Outline of covariates matrix calculation

We would like to find a method to connect the outcomes and covariates to return a beter clustering result.

One way is to add covariates in the linear mixed effect model, which is:

old:

$$Y = S(\beta + b) + \epsilon$$

where

- Y is the outcome. The dimension is $n * 1$.
- S is the design matrix for the orthogonal polynomials voer time. The dimension is $n * p$. n is the number of observations in total.
- β presents the fixed effects in the model. Dimension $(p, 1)$. Matrix b and ϵ present the random effect in the model. Dimension of b is $(p, 1)$. Dimension of ϵ is $(n, 1)$

We can re-write it into: $Y = X\beta + \epsilon^{\star}$, where $\epsilon^{\star} = Sb + \epsilon$ shows the random effect and $\epsilon^{\star} \sim N_n(0, V)$. $V = SGS^T + R$. G is the covariates matrix for b. R is the variance matrix for ϵ.

new:

If we add the covarients as a random effect into the LME model, then:

$$Y = S(\beta + b + \gamma AX) + \epsilon$$

We can assume that:

- $b, \gamma AX, \epsilon$ are all independent
- γAX, whose dimension is (q, 1), is from a multivariate normal distribution.

We can also re-write it into: $Y = X\beta + \epsilon^{\star\star}$, where $\epsilon^{\star\star} = Sb + S\gamma AX + \epsilon$ shows the random effect.

I feel they have very similar formats. We can still calculate $f(x)$ in the same way with a different LME.

Algorithm

- Initinal matrix A
- Fit linear mixed effect model with matrix A
- Run convexity-based clustering method
- Compute the purity

Goal

The goal is to get the matrix with the max purity.

LME

The old lme function in R is:

```
fit1 <- lmer(y ~ t1 + I(t1^2) + (t1+I(t1^2)|subj), data = dati, REML = FALSE)
```

We can just add the random effect of covariates inside the model:

```
fit2 <- lmer(y ~ t1 + I(t1^2) + (t1+I(t1^2)|subj) +
               (t1+I(t1^2)|AX), data = dati, REML = FALSE)
```

Therefore, keep everything the same instead of the linear mixed effect model in the *cvxclustr* function.

In this scenario, we can wrap the above into a function:

$$Purity = g(A)$$

Just input the matrix A to combine the covariates X, then we can return a purity.

Therefore, our question becomes to calculate $\underset{A}{\text{argmax}} g(A)$. We may try to use Newton Raphson method in high dimension to solve it.

Newton Raphson method in high dimension

The extreme value for F(X) in the interval (a,b) can be calculated by:

$$X^{(i+1)} = X^{(i)} - H^{-1}\nabla F$$

where,

$$\nabla F = \left[\frac{\partial F}{\partial X_1}, \frac{\partial F}{\partial X_2}, ..., \frac{\partial F}{\partial X_p}\right]^T$$

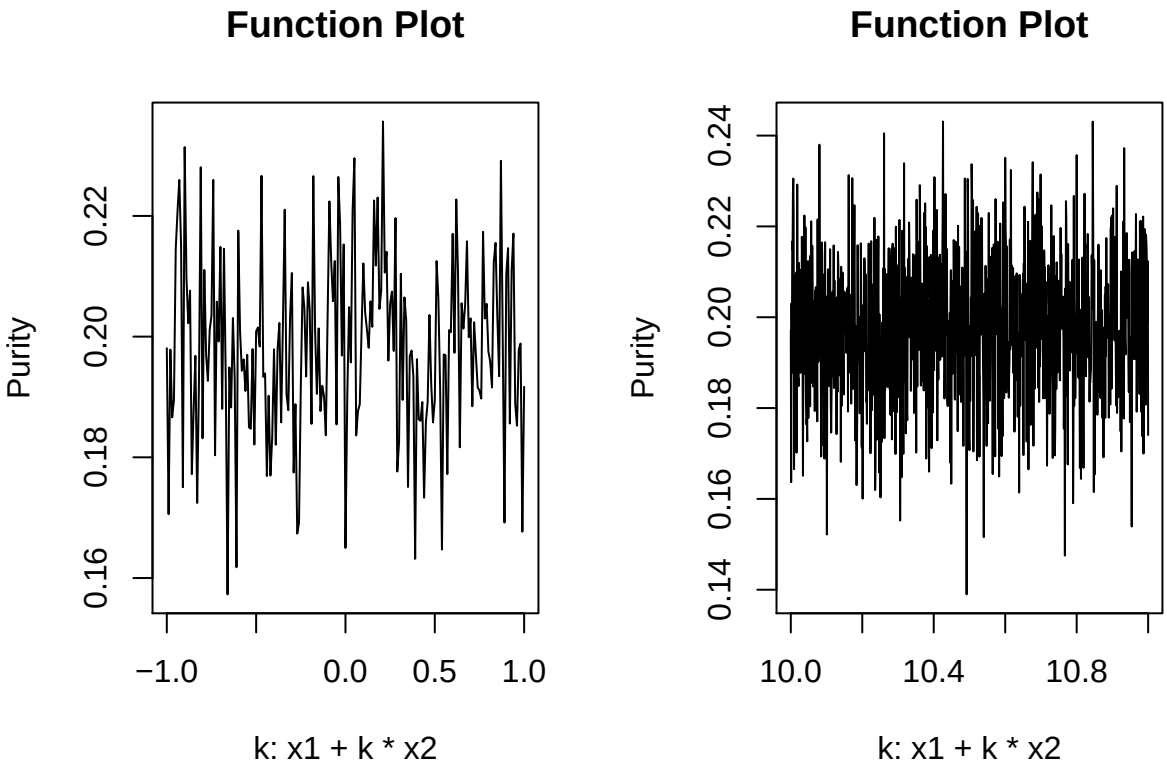
$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 F}{\partial X_1^2} & \frac{\partial^2 F}{\partial X_1 \partial X_2} & \cdots & \frac{\partial^2 F}{\partial X_1 \partial X_p} \\ \frac{\partial^2 F}{\partial X_2 \partial X_1} & \frac{\partial^2 F}{\partial X_2^2} & \cdots & \frac{\partial^2 F}{\partial X_2 \partial X_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 F}{\partial X_p \partial X_1} & \frac{\partial^2 F}{\partial X_p \partial X_2} & \cdots & \frac{\partial^2 F}{\partial X_p^2} \end{pmatrix}$$

If the iteration does not coverage, then the extreme value should be at the edge of the interval.

Back to our scenario, purity is bounded by [0,1]. Therefore, the g(A) should have a max value, which is not at the boundary, since the boundary of A is infinity.

However, I iterated the Newton method for 100 times, the function was very hard to coverage.

Here are the plots about what does the function look like. (at input around c(1,1) and around c(1,10))



It seems that this is not a good function. We probability cannot get a max value through Newton method.