HW 1

Question 1: EM for the PANSS data

Load the `Surrogate` package in R and load the dataset `Schizo_PANSS`:

```
library(Surrogate)
data("Schizo_PANSS")
```

The dataset combines five clinical trials aimed at determining if risperidone decreases the Positive and Negative Syndrome Score (PANSS) over time compared to a control treatment for patients with schizophrenia. These are longitudinal trials where patients are assessed at weeks 1, 2, 4, 6 and 8 after being assigned to a treatment arm.

Each row in the dataset is a different trial participant, and `Week1`, `Week2`, `Week4`, `Week6`, `Week8` records the change in PANSS from baseline. The variable `Treat` represents whether the patient was enrolled in the control (-1) or if the patient was in the risperidone arm (1).

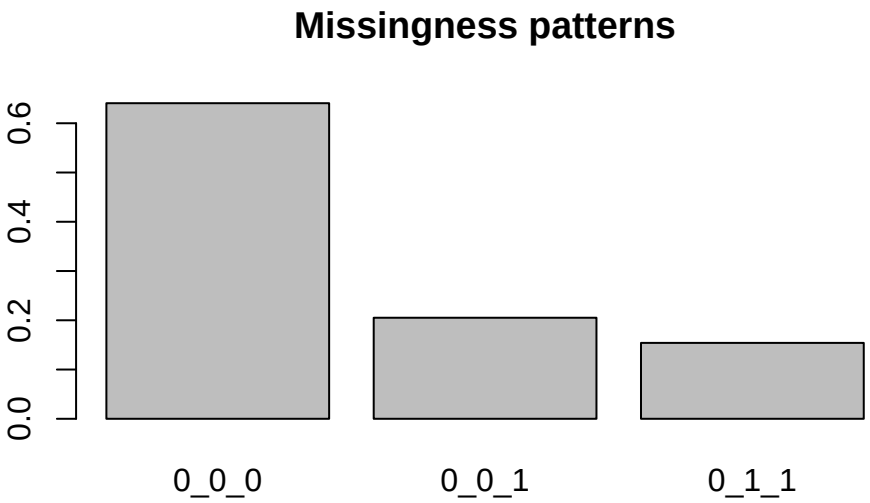
Subset the data to include only `Week1`, `Week4` and `Week8`, and including only complete cases, dropouts between week 4 and week 8, and dropouts between weeks 1 and weeks 4.

```
hw_data <- Schizo_PANSS[,c("Id","Treat","Week1","Week4","Week8")] |>
  subset((!is.na(Week1) & !is.na(Week4) & !is.na(Week8))
         | (!is.na(Week1) & !is.na(Week4) & is.na(Week8))
         | (!is.na(Week1) & is.na(Week4) & is.na(Week8)))
```

Double-checking we did the subsetting correctly:

```
gen_miss_patterns <- function(mat) {
  patterns <- mat |>
    is.na() |>
    apply(2,as.integer) |>
    apply(1, \(row) {
      paste(row,collapse = "_")
    })
  return(patterns)
```

```
}
y_data <- hw_data[,c("Week1","Week4","Week8")]
miss_patterns <- gen_miss_patterns(y_data)
tab_miss <- table(miss_patterns)
prop_miss <- prop.table(tab_miss)
prop_miss <- sort(prop_miss,decreasing = TRUE)
barplot(prop_miss, main = "Missingness patterns")
```



Of the remaining data, 20% of cases dropped out before the final week, and 15% of cases dropped out between weeks 1 and 4.

We're going to fit the following model to this dataset:

$$y_i \mid \text{Treat}_i \sim \text{Normal}(\mu + \beta_1 \text{Treat}_i + \beta_2 t + \beta_3 \text{Treat}_i t, \Sigma)$$

where t is the vector $1, 4, 8$ indicating at what time points the measurements were taken and μ is a scalar mean, under the somewhat dubious assumption of ignorable dropout.

In order to do so, we can use an Expectation-Conditional-Maximization algorithm:

1. Initialize with $\beta^{(1)}, \Sigma^{(1)}$, and a value ϵ
2. For $t = 1, 2, ...$
    a. Compute $\mathbb{E}\left[y_i \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$ and $\mathbb{E}\left[y_i y_i^T \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$ for all i
    b. Update $\beta^{(t)}$ to $\beta^{(t+1)}$

$$\beta^{(t+1)} = \left(\sum_i X_i^T (\Sigma^{(t)})^{-1} X_i\right)^{-1} \sum_i X_i^T (\Sigma^{(t)})^{-1} \mathbb{E}\left[y_i \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$$

    c. Update $\Sigma^{(t)}$ to $\Sigma^{(t+1)}$

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_i \mathbb{E}\left[(y_i - X_i\beta^{(t+1)})(y_i - X_i\beta^{(t+1)})^T \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$$

d. If $Q(\beta^{(t+1)}, \Sigma^{(t+1)} \mid \beta^{(t)}, \Sigma^{(t)}) - Q(\beta^{(t)}, \Sigma^{(t)} \mid \beta^{(t)}, \Sigma^{(t)}) < \epsilon$, stop, otherwise, return to step a.

Compute $Q(\beta, \Sigma \mid \beta^{(t)}, \Sigma^{(t)})$ as:

$$-\frac{n}{2}\log\det\Sigma - \frac{1}{2}\text{tr}\left(\left(\sum_i \mathbb{E}_{y_{i(1)}|y_{i(0)},\beta^t,\Sigma^t}\left[(y_i - X_i\beta)(y_i - X_i\beta)^T\right]\right)\Sigma^{-1}\right)$$

Part a

Fit your model to simulated data

```
set.seed(123)
n <- 1000
p <- 5
K <- 3
X <- list()
beta <- rnorm(p)
L <- matrix(rnorm(K * K),K,K)
Sigma <- L %*% t(L)
phi <- rnorm(p)
y <- list()
### d will hold our indicators for which group each patient is in.
### d = 1: dropout between weeks 1 and 4
### d = 2: dropout between weeks 4 and 8
### d = 3: all values are observed
d <- rep(NA_integer_,n)
for (i in 1:n) {
  X[[i]] <- matrix(rnorm(p * K),K,p)
  y[[i]] <- X[[i]] %*% beta + MASS::mvrnorm(mu = rep(0,K), Sigma = Sigma)
  logit_p_m <- X[[i]] %*% phi + c(-2, -1, 1)
  p_m <- exp(logit_p_m) / sum(exp(logit_p_m))
  d_i <- rmultinom(1, 1, p_m)
  d[i] <- which(as.logical(d_i))
  y[[i]] <- y[[i]][1:d[i]]
}
```

For this model, we will assume that even if you drop out, we still observe your covariates for all time points.

Use the expressions from lecture 11's notes to derive asymptotic standard errors from the observed information for the observed data for your estimates for beta.

This is fairly involved, so let's work through the terms we'll need for the standard errors.

$$I(\theta^\star \mid Y_{(0)} = \tilde{y}_{(0)}) = -\nabla^2_\theta Q(\theta \mid \theta^\star) \mid_{\theta=\theta^\star} \\ - \mathbb{E}_{Y_{(1)}|Y_{(0)}=\tilde{y}_{(0)}} \left[\nabla_\theta \ell_Y(\theta \mid Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) \ell_Y(\theta \mid Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)})^T\right] \mid_{\theta=\theta^\star}$$

Let's start with the score function of the complete data log-likelihood, or the second expression above. We know that $\ell_Y(\beta, \Sigma \mid Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)})$ is:

$$-\frac{n}{2}\log\det\Sigma - \frac{1}{2}\text{tr}\left(\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}\right)$$

We'll need the differential of ℓ_Y. From lecture 3, we have

$$\mathrm{d}\log\det\Sigma = \text{tr}(\Sigma^{-1}\mathrm{d}\Sigma)$$

$$-\frac{n}{2}\text{tr}(\Sigma^{-1}\mathrm{d}\Sigma) + \frac{1}{2}\text{tr}\left(\left(\sum_i (X_i\mathrm{d}\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}\right) \\ + \frac{1}{2}\text{tr}\left(\left(\sum_i (y_i - X_i\beta)(X_i\mathrm{d}\beta)^T\right)\Sigma^{-1}\right) - \frac{1}{2}\text{tr}\left(\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\mathrm{d}\Sigma^{-1}\right)$$

We will repeatedly use the fact that

$$\text{tr}(AB) = \text{tr}(BA),\ \text{tr}(A) = \text{tr}(A^T)\ \text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$$

This implies that for $B = B^T$:

$$\begin{aligned}\text{tr}(A^TB) &= \text{tr}((B^TA)^T) \\ &= \text{tr}(B^TA) \\ &= \text{tr}(AB^T) \\ &= \text{tr}(AB)\end{aligned}$$

We then get

$$-\frac{n}{2}\text{tr}(\Sigma^{-1}\mathrm{d}\Sigma) + \text{tr}\left(\left(\sum_i (y_i - X_i\beta)(X_i\mathrm{d}\beta)^T\right)\Sigma^{-1}\right) - \frac{1}{2}\text{tr}\left(\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\mathrm{d}\Sigma^{-1}\right)$$

We can further simplify the middle term:

$$\begin{aligned}\text{tr}\left(\left(\sum_i (y_i - X_i\beta)(X_i\mathrm{d}\beta)^T\right)\Sigma^{-1}\right) &= \text{tr}\left(\Sigma^{-1}\left(\sum_i (y_i - X_i\beta)(X_i\mathrm{d}\beta)^T\right)\right) \\ &= \sum_i \text{tr}\left(\Sigma^{-1}(y_i - X_i\beta)(X_i\mathrm{d}\beta)^T\right) \\ &= \sum_i \text{tr}\left(X_i^T\Sigma^{-1}(y_i - X_i\beta)\mathrm{d}\beta^T\right) \\ &= \text{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\mathrm{d}\beta^T\right)\end{aligned}$$

And we also have that $\mathrm{d}\Sigma^{-1} = \Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}$, so the full first differential is:

$$-\frac{n}{2}\mathrm{tr}(\Sigma^{-1}\mathrm{d}\Sigma) + \mathrm{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\mathrm{d}\beta^T\right) \\ -\frac{1}{2}\mathrm{tr}\left(\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\right)$$

We can also combine the first and third terms:

$$\begin{aligned}
&-\frac{n}{2}\mathrm{tr}(\Sigma^{-1}\mathrm{d}\Sigma) - \frac{1}{2}\mathrm{tr}\left(\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\right) \\
&\quad = -\frac{n}{2}\mathrm{tr}(\Sigma^{-1}\mathrm{d}\Sigma) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right) \\
&\quad = -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\mathrm{d}\Sigma\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\right)
\end{aligned}$$

This leads to a final first differential of ℓ_Y:

$$-\frac{1}{2}\mathrm{tr}\left(\mathrm{d}\Sigma\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right) + \mathrm{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\mathrm{d}\beta^T\right)$$

The gradient with respect to β is:

$$\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)$$

and the gradient with respect to $\mathrm{vec}(\Sigma)$ is:

$$-\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right)$$

Now we need to compute the following matrix:

$$\begin{bmatrix} I_{\beta\beta} & I_{\beta\Sigma} \\ I_{\beta\Sigma}^T & I_{\Sigma\Sigma} \end{bmatrix}$$

$$I_{\beta\beta} = \mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)^T\right]$$

$$I_{\beta\Sigma} = \mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[-\frac{1}{2}\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\mathrm{vec}\left(\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right.\right.$$

$$I_{\Sigma\Sigma} = \mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[\frac{1}{4}\text{vec}\left(\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right)\right.$$
$$\left.\times \text{vec}\left(\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right)^T\right]$$

The expression for $I_{\beta\beta}$ simplifies to:

$$I_{\beta\beta} = \sum_i \sum_j X_i^T(\Sigma^\star)^{-1}\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[(y_i - X_i\beta^\star)(y_j - X_j\beta^\star)^T\right](\Sigma^\star)^{-1}X_j$$

The expression

$$\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[-\frac{1}{2}\left(\sum_i X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\right)\text{vec}\left((\Sigma^\star)^{-1}\left(n\Sigma^\star - \left(\sum_i (y_i - X_i\beta^\star)(y_i - X_i\beta^\star)^T\right)\right)\right.\right.$$

will simplify because vec is a linear operator, and the term

$$-\frac{1}{2}\left(\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[\sum_i X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\right]\right)\text{vec}\left((\Sigma^\star)^{-1}(n\Sigma^\star)(\Sigma^\star)^{-1}\right)^T$$

is zero because the $\beta^\star$ solves the expected score equation for β:

$$\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[\sum_i X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\right] = 0$$

However, this term:

$$\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[-\frac{1}{2}\left(\sum_i X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\right)\text{vec}\left((\Sigma^\star)^{-1}\left(\sum_i (y_i - X_i\beta^\star)(y_i - X_i\beta^\star)^T\right)(\Sigma^\star)^{-1}\right)^T\right]$$

won't simplify very nicely, though we can use the fact that vec is a linear operator and the following identities:

$$\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B),\ \text{vec}(A + B) = \text{vec}(A) + \text{vec}(B)$$

to give:

$$\mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[-\frac{1}{2}\left(\sum_i X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\right)\sum_i \text{vec}\left((y_i - X_i\beta^\star)(y_i - X_i\beta^\star)^T\right)^T\Sigma^{-1}\otimes\Sigma^{-1}\right]$$

which further simplifies:

$$I_{\beta\Sigma} = \frac{1}{2}\left(\sum_i \sum_j \mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[X_i^T(\Sigma^\star)^{-1}(y_i - X_i\beta^\star)\text{vec}\left((y_j - X_j\beta^\star)(y_j - X_j\beta^\star)^T\right)^T\right]\Sigma^{-1}\otimes\Sigma^{-1}\right)$$

As for the lower diagonal block, the only nonzero term will be the conditional expectation of this:

$$\frac{1}{4}\text{vec}\left(\Sigma^{-1}\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}\right)\text{vec}\left(\Sigma^{-1}\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}\right)^T$$

Using the following identities:

$$\text{vec}(ABC) = (C^T\otimes A)\text{vec}(B),\ \text{vec}(A+B) = \text{vec}(A) + \text{vec}(B)$$

Gives

$$I_{\Sigma\Sigma} = \frac{1}{4}(\Sigma^\star)^{-1}\otimes(\Sigma^\star)^{-1}D(\Sigma^\star)^{-1}\otimes(\Sigma^\star)^{-1}$$

where D is:

$$D = \sum_i\sum_j \mathbb{E}_{Y_{(1)}|Y_{(0)},\beta^\star,\Sigma^\star}\left[\text{vec}((y_i - X_i\beta^\star)(y_i - X_i\beta^\star)^T)\text{vec}((y_j - X_j\beta^\star)(y_j - X_j\beta^\star)^T)^T\right]$$

These expressions are pretty complicated; I would attack this with a Monte Carlo estimator by simulating from the conditional distribution:

$$y_i \mid y_{i(0)}, \beta^\star, \Sigma^\star$$

and taking the sample average of the draws for each of the above expressions.

On to the expression for the second derivative of $Q(\theta \mid \theta^\star)\mid_{\theta=\theta^\star}$. We start with an expression for the first differential from above:

$$-\frac{1}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right) + \text{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}(y_i - X_i\beta)\right)\text{d}\beta^T\right)$$

Now we take the second derivatives:

$$\begin{aligned}
&-\frac{1}{2}\text{tr}\left(\text{d}\Sigma\text{d}\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right)\\
&-\frac{1}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\left(n\text{d}\Sigma - \text{d}\left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\Sigma^{-1}\right)\\
&-\frac{1}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\left(n\Sigma - \left(\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\right)\text{d}\Sigma^{-1}\right)\\
&+\text{tr}\left(\left(\sum_i X_i^T\text{d}\Sigma^{-1}(y_i - X_i\beta)\right)\text{d}\beta^T\right)\\
&-\text{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}X_i\text{d}\beta\right)\text{d}\beta^T\right)
\end{aligned}$$

Before we evaluate these, we can see the first and third expressions will be zero because the first order conditions at which we're evaluating Q will solve:

$$n\Sigma^\star - \sum_i (y_i - X_i\beta^\star)(y_i - X_i\beta^\star)^T = 0$$

The second expression will become:

$$-\frac{1}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\left(n\text{d}\Sigma + 2\left(\sum_i (y_i - X_i\beta)\text{d}\beta^T X_i^T\right)\right)\Sigma^{-1}\right)$$

which simplifies to

$$-\frac{n}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\text{d}\Sigma\right) - \text{tr}\left(\sum_i X_i^T\Sigma^{-1}\text{d}\Sigma\Sigma^{-1}(y_i - X_i\beta)\text{d}\beta^T\right)$$

The fourth line of the second derivative is:

$$\text{tr}\left(\left(\sum_i X_i^T\Sigma^{-1}\text{d}\Sigma\Sigma^{-1}(y_i - X_i\beta)\right)\text{d}\beta^T\right)$$

which cancels with the second term above.

We're left with the following:

$$-\frac{n}{2}\text{tr}\left(\text{d}\Sigma\Sigma^{-1}\text{d}\Sigma\Sigma^{-1}\right) - \text{tr}\left(\text{d}\beta^T\left(\sum_i X_i^T\Sigma^{-1}X_i\right)\text{d}\beta\right)$$

Thus, the Hessian of the Q function evaluated at $\Sigma^\star, \beta^\star$ is

$$\begin{bmatrix} -\sum_i X_i^T(\Sigma^\star)^{-1}X_i & 0 \\ 0 & -\frac{n}{2}(\Sigma^\star)^{-1}\otimes(\Sigma^\star)^{-1} \end{bmatrix}$$

Compute marginal asymptotic confidence intervals for each element of beta and test whether the true values lie in those intervals.

Report how many steps it took for your model to converge.

Part b

Now fit your model to the real data, report the MLEs and the standard errors for your β coefficients.

Part c

Instead of ECM, change your algorithm to an EM algorithm which will involve running your maximization to convergence for each EM step.

1. Initialize with $\beta^{(1)}, \Sigma^{(1)}$, and a value ϵ
2. For $t = 1, 2, ...$
    a. Compute $\mathbb{E}\left[y_i \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$ and $\mathbb{E}\left[y_i y_i^T \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$ for all i
    b. For $s = 1, 2, ...$
        i. At $s = 1$, set $\beta^{(s)} = \beta^{(t)}, \Sigma^{(s)} = \Sigma^{(t)}$
        ii. Update $\beta^{(s)}$ to $\beta^{(s+1)}$

$$\beta^{(s+1)} = \left(\sum_i X_i^T (\Sigma^{(s)})^{-1} X_i\right)^{-1} \sum_i X_i^T (\Sigma^{(s)})^{-1} \mathbb{E}\left[y_i \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$$

        iii. Update $\Sigma^{(s)}$ to $\Sigma^{(s+1)}$

$$\Sigma^{(s+1)} = \frac{1}{n} \sum_i \mathbb{E}\left[(y_i - X_i\beta^{(s+1)})(y_i - X_i\beta^{(s+1)})^T \mid y_{i(0)}, \beta^{(t)}, \Sigma^{(t)}\right]$$

        iv. Iterate until $\beta^{(s)}, \Sigma^{(s)}$ reach a stationary point
    c. Set $\beta^{(t+1)} = \beta^{(s)}, \Sigma^{(t+1)} = \Sigma^{(s)}$
    d. If $Q(\beta^{(t+1)}, \Sigma^{(t+1)} \mid \beta^{(t)}, \Sigma^{(t)}) - Q(\beta^{(t)}, \Sigma^{(t)} \mid \beta^{(t)}, \Sigma^{(t)}) < \epsilon$, stop, otherwise, return to step a.

Fit this model to your simulated dataset, and report how many steps it took for this EM algorithm to converge versus the ECM algorithm.