Identifiability or
"When is your model too big for your data"

Mixed or longitudinal model for ith cluster

$$\underset{\sim}{Y}_i = X_i \underset{\sim}{\gamma} + Z_i \underset{\sim}{u}_i + \underset{\sim}{\varepsilon}_i$$

$\underset{\sim}{u}_i \sim N(\underset{\sim}{0}, G)$ indep. of $\underset{\sim}{\varepsilon}_i \sim N_{n_i}(\underset{\sim}{0}, R_i)$

$i = 1, \dots, K$

For the fixed part of the model, i.e. $X_i$'s, the question is similar to 3330.
You need to consider the entire X matrix

$$\underset{\sim}{Y} = X\underset{\sim}{\gamma} + Z\underset{\sim}{u} + \underset{\sim}{\varepsilon}$$

where $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix}$, $Z = \begin{bmatrix} Z_1 & 0 & & \\ 0 & Z_2 & & \\ & & \ddots & 0 \\ & & 0 & Z_K \end{bmatrix}$,

$\underset{\sim}{u} = \begin{bmatrix} \underset{\sim}{u}_1 \\ \vdots \\ \underset{\sim}{u}_K \end{bmatrix}$, $\underset{\sim}{\varepsilon} = \begin{bmatrix} \underset{\sim}{\varepsilon}_1 \\ \vdots \\ \underset{\sim}{\varepsilon}_K \end{bmatrix}$

- Use: vif(fit) in the 'car' package to get collinearity diagnostics.
- DO NOT MECHANICALLY APPLY A SIMPLISTIC RULE TO DECIDE WHAT TO DO!!

- UNDERSTAND THE REASON FOR COLLINEARITY BEFORE ACTING.

  e.g. Important confounders may be highly collinear with a causal variable. If you drop the confounder you will have a biased estimate of the causal effect.

- NEVER FOLLOW RECIPES WITHOUT UNDERSTANDING THE CONSEQUENCES OF YOUR CHOICES.

REMINDER:

PRIME DIRECTIVES:

1) KNOW THE PURPOSE OF YOUR ANALYSIS:
   - a) CAUSAL / EXPLANATORY
   - versus b) FINDING A PREDICTIVE ALGORITHM
   - c) DESCRIPTIVE (often an excuse for ignoring the real purpose)

   what difference does it make?

2) THE NATURE OF YOUR DATA
   - NOT JUST WHAT IT LOOKS LIKE BUT HOW WAS IT OBTAINED?
   - WAS THERE:
     - RANDOM ASSIGNMENT?
     - RANDOM SELECTION?

   what's the difference and why does it matter?

---

Example: - Longitudinal study with 2 time points $T = -1, 1$

Suppose we fit:

lme( $Y \sim 1 + X$ , data, random = $\sim 1 + X$ | id)

$$\underset{\sim}{Y}_i = X_i \underset{\sim}{\gamma} + Z_i \underset{\sim}{u}_i + \underset{\sim}{\varepsilon}_i$$

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & +1 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ 1 & +1 \end{bmatrix} \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{bmatrix}$$

$$\begin{pmatrix} u_{i0} \\ u_{i1} \end{pmatrix} \sim N\left( \underset{\sim}{0}, \begin{bmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{bmatrix} \right)$$

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim N\left( \underset{\sim}{0}, \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

---

The variance model has 4 independent parameters

$$g_{00},\ g_{01} = g_{10},\ g_{11},\ \sigma^2$$

They are not directly observable and can only be estimated through the observable

$$\mathrm{Var}(\underset{\sim}{Y}_i) = \mathrm{Var}(X_i \underset{\sim}{\gamma} + Z_i \underset{\sim}{u}_i + \underset{\sim}{\varepsilon}_i)$$

$$= Z_i G Z_i' + \sigma^2 I$$

$$\begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & +1 \end{bmatrix} \begin{bmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & +1 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} g_{00} - 2g_{01} + g_{11} + \sigma^2 & g_{00} - g_{11} \\ g_{00} - g_{11} & g_{00} + 2g_{01} + g_{11} + \sigma^2 \end{bmatrix}$$

Problem?

$$\begin{pmatrix} \text{3 observable} \\ \text{functions} \\ \text{of variance} \end{pmatrix} = f\begin{pmatrix} \text{4 independent} \\ \text{parameters} \end{pmatrix}$$

So: model is overparametrized and there is no hope of solving for unique optimal values of the G and R parameters.

In practice:

This might not matter if your goal is to estimate $\underset{\sim}{\gamma}$ since $Var(\hat{\gamma})$ depends on V

$$\underset{\sim}{Y} = X\gamma + Z\underset{\sim}{u} + \underset{\sim}{\varepsilon}$$

$$Y = X\gamma + \underset{\sim}{\delta}$$

$$Var(\underset{\sim}{\delta}) = ZGZ^T + R = V$$

GLS est $\quad (X'V^{-1}X)(X'V^{-1}Y)$