Correct specification of design matrices in Linear Mixed effects Models: Tests with graphical representation. Supplementary material

Jakob Peterlin, Nataša Kejžar, Rok Blagus

Contents

1 Proofs of theoretical results

In this section the proofs of theoretical results are given. First we prove that $\mathrm{P} f_t^O = 0$ holds even if the random vectors $\boldsymbol{\epsilon}_i$ are not multivariate normal if the matrix $\boldsymbol{Z}_i$ has only one column, which is a multiple of $\mathbf{1}$.

Fix $\boldsymbol{\nu}_i$ and therefore also n_i. Let the permutation matrix $\boldsymbol{\Pi}$ be defined as $\{\boldsymbol{\Pi}\}_{i+1,i} = 1, i = 1, \dots n_i - 1$, $\{\boldsymbol{\Pi}\}_{1,n_i} = 1$. In other words the permutation matrix $\boldsymbol{\Pi}$ is a cyclical permutation of the standard basis. Let the parameter λ be equal to $\lambda = \infty$.

Fix D_i. Now observe that it holds for any $l = 1, \dots, n_i$

$$\begin{aligned} E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \boldsymbol{e}_i^O \big| D_i, \boldsymbol{y}_i^O\big) =& E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \boldsymbol{P}_i \boldsymbol{\epsilon}_i \big| D_i, \boldsymbol{b}_i, (\boldsymbol{I} - \boldsymbol{P}_i)\, \boldsymbol{\epsilon}_i\big) \\ =& E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \boldsymbol{P}_i \boldsymbol{\Pi}^l \boldsymbol{\epsilon}_i \big| D_i, \boldsymbol{b}_i, (\boldsymbol{I} - \boldsymbol{P}_i)\, \boldsymbol{\epsilon}_i\big). \end{aligned} \tag{1}$$

The first equality follows from the definition of $\boldsymbol{e}_i^O$ and $\boldsymbol{y}_i^O$ and the second equality follows from the fact that the components of $\boldsymbol{\epsilon}_i$ are i.i.d.

Furthermore it holds that because $\boldsymbol{Z}_i$ contains a column that is collinear with $\mathbf{1}$, that

$$\sum_{l=1}^{n_i} \boldsymbol{P}_i \boldsymbol{\Pi}^l \boldsymbol{\epsilon}_i = \mathbf{0}. \tag{2}$$

And therefore by (1) and (2)

$$\begin{aligned} n_i E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \boldsymbol{e}_i^O \big| D_i, \boldsymbol{y}_i^O\big) =& \sum_{l=1}^{n_i} E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \boldsymbol{P}_i \boldsymbol{\Pi}^l \boldsymbol{\epsilon}_i \big| D_i, \boldsymbol{b}_i, (\boldsymbol{I} - \boldsymbol{P}_i)\, \boldsymbol{\epsilon}_i\big) \\ =& E\left(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, t\big)^T \left(\sum_{l=1}^{n_i} \boldsymbol{P}_i \boldsymbol{\Pi}^l \boldsymbol{\epsilon}_i\right) \middle| D_i, \boldsymbol{b}_i, (\boldsymbol{I} - \boldsymbol{P}_i)\, \boldsymbol{\epsilon}_i\right) \\ =& \mathbf{0}. \end{aligned}$$

Next we show a counterexample where in the absence of normality $\mathrm{P} f_t^O = 0$ does not hold if $\boldsymbol{Z}_i$ has only one column, which is not a multiple of $\mathbf{1}$. Let $\beta = 1.35$, $n_i = 2 = n_{\max}$ and let the matrices $\boldsymbol{X}_i \in \mathbb{R}^{2\times 1}$ and $\boldsymbol{Z}_i \in \mathbb{R}^{2\times 1}$ be equal and be given by

$$\boldsymbol{X}_i = \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix},$$

where $X_{i,1}$ and $X_{i,2}$ are i.i.d. discrete random variables, defined by

$$P(X_{i,1} = 2) = \frac{1}{2},$$
$$P(X_{i,1} = 1) = \frac{1}{2}.$$

Additionally let the i.i.d. random variables b_i be independent of $\boldsymbol{X}_i, \boldsymbol{Z}_i$ and distributed as a uniform random variable in the interval $[-\frac{1}{2}, \frac{1}{2}]$. Let ϵ_i be i.i.d. random variables that are independent of $\boldsymbol{X}_i, \boldsymbol{Z}_i, b_i$ and distributed uniformly on $(-\frac{1}{2}, \frac{1}{2})$. Let $\lambda = \infty$; recall that in this case for any $y \in \mathbb{R}^2$, $\boldsymbol{\sigma}_\lambda(y, t)$ is equal to $[a, b]$, where a, b are either zero or one and where we used the notation $[x, y]$ to denote a row vector, $[x, y] \in \mathbb{R}^{1\times 2}$.

On the figure 1 we can see, marked by a cross, every possible value of $\boldsymbol{X}_i\beta$ shown in a separate plot. The corresponding possible values of $\boldsymbol{y}_i^O$ are shown as the red line. The gray area corresponds to all possible values of $\boldsymbol{\epsilon}_i$, the yellow area corresponds to the area where the value of the second component of the vector $\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, 1\big)$ is equal to 1, the green area corresponds to the area where the first component of $\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, 1\big)$ is equal to 1 and the blue area corresponds to the area where both components of $\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O, 1\big)$ are equal to 1.

We will now prove that in this case

$$\mathrm{P} f_1^O \neq 0.$$

We will do this by calculating the integral. From the calculation that follows and the countinuity of $t \mapsto \mathrm{P} f_t^O$ in this example it is clear that the inequality $\mathrm{P} f_t^O \neq 0$ holds in at least a small neighborhood of 1.

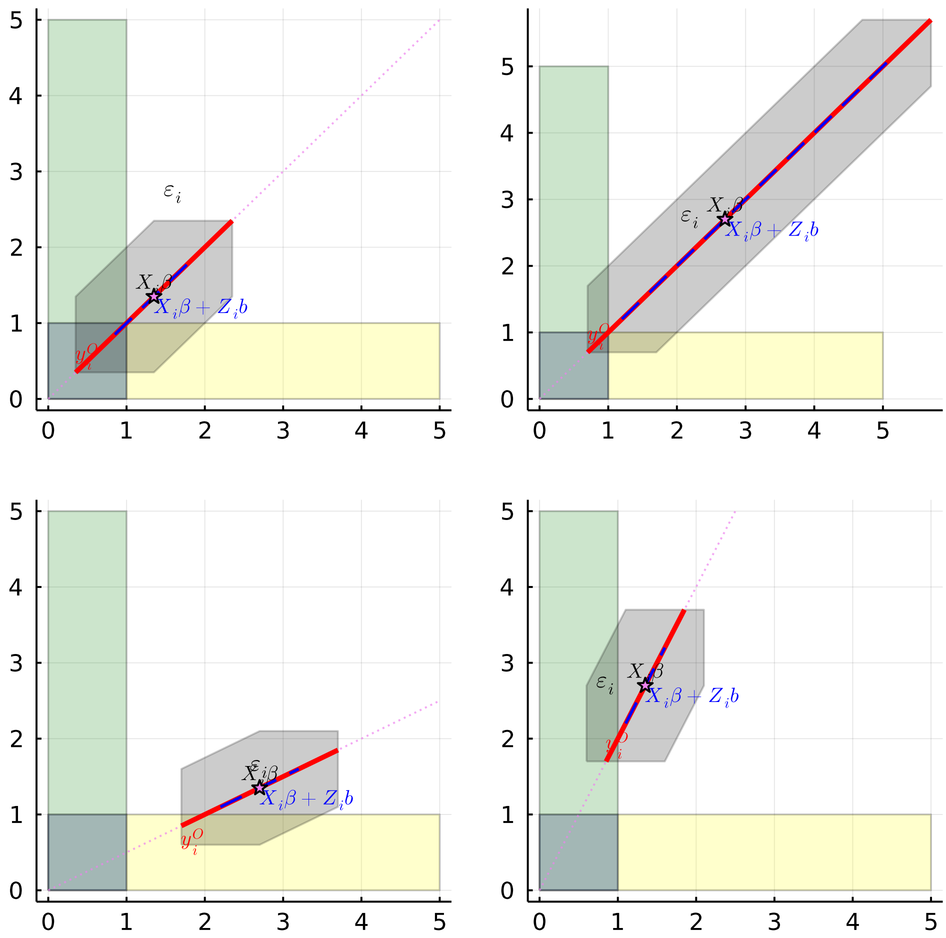


Figure 1: Four plots that are showing the possible values of $\boldsymbol{y}_i^O$, conditionally on the values of $\boldsymbol{X}_i$.

Using the law of total expectation we get

$$\begin{aligned}
\mathrm{P}\, f_1^O =& P(X_1=1,X_2=1)E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O,1\big)\boldsymbol{e}_i^O\big|X_1=1,X_2=1\big)\\
&+P(X_1=2,X_2=2)E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O,1\big)\boldsymbol{e}_i^O\big|X_1=2,X_2=2\big)\\
&+P(X_1=2,X_2=1)E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O,1\big)\boldsymbol{e}_i^O\big|X_1=2,X_2=1\big)\\
&+P(X_1=1,X_2=2)E\big(\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^O,1\big)\boldsymbol{e}_i^O\big|X_1=1,X_2=2\big)\\
=&\frac{1}{4}E\big([1,1]\cdot\boldsymbol{e}_i^O\big|X_1=1,X_2=1\big)\\
+&\frac{1}{4}E\big([1,1]\cdot\boldsymbol{e}_i^O\big|X_1=2,X_2=2\big)\\
+&\frac{1}{4}E\big([0,1]\cdot\boldsymbol{e}_i^O\big|X_1=2,X_2=1\big)\\
+&\frac{1}{4}E\big([1,0]\cdot\boldsymbol{e}_i^O\big|X_1=1,X_2=2\big)
\end{aligned} \tag{3}$$

The justification for the last equality is clear as soon as one realizes that the four summands correspond to the four plots in the figure 1.

Now observe that each component of the projection of $\boldsymbol{e}_i^O$ on the space spanned by $\boldsymbol{P}_i$ is symmetric around 0. Therefore the first and the second summand are equal to zero. This might also help us to understand the case when the matrices $\boldsymbol{Z}_i$ are always assumed to have only one column, which is collinear with $\mathbf{1}$.

Note that the third and the fourth summand in (3) are equal. We will therefore calculate only the third summand. We will do this by additionally conditioning on b_i first, and then integrate by b_i. The area which we are integrating is shown, for 4 possible values of b_i, on figure 2 in gray (the values of ϵ_i over which we are not integrating are shown as red area).

For the ease of calculation we will now scale our coordinates and use the same coordinates as in figure 2. Note that in this way, because the components of $\boldsymbol{\epsilon}_i$ are i.i.d. and uniform, we are integrating the second component of the projection of the triangle shown in gray, on the vector given by $y=-2x+1$.

Note that the value of b_i determines the point at which the hypothenuse of this triangle intersects the y axis. Define the y coordinate of this intersection as y_b.

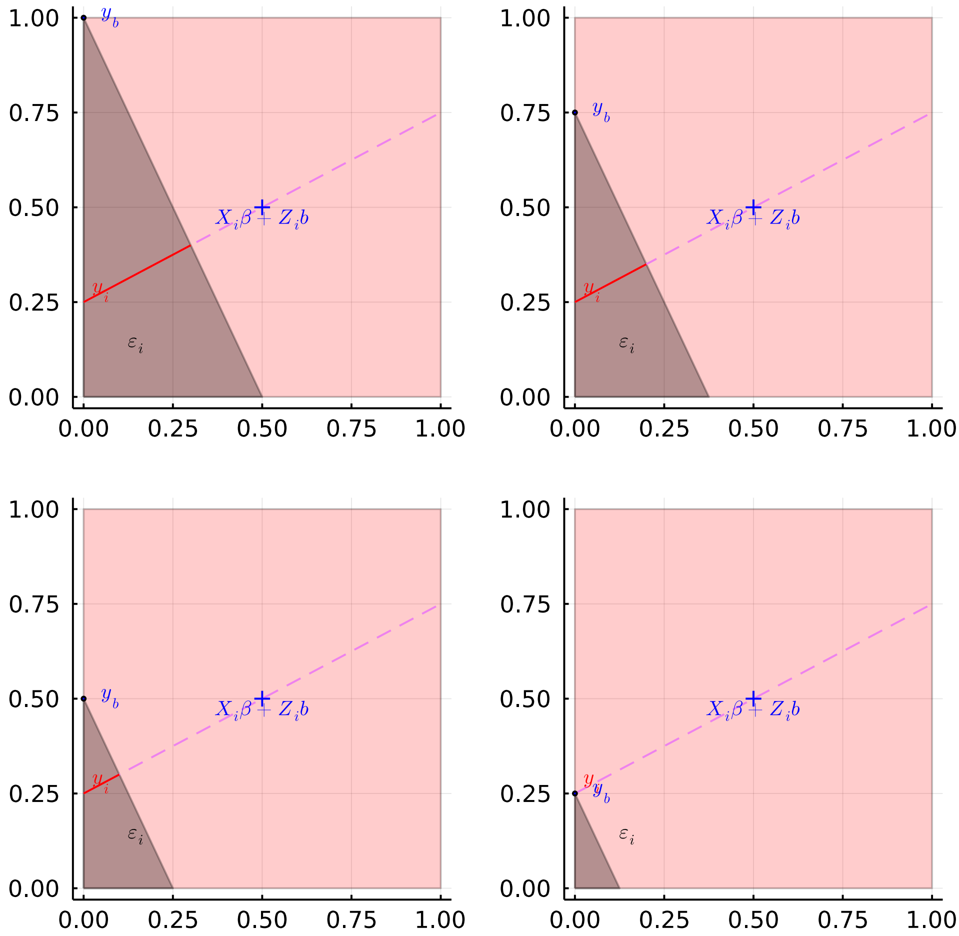


Figure 2: Plot showing various different areas of integration for the third summand of (3), conditionally on a few different values of b_i. The corresponding values of b_i are $-\frac{1}{2}, -\frac{5}{12}, -\frac{1}{3}$ and $-\frac{1}{4}$.

Observe also that the projection of all points for which $y < \frac{y_b+1}{5}$ is negative and non-negative otherwise.

Define the function d_b as

$$d_b(y) = \begin{cases} \sqrt{5}\frac{y-y_b}{4}, & y > \frac{1}{5}\left(y_b+1\right), \\ \sqrt{5}y, & y \leq \frac{1}{5}\left(y_b+1\right), \end{cases}$$

For a fixed b_i, the integral is therefore equal to

$$\begin{aligned} E\big([0,1]\boldsymbol{e}_i^O \big| X_1 = 2, X_1 = 1, b_i\big) &= \int_0^{y_b} \left(d_b(y) - \frac{1}{5}\left(y_b+1\right)\right) dy \\ &= \frac{\sqrt{5}}{100}\left(14y_b^2 - 12y_b - 1\right) \end{aligned}$$

We can now calculate the desired integral, by integrating over $y_b \in [\frac{1}{4}, 1]$ and taking into account that this corresponds to $b_i \in [-\frac{1}{2}, -\frac{1}{4}]$, we get

$$\begin{aligned} E\big([0,1]\cdot\boldsymbol{e}_i^O \big| X_1 = 2, X_2 = 1\big) &= \frac{1}{4}\int_{\frac{1}{4}}^{1} \frac{\sqrt{5}}{100}\left(14y_b^2 - 12y_b - 1\right) dy_b \\ &= \sqrt{5}\frac{57}{12800}. \end{aligned}$$

The (3) is therefore equal to

$$\mathrm{P}\, f_1^O = \sqrt{5}\frac{57}{25600}$$

Next we give the proofs of the theoretical results presented the paper.

Proof of Theorem 1. Observe that the domain of functions in $\mathcal{F}^O$ and $\mathcal{F}^F$ — that is $\mathcal{X}$ can be thought of as a subset of $\mathbb{R}^{1+n_{\max}(k+m+1)}$.

Let $\mathcal{I}_t^O$ denote the set of all values of X_i such that at least one component of $\boldsymbol{y}_i^O$ is less than t and let $\mathcal{I}_t^F$ denote all possible values of X_i such that at least one component of $\boldsymbol{y}^F$ is less than t. For $t \in \mathbb{R}$ define also the sets $\mathcal{Y}_t^O$ and $\mathcal{Y}_t^F$ as

$$\begin{aligned} \mathcal{Y}_t^O &= \left\{f_t^O(X), X \in \mathcal{I}_t^O\right\}, \\ \mathcal{Y}_t^F &= \left\{f_t^F(X), X \in \mathcal{I}_t^F\right\}. \end{aligned}$$

Remember that a subgraph Γf of function $f : X \to \mathbb{R}$, where X is a domain of f, is defined as

$$\{(x,t), x \in X, t \in \mathbb{R}, t < f(x)\}.$$

Let S^C denote the complement of a set S. When $\lambda < \infty$, the subgraphs Γf_t^O and Γf_t^F of f_t^O and f_t^F are equal to

$$\Gamma f_t^O = \left(\mathcal{I}^O_{t+\frac{1}{\lambda}} \times \mathcal{Y}^O_{t+\frac{1}{\lambda}}\right) \bigcup \left(\left(\mathcal{I}^O_{t+\frac{1}{\lambda}}\right)^C \times \mathbb{R}_-\right)$$
$$\Gamma f_t^F = \left(\mathcal{I}^F_{t+\frac{1}{\lambda}} \times \mathcal{Y}^F_{t+\frac{1}{\lambda}}\right) \bigcup \left(\left(\mathcal{I}^F_{t+\frac{1}{\lambda}}\right)^C \times \mathbb{R}_-\right).$$

When $\lambda = \infty$, the subgraphs Γf_t^O and Γ_t^F are equal to

$$\Gamma f_t^O = \left(\mathcal{I}_t^O \times \mathcal{Y}_t^O\right) \bigcup \left(\left(\mathcal{I}_t^O\right)^C \times \mathbb{R}_-\right),$$
$$\Gamma f_t^F = \left(\mathcal{I}_t^F \times \mathcal{Y}_t^F\right) \bigcup \left(\left(\mathcal{I}_t^F\right)^C \times \mathbb{R}_-\right).$$

Since for $s \leq t$ it holds that $\mathcal{I}_s^O \subseteq \mathcal{I}_t^O$ and $\mathcal{I}_s^F \subseteq \mathcal{I}_t^F$, hence if we have at least two distinct values of $\boldsymbol{e}_i^F$, the VC indexes of $\mathcal{F}^O$ and $\mathcal{F}^F$ are therefore equal to 2, otherwise the class has only one function and the desired result follows from the CLT. Let the functions $F^O : \mathcal{X} \to \mathbb{R}$ and $F^F : \mathcal{X} \to \mathbb{R}$ be defined as

$$F^O(X_i) = \left\|\boldsymbol{e}_i^O\right\|_1,$$
$$F^F(X_i) = \left\|\boldsymbol{e}_i^F\right\|_1.$$

These two functions are square integrable envelopes of families $\mathcal{F}^O$ and $\mathcal{F}^F$. Therefore by Theorem 2.6.7 and 2.5.2 from van der Vaart and Wellner (1996) the classes of functions are P-Donsker.

$\square$

Proof of Theorem 2. Observe that

$$\hat{\boldsymbol{e}}_i = \boldsymbol{e}_i + \boldsymbol{X}_i\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right),$$

and because $\hat{\sigma} = \sigma + o_P(1)$ and by Slutsky's lemma

$$\hat{\boldsymbol{e}}_i^O = (1 + o_P(1)) \left(\boldsymbol{e}_i^O + \boldsymbol{P}_i \boldsymbol{X}_i \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) \right). \tag{4}$$

Observe that the function $(\sigma_., \boldsymbol{D}_.) \mapsto \left(\sigma_.^2 \boldsymbol{I} + \boldsymbol{Z} \boldsymbol{D}_. \boldsymbol{Z}^T \right)^{-\frac{1}{2}} = \boldsymbol{V}_.^{-\frac{1}{2}}$ is differentiable for scalars $\sigma_.$ and matrices $\boldsymbol{D}_.$ close to true parameters σ and $\boldsymbol{D}$. Therefore by Lemma 2.12 in van der Vaart (1998)

$$\hat{\boldsymbol{V}}_i = \boldsymbol{V}_i + \boldsymbol{E}_n,$$

where the matrix $\boldsymbol{E}_n$ has $o_P(1)$ functions for its entries. Hence

$$\hat{\boldsymbol{e}}_i^F = (1 + o_P(1)) \left(\boldsymbol{e}_i^F + \boldsymbol{V}_i^{-\frac{1}{2}} \boldsymbol{X}_i \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) \right). \tag{5}$$

The next part of the proof is the same for both processes, so we use the superscript ^{U} to mean either ^{O} or ^{F}. Let $\boldsymbol{M}_i^U$ be equal to $\boldsymbol{P}_i$ if $U = O$ and $\boldsymbol{M}_i^U$ be equal to $\boldsymbol{V}_i^{-\frac{1}{2}}$ if $U = F$. From (4) and (5) it follows that for a fixed $t \in \mathbb{R}$

$$\begin{aligned}
\hat{W}_n^U(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big)^T \hat{\boldsymbol{e}}_i^U \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big)^T \boldsymbol{e}_i^U \, (1 + o_P(1)) \\
+& \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)^T \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big) \, (1 + o_P(1)) \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big)^T \boldsymbol{e}_i^U \\
+& \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{m}(X_i) \right) \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big) \right) + o_P(1).
\end{aligned}$$

The last equality follows by the CLT for the first part and by assumption (A2), WLLN and CLT for the second part.

if $\lambda < \infty$, then every component of $\boldsymbol{y}_i^U \to \boldsymbol{\sigma}_\lambda \big(\hat{\boldsymbol{y}}_i^U, t \big)$ is differentiable, thus

$$\begin{aligned}
\hat{W}_n^U(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\sigma}_\lambda \big(\boldsymbol{y}_i^U, t \big)^T \boldsymbol{e}_i^U \\
+& \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{m}(X_i) \right) \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda \big(\boldsymbol{y}_i^U, t \big) \right) + o_P(1).
\end{aligned}$$

In the case that $\lambda = \infty$, the equality

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\sigma}_\lambda\big(\hat{\boldsymbol{y}}_i^U, t\big)^T\boldsymbol{e}_i^U = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^U, t\big)^T\boldsymbol{e}_i^U + o_P(1),$$

follows from equicontinuity of the families of functions $\mathcal{F}^O$ and $\mathcal{F}^F$, by a classic argument in this area, as found for example in the proof of Lemma B.2 in Hagemann (2017). The second equality, that is

$$(\mathbb{G}_n\,\boldsymbol{m})\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{M}_i^U\boldsymbol{\sigma}_\lambda\big(\hat{\boldsymbol{y}}_i^U, t\big)\right) = (\mathbb{G}_n\,\boldsymbol{m})\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{M}_i^U\boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^U, t\big)\right) + o_P(1),$$

follows by assumption (A3) and Lemma 2.12 in van der Vaart (1998).

From Slutsky's lemma it follows that

$$\hat{W}_n^U(t) = \mathbb{G}_n\, f^U + \mathbb{G}\boldsymbol{m}\,\mathrm{P}\boldsymbol{g}_t^U + o_P(1). \tag{6}$$

Note that the components of the function $t \mapsto \mathrm{P}\boldsymbol{g}_t^U$ are differentiable since either $\lambda < \infty$ or $\lambda = \infty$ and (A3) holds. By assumption (A1) this function is also bounded, therefore the process

$$t \mapsto \mathbb{G}_n\,\boldsymbol{m}\,\mathrm{P}\boldsymbol{g}_t^U,$$

is asymptotically ρ-equicontinuous, where ρ is some semimetric defined by the help of the limiting distribution of $\mathbb{G}_n\,\boldsymbol{m}$ and the previous paragraph. Note that any finite collection of marginals of this process also converges.

Now we use the same reasoning as in Theorem 19.23 in van der Vaart (1998) to conclude that, as a consequence of the continuous mapping theorem, the limit of the process $\hat{W}^U$ is

$$\hat{W}^U \rightsquigarrow \mathbb{G}\, f^U + \mathbb{G}\boldsymbol{m}\,\mathrm{P}\boldsymbol{g}_t^U.$$

□

Proof of Theorem 3. We use the same notation as in the proof of Theorem 2 - that is ^{U} to refer to either ^{O} or ^{F}, since the proofs for conditional convergence of both processes are very similar.

Observe that conditionally on data $X_1, X_2, \dots$

$$
\begin{aligned}
\hat{\boldsymbol{e}}_i^{U*} =& \xi_i \hat{\boldsymbol{e}}_i^U + \frac{1}{n} \hat{\boldsymbol{M}}_i^U \boldsymbol{X}_i \sum_{j=1}^n \boldsymbol{m}(X_j^*) \\
=& \xi_i \boldsymbol{e}_i^U + \frac{1}{n} \boldsymbol{M}_i^U \boldsymbol{X}_i \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j) \\
&+ \xi_i \hat{\boldsymbol{M}}_i^U \boldsymbol{X}_i \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) + \frac{1}{n} \left(\boldsymbol{M}_i^U - \hat{\boldsymbol{M}}_i^U\right) \boldsymbol{X}_i \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j) + o_P\left(n^{-\frac{1}{2}}\right).
\end{aligned} \tag{7}
$$

The above equality follows by the assumption (A2). The process $\hat{W}^U$ is therefore equal to

$$
\begin{aligned}
\hat{W}_n^{U*}(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)^T \boldsymbol{e}_i^U \\
&+ \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j)\right)^T \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)\right) \\
&+ \frac{1}{n} \sum_{i=1}^n \xi_i \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)^T \hat{\boldsymbol{M}}_i^U \boldsymbol{X}_i \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \\
&+ \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j)\right)^T \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \left(\boldsymbol{M}_i^U - \hat{\boldsymbol{M}}_i^U\right) \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)\right) + o_P(1) \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)^T \boldsymbol{e}_i^U \\
&+ \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j)\right)^T \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda\left(\hat{\boldsymbol{y}}_i^U, t\right)\right) + o_P(1).
\end{aligned}
$$

The equality above follows by WLLN. By the same arguments as in the proof of Theorem 2 it therefore follows that

$$
\begin{aligned}
\hat{W}_n^{U*}(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \boldsymbol{\sigma}_\lambda\left(\boldsymbol{y}_i^U, t\right)^T \boldsymbol{e}_i^U \\
&+ \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \xi_j \boldsymbol{m}(X_j)\right)^T \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{M}_i^U \boldsymbol{\sigma}_\lambda\left(\boldsymbol{y}_i^U, t\right)\right) + o_P(1).
\end{aligned} \tag{8}
$$

From the equality above it follows that the marginals of $\hat{W}_n^{U*}$ at times $t_1, \dots, t_r, r < \infty$ converge to the marginals of the process $\hat{W}_n^U$ at these times.

Note that since the i.i.d. random variables ξ_i have zero mean, variance 1 and there exists $r > 2$ so that $E(|\xi|^r) < \infty$ (by assumption (A4)) and by Theorem 1 the family $\mathcal{F}^U$ is Donsker. We can use the Theorem 2.9.6 from van der Vaart and Wellner (1996) to prove that

$$t \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \boldsymbol{\sigma}_\lambda\big(\boldsymbol{y}_i^U, t\big)^T \boldsymbol{e}^U \overset{*}{\rightsquigarrow} t \to \mathbb{G} f_t^U.$$

For the proof that the second part of the process (8) converges as a process, use the same arguments as in the proof of Theorem 2.

At the end use the same arguments as in the proof of Theorem 2 to argue that the process converges.

□

Proof of Theorem 4. The appropriate probability measure will be denoted by P. Remember that in our previous proofs $\mathrm{P} f_t^U = 0$ for every $f \in \mathcal{F}^U$. Note that this now still holds only for $\mathcal{F}^F$ and only under the hypothesis (H_2). In every other case define the following process

$$\begin{aligned} \hat{W}_n^U(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\boldsymbol{\sigma}_\lambda\big(\hat{\boldsymbol{y}}_i^U, t\big)^T \hat{\boldsymbol{e}}_i^U - \mathrm{P} f_t^U\right), \\ \hat{W}_n^{U*}(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left(\boldsymbol{\sigma}_\lambda\big(\hat{\boldsymbol{y}}_i^U, t\big)^T \hat{\boldsymbol{e}}_i^U - \mathrm{P} f_t^U\right) \\ & + \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \boldsymbol{m}(X_i)\right)^T \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i)^T \boldsymbol{M}_i \boldsymbol{\sigma}_\lambda\big(\hat{\boldsymbol{y}}_i^U, t\big)\right). \end{aligned} \tag{9}$$

Note that we can use the same arguments as in Theorem 2 and Theorem 3 to prove that the processes $\hat{W}_n^U$ and $\hat{W}_n^{U*}$, the later conditionally on data $\{X_i\}_{i=\mathbb{N}}$, converge weakly to the same limiting process.

Note also that conditionally on data $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \mathrm{P} f_t^U$ is zero. On the other hand, the expectation of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{P} f_t^U$ does not converge under any of (H_1), (H_2) and (H_3), except in the case of (H_2) and $U = F$. If we subtract these quantities from the processes in (9) and then multiply them by $\frac{1}{\sqrt{n}}$ it therefore follows that the first converges towards $\mathrm{P} f_t^U$ in probability and the second converges towards zero in probability.

$\square$

2 Potential extension to multilevel and/or crossed random effects

Assume that the entire vector of outcomes $\boldsymbol{y} \in \mathbb{R}^N$ is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{l=1}^{L} \boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)} \boldsymbol{b}^{(l)} + \boldsymbol{\epsilon}, \tag{10}$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times m}$ and $\boldsymbol{\beta} \in \mathbb{R}^m$ are as before a matrix of fixed effects and an unknown parameter, respectively, $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is a vector of i.i.d. random variables with mean zero and variance σ^2, which is as before an unknown parameter. The matrix $\boldsymbol{\Pi}^{(l)} \in \mathbb{R}^{N \times N}$ is a known permutation matrix, which can permute the random effects and therefore allows us to have many different kinds of structures of random effects, $\boldsymbol{Z}^{(l)} \in \mathbb{R}^{N \times \left(n^{(l)} k^{(l)}\right)}$ is known block diagonal matrix of random effects and the vector $\boldsymbol{b}^{(l)} \in \mathbb{R}^{\left(n^{(l)} k^{(l)}\right)}$ is a vector of vertically concatenated $n^{(l)}$ i.i.d. random vectors of length $k^{(l)}$ with mean $\mathbf{0}$ and variance $\boldsymbol{D}^{(l)}$. With this notation the model (1.1) can be written in the above way by setting

$L = 1, n^{(1)} = n, k^{(1)} = k, \mathbf{\Pi}^{(1)} = I, \boldsymbol{D}^{(1)} = \boldsymbol{D}$ and defining the following matrices

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{pmatrix},$$

$$\boldsymbol{Z}^{(1)} = \begin{pmatrix} \boldsymbol{Z}_1 & & \\ & \ddots & \\ & & \boldsymbol{Z}_n \end{pmatrix},$$

$$\boldsymbol{b}^{(1)} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_n \end{pmatrix}.$$

Observe that if we attempted to do a straightforward generalization of the theory presented in Section 6 and Section 7, we would have some issues. The biggest issue would be the fact that we assumed that the random elements X_i are independent. This assumption is crucial, because the proofs in Supplementary section 1 are all based on the theory from van der Vaart and Wellner (1996), which is based on the independence of random elements.

One way in which we can generalize our theory, solve the above problem and still consider the theory that is also applicable in practice is the following. As before, define first the following random matrices and vectors, for $l = 1, \ldots, L$.

$$
\begin{aligned}
\boldsymbol{V} =& \sigma^2 \boldsymbol{I} + \sum_{l=1}^{L} \boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)} \boldsymbol{D}^{(l)} \left(\boldsymbol{\Pi}^{(l)} Z^{(l)}\right)^T, \\
\boldsymbol{P}^{(l)} =& \boldsymbol{I} - \boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)} \left(\left(\boldsymbol{Z}^{(l)}\right)^T \boldsymbol{Z}^{(l)}\right)^+ \left(\boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)}\right)^T, \\
\boldsymbol{P} =& \frac{1}{\sigma} \prod_{l=1}^{L} \boldsymbol{P}^{(l)}, \\
\boldsymbol{y}^F =& \boldsymbol{X}\boldsymbol{\beta}, \\
\boldsymbol{e} =& \boldsymbol{y} - \boldsymbol{y}^F, \\
\boldsymbol{e}^F =& \boldsymbol{V}^{-\frac{1}{2}} \boldsymbol{e}, \\
\boldsymbol{e}^O =& \boldsymbol{P}\boldsymbol{e}, \\
\tilde{\boldsymbol{b}}^{(l)} =& \left(\boldsymbol{I} \otimes \boldsymbol{D}^{(l)}\right) \left(\boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)}\right)^T \left(\sigma^2 \boldsymbol{I} + \boldsymbol{\Pi}^{(l)} \boldsymbol{Z}^{(l)} \boldsymbol{D}^{(l)} \left(\boldsymbol{\Pi}^{(l)} Z^{(l)}\right)^T\right)^{-1} \boldsymbol{e}^{O(l)}, \\
\boldsymbol{e}^{O(l)} =& \prod_{k \neq l} \boldsymbol{P}^{(k)} \boldsymbol{e}, \\
\boldsymbol{y}^{O(l)} =& \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{b}}^{(l)}.
\end{aligned}
$$

As before, we use the $\hat{\cdot}$ to denote that the parameters in an expression are replaced with their estimators. Now define the following stochastic processes for $l = 1, \ldots, L$

$$
\begin{aligned}
\hat{W}_n^{F1} =& \frac{1}{\sqrt{n}} \boldsymbol{\sigma}_\lambda \left(\boldsymbol{y}^F, t\right)^T \boldsymbol{e}^F, \\
\hat{W}_n^{F2} =& \frac{1}{\sqrt{n}} \boldsymbol{\sigma}_\lambda \left(\boldsymbol{y}^F, t\right)^T \boldsymbol{e}^O, \\
\hat{W}_n^{O(l)} =& \frac{1}{\sqrt{n}} \boldsymbol{\sigma}_\lambda \left(\boldsymbol{y}^{O(l)}, t\right)^T \boldsymbol{e}^O.
\end{aligned}
$$

Because the joint normality and absence of correlation imply independence, we can use the process $\hat{W}_n^{F1}$ for checking the hypothesis $\boldsymbol{H}_0^F$ if we can assume that the random vectors in $\boldsymbol{b}^{(l)}, l = 1, \ldots, L$ are distributed as multivariate normal random vectors. Without the normality assumption we can also use the process $\hat{W}_n^{F1}$ to check the hypothesis $\boldsymbol{H}_0^F$ if the random effects are nested. Otherwise we can use the process $\hat{W}_n^{F2}$ to check the null hypothesis $\boldsymbol{H}_0^F$. The processes $\boldsymbol{W}_n^{O(l)}, l = 1, \ldots, L$ can be used to check the null hypothesis $\boldsymbol{H}_0^O$.

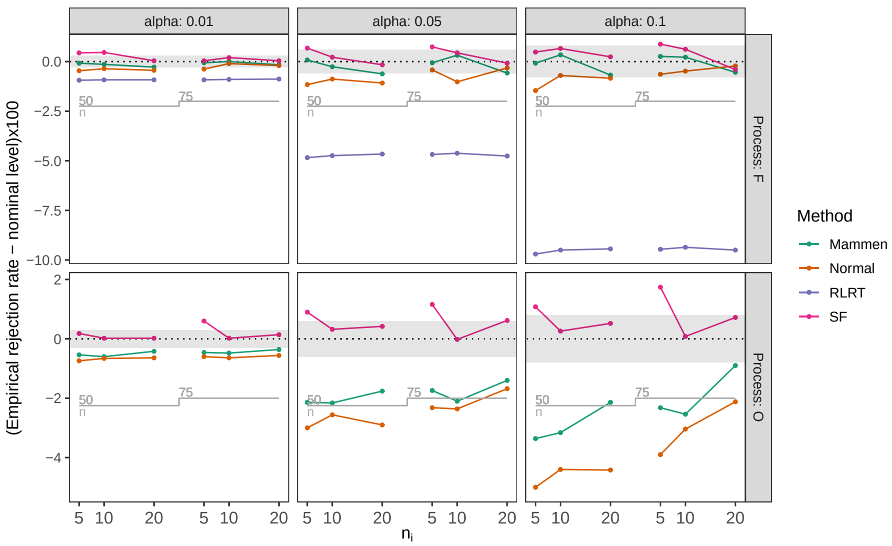
For proving the theory in a similar way as before, we have to rewrite these processes as a sum of two parts, as in Equation (6). The first part is equal to $\mathbb{G}_n f_t$, where $\mathbb{G}_n$ is defined as before, but may be based on a different sequence of random elements $\{X_i\}_{i\in\mathbb{N}}$. The second term is a consequence of estimating the parameter $\boldsymbol{\beta}$. Depending on how matrix $\boldsymbol{X}$ is generated we can then determine the appropriate random elements X_i, so that the first term converges. Note that the first term in the sum is then already based on the independent random variables. For the convergence of the second term and the convergence of the sum of these two terms we then need to make some additional assumptions. These additional assumptions might require the use of the theory for empirical stochastic processes based on the independent but not identically distributed random elements.

3 Additional simulation results

3.1 Size for relaxed normality assumption in error and random effect terms

The outcome was simulated from Eq. (12) in the main file specifying ϵ_{ij}, $b_{i,0}$ and $b_{i,1}$ independently from a centered gamma distribution with parameters shape and scale set to 1 and 2, respectively, and $\beta_3 = 0$. The fitted model was the same as the simulated model. The simulations were done for $n = 50, 75$ and $n_i = 5, 10, 20$.

Figure 3: Size under non-normal errors and random effects for different α (columns) and processes (rows) using the CvM test statistics; λ was chosen using the data-driven approach. Each figure shows *difference between the empirical rejection rate and the nominal level* (y axis) for different number of observations inside groups (x axis) and different number of groups (left part $n = 50$, right part $n = 75$). Colors specify different distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors. RLRT - restricted likelihood ratio test of Greven et al. (2008).



The proposed test for both processes and all distributions performed very similarly as in Example I in the main paper, showing robustness against non-normality (Figure 5). RLRT test was again too conservative.

3.2 Simulation results for the alternative definition of the function p

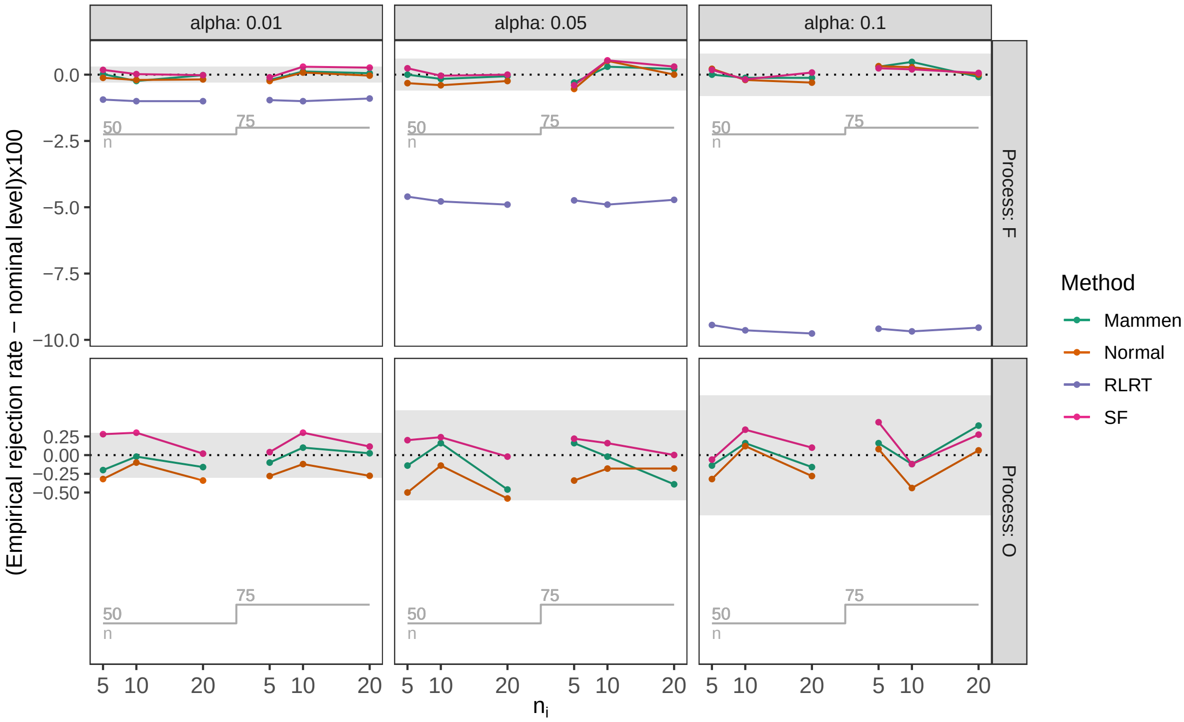
Here we report additional simulation results, using the following definition of the function p,

$$p(x) = \begin{cases} 8x^3\left(3x-2\right)+1, & x < \frac{1}{2}, \\ -8\left(x-1\right)^3\left(3x-1\right), & x \geq \frac{1}{2}. \end{cases} \tag{11}$$

considering different values of $\lambda = \{0.1, 0.5, 2, \infty\}$. The simulation setup was as described in the main text.

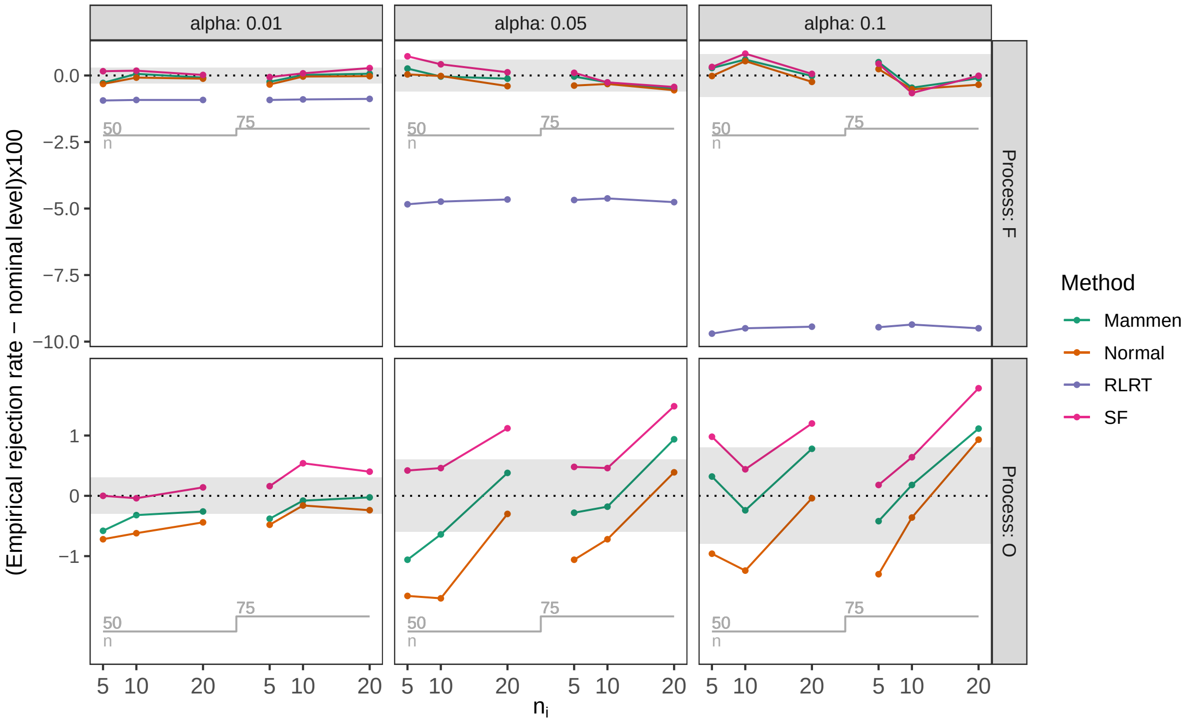
3.2.1 Example I - size under normal errors and normal random effects

Figure 4: Size under normal errors and normal random effects for different α (columns) and processes (rows) using the CvM test statistics; $\lambda = 2$. Each figure shows *difference between the empirical rejection rate and the nominal level* (y axis) for different number of observations inside clusters (x axis) and different number of clusters (left part $n = 50$, right part $n = 75$). Colors specify different distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors. RLRT - restricted likelihood ratio test of Greven et al. (2008).



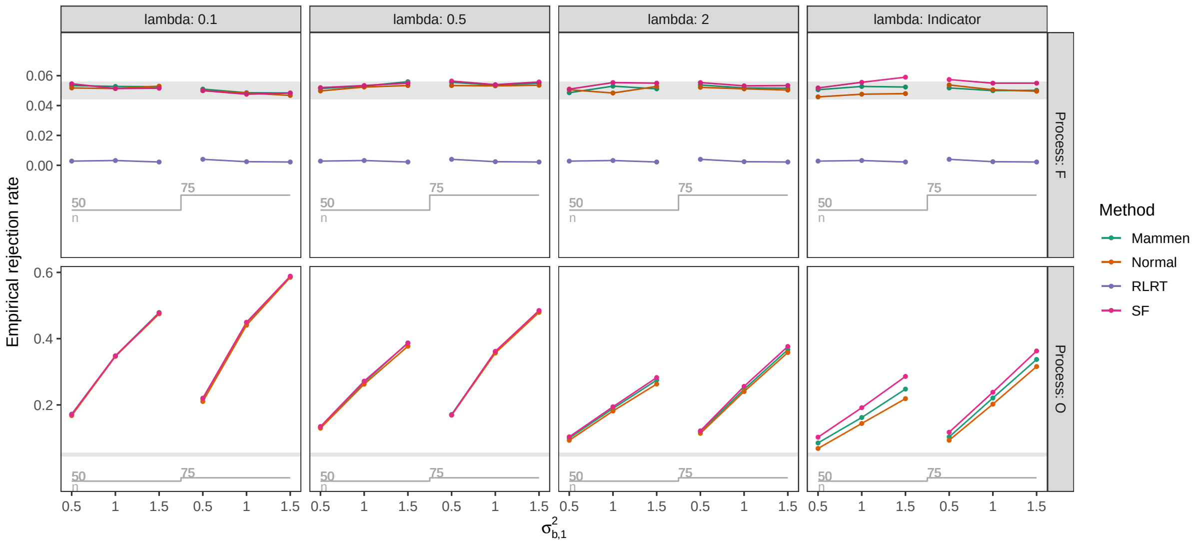
3.2.2 Example Ia - size for relaxed normality assumption in error and random effect terms

Figure 5: Size under non-normal errors and random effects for different α (columns) and processes (rows) using the CvM test statistics; $\lambda = 2$. Each figure shows *difference between the empirical rejection rate and the nominal level* (y axis) for different number of observations inside clusters (x axis) and different number of clusters (left part $n = 50$, right part $n = 75$). Colors specify different distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors. RLRT - restricted likelihood ratio test of Greven et al. (2008).



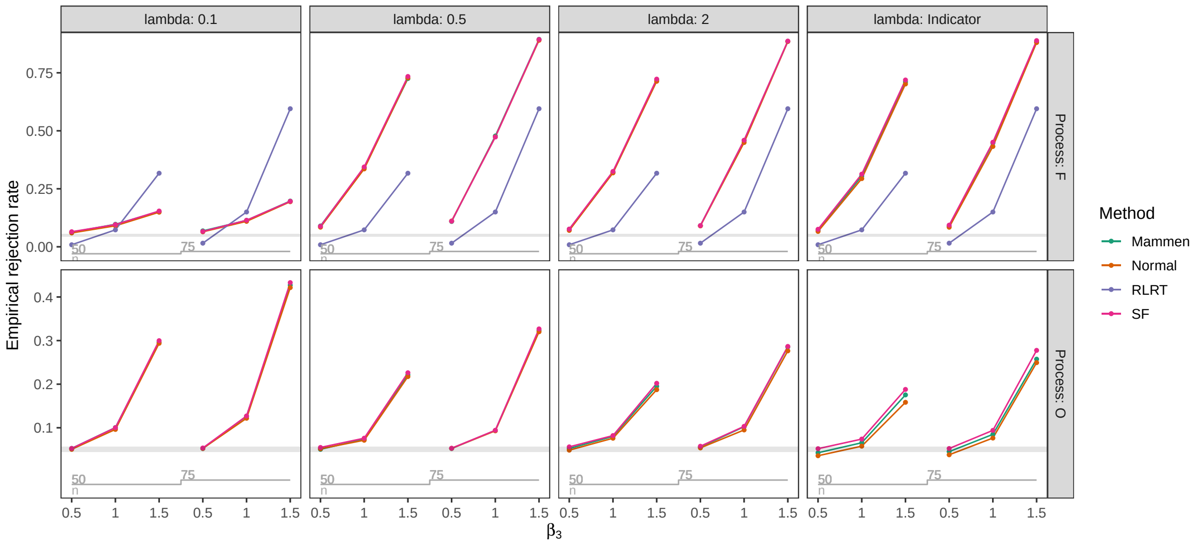
3.2.3 Example II - mis-specified random effects design matrix

Figure 6: Power under mis-specified random effects design matrix for different λ (columns) and processes (rows) using the CvM test statistics; $\alpha = 0.05$. Each figure shows *empirical rejection rate* (y axis) for different variability of the missed random effect $b_{i,1}$ (x axis) and different number of clusters (left part $n = 50$, right part $n = 75$). Colors specify different proposed distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors under the null. RLRT - restricted likelihood ratio test of Greven et al. (2008). For normal distribution and $\lambda = \infty$ the results for the F-process are equivalent to using the approach of Pan and Lin (2005).



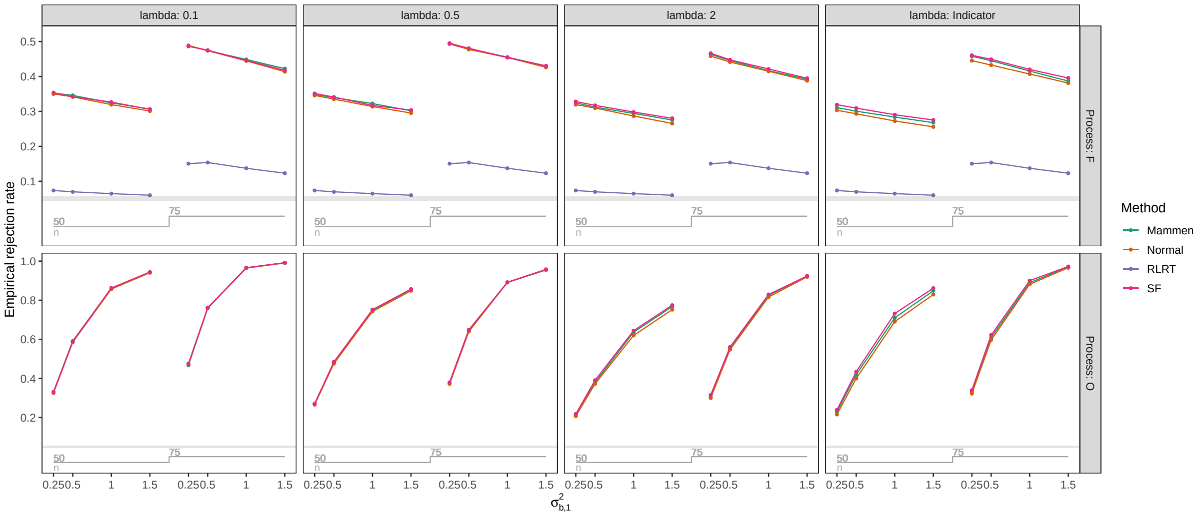
3.2.4 Example III - mis-specified fixed effects design matrix

Figure 7: Power under mis-specified fixed effects design matrix for different λ (columns) and processes (rows) using the CvM test statistics; $\alpha = 0.05$. Each figure shows *empirical rejection rate* (y axis) for different fixed effect β_3 (x axis) and different number of clusters (left part $n = 50$, right part $n = 75$). Colors specify different proposed distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors under the null. RLRT - restricted likelihood ratio test of Greven et al. (2008). For normal distribution and $\lambda = \infty$ the results for the F-process are equivalent to using the approach of Pan and Lin (2005).



3.2.5 Example IV - mis-specified fixed and random effects design matrices

Figure 8: Power under mis-specified fixed and random effects design matrices for different λ (columns) and processes (rows) using the CvM test statistics; $\alpha = 0.05$. Each figure shows *empirical rejection rate* (y axis) for different variability of the missed random effect $b_{i,1}$ (x axis) and different number of clusters (left part $n = 50$, right part $n = 75$). Colors specify different proposed distributions used in the proposed bootstrap approach. The shaded areas are simulation margins of errors under the null. RLRT - restricted likelihood ratio test of Greven et al. (2008). For normal distribution and $\lambda = \infty$ the results for the F-process are equivalent to using the approach of Pan and Lin (2005).



3.3 An example where smoothing improves the power for the *F-process*

The outcome was simulated from a linear mixed model with random intercepts

$$y_{ij} = -1 + 2X_{ij} + 0.75X_{ij}^4 + b_{i,0} + \epsilon_{ij}, \quad (12)$$

where $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$, but the fitted model assumed a linear effect of X. All quantities on the right side of (12) were simulated independently from each other: $X_{ij} \sim U(0,1)$, $b_{i,0} \sim N(0, 0.25)$, and $\epsilon_{ij} \sim N(0, 0.5)$. The number of groups was set to $n = 50$ and for each group we set $n_i = 10$. We use the same function p as in the main document, with $\lambda = \{0.1, 0.5, 1, 2, 3, 5, 7, 10, 50, \infty\}$, using also the data-driven approach to select the smoothing parameter. The p-values were estimated by using $M = 500$ random realizations of null approximations simulating ξ_i, $i = 1, \ldots, n$, from a Rademacher distribution. The p-values were simulated 500 times. Results for the percentage of rejected null hypotheses at $\alpha = 0.1$ are shown in Table 1. In contrast to the examples provided in the main document, smoothing can improve the power also for the *F-process*.

Table 1: Power (%) under a mis-specified random effects design matrix for different λ (columns) and processes (rows) using the CvM test statistics; $\alpha = 0.05$; opt - the λ chosen using the data-driven approach.

λ	opt	0.1	0.5	1	2	3	5	7	10	50	∞
F-process	73	76	73	69	73	73	73	72	72	72	72
O-process	35	40	36	30	35	34	32	32	33	32	32

4 Application

4.1 ANES cross-sectional example

The definitions of the covariates in the data analysis section are as in Peng and Lu (2012):

1. age: respondent's age.
2. gender: respondent's gender. 1 =`male`, 0 =`female`.
3. educ: respondent's education. 1 =`high school and higher`, 0 =`less than high school`.
4. christian: respondent's religious denomination. 1 =`Christian/Catholic`, 0 =`other`.
5. black: dummy variable for respondent's race where `white` is the reference group. 1 =`black`, 0 =`not`.
6. other: dummy variable for respondent's race. 1 =`other race`, 0 =`not`.
7. liberal: dummy variable for respondent's ideology where `conservative` is the reference group. 1 =`liberal`, 0 =`not`.
8. defence: the importance of defense spending issues to the respondent. 1 =`very important`, 0 =`not very important`.
9. death: respondent's attitude toward death penalty. 1 =`favor`, 0 =`oppose`.
10. democrat: dummy variable for respondent's party identification where `Republican` is the reference group. 1 =`Democrat`, 0 =`not`.
11. indep: dummy variable for respondent's party identification. 1 =`Independent`, 0 =`not`.
12. Iraq: Does the respondent approve the way President Bush handles Iraq war? 1 =`approve`, 0 =`disapprove`.

Table 2: Parameter estimates of the fixed effect coefficients for the fine-tuned model and the starting model on our data set (middle three columns) and Peng-Lu data set. There are slight differences in estimates due to difference in the version of original data set.

nlme estimates	final model			starting model			Peng and Lu (2012)		
Fixed effects (β)	Coefficient	Std.Error	p-value	Coefficient	Std.Error	p-value	Coefficient	Std.Error	p-value
(Intercept)	57.02	4.03	0.00	47.67	3.44	0.00	46.62	3.34	0.00
age	0.09	0.04	0.01	0.08	0.04	0.04	0.09	0.04	0.02
educ	-4.51	1.29	0.00	-5.41	1.32	0.00	-5.23	1.33	0.00
christian	8.67	1.94	0.00	7.97	1.77	0.00	7.62	1.76	0.00
black	12.59	7.99	0.12	-1.67	1.97	0.40	-2.52	2.01	0.21
other	5.24	1.97	0.01	4.81	2.00	0.02	4.59	1.97	0.02
liberal	-11.33	1.70	0.00	-10.16	1.48	0.00	-11.68	1.75	0.00
defence	-1.97	1.60	0.22	1.69	1.30	0.19	1.95	1.31	0.14
death	4.18	1.38	0.00	4.93	1.42	0.00	4.62	1.43	0.00
democrat	-34.22	3.21	0.00	-24.47	2.12	0.00	-25.05	2.06	0.00
indep	-22.48	3.10	0.00	-13.62	1.75	0.00	-14.53	1.71	0.00
Iraq	13.69	3.52	0.00	30.37	1.60	0.00	30.62	1.58	0.00
christian:black	-20.86	8.91	0.02						
defence:Iraq	8.71	2.53	0.00						
liberal:Iraq	5.89	3.18	0.06						
black:democrat	9.37	3.63	0.00						
indep:Iraq	14.11	3.63	0.00						
democrat:Iraq	16.25	4.70	0.00						
Model variance (σ^2)	408.18			451.87			442.26		

4.2 AIDS longitudinal example

As an additional example we apply the proposed methodology to a subset of data from the Multicenter AIDS Cohort Study (see Kaslow et al. (1987) for more details on the data). The data have been used by Andriyana et al. (2014) to illustrate P-splines quantile regression estimation in varying coefficient models for longitudinal data. We show that similar results as presented in Andriyana et al. (2014) for the median regression could be obtained by using appropriately specified LMM (the CD4 cell percentage for this data is fairly symmetrical with mean and median being very similar, 29.2 and 29, respectively) where our approach can be used to correctly determine the functional form of the two design matrices of the fitted LMM. The data are available in the R package **QRegVCM** (Andriyana et al., 2018).

The data contain repeated measurements of physical examinations and CD4 percentages of 283 homosexual men (in total 1817 measurements) who became HIV-positive between year 1984 and 1991. Let $CD4_{ij}$ denote the CD4 percentage for the jth measurement of individual i. Let T_{ij} denote the time at which the jth measurement for individual i was taken. With A_i, S_i and P_i we denote the age of subject i at HIV infection, his smoking status (1 for smoking and zero otherwise) and pre-HIV infection CD4 cell percentage, respectively.

We first fit the following model to the data

$$CD4_{ij} = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 T_{ij} + \beta_5 P_i + b_i, \qquad \text{(Model 1)}$$

We plot in Figure 9 the proposed empirical stochastic processes and in gray the corresponding $M = 500$ generated null processes according to *SF* approach, using the function p as defined in (8.1), where the smoothing parameter was chosen from the grid $\{0.5, 2, 5, 10, 1000\}$ using the data-driven approach. The processes are shown as step functions to point out that σ_λ was, for computational reasons, evaluated only at the distinct fitted values. The processes follow (from left to right): $\hat{W}_n^{F^s}(t)$, the *subset F-process*, $\hat{W}_n^F(t)$, the *F-process*, and $\hat{W}_n^O(t)$, the *O-process*. Several different *subset F-process* were considered: the subsets contained the estimated coefficients associated with variable T_{ij}, A_i and P_i yielding 3 *subset F-process* which were considered for Model 1. Of these, only the first *subset F-process* is shown in Figure 9.

The S shape observed for the *subset F-process* for the subset of the coefficients associated with variable T_{ij} ($p = 0.002$, Figure 9, row 1, column 1) is caused by a sequence of (mainly) positive (ordered) residuals, followed by a sequence of (mainly) negative residuals and finally a sequence of positive residuals, which implies that a quadratic term is missing (see also Lin et al. (2002)). The p-values based on the other two *subset F-process* were however not significant ($p = 0.88$ and 0.44 for the *subset F-process* including A_i and P_i, respectively). Therefore, we fit the following model to the data

$$CD4_{ij} = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 T_{ij} + \beta_5 T_{ij}^2 + \beta_6 P_i + b_i. \qquad \text{(Model 2)}$$

For Model 2 we considered 4 different *subset F-processes*: subset containing all estimated coefficients associated with variable T_{ij} (also the one with the term that suggests curvature in the variable), subset containing the estimated coefficient associated with variable A_i, subset containing the estimated coefficient associated with variable P_i and subset including the estimated coefficients for P_i and all estimated coefficients associated with variable T_{ij}. Only the later process is shown in Figure 9. The p-values for the first three *subset F-processes* were not significant, $p = 0.50$, 0.90 and 0.28, respectively. However, the p-value for the *F-process* was still significant at the 5% level ($p = 0.046$, Figure 9, row 2, column 2). When the p-values for the *subset F-processes* including the coefficient(s) of a single variable (including those that suggest curvature in the variable) are not significant, but the p-value for the *F-process* is significant, this implies that an interaction effect between the variables is missing from the model. This is confirmed by looking at the *subset F-process* which includes the estimated coefficients for P_i as well as all estimated coefficients associated with variable T_{ij}, where the p-value is significant at the 5% level ($p = 0.024$, Figure 9, row 2, column 1), which implies that the interaction between pre-HIV infection CD4 cell percentage and time is missing from the model. Therefore we fit the following model

$$CD4_{ij} = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 T_{ij} + \beta_5 T_{ij}^2 + \beta_6 P_i + \beta_7 T_{ij} P_i + \beta_8 T_{ij}^2 P_i + b_i. \qquad \text{(Model 3)}$$

Here we consider 2 *subset F-processes*: subset containing the estimated coefficient associated with variable A_i, subset containing the estimated coefficient associated with variable P_i and subset including the estimated coefficients for P_i and all estimated coefficients associated with variable T_{ij}. Inclusion of the interaction terms substantially improves the model's fit, with the p-values for both *subset F-processes* and for the *F-process* being larger than 0.05 ($p = 0.85$ for the subset including A_i, $p = 0.14$ for the subset including P_i and all the coefficients associated with T_{ij} (Figure 9, row 3, column 1) and $p = 0.28$ for the *F-process* (Figure 9, row 3, column 2)). The p-value obtained from the *O-process* is, however, still significant at the 0.05 level ($p = 0.002$, Figure 9, row 3, column 3), suggesting that the random effects design matrix is mis-specified. Therefore we include in the model also a random slope for T_{ij} and fit the following model

$$CD4_{ij} = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 T_{ij} + \beta_5 T_{ij}^2 + \beta_6 P_i + \beta_7 T_{ij} P_i + \beta_8 T_{ij}^2 P_i + b_{i1} + b_{i2} T_{ij}. \qquad \text{(Model 4)}$$

Inclusion of the random slope for T_{ij} substantially improves the model's fit (Figure 9, row 4, column 3) and all the p-values are larger than 0.05 ($p = 0.13$, 0.19, 0.20, and 0.56 for the 2 *subset F-processes* (subset including A_i and subset including P_i and all the coefficients associated with T_{ij} (the later is shown in Figure 9, row 4, column 1), the *F-process* (Figure 9, row 4, column 2) and the *O-process* (Figure 9, row 4, column 3), respectively).

Finally, in Figure 10 we show the association between the CD4 percentages and time (left panel, where the results are for smokers where the values of age and pre-HIV infection CD4 cell percentage were set to their respective medians) and the effect of pre-HIV infection CD4 cell percentage on CD4 percentages in time (right panel) with corresponding 95% confidence intervals as estimated by the Model 5. The results in general agree with those presented by Andriyana et al. (2014) for the median regression: the CD4 cell percentage for a 32.6-year-old smoker with pre-infection CD4 of 42.3 decreases in time, moreso earlier after the infection with HIV. The patients with larger pre-HIV infection CD4 cell percentage had larger CD4 cell percentage, but this effect decreased rapidly in time and after about 4 years after HIV infection the effect of pre-HIV infection CD4 cell percentage was no longer significant as judged by the estimated 95% confidence intervals.

Figure 9: Processes $\hat{W}_n^{F^s}(t)$ (left), $\hat{W}_n^F(t)$ (center) and $\hat{W}_n^O(t)$ (right) for different models (rows 1 to 4 correspond to Models 1 to 4, respectively), for the CD4 dataset.

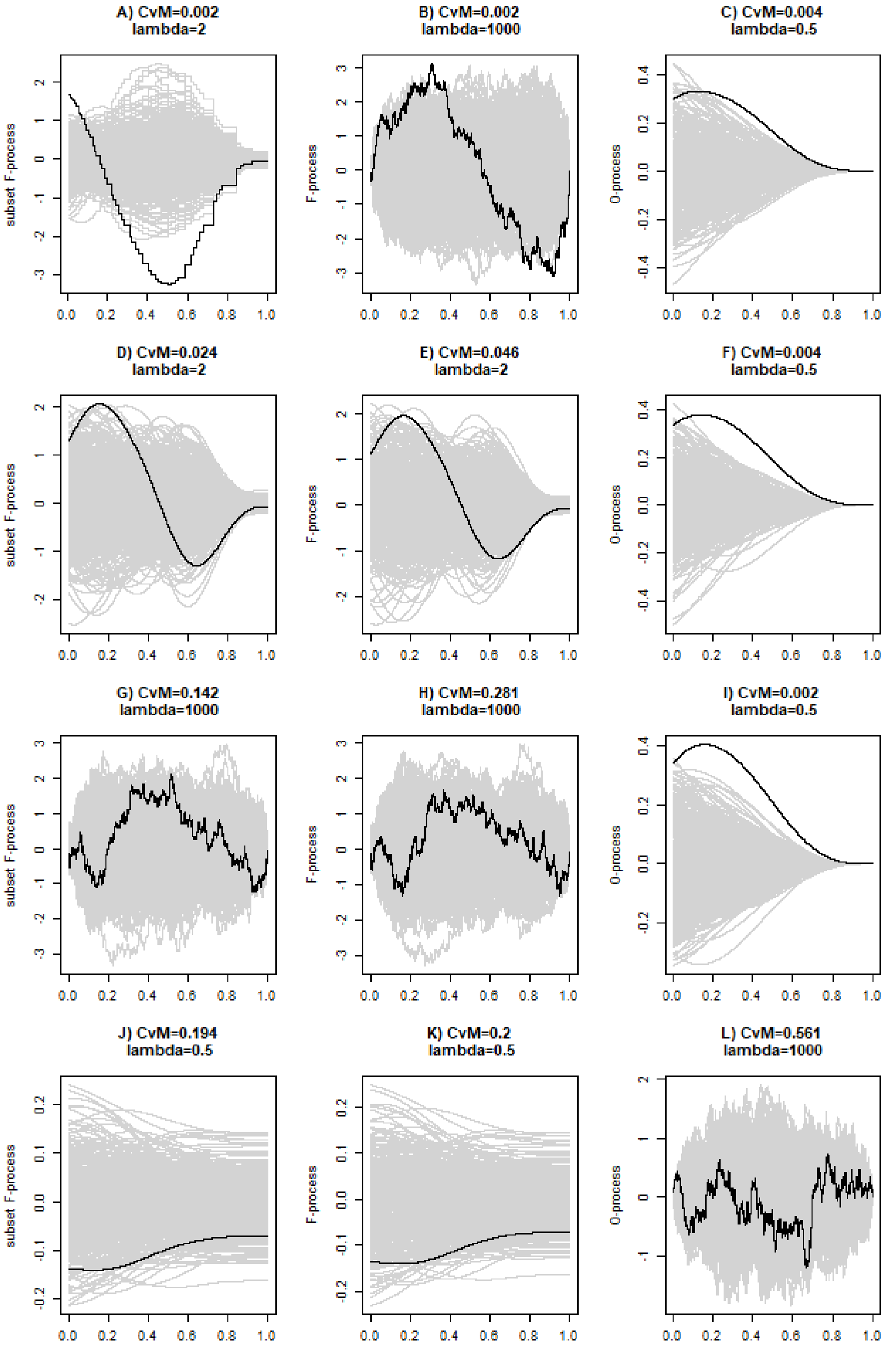
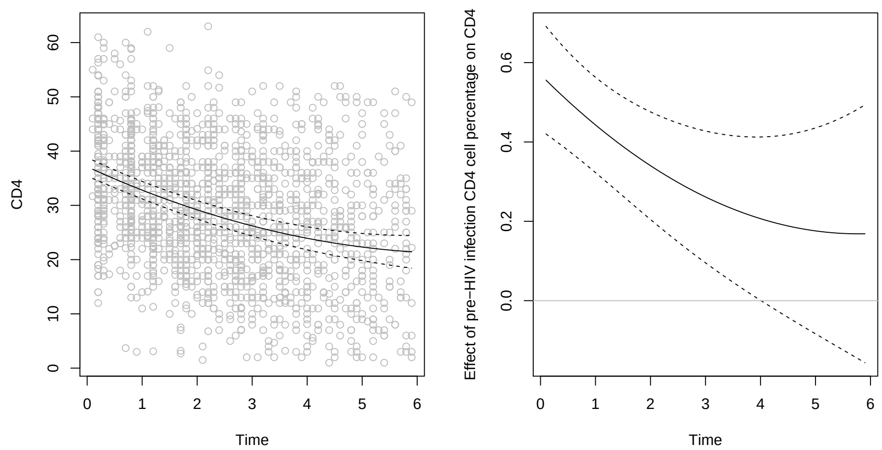


Figure 10: The association between the CD4 percentages and time - left panel (the results are for smokers where the values of age and pre-HIV infection CD4 cell percentage were set to their respective medians) and the effect of pre-HIV infection CD4 cell percentage on CD4 percentages in time - right panel as estimated by Model 5. The dashed lines are the corresponding 95% confidence intervals.



References

Andriyana Y, Gijbels I, Verhasselt A (2014) P-splines quantile regression estimation in varying coefficient models. Test 23(1):153–194

Andriyana Y, Ibrahim MA, Gijbels I, Verhasselt A (2018) QRegVCM: Quantile Regression in Varying-Coefficient Models. URL https://CRAN.R-project.org/package=QRegVCM, r package version 1.2

Greven S, Crainiceanu CM, Küchenhoff H, Peters A (2008) Restricted likelihood ratio testing for zero variance components in linear mixed models. Journal of Computational and Graphical Statistics 17(4):870–891, URL http://www.jstor.org/stable/25651233

Hagemann A (2017) Cluster-robust bootstrap inference in quantile regression models. Journal of the American Statistical Association 112(517):446–456, DOI 10.1080/01621459.2016.1148610, URL https://doi.org/10.1080/01621459.2016.1148610, https://doi.org/10.1080/01621459.2016.1148610

Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo J Charles R (1987) The multicenter AIDS cohort study: rationale, organization, and selection characteristics of the participants. American Journal of Epidemiology 126(2):310–318, DOI 10.1093/aje/126.2.310, URL https://doi.org/10.1093/aje/126.2.310, https://academic.oup.com/aje/article-pdf/126/2/310/197531/126-2-310.pdf

Lin DY, Wei LJ, Ying Z (2002) Model-checking techniques based on cumulative residuals. Biometrics 58(1):1–12, DOI 10.1111/j.0006-341X.2002.00001.x, URL http://dx.doi.org/10.1111/j.0006-341X.2002.00001.x

Pan Z, Lin DY (2005) Goodness-of-fit methods for generalized linear mixed models. Biometrics 61(4):1000–1009, DOI 10.1111/j.1541-0420.2005.00365.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00365.x

Peng H, Lu Y (2012) Model selection in linear mixed effect models. Journal of Multivariate Analysis 109:109–129, DOI https://doi.org/10.1016/j.jmva.2012.02.005, URL https://www.sciencedirect.com/science/article/pii/S0047259X12000395

van der Vaart AW (1998) Asymptotic Statistics. Cambridge University Press, DOI 10.1017/cbo9780511802256, URL https://doi.org/10.1017/cbo9780511802256

van der Vaart AW, Wellner JA (1996) Weak Convergence and Empirical Processes. Springer New York, DOI 10.1007/978-1-4757-2545-2, URL https://doi.org/10.1007/978-1-4757-2545-2