

Matrix Calculus

Sangmyung Ha

January 15, 2026

1 Definitions and Facts

1. Trace of $n \times n$ dimensional matrix $A = (a_{ij})_{i,j=1}^n$ is defined as

$$\text{tr}(A) = \sum_{i=1}^n a_{ii}$$

FACTS:

- (a) $\text{tr}(AB) = \text{tr}(BA)$ as long as AB and BA are well defined.
 - (b) Fact (a) implies $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$, i.e., cyclical permutations are allowed.
2. Let A be $m \times n$ dimensional matrix given by $A = (a_1, \dots, a_n)$. We define column-major vectorization of A as

$$\overline{\text{vec}}(A) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix},$$

and row-major vectorization of A as

$$\text{vec}(A) = \overline{\text{vec}}(A').$$

FACTS:

- (a) $\text{tr}(A'B) = \overline{\text{vec}}(A)' \overline{\text{vec}}(B) = \text{vec}(A)' \text{vec}(B)$ so that $\text{tr}(A'B)$ defines inner product of two matrices A and B .
- (b) The norm induced by the trace inner product is called Frobenius norm and is given by

$$\|A\|_F = \sqrt{\text{tr}(A'A)}$$

3. Kronecker product of two vectors $a = (a_1, \dots, a_m)' \in \mathbb{R}^m$ and $b = (b_1, \dots, b_p)' \in \mathbb{R}^p$ is defined as mp dimensional vector

$$a \otimes b = \begin{pmatrix} a_1 b \\ \vdots \\ a_m b \end{pmatrix},$$

i.e., Kronecker product defines lexicographic elementwise multiplication.

FACTS:

- (a) Kronecker product is reordering of outer products, i.e.,

$$\begin{aligned} \overline{\text{vec}}(ab') &= b \otimes a \\ \text{vec}(ab') &= a \otimes b, \end{aligned}$$

i.e., $(a \otimes b)$ defines lexicographic reordering of ab' .

- (b) Let $\circ$ denote tensor product defined as

$$(a \circ b \circ c) \in \mathbb{R}^{m \times p \times r},$$

where $a \in \mathbb{R}^m, b \in \mathbb{R}^p$, and $c \in \mathbb{R}^r$, i.e., (i, j, k) - *th* element of $(a \circ b \circ c)$ is given by $a_i b_j c_k$ and likewise for higher orders, e.g., $(a \circ b \circ c \circ d) \in \mathbb{R}^{m \times p \times r \times t}$. We call such higher-order array of numbers *tensors*. Kronecker product is lexicographic (backward) reordering of tensor products, i.e.,

$$\begin{aligned} \overline{\text{vec}}(a \circ b \circ c) &= (c \otimes b \otimes a) \\ \text{vec}(a \circ b \circ c) &= (a \otimes b \otimes c), \end{aligned}$$

where $\overline{\text{vec}}$ vectorizes tensors by enumerating from the first direction, vec vectorizes tensors by enumerating from the final direction.

4. Kronecker product of $m \times n$ and $p \times q$ dimensional matrices $A = (a_1, \dots, a_n)$ and $B = (b_1, \dots, b_q)$ is defined as

$$A \otimes B = \begin{pmatrix} a_1 \otimes b_1 & a_1 \otimes b_2 & \dots & a_n \otimes b_{q-1} & a_n \otimes b_q \end{pmatrix},$$

i.e., lexicographic Kronecker products of columns of A and B .

FACTS

- (a) $(A \otimes B)' = (A' \otimes B')$
- (b) $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$
- (c) $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
- (d) $(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D$

- (e) $(A \otimes B)(C \otimes D) = (AC \otimes BD)$
- (f) $\overline{\text{vec}}(ABC) = (C' \otimes A)\overline{\text{vec}}(B)$
- (g) $\text{vec}(ABC) = (A \otimes C')\text{vec}(B)$
- (h) $P_{A \otimes B} = P_A \otimes P_B$.

Facts (a)-(e) can be verified from the definition of Kronecker product. Fact (e) implies that Kronecker product of two matrices preserves tensor structure, i.e.,

$$(A \otimes B)(a \otimes b) = (Aa \otimes Bb) = \text{vec}((Aa)(Bb)') = \text{vec}((Aa) \circ (Bb)),$$

more generally,

$$(A \otimes B \otimes C)(a \otimes b \otimes c) = (Aa \otimes Bb \otimes Cc) = \text{vec}((Aa) \circ (Bb) \circ (Cc)).$$

Also,

$$(A_1 \otimes B_1)(A_2 \otimes B_2)(a \otimes b) = (A_1 A_2 a \otimes B_1 B_2 b) = \text{vec}((A_1 A_2 a) \circ (B_1 B_2 b)).$$

Fact (f) and (g) follows from Fact (e). Let the singular value decomposition of B be given by $B = UDV' = \sum_j \delta_j u_j v_j'$. We have

$$\begin{aligned} (A \otimes C')\text{vec}(B) &= (A \otimes C')\text{vec}\left(\sum_j \delta_j u_j v_j'\right) \\ &= \sum_j (A \otimes C')\delta_j \text{vec}(u_j v_j') \\ &= \sum_j \delta_j (A \otimes C')(u_j \otimes v_j) \\ &= \sum_j \delta_j (A u_j \otimes C' v_j) \\ &= \sum_j \delta_j \text{vec}(A u_j v_j' C) \\ &= \text{vec}\left(A \sum_j \delta_j u_j v_j' C\right) \\ &= \text{vec}(ABC) \end{aligned}$$

Fact (h) comes naturally from Facts (a), (b), (e). To see this,

$$\begin{aligned} P_{A \otimes B} &= (A \otimes B)((A \otimes B)'(A \otimes B))^{-1}(A \otimes B)' \\ &= (A \otimes B)((A' A)^{-1} \otimes (B' B)^{-1})(A \otimes B)' \\ &= P_A \otimes P_B. \end{aligned}$$

2 Differential

¹ Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be the function that we want to differentiate. We define the **differential** $df(x)$ of f at a point x as a linear function of $dx \in \mathbb{R}^n$ that satisfies

$$df(x) = \left(\frac{\partial f(x)}{\partial x'} \right) dx,$$

where $\frac{\partial f(x)}{\partial x'} = \left(\frac{\partial f_i(x)}{\partial x_j} \right)_{i,j=1}^{m,n} \in \mathbb{R}^{m \times n}$ is a matrix of partial derivatives.

Lemma 1. The differential is a linear operator and obeys the following rules:

- (1) $da = 0$
- (2) $d(x') = (dx)'$
- (3) $d(a'x) = a'dx$
- (4) $d(x + y) = dx + dy$
- (5) $d(x'y) = (dx)'y + x'dy$
- (6) $d \log x = \frac{1}{x} dx$
- (7) $d\left(\frac{1}{x}\right) = -\left(\frac{1}{x^2}\right)(dx)$

where x, y are variables and a is constant.

The goal is to first find the differential written as $df(x) = A(x)dx$ for some $A(x) \in \mathbb{R}^{m \times n}$. Then, we may deduce that $A(x) = \frac{\partial f(x)}{\partial x'}$.

2.1 Examples

1. Find differential of $f(\beta) = (y - X\beta)'(y - X\beta)$. Note that

$$\begin{aligned} df(\beta) &= d((y - X\beta)'(y - X\beta)) \\ &= (d(y - X\beta)')(y - X\beta) + (y - X\beta)'d((y - X\beta)) \\ &= -(d\beta)'X'(y - X\beta) - (y - X\beta)'X(d\beta) \\ &= -2(y - X\beta)'X(d\beta) \end{aligned}$$

where we've used Lemma 1(5) in the second equality and Lemma 1(4), (1), (2) in the third equality. Thus, the derivative $\frac{\partial f(\beta)}{\partial \beta'}$ is given by

$$\frac{\partial f(\beta)}{\partial \beta'} = -2(y - X\beta)'X$$

¹References: "Matrix Differential Calculus with Applications in Statistics and Econometrics", Textbook by Heinz Neudecker and Jan R. Magnus

2. Find differential of the likelihood function $\ell(\beta, \sigma^2)$ from the linear model

$$y = X\beta + \varepsilon$$

where $\varepsilon \sim \mathbb{N}(0, \sigma^2 I_n)$ and X is fixed. Note that $y \sim \mathbb{N}(X\beta, \sigma^2 I_n)$ so that

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta).$$

The differential is given by

$$d\ell(\beta, \sigma^2) = -\frac{n}{2} d \log \sigma^2 - \frac{1}{2} d \left(\frac{1}{\sigma^2} \right) (y - X\beta)'(y - X\beta) + \frac{1}{\sigma^2} (y - X\beta)' X (d\beta).$$

Note that we have

$$\begin{aligned} d \log \sigma^2 &= \frac{1}{\sigma^2} d\sigma^2 \\ d \left(\frac{1}{\sigma^2} \right) &= - \left(\frac{1}{(\sigma^2)^2} \right) (d\sigma^2), \end{aligned}$$

so that

$$\begin{aligned} d\ell(\beta, \sigma^2) &= -\frac{n}{2} d \log \sigma^2 - \frac{1}{2} d \left(\frac{1}{\sigma^2} \right) (y - X\beta)'(y - X\beta) + \frac{1}{\sigma^2} (y - X\beta)' X (d\beta) \\ &= -\frac{n}{2} \frac{1}{\sigma^2} d\sigma^2 + \frac{1}{2} \left(\frac{1}{(\sigma^2)^2} \right) (d\sigma^2) (y - X\beta)'(y - X\beta) + \frac{1}{\sigma^2} (y - X\beta)' X (d\beta) \\ &= \left(-\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \left(\frac{1}{(\sigma^2)^2} \right) (y - X\beta)'(y - X\beta) - \frac{1}{\sigma^2} (y - X\beta)' X \right) \begin{pmatrix} d\sigma^2 \\ d\beta \end{pmatrix}. \end{aligned}$$

Thus, we have

$$\begin{aligned} \frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \left(\frac{1}{(\sigma^2)^2} \right) (y - X\beta)'(y - X\beta) \\ \frac{\partial \ell(\beta, \sigma^2)}{\partial \beta'} &= \frac{1}{\sigma^2} (y - X\beta)' X. \end{aligned}$$

Solving for $(\hat{\beta}, \hat{\sigma})$, we obtain

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})'(y - X\hat{\beta}) \end{aligned}$$

3. (Serially Correlated Errors)

Find differential of the likelihood function $\ell(\beta, \rho, \sigma^2)$ from the linear model

$$\begin{aligned} y_t &= x_t' \beta + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t \end{aligned}$$

where $\varepsilon_t \sim \mathbb{N}(0, \sigma^2)$, $u_0 = 0$, and (x_t) is fixed.

$$\begin{aligned} p((y_t)_{t=1}^n | \beta, \rho, \sigma^2) &= p((u_t)_{t=1}^n | \beta, \rho, \sigma^2) \\ &= \prod_{t=1}^n p(u_t | (u_s)_{s=1}^{t-1}, \beta, \rho, \sigma^2), \end{aligned}$$

so that

$$\begin{aligned} \ell(\beta, \rho, \sigma^2) &= \sum_{t=1}^n \log(p(u_t | (u_s)_{s=1}^{t-1}, \beta, \rho, \sigma^2)) \\ &= \sum_{t=1}^n \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (u_t - \rho u_{t-1})^2 \right) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^n (u_t - \rho u_{t-1})^2 \end{aligned}$$

where $u_t = y_t - x'_t \beta$, $u_0 = 0$. The differential is given by

$$\begin{aligned} d\ell(\beta, \rho, \sigma^2) &= -\frac{n}{2} d\log(\sigma^2) - d\left(\frac{1}{2\sigma^2}\right) \sum_{t=1}^n (u_t - \rho u_{t-1})^2 - \frac{1}{2\sigma^2} \sum_{t=1}^n d(u_t - \rho u_{t-1})^2 \\ &= -\frac{n}{2} \frac{1}{\sigma^2} (d\sigma^2) + \frac{1}{2} \left(\frac{1}{(\sigma^2)^2}\right) \sum_{t=1}^n (u_t - \rho u_{t-1})^2 (d\sigma^2) \\ &\quad - \frac{1}{2\sigma^2} \sum_{t=1}^n 2(u_t - \rho u_{t-1}) (du_t - u_{t-1}(d\rho) - \rho du_{t-1}) \\ &= -\frac{n}{2} \frac{1}{\sigma^2} (d\sigma^2) + \frac{1}{2} \left(\frac{1}{(\sigma^2)^2}\right) \sum_{t=1}^n (u_t - \rho u_{t-1})^2 (d\sigma^2) \\ &\quad - \frac{1}{2\sigma^2} \sum_{t=1}^n 2(u_t - \rho u_{t-1}) (-x'_t d\beta - u_{t-1}(d\rho) + \rho x'_{t-1} d\beta) \\ &= \left(\frac{1}{\sigma^2} \sum_{t=1}^n (u_t - \rho u_{t-1})(x'_t - \rho x'_{t-1})\right) d\beta + \left(\frac{1}{\sigma^2} \sum_{t=1}^n (u_t - \rho u_{t-1})u_{t-1}\right) d\rho \\ &\quad - \left(\frac{n}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{t=1}^n (u_t - \rho u_{t-1})^2\right) d\sigma^2 \end{aligned}$$

so that the derivatives are given by

$$\begin{aligned}\frac{\ell(\beta, \rho, \sigma^2)}{d\beta'} &= \frac{1}{\sigma^2} \sum_{t=1}^n (u_t - \rho u_{t-1})(x'_t - \rho x'_{t-1}) \\ \frac{\ell(\beta, \rho, \sigma^2)}{d\rho} &= \frac{1}{\sigma^2} \sum_{t=1}^n (u_t - \rho u_{t-1})u_{t-1} \\ \frac{\ell(\beta, \rho, \sigma^2)}{d\sigma^2} &= - \left(\frac{n}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{t=1}^n (u_t - \rho u_{t-1})^2 \right).\end{aligned}$$

The MLE are characterized by

$$\begin{aligned}\sum_{t=1}^n \left((y_t - \hat{\rho}y_{t-1}) - (x_t - \hat{\rho}x_{t-1})' \hat{\beta} \right) (x_t - \hat{\rho}x_{t-1}) &= 0 \\ \sum_{t=1}^n (\hat{u}_t - \hat{\rho}\hat{u}_{t-1})\hat{u}_{t-1} &= 0 \\ \frac{1}{2(\hat{\sigma}^2)^2} \sum_{t=1}^n (\hat{u}_t - \hat{\rho}\hat{u}_{t-1})^2 &= \frac{n}{2\hat{\sigma}^2},\end{aligned}$$

where $\hat{u}_t = y_t - \hat{\beta}x_t$. Solving the equations further, we have

$$\begin{aligned}\hat{\beta} &= \left(\sum_{t=1}^n (x_t - \hat{\rho}x_{t-1})(x_t - \hat{\rho}x_{t-1})' \right)^{-1} \left(\sum_{t=1}^n (x_t - \hat{\rho}x_{t-1})(y_t - \hat{\rho}y_{t-1}) \right) \\ \hat{\rho} &= \left(\sum_{t=1}^n \hat{u}_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^n \hat{u}_{t-1}\hat{u}_t \right) \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{t=1}^n (\hat{u}_t - \hat{\rho}\hat{u}_{t-1})^2.\end{aligned}$$

Note that $(\hat{\beta}, \hat{\rho}, \hat{\sigma}^2)$ cannot be solved in closed form but we may iterate the three equations until convergence. Conditional on ρ , $(\hat{\beta}, \hat{\sigma}^2)$ are maximizers. Conditional on $\hat{\beta}$, $\hat{\rho}$ solves a least squares problem. The each step of iteration increases the likelihood (because each solves a maximization problem given the others) and convergence is guaranteed. Question: is $\hat{\beta}$ consistent?

4. (Panel Model with Group-Specific Heteroskedasticity)

Consider the panel data model

$$y_{ij} = x'_{ij}\beta + \varepsilon_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, J,$$

where the regressors $x_{ij} \in \mathbb{R}^k$ are fixed and

$$\varepsilon_{ij} \sim \mathbb{N}(0, \sigma_j^2), \quad \sigma_j^2 > 0,$$

with independence across i and j . The variance is allowed to differ across groups j but is constant within each group. The joint log-likelihood function is

$$\ell(\beta, \{\sigma_j^2\}_{j=1}^J) = -\frac{1}{2} \sum_{j=1}^J \left[n_j \log(2\pi) + n_j \log \sigma_j^2 + \frac{1}{\sigma_j^2} \sum_{i=1}^{n_j} (y_{ij} - x'_{ij}\beta)^2 \right].$$

The differential is given by

$$d\ell(\beta, \{\sigma_j^2\}_{j=1}^J) = -\frac{1}{2} \sum_{j=1}^J \left[n_j \frac{1}{\sigma_j^2} d\sigma_j^2 - \left(\frac{1}{(\sigma_j^2)^2} \right) (d\sigma_j^2) \sum_{i=1}^{n_j} (y_{ij} - x'_{ij}\beta)^2 - \frac{1}{\sigma_j^2} \sum_{i=1}^{n_j} 2(y_{ij} - x'_{ij}\beta)x'_{ij} d\beta \right]$$

Differentiating with respect to β yields

$$\frac{\partial \ell}{\partial \beta'} = \sum_{j=1}^J \frac{1}{\sigma_j^2} \sum_{i=1}^{n_j} (y_{ij} - x'_{ij}\beta) x'_{ij}.$$

Setting the derivative equal to zero gives

$$\sum_{j=1}^J \sum_{i=1}^{n_j} \frac{1}{\sigma_j^2} x_{ij} (y_{ij} - x'_{ij}\beta) = 0.$$

Thus, the MLE of β is given by

$$\hat{\beta} = \left(\sum_{j=1}^J \frac{1}{\hat{\sigma}_j^2} \sum_{i=1}^{n_j} x_{ij} x'_{ij} \right)^{-1} \left(\sum_{j=1}^J \frac{1}{\hat{\sigma}_j^2} \sum_{i=1}^{n_j} x_{ij} y_{ij} \right).$$

Differentiating with respect to σ_j^2 gives

$$\frac{\partial \ell}{\partial \sigma_j^2} = -\frac{n_j}{2\sigma_j^2} + \frac{1}{2(\sigma_j^2)^2} \sum_{i=1}^{n_j} (y_{ij} - x'_{ij}\beta)^2.$$

Setting this equal to zero yields

$$\hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - x'_{ij}\hat{\beta})^2, \quad j = 1, \dots, J.$$

This is the maximum likelihood estimator of the group-specific variance. In general, this system has no closed-form joint solution, but it can be solved by iterating between the update for β and the variance updates for σ_j^2 until convergence. Question: is $\hat{\beta}$ consistent?

3 Matrix Calculus

For general matrix to matrix function $F : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times q}$, the partial derivative may be collected into a tensor $\frac{\partial F(X)}{\partial X} = (\frac{\partial F_{kl}(X)}{\partial X_{ij}})_{i,j,k,l=1}^{m,n,p,q} \in \mathbb{R}^{m \times n \times p \times q}$ whose (i, j, k, l) -th element collects the partial derivative $\frac{\partial F_{kl}(X)}{\partial X_{ij}}$. We may also collect the partial derivatives in matrices: $\frac{\partial \text{vec}(F(X))}{\partial \text{vec}(X)'} \text{ or } \frac{\partial \overline{\text{vec}}(F(X))}{\partial \overline{\text{vec}}(X)'} \in \mathbb{R}^{pq \times mn}$. It is customary to deal with the column-wise matrixed version:

$$\frac{\partial \overline{\text{vec}}(F(X))}{\partial \overline{\text{vec}}(X)'} \in \mathbb{R}^{pq \times mn},$$

which can be obtained using the same method as in the vector case by considering $\overline{\text{vec}}(F(X))$ as vector-to-vector function $\overline{\text{vec}}(F) : \mathbb{R}^{mn} \rightarrow \mathbb{R}^{pq}$.

Some of the useful rules of the differential of matrices are given below:

Lemma 2. The differential is a linear operator and obeys the following rules:

(1)

$$dA = 0, \quad d(aX) = adX, \quad d(X') = (dX)', \quad d(X+Y) = dX+dY, \quad d(XY) = (dX)Y+X(dY)$$

(2) For square matrix X ,

$$d\text{tr}(X) = \text{tr}(dX)$$

(3) For nonsingular X ,

$$d|X| = |X|\text{tr}(X^{-1}dX), \quad dX^{-1} = -X^{-1}(dX)X^{-1}$$

and if $|X| > 0$,

$$d \log |X| = \text{tr}(X^{-1}dX)$$

3.1 Examples

1. Find differential of the likelihood function $\ell(\beta, \Omega)$ from the linear model

$$y = X\beta + \varepsilon$$

where $\varepsilon \sim \mathbb{N}(0, \Omega)$ and X is fixed. Note that $y \sim \mathbb{N}(X\beta, \Omega)$ so that

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Omega| - \frac{1}{2} (y - X\beta)' \Omega^{-1} (y - X\beta).$$

The differential is given by

$$\begin{aligned}
d\ell(\beta, \Omega) &= -\frac{n}{2}d\log|\Omega| - (y - X\beta)' \Omega^{-1} d(y - X\beta) - \frac{1}{2}(y - X\beta)' (d\Omega^{-1}) (y - X\beta) \\
&= -\frac{1}{2}\text{tr}(\Omega^{-1}d\Omega) + (y - X\beta)' \Omega^{-1} X(d\beta) + \frac{1}{2}(y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (y - X\beta) \\
&= -\frac{1}{2}\text{tr}(\Omega^{-1}d\Omega) + (y - X\beta)' \Omega^{-1} X(d\beta) + \frac{1}{2}\text{tr}((y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (y - X\beta)) \\
&= (y - X\beta)' \Omega^{-1} X(d\beta) + \frac{1}{2}\text{tr}((- \Omega^{-1} + \Omega^{-1} (y - X\beta)(y - X\beta)' \Omega^{-1}) (d\Omega)) \\
&= (y - X\beta)' \Omega^{-1} X(d\beta) + \frac{1}{2}\text{vec}(- \Omega^{-1} + \Omega^{-1} (y - X\beta)(y - X\beta)' \Omega^{-1})' \text{vec}(d\Omega).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta'} &= (y - X\beta)' \Omega^{-1} X \\
\frac{\partial \ell(\beta, \Omega)}{\partial \text{vec}(\Omega)'} &= -\frac{1}{2}\text{vec}(- \Omega^{-1} + \Omega^{-1} (y - X\beta)(y - X\beta)' \Omega^{-1})'.
\end{aligned}$$

Matricizing $\frac{\partial \ell(\beta, \Omega)}{\partial \text{vec}(\Omega)'}$, we have

$$\frac{\partial \ell(\beta, \Omega)}{\partial \Omega} = -\frac{1}{2}(- \Omega^{-1} + \Omega^{-1} (y - X\beta)(y - X\beta)' \Omega^{-1}).$$

Solving the FOC, we obtain

$$\begin{aligned}
\hat{\beta} &= (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} y \\
\hat{\Omega} &= (y - X\hat{\beta})(y - X\hat{\beta})'.
\end{aligned}$$

However, the equations are not well defined because $\hat{\Omega} = (y - X\hat{\beta})(y - X\hat{\beta})'$ is rank one and hence nonsingular. Note that the number of parameters that we need to estimate $\frac{n(n+1)}{2} + m$ far exceeds the number of observations n . We need to impose further structure on Ω to be able to obtain a sensible estimator.