# Empirical Likelihood-based Analysis of Variance Component in Linear Mixed-Effects Models

Jingru Zhang, Haochang Shou and Hongzhe Li

University of Pennsylvania

## Introduction

Linear mixed-effects (LME) models are widely used in analyzing repeated measurement and longitudinal data. Although statistical inference of the fixed effects is well studied, inference of the variance component is rarely explored, which often requires strong distributional assumptions on the random effects and errors.

**Question:** How to do distribution-free inference of the variance component in LME models?

## Problem setup

- $n$ subjects.
  For the $i$th subject, $n_i$ repeated measurements.
  For each repeated measure, data are collected at time $t = s_1, s_2, \cdots, s_m$.
- For the $i$th subject at time $t$, we observe
  a response vector $y_i(t) \in R^{n_i}$,
  an $n_i \times p$ design matrix $X_i$ for the fixed effects $\beta(t) \in R^p$,
  $d$ $n_i \times n_i$ semi-positive design matrices $\Phi_{iq}$ $(q = 1, \cdots, d)$ for the variance components $\theta^*(t) \in (R_+ \cup \{0\})^d$.

## LME model

A general setting of the linear mixed-effects model:

$$y_i(t) = X_i\beta(t) + r_i(t),\ i = 1, \cdots, n,$$

where $r_i(t) \in R^{n_i}$ is a zero-mean random variable with variance $H_i(\theta^*(t))$.

We consider $H_i(\theta^*(t))$ with a linear structure, i.e.,

$$H_i(\theta^*(t)) = \sum_{q=1}^{d} \theta_q^*(t)\Phi_{iq},$$

$$\theta^*(t) = (\theta_1^*(t), \cdots, \theta_d^*(t))^T \doteq (\theta_1^*(t), \theta_{(1)}^*(t)^T)^T.$$

- In this general setting, we do not specify any distribution for the data.
- The data $y_i(t)$ are independent over different subjects $i$, while they are allowed to be non-independent over $t$.

## Testing problems

1. Local testing problem $H_0 : \theta_1^*(t) = \theta_1^0(t)$ at a given $t$.
2. Global testing problem $H_0 : \theta_1^*(t) \equiv \theta_1^0$, $t \in [t_1, t_2]$.

## Local test

- When $\beta(t)$ is unknown, suppose $\hat{\beta}(t)$ is an unbiased estimator of $\beta(t)$ based on all the data.
- Let $R_i(t) = r_i(t)r_i(t)^T$. Since $var(r_i(t)) = H_i(\theta^*(t))$, we have

$$R_i(t) = H_i(\theta^*(t)) + \delta_i(t) = \sum_{q=1}^{d} \theta_q^*(t)\Phi_{iq} + \delta_i(t),$$

  where $E(\delta_i(t)) = 0$ and $\text{var}(\delta_i(t))$ exists.
- For $i = 1, \cdots, n$, let

$$\hat{r}_i(t) = y_i(t) - X_i\hat{\beta}(t) = r_i(t) + X_i(\beta(t) - \hat{\beta}(t)),$$
$$\hat{R}_i(t) \doteq \hat{r}_i(t)\hat{r}_i(t)^T = H_i(\theta^*(t)) + \delta_i(t) + \hat{\epsilon}_i(t),$$

- Let $\Xi$ be a $d \times d$ symmetric matrix with the $(k, l)$th element $\Xi_{kl} = \sum_{i=1}^n \text{tr}(\Phi_{ik}\Phi_{il})$.
  For each $t$, let $\hat{\Upsilon}(t)$ be a $d$-dimensional vector with the $k$th element $\hat{\Upsilon}_k(t) = \sum_{i=1}^n \text{tr}(\Phi_{ik}\hat{R}_i(t))$.
- We define

$$\hat{Z}_i(\theta_1(t)) = \text{tr}\left(\Phi_{i1}\left(\hat{R}_i(t) - \Phi_{i1}\theta_1(t) - \sum_{q=2}^{d}\hat{\theta}_q(t)\Phi_{iq}\right)\right),$$

  where

$$\hat{\theta}_{(1)}(t) \doteq (\hat{\theta}_2(t), \cdots, \hat{\theta}_q(t))^T = (\Xi^{-1})_{-1}^T\hat{\Upsilon}(t).$$

- The empirical likelihood ratio is defined by

$$S(\theta_1(t)) = \max_{p_i}\left\{\prod_{i=1}^{n}(np_i) | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\hat{Z}_i(\theta_1(t)) = 0\right\}.$$

## Global test

- A maximally selected empirical likelihood ratio statistic:

$$\Gamma = \sup_{t\in[t_1,t_2]}\left[\hat{c}_n(\theta_1^0)\left(-2\log\frac{S(\theta_1^0)}{S(\hat{\theta}_1(t))}\right)\right].$$

- Rewrite $\Xi$ as $\Xi = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22}\end{pmatrix}$ with $E_{11}$ being a scalar.
  Let $F = E_{22}^{-1}E_{21} = (F_1, \cdots, F_{d-1})^T$.
- We can show that

$$\Gamma = \sup_{t\in[t_1,t_2]} ER(t) + o_p(1),$$

  where

$$ER(t) = \begin{cases} \frac{(n^{-1/2}\sum_{i=1}^n \hat{D}_i(t))^2}{\hat{\nu}_{1n}^2(t)} I(\sum_{i=1}^n \hat{D}_i(t) \geq 0), & \text{if } \theta_1^0 = 0, \\ \frac{(n^{-1/2}\sum_{i=1}^n \hat{D}_i(t))^2}{\hat{\nu}_{1n}^2(t)}, & \text{if } \theta_1^0 > 0. \end{cases}$$

  Here,
  $\hat{D}_i(t) = \alpha^{-1}\left\langle \Phi_{i1} - \sum_{q=1}^{d-1} F_q\Phi_{iq+1}, \hat{R}_i(t) - \theta_1^0\Phi_{i1}\right\rangle$
  are asymptotically independent.
- For each permutation $g$ $(g = 1, \cdots, G)$, let

$$ER^{(g)}(t) = \begin{cases} \frac{(n^{-1/2}\sum_{i=1}^n \hat{D}_i(t)\xi_i^{(g)})^2}{\hat{\nu}_{1n}^2(t)} I(\sum_{i=1}^n \hat{D}_i(t)\xi_i^{(g)} \geq 0), & \text{if } \theta_1^0 = 0, \\ \frac{(n^{-1/2}\sum_{i=1}^n \hat{D}_i(t)\xi_i^{(g)})^2}{\hat{\nu}_{1n}^2(t)}, & \text{if } \theta_1^0 > 0, \end{cases}$$

$$\Gamma^{(g)} = \sup_{t\in[t_1,t_2]} ER^{(g)}(t),$$

  where $\xi_i^{(g)}$ are i.i.d. standard normal distributed.
- The $p$-value of $\Gamma$ can be approximated by

$$\hat{p} = \frac{1}{G}\sum_{g=1}^{G} I(\Gamma^{(g)} > \Gamma).$$

## Theorem

**Condition 1.** As $n \to \infty$, $P(0 \in ch\{\hat{Z}_1(\theta_1^0(t)), \cdots, \hat{Z}_n(\theta_1^0(t))\}) \to 1$, where $ch\{\}$ is the convex hull.
**Condition 2.** The expectation $E\|r_i(t)\|_2^{4+\gamma_1}$ are bounded uniformly for some $\gamma_1 > 0$.
**Condition 3.** $E(\hat{\epsilon}_i(t)) = O(n^{-\gamma_2/2})$; $\text{cov}(r_i(t)r_i(t)^T, \hat{\epsilon}_j(t)), \text{cov}(\hat{\epsilon}_i(t), \hat{\epsilon}_j(t)) = O(n^{-\gamma_2})$, $i \neq j$, for some $\gamma_2 > 1$.

Let $\hat{\theta}_1(t) = \arg\max_{\theta_1(t)\geq 0} S(\theta_1(t))$. Let $\hat{c}_n(\theta_1^0(t)) = \hat{\nu}_{2n}^2(\theta_1^0(t))/\hat{\nu}_{1n}^2(\theta_1^0(t))$, where $\hat{\nu}_{1n}^2(\theta_1^0(t))$ is a consistent estimator of the asymptotic variance of $n^{-1/2}\sum_{i=1}^n \hat{Z}_i(\theta_1^0(t))$ and $\hat{\nu}_{2n}^2(\theta_1^0(t)) = n^{-1}\sum_{i=1}^n \hat{Z}_i^2(\theta_1^0(t))$.
If $\theta_{(1)}^*(t) \in R_+^{d-1}$, then under Conditions 1–3, as $n \to \infty$,

$$\hat{c}_n(\theta_1^0(t))\left(-2\log\frac{S(\theta_1^0(t))}{S(\hat{\theta}_1(t))}\right) \xrightarrow{d} \mathcal{X}_1^2$$

when $\theta_1^0(t) > 0$, and

$$\hat{c}_n(0)\left(-2\log\frac{S(0)}{S(\hat{\theta}_1(t))}\right) \xrightarrow{d} U_+^2,$$

where $U \sim N(0, 1)$ and $U_+ = \max(U, 0)$.

## Real application

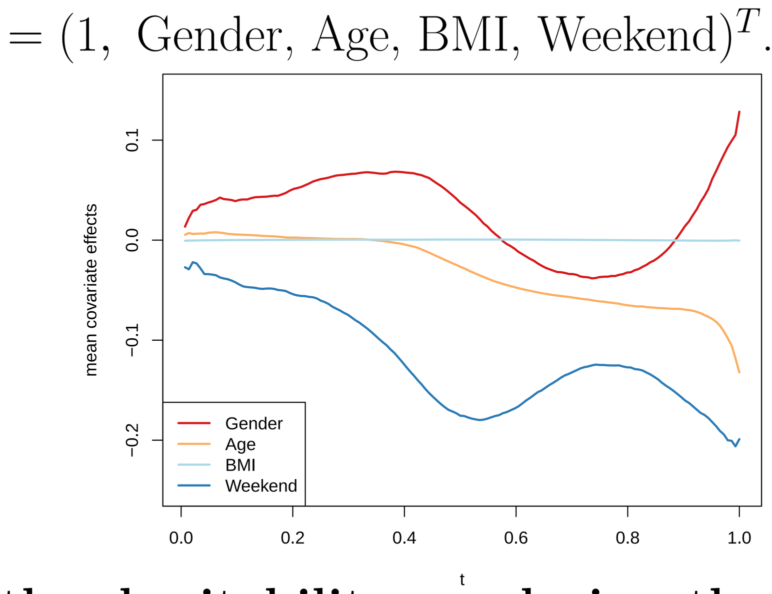
- 298 healthy twins: 126 monozygotic (MZ) twins and 172 dizygotic (DZ) twins.
- The subjects wore actigraphy to track their physical activities for 2 weeks.
- We rescaled and transformed the minute-level ENMO values (1440-dimensional vector per day) as follows:
  For the $j$-th measurement (day) from the subject $i$, the raw data $\xi_{ij} = (\xi_{ij1}, \cdots, \xi_{ij1440})^T$ from the wearable device were transformed by using

$$\tilde{\xi}_{ij} = \log(9250 \cdot \xi_{ij} + 1).$$

- Define the $t$-quantile of activity counts by

$$y_{ij}(t) = \tilde{\xi}_{ij}^{[1440\cdot t]}, \quad t = 1/144, 2/144, \cdots, 144/144,$$

  where $\tilde{\xi}_{ij}^{[s]}$ denotes the $s$-th order statistic of $\tilde{\xi}_{ij}$.
- $x_{ij} = (1,\ \text{Gender, Age, BMI, Weekend})^T$.




**In the heritability analysis, the linear variance structure can be constructed straightforwardly. Whether there is significant genetic effects is of most interest.**

1. Local test $H_0 : \theta_1^*(t) = 0$.

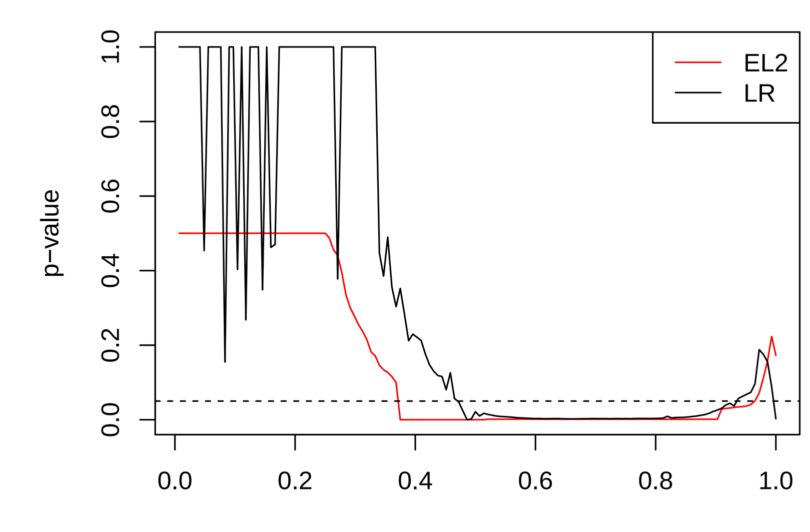



Figure 1: The $p$-values of the proposed local test (EL2) and the likelihood ratio test (LR). The null hypothesis is rejected if $t \in [0.375, 0.958]$ for EL2 and $t \in [0.472, 0.931]$ for LR at the 0.05 significance level.

2. Global test $H_0 : \theta_1^*(t) \equiv 0,\ t \in [0, 1]$.
   The $p$-value is 0 when applying the proposed global test (gEL2).
   We further examine the interval of heritable percentile ranges by setting the scanning lengths 8, 9, 10, 11, 12, and apply gEL2 to the candidate intervals.
   The proposed gEL2 identifies the heritable interval of percentiles between $t \in [0.354, 0.903]$.