T \mu} \prod_{i=1}^n E[e^{l_i Y_i}] \\ &= e^{t^T \mu} \prod_{i=1}^n e^{l_i^2/2} \\ &= \exp\left(\mu^T t + \frac{1}{2} l^T l\right) \\ &= \exp\left(\mu^T t + \frac{1}{2} t^T \Sigma t\right) \end{aligned}$$

While  $Z = aY + c$ , then we have the MGF for  $Z$

$$\begin{aligned} M_Z(t) &= E[e^{t^T Z}] = E[e^{t^T(aY + c)}] = e^{t^T c} E[e^{t^T aY}] \\ &= e^{t^T c} E[e^{(t^T a)Y}] \\ &= e^{t^T c} \exp\left(\mu^T (a^T t) + \frac{1}{2} (a^T t)^T \Sigma (a^T t)\right) \\ &= \exp\left(t^T (a\mu + c) + \frac{1}{2} t^T a \Sigma a^T t\right) \end{aligned}$$

So  $E(Z) = (a\mu + c)$ ,  $Var(Z) = a\Sigma a'$ , and  $Z \sim Normal(a\mu + c, a\Sigma a')$ .

## 2 Problem 2 - Develop a theoretical understanding of REML estimation

$Y_i \sim MVN(X_i\beta, \Sigma)$

- (a) Read "Bayesian inference for variance components using only error contrasts" by Harville(1974).

In the paper, an alternative method for deriving the likelihood function for error contrasts are provided using Bayesian method. I would like to walk through the derivation of the likelihood here in the homework. The error contrasts are the product of  $n - p$  independent normal distributions,

$$\begin{aligned} M &= X(X'X)^{-1}X', \quad \Sigma = \Sigma(\theta) \\ I - M &= AA', \quad A'A = I, \quad \text{Cholesky decomposition} \\ A'y &\sim MVN(A'X\hat{\beta}, A'\Sigma A), \quad Var(A'y) = A'Var(y)A = A'\Sigma A = \Sigma \\ f_w(A'y|\theta, \hat{\beta}) &= (2\pi)^{-\frac{n-p}{2}} \left(|\Sigma|\right)^{-1/2} \exp\left(-\frac{n-p}{2}(A'y - A'X\hat{\beta})'(\Sigma)^{-1}(A'y - A'X\hat{\beta})\right) \end{aligned}$$

where  $f_w(A'y|\theta, \hat{\beta})$  is the likelihood function of  $\theta, \hat{\beta}$ , and we could use Bayesian method to find the distribution of  $\hat{\beta}$  and further integrate out  $\hat{\beta}$  to get the distribution of  $f_w(A'y|\theta)$

Let  $\Sigma = QQ'$ , then we have the orthogonal projection matrix onto  $C(Q^{-1}X)$

$$\begin{aligned} P &= Q^{-1}y\left((Q^{-1}y)^T(Q^{-1}y)\right)^{-1}(Q^{-1}y)' = Q^{-1}y(y^TQ'^{-1}Q^{-1}y)^{-1}(Q^{-1}y)^T \\ &= \Sigma^{-1}y(y^T\Sigma^{-1}y)^{-1}y^T \end{aligned}$$

The estimate of  $\hat{\beta}$  could be denoted as

$$\begin{aligned} G &= \Sigma^{-1}y(y^T\Sigma^{-1}y)^{-1}, \quad G'Y = \hat{\beta} \\ G'y &\sim MVN(\hat{\beta}, G'\Sigma G), \quad Var(G'y) = G'Var(y)G = G'\Sigma G = (Y'\Sigma^{-1}Y)^{-1} \end{aligned}$$

We also have the relationship as below, it shows that the  $f_w(\beta, \theta)$  are normal distribution, and the quadratic form could be decomposed into  $f_w(\theta, \hat{\beta})$  and  $f_{\hat{\beta}}(\beta)$ .

$$\begin{aligned} p(\beta|Y, \Sigma) &\propto \exp\left(-\frac{1}{2}(Y - X\beta)' \Sigma^{-1}(Y - X\beta)\right) \\ (Y - X\beta)' \Sigma^{-1}(Y - X\beta) &= (Y - X\hat{\beta})' \Sigma^{-1}(Y - X\hat{\beta}) + (\beta - \hat{\beta})'(X\Sigma^{-1}X)(\beta - \hat{\beta}) \\ p(\beta|Y, \Sigma) &\propto \exp\left((Y - X\hat{\beta})' \Sigma^{-1}(Y - X\hat{\beta}) + (\beta - \hat{\beta})'(X\Sigma^{-1}X)(\beta - \hat{\beta})\right) \\ &\propto \exp\left((\beta - \hat{\beta})'(X\Sigma^{-1}X)(\beta - \hat{\beta})\right) \end{aligned}$$

Also the paper provided the derivation using distribution of  $A'y$ , which is  $f_w(\theta, \hat{\beta})$  and distribution for  $G'y$ , which is  $f_{\hat{\beta}}(\beta)$ .

$$\begin{aligned}
f_w(A'y|\theta) &= \int f_w(A'y|\theta, \hat{\beta}) f_{\hat{\beta}}(G'y|\theta, \beta) d\beta \\
&= \left( \det(Y'Y) \right)^{1/2} \int f_y(y|\theta, \beta) d\beta \\
&= (2\pi)^{-\frac{1}{2}(n-p)} \left( \det(Y'Y) \right)^{1/2} \left( \det(\Sigma) \right)^{-1/2} \left( \det(Y'\Sigma Y) \right)^{-1/2} \\
&\quad \exp \left( -\frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta}) \right)
\end{aligned}$$

Overall, this paper provided a Bayesian method to get the marginal distribution of  $\Sigma(\theta)$ , so that we can use the maximum likelihood method to estimate.

- (b) Read the paper "A Conditional Derivation of Residual Maximum Likelihood" by Verbyla (1990).

This paper reviewed the Bayesian method in deriving likelihood function in Harville's paper and provided an alternative method in doing transformation without  $\Sigma$ . The approach is to construct the conditional distribution in matrix form. The estimate of  $\hat{\beta}$  is based on the conditional distribution  $Y_1|Y_2$  which based on the orthogonal projection matrix onto  $C(Q^{-1}X)$ .

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

And the REML log-likelihood for  $\sigma^2$  and  $\gamma$  is the marginal log-likelihood based on  $Y_2$ . The approach used the standard decomposition of the joint distribution of  $L'Y$  into conditional and marginal distributions. The idea of residual maximum likelihood as a marginal likelihood is actually the idea of the use of a conditional likelihood.

- (c) Derive the REML log-likelihood for  $\Sigma Y \sim X\beta + \epsilon$ , where  $\epsilon \sim N(0, \sigma^2\Omega(\gamma))$  Let  $L_1(n \times p)$  and  $L_2(n \times (n-p))$  satisfying  $L_1'X = I_p$  and  $L_2'X = 0$ . Also

$$L'Y = [L_1', L_2']^T Y = [Y_1, Y_2]^T$$

Thus,

$$\begin{aligned}
Y &= [Y_1, Y_2]^T \sim N\{\mu, \Sigma\} \\
\mu &= \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \quad \Sigma = \sigma^2 \begin{pmatrix} L_1'\Omega L_1 & L_1'\Omega L_2 \\ L_2'\Omega L_1 & L_2'\Omega L_2 \end{pmatrix},
\end{aligned}$$

From the paper,  $Y_2 \sim N(0, \sigma^2 L_2'\Omega L_2)$ , and the REML log-likelihood for  $\sigma^2$  and  $\gamma$  is the marginal log-likelihood based on  $Y_2$  and is

$$\begin{aligned}
\text{Log}L(\sigma^2, \gamma, Y_2) &= -\frac{1}{2} \left( (n-p) \log \sigma^2 - \log \det(L_2'\Omega L_2) - \frac{Y_2'(L_2'\Omega L_2)^{-1}Y_2}{\sigma^2} \right) \\
&= -\frac{1}{2} \left( (n-p) \log \sigma^2 - \log \det(\Omega) - \log \det(X'\Omega^{-1}X) - \frac{Y'PY}{\sigma^2} \right) \\
P &= L_2(L_2'\Omega L_2)^{-1}L_2' = \Omega^{-1} - \Omega^{-1}X(X'\Omega X)^{-1}X'\Omega^{-1}
\end{aligned}$$

where we kept only the terms associated with  $\sigma^2$  and  $\gamma$ . We will use the following property in deriving the MLE

$$\begin{aligned}
tr[x'Ax] &= tr[xx'A], \\
\partial_A \log|A| &= A^{-1}, \\
\partial_A tr(BA) &= B, \text{ if } A \text{ is symmetric} \\
LogL(\sigma^2, \gamma, Y_2) &= -\frac{1}{2} \left( (n-p) \log \sigma^2 + \log|\Omega^{-1}| - \log|tr(X'X'\Omega^{-1})| \right. \\
&\quad \left. - \frac{Y' \left[ \Omega^{-1} - \Omega^{-1} X (X' \Omega X)^{-1} X' \Omega^{-1} \right] Y}{\sigma^2} \right) \\
\partial_{\Omega^{-1}} Logf &= -\frac{1}{2} \Omega + \frac{1}{2} \Omega + \frac{1}{2\sigma^2} Y Y' - \frac{1}{2\sigma^2} Y Y' \Omega^{-1} X (X' \Omega X)^{-1} = 0 \\
\partial_{\sigma^2} Logf &= -\frac{(n-p)}{2\sigma^2} - \frac{Y' P Y}{2\sigma^4} = 0
\end{aligned}$$

We can solve  $\sigma^2$  and  $\gamma$  by above equations.

- (i) Verify Verbyla's representation is equivalent to equation (3) from Harville's paper. By looking at the equation(3) from Harville's paper

$$\begin{aligned}
f_w(A'y|\theta) &= (2\pi)^{-\frac{(n-p)}{2}} \left( \det(X'X) \right)^{1/2} \left( \det(\Sigma) \right)^{-1/2} \left( \det(Y'\Sigma Y) \right)^{-1/2} \\
&\quad \exp \left( -\frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta}) \right)
\end{aligned}$$

Let  $\Sigma = \sigma^2 \Omega$ , then

$$\begin{aligned}
\det(\Sigma) &= \det(\sigma^2 \Omega) = (\sigma^2)^{n-p} \det(\Omega), \quad \Sigma_{n-p \times n-p} \\
\det(\Sigma)^{-1/2} &= \det(\sigma^2 \Omega)^{-1/2} = (\sigma^2)^{-(n-p)/2} \det(\Omega)^{-1/2}
\end{aligned}$$

So, also need to show that  $\exp \left( -\frac{1}{2} \frac{Y' P Y}{\sigma^2} \right)$  is the same as  $\exp \left( -\frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta}) \right)$

$$\begin{aligned}
\exp \left( -\frac{1}{2} \frac{Y' P Y}{\sigma^2} \right) &= \exp \left( -\frac{1}{2} Y' \left[ \Sigma^{-1} - \Sigma^{-1} X (X' \Sigma X)^{-1} X' \Sigma^{-1} \right] Y \right) \\
&= \exp \left( -\frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta}) \right), \quad (X' \Sigma X)^{-1} X' \Sigma^{-1} Y = \hat{\beta}
\end{aligned}$$

From above equation, we can see that the two equations are the same from the two papers.

- (ii) Explain how the REML estimator of  $\beta$  in this setting can be viewed as a type of "plug-in" generalized least squares estimator where the weight matrix is equal to the REML estimate for  $\Omega$ .

The REML log-likelihood does not contain  $\beta$ , we could use the REML log-likelihood to estimate the  $\Omega$ , denoted by  $\hat{\Omega}$ . Furthermore, we can estimate the  $\beta$  by the following

$$\hat{\beta} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} Y$$

In the weighted least square estimator,

$$\tilde{\beta} = (X' W X)^{-1} X' W Y, \quad W = \text{diag}\{w_1, \dots, w_n\}$$

Thus, the REML estimate of  $\beta$  could be considered as WLS estimate with "plug-in"  $W = \hat{\Omega}^{-1}$ .

- (d) Now, consider the non-regression case where each random vector  $Y_i$  has mean  $\mu$ . Note that we may formulate this case as a regression model with matrix  $X_{N \times n}$  defined by

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} I_n \\ I_n \\ \vdots \\ I_n \end{pmatrix}$$

and  $\beta = \mu$ . Show that the REML estimate for  $\Sigma$  maximizes the log-likelihood given by

$$\log l_{REML}(\Sigma) \propto \frac{N-1}{2} \log(|\Sigma^{-1}|) - \frac{1}{2} \text{tr}[(N-1)S\Sigma^{-1}]$$

Proof: From above, we have the REML log-likelihood as following:

$$\begin{aligned} \ln_{REML}(\beta) &\propto \frac{1}{2}(\log |\Omega^{-1}|) - \frac{1}{2}(\log |X' \Omega^{-1} X|) - \frac{1}{2}(Y - X\hat{\beta})^T \Omega^{-1} (Y - \hat{\beta}) \\ |\Omega| &= |\Sigma|^N, \quad X' \Omega^{-1} X = \sum_{i=1}^N X' \Sigma^{-1} X \\ \ln_{REML}(\beta) &\propto \frac{N}{2}(\log |\Sigma^{-1}|) - \frac{N}{2}(\log |X' \Sigma^{-1} X|) - \frac{1}{2}(\beta - \hat{\beta})^T X^T \Omega^{-1} X (\beta - \hat{\beta}), \quad X = I \\ &\propto \frac{N}{2}(\log |\Sigma^{-1}|) - \frac{1}{2}(\log |N\Sigma^{-1}|) - \frac{1}{2} \text{tr} \left[ \sum_{i=1}^N (Y_i - X_i \hat{\beta})^T \Sigma^{-1} (Y_i - X_i \hat{\beta}) \right], \\ &\propto \frac{N}{2}(\log |\Sigma^{-1}|) - \frac{n}{2} \log N - \frac{1}{2}(\log |\Sigma^{-1}|) - \frac{1}{2} \text{tr} \left[ \sum_{i=1}^N (Y_i - \bar{Y}_i)(Y_i - \bar{Y}_i)^T \Sigma^{-1} \right] \\ &\propto \frac{N-1}{2} \log(|\Sigma^{-1}|) - \frac{1}{2} \text{tr}[(N-1)S\Sigma^{-1}] \end{aligned}$$

We prove the REML log-likelihood function

- (e) Derive the REML estimate for  $\Sigma$  is the same as  $S$ .

We will take the first derivative of the REML log-likelihood function as in part(d)

$$\begin{aligned} \partial_{\Sigma^{-1}} \ln_{REML}(\beta) &= \frac{N-1}{2} \Sigma - \frac{N-1}{2} S = 0 \\ \Sigma &= S \end{aligned}$$

### 3 Problem 3

Objective: Develop understanding of applying the EM Algorithm by deriving the REML estimator for using an EM approach.

Let  $Y_i, i = 1, \dots, N$ , be i.i.d. random sample of  $n \times 1$  random vectors with  $Y_i \sim MVN(\mu, \Sigma)$  for positive definite  $\Sigma$ . We will apply the EM algorithm to estimate  $\Sigma$  and observe the estimator is equal to that obtained in problem 2.

- (a) Write the complete data log-likelihood  $l(\Sigma|\mu, Y)$ . Note that we are writing the expression with  $\mu$  on the RHS simply to convey that we are treating it effectively as missing data for this exercise.

The complete data log-likelihood, treating  $\mu$  as missing,

$$\begin{aligned} p(\Sigma|\mu, Y_i) &= (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y_i - \mu)^T \Sigma^{-1}(Y_i - \mu)\right) \\ \ln(\Sigma|\mu, Y_i) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(Y_i - \mu)^T \Sigma^{-1}(Y_i - \mu) \\ \ln(\Sigma|\mu, Y) &= -\frac{Nn}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \sum_{i=1}^N \frac{1}{2}(Y_i - \mu)^T \Sigma^{-1}(Y_i - \mu) \end{aligned}$$

- (b) Verify the conditional distribution for the missing data  $\mu$  given the observed data  $Y_i, i = 1, \dots, N$  is given by

$$\mu|Y, \Sigma^{(t)} \sim \text{Normal}\left(\bar{Y}, \frac{\Sigma^{(t)}}{N}\right)$$

Prove: We know that the conditional distribution of  $\mu$

$$f(\mu|\Sigma, Y) = \frac{f(\mu, Y|\Sigma)}{f(Y|\Sigma)} \propto f(\mu, Y|\Sigma)$$

When  $\Sigma$  is known, we will have the distribution of  $\mu$

$$f(\mu, Y|\Sigma) \propto \prod_{i=1}^N \exp\left(-\frac{1}{2} \sum_{i=1}^N (Y_i - \mu)^T \Sigma^{-1}(Y_i - \mu)\right)$$

We will write the distribution in terms of known Y statistics and  $\Sigma^{(t)}$

$$\begin{aligned}
f(\mu, Y|\Sigma) &\propto \prod_{i=1}^N \exp\left(-\frac{1}{2} \sum_{i=1}^N (Y_i^T \Sigma^{-1} Y_i - 2\mu^T \Sigma^{-1} Y_i + \mu^T \Sigma^{-1} \mu)\right) \\
&\propto \prod_{i=1}^N \exp\left(-\frac{1}{2} \sum_{i=1}^N (\mu^T \Sigma^{-1} \mu - 2\mu^T \Sigma^{-1} Y_i)\right) \\
&\propto \prod_{i=1}^N \exp\left(-\frac{1}{2} (N\mu^T \Sigma^{-1} \mu - 2N\mu^T \Sigma^{-1} \bar{Y}_i)\right) \\
&\propto \prod_{i=1}^N \exp\left(-\frac{1}{2} ((\mu - \bar{Y}_i)^T (N\Sigma^{-1}) (\mu - \bar{Y}_i))\right) \\
&\sim N\left(\bar{Y}_i, \frac{\Sigma}{N}\right)
\end{aligned}$$

We could replace  $\Sigma$  by  $\Sigma^{(t)}$ , thus we have the proof.

(c) Compute  $Q(\Sigma|\Sigma^{(t)})$  and derive the  $\Sigma^{(t+1)}$  as a function of  $\Sigma^{(t)}$ .

$Q(\Sigma|\Sigma^{(t)})$  is the expectation of  $\ln(\Sigma|\mu^{(t)}, Y)$

$$\begin{aligned}
Q(\Sigma|\Sigma^{(t)}) &= E\left[\ln(\Sigma|\mu, Y)\right] \\
&= E\left[-\frac{Nn}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \sum_{i=1}^N \frac{1}{2} (Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)\right] \\
&\propto -\frac{N}{2} E\left[\log |\Sigma|\right] - \frac{1}{2} E\left[\sum_{i=1}^N (Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)\right] \\
&\propto -\frac{N}{2} E\left[\log |\Sigma|\right] - \frac{1}{2} E\left[\sum_{i=1}^N Y_i^T \Sigma^{-1} Y_i - 2Y_i^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu\right]
\end{aligned}$$

We have

$$\begin{aligned}
E[\mu] &= \bar{Y}_i \\
E[\mu^T \Sigma^{-1} \mu] &= E\left[\text{tr}\left(\mu \mu^T \Sigma^{-1}\right)\right] \\
&= \text{tr}\left(E\left[\mu \mu^T \Sigma^{-1}\right]\right) = \Sigma^{-1} \text{tr}\left(\text{Var}(\mu) + E[\mu]^2\right) \\
&= \text{tr}\left[\left(\frac{\Sigma^{(t)}}{N} + \bar{Y}^{\otimes 2}\right) \Sigma^{-1}\right]
\end{aligned}$$

Then we have

$$\begin{aligned}
Q(\Sigma|\Sigma^{(t)}) &\propto -\frac{N}{2}E\left[\log|\Sigma^{(t)}|\right] - \frac{1}{2}E\left[\sum_{i=1}^N Y_i^T \Sigma^{-1} Y_i - 2Y_i^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu\right] \\
&\propto -\frac{N}{2}\log|\Sigma^{(t)}| - \frac{1}{2}\sum_{i=1}^N Y_i^T \Sigma^{-1} Y_i + \sum_{i=1}^N Y_i^T \Sigma^{(t)-1} \bar{Y}_i + \frac{N}{2}\text{tr}\left(\frac{\Sigma^{(t)}}{N}\Sigma^{-1} + \bar{Y}^{\otimes 2}\Sigma^{-1}\right) \\
&\propto -\frac{N}{2}\log|\Sigma^{(t)}| - \frac{1}{2}\sum_{i=1}^N Y_i^T \Sigma^{-1} Y_i + \sum_{i=1}^N Y_i^T \Sigma^{(t)-1} \bar{Y}_i + \frac{N}{2}\text{tr}\left(\frac{\Sigma^{(t)}}{N}\Sigma^{-1} + \bar{Y}^T \Sigma^{-1} \bar{Y}\right) \\
&\propto -\frac{N}{2}\log|\Sigma^{(t)}| - \frac{1}{2}\text{tr}\left[\sum_{i=1}^N (Y_i - \bar{Y})^T \Sigma^{-1} (Y_i - \bar{Y}) + \Sigma^{(t)}\Sigma^{-1}\right] \\
&\propto -\frac{N}{2}\log|\Sigma^{(t)}| - \frac{1}{2}\text{tr}\left[\left\{\sum_{i=1}^N (Y_i - \bar{Y})(Y_i - \bar{Y})^T + \Sigma^{(t)}\right\}\Sigma^{-1}\right] \\
&\propto -\frac{N}{2}\log|\Sigma^{(t)}| - \frac{1}{2}\text{tr}\left[\{(N-1)S + \Sigma^{(t)}\}\Sigma^{-1}\right]
\end{aligned}$$

The  $\Sigma^{(t+1)}$  could be derived by MLE of  $Q(\Sigma|\Sigma^{(t)})$

$$\begin{aligned}
\partial_{\Sigma} Q(\Sigma|\Sigma^{(t)}) &= -\frac{N}{2}\Sigma^{(t)} - \frac{1}{2}\{(N-1)S + \Sigma^{(t)}\} = 0 \\
\Sigma &= \frac{N-1}{N}S + \frac{1}{N}\Sigma^{(t)} \\
\Sigma^{(t+1)} &= \frac{N-1}{N}S + \frac{1}{N}\Sigma^{(t)}
\end{aligned}$$

The limiting of  $\Sigma^{(t+1)}$

$$\begin{aligned}
\Sigma^{(t+1)} &= \frac{N-1}{N}S + \frac{1}{N}\Sigma^{(t)} \\
\Sigma^{(t+k)} &= \left(\frac{N-1}{N} + \frac{N-1}{N}\frac{1}{N} + \dots + \frac{N-1}{N}\frac{1}{N}^{k-1}\right)S + \frac{1}{N}^k \Sigma^{(t)} \\
\Sigma^{(t+k)} &\xrightarrow{k \rightarrow \infty} S \frac{\frac{N-1}{N}}{1 - \frac{1}{N}} = S
\end{aligned}$$

So  $\hat{\Sigma}_{MLE} = S$